Structural and Biochemical Investigation of Protein-RNA Interactions

Daniel T. Peters

PhD

University of York

Biology

April 2014

Abstract

Non-coding RNAs (ncRNAs) are nucleic acids that do not code for protein. Rather, they have evolved highly specialised secondary structures and catalytic mechanisms that place them at the heart of regulating gene expression. The function of ncRNAs is often mediated or dependent on their interactions with RNA binding proteins. The study of both the structure and function of these proteins is crucial for understanding the biological role of the protein-RNA complexes.

In this thesis, the structure and function of two RNA binding proteins: Lin28 and dihydrouridine synthase C (DusC) were investigated using X-ray crystallography and biophysical techniques. In both systems, the specific recognition of target molecules is important for function. The aim of the study was therefore to use structural and functional data to elucidate the molecular basis of these protein-RNA interactions. There are three main findings: (1) specific recognition of microRNAs by Lin28 is dependent on the interaction of the Zinc Knuckle domain of the protein with a 3' GGAG motif; (2) non-specific, electrostatic interactions between the cold-shock domain of Lin28 and RNA suggest a transcriptome scanning mechanism for recognising Lin28 targets; and (3) modification of specific nucleotide positions within tRNA by DusC is dependent on the orientation in which the tRNA is bound, which is determined by minor changes in the protein structure.

These findings have helped to elucidate the mechanisms, and hence biological functions, of these RNA binding proteins. Both proteins have been previously associated with cancer. Through greater understanding of the molecular basis of these protein-RNA interactions, the production of novel therapeutic agents can be informed, which can help to combat disease. This data will therefore aid future efforts to treat and prevent the cancers caused by the aberrant actions of these RNA binding proteins.

Table of Contents

List of Figures

Acknowledgements

I'd like to thank my supervisors, Fred and Mark, for all their help, support, and advice during my PhD project. I'd like to thank Huw for teaching me the basics of how to solve crystal structures, and Sandra, whose advice on the interpretation of binding data was invaluable for the completion of this project. In addition, I'd like to thank Oleg, James, Rob and Fiona, whose work on Lin28 and DusC provided a solid foundation on which to base my research. Thanks also to Andrew, Maria, Sally and Simon for providing and maintaining the lab space and equipment. I'd like to thank the members of the Antson and Coles groups, as well as the other PhD students in the department for their support, and for providing such a pleasant working environment in which to do my work. Last, but not least, I'd like to thank my parents and my sister Rosie for their constant support during times which have not always been stress-free. Without all of the people mentioned above I could not have completed this project, and you have my gratitude for all the help you have given me over the past few years.

Declaration

I declare that this thesis is a presentation of original work and that I am the sole author. All sources are acknowledged as references. Copyright permission has been obtained for all figures reproduced or adapted from existing publications. Parts of Chapter 3, Chapter 4 and Chapter 5 are being written into manuscripts for publication, and are undergoing refinement before submission. This work has not previously been presented for an award at this, or any other, University.

Chapter 1 : Introduction

RNA is a highly versatile biomolecule, well adapted for its roles in the cell. It is composed of a sequence of nucleotides, each formed from a nucleobase (canonically A, U, C or G), a ribose sugar and a phosphate. Whilst chemically simplistic compared to the diversity of the 20 standard amino acid side-chains, these components allow RNA to fulfil several biological roles more efficiently than their protein counterparts.

The presence of bases that can form complementary hydrogen bonds to the bases of both RNA and DNA molecules means that RNA can easily recognise particular sequences in a specific manner. What sets RNA apart from DNA, however, is the 2' OH group of the ribose sugar, which is absent in deoxyribose. Whilst this group makes RNA less stable than DNA, it greatly expands its catalytic repertoire. In addition, in contrast to DNA, which forms the stable double helix structure, RNA is often single stranded, and folds into complex tertiary structures. These structures can influence the catalytic properties of an RNA molecule and enable it to fulfil a variety of roles in the cell.

Messenger RNA, or mRNA, is one form of RNA. mRNA is transcribed from DNA and its sequence of bases encodes a polypeptide chain. This is known as coding RNA, and contrasts from non-coding RNA (ncRNA) that is transcribed, but not translated into a polypeptide. The ability of ncRNAs to recognise specific mRNA and DNA molecules, combined with their diverse range of structures and the ability to perform catalytic functions, means they are ideal for both regulating and effecting the processes of transcription and translation in the cell.

1.1.1 ncRNAs as regulators: miRNAs

Discovered in 1993 in *C. elegans*,[1] microRNAs (miRNAs) are perhaps the best known of all the non-coding RNAs. They are short ~22nt segments of RNA that post-transcriptionally repress the expression of their target genes. The genomic origins of miRNA are diverse, with some being found at distinct genomic loci, some present in clusters and others found in the introns of mRNA transcripts [2].

1.1.1.1 miRNA biogenesis

The biogenesis of miRNAs is a multi-step process (**Fig. 1.1)**. In the first step, primary miRNAs (pri-miRNAs) are produced. Pri-miRNAs are long transcripts containing one or multiple miRNA sequences and are transcribed by RNA polymerase II [3], but can also refer to the mRNA transcripts in the case where the miRNA sequence is located within an intron [4, 5]. dsRNA hairpin segments within the pri-miRNA are then recognized by the DGCR8 component of the microprocessor complex, [5, 6] and are released from the pri-miRNA following cleavage by the Drosha RNaseIII enzyme, which constitutes the second part of the microprocessor [2, 5]. These ~60-70nt hairpins are called precursor, or pre-miRNAs and are exported from the nucleus to the cytoplasm via the Exportin5 transport receptor protein [2, 5].

Once in the cytoplasm, the pre-miRNAs are recognized by the Dicer complex. In mammals, this consists of the Dicer RNaseIII type enzyme, in complex with two other proteins: PACT and TRBP, which are important in generating Dicer specificity [5]. The Dicer enzyme recognizes the double stranded region of the pre-miRNA and cleaves it, removing a region known as the terminal loop, and resulting in an RNA duplex ~22nt in length, with the 5’phosphate and ~2nt 3’ overhang that is the signature of the RNaseIII type enzymes [2, 5, 7]. This duplex consists of the mature miRNA in complex with its antisense strand, which is called the miRNA* [2, 5].

The miRNA* seems to be peeled away from the miRNA and degraded, and so the current model of miRNA silencing involves the separation of this duplex, where the miRNA is loaded into the RNA Induced Silencing Complex, or RISC (although in some cases the miRNA* might also be functional) [2, 8]. The RISC contains several proteins with the key effector being a member of the Argonaute family of proteins. The function of the miRNA in the RISC then is to act as a target specifier – directing the RISC towards the mRNAs which are to have their expression repressed [2].

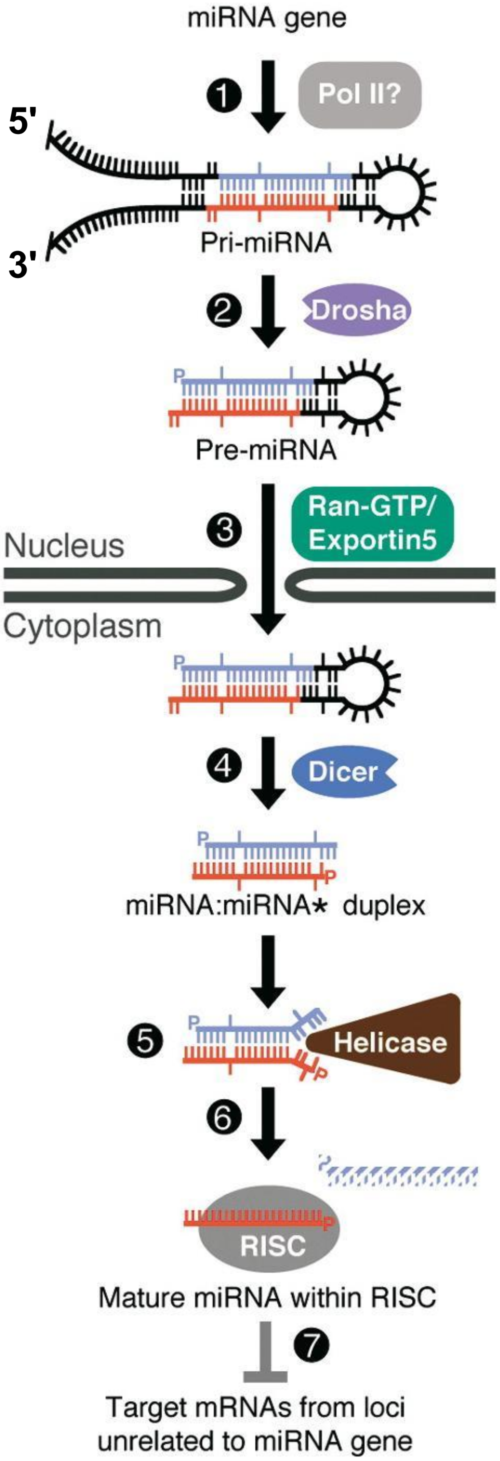


Figure 1.1: The biogenesis of mammalian miRNAs.

Pri-miRNAs are transcribed from DNA by PolII and cleaved by the drosha component of the microprocessor to produce pre-miRNAs. The pre-miRNA is then exported to the cytoplasm from the nucelous to be cleaved by Dicer. The strands of the mature miRNA duplex are separated by a helicase and one of the strands loaded into the RISC. This targets the RISC to various mRNAs which prevents their translation. miRNA is coloured in red, miRNA* is coloured in blue, and the terminal loop region is coloured in black. Adapted from Bartel (2004).

1.1.1.2 The effect of miRNAs on gene expression

The prevention of translation by the RISC can take three forms: direct cleavage of mRNA, inhibition of translation by the ribosome, or mRNA destabilization and degradation [2, 9, 10]. It is thought that the method of post-transcriptional repression chosen by the cell is dependent on the degree of complementarity between the miRNA and target mRNA sequence [2]. Plant miRNAs often have perfect complementarity between their miRNAs and target mRNA sequences and mRNA cleavage is mediated by an Argonaute protein [11]. In animals, however, the degree of complementarity is low, and usually confined to a 7-8nt region of the miRNA called the seed region [5, 10], which defines both the targets and family of the miRNA [12]. For animals, the most common methods adopted are translational repression and mRNA destabilization [12].

In the translational repression method, the prevailing model is that once a miRNA loaded RISC (miRISC) is bound to the 3'untranslated region (UTR) of the target mRNA, it is then able to also bind the 5' 7-methylguanosine cap. This occurs through its AGO2 Argonaute protein, which displays some similarity with the eIF4E eukaryotic translation initiation factor. In such a system, the miRISC would therefore prevent the translation initiation complex from forming, and hence prevent the translation of the RNA. Although this model is appealing, it is possible that the miRISC could also interfere at other stages of translation, either instead of, or in tandem with, this process [10].

In contrast, the exact mechanism used by the miRISC to destabilize and degrade mRNA is not known. Binding of the miRISC leads to deadenylation of the mRNAs, which allows them to be degraded by the exosome, but how the miRISC causes this is not clear. It is possible that binding of the miRISC to the 5' cap could disrupt the normally circular structure of the mRNA, which would make the 3' poly(A) tail more likely to degrade, but the order in which the different mechanisms of miRISC based repression occur has not been elucidated [10].

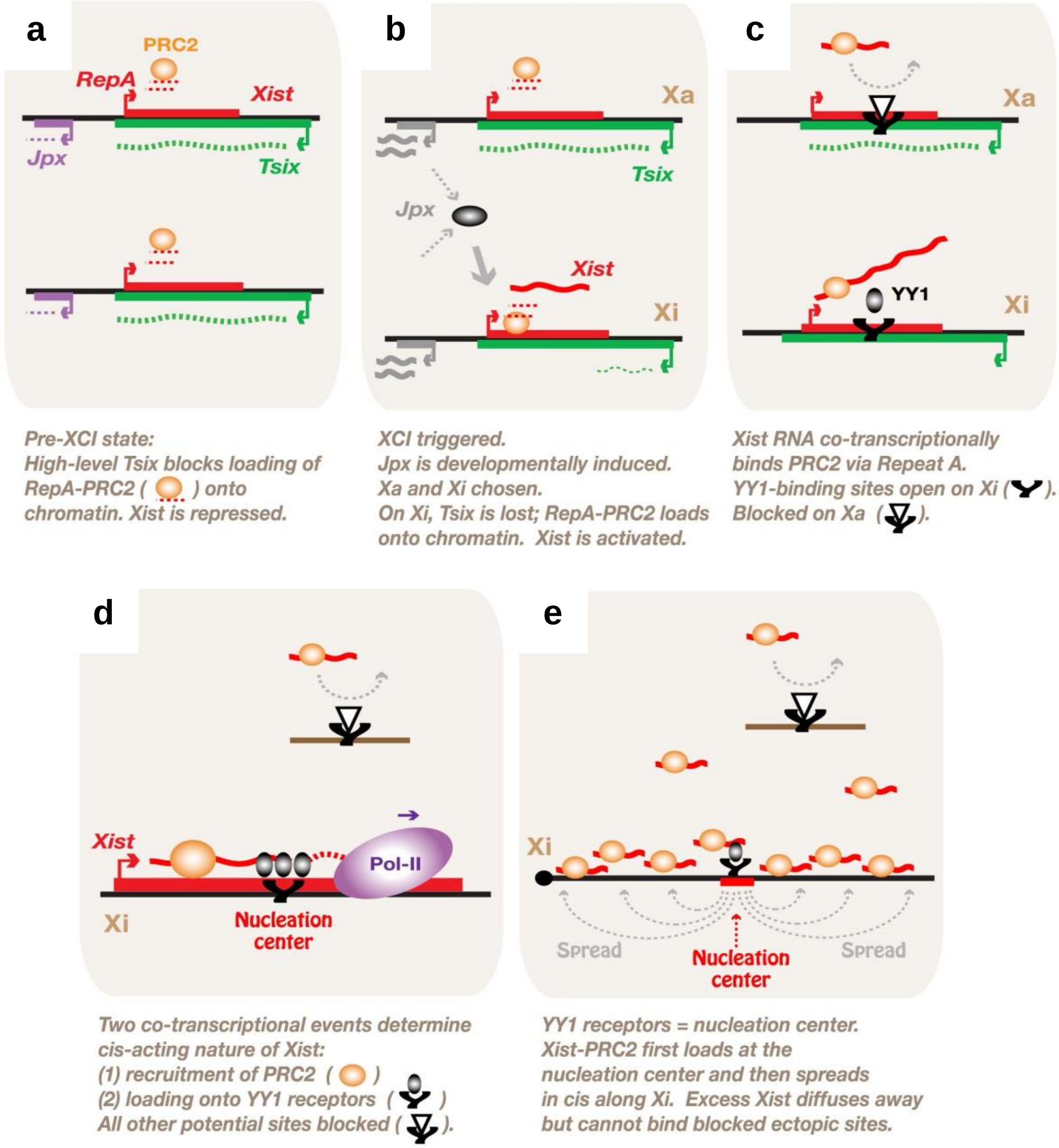
1.1.1.3 Biological consequences

The advantage to the cell of the miRNA system of gene expression control is in its versatility. As mentioned above, only the 7-8nt seed sequence is important in defining the target of a miRNA, and so each miRNA has a vast number of potential targets [12, 13]. In addition, the effect of a miRNA on its targets is modest [14], and to ensure effective repression, multiple miRNA target sites are needed per mRNA [10, 12]. The result of this is that miRNAs can be used to fulfill multiple regulatory functions in the cell. One such function is that of a switch, where a miRNA has a strong effect on the target mRNA. Here, the production of the miRNA can be used to "turn off" the expression of a gene. This is the classical view of miRNA function, but if the miRNA is already present, such thinking can lead to an alternate view, whereby a miRNA could act as a "failsafe", providing a redundant level of protection against the expression of a particular gene, which is not meant to be expressed at a given time. miRNAs can also act as a "tuning" system. By only reducing the expression levels of protein from a particular gene by a small amount, a miRNA can help to optimize the quantities of protein in a cell to ensure maximum efficiency [12]. miRNAs are therefore powerful regulators of gene expression, and their ability to target multiple mRNAs can place them at the centre of complex regulatory networks, resulting in numerous important biological outcomes. This function is due to the properties of RNA; the ability of a miRNA to "read" the sequences of multiple mRNAs allows it to help decode the transcriptome for the RISC proteins, and thus control gene expression at a post-transcriptional level.

1.1.2 ncRNAs as regulator: long ncRNAs

There are currently two classes of ncRNA that are known to influence gene expression at the transcriptional level: lncRNAs and piRNA [15]. Long non-coding RNAs (lncRNAs) are defined as transcripts above 200 nucleotides in length that do not encode a functional protein and are known to control gene expression levels through a variety of different mechanisms [16, 17].

Perhaps the best-known lncRNA is the Xist RNA, which has been well studied due to its role in X chromosome inactivation (**Fig. 1.2**) [17-20]. Female mammals possess two X chromosomes, one maternal and one paternal, and, in order to control the levels of gene dosage, one of these two chromosomes is randomly inactivated [21, 22]. Present on the X chromosomes is a region known as the X inactivation centre (Xic), from which several important lncRNAs, including Xist, are transcribed. [22-24] Basal levels of Xist are found spread throughout the nucleus [25], and are countered by another lncRNA, Tsix, which is transcribed from the antisense strand of the Xist gene [26-28]. At the beginning of X Chromosome Inactivation (XCI), the expression of Tsix RNA is downregulated from the X chromosome that is to become inactive (Xi) [28]. Tsix RNA is therefore lost from the future Xi, but still present on the future Xa [28]. This allows a short RNA transcript of the Repeat A region of Xist, known as RepA, to be transcribed. This in turn recruits the polycomb complex, PRC2, to methylate the histones of the Xist promoter on the future Xi chromosome [29, 30], allowing it to be activated by the developmentally timed lncRNA known as Jpx, which causes an upregulation of Xist. [29-31] The Xist transcripts are also able to bind PRC2 through the Repeat A motif they contain, [30] but in order to enable the silencing of the X chromosome genes, this protein:RNA complex must be tethered to the target sequence [25]. This is achieved by the binding of a downstream site on the DNA by the YY1 protein, which is able to bind to both DNA and RNA, and so is able to tether the Xist/PRC2 complex co-transcriptionally to the Xic site. This nucleation step then allows further Xist/PRC2 silencing complexes to spread across the chromosome and hence prevent gene expression from the now inactive X chromosome by an as-yet undiscovered mechanism. Alternative sites then cannot be bound by Xist/PRC2 complexes, because they either lack the presence of YY1, or because they are prevented from interacting by Tsix [25].



Pre-XCI state:
High-level Tsix blocks loading of RepA-PRC2 (●) onto chromatin. Xist is repressed.

XCI triggered.
Jpx is developmentally induced.
Xa and Xi chosen.
On Xi, Tsix is lost; RepA-PRC2 loads onto chromatin. Xist is activated.

Xist RNA co-transcriptionally binds PRC2 via Repeat A.
YY1-binding sites open on Xi (Y).
Blocked on Xa (▽).

Two co-transcriptional events determine cis-acting nature of Xist:
(1) recruitment of PRC2 (●)
(2) loading onto YY1 receptors (Y)
All other potential sites blocked (▽).

YY1 receptors = nucleation center.
Xist-PRC2 first loads at the nucleation center and then spreads in cis along Xi. Excess Xist diffuses away but cannot bind blocked ectopic sites.

Figure 1.2 Mechanism of X-chromosome inactivation by Xist lncRNA.

(a) RepA-PRC2 is prevented from binding chromatin by the Tsix lncRNA. (b) During XCI, Tsix is lost from Xi and RepA-PRC2 methylates the Xist promoter, allowing its activation by Jpx. (c) and (d) Xist lncRNA binds PRC2 and is tethered to Xi by YY1. (e) Further Xist/PRC2 complexes spread along Xi in cis, resulting in methylation of Xi. Xa remains unaffected as its YY1 site is blocked by Tsix. Expression from Xi is therefore silenced through this mechanism. From Jeon et al. (2011)

Other lncRNAs are also known to act as tethers and recruit polycomb complexes to other genes, epigenetically silencing them [15, 17]. One problem a cell faces is in ensuring that only a specific allele or locus is targeted by the silencing machinery. The properties of lncRNA make it ideally suited to this task. While a protein must first exit the nucleus and be translated, and can only recognize short segments of DNA that may be present multiple times within a genome; RNA may hybridize with DNA co-transcriptionally, and retain its positional information. In addition, the length of the lncRNA allows for some mobility even while the RNA is tethered, and hence enables it to recruit proteins to its tethering site [17]. In this manner, a lncRNA can recruit histone methylation proteins to a specific site in the genome, and ensure epigenetic silencing.

1.1.3 ncRNAs as regulators: piwi-associated ncRNAs

Piwi-associated ncRNAs, also known as piRNAs, are another class of ncRNAs that can influence the epigenetic state of target genes. They are only found in germline cells, and cells located nearby, and are known to play a vital role in protecting the germline from transposable elements [15, 32]. In *Drosophila*, this is achieved at the post-transcriptional level through what has been dubbed the “ping-pong” mechanism [32, 33] (**Fig. 1.3**). Briefly, piRNA clusters which contain antisense transposon sequences are transcribed and bound by the Piwi clade of Argonaute proteins, which are then targeted to sense transcripts of the corresponding transposon, that are, in turn, cleaved by the protein. The resultant sense RNA is then bound by Ago3, another Piwi clade protein, which again cleaves the RNA. This sense complex can then bind another piRNA cluster transcript and cleave it in the same way, and thus produce more piRNAs. Through such a method, therefore, transposon mobility can be prevented by a positive amplification loop.

Mammalian piRNAs repress transposon mobility by an alternative method. The piwi clade proteins in mice are known as MIWI, MIWI2 and MILI. MIWI is expressed in the pachytene stage of meiosis, but its function is currently unknown [15, 34]. MIWI2 and MILI, on the other hand, are expressed in the pre-pachytene stage [34], where the majority of bound piRNAs correspond to transposon sequences [35]. These proteins also use the ping-pong mechanism of amplification, but in addition are able to prevent transposon mobility by transcriptional gene silencing [15, 34]. Here, although the exact molecular mechanism of silencing remains unknown [15, 32], a link has been established to the DNT3A and DNT3B DNA methyltransferase enzymes [34]. It has therefore been postulated that the MIWI2 and MILI proteins use the piRNAs generated from transposon mRNA transcripts to guide them towards the genomic loci encoding these transcripts through base pairing interactions [15]. Once bound, the complexes could repress the expression of these elements by recruiting DNA methylases to these regions. It is also speculated that there could be an interaction between these complexes and histone methylating proteins which would also help to silence the transposable elements [15, 32].

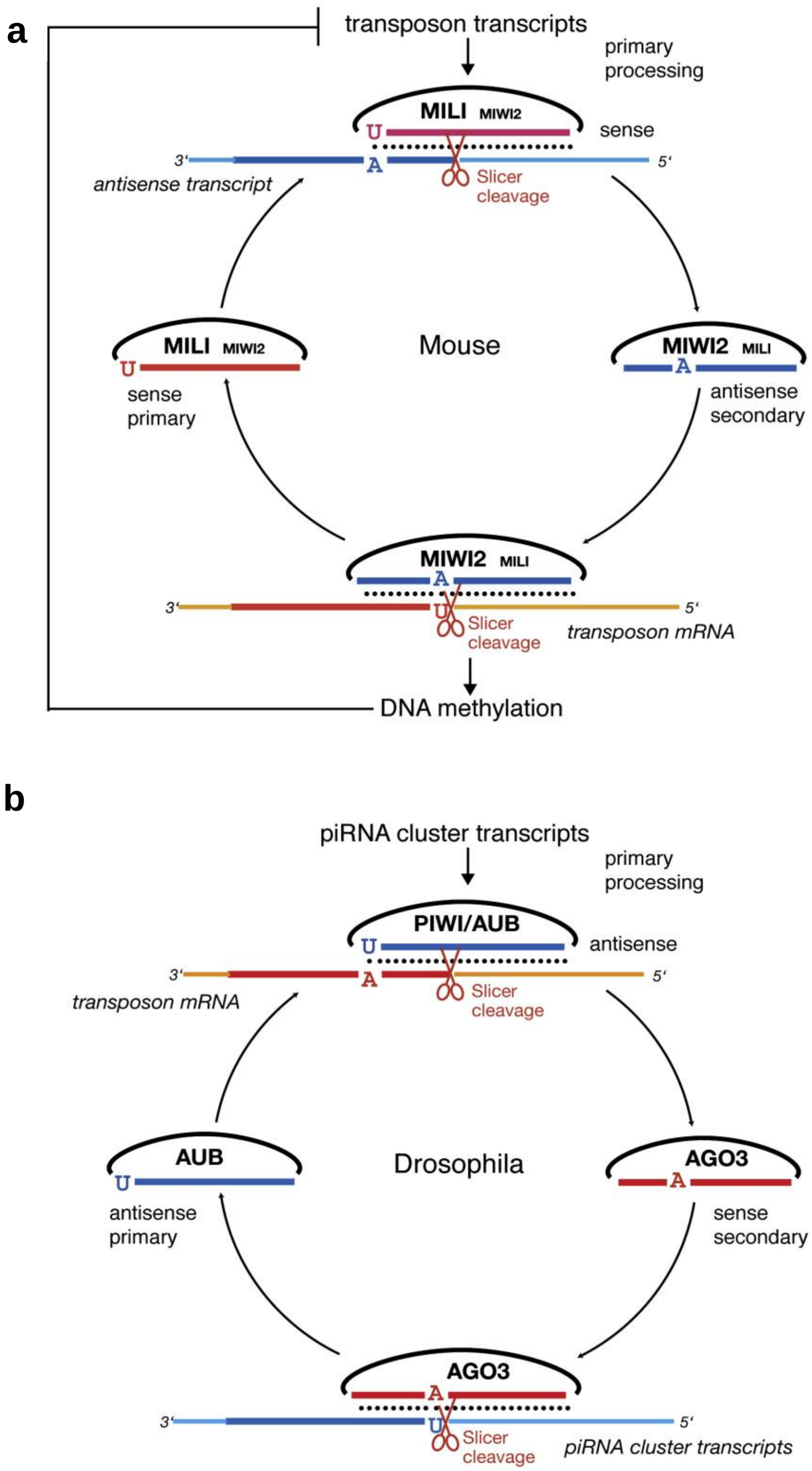


Figure 1.3: Ping-Pong mechanisms of piRNA mediated genome defence.

(a) In mammals, MILI uses an RNA derived from transposon transcripts to target a piRNA. The piRNA is cleaved, and the cleaved piRNA segments used to target MIWI2 to transposon mRNA. This mRNA is then also cleaved and the cleaved portion used by MILI to start a new round of the cycle. MIWI2 can also use the cleaved piRNA to direct DNA methylation of transposon genes. (b) In flies, a similar mechanism is used with the Piwi clade proteins AUB, PIWI and AGO3 in place of MILI and MIWI2. Adapted from Aravin et al. (2008)

The advantage of using RNA as a defence mechanism against these transposable elements is in its flexibility. The major challenges faced by cells in preventing the genomic instability caused by such elements are that they can be diverse in sequence, and must be distinguished from endogenous genes. The elegance of this piRNA system is that it targets transposons based on one of their major defining properties – their mobility within a genome. Once a transposon has inserted itself within a piRNA cluster, it is targeted by the silencing machinery at multiple levels of gene expression. The ping-pong mechanism of amplification removes the transposon mRNA by cleavage whilst at the same time enhancing the effect of the piRNA machinery. The direction of methylase enzymes towards the loci encoding the transposons then provides a longer-term solution for preventing their mobility, by silencing their transcription. piRNAs and their associated proteins therefore form an immune system which is able to defend the germline against the destabilizing action of these mobile elements [32].

1.1.4 ncRNAs as a defence mechanism: CRISPRs and crRNAs

Prokaryotes and archaea have also developed a system of using ncRNAs as a defence mechanism. crRNAs are transcribed from Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) loci, and act as a defence against infection by foreign DNA. CRISPR loci consist of a leader sequence followed by a series of direct repeats 20-50bp in length separated by unique spacer sequences [36, 37]. These spacer sequences originate from foreign plasmids and phages that the organism has previous encountered, and act as a genetic "memory bank" [37] of previous infections. The CRISPR system is therefore often referred to as an RNA mediated adaptive immune system [37].

The mechanism of CRISPR mediated immunity has three stages (**Fig. 1.4**), each mediated by Cas (CRISPR associated) proteins, whose genes are often found adjacent to CRISPR loci. The three stages are: (1) integration of genetic information into the CRISPR locus; (2) biogenesis of crRNAs; and (3) disruption of targets. In addition, there are three different types of CRISPR-Cas system which must complete these processes, known as types I, II and III. The differences between these systems is mainly due to the different Cas proteins involved in each stages, and organisms often utilize more than one type of CRISPR-Cas system [36].

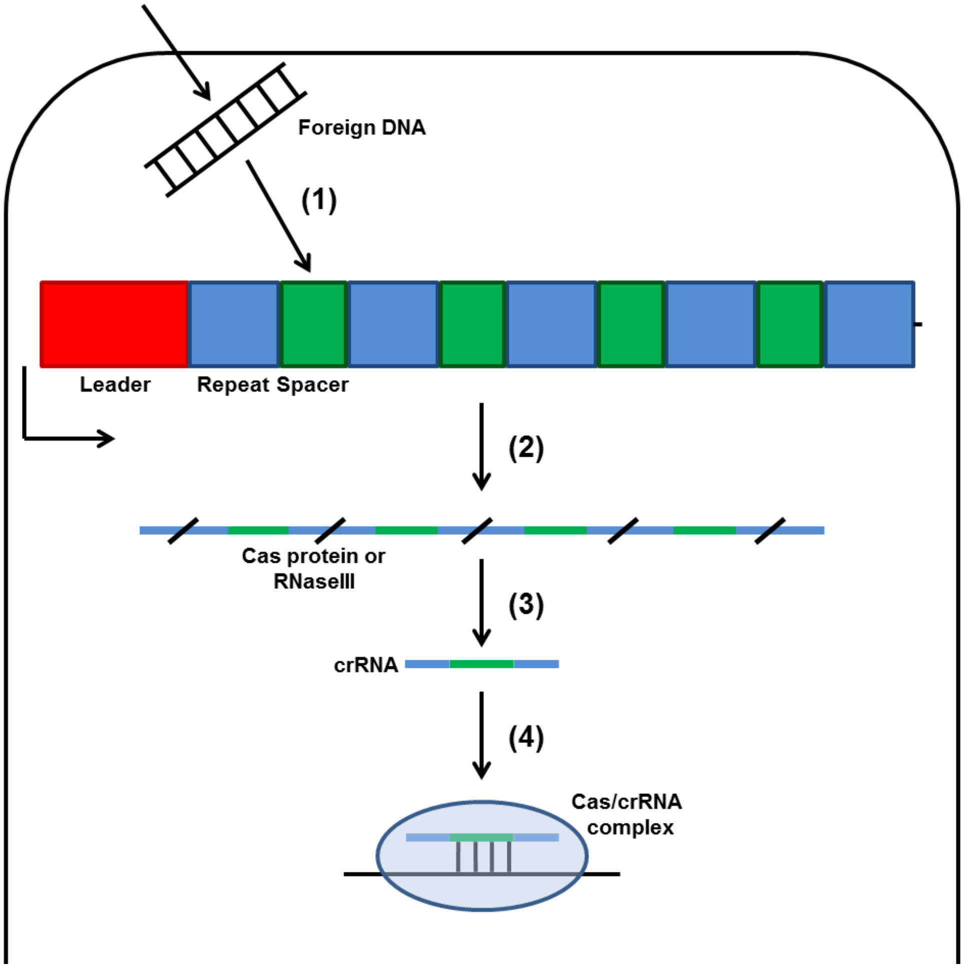


Figure 1.4: General mechanism of CRISPR-Cas system.

(1) When foreign DNA (from a plasmid or bacteriophage) is inserted into a bacterium that contains the CRISPR-Cas system, it is incorporated as a new spacer between two repeat sequences at the 5' end of the CRISPR locus. (2) The CRISPR locus is transcribed as a long RNA. (3) The RNA transcript is cleaved either by Cas proteins or RNaseIII enzymes to produce the short crRNAs, which are comprised of the spacer region and parts of the flanking repeat regions. (4) Target interference is mediated by Cas proteins, which are guided to the DNA targets via base pairing between the crRNA and the target DNA. This occurs as the spacer region of the crRNA is complementary to the target, but the flanking repeat regions are not.

In the first stage, a short segment of foreign DNA is inserted into the CRISPR locus. Although this occurs in all CRISPR-Cas systems, it has only been demonstrated experimentally for the type II system. In this system, new sequences are integrated at the start of the CRISPR locus near the leader sequence. The repeat sequence is duplicated for each new DNA segment added in order to conserve the architecture of the CRISPR locus. The integration of DNA into the CRISPR locus is dependent on Cas7 for type II systems. It is not currently known which Cas proteins provide this function in type I and type III systems, but several studies have suggested Cas1 and Cas2 as potential candidates [36].

The activity of the CRISPR-Cas system is dependent on crRNAs. A long RNA transcript is first produced from the CRISPR locus, which is cleaved in the middle of each repeat section to produce the short ~60nt crRNAs corresponding to the spacer sequences flanked by the remnants of the repeat segments. The Cas protein responsible for this cleavage in type I and III systems is the Cas6e endoribonuclease, although it is currently unknown which Cas protein is responsible for cleavage in type II systems. Recent evidence has suggested a role for RNaseIII in crRNA processing, implying that it may act as an extra host-dependent factor responsible for crRNA biogenesis. Alternatively, it has been shown that Cas9 is required in vivo for crRNA processing, although its precise enzymatic role has yet to be defined [36].

Following their biogenensis, the crRNAs associate with other Cas proteins to form large RNPs, and guide them towards target foreign DNA. These are recognised through base-pairing interactions between the crRNA and either the sense or antisense strand of the target DNA [36]. The efficiency of this base-pairing interaction is dependent on the binding of a high affinity site on the 5’ end of the spacer sequence, which acts similarly to the seed sequence of miRNAs [36, 38] (discussed in *Section 1.1.1.2*). Once the target has been recognised and bound by the RNP, both DNA strands are cleaved, preventing any further activity [36]. Crucially, this cleavage is dependent on the degree of complementarity between the crRNA and the target DNA. This is important for distinguishing between self and non-self DNA. The spacer region of the crRNA will be complementary to both the foreign DNA and the spacer in the CRISPR locus, but the remnants the flanking repeat regions will only be complementary to the CRISPR spacer. Full complementarity protects against cleavage by the Cas proteins, and so only the foreign DNA will be cleaved [39]. The CRISPR-Cas system therefore resembles both the miRNA system - where base-pairing between an ncRNA and target is facilitated through the use of a seed sequence and leads to target selection and interference - and the piRNA system - where ncRNA acts as a defence against harmful nucleic acid species. As in the eukaryotic systems, the ability of crRNAs to recognise their targets through base-pairing interactions, and hence guide the effector Cas proteins to these targets, is important for their function.

1.1.5 ncRNAs in mRNA processing

Although not directly linked to regulation of gene expression, another type of ncRNA, small nuclear RNAs (snRNAs), do play a role in the processing of mRNA transcripts. After transcription, mRNAs must be capped, spliced and polyadenylated before their export to the cytoplasm [40]. The splicing process, that is, removing certain introns from the mRNA sequence, is catalysed by a dynamic molecular machine called the spliceosome. The spliceosome has many constituent parts, including a highly changeable array of proteins, involving five protein:snRNA complexes called snRNPs as the major components [41, 42]. Namely, these are U1, U2, U4/U6 and U5, which each contain a snRNA, seven Sm proteins and varying numbers of proteins that are specific for each snRNP.

In order to remove an intron, the spliceosome must catalyse two transesterification reactions, leading to the removal of a lariat intron, and the ligation of the 5’ and 3’ exons [41]. The canonical mechanism by which the spliceosome achieves this is as follows (**Fig. 1.5**): (1) U1 snRNP binds the 5' splice site, while protein factors interact with the branch site (BS) and poly-pyrimidine tract (PPT) to form what is known as the E complex. (2) The U2 snRNP then binds to the BS to form the A complex. (3) The U4/U6.U5 tri-snRNP then attaches to form the B complex, or pre-catalytic spliceosome. (4) Rearrangements in the spliceosome structure result in the destabilization of the U1 and U4 snRNPs, which dissociate, allowing the spliceosome to adopt the active form of the B complex. (5) Catalysis of the first transesterification reaction is initiated by the interaction of the Prp2 RNA helicase with the spliceosome, which results in the C complex. (6) The C complex then catalyses the second transesterification reaction to complete splicing. (7) The spliceosome components dissociate and rearrange before taking part in another splicing reaction [42].

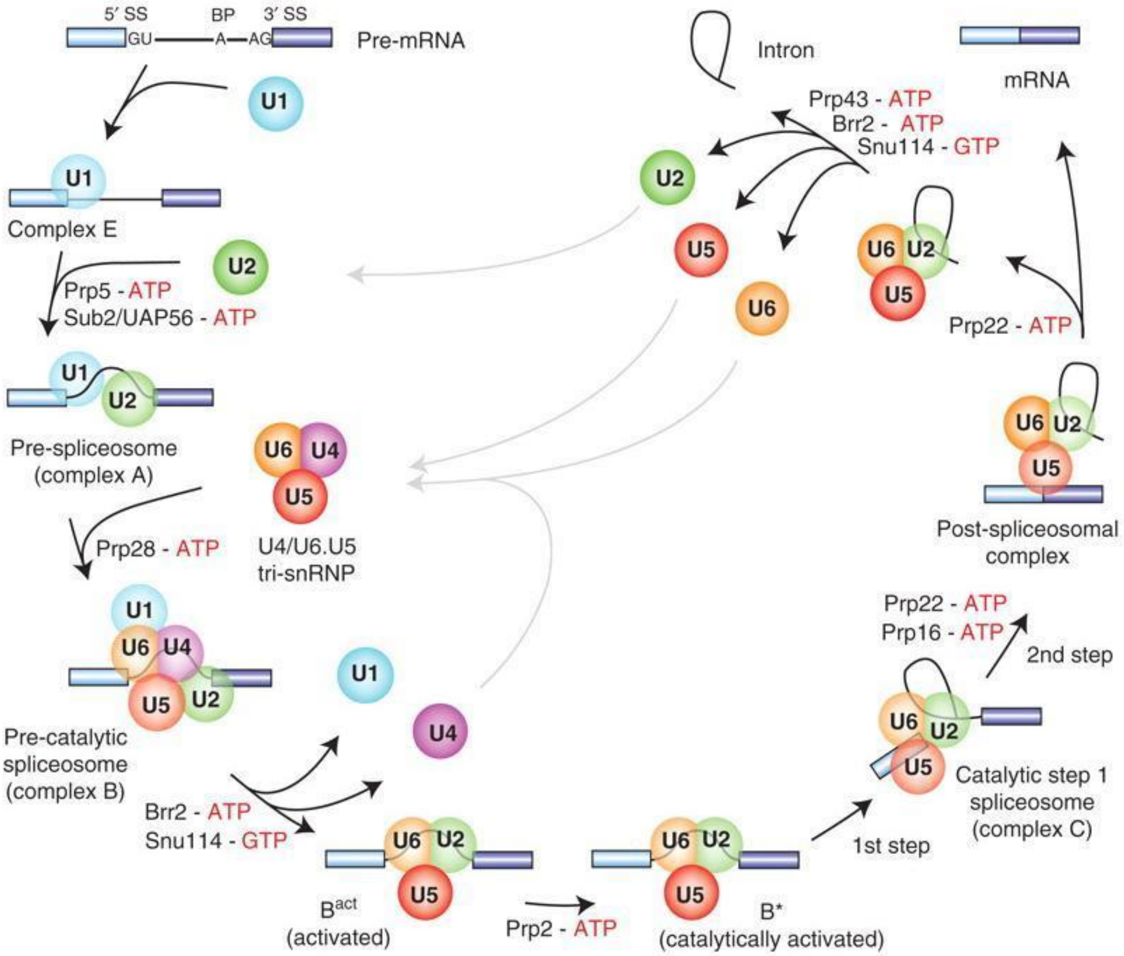


Figure 1.5: Overview of the splicing mechanism.

The U1 snRNP binds the 5' splice site of a pre-mRNA, forming the E complex. The U2 snRNP then binds the BS, forming the A complex. The U4/U6.U5 tri-snRNP then binds to form the pre-catalytic spliceosome, known as complex B. A rearrangement then takes place: U1 and U4 are destabilised and dissociate, forming the active B complex. The Prp2 RNA helicase interacts wih the active B complex and facilitates the first catalytic step, resulting in the C complex. This complex catalyses the second step, forming the spliced mRNA, an intron lariat and the free U2, U5 and U6 snRNPs. Extra protein factors that facilitate each of these steps are highlighted next to the arrows depicting the progression from each step to the next. Adapted from Will and Luhrman (2010)

Whilst the large number of protein factors involved are key components of the spliceosome and its function, the role of RNA in facilitating splicing resides mainly in its ability to form base pairs, and adopt a large range of different conformations [42]. Base pairing is important for enabling U1 snRNPs to recognize the 5' splice site, and in allowing U2 to bind to the BS. Upon U2 binding, the branch point adenine is bulged out from the duplex, which primes the 2' hydroxyl group for the first catalytic step. Base pairing contacts also help to stabilize the U4/U6 snRNP complex, and the reorganization of the hydrogen bonding network upon the U4/U6.U5 tri-snRNP binding of the A complex causes U6 to base pair with the 5' splice site, resulting in the displacement of U1 from this site, and its subsequent destabilization [41, 42]. The U2/U6 interaction is also dynamic, and it is possible that the base pairing interactions that dictate the conformations adopted by these snRNPs change during splicing [42]. With these factors in mind, it is clear that the properties of RNA are crucial in determining the dynamic nature and function of the spliceosome.

1.1.6 rRNA and tRNA: The RNA components of translation

The final role of ncRNAs in the cell are as components of the translation machinery, the ribosomal (r)RNAs and the transfer (t)RNAs. The ribosome is the molecular machine responsible for catalyzing the synthesis of polypeptide chains; it is composed of a mixture of rRNA and protein, with rRNA as the major component. It contains three tRNA-binding sites, named A,P and E for aminoacyl, peptidyl and exit. The extension of the polypeptide chain is catalysed by the progression of the ribosome across the mRNA transcript, with the aminoacyl-tRNA:mRNA complexes cycling through the different sites, each promoting a different stage of the reaction. The role of rRNA in this process is two-fold; it ensures fidelity of the tRNA:mRNA interaction and also acts as a catalyst. The catalytic function of the rRNA in this case is in positioning the α-amino group of the amino acid bound by the A site tRNA in such a way as to prime its reaction with the carboxy group of the tRNA-bound amino acid located in the P site [43].

1.1.6.1 rRNA as a determinant of fidelity

rRNA also has a major role in determining the fidelity of translation. The functional challenge facing the ribosome is that the difference in energy between cognate and non-cognate mRNAs is very small, and fidelity is of critical importance, as a mistranslation of the mRNA could result in an inactive protein. To overcome this challenge, the ribosome contains several nucleotides that help to distinguish between Watson-Crick, and non-Watson-Crick base pairs between the codon and anticodon.

The structure of the *Thermus thermophilus* small ribosomal subunit shows three nucleotides that change conformation upon the binding of tRNA in the A site. These nucleotides interact with the minor groove of a double stranded helical segment that forms between the codon and anticodon and form a hydrogen bonding network. Importantly, for the first two nucleotides, these interactions can only take place in the presence of a cognate Watson-Crick base pair, irrespective of which type of base pair it is (i.e. A-T, C-G). This is not the case for the third "wobble" position, explaining the lack of specificity at this position. This mechanism, along with an additional "proofreading" mechanism whereby non-cognate tRNAs are more likely to dissociate during the accommodation step of translation, explains how the ribosome maintains translation fidelity [43]. As before, it is due to the structural and chemical properties that RNA possesses.

1.1.6.2 The role of tRNA in translation

The major advantage of using RNA as the means of amino acid transfer to the ribosome is in its ability to decode the codon sequence through specific Watson-Crick base pairs to ensure that only the correct amino acid can be incorporated into a polypeptide chain at a particular position. In addition, there are numerous stabilising RNA:RNA interactions that take place between the ribosome and the tRNA [43].

In order to function correctly, tRNA must adopt a complex tertiary structure so that it can be recognized both by the ribosome and also the aminoacyl-tRNA synthetase enzymes. The secondary structure of tRNA is often presented as a classical “cloverleaf” consisting of three stem loops and the acceptor stem (**Fig. 1.6**), that are each characterized by a particular sequence motif. The four arms of the tRNA cloverleaf are thus named the CCA acceptor stem, the Dihyrouridine loop (D-loop), the anticodon stem loop and the TΨC loop.

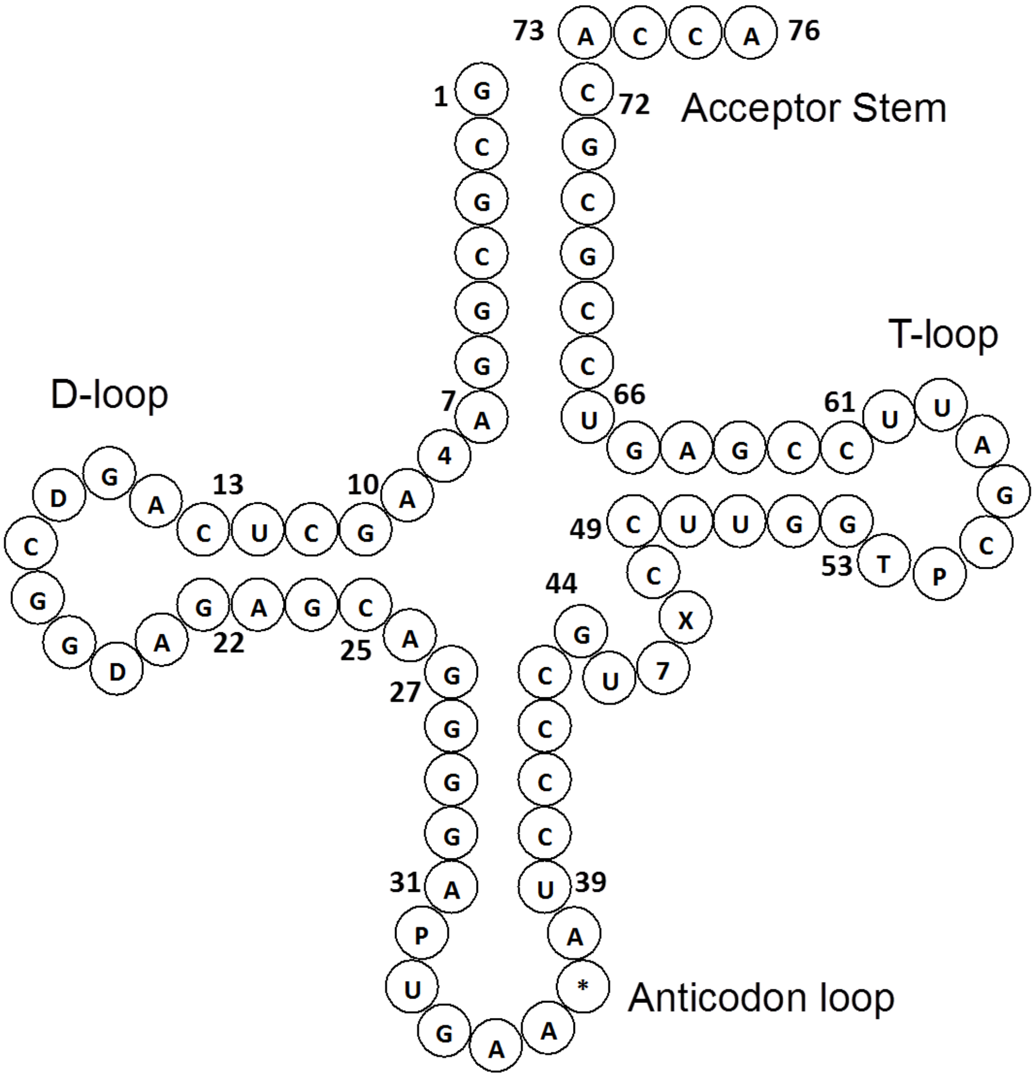


Figure 1.6: Secondary structure of tRNA^{phe}.

Modified nucleotides are depicted as follows: 4 = 4-thiouridine, D = dihydrouridine, P = Pseudouridine, * = 2-methylthio-N6-isopentenyladenosine, X = 3-(3-amino-3-carboxypropyl)uridine, T= 5-methyluridine.

1.1.6.3 Structural features of tRNAs

The acceptor stem contains the 5' and 3' termini of the tRNA. The 3' terminus is longer than the 5' by four nucleotides, ending in the conserved CCA motif [44]. The terminal adenine of this motif is the site of amino acid coupling to the tRNA by the aminoacyl-tRNA synthetase enzymes [45], hence why this arm is called the acceptor stem. At the opposite end of the tRNA is the anticodon stem loop, which contains the trio of bases used to specifically recognise the codon sequence in the mRNA.

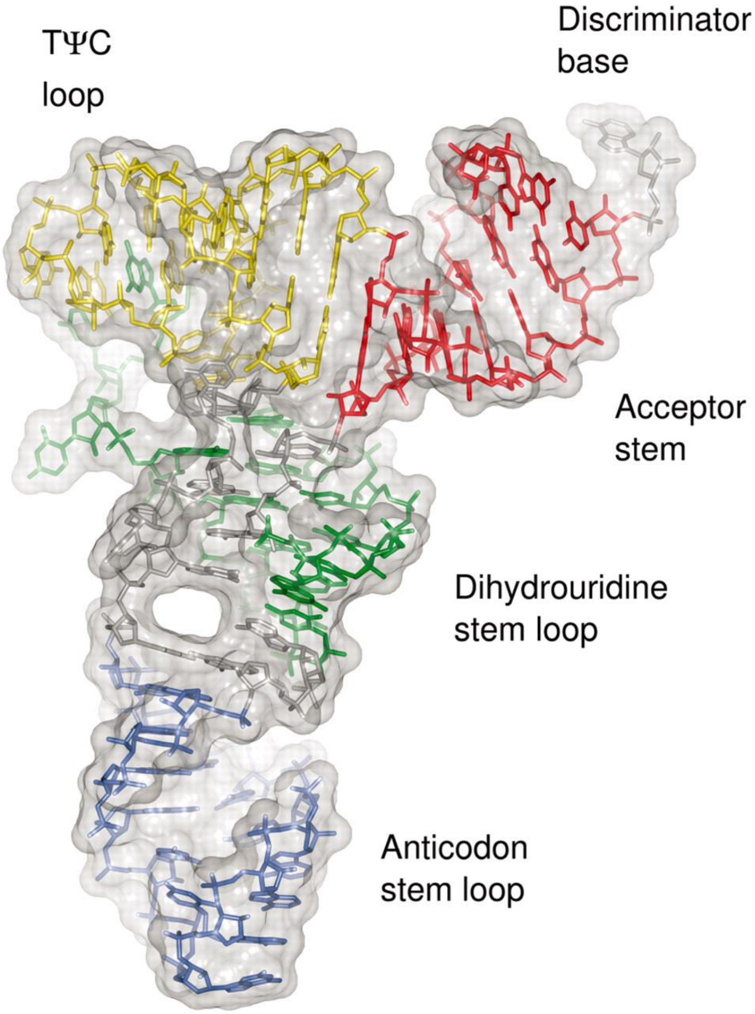


Figure 1.7: The crystal structure of unmodified tRNAphe.

This structure shows the classical "L-shaped" tertiary structure, with the relevent arms highlighted. Adapted from Byrne et al. (2010).

The other two arms are named after the modified nucleotides that feature prominently in their sequences, namely dihydrouridine in the D-loop and pseudouridine (Ψ) in the TΨC loop. The tertiary structure of tRNA adopts a highly folded L-shape (**Fig. 1.7**) rather than the cloverleaf structure and in order to form this structure several non-canonical RNA:RNA interactions must form. The modification of tRNA nucleotides can alter the local properties of the structure, and, although modifications do not significantly alter the overall structure of tRNA, their presence has several structural implications that help to optimize the biological performance of the tRNAs [46-48]. Consequently, tRNA molecules are the most heavily modified of all the RNA types [44, 49-51].

1.1.7 RNAs as controllers and mediators of transcription

What has been shown in this section, therefore, is that RNA is a very adaptable molecule that is well suited to fulfilling a wide variety of roles in the cell. The fact that RNA has an OH group at the C2’ position instead of H as in DNA, means it is more chemically active, and can more readily take part in catalytic processes. In addition, the ability of RNA to “read” both DNA and other RNA sequences through base pairing interactions enable it to not only serve useful purposes during translation by providing sequence fidelity to the process, but also act as a decoder for the more chemically diverse proteins, which can then perform their functions in a sequence specific manner. Finally, the ability of RNAs to hydrogen bond to themselves, sometimes with the assistance of nucleobase modifications, enables them to adopt a range of complex structures, which further diversifies their usefulness as catalytic entities within the cell.

1.2 RNA-binding proteins are key components of RNA mediated gene expression control

Many RNA molecules interact with and are controlled by proteins that bind them and influence their structure, stability and function. For example, neither the lncRNAs nor piRNAs can complete their functions without the action of the PRC2 methylation complex, which is composed of protein. Without the action of the microprocessor, Dicer and RISC protein complexes, miRNAs could neither exist nor influence gene expression levels.. Without the proteins that comprise the snRNPs of the spliceosome, or those present in the ribosome, the core processes of transcription and translation could not take place. The CRISPR-Cas system requires several different Cas proteins as effectors at each stage of its mechanism. Finally, for tRNAs to function, both aminoacyl-tRNA synthetase and nucleotide modification enzymes need to be present. With such a range of functions that need to be completed, RNA binding proteins must have evolved with a variety of RNA recognition, binding and catalytic structural domains. This section is therefore devoted to examining how the structures of RNA binding domains are adapted toward their various functions.

1.2.1 Structural principles of RNA binding proteins

1.2.1.1 RNA is bound by the action of several forces

RNA is composed of three elements: phosphate groups, ribose sugars, and nucleobases. Binding of a protein to RNA is often the result of a mixture of both specific and non-specific interactions. Specific interactions take the form of a hydrogen bonding network formed between the amino acids of a protein and the RNA bases, where the hydrogen bonds are dependent on the existence of a particular base at that position. If this base is not present, an incomplete hydrogen bonding network will form and binding will be weaker. Non-specific interactions can be roughly subdivided into three types: hydrophobic stacking interactions between the RNA bases and aromatic amino acids in the protein, hydrogen bonds between the protein and the sugar-phosphate backbone of the RNA, and electrostatic interactions between the negatively charged phosphate groups in the RNA backbone and positively charged amino acids in the protein. These interactions are important as they can allow a protein to recognize the shape and structure of the RNA, and can be used to strengthen complexes either in addition to sequence specific contacts, or in cases where it is more important to bind a certain RNA fold rather than a specific sequence.

1.2.1.2 RNA Binding Proteins exploit modularity

Whilst the variety of RNA binding protein (RBP) functions would imply a similar diversity in structure, there are relatively few types of RNA binding domain. How then, is it possible to generate the structural diversity needed to bind such a wide range of RNA substrates, each with different sequences, lengths and structures? To fulfill their functions, RBPs exploit a property called modularity – using combinations of RNA binding domains to create an extended binding surface which can have different properties without the need for evolving new types of domain (**Fig. 1.8**). Evolutionarily this is advantageous as new RBPs can be created easily by gene duplication, resulting in different RNA binding surfaces with different functions [52]. In addition, most RNA binding domains can only recognise a short segment of RNA, thereby limiting both specificity and affinity. By using several RNA binding domains at once, the protein can increase both of these factors by extending the length of RNA that can be recognized [52, 53]. The structural examination of RBPs therefore involves determining which classes of RNA binding domains they contain, and how these link together in order to create the RNA binding surface.

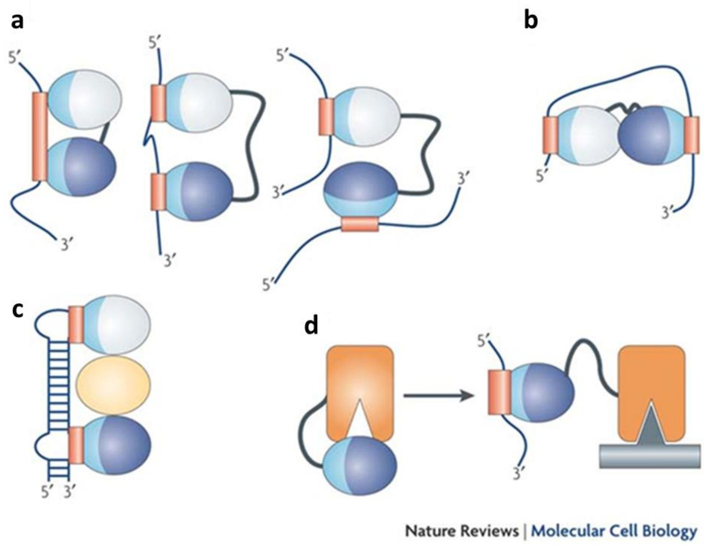


Figure 1.8: RNA Binding Proteins exploit modularity for multiple effects.

(a) Combining two RNA binding domains with a flexible linker allows longer or more distal sites to be recognised by the RBP. (b) RNAs can be made to adopt certain topologies upon RBP binding. (c) Binding of multiple domains can allow other modules in the protein to be positioned properly. (d) The RNA binding domains can help determine target specificity for enzymatic domains. RNA is shown as lines with red binding sites, RNA binding domains as blue elipses, and other protein domains as orange shapes. Adapted from Lunde et al. (2007).

1.2.1.3 Importance of the linker regions in RBPs

The tethering of several RNA binding domains together places emphasis on the linker regions, which, although not necessarily involved in binding, do serve important functions in creating the RNA binding surface. Primarily, it is the length of the linker region which produces these effects. A longer flexible linker region between two RNA binding domains allows them to accommodate a wider range of targets, whereas shorter, less flexible linkers are more suited to providing an extended RNA binding platform [52, 53]. In addition to this, the linker regions can also greatly influence the affinity of an interaction. Here, the tethering of a second domain to the first limits its radius of movement, and so can increase its effective concentration in the area surrounding an RNA once the first domain is bound, increasing the affinity of the interaction compared to that of the two free domains. This is highly dependent on the length of the linker, such that at 60 residues there is little difference between the multidomain protein and the individual domains, but at shorter lengths will lead to increases in affinity compared to the free domains [52-54]. Finally, the above two effects assume the linker is flexible, but linker regions can also become ordered upon RNA binding, and participate in the interaction [52, 53].

1.2.2 RNA Binding Domains: RNA Recognition Motif (RRM)

The RRM (also referred to as the RNA Binding Domain, RBD, or Ribonucleoprotein Domain, RNP) is the most ubiquitous type of RNA binding domain found in higher vertebrates and adopts a $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ fold, where the four β strands stack against the two α helices. Single stranded RNAs are preferred and the RNA binding surface of this domain is on the β sheet face. In this mode, a typical RRM can bind between 2 and 8 nucleotides using a series of aromatic residues found in the first and third β strands, which stack with the 5' and 3' nucleotides (**Fig. 1.9**). In addition, between one and three connecting loops are also involved and are critical in allowing the interaction to take place. As the binding site size is small, RRMs are often modified slightly, either with extensions to the N or C termini, or by the inclusion of a linker region. They are also an excellent example of the modularity described above: RRMs are often found in tandem, where short linkers allow them to extend their RNA binding surface and increase the affinity of the protein for the RNA by providing extra interactions between the two components [55, 56]. Thus, as described above, multiple weak interactions can be combined to bind to longer or differently shaped RNA binding partners with high affinity and specificity. This therefore makes the RRM highly adaptable and able to recognise a wide range of targets making them well adapted to their use in proteins involved in mRNA processing and transport [57].

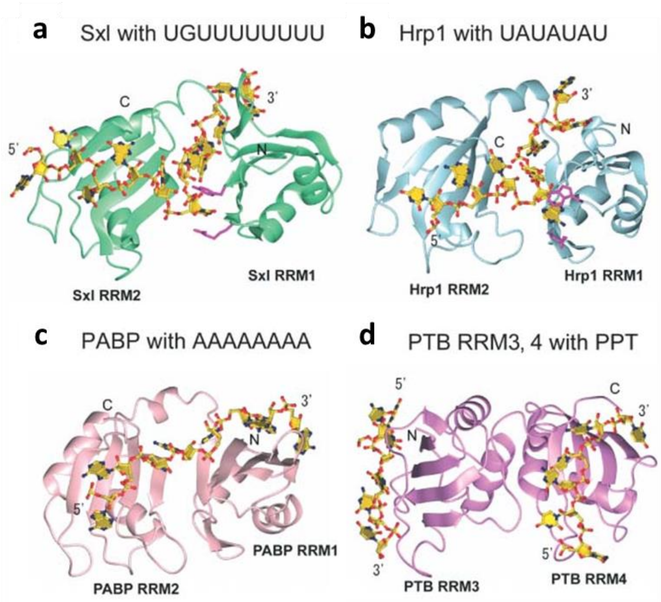


Figure 1.9: Examples of tandem RRM domains recognising RNA substrates.

The $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ fold is visible in each case, and the ssRNA can be seen to bind across the surface of the β sheets, with extra interactions provided from the loop regions of the protein. Long sequences, or multiple sites, can be bound by the tandem RRMs. (a) shows the Sxl protein's RRM1 and 2 in complex with a UGUUUUUUUU oligonucleotide (PDB: 1BZF), (b) is the Hrp1 protein's RRM1-RRM2 in complex with UAUAUAU RNA (PDB: 2CJK), (c) is of the PABP protein's RRM1-RRM2 binding AAAAAAAA RNA (PDB: 1CVJ) and (d) is the RRM3-RRM4 of the PTB protein in complex with a poly-pyrimidine tract oligonucleotide (PPT) (PDB: 2EVZ). Adapted from Muto and Yokoyama (2012).

1.2.3 RNA Binding Domains: Cold-Shock Protein Domain (CSD)

Another common RNA binding domain is the CSD, which is named after the cold shock proteins originally discovered in bacteria. These proteins were found to act as single stranded RNA binding proteins [58], preventing mRNA from misfolding in cold conditions and allowing translation to proceed [58-60]. Eukaryotic homologues contain the CSD, but instead use its ssRNA binding activity to achieve different functions via alternate mechanisms [61].

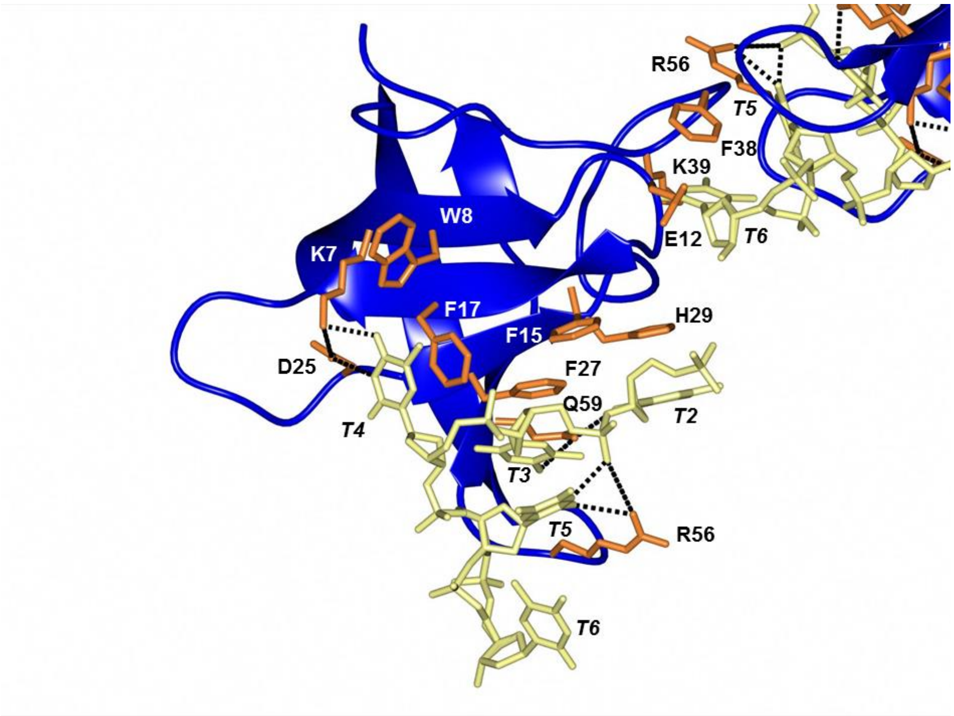


Figure 1.10: Ribbon diagram of CspB from *B.subtilis*, in complex with hexathymidine.

Protein is shown in blue with interacting sidechains shown in orange cylinders. Hexathymidine is shown in yellow cylinders. The DNA is bound through both stacking interactions and hydrogen bonds (dashed lines). A second symmetry related CspB/dT6 complex is shown in the top right to depict the interactions of T6 and T5 with the protein. Residue numbers are shown, along with the nucleotide numbers in italics. PDB code: 2ES2

The CSD adopts the OB fold, and forms a characteristic β-barrel structure, constructed from five β-strands [61, 62]. A 1.78Å crystal structure of the cold shock protein CspB from *Bacillus subtilis* shows how CSDs bind their RNA targets by analysing the binding of the protein to dT6 (hexa-thymidine, **Fig.1.10**). The CspB binding surface consists of a preformed hydrophobic binding platform composed of aromatic residues surrounded by positively charged residues. Phenylalanine side chains stack in between the nucelobases of the dT6, while a hydrogen bonding network forms between basic and polar side chains from around the hydrophobic platform and nucleobase groups. Replacement of hydrophobic residues with alanine resulted in 55-190 fold increases in $K_{d.}$ In contrast, replacing the polar residues resulted in either no increase, or a smaller, 12-fold increase in K_d. High concentrations of salt do not disrupt the CspB-dT6 interaction. These data therefore indicate that the CspB-Nucleic Acid interaction is largely hydrophobic in nature [63]. This suggests that the positively charged residues found around the platforms help to overcome the repulsion caused by the negatively charged nucleotides encountering the acidic CspB [64]. This binding activity can be utilized by eukaryotic proteins which employ the CSD in functions such as regulation of transcription, mRNA export and stability, splicing and miRNA biogenesis [61].

1.2.4 RNA Binding Domains: Zinc Fingers (ZnF)

One of the best-known nucleic acid binding domains is the ZnF. The classical ZnF was initially discovered in transcription factors and they have been well characterised as DNA binding domains [65], although they also bind RNA [66]. The classical ZnF contains a β hairpin and an α helix, with a Zinc atom tetrahedrally coordinated by a variation on a motif of cysteins and histidines e.g. CCCC, CCHH, CCHC or CCCH [66], although the term has been used to describe any small domain stabilised by a Zinc atom [67].

There are therefore many folds that can be described as ZnFs, with a large variety of structures [67]. Correlated with this is the large array of functions that ZnFs take part in; many are found in transcription factors, but other examples include the YY1 protein involved in XCI (*Section 1.1*) and the retroviral nucleocapsid proteins [68].

The HIV-1 nucleocapsid protein contains two tandem CCHC type Zinc Knuckles (ZnK) and binds the third of four stem loops (SL3) in the unspliced Ψ RNA during retroviral genome recognition and packaging [69] (**Fig. 1.11**). Binding is achieved through the placement of an alpha helix of the protein into the major groove of the RNA helix. Exposed guanosine bases of the GGAG tetraloop of the SL3 are incorporated into hydrophobic clefts formed by residues from each of the two ZnKs. Hydrogen bonding interactions also take place between the residues of the ZnK and the exposed guanosines of the tetraloop in a manner reminiscent of Watson-Crick base pairing, generating specificity. Also present are various basic residues that can form either hydrogen bonds or electrostatic contacts with the phosphodiester bonds of the RNA. It should be noted, however, that this mechanism of binding differs greatly from CCHH type Zinc Fingers often found in eukaryotic transcription factors, which normally bind mainly through side chains of residues found in alpha helices [69-71]. The ZnF family is therefore both structurally and functionally diverse, and once again use modularity to assist in binding their targets [67, 68].

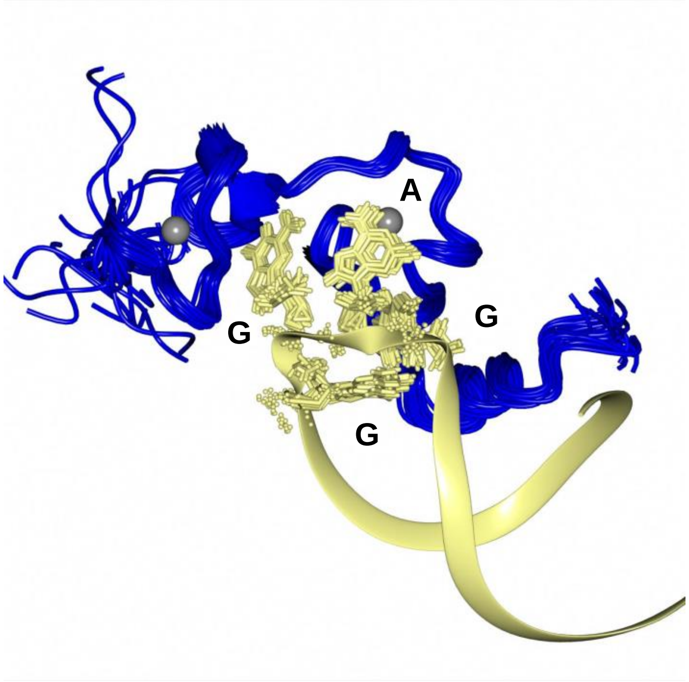


Figure 1.11: Solution structure of the HIV-1 nucleocapsid protein.

Zinc Knuckles bound to the GGAG RNA tetraloop of the SL3 stem loop recognition element of the ψ site. RNA is shown in yellow and protein in blue, with Zn^{2+} ions depicted as grey spheres. PDB code 1A1T.

1.2.5 RNA Binding Domains: Double Stranded RNA Binding Domain (dsRBD)

The classical dsRBD differs from the above domain types in that it is mainly non-specific in nature and is used to differentiate between RNAs based on their structure rather than their sequence [52, 72]. The general architecture of the protein is of an $\alpha_1\beta_1\beta_2\beta_3\alpha_2$ structure where the N and C terminal α helices pack against the three stranded antiparallel β sheet. 16bp of dsRNA is then recognized by three regions of the protein (**Fig. 1.12**). In the first region, the N terminal α helix interacts with the minor groove of the RNA and forms direct and water mediated hydrogen bonds to bases and the 2’ OH ribose groups. In the second region, a protein loop interacts with the adjacent minor groove, resulting in the formation of more hydrogen bonds. Finally, in the third region, the C terminal α helix bridges across the major groove of the RNA, with several amino acids forming a further hydrogen bonding network. For this interaction to take place, the major groove of an RNA double helix must be recognised. Similarly, the hydrogen bonds between the protein and the 2'OH of the ribose sugars in the other two regions mean that the dsRBD binds only dsRNA and not dsDNA. The majority of hydrogen bonds in this case are between the protein and the sugar-phosphate backbone, and contacts with bases tend to be water mediated and so not sequence specific. In such a way therefore, the dsRNA is perfectly adapted to recognizing dsRNA in a non-specific manner [72, 73], and multiple dsRBDs can be used to specify for particular dsRNA structure types [52]. This function of dsRBDs is exploited by a variety of proteins for different cellular roles including RNA deaminase enzymes, viral defense proteins [74], the tRNA modification enzyme hDus2L [75, 76], as well as by TRBP and Dicer in miRNA biogenesis as described above (*Section 1.1.4.1*) [74].

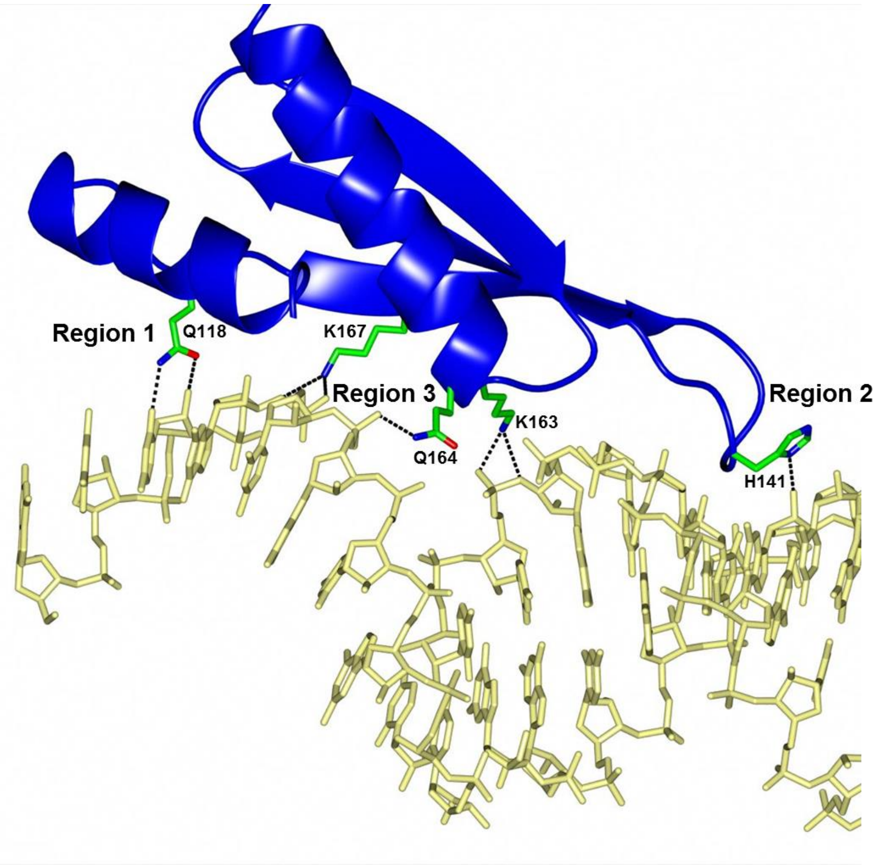


Figure 1.12: Crystal structure of the dsRBD from the Xlrbpa2 protein, in complex with dsRNA.

All three regions of the domain are visible in blue, and the domain interacts with 2 successive minor grooves of the dsRNA (yellow). Direct hydrogen bonds between the protein and RNA are depicted as dashed lines. Interacting protein side chains are coloured by atom and shown as sticks. Regions 1 and 2 can be seen contacting the minor grooves of the dsRNA whilst region 3 is inserted within the major groove. PDB code 1DI2.

1.2.6 The importance of RNA binding protein structure

Several key types of RNA binding domain highlight the importance of protein structure with regards to the biological function of the protein:RNA complexes in gene expression systems. By combining several of these domains, RBPs can recognise a myriad of targets and perform a large variety of functions. Due to the flexibility of RNA as a mediator of gene expression, this structural and functional diversity of RBPs is critical for allowing them to interact with and either facilitate or regulate the actions of the different types of both coding and non-coding RNA.

To understand the structure/function of RBPs, this thesis will focus on two different RBPs: the Lin28 miRNA binding protein and the DusC (Dihydrouridine synthase C) tRNA modification protein.

1.3 Lin28: A Master Regulator of Gene Expression

One well studied RNA binding protein is the eukaryotic Lin28 protein. Highly conserved throughout both vertebrates and invertebrates, Lin28 (or Lin-28 in *C. elegans*) is composed from a unique combination of RNA binding domains: an N terminal cold-shock domain and two tandem retroviral type CCHC Zinc Knuckles (forming the ZnK domain), connected through a flexible linker region [77] (**Fig. 1.13**). Its best studied function is as a regulator of miRNA biogenesis, where it prevents the maturation of the let-7 family of miRNAs [78-81], however recently Lin28 has been shown to be a master regulator of gene expression, with an alternative role in mRNA binding and regulatory function [78]. Lin28 roles in multiple biological processes including embryonic development, maintenance of pluripotency and tumour formation and growth [82].

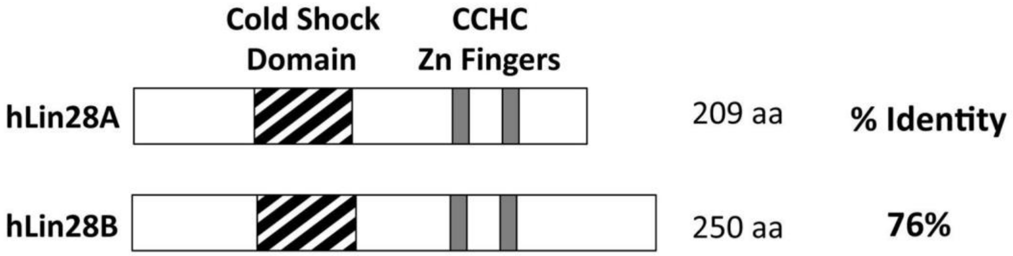


Figure 1.13: Domain structure of human Lin28A and Lin28B.

Each protein consists of one Cold Shock Domain (CSD) and two tandem Zinc Knuckles in the ZnK domain, connected by a flexible linker region. The major difference between the paralogs is the extended C-terminus present in hLin28B. Adapted from Piskounova et al. (2011).

1.3.1 The Lin28/let-7 axis

1.3.1.1 The let-7 family of miRNAs

Let-7 was first discovered in *C. elegans* where its mutation was lethal to developing worms, and since then has been intensively studied due to its important developmental roles, where its expression is largely correlated with the differentiation state of the cells it is present in [83, 84]. The let-7 family of miRNAs in humans consists of 13 miRNAs related through their seed sequences, which are highly conserved throughout both vertebrates and invertebrates [84] (**Fig. 1.14**). In addition to its role in cellular differentiation, let-7 miRNAs also target several potent oncogenes, notably the RAS proteins K-RAS, N-RAS and H-RAS; MYC, and HMGA2. This gives let-7 miRNAs an additional function as a tumour suppressor and maintenance of phenotype [83].

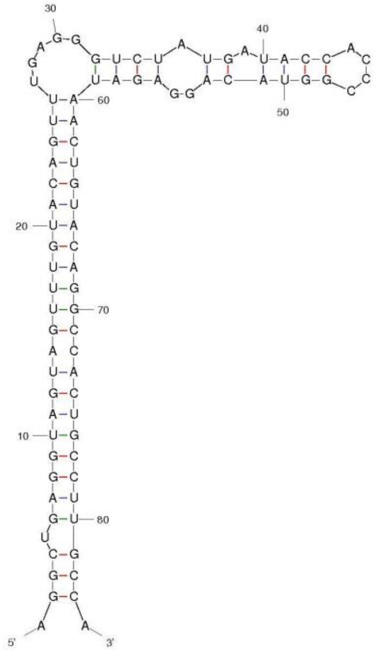


Figure 1.14: Secondary structure of the human pre-let-7g.

The lowest energy structure as predicted by MFOLD is shown.

1.3.1.2 Lin28 blocks let-7 maturation

The Lin28/let-7 interaction was discovered as a protein that bound agarose beads conjugated to pre-let-7g incubated with P19 embryonic carcinoma cell extracts [79]. This protein was Lin28, whose homologues had been previously shown to be important in the development of *C. elegans* [77] as well as several other processes [85, 86], mirroring the action of let-7 miRNAs [83]. Further experiments revealed that expression of Lin28 correlated with a block in let-7 processing, and demonstrated that Lin28 was responsible for binding to the let-7 miRNAs during their biogenesis and preventing their maturation [79, 80, 87]. Intriguingly, Lin28 protein is itself downregulated by let-7 miRNAs through binding sites on its 3'UTR [88]. Therefore, Lin28 and let-7 are part of a double negative feedback loop, where expression of one lowers the expression of the other, and each has the opposite biological effect. This led to the establishment of the concept of a Lin28/let-7 axis, where different phenotypic effects would be exhibited depending on whether let-7 or Lin28 was dominant in a particular cell at a moment in time [78, 82, 89].

The detailed mechanism of Lin28 protein's function has been elucidated in mammalian cells (**Fig. 1.15**). The two mammalian paralogues of Lin28, Lin28A and Lin28B, were found to inhibit let-7 biogenesis by two distinct mechanisms [81]. Lin28A is mainly found in the cytoplasm, where it can interact with the pre-let-7 miRNA (**Fig 1.14**). Once bound, a terminal uridyl transferase (TUTase), either TUT4 (Zcchc11) or TUT7 (Zcchc6) [90-92], is recruited to the complex and adds ~14 uracil nucleotides to the 3'end of the let-7 [93]. The addition of these uracils then signals for the degradation of the let-7 pre-miRNA by the Dis3L2 nuclease [94, 95], thereby preventing the cleavage of the precursor miRNA by Dicer to form mature let-7 miRNAs. In contrast, Lin28B is found in the nucleolus, due to the presence of localization signals in its C terminal and inter-domain sequence (**Fig. 1.13**). This enables it to bind to the pri-let-7 transcripts and sequester them in the nucleolus, and so no interaction with the microprocessor can occur [81]. If the primary transcript is not cleaved, no maturation step occurs and thus mature let-7 levels are downregulated. Through these mechanisms, the two Lin28 paralogs prevent mature let-7 miRNA production, resulting in increased let-7 target gene expression [96].

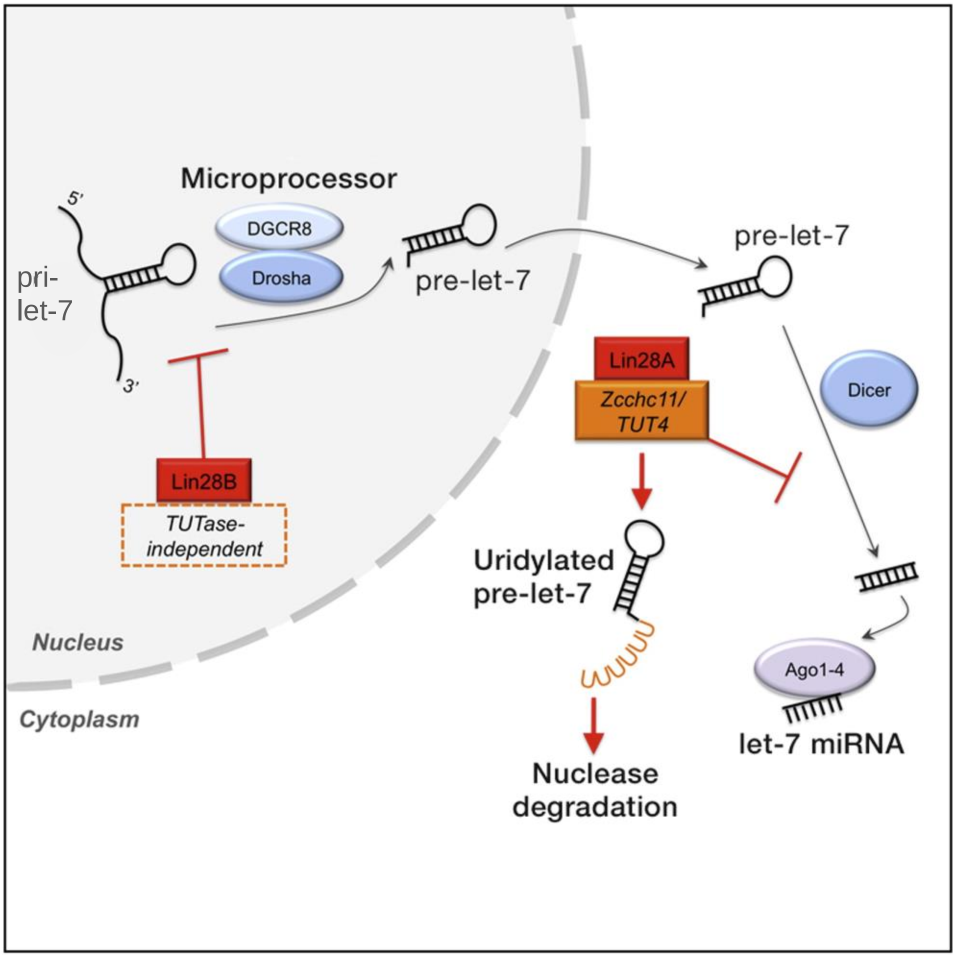


Figure 1.15: The differential roles of Lin28A and B in the inhibition of let-7 biogenesis.

Lin28B is localised to the nucleolus and prevents cleavage of the primary miRNA transcript by the microprocessor. In contrast, Lin28A is cytoplasmic, and directs the uridylation and subsequent degradation of the pre-let-7. If neither paralog is expressed, the miRNA undergoes biogenesis as normal. Adapted from Piskounova et al. (2011).

1.3.1.3 Biological consequences of Lin28 expression: Pluripotency

Lin28 expression is associated with the maintenance of pluripotency, a function it performs by utilizing both let-7 dependent and independent mechanisms [97]. Lin28 is thus one of four factors that can be used to generate induced pluripotent stem cells (iPS cells) [86], where it acts by accelerating the reprogramming process [98].

1.3.1.4 Biological consequences of Lin28 expression: Misregulation

Pluripotent cells have the capacity to self-renew indefinitely unlike normal somatic cells, a property also found in cancer cells. Lin28 regulates pluripotency and is found to be aberrantly expressed in ~15% of all cancers and is associated with tumour aggressiveness and advanced stage disease [99], and so has a key role in maintaining the tumour. This concept of pluripotency links Lin28 to the cancer stem cell (CSC) model of tumorigenesis. In human ovarian tumour samples, Lin28 expression was found to correlate with the proportion of ALDH1 (a marker of CSCs) positive cells within a tumour population [100], and Lin28B was found to be necessary for lung CSC proliferation and growth [101]. A potential model for Lin28's misregulation in CSCs is that aberrantly expressed/constitutively active oncogenes induce expression of Lin28A [102, 103] or Lin28B [103, 104], reducing mature let-7 levels. The reduction in let-7 induces further oncogenes, including Lin28, and inhibiting let-7 induced differentiation, hence the cell would revert to a less differentiated, self-renewing, stem-cell like state [99]. These CSCs are essential for the growth of the tumour according to the CSC hypothesis [105].

1.3.1.5 Biological consequences of Lin28 in Development

Lin28 was first discovered as a heterochronic gene in *C. elegans,* and so is required for the proper development of the organism [77]. In addition Lin28 has been shown to be important in determining the developmental timing and growth of both mice [106] and humans [107, 108], as well as in other metazoa [109, 110]. Lin28 can bind both miRNAs and mRNAs and has been shown to have a role in cellular development through both let-7 dependent and independent mechanisms. Lin28 regulates the development of primordial germ cells from embryonic stem cells by removing the let-7 mediated suppression of the Blimp1 transcription factor [111]. Overexpression of Lin28 in chondrocytes resulted in impaired proliferation through downregulation of let-7, leading to skeletal growth defects in mice. This observation demonstrated that suppression of Lin28 is important for allowing normal skeletal development, which requires mature let-7 miRNAs be present in order to downregulate the expression of their target genes [112]. Expression of Lin28 is also important for enabling the differentiation of myoblasts into myotubules through the stabilisation of IGF-2 mRNA [113].

The role of the Lin28/let-7 axis is more complicated in neuronal development. Lin28 is required for the Sox2 dependent proliferation of neuronal precursor cells (NPCs). This occurs through repression of let-7 biogenesis. Expression of let-7 in these cells prevents NPC proliferation and neuronal differentiation, therefore Lin28 expression allows NPC proliferation, as well as differentiation towards a neuronal cell type [114]. In addition, constitutive expression of Lin28 in P19 embryonic carcinoma cells differentiated down a neuronal-glial lineage resulted in blockage of glial differentiation and promotion of neuronal differentiation. A CSD was not required for blockade of gliogenesis, whereas the ZnK domain was required for both functions. The Lin28 dependent upregulation in the expression of several genes involved in neuronal differentiation was observed before the block in let-7 biogenesis. It was therefore concluded that Lin28 is able to promote neuronal differentiation through a let-7 independent pathway. Inhibition of gliogenesis correlated with the block in let-7 biogenesis, although it could not be determined whether gliogenesis is let-7 dependent [115]. The Lin28/let-7 axis is therefore an important factor in the development of various different cell types, although currently there is no clear general model of how the expression of Lin28 or let-7 results in particular cell types. It is likely that the outcome of Lin28 or let-7 expression is dependent on its cellular context.

1.3.1.6 Biological consequences of Lin28 in Metabolism and Tissue Regeneration

The role of the Lin28/let-7 axis in metabolic regulation comes through control of glucose uptake. Lin28 expression in adult mice was associated with insulin sensitivity, glucose tolerance and resistance to diabetes, whereas induction of let-7 led to the opposite effect [116]. The mechanism behind the link between Lin28/let-7 and diabetes is still not entirely clear [117], but it is known to be mediated by the Insulin-PI3K-mTOR pathway [116].

Lin28 is also involved in wound healing and tissue regeneration. It was observed that transgenic mice with a doxycycline inducible Lin28A increased levels of hair regrowth, digit repair and pinnal (ear) tissue regeneration, indicating that Lin28 might regulate the wound repair process. This was found to occur through let-7 dependent and independent mechanisms, and partly through molecular bioenergetics [118]. In addition, Lin28 has been shown to be required for retinal regeneration after injury in Zebrafish [119]. Lin28 in the retina was found to be regulated by the Ascl1a transcription factor, and induced the dedifferentiation of Müller glial cells by reducing let-7 levels and thus increasing the expression of regeneration and pluripotency genes that were repressed by the miRNA.

The biological functions of Lin28 are key and act on numerous different systems and tissues in the body. It is therefore of biological and clinical significance to comprehend the molecular mechanisms of Lin28 structure and function in detail in order to further our understanding of its role in these systems, and also aid in the design of inhibitors that may help to prevent or at least mitigate the effects of aberrant Lin28 expression in cancer. To this end, the induction of let-7 either artificially [99] or through the use of drugs [120, 121], results in suppression of cancerous phenotypes and implies that an appropriate Lin28 inhibitor could be clinically useful in helping to treat aggressive cancers.

1.3.2 Lin28 Structures

Three-dimensional structures of Lin28 on its own, and in complex with a nucleic acid have been determined. Three sets of structures have been documented to date in the literature: (1) Mouse Lin28A in complex with let-7d, let-7f-1 and let-7g [122], (2) The CSD of *Xenopus* Lin28B on its own and in complex with hexa- and hepta-thymidine, as well as the apo human Lin28B CSD [123], and (3) the solution structure of the ZnK domain of the human Lin28A [124].

1.3.2.1 Complexes of mLin28A with preE-let-7 miRNAs

Structures containing both domains were reported previously [122]. By truncating the interdomain linker of the mouse Lin28A protein, co-crystals of the protein with the preE segments of three let-7 family members: let-7d, let-7f-1 and let-7g were obtained that diffracted to 2.9Å, 2.8Å and 2.0Å respectively. The preE, or pre-element, is analogous to the terminal loop that Lin28 had already been shown to interact with [80].

The structural data, in combination with biochemical data, demonstrated that the separate domains of Lin28 (CSD and ZnK) contact two separate ssRNA sites of the miRNA. The CSD binds to the stem loop of the preE segment whilst the ZnK domain interacts with a conserved GGAG motif present on the 3' of the preE (**Fig. 1.16**), in agreement with the results of previous biochemical studies [92].

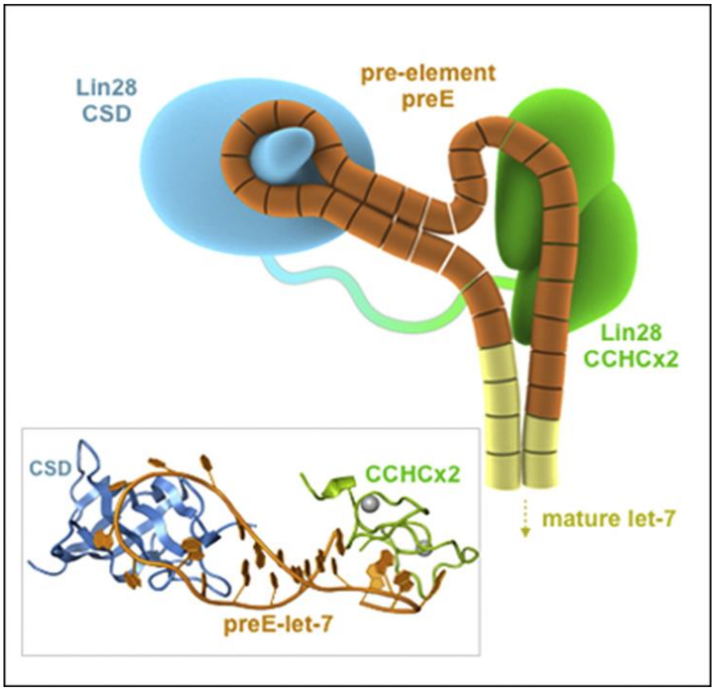


Figure 1.16: Lin28 interacts with two independent regions of let-7 miRNAs.

The cartoon model shows that the CSD (blue) contacts the loop region whilst the ZnK domain (green) interacts with the single stranded GGAG motif. The two domains are connected by a flexible linker region, which could allow them to recognise different length RNA sequences. The crystal structure is shown in a box at the bottom left of the figure. Adapted from Nam et al. (2011)

The interactions of the CSD with the preE loop showed that the loop wraps around a protrusion from the surface of the CSD in what is described as a “necktie” [122]. The bases of the loop are inserted into several hydrophobic pockets in the CSD where they stack against aromatic amino acid side chains (**Fig. 1.17**). In addition, a hydrogen bonding network exists between the bases and the protein. It was proposed that this system of hydrogen bonds, as well as steric factors within the pocket, define an ideal binding substrate of sequence NGNGAYNNN, where N is any nucleotide and Y is any pyrimidine [122].

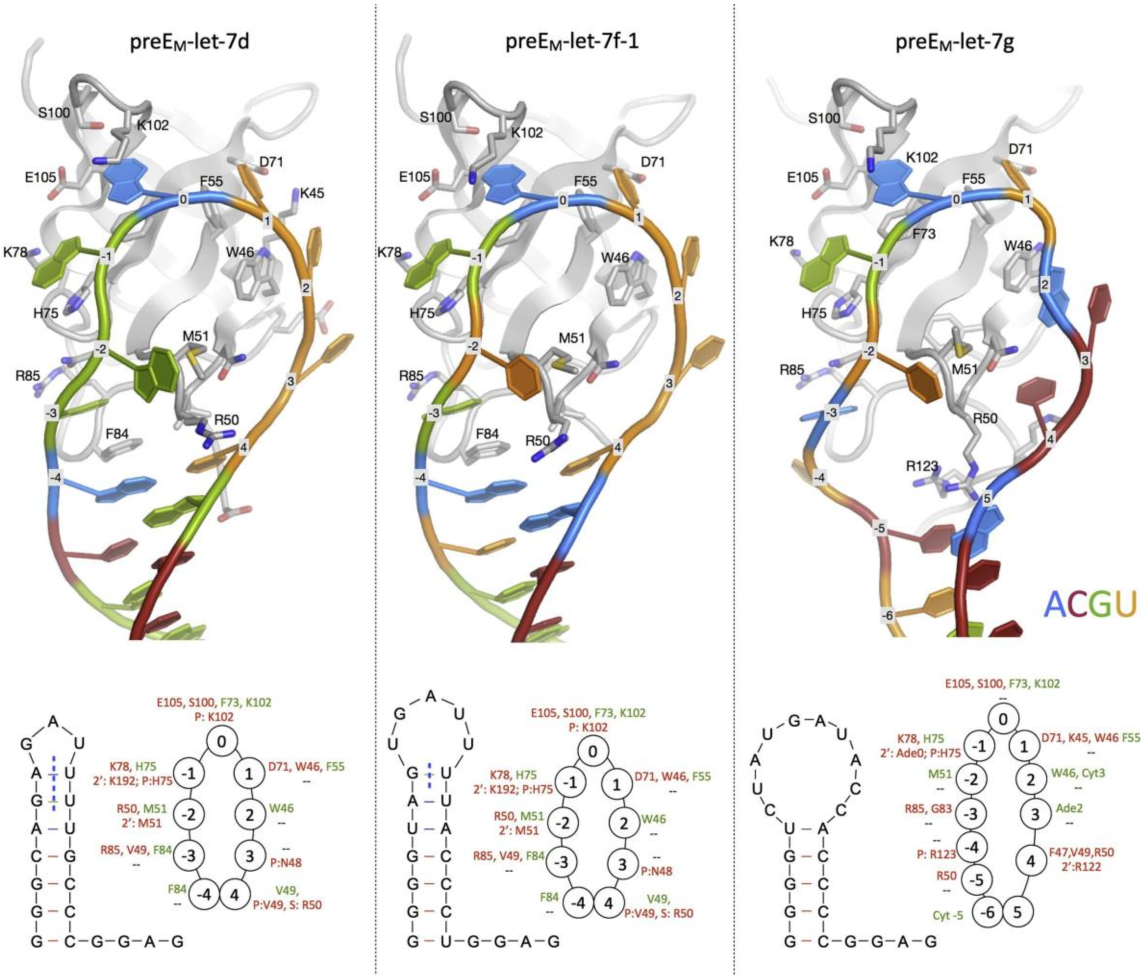


Figure 1.17: Interactions of the CSD with the terminal loop of let-7 miRNAs.

RNA bases are coloured as follows: A in blue, C in red, U in orange and G in green, with the CSD in white ribbons, and oxygen and nitrogen atoms of the side chains depicted in red and blue respectively. The RNA is bound by a mixture of stacking and hydrogen bonding interactions, which are detailed in below each structure, with stacking residues represented in green and residues forming hydrogen bonds shown in red. For each nucleotide position, the top line represents base contacts and the bottom line corresponds to interactions with the sugar-phosphate backbone. The secondary structure of each miRNA crystallised is also shown. Adapted from Nam et al. (2011).

The ZnK domain utilizes a vast hydrogen bonding network to maximize its interaction with a small motif [122]. The residue Y140 also makes a key interaction here by stacking in between the final A and G bases, resulting in a distinctive kink in the RNA structure (**Fig. 1.18**). It is of note that the conformation of the ZnK domain of Lin28 is highly divergent from that of the HIV nucleocapsid protein mentioned previously [69], and is likely to be unique to Lin28 [122].

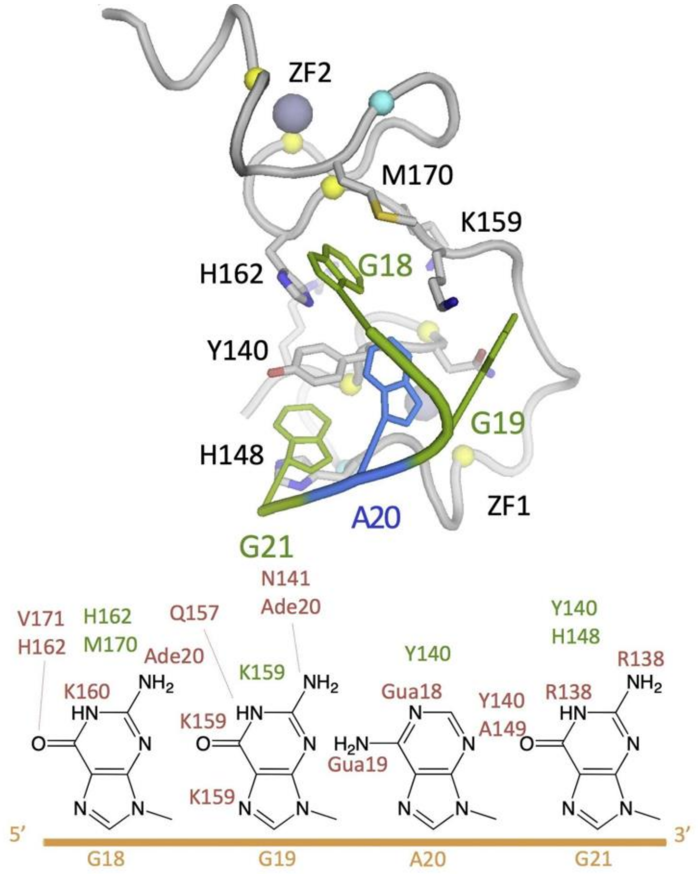


Figure 1.18: Interactions of the ZnK domain with let-7 miRNAs.

RNA nucleotides G and A are shown in green and blue respectively, cysteine residues as yellow spheres, histidines residues as cyan spheres, and zinc ions as grey spheres. Below, the hydrogen bonding interactions (red) and stacking interactions (green) are displayed. The residue Y140 can be seen to make a stacking interaction in this figure, and induce a distinctive kink in the RNA backbone. Adapted from Nam et al. (2011).

There are several other intriguing observations in these initial structures. The first is that the crystal structures formed domain swapped dimers, where each Lin28 molecule contacts two preE-let-7 miRNAs, to form a 2:2, protein:RNA complex. AUC experiments showed however that the complex should be monomeric, and so the proposed structure of the 1:1 complex was partly generated by modelling [122]. In addition to this, the conformation of the let-7g RNA seen in the crystal structures was not the most energetically stable secondary structure predicted to form by MFOLD [125], implying that the Lin28 either alters the RNA conformation, or preferentially binds one conformation over the other. Finally, the results of NMR relaxation experiments proved that the linker region between the two Lin28 domains is flexible, which would allow Lin28 to bind a range of substrates with different length. A final important conclusion is that binding of Lin28 to the preE-let-7 miRNA in the mode described in the paper would occlude the binding of Dicer to the full pre-let-7 miRNA, and hence prevent its maturation [122], presumably in conjunction with the other mechanisms mentioned previously [81].

1.3.2.2 Structures of the Lin28B CSD

Further structural data on Lin28 came from the structure of the CSD [123], from the *Xenopus* Lin28B protein, solved in the apo form and in complex with hexa-/hepta-thymidine segments of DNA. In addition, the structure of the human Lin28B CSD was also determined. The rationale for co-crystallising with the short poly-dT nucleotides came from previous studies of the bacterial CspB [63], and from biochemical analysis, that showed preferential binding to these types of oligonucleotide [123].

The structures of the human and *Xenopus* apo CSDs were determined to 1.95Å and 1.06Å respectively, and revealed that, similar to the bacterial CspB mentioned earlier, the domain adopted the β_5 barrel characteristic of the OB fold, containing a hydrophobic nucleic acid-binding surface [123]. This binding site is surrounded by polar and positively charged residues, giving the protein an amphipathic character due to the negatively charged surface on the opposing face of the protein. These structural characteristics make the Lin28 CSD more similar to the bacterial cold shock proteins, rather than their eukaryotic homologues. There was very high similarity between the human and *Xenopus* CSD crystal structures, which were almost identical with a backbone Cα atom RMSD of 0.2Å [123].

These structures revealed in depth how the Lin28B CSD is able to bind nucleic acids using this pre-formed platform. The DNA nucleotide bases point inwards to stack inside various hydrophobic pockets within the CSD [123], similar to the mLin28A structures [122]. The first nucleobase of the dT7 oligomer formed a 3 membered stacking interaction with F77 and the second T nucleobase. As well as taking part in this stack, T-2 was found inside a pocket defined by the residues Q69, V42, F40 and R78. The R78 then makes a total of 3 hydrogen bonds to T-2, causing it to be specifically recognized, with the addition of a water mediated hydrogen bond between the T-2 and G76.The interactions between the CSD and T-3 are not discernible from this structure as it contacts another CSD in the asymmetric unit, occupying binding site 2. T-4 is mainly bound by stacking interactions, while T-5 is specifically hydrogen bonded by F97 and S93. Finally, T-6 is recognized through hydrogen bonds to K38 and W39 and T-7 is also stabilized by stacking interactions (**Fig. 1.19**) [123].

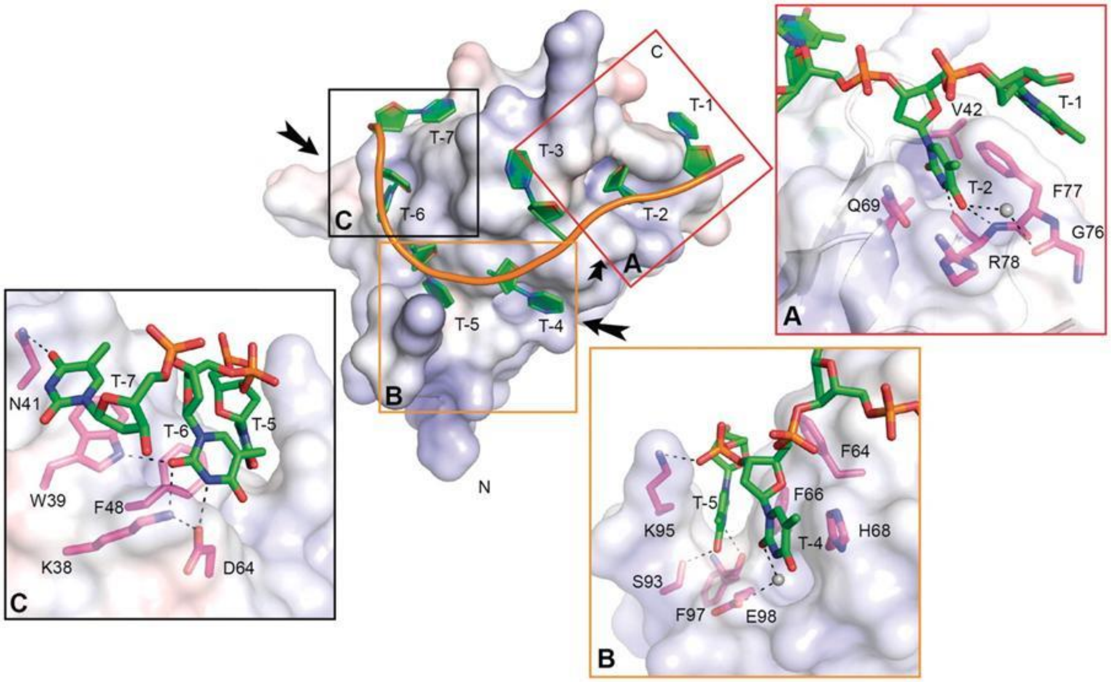


Figure 1.19: The CSD of Lin28B binds dT7.

The electostatic surface of the protein is displayed as a range from -10kT (red) to +10kT (blue), and the oligonucleotides can be seen to bind across the protein surface. A, B and C show detailed close ups of the interactions of each of the three binding subsites, where the majority of contacts are formed through hydrogen bonding and stacking interactions between protein side chains and the bases of the DNA. Adapted from Mayr et al. (2011)

The structures presented in this report, alongside other biochemical data, therefore indicated the preference of the CSD for poly-pyrimidine oligonucleotides, as well as suggesting a lack of sequence specificity for let-7 RNAs [123]. Overall, the structural data supported the model of Lin28-RNA interaction described previously [122].

1.3.2.3 Structure of the ZnK domain

In addition to the above structure, a solution structure of the human Lin28 ZnK domain bound to an AGGAGAU oligoribonucleotide was reported [124]. Here, the individual zinc knuckles comprising the ZnK domain were seen to adopt distorted β-hairpin structures, with Zinc ions co-ordinated by CCHC residues, and a proline rich linker between them. In this structure, the second zinc knuckle interacts with the 5’A and G, whereas the first zinc knuckle makes contacts with the following G and A. Both G nucleobases are recognized specifically through hydrogen bonds between the Waston-Crick edge and the backbone amide groups of residues Y140, A149 and V171, and the carbonyl groups of R138 and K160. Stacking interactions were also observed for these bases, as well as for A1 (**Fig. 1.20**). This, in combination with mutational analysis, led to the proposal that the Lin28 ZnK domain specifically recognizes a NGNNG motif present within let-7 terminal loop segments [124].

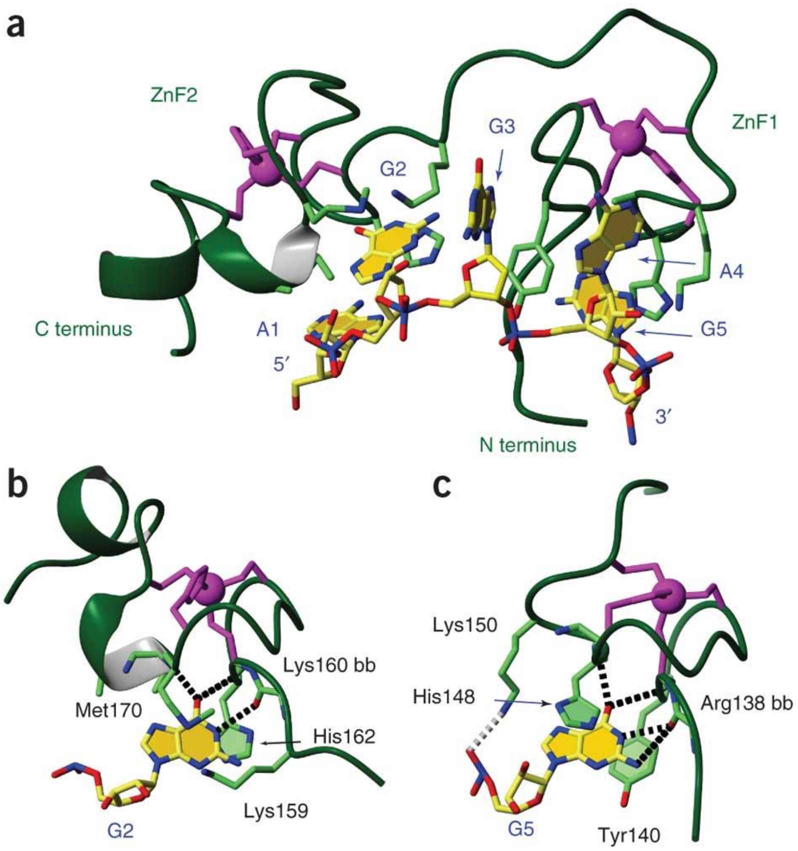


Figure 1.20: Solution structure of the Lin28 ZnK domain bound to AGGAGAU RNA.

(a) Overall structure shows the zinc ions as purple spheres. More detailed views showing hydrogen bonding networks can be seen in (b) and (c) for ZnK2 and ZnK1 respectively. Adapted from Loughlin et al. (2011)

1.3.3 Let-7 independent functions of Lin28

As mentioned in **Section 1.3.1**, several of Lin28's biological functions have been ascribed to let-7 independent pathways. Recently, there has been an influx of data on Lin28's function as an mRNA binding protein, and how it is able to influence translation, showcasing its importance as a master regulator of gene expression in the cell [78].

1.3.3.1 Translation Stabilisation

The first pieces of evidence that described Lin28's mRNA binding function came from studies that showed the association of Lin28 with polysomes in the cell, implying that Lin28 is able to associate with actively translating mRNAs [126]. Subsequently, various studies showed Lin28 binding to the mRNA transcripts of IGF-2 [113], Oct4 [127], Histone H2A [128], Cyclins A and B and cdk4 [129], and stabilizing their translation. The exact mechanism of binding is still unclear, although several reports postulate the idea of a Lin28 Response Element [127, 128, 130, 131], or LRE. Although there was no obvious consensus sequence present in the LRE's that seems to define Lin28 binding, a small structural motif, consisting of a bulged adenine nucleotide flanked by two G-C base pairs, was found to be conserved between the mRNA secondary structures, and its mutation caused a drop in the affinity of Lin28 for the Oct4 LRE, and also prevention of translation stabilization *in vivo* [130]. Once bound, Lin28 recruits RNA helicase A (RHA) which is the key factor in Lin28 mediated translational stabilization [132]. How the RHA fulfills this function is still being debated, but it could either be due to removal of mRNA secondary structures [132] or the disruption of miRNA bound RISC complexes targeting the mRNA, hence resulting in the relief of miRNA mediated translational repression [133, 134].

1.3.3.2 mRNA targets of Lin28

Recently, attempts have been made to quantify the number of Lin28 mRNA targets. The first of these observed Lin28s interaction with over 6000 RNAs in both embryonic stem cells and in somatic cells overexpressing Lin28. Intriguingly, many of these transcripts were those of RNA binding proteins, especially those involved in regulating splicing (e.g FUS/TLS, hnRNP F etc.). Further experiments confirmed that both Lin28A and Lin28B were involved directly in stabilizing the translation of these splice factor genes. Intriguingly, in contrast to the results mentioned above, this study found that the GGAG motif, known to be important in let-7 miRNA binding, was the motif that specified Lin28 binding to these mRNAs, and that Lin28 binding was enriched in both exons and 3'UTR regions of these transcripts. Additionally, the interaction of Lin28 with several non-let-7 pre-miRNAs was also observed [135].

A second study revealed more information about the Lin28A/mRNA interaction using a similar cross-linking immunoprecipitation/sequencing assay (CLIP-seq), this time by immunoprecipitating Lin28A/RNA complexes from mouse embryonic stem cells [136]. Again, Lin28 was seen to interact with a large number of mRNAs, which were determined to be the major target, as well as miRNAs, and that the GGAG motif was significantly enriched, signifying its importance as the Lin28 binding site. In contrast to the previous study however, this report observed that Lin28 was suppressing, rather than stabilizing the translation of the bound mRNAs, by preventing the association of them with the ribosome. Even more surprisingly, several pieces of evidence revealed that the majority of Lin28 targets were mRNAs targeted to the endoplasmic reticulum (ER), where the Lin28A appeared to localize. It was therefore concluded that Lin28s mRNA binding function was to suppress the translation of ER targeted mRNAs [136].

The results of a third study correlated more with those of the first. Using another CLIP-seq based technique the stabilization of mRNAs was seen as the major effect of Lin28A or Lin28B induction in HEK293K cells [137]. In this study, Lin28A was observed to bind and regulate ~1800 mRNA transcripts compared to ~3800 for Lin28B. Of these mRNA targets, the most frequently targeted were those involved in nuclear processing, cell cycle genes, splice factors and RBPs. Intriguingly, these results differed from both of the above reports as, although some Lin28 binding sites contained the GGAG motif, the most common motif was AYYHY, where Y is any pyrimidine and H is a C,U or A nucleotide [137].

Therefore, Lin28 binds and regulates the translation of a large number of mRNA targets, amongst other types of RNA. There is still, however, much debate over the exact functions, mechanisms, targets and binding sites that are recognised by Lin28.

1.4 Dihydrouridine Synthase Enzymes Bind and Modify Specific uridines in tRNA

RNA binding proteins are involved in modifying specific nucleotide positions in tRNA molecules. One such group of enzymes is known as the Dihydrouridine Synthases (Dus), which are responsible for the modification of uracil to dihydrouridine [138].

1.4.1 The dihydrouridine modification

Dihyrdouridine (see *Section 1.1.5.3*) is one of the most common tRNA modifications and is mostly present in the D-loop of tRNA, which takes its name from the abundance of this modified nucleoside [44, 49]. In bacteria, uracil bases at positions 16, 17, 20 and 20a can be modified to dihydrouridine, whereas in yeast positions 20b and 47 can also be targeted by the enzyme [44, 49].

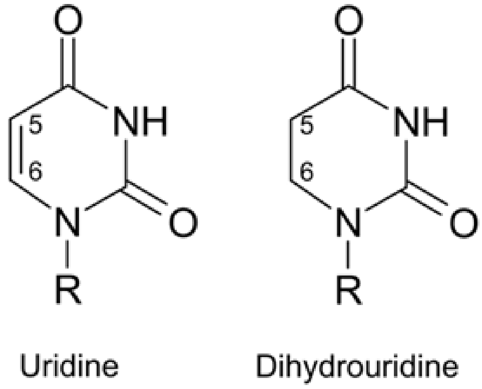


Figure 1.21: Reduction of uridine to dihydrouridine.

The modification takes place across the C5-C6 double bond, through the addition of two hydrogen atoms. (Dr. Rob Byrne, Antson group).

The conversion of uracil to dihydrouridine occurs through the reduction of the C5-C6 double bond (**Fig. 1.22**), which results in two key changes to the local tRNA structure [139]. The first of these effects is that after the reduction of this double bond, the base is no longer planar, and so can no longer take place in any stabilizing stacking interactions. Secondly, the presence of the dihydrouridine base shifts the preferred conformation of the ribose sugar from the C3’ endo conformation to the C2’ endo conformation, which is inherently more flexible [139]. The overall effect of this modification, therefore, is to increase the flexibility of the local tRNA structure, which can facilitate the formation of tertiary interactions and helps to “fine tune” the overall structure of tRNA [139]. This is in contrast to several other nucleotide modifications that increase the stability of the local tRNA structures, with the differences in abundance being seen in both thermophillic and psychrophillic bacteria, where the other modifications [140] and dihydrouridine [141] are seen in increased relative amounts respectively.

1.4.2 Dus Structure and Function

1.4.2.1 Dus Enzyme Mechanism

The function of the Dus enzymes is therefore to catalyse the reduction of the C5-C6 double bond. The mechanism for this reaction has come from both biochemical studies in yeast [142] and also from structural studies on the Dus from *T. thermophilus* [143]. In this mechanism, the key components are a conserved cysteine residue, and an FMN prosthetic group, which together define the active sites of the enzymes [142, 144]. Initially, a hydride from N5 of the reduced FMN attacks C6 of the uracil to be modified. This causes C5 to become a nucleophile and attack the distal hydride of the active site cysteine, resulting in dihydrouridine and oxidized FMN [142, 143], which must then be regenerated by NADPH [142] (**Fig. 1.23**).

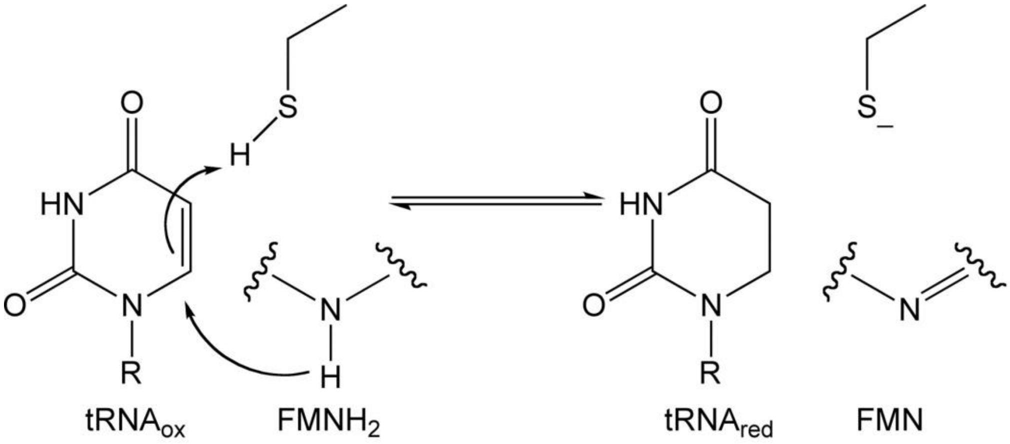


Figure 1.22: General reaction mechanism of Dus enzymes.

The hydride of N5 of the FMN attacks C6 of the active site uracil. This causes the electrons of the double bond to transfer to C5, which then attacks the distal hydride of the active site residue C93. This results in an oxidised FMN co-factor, as well as the dihydrouridine product. From Ryder et al. (2009).

1.4.2.2 The Problem of Specificity

The functional problem that the Dus enzymes face was highlighted upon the discovery that each enzyme had different, non-overlapping, modification position specificity – as in each would specifically only modify certain uracil positions in the tRNA structure [138]. In yeast, these specificities have been determined with proteins Dus1p, Dus2p, Dus 3p and Dus4p modifying positions U16/17, U20, U47 and U20a/b respectively [145]. For bacteria however, where the Dus family consists of DusA, DusB and DusC enzymes [144], only the specificity of DusA has so far been elucidated, which is for the U20 position [138]. How then do the Dus enzymes specifically recognise and modify uridines at only one or two positions of multiple tRNA molecules with divergent sequences [44, 49]?

The Dus enzymes contain a common catalytic core consisting of a TIM barrel domain. Several Dus enzymes, including the human hDus2L protein, also contain an extra dsRBD. However, although sequence alignments reveal that there are differences in the predicted secondary structures of the different enzymes, it is not possible to account for the different specificities of these enzymes from sequence alone [144].

1.4.2.3 Structures of Dus Enzymes

To date, there exist three crystal structures of Dus enzymes: the *T. maritima* Dus [146], the Dus from *T. thermophilus* [143] and the DusC enzyme from *E. coli* [147]. Of these, the *T. thermophilus (Tt)* structure is the most enlightening as structures of both the unbound and tRNA^{phe} bound states are available (**Fig. 1.24**).

The *Tt*Dus consists of the TIM barrel catalytic domain with the FMN prosthetic group bound in the centre, in addition to an extended C-terminal helical domain [143]. Examination of the surface of the protein revealed a positively charged groove, in which the D, TΨC and anticodon stems of the tRNA sit. To stabilize the complex, a hydrogen bonding network is formed between the protein and the aforementioned segments of

RNA, so that it is held in such a way so that the substrate uracil (in this case U20) is positioned into the active site to stack on top of the isoalloxanthine ring of the FMN.

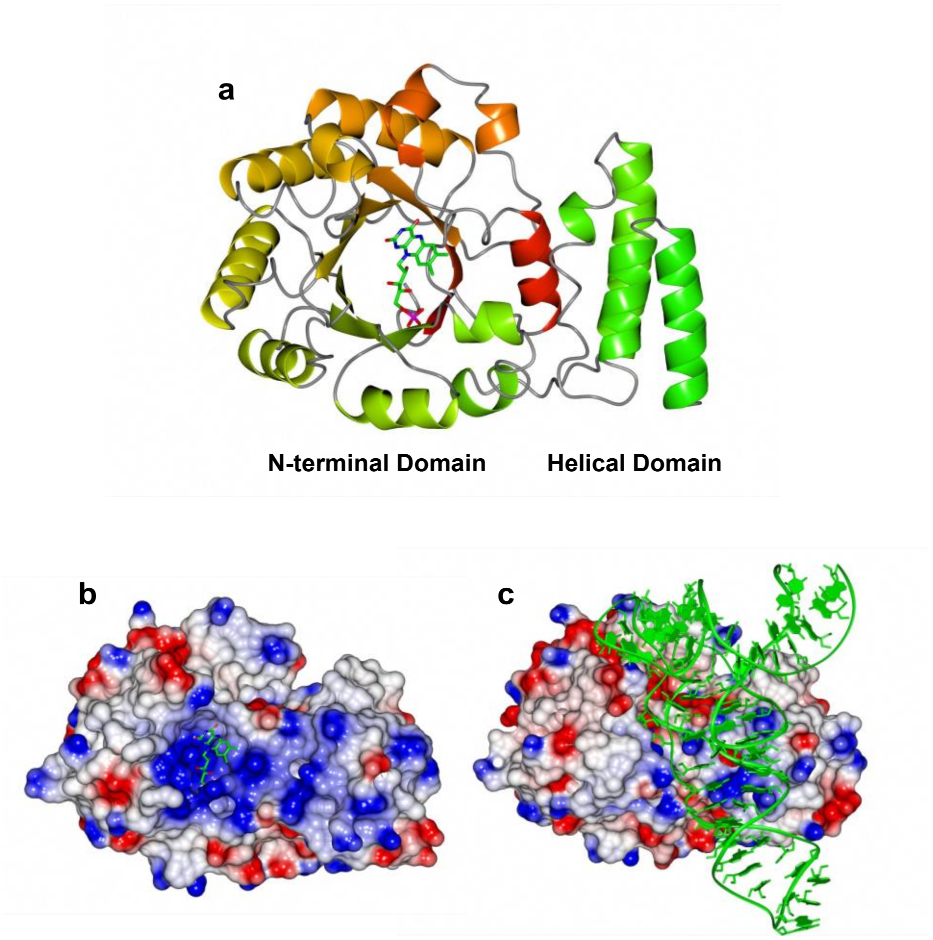


Figure 1.23: Crystal structure of *Thermus thermophilus* dihydrouridine synthase.

(a) Ribbon diagram of the unbound *Tt*Dus (PDB code: 3B0P), with the N-terminal and helical domains highlighted. FMN is shown as cylinders, coloured by atom type (b) The electrostatic surface of the *Tt*Dus reveals a positively charged groove, into which the tRNA^{phe} binds, as shown in (c, PDB code: 3B0V). The surface is coloured from red (-1V) to blue (+1V). tRNA is shown in green.

There are, however, two anomalies in this structure. First, a covalent bond is formed between the active site cysteine and the dihydrouridine of the co-purified substrate tRNA [143], which indicates that the enzyme here is inactive, and so may represent a non-native state. Second, the existence of a region of "mystery density" present in the active site, which could not be assigned to any of the buffer components [143]. This density also appeared in the structure of the *E. coli* DusC enzyme [108]. Although it was proposed that this density corresponds to an as-yet unidentified co-factor, and that it is this co-factor which is the major determinant of position specificity in the Dus enzymes, until this factor is identified and the density assigned, no definitive conclusions can be reached as to how position specificity is generated in the Dus enzymes.

1.4.2.4 Clinical Relevance of the Dus Enzymes

Whilst the biochemical function of the Dus enzymes is now known, the evolutionary advantage their presence confers to an organism is subtle, as an *E. coli* strain with all three Dus enzymes deleted showed no obvious growth defects [138]. Intriguingly though, a link between hDus2L expression and cancer has been seen in humans [75]. In this study, the increased expression of hDus was correlated with increased dihydrouridine levels, and with non-small cell lung cancer (NSCLC) tumour growth and poor prognosis. This implies that studies of Dus enzymes could have clinical significance, as hDus2 inhibitors could be potentially used to prevent or reduce tumour growth. However, at this stage it remains to be seen whether further links between Dus expression and cancer exist, and what is the exact mechanism by which Dus promotes carcinogenesis.

1.5 Outstanding Questions

1.5.1 The Structure/Function Relationship of Lin28

There are a number of unanswered questions and contradictions in the literature about the Lin28 protein. One of the major problems has been the determination of an accurate affinity of Lin28 towards let-7 miRNA. Results in the literature vary over a 10,000 fold range, from between 0.15nM [148] to 2.1µM [80] for binding of Lin28A to pre-let-7g. This problem could be related to the uncertainty about the stoichiometry of the Lin28/let-7 complex. Stoichiometries of 1:1 [122], 2:1 [123, 137, 148] and 3:1 [137] of Lin28 per let-7 RNA have been reported.

Following on from this, there is a debate about the relative importance of the two domains – chiefly as to whether they are equally important for binding [122], or whether they perform different functions, for example with the CSD as the major determinant of affinity, and the ZnK as the determinant of specificity [123]. This also feeds in to the need for more evidence for the current mechanisms of Lin28's binding to both let-7 miRNAs, and to target mRNAs. Additionally, the novel targets [92, 135-137] of Lin28 will need to be experimentally verified and characterized.

Finally, one of the key questions that remains to be answered is that, if Lin28 has such a vast array of RNA targets (mRNA, miRNA and others), how does it recognise them specifically and efficiently out of the entire transcriptome? The GGAG binding segment, which, as mentioned several times previously, appears to be critical for Lin28 binding, is rather short, and so there must be mechanisms in place in order to ensure Lin28 only specifically recognizes the RNA sequences that it must bind. As this site is so small, an additional question is how Lin28 is able to efficiently search through the transcriptome in order to find such sites within a reasonable timeframe. These are currently the most important questions that need to be resolved in order to further the understanding of the molecular mechanisms used by Lin28 to achieve its plethora of biological functions. Answers to these questions are necessary for advancing the different fields in which Lin28 fulfils a vital role, and also in aiding the development of clinically relevant inhibitors that could ameliorate the negative carcinogenic and diabetic effects of aberrant Lin28 expression.

1.5.2 Determinants of Dus Enzyme Specificity

From a structural perspective, the challenge the Dus enzymes face is huge: using the same protein fold, they recognise only one or two specific uridines on multiple tRNA molecules, where each is structurally similar, but not identical, and diverges significantly in sequence. We know so far that the Dus enzymes contain a conserved catalytic domain fold, and the catalytic mechanism appears to be conserved not only between bacterial enzymes, but between bacteria and yeast. The question that remains, therefore, is how do Dus enzymes recognise and modify specific uracil positions across tRNAs, and modify them using a conserved mechanism without changing their overall fold? The answer to this question will bring insights into how these enzymes function, which could potentially be significant in the treatment of NSCLC.

1.6 Aims

The aims of this thesis are to delineate the mechanisms that determine the specificity of the human Lin28 and *E. coli* DusC RNA binding proteins toward their target RNA sequences. The main approach used in this thesis combines X-ray crystallography with several biochemical and biophysical techniques. The results of these studies should answer some of the outstanding questions about these systems, where the generation of target specificity is of great importance for function. In addition, the misregulation of both systems has been implicated in disease, and so relevant molecular details are relevant to human health.

Chapter 2 : Materials and Methods

2.1 Molecular Cloning

2.1.1 PCR

The polymerase chain reaction (PCR) is used to amplify specific DNA sequences from a larger context. To conduct a PCR reaction, several components are needed: a thermostable DNA polymerase, an appropriate buffer solution which contains magnesium ions, the template DNA containing the sequence to be amplified, and two short DNA oligonucleotides complementary to the regions that flank the sequence to be amplified, which are called the forward and reverse primers. Primer design is of utmost importance for performing a successful PCR, as they determine where the polymerase will begin its activity. In addition, they can be designed to contain extra "adaptor" elements, for example a restriction enzyme site, which may aid in future cloning steps. Poor primer design can lead to non-specific amplifications, or no amplification at all [149].

A typical PCR has three stages. In the first stage, high temperature (≥92°C) is used to melt the double stranded DNA, producing two single strands. The next stage is repeated through multiple cycles depending on how much product is required. First, a high temperature is used to melt the dsDNA. The reaction mix is then cooled to ~5°C below the melting temperature of the primers, which allows them to anneal to the DNA strands. The selection of this temperature is critical for determining the specificity of the amplification. Finally, the reaction mix is heated to 72°C, the optimum temperature for polymerase activity, in order to extend the primers. After several of these amplification cycles, a final extension is then used in order to allow the completion of all ongoing reactions [149].

The necessity for the polymerase to be thermostable comes from the high temperatures required to denature the double stranded DNA. Two types of polymerase are used in these reactions, non-proofreading polymerases, such as *Taq* polymerase, and proofreading polymerases, such as *Pfu* polymerase. The difference here lies in whether the enzyme contains an exonuclease domain, which can sense when an incorrect nucleotide has been incorporated and remove it, resulting in higher fidelity amplification [150].

2.1.1.1 Standard PCR

In this project, standard PCRs were conducted using Phusion© proofreading polymerase, which consists of a *Pfu* type polymerase fused with a processivity enhancing domain for increased speed and fidelity [142]. Each reaction had a total volume of 20µL composed of: 20units Phusion Polymerase (N.E.B), 0.31mM dNTPs, 1-5ng/µL template DNA, 1x Phusion Buffer (N.E.B) and 0.5µM forward and reverse primers, topped up with MilliQ water. For this PCR, the following cycling conditions were used:

1. Initial denaturation, 98°C, 3min
2. Denature, 98°C, 30s
3. Anneal, 50-60°C, 45s
4. Extension, 72°C, 30s/kb to be amplified
5. Repeat steps 2-4 25 times
6. Final extension, 72°C, 10min
7. Hold at 4°C

Products were checked by agarose gel electrophoresis (see **Section 2.1.7**). To linearise the vector by PCR, the same mix and program was used, but the primers were designed to amplify the vector outside of the multiple cloning site.

2.1.1.2 Mutagenesis PCR

For mutagenesis, the above protocol for PCR by proofreading polymerase was adapted. Firstly, overlapping primers of ~30nt in length were designed for the 5’ and 3’ strands which were identical to the plasmid sequence apart from the desired mutation. In addition, 3% DMSO was added to the reaction mix. Other than these additions, the same mixture and PCR program was followed as in the general case. Before transformation, samples were incubated at 37°C for 1 hour with 1μL of 20000 units/mL Dpn1 enzyme (N.E.B), to digest template DNA.

2.1.1.3 Colony PCR

To check whether transformed bacteria contained the plasmid of interest, colony PCR was performed. Typically, the reaction mix with a total volume of 50μL comprised: 0.025μL units of DreamTaq polymerase (Fermentas), 0.25mM dNTPs, 1x DreamTaq buffer (Fermentas), one colony and 0.2μM forward and reverse primers, made up to the total volume with MilliQ water. The program used for amplification was:

1. Initial denaturation, 95°C, 3min
2. Denature, 95°C, 30s
3. Anneal, 50-60°C, 45s
4. Extension, 72°C, 1 min/kb to be amplified
5. Repeat steps 2-4 25 times
6. Final extension, 72°C, 10min
7. Hold at 4°C

Products were then checked by agarose gel electrophoresis (see **Section 2.1.7**).

2.1.2 Standard Ligation

To generate constructs, it is necessary to insert DNA sequences generate by PCR into vectors containing the factors needed for replication of the DNA and expression of the protein. This can be done either through ligation-dependent, or ligation-independent means (**Section 2.1.3**). In the ligation dependent pathway, two steps are necessary. First, both the insert and the vector (plasmid DNA in this project) must be digested with restriction enzymes in order to generate complementary “sticky ends” [151]. They must then be mixed together in an appropriate ratio, and joined through the action of a DNA ligase enzyme. The new plasmid can then be amplified by transforming bacteria. It can then be purified, and used for further applications.

2.1.2.1 Restriction Digests

Standard restriction digests were performed in 10µL total volumes. The appropriate amount of DNA was digested with 0.5µL of each restriction enzyme (5-10 units in total), supplemented with 100µg/mL of BSA and 1µL of a 10X N.E.B buffer, where each enzyme would have maximum activity. Volumes were adjusted to 10µL using MilliQ water. Once prepared, the solution was mixed using a vortexer, and incubated at 37°C for 1hr.

2.1.2.2 Ligation

Ligations were performed in a 10µL volume, with a three-fold molar excess of purified insert, against the gel purified vector (**Section 2.1.6**). DNA was ligated in 10µL reaction volumes comprised 3 units of T4 ligase enzyme and 1x T4 ligase buffer from Fermentas and the volume adjusted to the total using MilliQ water. The solution was then incubated overnight at 16°C. 5µL was then used to transform 50µL of competent *E. coli* cells (see **Section 2.1.8**).

2.1.3 In-Fusion® Cloning

In some cases, the In-Fusion® system of ligation independent cloning (Clontech) was used to prepare constructs. In this system, vector was first linearized by PCR, as described above. The vector sample was then digested for 1 hour at 37°C using 1µL of 20000 units/mL Dpn1 restriction enzyme (N.E.B) to remove template DNA. The insert was then prepared using regular PCR, with primers designed to contain an identical 15bp sequence to the linearized vector. Both insert and vector were purified using PCR purification and gel purification respectively (**Section 2.1.6**). The insert and vector sequences were then mixed in a molar ratio calculated by the Clontech In-Fusion molar ratio calculator, along with 2µL 5x In-Fusion enzyme premix from Clontech, and made up to a final volume of 10µL with MilliQ water. The mix was then heated to 50°C for 15 minutes, and 5µL used to transform competent XL-10 Gold *E. coli* cells for plasmid amplification.

2.1.4 Sequencing

Sequencing reactions were performed by GATC-Biotech. Samples were prepared to contain between 30 and 100ng total of purified plasmid DNA. Sequencing was performed in a total volume of 20µL, and using universal primers provided by the company.

2.1.5 Plasmid Purification

Transformed *E. coli* cloning strain bacteria were used to inoculate 5mL volumes of LB media, supplemented with a relevant antibiotic, and grown overnight at 37°C. The cells were harvested by centrifuging the culture at full speed in a desktop centrifuge. Plasmids were then purified using QIAGEN miniprep kits, according to the manufacturer’s protocol.

2.1.6 Gel and PCR Purifications

Purification of digested vectors from agarose gels, and of PCR products was performed using with the appropriate kits from QIAGEN, and conducted according to the manufacturer’s protocol.

2.1.7 Agarose Gel Electrophoresis

It is important to monitor the steps involved in producing a particular construct. The most effective way of doing this is to use gel electrophoresis – whereby molecules are separated on the basis of their size, shape and charge. Due to the phosphate groups of the DNA backbone, all DNA molecules will be negatively charged, and thus migrate from a negative electrode to a positive one when an electric field is applied. By placing

DNA inside a porous gel, the larger DNA fragments will migrate more slowly than the smaller fragments, which can fit more effectively inside the pores of the gel, and so the different lengths can be separated when an electric field is applied across the gel. By using an intercalating dye, such as SYBRsafe or ethidium bromide, the DNA species can then be observed using UV light. It is important to note that several factors other than size can also affect DNA migration, such as supercoiling, and these factors must be taken into consideration when analyzing results [152].

Gels were prepared by dissolving solid agarose (Sigma) to a final concentration of 1% w/v in 1x TAE buffer and adding SYBRsafe dye (Invitrogen) to a final concentration of 0.5x. The gel was then left to set at room temperature. Samples were mixed with Novagen 6x DNA Gel Loading Buffer at 1x, and loaded into the wells of the gel, which was placed in a gel tank filled with 1x TAE buffer. Electrophoresis was conducted using a constant voltage of 120V for ~45min. The results were visualized using a Syngene bioimager with transilluminator.

50x TAE buffer (1L): 242g TRIS-HCl, 100mL 0.5M EDTA (pH 8.0), 57.1mL Glacial Acetic Acid.

2.1.8 Transformation of Bacterial Cells

In order to efficiently amplify constructed plasmids, and also to allow the proteins encoded by such plasmids to be expressed, they must first be used to transform relevant *E. coli* strains which contain the additional factors necessary for either process. The heat-shock method of transformation was used in each case.

Bacterial cloning strains used in this project are: DH5α, XL1-blue and XL10-gold. Protein Expression strains used were: BL-21 (DE3) pLysS, B834 (DE3), Rosetta (DE3) pLysS and Rosetta2 (DE3). Each strain was chemically transformation competent and was thawed on ice for 15 minutes from storage at -80°C. Once thawed, 1-5µL of DNA corresponding to masses of 1-100ng, were added to the cells, which were left to incubate on ice for 30 minutes. Following this, cells were heat-shocked at 42°C for 1 minute, before being cooled on ice for a further 5 minutes. 150µL of LB medium was then added to the suspension, which was then incubated for 1hr at 37°C with shaking. The cells were then spread onto agar plates supplemented with the relevant antibiotic, and left to grow overnight at 37°C.

2.2 Protein Production and Purification

Protein expression was controlled by the lac inducible system in all cases. In this system, promoters are placed under the control of a lac operator site, to which the lac repressor protein will bind and prevent transcription (and subsequent translation). Addition of the lactose analog, IPTG (Isoproyl-β-D-thiogalactoside), which cannot be metabolized, will prevent the repressor from binding the operator, and thus de-repress the expression of the recombinant protein encoded by the plasmid [153].

The timing of induction is important for obtaining maximal yields of protein. If induced too early, such as in the lag phase, cells will focus expressing the construct rather than growth, and, if induced too late, such as in the stationary phase, cells will be limited by the resources available to them. It is therefore useful to compromise and induce during the log phase, usually when the optical density (OD) of the culture is roughly 0.6.

Some expression strains contain extra elements to increase the efficiency of protein production. If the expression strain transformed have a DE3 positive genotype, and the construct used for transformation encodes a T7 promoter upstream of the sequence to be expressed, then the T7 RNA polymerase will transcribe this sequence rather than the *E. coli* RNA polymerase. In this case, addition of IPTG induces the expression of the T7 polymerase, which in turn will transcribe the recombinant gene sequence. As the T7 polymerase is more efficient than the endogenous enzyme, the expression level of the construct will be increased. In addition, only the recombinant gene is under the control of the T7 promoter so the T7 polymerase will only transcribe this gene. Therefore, the recombinant gene does not have to compete against other sequences for transcription by the polymerase [153].

In addition to the DE3 element, some bacterial expression strains contain the pLysS plasmid, which encodes the T7 lysozyme. This inhibits T7 RNA polymerase until induction and hence provides tighter control over when constructs are expressed. Rosetta expression strains contain another plasmid that encodes tRNAs that recognise codons not frequently used by *E. coli*, and so increase the translation efficiencies of sequences containing these codons. The choice of both plasmid and expression strain are therefore key decisions when attempting to produce high yields of recombinant proteins [153].

2.2.1 Expression Testing

2.2.2 Protein Overexpression

BL-21 (DE3), pLysS, B834 (DE3), Rosetta (DE3), pLysS and Rosetta2 (DE3) strains were used to express constructs. 5mL of LB medium supplemented with a relevant antibiotic, was inoculated with a transformed colony, and grown overnight at 37°C. 60µL of overnight culture was then used to inoculate 3mL of LB medium supplemented with the relevant antibiotic. These cultures were grown at 37°C till OD_{600} ~0.6 when they were induced with IPTG at a final concentration of 1mM. After induction, the cultures were grown for either 4 hours at 37°C, or overnight at 16°C. 1mL samples of culture were taken before induction and after expression and centrifuged at maximum speed in a bench top centrifuge. The supernatant was removed, and the cell pellet resuspended in 100µL of test buffer. The sample was then sonicated using 3x 0.8s pulses at full power, and a 10µL sample taken for SDS-PAGE analysis. The remainder of the sample was centrifuged at maximum speed in a bench top centrifuge and a 10µL sample removed from the supernatant to analyse the soluble protein fraction by SDS-PAGE.

E. coli expression cells (B834/Rosetta2) were transformed with the relevant construct. Selected colonies were grown in small culture volumes (10 or 50mL) in LB media with relevant antibiotic overnight at 37°C. 10-12.5mL of overnight culture was then used to inoculate 750ml of LB media supplemented with relevant antibiotic. These larger cultures were grown to an OD_{600} of ~0.6 when they were induced with 1mM IPTG.

After induction, cultures were grown overnight at 16°C. Finally, cells were harvested by centrifugation at 5000 x g and either stored at -20°C or used immediately for purification.

2.2.3 Standard Purification Procedure

Once proteins have been overexpressed, it is necessary to purify them from endogenous *E. coli* proteins. Therefore, techniques are chosen that separate out proteins based on their chemical properties, such as size, shape and charge.

The initial stage in a purification is usually affinity chromatography, as it is an efficient way to remove many contaminants from the sample in one step. Proteins are often conjugated to a tag that will bind specifically, but reversibly, to a stationary phase. Immobilized protein can then be washed with buffer to remove weakly bound contaminants before elution. In this way, the lysate can be selectively depleted of the protein of interest, which is then further purified from any remaining contaminants by other methods, such as size exclusion chromatography (gel filtration), which retards the motion of smaller species, enabling the separation of proteins based on their size and shape.

Standard His tagged proteins were purified as follows. Cell pellets were resuspended in lysis buffer with the addition of 0.5µg/µL leupeptin, 0.7µg/µL pepstatin and 1mM AEBSF as protease inhibitors. The solution was then left at 4°C for ~20 minutes. To lyse the cells, the solution was sonicated at 16µm amplitude for 30s, before a 2-minute rest on ice to prevent heating of the sample. To clear the lysate of cellular debris, the sample was centrifuged at maximum speed in a Sorvall centrifuge, SS34 rotor, for 50 minutes. A 5mL HisTrap column (GE Healthcare) was then equilibrated using 5 column volumes (CV) of MilliQ water, followed by 2CV of Elution buffer, and 5CV lysis buffer. The lysate was then loaded onto the column using a peristaltic pump at a low flow rate and the flow through collected. The column was then attached to an Äkta FPLC purifier, and the column washed with 6CV lysis buffer. A gradient of elution buffer, from 0-100% over 11CV was then applied to the column and elution of the protein detected using UV absorbances at 260 and 280nm.

Fractions containing the protein of interest were then checked by SDS-PAGE and concentrated using Amicon concentrators with a relevant molecular weight cut-off (MWCO, usually either 10kDa or 30kDa) to a final volume of <0.5mL. A gel filtration column, containing either superdex S75 or S200 resin as the stationary phase (GE Healthcare), was then equilibrated in gel filtration buffer. The sample was applied to this column and eluting protein again detected by 260nm and 280nm UV absorption. Appropriate fractions were again tested by SDS-PAGE and concentrated, before flash-freezing small aliquots in liquid nitrogen and storing at -80°C.

To produce several of the proteins in this study, this protocol was modified. For more details, see *Chapter 3*.

Lysis Buffer: 50mM TRIS-HCl pH 7.5, 250mM NaCl, 20mM Imidazole, 2mM DTT

Elution Buffer: 50mM TRIS-HCl pH 7.5, 250mM NaCl, 500mM Imidazole, 2mM DTT

Gel Filtraion Buffer: 10mM TRIS-HCl pH 7.5, 250mM NaCl, 2mM DTT

2.2.4 SDS-PAGE

SDS-PAGE was used to monitor purification processes. Unlike DNA, proteins have no standard charge and their shapes have more variation. To overcome this problem, sodium dodecyl sulphate (SDS), along with a reducing agent such as β-mercaptoethanol or DTT, is added to the samples, which are subsequently boiled. This causes the proteins to denature, and the negatively charged SDS then associates with the polypeptide proportionally to the length of the polypeptide chain. This results in denatured, negatively charged polypeptides which can then migrate through a gel (in this case composed of polyacrylamide) towards the positive electrode when an electric field is applied. The polypeptides migrate at speeds proportional to their size and charge and so will separate out, and can be visualized by using protein sensitive dyes.

Most SDS-polyacrylamide (SDS-PAGE) gels use a discontinuous buffer system to improve resolution. A porous “stacking” gel sits atop a less porous “resolving” gel, each containing a different buffer composition, which also differs from that of the running buffer that the gel is placed into. In the buffer system used in this study, the chloride ions of the gel buffers migrate more quickly through the stacking gel, with the glycine of the running buffer trailing behind, and the protein samples in the middle of the two. At the boundary between the stacking gel and the resolving gel, the polypeptides become stacked, and are present in high concentration within a thin region. At this position, the pH of the buffer system also changes and causes the glycine to ionize and enter the resolving gel at the same speed as the chloride ions, and the two migrate together. The protein samples are then free to migrate through the stacking gel as a continuous band in a zone of constant voltage, and thus be separated based on their size [154].

Stacking gels were produced by mixing 3.2mL MilliQ water with 1.3mL stacking buffer, 0.5mL 30% acrylamide, 25µL 10% APS and 8µL TEMED, as well as 10µL 0.01% w/v Bromophenol blue for visualization of the gel. Resolving gels were made to 12.5% and were composed of 3.2mL MilliQ water, with 2.5mL resolving buffer, 4.2mL 30% acrylamide, 75µL 10% APS and 8µL TEMED. The set gels were placed in a gel tank filled with gel running buffer. Samples were mixed with SDS sample buffer, heated to 98°C for 5 minutes, centrifuged at maximum speed in a benchtop centrifuge for 1 minute and added to the gel wells by pipette. Electrophoresis was performed using a constant voltage of 200V for ~50min. Protein bands were detected by staining the resolving gel with hot coomassie blue staining solution for 1-2 minutes, before washing with water, and left in hot destain solution overnight. The gels were then imaged using a Syngene Bioimager.

4x Sample Buffer (10mL): 1.2mL 0.5M TRIS-HCl (pH 6.8), 2mL 10% SDS, 1mL 50% glycerol, 1mL 0.5M 0.1% w/v Bromophenol blue, 1mL 0.5M β-mercaptoethanol, 4.8mL deionised water.

Stacking Gel Buffer: 0.5M TRIS-HCl (pH 6.8), 0.4% SDS

Resolving Gel Buffer: 1.5M TRIS-HCl (pH 8.8), 0.4% SDS

Gel Running Buffer (500mL): 125mL 4x Running Buffer Mix, 370mL deionized water, 5mL 10% SDS

4x Running Buffer Mix (10L): 576g Glycine, 120g TRIS-HCl,

Staining Solution (1L): 250mL Propan-2-ol, 100mL Glacial acetic acid, 650mL deionised water, 2g Coomassie Brilliant Blue R dye.

Destain Solution (5L): 250mL Propan-2-ol, 350mL Glacial acetic acid.

2.2.5 Native PAGE

Native poly-acrylamide gels were made to 10% by mixing 3.3mL 30% acrylamide with 2.5mL 5x TB buffer, ~75µL 10% APS, 8µL TEMED and 4.1mL MilliQ water. The gel was paced in a gel tank filled with 0.5x TB buffer. Samples were supplemented with 10% glycerol and added into the wells by pipette. Electrophoresis was conducted by using a constant voltage of 80V for between 80 minutes and 2 hours. During this time, the gel tank and buffer system were cooled using a flow of water. To visualize nucleic acid bands, the gel was first placed in a solution containing 50mL 0.5x TB buffer supplemented with 2µL 10mg/mL Ethidium Bromide and left to stain for 20 minutes. The gel was then washed with deionized water, and imaged using UV light in a Syngene Bioimager with transilluminator. Following this, protein bands were visualized by staining the resolving gel with hot coomassie blue staining solution (see **Section 2.2.4**) for 1-2 minutes, before washing with water, and left in hot destain solution (see **Section 2.2.4**) overnight. The gels were then imaged using a Syngene Bioimager.

5x TB buffer (1L): 54g TRIS-HCl, 27.5g Boric Acid.

2.3 Biochemical Methods: Differential Scanning Fluorimetry

Differential Scanning Fluorimetry (DSF), also called the Thermofluor assay, uses changes in the fluorescence intensity emitted from a dye to measure protein unfolding. The basis for this fluorescence change is due to the change in environment of the dye. The dye of choice for these experiments is called SYPRO® orange, which has the best signal to noise ratio. Its fluorescence is quenched in an aqueous environment, but increases greatly in a hydrophobic environment. This means that, when a protein unfolds with increasing temperature and the hydrophobic core is exposed, the dye can associate with these hydrophobic regions. This results in an increase in fluorescence intensity at the maximal emission wavelength. Therefore, protein unfolding with temperature can be measured as a function of the increase in fluorescence intensity at 610nm, which is the emission maximum of the SYPRO® orange dye. The experiment can be conducted in a Q-PCR machine, which can both change the sample temperature, and conduct the fluorescence measurements. This means that samples can be prepared either in standard PCR tubes, or in 96-well plates [155].

The most useful property to obtain from these experiments is the melting temperature of the protein (T_m), which provides information about the stability of the protein as it is the point at which the concentrations of folded and unfolded protein are equal. The binding of a ligand to protein tends to cause an increase in the T_m, in a manner dependent upon the concentration of the ligand, and also the affinity of the protein for the ligand [155, 156]. Therefore, ligand binding can be measured as a function of the increase in melting temperature of the protein when ligand is added in identical conditions. It is important to note, however, that it is difficult to make comparisons of the affinity between different ligands based solely on the change in T_m, as different modes of binding will produce different ΔT_ms due to the different relative contributions of the entropic and enthalpic factors that govern the stability of the complex. Similarly, the same change in melting temperature could be obtained by different binding mechanisms. Therefore, ligands bound with similar affinity might result in different ΔT_ms, and ligands bound with different affinities might be bound with similar ΔT_ms [155].

The T_m can be defined either by the midpoint of the fitted curve, or by the inflection point. The curves produced in DSF experiments are most often fit to the Bolzmann equation (*Equation 1*) in order to determine the T_m, which in this case is equal to both the midpoint of the curve, and the inflection point. However, a better fit to the curves can be obtained by introducing an extra parameter, known as the asymmetry factor *c*, into the equation, which was then called the sigmoid-5 equation (*Equation 2*) [157]. In this equation, the T_ms derived from the inflection point and midpoint were found to be different, with the inflection point determination returning higher values. The DSF curves produced in this study were fitted to the sigmoid-5 equation [157] using the freely available MTSA program for the Matlab software package (Mathworks), and the T_ms reported here correspond to the T_m values defined by the inflection point.

Equation 1:

$$\gamma(T) = min + \frac{\max - min}{1 + e^{(\frac{Tm - T}{a})}}$$

Where *γ(T)* is the fluorescence at a particular temperature, *min* is the minimum fluorescence value and *max* is the maximum fluorescence value. T_m is the melting temperature of the protein and *a* is the Hill slope of the curve.

Equation 2:

$$\gamma(T) = min + \frac{\max - min}{(1 + e^{\left(\frac{T - x}{a}\right)})^{c}}$$

Where the parameter values are the same as in *Equation 1*, with the addition of the asymmetric operator *c*. In this case, the inflection point is defined by *Equation 3*.

Equation 3:

$$T_{\mathrm{m}} = T - a * \ln(\frac{1}{c})$$

2.4 Biochemical Methods: Fluorescence Anisotropy

Fluorescence anisotropy uses linearly polarized light to provide information on how quickly a particular fluorophore is rotating in solution. For the study of biomolecular interactions, the fluorophore is attached onto a particular component, such as a protein or synthetic RNA. Vertically polarized light, with a wavelength similar to the maximum excitation wavelength of the fluorophore, is then used to excite the fluorophore. The fluorophore will then emit polarized light in the same direction, the insensity of which can be measured by passing the emitted light through a vertical polarizer before detection. However, the fluorophore and attached molecule will rotate and tumble in solution, due to Brownian motion, and this alters the orientation of the emitted light, meaning it becomes depolarized. The property anisotropy (r) is therefore defined as the ratio of the intensity of polarized light in the vertical direction, compared to the total light intensity. The total light intensity is measured by using a vertical polarizer for excitation of the fluorophore, and a horizontal polarizer for the emission detector. Anisotropy is therefore mathetically defined in *Equation 4:*

Equation 4:

$$r = \frac{I_{\mathrm{vv}} - I_{\mathrm{vh}}}{I_{\mathrm{vv}} + 2I_{\mathrm{vh}}}$$

where I_{vv} is the emission intensity using vertical polarisers for both excitation and emission, I_{vh} is the emission intensity when using a vertical polarizer for excitation and a horizontal polarizer for emission. Here, the total light intensity is defined as $I_{vv}+2I_{vh}$ to account for the intensity of light in all three axes (x, y and z), where the two horizontal axes are symmetrical [158].

The anisotropy term is useful for biochemical applications as in a dilute, non-viscous solution, its defining property is how quickly the molecule is rotating in solution. This is itself dependent on the size and shape of the fluorophore, so that a fluorescently tagged RNA on its own will have a lower anisotropy than the same RNA when bound by a bulky protein. A complex which has a larger mass and radius will rotate more slowly in solution, so it will cause less depolarization and have a higher anisotropy reading – that is, the vertically polarized component will be larger relative to the total light intensity for the complex compared to the free fluorescent ligand. Therefore, binding can be studied as a function of the increase in the anisotropy of the fluorescent component when the concentration of the non-fluorescent component is increased [158].

There are two instrumental setups that can be used to measure anisotropy, known as the L-format (**Fig. 2.1**) and T-format (**Fig. 2.2**) setup. These are named after the paths the excitation and emission waves travel, so that in an L-format the emission intensities are measured perpendicular to the direction of the incoming excitation beam, with the T-format looking similar but with the addition of an extra detector on the opposite side of the spectrometer to the first emission detector. In the L-format, only one intensity can be measured at a time, and so the polarizers are automated to move between vertical and horizontal positions. In this case, an extra factor, called the G factor, has to be introduced to account for differences in instensity caused by the different transmission effecicences of the horizontal and vertical polarizer orientations. The G factor can be defined as such:

Equation 5:

$$G = I_{\mathrm{hv}}/I_{\mathrm{hh}}$$

so that it is calculated using the ratio of horizontally and vertically polarized light intensities when the excitation beam is horiztonally polarized. This horizontal excitation lies in the direction of measurement, and so anisotropy will not be seen and any discrepancies between the vertical and horizontal emission intensities must be due to differences between the transmission efficiencies in each orientation. For an L-format spectrometer, the anisotropy *Equation 4* must be modified to account for this variance, and so in practice is defined as shown [158]:

Equation 6:

$$r = \frac{I_{\mathrm{vv}} - GI_{\mathrm{vh}}}{I_{\mathrm{vv}} + 2GI_{\mathrm{vh}}}$$

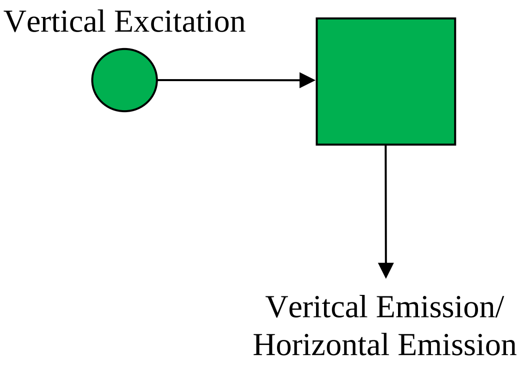


Figure 2.1 Top down view, L-format spectrometer.

The sample is shown in the cuvette in the centre. The excitation laser source is shown on the left as a circle. Light directions are depicted as arrows, and polarizer conformations noted.

For T-format spectrometers, no polarizer movement is needed, and so it is not necessary to involve the G factor. Instead, it is important to ensure that the gain on each emission detector is set accurately. The gain of each channel is therefore calibrated against the free fluorophore, and set so that the polarization (which is related to anisotropy) of the free fluorophore is equal to a previously determined value (e.g. 35mP for Fluorescein).

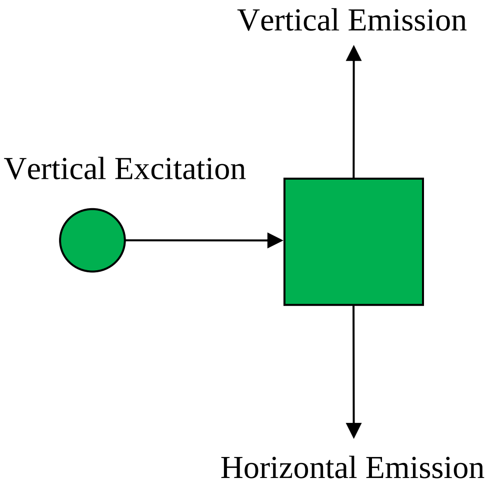


Figure 2.2: Top down view, T-format spectrometer.

The sample is shown in the cuvette in the centre. The excitation laser source is shown on the left as a circle. Light directions are depicted as arrows, and polarizer conformations noted.

2.4.1 Data Fitting

The fluorescence anisotropy experiments result in ligand binding curves where the anisotropy increases with protein concentration until the binding is saturated and maximal binding is achieved (*Bmax*). In order to describe the affinity of ligand binding, a property called the dissociation constant (*K_d*) is often calculated. The K_d of an interaction is the ratio of the product of the concentrations of the individual components of a complex to the concentration of the complex at equilibrium; and by extension is also the ratio of the rate of the reverse reaction (complex dissociation) over the forward reaction (complex formation) so that for the system:

$$[A] + [B] \underset{k_{rev}}{\overset{k_{for}}{\rightleftharpoons}} [AB]$$

where species *A* and *B* interact in a 1:1 stoichiometric ratio to produce complex *AB*,

Equation 7:
$$K_{\mathrm{d}} = \frac{[A][B]}{[AB]} = \frac{k_{\mathrm{rev}}}{k_{\mathrm{for}}} = \frac{1}{K_{\mathrm{a}}}$$

where K_a is the association constant, the reciprocal of K_d. The total amount of *A* in a particular experiment ($[A_T]$) is present in both bound *AB* and unbound *A* states so that

Equation 8:
$$[A_{\mathrm{T}}] = [A] + [AB]$$

and so the bound fraction of *A* can be defined by combining *Equations 7* and *8* to show:

Equation 9:
$$\frac{[AB]}{[A_{\mathrm{T}}]} = \frac{Ka[A][B]}{[A]+Ka[A][B]} = \frac{[B]}{Kd+[B]}$$

If the total concentration of *B* is in a large excess of *A*, $[B_{free}] \approx [B_{total}]$. In such a system therefore, the curve obtained from a binding experiment can be fit by *Equation 10*, where *y* is equal to the fraction of bound ligand.

Equation 10:
$$y = \frac{Bmax\,[x]}{K_{\mathrm{d}}+[x]}$$

where *y* is the fraction bound when the concentration of protein is equal to [*x*] and *Bmax* is value at which all free ligand is bound (where the curve plateaus) [159]. To accurately fit data from a fluorescence anisotropy experiment, an extra term must be introduced as the minimum anisotropy (that of the free fluorescent ligand), is not zero. This results in a modified form of *Equation 10*, shown below:

Equation 11: $$y = Amin + \frac{(Amax - Amin) * [x]}{K_{\mathrm{d}} + [x]}$$

The anisotropy, *y*, is directly related to the fraction of ligand bound, *Amax* is the maximum anisotropy and *Amin* the minimum anisotropy. The values of K_d, *Amax* and *Amin* can be obtained by fitting this equation to the data by non-linear regression using an appropriate software package.

One problem with *Equation 11* is that it does not take into account the difference between *[B_{total}]* and free *[B]*which may make it non-ideal for accurately fitting data and deducing K_d values from experimental data. Using the same principle as in *Equation 8*, we can substitute *[B]* for *[B_T]*. From this equation a quadratic equation can be derived which more accurately describes the system in question [159]:

Equation 12: $$[AB]^2 - ([A_{\mathrm{T}}] + [B_{\mathrm{T}}] + K_{\mathrm{d}})[AB] + [A_{\mathrm{T}}][B_{\mathrm{T}}] = 0$$

Therefore:

Equation 13: $$[AB] = \frac{\left(([A_{\mathrm{T}}]+[B_{\mathrm{T}}]+K_{\mathrm{d}}) - \sqrt{([A_{\mathrm{T}}]+[B_{\mathrm{T}}]+K_{\mathrm{d}})^2 - 4[A_{\mathrm{T}}][B_{\mathrm{T}}]}\right)}{2}$$

Similarly to *Equation 10*, this equation can then be further modified by the addition of several other parameters that take into account properties in the fluorescence anisotropy system. Initially, both sides are divided by *[A_T]*, to obtain the fraction of bound fluorescent ligand. Next, the term *Amin*, which reflects the anisotropy of the free ligand, is introduced as it is a non-zero value. This term allows the fit to start from the first data point, which will have the minimum *y* value. The whole equation is then relativized by introducing the maximum anisotropy value, *Amax*. This results in the below equation, where *[A_T]* is referred to as c, $[B_T]$ is referred to as $[x]$, and y is equal to the anisotropy at $[x]$, which is related to the fraction of bound ligand*:*

$$\textit{Equation 14}: \quad y = Amin + (Amax - Amin) * \frac{\left((c+[x]+K_{\mathrm{d}}) - \sqrt{(c+[x]+K_{\mathrm{d}})^2 - 4c[x]}\right)}{2c}$$

Equation 14 was used for the final data fitting procedures to produce accurate fits of the data. It is important to note, however, that this equation only applies to complexes with 1:1 stoichiometry, and does not take into account multiple binding sites.

2.5 Biochemical Methods: SEC-MALLS

The oligomeric state of a protein:RNA complex can be determined from the molecular mass, which is measured using size-exclusion chromatography, coupled to a multiple angle laser light scattering array (SEC-MALLS).

SEC-MALLS determines molecular weight by measuring how light responds when interacting with macromolecules. An incident light wave will scatter when encountering a macromolecule, where the intensity of the scattering at a particular scattering angle is proportional to the molecular weight of the species [160]. In addition, the intensity is proportional to concentration of scattering entities in solution, i.e. the more molecules in solution, the greater the amount of scattering [161]. As molecules move in solution by Brownian motion, scatterers that are not joined together (as in, moving together in solution) will scatter waves which are out of phase with each other, meaning the intensity is less than completely additive. Conversely, when two molecules are joined in a higher order state, such as a dimer or aggregate, their waves will be coherent and scatter in phase, making their intensities additive. Therefore, for a given concentration of macromolecule, higher order oligomeric states will scatter light to an intensity that is multiple of the intensity of the monomer, equivalent to the oligomeric state, so a dimer will have scattered light of twice the intensity as the monomer [161]. Once it is known how a species refracts light (known as the specific differential refractive index, or *dn/dc* value, 0.186 for proteins [162]), its concentration (gained by refractometry measurements) and the light scattering intensities at different angles, the molecular weight of the species can be deduced [163]. This is done by fitting the results to the Rayleight-Debye-Gans equation adapted by Zimm [160, 163] using the ASTRA software package [163].

Equation 15: $$\frac{K^*c}{R(\theta,c)} = \frac{1}{M_{\mathrm{w}}P(\theta)} + 2A_2c$$

Here, c is concentration of the species, M_w the average molecular weight of the species, $P(\theta)$ the angular dependence of the scattered light, A_2 is the second virial coefficient in the virial expansion of osmotic pressure (which accounts for non-ideal solutions) and $R(\theta,c)$ the excess Rayleigh ratio at an angle θ and concentration c, which is directly proportional to the amount of extra intensity of the light scattered by the species over the solute. K^* represents a constant *[160, 163]* defined as shown:

Equation 16: $$4\pi^2(dn/dc)^2n_0{}^2/(N_{\mathrm{a}}\lambda_0)^4$$

where N_a is Avogadro’s number, n_0 is the refractive index of the solvent and λ_0 is the wavelength of the incident light in a vacuum.

The role of the size exclusion column in SEC-MALLS experiment is therefore to separate molecules based on their size before each fraction reaches the MALLS instrumentation. This is useful as if multiple species enter the detector together, the reading is for the average molecular weight (as detailed in *Equation 17*). Therefore the elutant species are separated out into peaks containing molecules of similar sizes and shapes. The sample purity and resolution of separation are therefore of utmost importance for acquiring accurate results.

Another factor which must be considered is which *dn/dc* value is chosen for molecular weight determination [160]. Whilst there are set values for protein and nucleic acids, the exact values for protein-nucleic acid complexes will be different as they are a mix of these two species and so will have an intermediate *dn/dc* value, dependent on the relative sizes and stoichiometries of the components in the complex.

2.6 Structural Methods: X-ray Crystallography

Due to accuracy of structural information, X-ray crystallography is often the method of choice in structural biology. When molecules arranged in a crystal lattice are illuminated with an X-ray beam, the X-rays will be scattered by the electrons surrounding each atom. This is known as diffraction [164].

Scattering of X-ray photons can occur either coherently or incoherently. In coherent scattering, once the photon has been absorbed by the atom, another photon of the same wavelength and frequency is emitted. Conversely, in incoherent scattering, several lower energy photons (i.e. lower frequency, longer wavelength) are released instead. This can cause the release of energy in the molecule, which damages it [164].

In a 3D crystal lattice, the molecules are arranged in a repetitive fashion. The unit cell is the space containing the minimum number of atoms from which the entire crystal can be built by translational repetition. Moving along the crystal by the entire length of one of the edges of the unit cell will result in a unit cell that is indistinguishable from the one seen before the movement took place. The edges of the unit cell in real space are given the dimensions *a*,*b* and *c* [164].

There are therefore planes within the lattice where the points from each unit cell are equivalent, called Bragg planes. The position of these lattice planes are denoted by the miller indices *h*, *k* and *l*, one for each dimension *a*, *b* and *c*. The integer value of each index is a fraction of an edge of the unit cell, so that *h* = 1 is the lattice plane between unit cells, *h* = 2 is the lattice plane which bisects the unit cell axis *a* in half, *h* = 3 is the lattice plane that cuts the unit cell axis *a* into thirds, etc [164].

The angles at which the waves diffracted from a set of lattice planes are all in phase are related to the distance between the different planes and the wavelength of the incident beam. The Bragg equation (*Equation 17*) shows that for the scattered waves to be in phase, the path difference between the waves scattered from each lattice plane must be an integer (n) number of wavelengths (λ), and so the scattering angle θ will change depending on the distance (d) between the planes (**Fig. 2.3**) [164]:

Equation 17: $$n\lambda = 2d_{h,k,l}\sin\theta$$

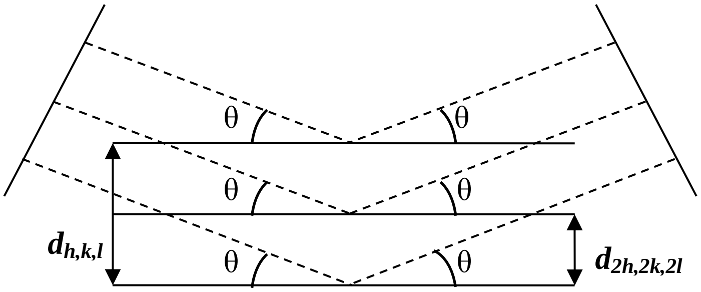


Figure 2.3: Diagram representing Bragg's law, which relates the path length and scattering angle from difference lattice planes to a whole number of wavelengths.

Here, lattice planes are spaced either *h,k,l* or *2h,2k,2l* apart. The difference in path length between the lattice planes is a whole number of wavelengths, meaning the scattering angle changes depending on the distance, *d*, between the planes, according to *Equation 17*.

Waves that diffract from atoms on the Bragg planes will scatter completely in phase, and atoms located half way between the planes will scatter completely out of phase, cancelling out the scattered beam. Atoms in other positions between the planes will scatter X-rays out of phase by an amount dependent on the distance between the atom and the nearest plane. The phase of the scattered wave therefore contains information on the position of atoms within the planes [165].

When the conditions for diffraction from a particular set of Bragg planes are met, then diffraction is multiplied, as each plane scatters in phase, and a reflection can be visualized on a detector. Reflections are referred to by the miller indices of the Bragg planes from which the wave that produced the spot diffracted [165].

The phase difference between the incident wave and the diffracted wave is dependent on the dot product of the vector describing the position of a set of electrons (*r*) relative to the origin, and the diffraction vector (*s*), which lies perpendicular to the Bragg planes and is equal to 1/*d*. This makes the phase difference equal to $2\pi s \cdot r$ (**Fig 2.4**) [165].

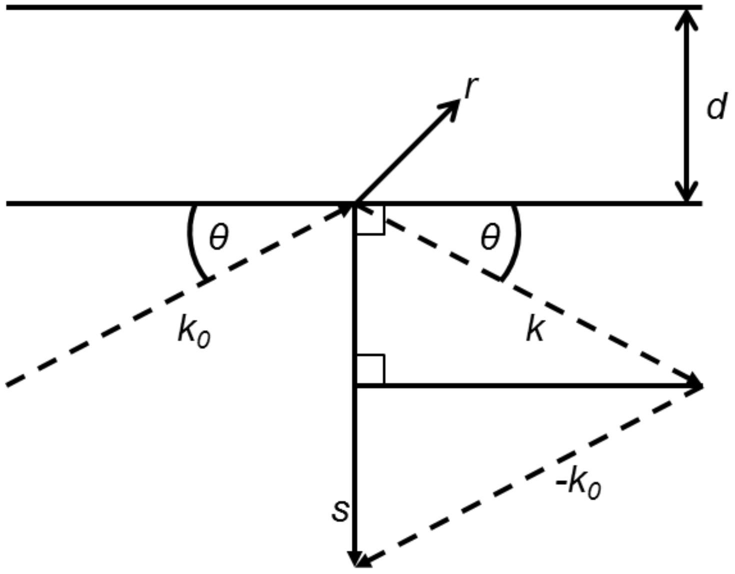


Figure 2.4: Definition of the diffraction vector, *s*.

The vector describing the incident X-ray is termed k_0, with a magnitude of 1/λ, which diffracts from the Bragg planes with an angle of θ to give a beam vector of *k* with the same magnitude. A wave diffracting from electron *r* between the planes will have the incident vector $k_0\cdot r$ and the diffracting wave vector $k\cdot r$. The phase difference between is $-2\pi(k0\cdot r - k\cdot r)$ which equals $2\pi(k - k_0)\cdot r$. If $s = k - k_0$, then the phase difference is $2\pi s\cdot r$. *s* lies perpendicular to the Bragg planes and has a magnitude of *1/d*, where *d* is the spacing between the Bragg planes.

Each diffracted wave can be described by a structure factor, which is a complex number containing information on the amplitude and phase of the wave. The structure factor that would define diffraction from a single electron, at position *r* [165] is as follows:

Equation 18: $$F_{(s)} = 1e \exp \quad [2\pi i s \cdot r]$$

where the diffracted wave has an amplitude of 1e and a phase of $2\pi is\cdot r$. For multiple electrons present at different points within a set of planes, the equation is extended as follows for *j* electrons [165]:

Equation 19: $$F_{(s)} = \sum_j exp\left[2\pi i s \cdot r_j\right]$$

By considering the positions of the electrons within the Bragg planes as a continuous function, known as electron density, the above equation can be modified [165] as such:

Equation 20: $$F_{(s)} = \int_{space} \rho_{(r)}\, exp[2\pi i s \cdot r] dr$$

This structure factor represents the fourier transform of the electron density within a set of Bragg planes [165]. In order to determine the position of electron density within a unit cell, it is useful to think in terms of reciprocal space [166]. The unit cell in reciprocal space is defined by axes taking the form of the three diffraction vectors that result from the Bragg planes in real space. $a*$ is perpendicular to planes b and c, $b*$ is perpendicular to ac and $c*$ is perpendicular to ab. The unit cell axis $a*$ is the diffraction vector from the 1 0 0 planes and its length is the reciprocal of the distance between each set of *bc* planes. In reciprocal space, the diffraction vector, *s* [166], is defined as:

Equation 21: $$s = ha^* + kb^* + lc^*$$

Where *h*, *k* and *l* are the miller indices of the Bragg planes. In this system, for the reflection *h,k,l* = 1 0 0, *s* = $a*$, etc. In terms of real space, a position in the unit cell can be defined by fractions of the unit cell axes in three dimensions [166], as follows:

Equation 22: $$r = xa + yb + zc$$

where *x*, *y* and *z* are fractional distances and *a*, *b* and *c* are the real space axes of the unit cell. It follows that the vector product of *s* and *r* [166] is:

Equation 23: $$s \cdot r = (ha^* + kb^* + lc^*) \cdot (xa + yb + zc)$$

This can be simplified [166] to:

Equation 24: $$s \cdot r = hx + ky + lz$$

By incorporating this into the structure factor equation above (*Equation 20*), the relationship between electron density and its position in terms of fractional unit cell coordinates can be established [166]:

Equation 25: $$F_{(hkl)} = \int_{cell} \rho_{(x,y,z)}\, exp[2\pi i(hx + ky + lz)]dx, y, z$$

The structure factor can also be described in atomic terms with the atomic scattering factor, f [166]. This factor describes the scattering amplitude of different atoms and is dependent on the Bragg angle of the X-ray beam, and the atomic number of the atom [167]. The structure factor in atomic terms is described below for the j^{th} atom [166]:

Equation 26: $$F_{(hkl)} = \sum_j f_j \exp \quad [2\pi i\left(hx_j + ky_j + lz_j\right)]$$

Simplifying this equation, for a reflection *hkl,* the structure factor is described [164] as:

Equation 27: $$F_{\mathrm{hkl}} = |F_{hkl}|\exp \quad [i\Phi_{hkl}]$$

where $|F_{hkl}|$ is the amplitude of the wave and Φ_{hkl} is the phase. The distribution of electron density across the unit cell can then be determined by the summation of all observed structure factors [164] as follows:

Equation 28: $$\rho(x, y, z) = \frac{1}{V}\sum_{hkl} F_{hkl} exp \quad [-2\pi i(hx + ky + lz)]$$

where V is the volume of the unit cell. Electron density is calculated for each coordinate, defined by X,Y and Z, in the unit cell, These coordinates are defined in *Equation 31* [164]:

Equation 29: $$X = xa \quad Y = yb \quad Z = zc$$

However, due to the position of the crystal relative to the beam, not all planes will be in the diffracting position. It is therefore necessary to rotate the crystal in order to measure the intensity of all possible reflections so that the electron density map can be calculated, and 3D model built [164]. As all unit cells that are being exposed contribute to the diffraction, the few molecules that have radiation damage and regions of disorder will be averaged out by the intact ordered regions within the crystal (providing they are in excess).

2.6.1 The Phase Problem

The major problem facing crystallographers is that information about relative phases of individual reflections cannot be experimentally measured. Therefore, phases must be determined by alternate methods in order to produce an electron density map. This can be done either by using the anomalous scattering of heavy metal atoms, or by using information from similar protein molecules for which structures have already been determined, in a process known as molecular replacement [164].

Once an estimate of the phases has been obtained, a model can be built. However, as the model is an interpretation of the electron density map, it may not be entirely accurate and must be checked and corrected. Therefore, an inverse Fourier transform is calculated from the model, so that the calculated structure factor amplitudes (from the model, $|F_{calc}|$, $|F_c|$) can be compared to those derived experimentally from the diffraction experiment ($|F_{obs}|$, $|F_o|$). There are two types of electron density maps used for model building: the F_o-F_c map, known as the difference density, which is used for identifying features observed but not present in the model (or vice-versa) and the $2F_o$-F_c map which is calculated using experimental amplitudes $|F_o|$ with the addition of the F_o-F_c difference for reducing bias. This is necessary as phase information is derived from the model, rather than the experimental data, and so the $2F_o$-F_c map is used to strengthen the effect of the observed data. In practice, these maps are calculated with the use of additional weighting terms, to further reduce bias from model derived phases [164].

Once a first model has been built, it must be refined. During this process positions of the atoms in the structure are altered to minimize the difference between F_o and F_c while at the same time minimising difference between model’s geometrical parameters (such as interatomic distances and angles) and their ideal values. This then further improves the calculated phases, and hence the electron density map, which in turn allows a more accurate model to be built. The progress of refinement is monitored by calculating R (*Equation 30*) and R*free* (*Equation 31*) factors [164], shown below:

Equation 30:

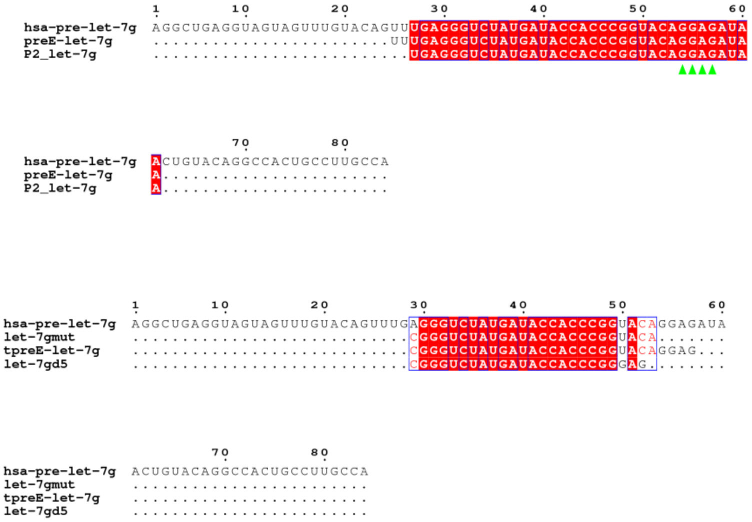
$$R = \frac{\sum_{hkl}\left||F_{\mathrm{o}}| - |F_{\mathrm{c}}|\right|}{\sum_{hkl}|F_{\mathrm{o}}|}$$

Equation 31:

$$R_{free} = \frac{\sum_{test\ set}\left||F_{\mathrm{o}}| - |F_{\mathrm{c}}|\right|}{\sum_{test\ set}|F_{\mathrm{o}}|}$$

The R factor represents the average difference between the observed and calculated structure factor amplitudes as a fraction of an average amplitude. This factor gives an estimate of the total amount of error in the model. If the model was completely random, the R-factor would be expected to be higher than 0.5. Fully refined models usually have R factors of <0.25. A problem with the R factor is that it does not take the bias of the model into account, in that, although the refinement process minimizes differences between Fo and Fc, it does not directly minimize difference between the true and calculated positions of atoms. The R_{free} factor is therefore a more useful measure for validating models. Typically ~1000 reflections are omitted from the refinement so that they do not participate in the process of minimizing the Fo-Fc difference. These omitted reflections are used exclusively for calculating the R_{free}. The R_{free} can therefore be used as an unbiased indicator of the validity of the refinement process [164].

2.7 RNA sequences and secondary structures



```
                 1        10        20        30        40        50        60
hsa-pre-let-7g   AGGCUGAGGUAGUAGUUUGUACAGUUUGAGGGUCUAUGAUACCACCCGGUACAGGAGAUA
preE-let-7g      ........................UUUGAGGGUCUAUGAUACCACCCGGUACAGGAGAUA
P2_let-7g        ..........................UGAGGGUCUAUGAUACCACCCGGUACAGGAGAUA

                          70        80
hsa-pre-let-7g   ACUGUACAGGCCACUGCCUUGCCA
preE-let-7g      A.......................
P2_let-7g        A.......................


                 1        10        20        30        40        50        60
hsa-pre-let-7g   AGGCUGAGGUAGUAGUUUGUACAGUUUGAGGGUCUAUGAUACCACCCGGUACAGGAGAUA
let-7gmut        ............................CGGGUCUAUGAUACCACCCGGUACA.......
tpreE-let-7g     ............................CGGGUCUAUGAUACCACCCGGUACAGGAG...
let-7gd5         ............................CGGGUCUAUGAUACCACCCGGGAG........

                          70        80
hsa-pre-let-7g   ACUGUACAGGCCACUGCCUUGCCA
let-7gmut        ........................
tpreE-let-7g     ........................
let-7gd5         ........................
```

Figure 2.5: Sequence alignments of the full human pre-let-7g sequence with the sequences used in this study.

Fully conserved nucleotides are highlighted in red. The conserved GGAG motif is highlighted by green triangles in the first alignment.

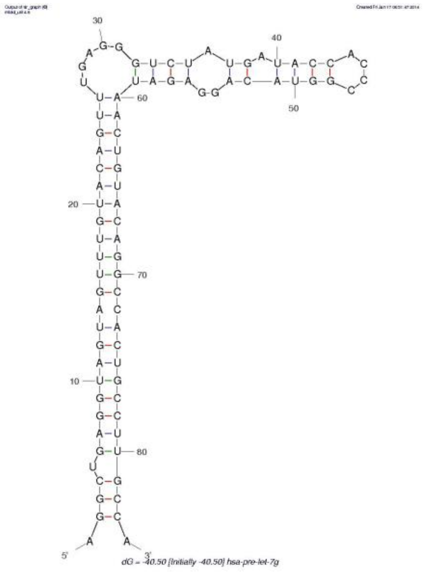


Figure 2.6: Lowest energy secondary structure of human pre-let-7g as predicted by MFOLD.

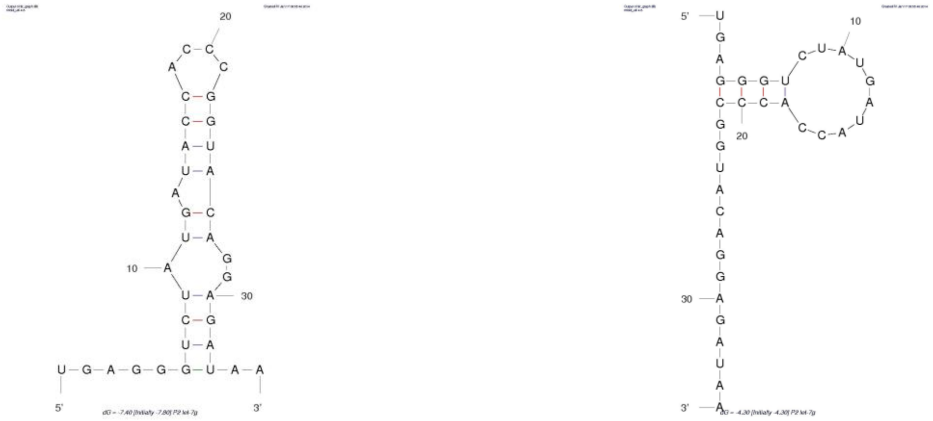

5'
3'
dG = -7.40 [Initially -7.80] P2 let-7g
dG = -4.30 [Initially -4.30] P2 let-7g

Figure 2.7: Two lowest energy secondary structures of P2 let-7g as predicted by MFOLD.

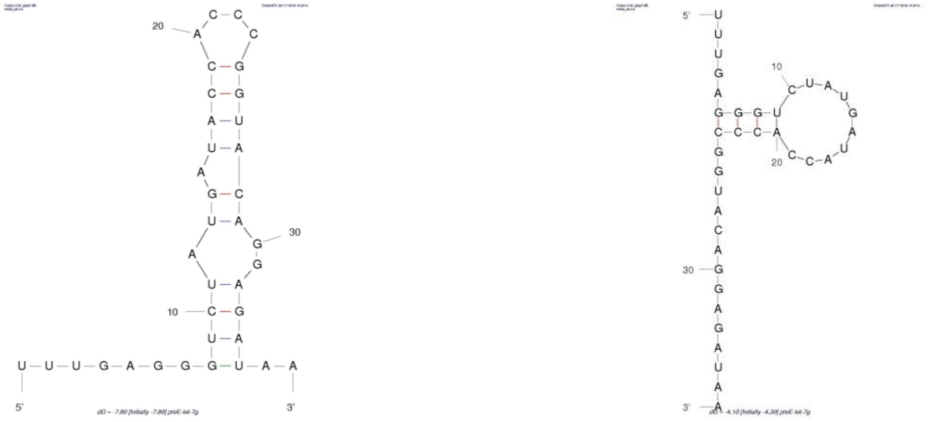

5'
3'
dG = -7.80 [Initially -7.80] preE-let-7g
dG = -4.10 [Initially -4.30] preE-let-7g

Figure 2.8: Two lowest energy secondary structures of preE-let-7g as predicted by MFOLD.

The fluorescent variant of this sequence had a 5' conjugated fluorescein fluorophore.

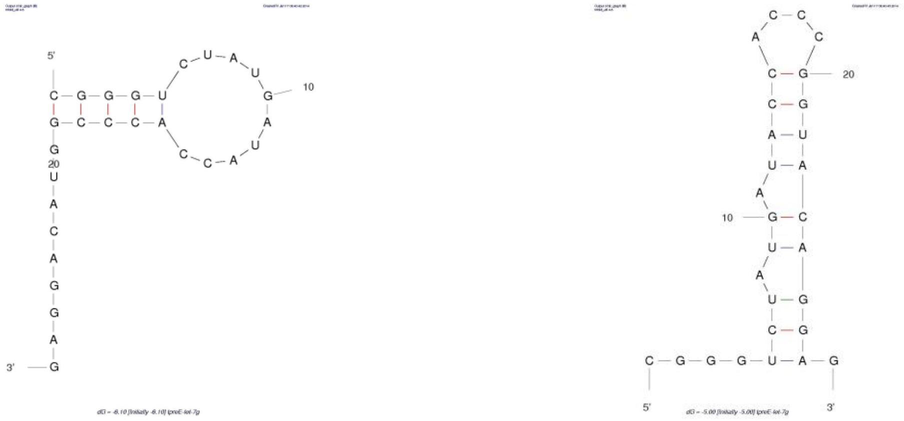

5'
3'
dG = -6.10 [Initially -6.10] tpreE-let-7g
dG = -5.00 [Initially -5.00] tpreE-let-7g

Figure 2.9: Two lowest energy secondary structures of tpreE-let-7g as predicted by MFOLD. The C of the 5' terminus was mutated from a G present in the wild-type sequence to strengthen the stem region.

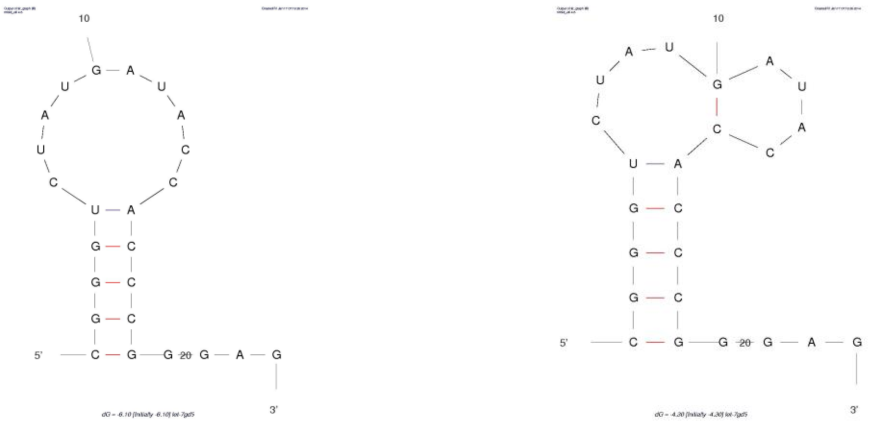

dG = -6.10 [Initially -6.10] let-7gd5
dG = -4.30 [Initially -4.30] let-7gd5

Figure 2.10: Two lowest energy secondary structures of let-7gΔ5 as predicted by MFOLD.

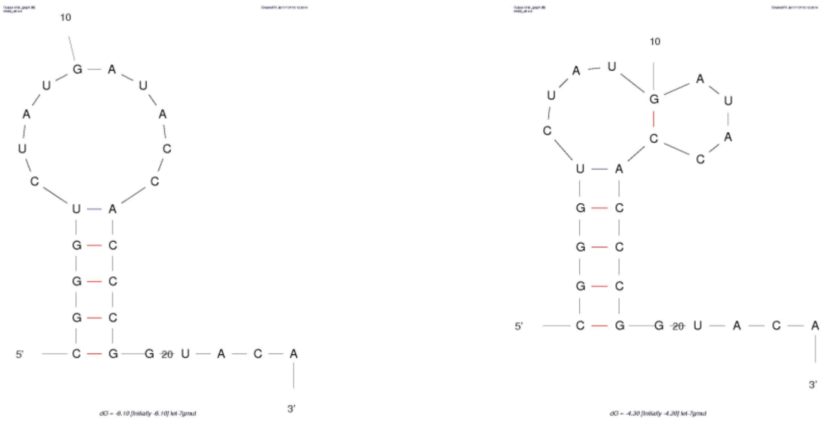

dG = -6.10 [Initially -6.10] let-7gmut
dG = -4.30 [Initially -4.30] let-7gmut

Figure 2.11: Two lowest energy secondary structures of let-7gmut as predicted by MFOLD.

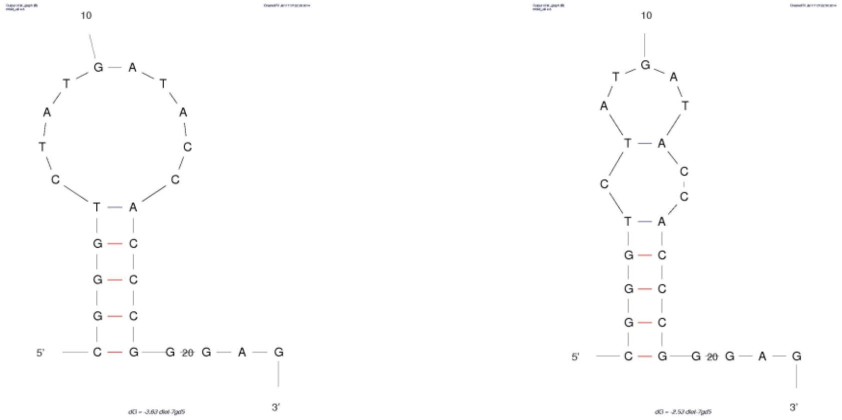


Figure 2.12: Two lowest energy secondary structures of dlet-7gΔ5 as predicted by MFOLD.

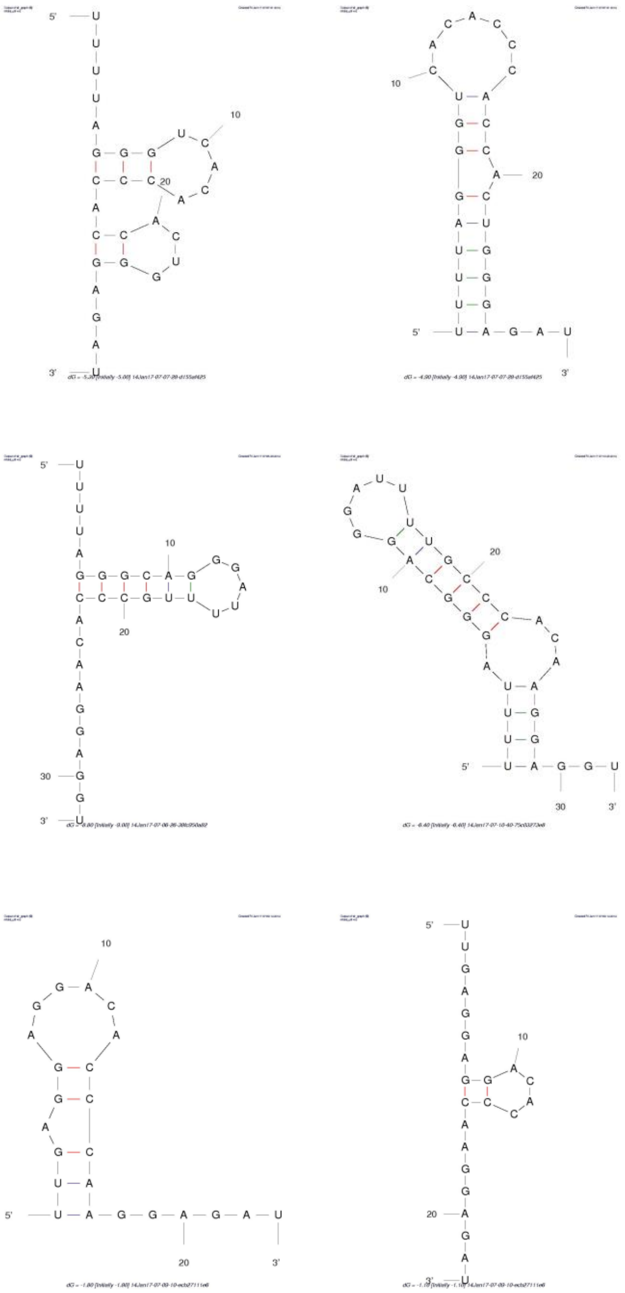

dG = -5.30 [initially -5.00] 14Jan17-07-07-29-d155af425
dG = -4.90 [initially -4.90] 14Jan17-07-07-29-d155af425
dG = -9.00 [initially -9.00] 14Jan17-07-06-26-39fc950a92
dG = -6.40 [initially -6.40] 14Jan17-07-10-40-75c03273e8
dG = -1.80 [initially -1.80] 14Jan17-07-09-10-ecb27111e6
dG = -1.10 [initially -1.10] 14Jan17-07-09-10-ecb27111e6

Figure 2.13: Secondary structures of let-7-A1 (top), let-7d (middle) and let-7e (bottom) oligonucleotides used in this study as predicted by MFOLD. Each is derived from the human pre-miRNA sequence and the two lowest energy structures are shown for each.

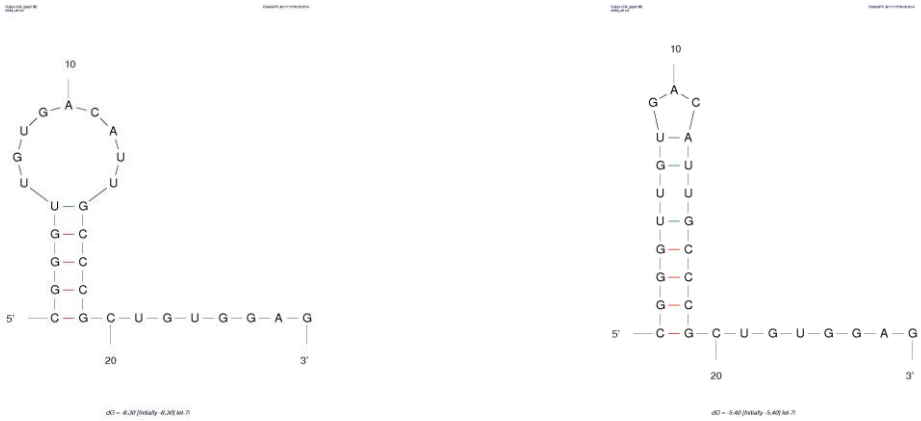
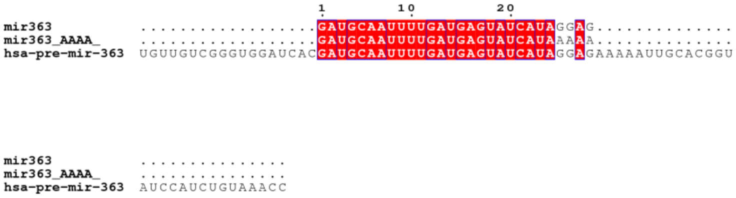

10
G
A
C
U
A
G
U
U
U
U
G
G
C
G
C
G
C
5'
C
G
C
U
G
U
G
G
A
G
20
3'
dG = -6.30 [Initially -6.30] let-7i
10
G
A
C
U
A
G
U
U
U
U
G
G
C
G
C
G
C
5'
C
G
C
U
G
U
G
G
A
G
20
3'
dG = -5.40 [Initially -5.40] let-7i

Figure 2.14: Two lowest energy secondary structures of let-7i as predicted by MFOLD.



```
                                     1        10        20
mir363            ..................GAUGCAAUUUUGAUGAGUAUCAUAGGAG..............
mir363_AAAA_      ..................GAUGCAAUUUUGAUGAGUAUCAUAAAAA..............
hsa-pre-mir-363   UGUUGUCGGGUGGAUCACGAUGCAAUUUUGAUGAGUAUCAUAGGAGAAAAAUUGCACGGU


mir363            ...............
mir363_AAAA_      ...............
hsa-pre-mir-363   AUCCAUCUGUAAACC
```

Figure 2.15: Sequence alignment of the full human pre-mir-363 sequence with the sequences used in this study. Fully conserved nucleotides are highlighted in red.

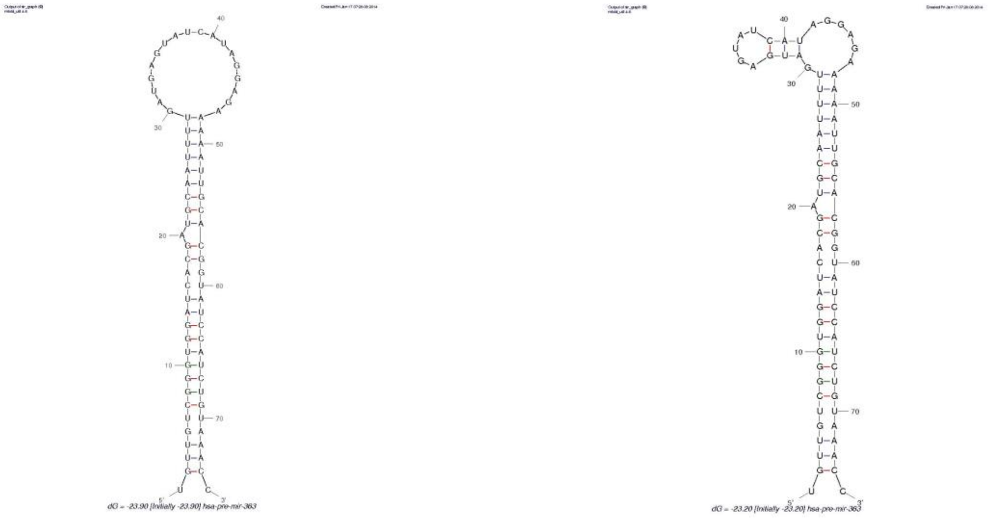

40
30
20
10
50
60
70
5'
3'
dG = -23.90 [Initially -23.90] hsa-pre-mir-363
dG = -23.20 [Initially -23.20] hsa-pre-mir-363

Figure 2.16: Two lowest energy secondary structure of human pre-mir-363 as predicted by MFOLD.

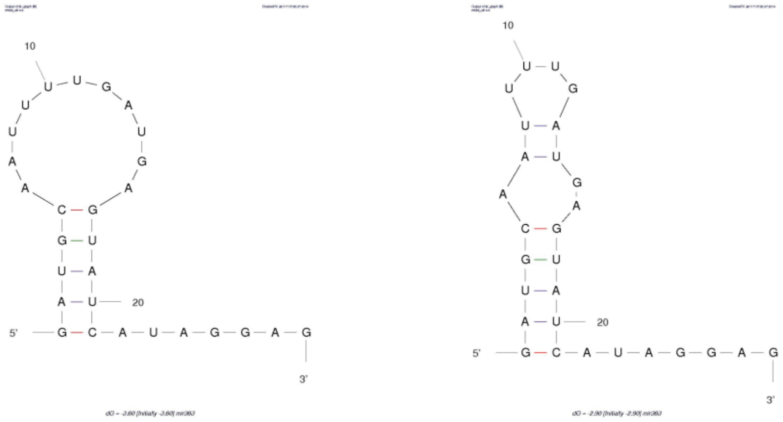

10
20
5'
3'
dG = -3.60 [Initially -3.60] mir363
dG = -2.90 [Initially -2.90] mir363

Figure 2.17: Two lowest energy secondary structures of mir363 as predicted by MFOLD.

The fluorescent variant of this sequence had a 5' conjugated fluorescein fluorophore.

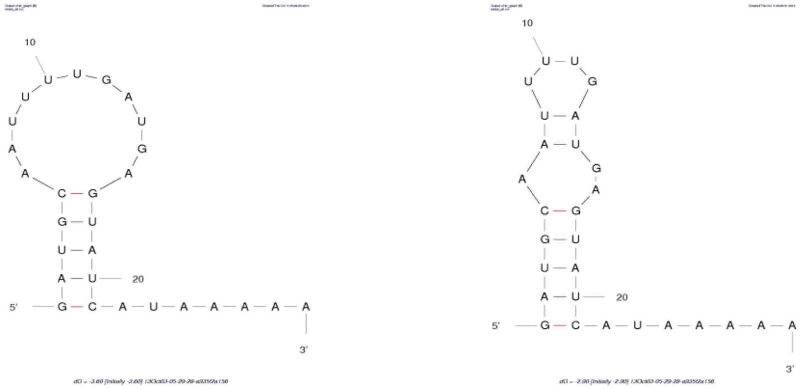

10
20
5'
3'
dG = -3.60 [Initially -3.60] 130ct03-05-29-28-a93592a156
dG = -2.90 [Initially -2.90] 130ct03-05-29-28-a93592a156

Figure 2.18: Two lowest energy secondary structures of mir363 as predicted by MFOLD.

The fluorescent variant of this sequence had a 5' conjugated fluorescein fluorophore.

Chapter 3 : Preparation and Purification of Stable Recombinant Lin28 Proteins

3.1 Introduction

To answer the outstanding questions about the structure and function of Lin28, and to understand molecular interactions of Lin28 with RNA, structural and biochemical data are needed. These approaches require milligram quantities of protein. The aim of this chapter is to describe the development of methods that enable the production of Lin28 protein in large quantities, for use in biochemical assay systems, as well as for the determination of the structure of Lin28/nucleic acid complexes by X-ray crystallography.

3.1.1 Recombinant Lin28 proteins aggregate due to non-specific DNA contamination

This section briefly describes experiments performed by Dr. Oleg Kovalevskiey (Antson group, unpublished data), revealing the challenges of producing recombinant Lin28 protein. Initially, Lin28 was expressed from a standard pET28(a) vector in *E. coli* cells and purified by Ni^{2+} immobilized metal affinity chromatography (IMAC). Lin28 fractions were heavily contaminated with nucleic acid, as determined by the high A_{260}/A_{280} UV absorbance ratio, and eluted in the void volume of a gel filtration column, which demonstrated the protein was forming high molecular weight aggregates complexed non-specifically with nucleic acid. The same effect was seen with Lin28TT truncated termini protein (residues 37-180). The CSD (residues 37-113) could not be produced in a soluble form. In contrast the ZnK domain (residues 137-180) did not aggregate, and could be purified by Ni^{2+} IMAC/gel filtration. This implies the aggregation seen with the full length and truncated Lin28 proteins occurs through the CSD. However, gel shift analysis showed no binding of the ZnK to the let-7g terminal loop segment. In addition, it could not be determined if the Zn^{2+} ions responsible for the fold of the domain were still bound, and so this protein was not deemed suitable for further study.

Therefore, in order to purify Lin28, nucleic acid would have to be removed. Multiple buffer conditions were tested; the concentration of NaCl in the IMAC wash buffer was varied, different reducing agents and detergent were added, but nucleic acid remained bound in each case. It was then attempted to remove nucleic acid by denaturing and refolding the protein. However, dialysis of the purified Lin28 aggregates against buffers containing stepwise dilutions of 8M urea resulted in a final protein sample that could not bind the let-7g terminal loop RNA, suggesting the protein did not refold correctly.

To test whether the bound nucleic acid was RNA or DNA, aggregates were probed by digestion with RNase free DNaseI, and samples analysed by agarose gel. A smear, rather than the high molecular weight band previously seen was observed only following digestion with DNaseI. This suggested the nucleic acid causing the aggregation was DNA rather than RNA.

Lin28TT aggregates were then immobilized on Ni^{2+} beads, and incubated with DNaseI. Almost no Lin28 protein was eluted, indicating that it is not stable without the presence of bound nucleic acid. New strategies were therefore needed to produce sufficient quantities of pure Lin28, which would have to take into account three factors: the removal of non-specifically bound nucleic acid, the stability of the protein in the absence of nucleic, and the maintenance of the integrity of the ZnK domain, which could lose bound Zn^{2+} ions, potentially leading to unfolding.

3.1.2 Large affinity tag fusions can aid protein production

His-tags are often used to purify recombinant proteins, as they allow simple IMAC purification, and, due to their small size, do not tend to interfere with biochemical assays or crystallization [168]. Therefore, it is not always necessary to cleave the tag after purification. This is advantageous as cleavage can be challenging [169]. In contrast, large affinity tags such as glutathione-S-transferase (GST) and maltose binding protein (MBP) are often cleaved, as the flexible linker that connects the protein and tag can interfere with crystal growth, or, in the case of an NMR approach, make the protein too large for structural studies [168].

The major advantage of fusion to a large affinity tag is the increase in solubility conferred to the passenger protein when conjugated to the C terminus of the tag. It has been suggested this effect is due to the recruitment of chaperones to the affinity tag as it is translated, which subsequently interact with the passenger protein, stabilizing its translation [168, 169]. Large affinity tags have therefore been demonstrated to be effective for producing soluble, active passenger proteins [170, 171].

In addition, large affinity tags can assist in obtaining protein crystals for structural analysis. The structures of MBP and GST are known, and it is possible that replicating the conditions used in their crystallization could also drive the crystallization of the passenger protein, providing it is small enough not to hinder the crystallization process. Additionally, the availability of these structures allows fusion protein structures to be solved by molecular replacement. In order to take advantage of these properties however, the interprotein linker must first be removed or made ridged in order to limit the flexibility of the fusion protein, as otherwise this could interfere with crystallization [168].

3.1.3 GST Fusion Proteins

The GST affinity tag sequence is from the parasite *Schistosoma japonicum.* The biological function of GST protein is to attach the glutathione tripeptide (GSH) to electrophilic toxins, preventing damage to the organism. The 2.4Å structure of the protein has been determined (**Fig. 3.1a**), and reveals two subunits of a short αβ N terminal domain linked to a larger α helical C terminal domain through a 6-residue linker region, arranged in a dimer. The dimer interface is formed by a narrow, 40 Å groove lined with polar residues and dominated by two salt bridges between D77/R89 and E51/R136. Several hydrophobic contacts also stabilize this interaction [172].

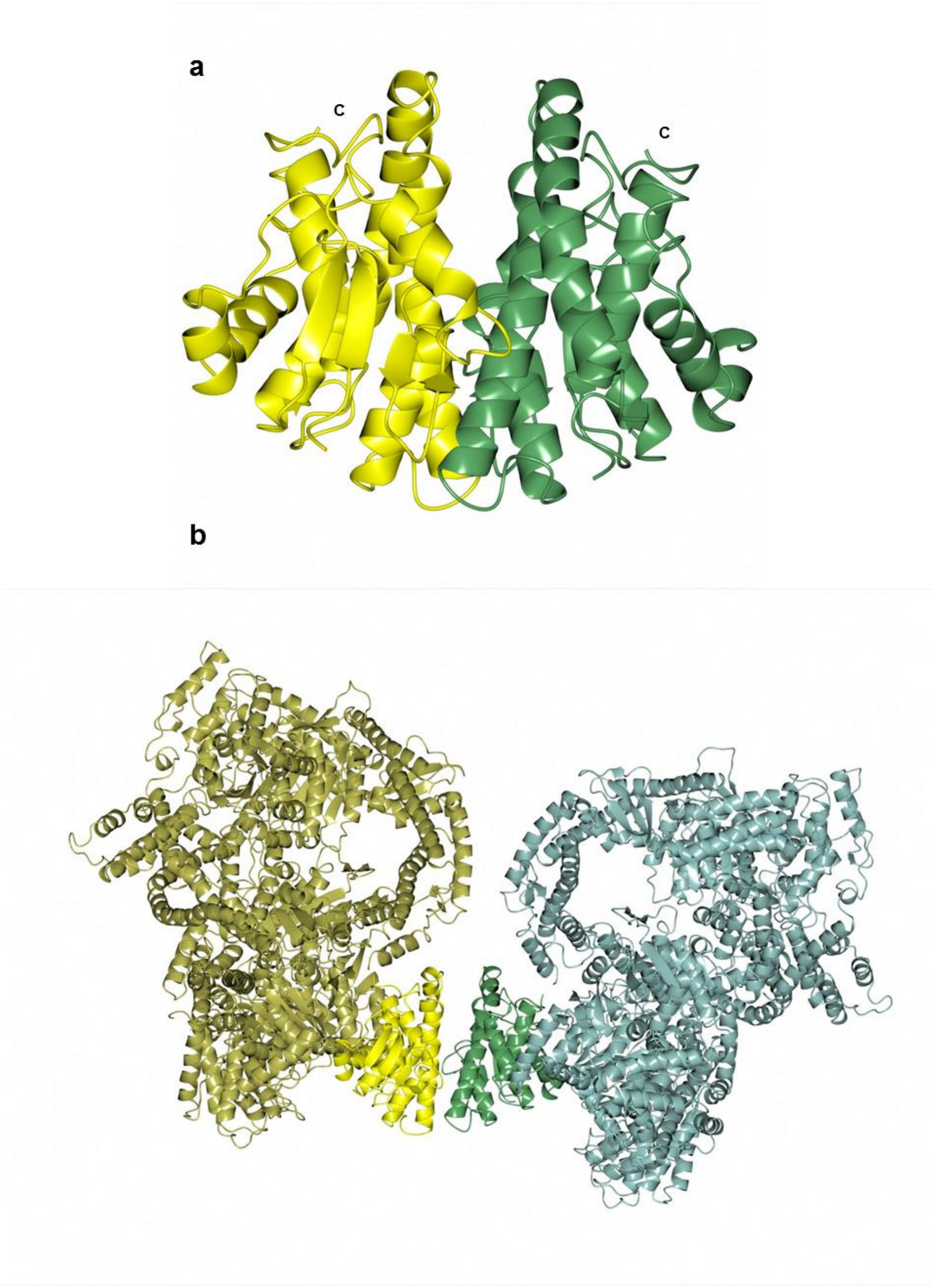


Figure 3.1: Ribbon diagrams of GST protein alone and a GST fusion with another protein.

(a) The structure of GST from *S.japonicum,* The monomers of the dimer are shown in yellow and green. The C termini used for fusions are highlighted. PDB code: 1UA5. (b) Structure of a GST fusion with a dynein motor domain. The GST dimer is coloured as in (a), with the fused C terminal dynein domains coloured in gold and sea-green. PDB code: 4AKG.

The majority of GST fusion protein structures are fusions with small domains, as these can fit into the gaps between adjacent GST molecules in the crystal [173], but the structure of the motor domain of dynein fused to GST demonstrates that this technique could also aid the crystallisation of larger proteins (**Fig. 3.1b**) [174]. Fusing Lin28 or its domains to GST could therefore be advantageous. Fusion proteins could be purified by glutathione affinity chromatography, which would eliminate the possibility of Ni^{2+} exchanging with the bound Zn^{2+} ions of the ZnK domain during IMAC, leading to the unfolding of the domain. Additionally, the increased solubility conferred by GST may assist in keeping the proteins in solution after removal of nucleic acids. Finally, GST may act as a driver of crystallization. Furthermore, the structure could be solved by molecular replacement with GST.

3.1.4 MBP Fusion Proteins

Compared with other affinity tags, MBP has been demonstrated to have a better success rate in solubilizing challenging proteins [170, 175] and as a driver of crystallization [168]. *E. coli* MBP is comprised of two globular domains (**Fig. 3.2**): the N terminal domain, and the C terminal domain, which is subdivided into the C1 and C2 domains. Both the N and C terminal domains have a central core composed of a 5-stranded β sheet with two α helices on one side and three on the other. The two domains are separated by a groove, and residues from both domains are required to form the maltose binding site. In contrast to GST, MBP is monomeric [176].

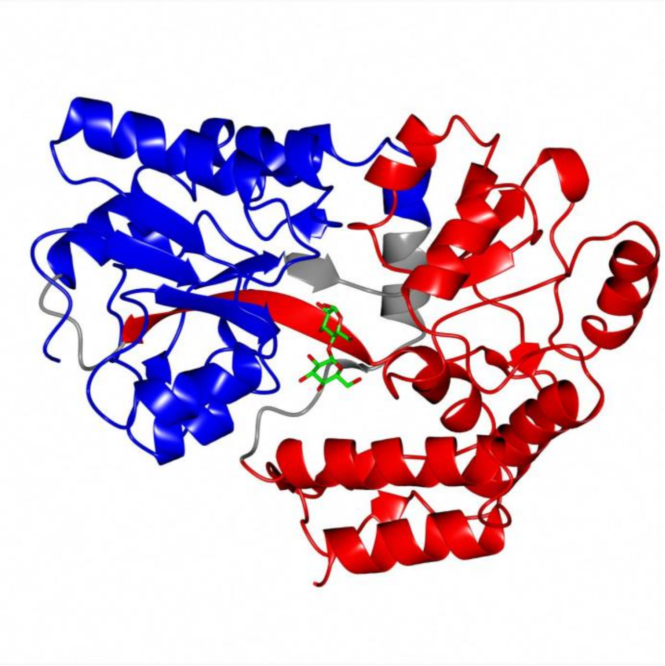


Figure 3.2: Ribbon diagram of *E. coli* MBP with bound maltose.

The N-terminal domain is in blue, the C-terminal domain is in red, and other residues are in grey. Bound maltose (sticks) is coloured by atom. PDB code: 1ANF.

The first structure of an MBP fusion protein was of the ectodomain of the gp21 protein, a transmembrane protein involved in human T-cell leukaemia virus type 1 (HTLV1) pathogenicity [177]. The purification and crystallization of gp21 had proven challenging, as it had low solubility when expressed in *E. coli*. Fusion of the gp21 ectodomain with MBP resulted in large quantities of soluble protein. However, crystals only grew when the interdomain linker was truncated to three alanine residues, which were chosen to stabilize the connection between the MBP C-terminal α-helix and the initial helix of gp21. In addition, several charged residues at the MBP C-terminus were mutated, to avoid charge-charge repulsion upon trimerization of gp21 [178]. From these crystals, the structure was determined to a resolution of 2.5Å by molecular replacement with MBP [177] (**Fig. 3.3a**). Importantly, the oligomerisation of gp21 was not inhibited by this fusion [177, 178] so biologically relevant conclusions could be made from the structure of the fused construct.

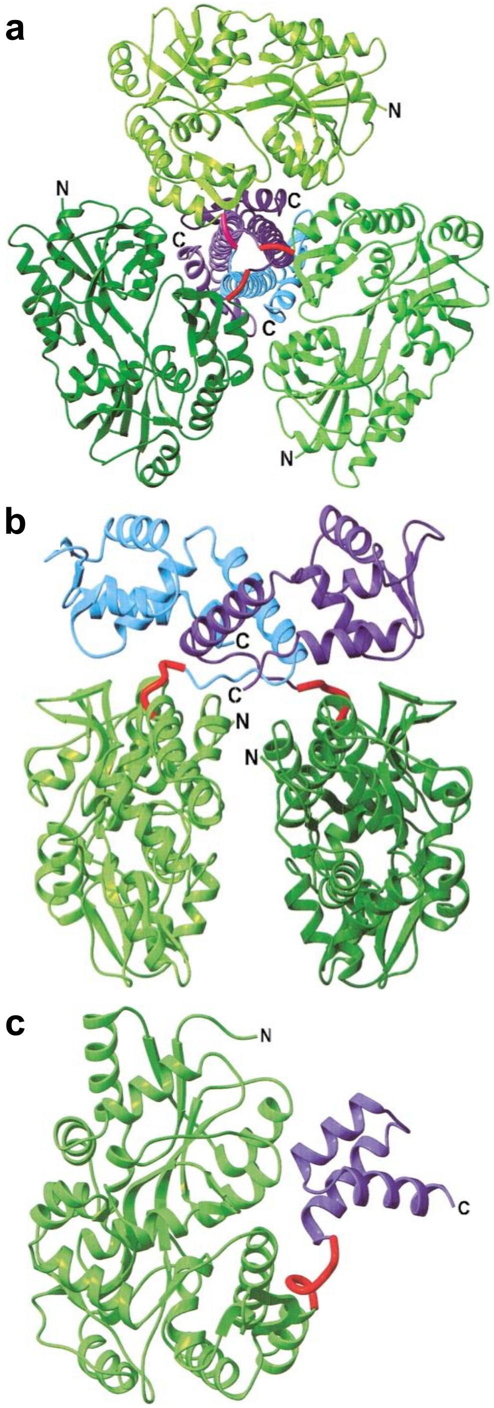


Figure 3.3: Structures of MBP several fusion proteins.

(a) Structure of MBP-gp21. The fusion proteins form a trimer through the gp21 passenger protein. gp21 subunits are coloured in blue and purple. (b) Structure of the MBP-SarR dimer, with SarR coloured in blue and purple. (c) Structure of the MBP-MATa1 protein, with MATa1 coloured in purple. MBP proteins are coloured in different shades of green and linker regions in red. N and C termini are labelled. Adapted from Smyth et al. 2003.

Similarly, the structure of SarR, a transcription factor from *S. aureus,* was also obtained through fusion with MBP through a truncated linker [179] (**Fig 3.3b**). The SarR formed a homodimer, and the structure could again be solved by molecular replacement with MBP to a resolution of 2.3Å.

Fusion with MBP was essential for producing crystals of MATa1 protein from yeast, which is part of a heterodimer that binds DNA [180]. It was necessary to obtain the structure of MATa1 in the absence of DNA and MATa2 (the second part of the heterodimer), but no crystals could be obtained, and the levels of protein expression were low (0.8mg/L). Fusion with MBP through a five alanine linker, with mutation of the charged C terminal MBP residues as before [178], increased protein expression levels by a factor of ~38, and greatly increased the chance of crystal formation [180]. The structure was solved at 2.1Å and 2.3Å resolution by molecular replacement with MBP (**Fig. 3.3c**).

These results suggest that the fusion of Lin28 or its domains to MBP could increase initial expression levels in *E. coli*, and maintain protein solubility during purification. If the linker region is truncated, the MBP tag could also act as a facilitator of crystallisation, and allow structure determination by molecular replacement with MBP.

3.2 Materials and Methods

3.2.1 Cloning and Plasmids: pGP constructs

GST-Lin28 fusion proteins were produced from a modified pGEX-6p-3 vector (see **Fig. 3.4**). Initially, a second BamHI restriction site was introduced by site directed mutagenesis at position 894, and the plasmid digested at both sites, purified by agarose gel electrophoresis, and ligated with T4 DNA ligase. This removed the DNA sequence corresponding to the cleavable linker region in the protein, so that constructs produced from this vector have a short, uncleavable linker. The plasmid was named pGP (pGEX Prepared).

The Lin28 sequences were generated by the PCR amplification of a synthetic, codon optimized sequence encoding Lin28A from *Homo sapiens* (accession number: NP_078950.1, GeneArt®), spanning residues 1-209 for the full length construct, and residues 37-180 for the truncated termini (TT) construct. These sequences and the pGP vector were then digested by BamHI and either XhoI or SalI restriction enzymes, and ligated with T4 DNA ligase to produce the final GST tagged contructs.

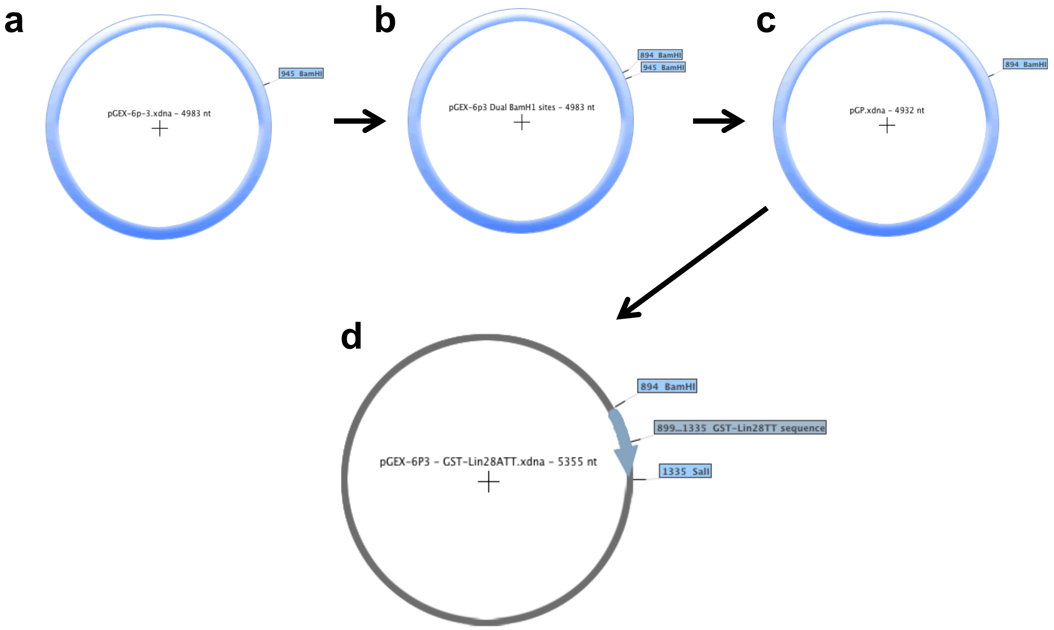

a
945 BamHI
pGEX-6p-3.xdna - 4983 nt
b
894 BamHI
945 BamHI
pGEX-6p3 Dual BamH1 sites - 4983 nt
c
894 BamHI
pGP.xdna - 4932 nt
d
894 BamHI
899..1335 GST-Lin28TT sequence
1335 SalI
pGEX-6P3 - GST-Lin28ATT.xdna - 5355 nt

Figure 3.4: Construction of the GST tagged Lin28 fusion proteins.

(a) The original pGEX-6p-3 plasmid containing the GST tag. (b) An extra BamH1 restriction site was introduced by site directed mutagenesis (c) The fragment between the two BamH1 sites was removed and the plasmid ligated. (d) The codon optimized Lin28 sequence amplified by PCR was inserted between the BamH1 and Sal1 sites to produce the final construct, which on induction would produce the GST-Lin28ATT fusion protein.

3.2.2 Cloning and Plasmids: pMBP Constructs

The plasmid encoding the His-MBP-4A-Lin28TT fusion protein was produced from the pETFF_2 vector provided by the York Technology Facility. The sequence encoding residues 32-187 of Lin28A from *Homo sapiens* (accession number: NP_078950.1, GeneArt$^{®}$) was amplified by PCR from the synthetic, codon optimized DNA sequence. The pETFF_2 vector was linearized by PCR and purified by agarose gel electrophoresis. The Lin28 sequence could be inserted into the pETFF_2 vector using the InFusion® system and the product transformed into *E. coli*.

3.2.3 Cloning and Plasmids: His Tagged Construct

The plasmid encoding the His-Lin28TT protein was previously prepared by Dr. Elena Blagova (Antson group). Briefly, a synthetic, codon-optimised DNA sequence corresponding to the Lin28A protein from *Homo sapiens* (accession number: NP_078950.1) was produced by GeneArt®, and the sequence spanning residues 37-180 amplified by PCR and subcloned into a YSBL LIC- vector by Ligation Independent Cloning (LIC).

3.2.4 Protein Expression

Plasmids encoding Lin28 constructs were transformed into *E. coli* expression cells. His-tagged and GST fusion plasmids were transformed into the B834 (DE3) expression strain, whereas the His-MBP-4A-Lin28TT encoding plasmid was transformed into the Rosetta2 (DE3) expression strain. Expression of the proteins proceeded as detailed in *Chapter 2*. For MBP tagged constructs the LB media was supplemented with 50µM $ZnCl_2$.

3.2.5 GST Fusion Protein Purification

Cell pellets were resuspended in Lysis Buffer (50mM TRIS pH 7.5, 2M NaCl, 0.5% Polyethylenimine (PEI), 10% sucrose, 2mM DTT) and lysates were prepared as detailed in *Chapter 2.*

Next, to precipitate DNA, the sample was diluted 4x using Salt Free Buffer (50mM TRIS pH 7.5, 0.5% Polyethylenimine (PEI), 10% sucrose, 2mM DTT). This reduced the NaCl concentration to 500mM and precipitated the DNA/PEI. The lysate was then clarified by centrifugation at 26892xg in a Sorvall centrifuge, SS34 rotor, for 50 minutes.

To remove excess PEI, solid ammonium sulphate was added slowly at 4°C to a final concentration of 60% saturation. The protein precipitate was then collected by centrifugation at 26892xg in a Sorvall centrifuge, SS34 rotor, for 50 minutes. The pellets were then resuspended in 50mM TRIS pH 7.5, 150mM NaCl, 10% w/v sucrose and the UV absorbance checked to determine if nucleic acid had been removed. The solution was then applied to 2x 1mL GSTrap columns (GE Healthcare) that had been connected in tandem and pre-equilibrated in the same buffer, using a peristaltic pump connected in such a way so that the flow through was reapplied to the column. This setup was then left at 4°C overnight. Subsequently, the columns were washed with 8 column volumes of the same buffer to elute non-specifically bound proteins. The GST fusion proteins were then eluted using buffer containing 50mM TRIS pH 8.5, 150mM NaCl, 2mM DTT, 10% w/v sucrose and 50mM GSH. The presence of the eluted protein was then verified using the Bradford assay, as GSH absorbs within in the UV region.

Finally, protein containing solutions were pooled and concentrated before being loaded on to a S75 10/30 (GST-CSD) or 26/60 (GST-Lin28A/GST-Lin28ATT) gel filtration column (GE Healthcare) equilibrated in 10mM TRIS pH 7.5, 500mM NaCl, 5% sucrose and 2mM DTT. The presence of the protein was checked by SDS-PAGE, with relevant fractions pooled, concentrated and flash frozen in liquid nitrogen for storage at -80°.

3.2.6 Purification of MBP tagged proteins

Cell pellets were resuspended in lysis buffer containing 50mM TRIS pH 7.5, 150mM NaCl, 2mM β-mercaptoethanol, 50μM $ZnCl_2$ and 20mM imidazole. The cell lysate was then prepared as described in *Chapter 2*, and applied at 4°C to a 5mL HisTrap column that had been charged with Zn^{2+} ions.

The column was washed at a flow rate of 0.1-0.2mL/min with buffer containing 50mM MES pH 6.0, 1M NaCl, 2mM β-mercaptoethanol, 50μM $ZnCl_2$ and 20mM imidazole at 4-6°C overnight (~16hrs). Proteins were eluted from the column with a buffer containing 50mM MES pH 6.0, 1M NaCl, 2mM β-mercaptoethanol, 50μM $ZnCl_2$ and 500mM imidazole. Fractions were checked by SDS-PAGE and those containing MBP fusion proteins pooled and concentrated using 30kDa MWCO concentrators (Amicon).

The concentrated eluate was then applied to either a S200 10/30 (His-MBP-4A-Lin28ATT) or 16/60 (His-MBP-4A-Lin28ATT/His-MBP-4A-CSD) gel filtration column (GE Healthcare) equilibrated with buffer containing10mM MES pH6, 1M NaCl, 2mM β-mercaptoethanol and 50μM $ZnCl_2$. Fractions containing MBP fusion proteins were pooled and concentrated using 30kDa MWCO concentrators (Amicon). Finally, samples were dialysed using a slide-a-lyzer (Pierce) against 2x500mL of buffer containing 10mM TRIS pH7.5, 150mM NaCl, 50μM $ZnCl_2$ and 2mM β-mercaptoethanol for the His-MBP-4A-Lin28ATT fusion protein. In the case of the His-MBP-4A-Lin28ACSD protein, samples were desalted following gel filtration, resulting in a final buffer composition of 10mM MES pH6, ~100mM NaCl, 2mM β-mercaptoethanol, 50μM $ZnCl_2$. The protein was then aliquoted, flash frozen in liquid N_2, and stored at -80°.

3.3 Results

3.3.1 Purification of GST-Lin28 Fusion Proteins

Fusions of Lin28 with GST were generated in order to produce sufficient quantities of protein for structural and biochemical analysis. Three constructs were produced, with Lin28 fused to GST (residues 1-213) through a single serine linker: GST-Lin28A with Lin28A residues 1-209, GST-Lin28TT with Lin28A residues 37-180, and GST-CSD with Lin28A residues 37-113.Testing the expression of the GST-CSD fusion protein in different bacterial expression strains, as well as at different temperatures, revealed the protein was optimally expressed in B834 *E. coli* cells overnight at 16°C. However, during a standard purification protocol (GSH affinity column/analytical gel filtration, **Fig 3.5**), only one high molecular weight species was observed, eluting in the void volume of the gel filtration column (**Fig. 3.5 b,c**). Analysis of this fraction by SDS-PAGE revealed that the major species was GST-CSD, but the A_{260}/A_{280} ratio of the sample was very high, indicating the presence of nucleic acid contaminants. Hence it was concluded that, in standard conditions, GST-CSD forms high molecular weight aggregates due to non-specific nucleic acid contamination.

To prevent this, the aggregated sample was applied to the column again in the presence of 1M NaCl. The aggregates were somewhat disrupted, revealing an extended curve on the gel filtration profile, but with a high A_{260}/A_{280} ratio (**Fig. 3.5d**), indicating that nucleic acids had not been removed.

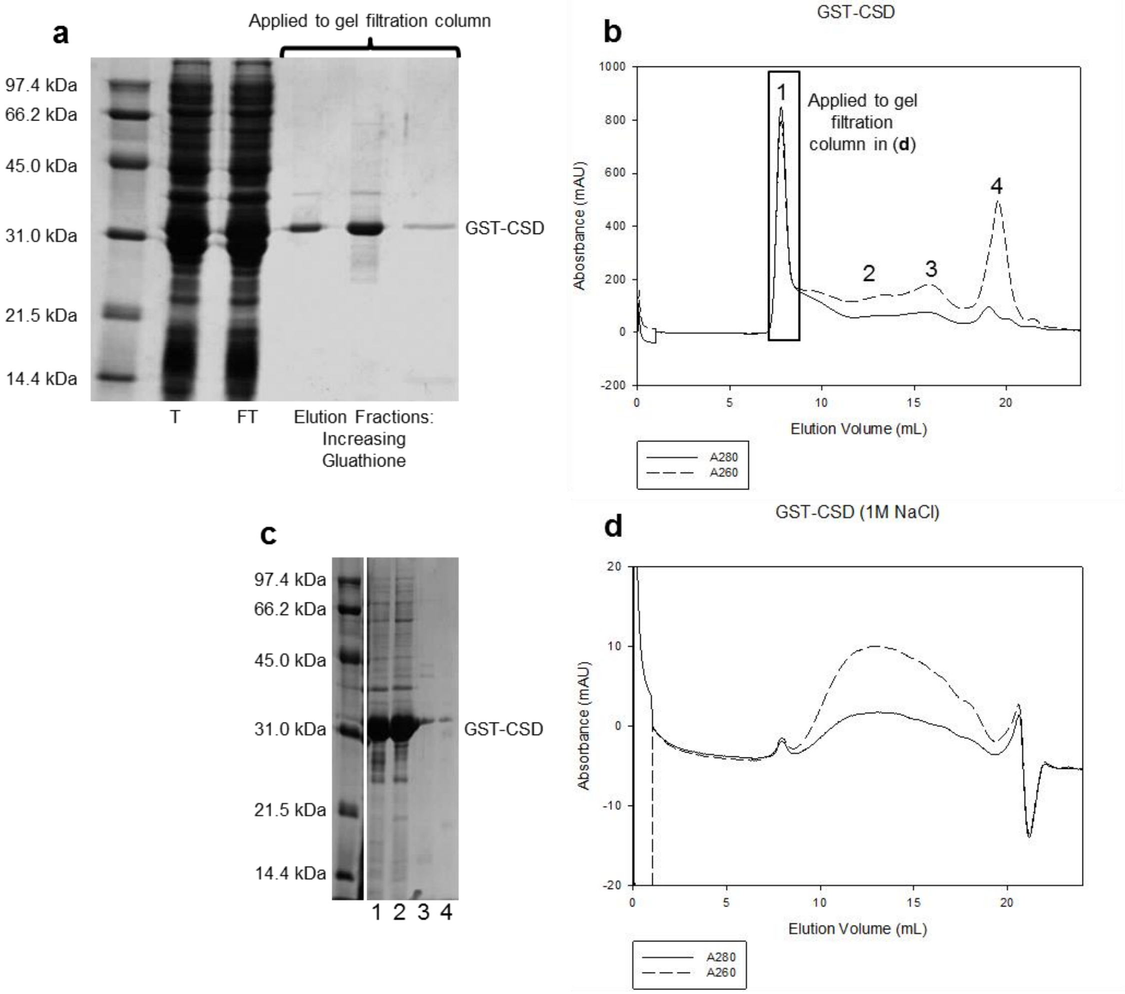


Figure 3.5: Purification of GST-CSD fusion proteins

(a) SDS-PAGE analysis of fractions eluting from the GSTrap affinity column. Fractions from lanes 4-6 were pooled and applied to the analytical gel filtration column shown in (b). (b) gel filtration elution profile of GST-CSD sample. (c) SDS-PAGE analysis of gel filtration fractions. (d) gel filtration elution profile of GST-CSD protein from sample 1 shown in (b) ran in a high salt buffer. Fractions from this gel filtration were too low in concentration to visualise by SDS-PAGE.

Polyethyleneimine (PEI) is a positively charged compound that binds nucleic acid and precipitates it from solution. Cells expressing GST-Lin28ATT were divided up and resuspended in buffer solutions containing 0.5M, 1M and 2M NaCl concentrations. Each of these aliquots was then further subdivided and different concentrations of PEI added to each. Lin28 remained in solution at all concentrations of salt and PEI. The 0.4% PEI samples were then desalted by buffer exchange and purified by glutathione affinity. The resulting eluates all contained Lin28, but concentrations determined by the Bradford assay were much higher than expected from the intensity of the bands seen by SDS-PAGE. This indicated that PEI was still present, as it strongly interacts with the Bradford reagent.

An optimized procedure was then developed. *E. coli* expressing GST-Lin28ATT were pelleted by centrifugation and resuspended in buffer containing 2M NaCl and 0.5% PEI and lysed by sonication. No precipitate appeared as PEI does not bind nucleic acid in high salt conditions. Therefore, the solution was diluted in a similar buffer without NaCl, until the total NaCl concentration was equal to 0.5M. During the dilution, a cloudy white precipitate appeared, and was removed by centrifugation. To remove the excess PEI, solid ammonium sulphate was added to 60% saturation. This precipitates out all proteins in solution, and the precipitate was collected by centrifugation and resuspended in fresh buffer, to produce a protein solution free from the presence of PEI. GST-Lin28TT protein remained in solution (**Fig. 3.6a**) with a UV absorbance of A_{260}/A_{280} ratio ~0.7, showing nucleic acid contaminants had been removed from the sample.

GST-Lin28ATT could then be purified from this solution by glutathione affinity chromatography followed by preparative gel filtration, which results in a single major peak eluting after the void volume (**Fig. 3.6 b,c,d**). The fractions containing this peak were pooled, concentrated and frozen in liquid N_2 for storage. This procedure was hence able to remove nucleic acid by disrupting the aggregates and precipitating out the contaminants in a two-step process prior to the protein purification steps, and, therefore, was used to purify GST-Lin28A (**Fig. 3.7**) in a similar fashion.

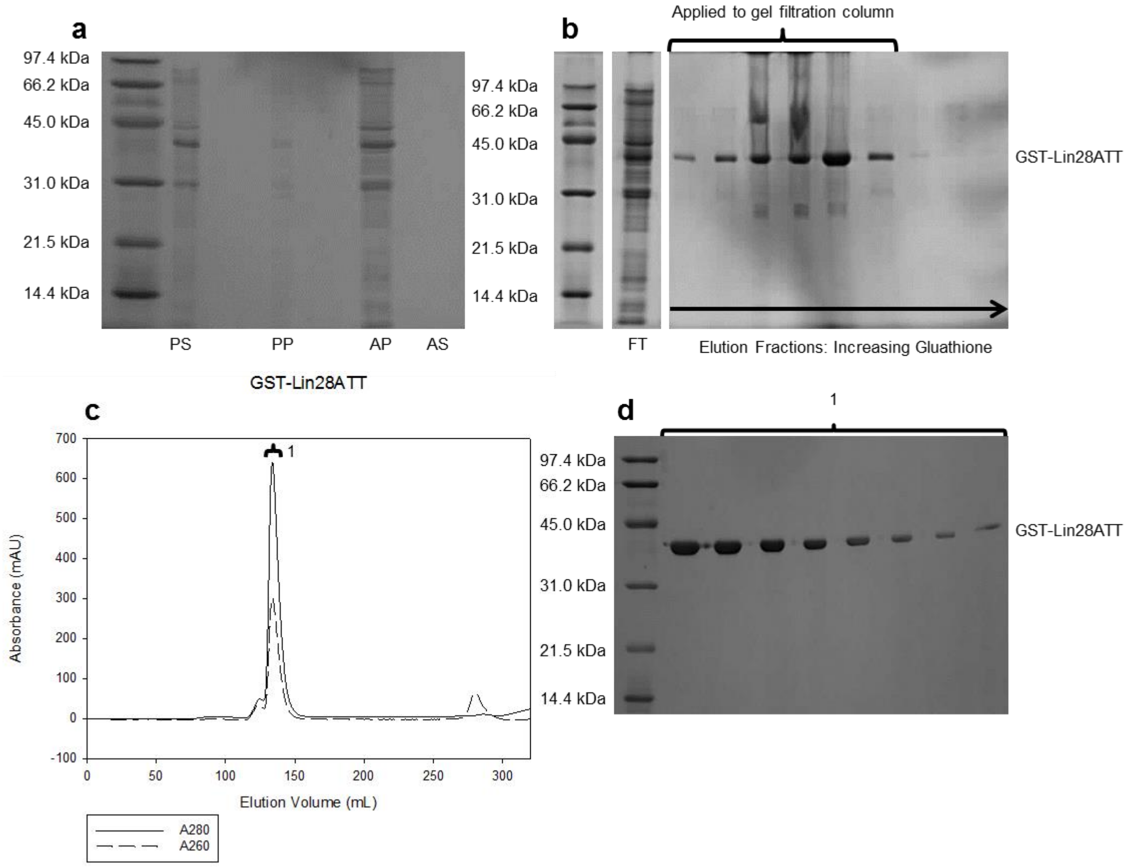


Figure 3.6: Purification of GST-Lin28ATT fusion proteins

(a) SDS-PAGE analysis of the precipitation steps used to remove nucleic acid from the GST-Lin28ATT sample. PS – PEI supernatant, PP – PEI pellet, AP – ammonium sulphate pellet, AS – ammonium sulphate supernatant. (b) SDS-PAGE analysis of fractions eluting from the GSTrap affinity column. Fractions from lanes 3-8 were pooled and applied to the preparative gel filtration column shown in (c). (c) gel filtration elution profile of GST-Lin28ATT sample. (d) SDS-PAGE analysis of gel filtration fractions.

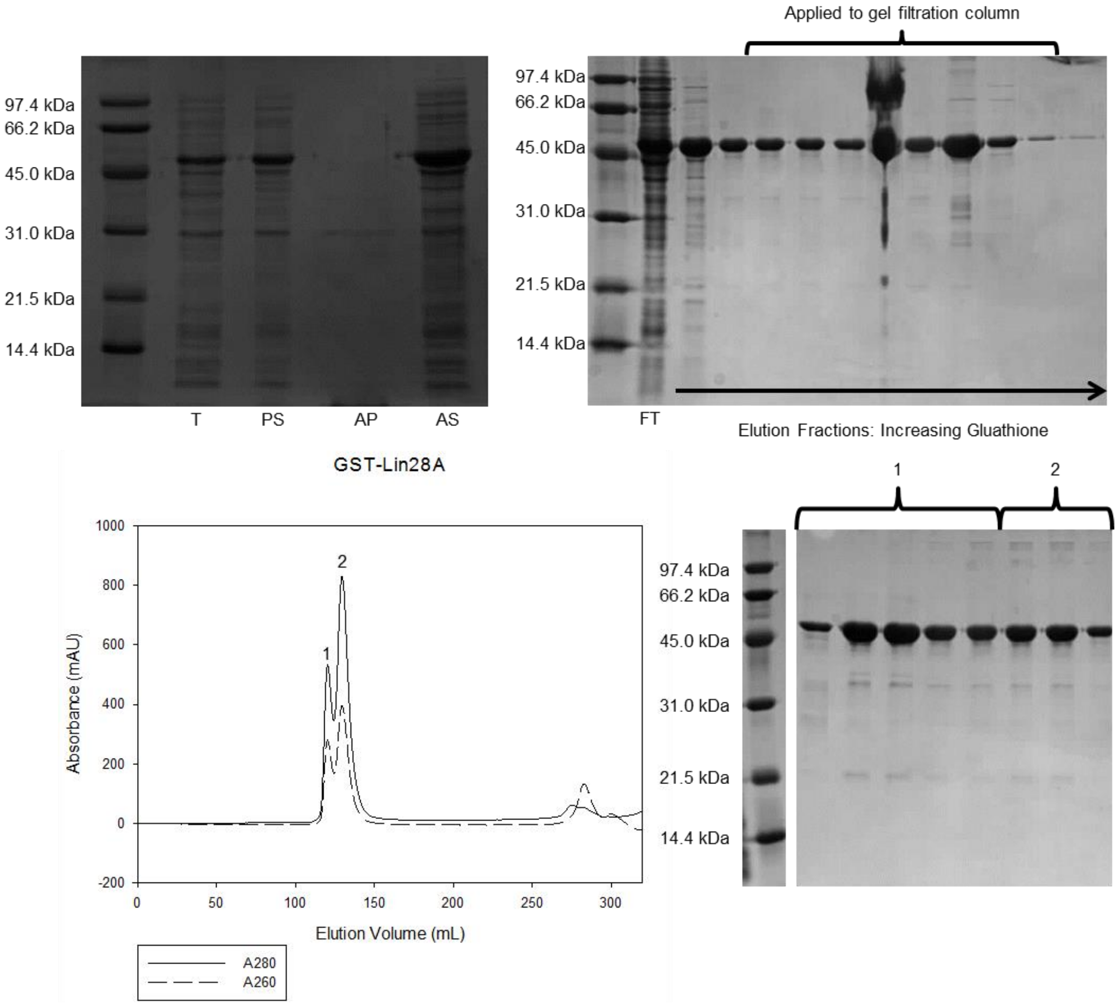


Figure 3.7: Purification of GST-Lin28A fusion proteins

(a) SDS-PAGE analysis of the precipitation steps used to remove nucleic acid from the GST-Lin28ATT sample. PS – PEI supernatant, AP – ammonium sulphate pellet, AS – ammonium sulphate supernatant. (b) SDS-PAGE analysis of fractions eluting from the GSTrap affinity column. Fractions from lanes 3-8 were pooled and applied to the preparative gel filtration column shown in (c). (c) gel filtration elution profile of GST-Lin28A sample. Elution fractions were analysed by SDS-PAGE shown in (d).

3.3.2 Purification of His-MBP-4A-Lin28A proteins

The GST fusion proteins were useful for initial biochemical tests on Lin28 (see *Chapter 4*), but were not optimal for in depth functional investigation. The GST tag necessary for maintaining the stability of Lin28 both throughout the precipitation protocol and in the absence of nucleic acids, forms a dimer, which limited the use of the fusion protein in testing the affinity and oligomeric states of Lin28 complexes. Additionally, as the precipitation protocol is a harsh procedure, it could not be determined if the ZnK domain was still intact with both Zn^{2+} ions still bound by the protein. Therefore, a new method was developed for removing nucleic acid, and a series of MBP tagged fusion proteins were produced, which would be more appropriate for biochemical analysis of Lin28. An N terminal His tag was added to facilitate purification due to the low affinity of MBP for amylose, which limits the effectiveness of purification by affinity chromatography [169].

Two constructs were generated: His-MBP-4A-Lin28ATT, containing Lin28A residues 32-187 fused to His-MBP (residues 25-396) through a four alanine linker, and His-MBP-4A-Lin28ACSD, which contained residues 32-127 of Lin28A fused to MBP in the same way.

Initial expression tests revealed that His-MBP-4A-Lin28ATT (Lin28TT) proteins were optimally expressed in Rosetta2 *E. coli* cells, grown at 16°C overnight. After expression in a large scale culture, a standard purification (Ni^{2+} IMAC/gel filtration, **Fig. 3.8**) of the Lin28TT protein was attempted. As with previous constructs, only large aggregates could be seen by analytical gel filtration, with a high A_{260}/A_{280} ratio indicating the presence of nucleic acid contaminants (**Fig. 3.8 b,c**).

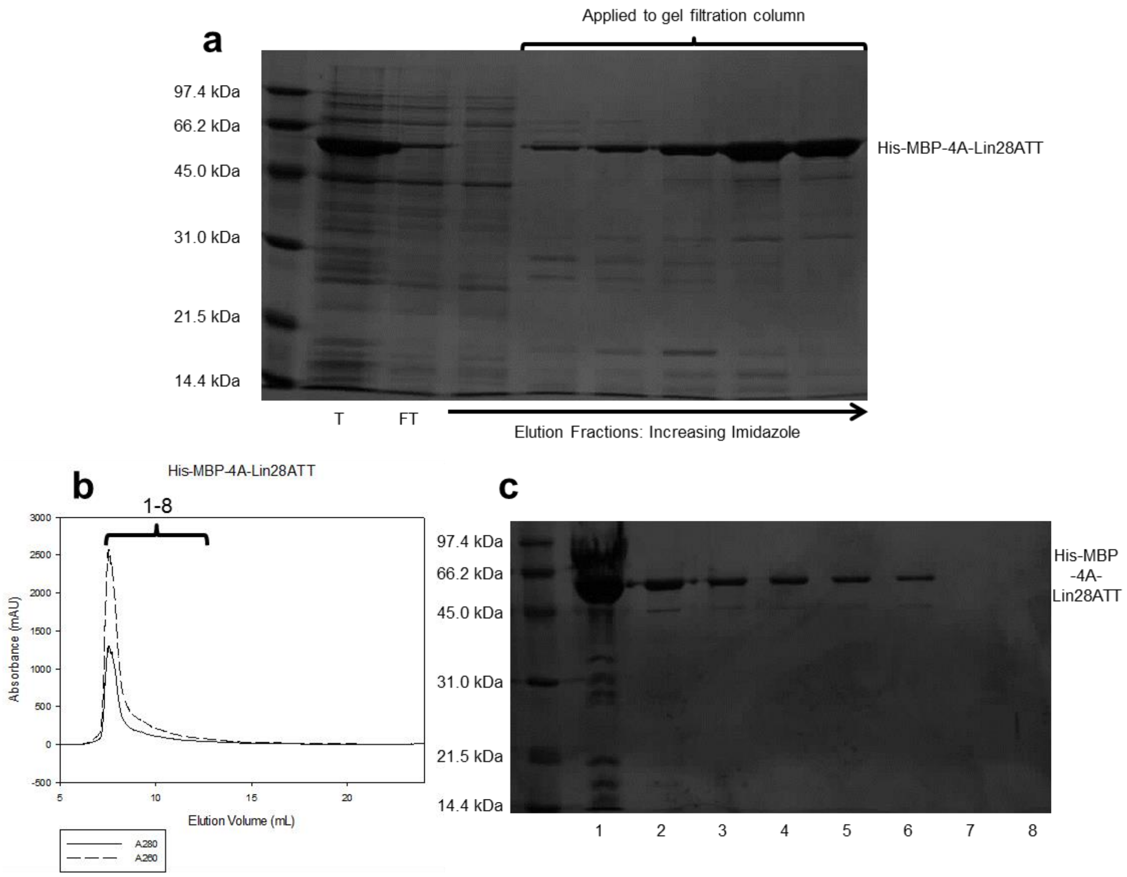


Figure 3.8: Purification of His-MBP-4A-Lin28ATT.

(a) SDS-PAGE analysis of the fractions eluting from the Ni^{2+} IMAC column. T = Total soluble protein, FT = column flow-through (unbound protein). Fractions from lanes 5-9 were pooled and applied to the analytical gel filtration column shown in (b). (b) Elution profile from gel filtration column. Elution fractions were analysed by SDS-PAGE, shown in (c).

To combat this, the purification procedure was modified. The protein was expressed in LB media supplemented with 50μM $ZnCl_2$. Following harvesting, the cells were resuspended in a mild buffer containing 50mM TRIS pH 7.5 and 150mM NaCl, and lysed by sonication. The lysate was then applied to a HisTrap column charged with Zn^{2+} ions. The column was then washed with buffer containing 1M NaCl in pH6 buffer at a low flow rate overnight (~16hrs). This successfully removed the bound nucleic acid; when Lin28 was eluted and applied to a preparative gel filtration column, only one major peak was observed which had a low A_{260}/A_{280} ratio of ~0.7 (**Fig. 3.9**).

To maintain the integrity of the ZnK domain, media and all buffers were supplemented with 50μM $ZnCl_2$. In addition, the use of the Zn^{2+} IMAC meant that the protein could still be purified via its His tag but no Ni^{2+} ions form the column could replace the Zn^{2+} in the ZnK domain. All steps were performed at temperatures of between 4-6°C to keep the protein stable throughout the process.

This procedure produced stable, nucleic acid-free Lin28 fusion proteins in sufficient quantities for biochemical and structural investigation. His-MBP-4A-Lin28ACSD (CSD) protein could also be purified by this method (**Fig. 3.10**).

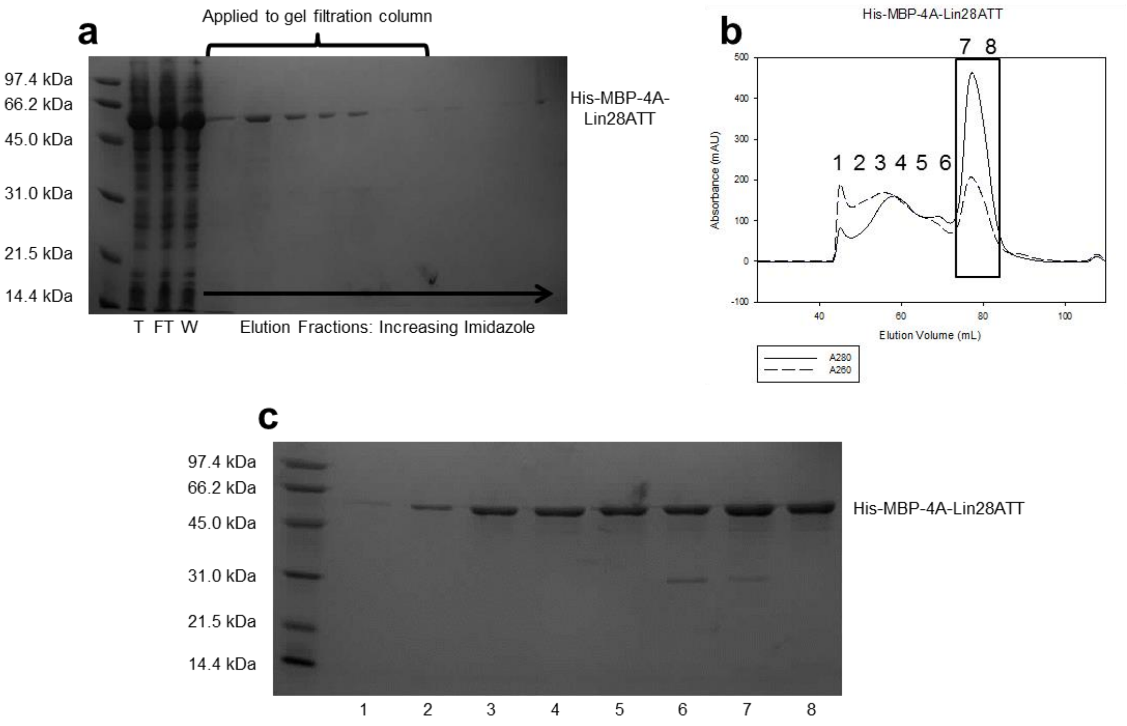


Figure 3.9: Purification of His-MBP-4A-Lin28ATT with long salt wash step.

(a) SDS-PAGE analysis of the fractions eluting from the Zn^{2+} IMAC column. T = Total soluble protein, FT = column flow-through (unbound protein), W = Wash sample eluting during long salt wash step. Fractions from lanes 5-11 were pooled and applied to the preparative gel filtration column shown in (b). (b) Elution profile from gel filtration column. Elution fractions were analysed by SDS-PAGE, shown in (c). The box denotes fractions pooled and used for downstream applications.

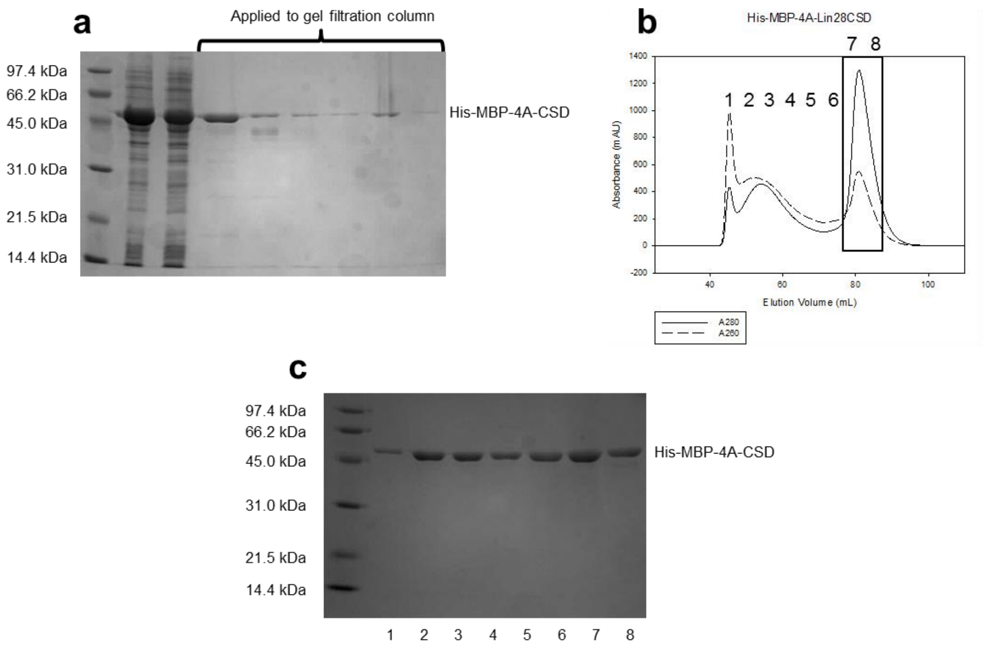


Figure 3.10: Purification of His-MBP-4A-CSD.

(a) SDS-PAGE analysis of the fractions eluting from the Zn^{2+} IMAC column. T = Total soluble protein, FT = column flow-through (unbound protein). Fractions from lanes 4-9 were pooled and applied to the preparative gel filtration column shown in (b). (b) Elution profile from gel filtration column. Elution fractions were analysed by SDS-PAGE, shown in (c). The box denotes fractions pooled and used for downstream applications.

3.3.3 Substitution Method for generating Lin28:nucleic acid complexes

Another method was developed allowing the quick output of pure, non-aggregated Lin28/nucleic acid complexes. A His-Lin28A-TT (residues 37-180) construct with an uncleavable His-tag was found to express well in BL-21 cells at 37°C for 4 hours (Dr. Elena Blagova, Antson group). A large-scale culture of these cells was grown and expression of the protein induced. Harvested cells were lysed by sonication and purified by Ni^{2+} IMAC (**Fig. 3.11a**). Eluted His-Lin28A-TT was then split into two parts. One part was applied to an analytical gel filtration column (**Fig. 3.11b**), whilst let-7gΔ5 RNA was added in a 1:1 molar ratio to the other part. The resulting mixture was concentrated, and left overnight at 4°C before application to the same analytical gel filtration column (**Fig. 3.11c**). The results showed that His-Lin28A-TT without added nucleic acid formed solely high molecular weight aggregate species, which elute at void volume (**Fig. 3.11b**), as before. However, when the RNA was added, an extra peak eluted much later than the aggregates (**Fig. 3.11c**). The high A_{260}/A_{280} ratio confirmed that nucleic acid was present in this fraction and SDS-PAGE revealed the only protein component was His-Lin28A-TT (**Fig. 3.11d**). Therefore, the specific RNA oligonucleotide is able to compete off the non-specifically bound nucleic acid contaminants to produce a smaller nucleic acid containing fraction, likely corresponding to a specific protein:RNA complex.

To further investigate this effect, unbound His-Lin28A-TT was concentrated and mir363, let-7i and dT20 oligonucleotides were added in 1:1 molar ratios, with samples left to incubate overnight at 4°C and loaded onto the analytical gel filtration column. Again, peaks corresponding to protein/nucleic acid complexes were observed for each oligonucleotide (**Fig. 3.12 a,b,c**). The substitution method therefore generates Lin28/nucleic acid complexes with non-fusion Lin28 proteins, in a stable and reproducible way. This technique enables the quick output of such complexes from purification directly into crystal screening.

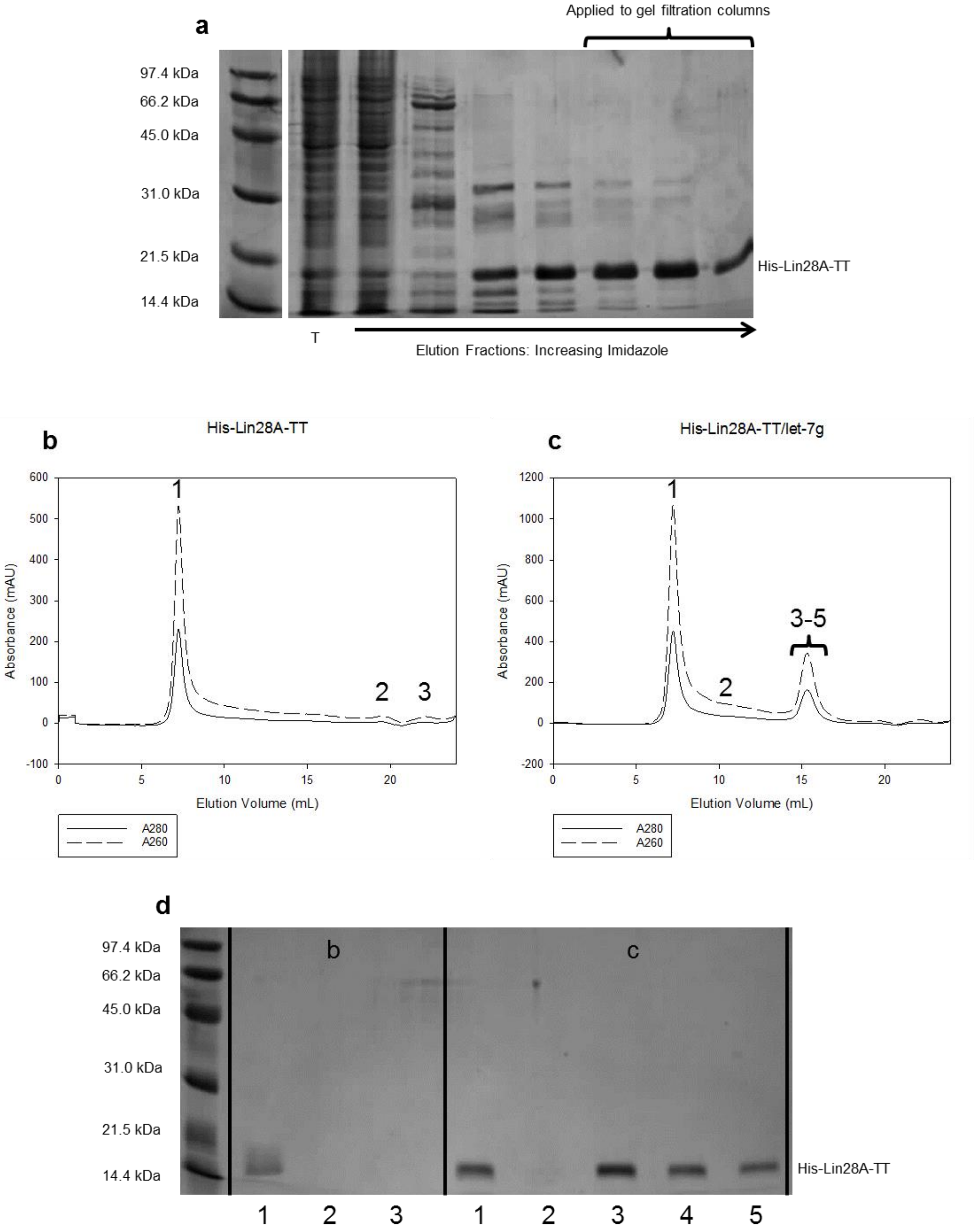


Figure 3.11: Purification of His-Lin28A-TT and its complex let-7g produced by the substitution method.

(a) SDS-PAGE analysis of the fractions eluting from the Ni^{2+} IMAC column. T = Total soluble protein. Fractions from lanes 3-9 were pooled and applied to the analytical gel filtration columns shown in (b) and (c). Gel filtration elution profiles of protein only (b) and the protein/let-7g mix (c) are shown, and eluting fractions analysed by SDS-PAGE in (d).

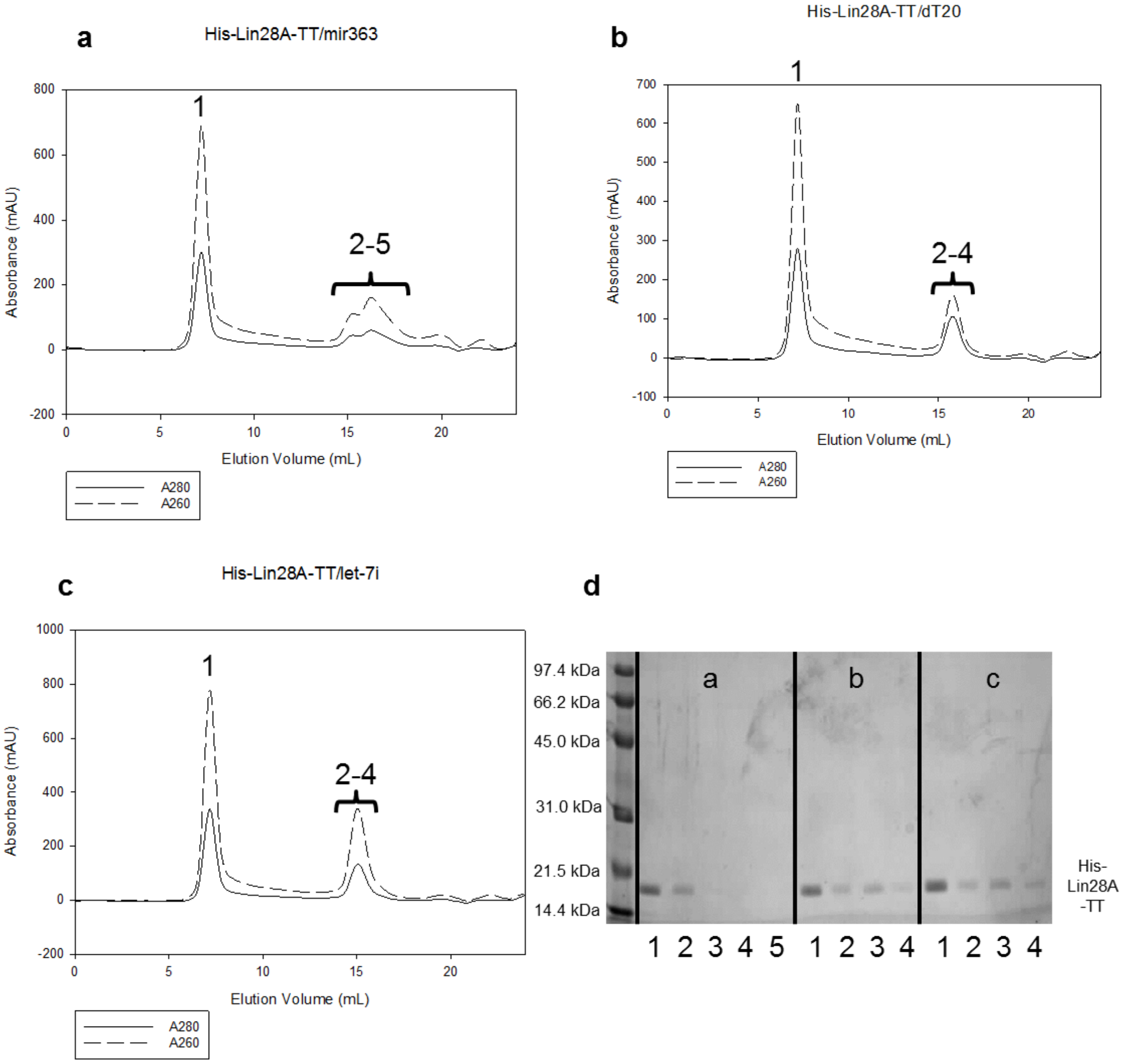


Figure 3.12: Purification of His-Lin28A-TT/nucleic acid complexes generated by the substitution method.

Gel filtration elution profiles of His-Lin28A-TT when mixed with mir363 (a), dT20 (b) and let-7i (c) oligonucleotides (b) are shown, and eluting fractions analysed by SDS-PAGE in (d).

3.4 Discussion

Purifying recombinant Lin28 is a challenging procedure for three reasons: (i) the aggregation caused by strong binding of nucleic acid contaminants, (ii) the instability of the protein in the absence of nucleic acid and (iii) the difficulty in keeping the ZnK domain intact. Each purification procedure must contain strategies to account for each of these effects.

3.4.1 Nucleic acid contaminants can be removed or replaced

The interaction between Lin28 and the nucleic acid contaminants is reversible in certain conditions, as it is at least partially electrostatic in nature; using high salt concentrations can disrupt the aggregates. Once a portion of the nucleic acid is unbound, it can either be removed by precipitation, or washed off with buffer. Alternatively, the addition of oligonucleotides that Lin28 binds allows complexes to form that are of lower molecular weight than the non-specific aggregates and thus can be purified by gel filtration. One caveat of this approach is that it is yet to be confirmed whether or not these peaks correspond to specific complexes between Lin28 and the added oligonucleotide. The fractions comprising these peaks will therefore need to be analyzed by native gel and SEC-MALLS. If a specific complex is formed, the substitution method provides another approach for removing non-specific nucleic acid contaminants from purified Lin28. Therefore, three purification approaches were developed in total: the precipitation protocol used to produce the GST fusion proteins, the method involving a long wash with high salt buffer used to purify the MBP fusion proteins, and the nucleotide substitution method used in the purification of His-Lin28A-TT, where non-specific nucleic acid is competed off with specific oligonucleotides.

3.4.2 Fusion of Lin28 to large affinity tags increases solubility and stability

Removing the nucleic acid is the most challenging step, as the Lin28/nucleic acid complex is highly stable. However, reversing this interaction often involves harsh methods that could destabilize the ZnK domain, and the protein appears to be unstable in the absence of bound nucleic acid.

The fusion of MBP or GST to Lin28 provides multiple advantages [168]. The expression levels and solubility of fusion proteins are higher, resulting in increased yields of protein. Additionally, the tag stabilizes the protein both throughout the purification procedure, and once nucleic acid has been removed. The large affinity tag also increases the molecular weight of the relatively small Lin28 protein, which allows the fusion proteins to be used in techniques such as fluorescence anisotropy and SEC-MALLS, where a relatively large difference in mass between binding partners is needed for obtaining high quality data. Finally, if crystals were obtained, the structure could be determined by molecular replacement with the relevant tag, as has been done previously [168, 173, 174, 177, 179, 180]. Therefore, the fusion proteins described are highly appropriate for investigating the structure and function of Lin28 proteins.

The oligonucleotide substitution technique does not remove nucleic acid, and so there is never a point where Lin28 is unbound, meaning stability can be maintained even without conjugation to a large affinity tag.

3.4.3 Maintenance of ZnK domain integrity

It is difficult to ascertain if the ZnK domain remains folded after purification. The fold is stabilised by bound Zn^{2+} ions that are coordinated by the CCHC motif of each zinc knuckle subdomain [122, 124] Depending on how tightly Zn^{2+} ions are bound, it may be important to ensure availability of Zn^{2+} ions, whilst avoiding competing buffer components that may disrupt Zn^{2+} binding. This maximizes the probability of retention of these ions by the ZnK domain during purification.

Each of the three purification procedures takes this into account. Purification of the MBP fusion proteins involved an IMAC stage that could potentially result in the exchange of bound Zn^{2+} ions for Ni^{2+} ions attached to the column. To prevent this, a HisTrap column with chelated Zn^{2+} ions instead of Ni^{2+} was used. This ensured that any Zn^{2+} exchanged could only be replaced by Zn^{2+}. In addition, every buffer was supplemented with $ZnCl_2$ to provide an excess of Zn^{2+} ions in case any were lost from the protein during purification. This problem was eliminated in the purification of the

GST fusion proteins, as GSH affinity chromatography is instead of Ni^{2+} IMAC, preventing the exchange of Ni^{2+} ions with the Zn^{2+} ions of the ZnK domain. In the course of nucleotide substitution procedure, it is possible that, because Lin28 molecules were bound to nucleic acid, Zn^{2+} ions were unlikely to dissociate from the protein. This should prevent the exchange of Zn^{2+} ions with Ni^{2+} ions from the column during Ni^{2+} IMAC.

The choice of reducing agent in the buffer is also an important factor. It has been shown that DTT can co-ordinate Zn^{2+} and inhibit proteins that rely on bound Zn^{2+} ions to perform their functions [181]. Buffers used in the purification of the MBP fusion proteins therefore contained 2mM β-mercaptoethanol as a reducing agent instead of DTT.

In conclusion, this chapter presents three alternative strategies for purifying stable recombinant Lin28 proteins that are free from nucleic acid contamination. Each of these strategies takes into account the three challenges detailed above, that must be overcome in order to produce uncontaminated Lin28. The proteins produced by the above methods resulted in homogeneous preparations suitable for structural and biophysical analysis.

Chapter 4 : Determinants of Lin28 Specificity

4.1 Introduction

Lin28 binds primary and precursor sequences of the let-7 family of miRNAs and prevents their biogenesis [79]. Through this process, the mRNA targets of let-7 are no longer repressed by the miRNA, and their expression is upregulated. Three let-7 targets, K-Ras, c-Myc and HMGA2, are potent oncogenes, and so the aberrant expression of Lin28 is associated with cancer [99]. The elucidation of the molecular mechanisms Lin28 uses to recognise and bind let-7 RNA could therefore inform the development of novel therapeutic agents that could combat disease.

In the prevailing mechanism, Lin28 and its paralog Lin28B prevent let-7 biogenesis by two distinct pathways [81]. Lin28 enhances the uridylation of precursor let-7 sequences by the terminal-U transferase enzyme Zcch11 [91-93], and is localised mainly to the cytoplasm [81]. The paralogs are highly similar in sequence, and biochemically interchangeable [92]. However, in contrast to Lin28A, Lin28B is localised to the nucleolus [81]. This is due to nuclear and nucleolar localisation signals in the Lin28B sequence. Lin28B was found to directly bind pri-let-7g *in vitro* by gel shift assay, with an apparent K_d of 0.5nM, close to the K_d of 0.6nM for the interaction of this RNA with Lin28A. However, RNA immunoprecipitation (RIP) of HeLa cells revealed that pri-let-7g is enriched in Lin28B extracts to a much greater extent than Lin28A extracts, and showed greater accumulation when Lin28B was transiently expressed rather than Lin28A. In addition, co-immunoprecipitation did not detect interaction between Zcchc11 and Lin28B. It was observed, however, that the microprocessor complex responsible for pri-let-7 maturation and cleavage could not enter the nucleolus. Therefore, it was concluded that Lin28 would bind pre-let-7 RNAs in the cytoplasm and recruit Zcchc11 to uridylate the RNA [81], signalling for its degradation by the Dis3L [95] nuclease, whereas Lin28B would bind and sequester pri-let-7 transcripts in the nucleolus, and prevent them from being cleaved by the microprocessor complex. By these two mechanisms, the biogenesis of let-7 miRNAs is inhibited [81].

4.1.1 Affinities of Lin28/let-7 interactions

Initially, interaction between Lin28 and let-7 RNA was characterized by gel shift analysis. His-Lin28 was expressed in *E. coli* BL-21 cells, and purified by standard procedures, and bound to pre-let-7g in a 50mM TRIS pH 7.6 buffer, including 100mM NaCl, 30μg yeast tRNA extract and β-mercaptoethanol as the reducing agent. Lin28 protein was mixed with pre-let-7g RNA, and a K_d of 2.1μM was deduced when the single site ligand binding equation (See *Chapter 2*, *Equation 10*) was fit to the data. This value compared to the 1.5μM K_d obtained with only the terminal loop of the miRNA (**Fig. 4.1**), and no binding was seen to the mature miRNA duplex. A conserved cytosine (C45 in human pre-let-7g) was identified within the terminal loop sequence which, when mutated to alanine, reduced the affinity of Lin28 for this sequence 20-fold, without altering the structure of the RNA. Both domains of Lin28 were necessary for binding. Therefore, it was concluded that Lin28 would bind the let-7 terminal loop though the conserved cytosine [80].

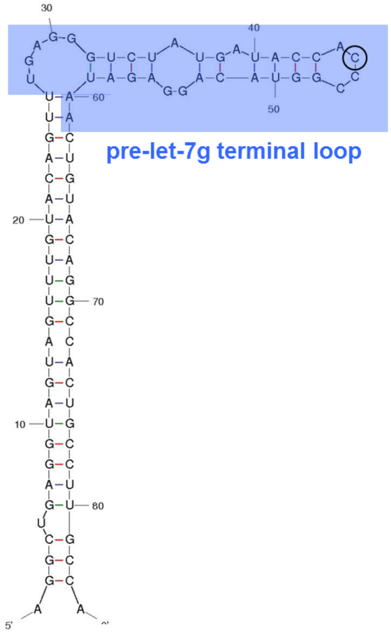


Figure 4.1: Secondary structure of the human pre-let-7g miRNA as predicted by MFOLD. The terminal loop segment is highlighted in blue.

Both pri- and pre-let-7 sequences contain the terminal loop. The 3500-fold difference in affinity between the reported K_d for the pre-let-7g interaction and that observed for the pri-let-7g interaction (see above section), was ascribed to the addition of the yeast tRNA to the buffer used to determine the Lin28/pre-let-7g K_d [81].

Further gel shift analysis resulted in K_d's of 0.15 and 0.13nM for the binding of full-length and terminal loop sequences of let-7g by mouse Lin28. Binding reactions were conducted in a solution containing 50mM TRIS pH 7.6, with 50mM NaCl, 10% glycerol, 0.05% NP-40 alternative detergent and 2mM DTT. Recombinant Lin28 in this case was expressed in BL-21 *E. coli* and purified by affinity to GSH resin, where it was washed and incubated with 5 units/mL S7 nuclease before elution and further purification by gel filtration. The binding data in this study were fit to a variant of *Equation 10* containing a Hill term that accounts for multiple co-operative binding events. A Hill coefficient of ~3 was calculated, which implies co-operative binding of up to three Lin28 molecules per let-7 terminal loop. In this case, it was determined that a G-rich bulge at the 5' end of the terminal loop (**Fig. 4.3**) was critical for Lin28 binding [148].

This G-rich bulge had enhanced levels of protection when human pre-let-7g was bound by Lin28 and digested with RNase enzymes implying this site is bound by Lin28. The affinity of the interaction was measured by gel shift, and K_d values of 0.88μM for pre-let-7g and 1.1μM for the terminal loop region were deduced when data were fit to *Equation 10*. The buffer used in this experiment containined 50mM TRIS pH 7.6 100mM NaCl, 0.07% β-mercaptoethanol, 1mM $Mg(OAc)_2$ and 12.5μg yeast tRNA. The protein used for analysis was a GST-Lin28 fusion, expressed in *E. coli* BL-21 cells. The RNAse protection assay also revealed that the structure of pre-let-7g melted upon Lin28 binding, suggesting Lin28 is able to modify the secondary structure of bound RNA sequences [112].

4.1.2 Lin28/Lin28B induce changes in pre-let-7 secondary structures

Further evidence supporting the ability of Lin28 to melt RNA hairpin structures came from the study of the CSD of *Xtr*Lin28B [123]. His-tagged Lin28B/CSD/ZnK constructs were expressed in *E. coli*, and purified by Ni^{2+}/Zn^{2+} affinity chromatography. After elution from the affinity column, the His-tag was cleaved by TEV protease and the proteins further purified by heparin column and gel filtration. 10μM $ZnSO_4$ was present in all buffers [123].

Affinity measurements revealed the preferential binding of heptameric single stranded polypyrimidine oligonucleotides by the *Xtr*Lin28B CSD [123]. DNA oligonucleotides were bound, with the optimal sequence of GTTTTTT. K_ds of 12nM and 26nM were determined for this interaction by fluorescence titration and ITC respectively. This compared to 35nM for GUUUUUU and 13nM for GUCAUAC RNA oligonucleotides. This sequence was taken from the pre-let-7f terminal loop and the binding affinity measured by ITC in a solution containing 20mM TRIS pH 8, 60mM KCl. This demonstrated the limited sequence specificity of the CSD, and revealed its site size of ~7nt [123], comparing with the 9-11nt loops bound in the mLin28/preE-let-7 crystal structures [122].

Binding of the ZnK domain only to AAGGAGAA and AAGGUGAA RNA oligonucleotides resulted in deduced K_d values of 45nM and 32nM respectively by ITC. No binding could be observed when this motif was mutated to AAAAAAAA [123].

The affinity of the *Xtr*Lin28B protein, containing both the CSD and ZnK domains but with truncated N- and C- termini, towards *Xtr*-pre-let-7f, was determined by gel shift assay [123], and returned a K_d of 1.6μM. *Equation 10* with the Hill term added was used to fit the data, and showed that the binding of this protein was co-operative with a Hill coefficient of 2.3, implying that ~2 molecules of Lin28B are binding per let-7f sequence [123].

A fluorescence quenching assay was then performed, which showed that binding of the *Xtr*Lin28B and its CSD could remodel the terminal loop of the *Xtr*-pre-let-7f [123]. In contrast, the ZnK domain on its own could not. Kinetic measurements then revealed that the full-length protein binds this loop in two phases – an initial fast step followed by a slower step. The CSD on its own produced only a monophasic curve, and only the fast step was observed. Therefore, it was concluded that the CSD of Lin28B binds first, and rearranges, or melts, the pre-let-7f terminal loop (**Fig. 4.2**). This exposes the GGAG motif, which is initially inaccessible to the ZnK domain, thus allowing the ZnK to bind. As the binding of Lin28B molecules is cooperative, it was suggested that the remodelling may be facilitated by a second Lin28B molecule [123].

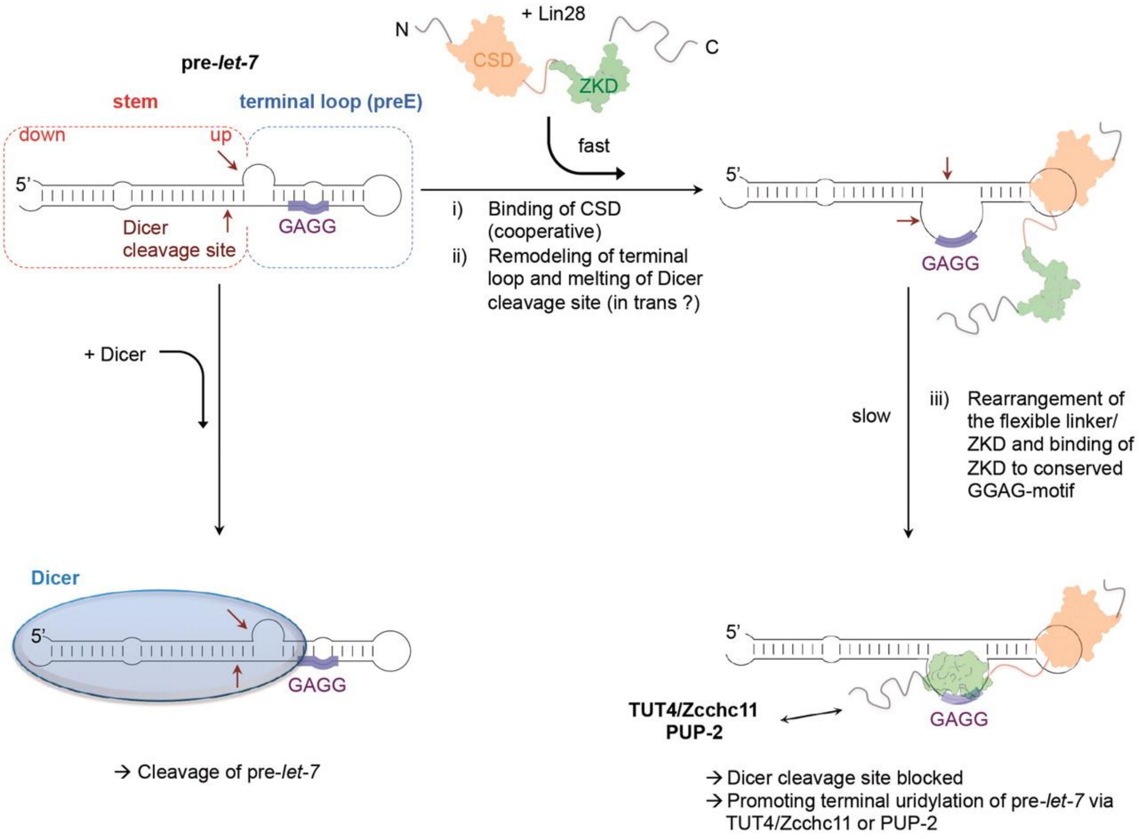

pre-*let-7*
stem
terminal loop (preE)
down
up
5'
Dicer cleavage site
GAGG
N
+ Lin28
C
CSD
ZKD
fast
i) Binding of CSD (cooperative)
ii) Remodeling of terminal loop and melting of Dicer cleavage site (in trans ?)
5'
GAGG
+ Dicer
slow
iii) Rearrangement of the flexible linker/ ZKD and binding of ZKD to conserved GGAG-motif
Dicer
5'
GAGG
→ Cleavage of pre-*let-7*
5'
TUT4/Zcchc11
PUP-2
GAGG
→ Dicer cleavage site blocked
→ Promoting terminal uridylation of pre-*let-7* via TUT4/Zcchc11 or PUP-2

Figure 4.2: Binding of *Xtr*-let-7g by *Xtr*Lin28B proceeds via a remodelling of the RNA structure. Adapted from Mayr et al. (2012).

4.1.3 Sequential addition of three Lin28 molecules per let-7 terminal loop

Recent data have observed up to three Lin28 molecules binding the terminal loop of pre-let-7g. Gel shift analysis demonstrated the presence of three different binding sites within the pre-let-7g terminal loop sequence: the G-rich bulge, internal loop (iloop) and tetraloop (**Fig. 4.3**). Binding of Lin28 to the terminal loop sequence gave a K_d of 0.13nM and a Hill coefficient of ~3, indicating that three Lin28 molecules were bound cooperatively. Mutations in any of these regions did not greatly decrease the affinity of Lin28 towards the RNA, and high affinity, 1:1 binding was observed with each region separately [182]. Similarly to the results of the *Xtr*Lin28B binding study [123], the CSD of human Lin28 was observed to melt the terminal loop sequence by two different biophysical assays. In contrast to that study, however, the ZnK domain was seen to bind with higher affinity to the terminal loop than the CSD alone. The formation of the 3:1 complex was also dependent on the concentration of protein used in the gel shift assay such that the higher order stoichiometries were only visible at higher protein concentrations [182].

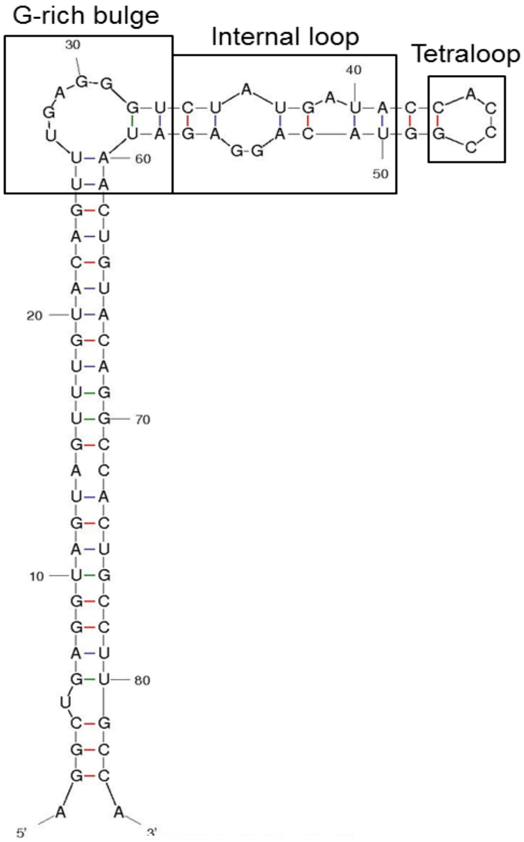


Figure 4.3: Secondary structure of the human pre-let-7g miRNA as predicted by MFOLD. The three proposed Lin28 binding sites are shown in boxes.

A model was proposed (**Fig. 4.4**) whereby a Lin28 molecule initially binds the G-rich bulge region, with the 5' GGAG motif bound by the ZnK domain. Binding of the CSD melts the stem of the terminal loop, and exposes the second 3' GGAG motif, facilitating the binding of a second Lin28 molecule through its ZnK domain. A third Lin28 molecule can then associate with the region in between the two sites by binding through its CSD only. This then causes the Lin28 bound to the first site to relocate the position of its CSD on the RNA. Similar findings were observed with other let-7 family members, and it was suggested this is the general mechanism for Lin28 binding to miRNA sequences [182].

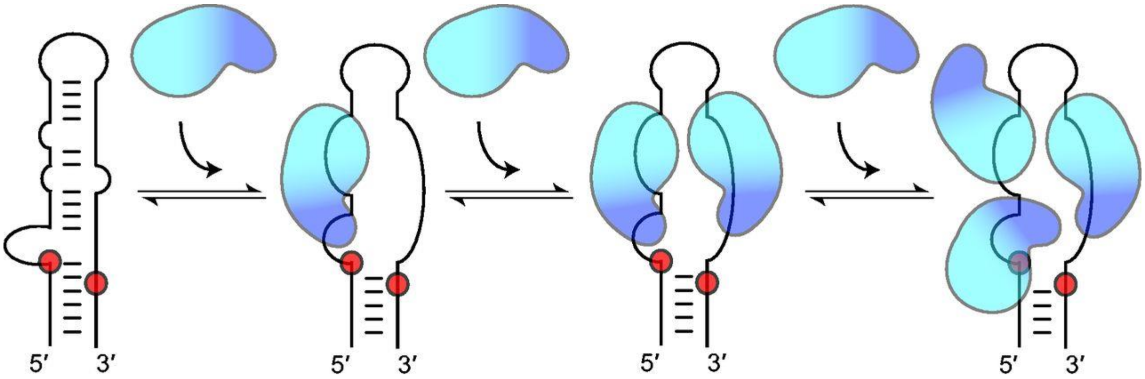


Figure 4.4: Binding of the pre-let-7g terminal loop by Lin28.

The ZnK domain of the first Lin28 molecule binds the 5' GGAG motif, and the CSD binds at a proximal location. This melts the stem loop exposing the 3'GGAG to be bound by the ZnK domain of a second Lin28 molecule, with its CSD binding nearby. The CSD from a third Lin28 molecule then binds, prompting a rearrangement of the CSD in the first molecule. Adapted from Desjardins et al. (2014).

4.1.4 Remaining questions

The mechanism by which Lin28 interacts with miRNA sequences is gradually becoming clearer. However, the observations of the stoichiometries and affinities of Lin28/let-7 complexes detailed above are highly variable. The affinity measurements for the Lin28/pre-let-7g interaction differ by ~10000-fold. The stoichiometries reported for Lin28/B complexes with let-7 sequences also encompass 1:1 [122, 183], 2:1 [123, 137, 148] and 3:1 [148, 182] states.

This variation is likely due to differences in the experimental conditions used in each study, as well as differences in how the proteins under investigation were produced. As discussed in the previous chapter, Lin28 expressed in *E. coli* cells is heavily contaminated with nucleic acid, and the production of stable, nucleic acid free Lin28 protein is non-trivial. Most of the studies involved expressing Lin28 in *E. coli* cells without sufficient precautions for removing nucleic acid and maintaining the presence of Zn^{2+} ions in the ZnK domain during purification. The differences in the reported affinities could therefore be due to differences in the protein samples used, rather than other factors, such as the addition of yeast tRNA to the sample buffer, as suggested previously [81]. In addition, only the study of the *Xtr*Lin28B protein involved quantitative biophysical measurements in addition to gel shift analysis. Whilst useful insights can be gained by gel shift analysis, image interpretation introduces problems not encountered by quantitative techniques. The binding equation used to determine the K_d is also critical for obtaining accurate data; a simplistic one-site binding model may not accurately describe the data for a system where multiple interactions take place. A key goal is therefore to accurately determine the affinities and stoichiometries of Lin28/let-7 complexes, in order to elucidate and refine the binding mechanism.

Lin28 is able to recognise many RNA sequences through the GGAG motif [136, 137]. This is a short motif compared to the size of the whole transcriptome. This poses the question of how Lin28 is able to specifically recognise and bind the subset of mRNA and miRNA target sequences out of the entire transcriptome through such a short motif in an efficient manner.

The fusion proteins produced in the previous chapter could be purified in high yields, and were stable and free from nucleic acid contaminants. This makes them amenable for biochemical analysis. Therefore, to address the questions of Lin28 specificity, affinity and stoichiometry, these fusion proteins will be probed using quantitative biophysical techniques.

4.2 Materials and Methods

4.2.1 DSF

DSF experiments were conducted using a Stratagene Mx300SP qPCR system (Agilent Technologies) and strips of 8, 200μL PCR tubes. GST or GST tagged protein was diluted to 0.5mg/ml in the buffer containing 10mM TRIS pH7.5, 150mM NaCl, 2mM DTT, 10% w/v sucrose, and 1x SYPRO® orange dye. 15 μL of the diluted protein was then mixed with 15 μL of the same buffer, or buffer and a 1.5x molar excess of nucleotide, and vortexed briefly, followed by a 1 minute centrifugation at 865 x g. Fluorescence readings were then taken at 492/610nm excitation/emission wavelengths every minute (corresponding to peaks of 517/585nm for SYPRO® orange), with a temperature increase of 1°C/min. The resultant curves were then analysed in MATLAB using the MTSA program [6] to find the inflection point of the curve, equivalent to the average melting temperature of the proteins in solution, either free or in complex. In the cases where more than one inflection point was present, only the data points corresponding to a single transition curve were used in the analysis.

4.2.2 Fluorescence Anisotropy: GST fusion proteins

Fluorescently-labelled oligonucleotides were ordered from either Dharmacon (RNA) or MWG (DNA). Nucleic acid was synthesized with a 5' fluorescein (RNA) or FITC (DNA) tag for use in the fluorescence anisotropy experiments. Each oligonucleotide was diluted to 50nM in a 500μL volume of 10mM TRIS pH7.5, 150mM NaCl and 5% sucrose, in a quartz cuvette. Fluorescence was measured using a HORRIBA Jobin Yvon fluorimeter with polarizers aligned in an “L” format, with both sets of polarizers alternating between horizontal and vertical positions to produce a total of 4 measurements (V_V, V_H, H_V, H_H). Anisotropy was calculated according to *Equation 6* (see *Chapter 2*). The RNA was then titrated with small aliquots of protein at a range of concentrations. Dissociation constants were calculated by non-linear regression analysis using the GraphPad Prism® software package according to *Equation 14*, with *c* constrained to 50 to reflect the 50nM concentration of fluorescent nucleotide.

4.2.3 Fluorescence Anisotropy: MBP fusion proteins

MBP tagged proteins were added to different concentrations in 200µL solutions in a 96 well black plate containing a buffer solution and 20nM fluorescent nucleotide. The buffer consisted of 20mM TRIS pH 7.5, 100mM NaCl, 10mM β-mercaptoethanol, 50µM ZnCl2 and 0.01% Tween20. Fluorescence readings were obtained with a BMG POLARstar Optima plate reader, with polarizers mounted in a "T" format for simultaneous reading of vertically and horizontally polarized light. Anisotropy was then calculated according to *Equation 6* (see *Chapter 2*). Dissociation constants (K_d) were then calculated by non-linear regression analysis using the GraphPad Prism® software package with one of the following equations. For situations where the Lin28 is in a 1:1 complex with the target nucleic acid, the standard quadratic binding equation was used (*Equation 14, Chapter 2*). In some cases, the two binding sites of Lin28 were (given the assumptions mentioned below) equivalent. In these cases, attempting to fit a two-site equation to the data would cause the variables to become co-dependent, indicating the equation is over-parameterized. *Equation 14* could be fit to the data well, however and any distinction between the affinities for each binding site could not be seen in the data. In most cases, the value of *c* was constrained to 20, reflecting the 20nM concentration of fluorescent nucleotide used in each well. Outliers were identified and removed automatically by the software package using the ROUT (Robust regression and Outlier removal) method [184]. Statistical values were calculated automatically by the regression software. All fits and residual plots are shown in Figures A1-9 and values determined by non-linear regression, as well as fitting statistics, are shown in Tables A1-4 (*see Appendix*).

4.2.4 SEC-MALLS

A Biosep SEC S3000 column (Phenomenex) pre-equilibrated with buffer was connected to a Dawn Helios II 18-angle light-scattering detector (Wyatt Technology). Lin28 samples were diluted to 2 mg/ml (~30-40μM) in the same buffer, or buffer combined with oligonucleotide. The eluting species were detected by measuring the UV absorbance at 280 nm, the concentration of the species was determined using an Optilab rEX refractometer (Wyatt Technology) and a refractive index increment of 0.18 mL/g was used for calculation of the molecular weight. For GST tagged proteins, a running solution contained 10mM TRIS pH 7.5, 500mM NaCl, 2mM DTT and 5% sucrose was used. For MBP tagged proteins, the running solution contained 20mM TRIS pH 7.5 and 250mM NaCl.

4.3 Results

4.3.1 GST-Lin28A is stabilised by let-7 miRNAs *in vitro*

DSF was used to determine whether GST fusion proteins interact with let-7 sequences. Full-length GST-Lin28 and constructs with truncated termini were mixed with SYPRO® orange dye and their melting curves were measured. A bi-phasic shift was observed for the protein on its own, with the first phase corresponding to the melting of Lin28 at 37°C and the second corresponding to the melting of GST at 52°C (**Fig. 4.5a**). When mixed in a 1:1 molar ratio with P2 let-7g RNA (see *Chapter 2*), the GST-Lin28 proteins had a monophasic shift with a melting temperature of ~52°C. The shift in melting temperature of the Lin28 curve from 37°C to 52°C implies that the Lin28 portion of the protein is being stabilised by the RNA in a manner consistent with binding. GST on its own does not bind let-7 RNA [112]. The experiments were repeated with other preE-let-7 family members: let-7a1, let-7d and let-7e (**Fig. 4.5b**). Shifts from 37°C to 52°C were seen, demonstrating that GST-Lin28 fusion proteins are stabilised by let-7 miRNAs in a manner consistent with binding.

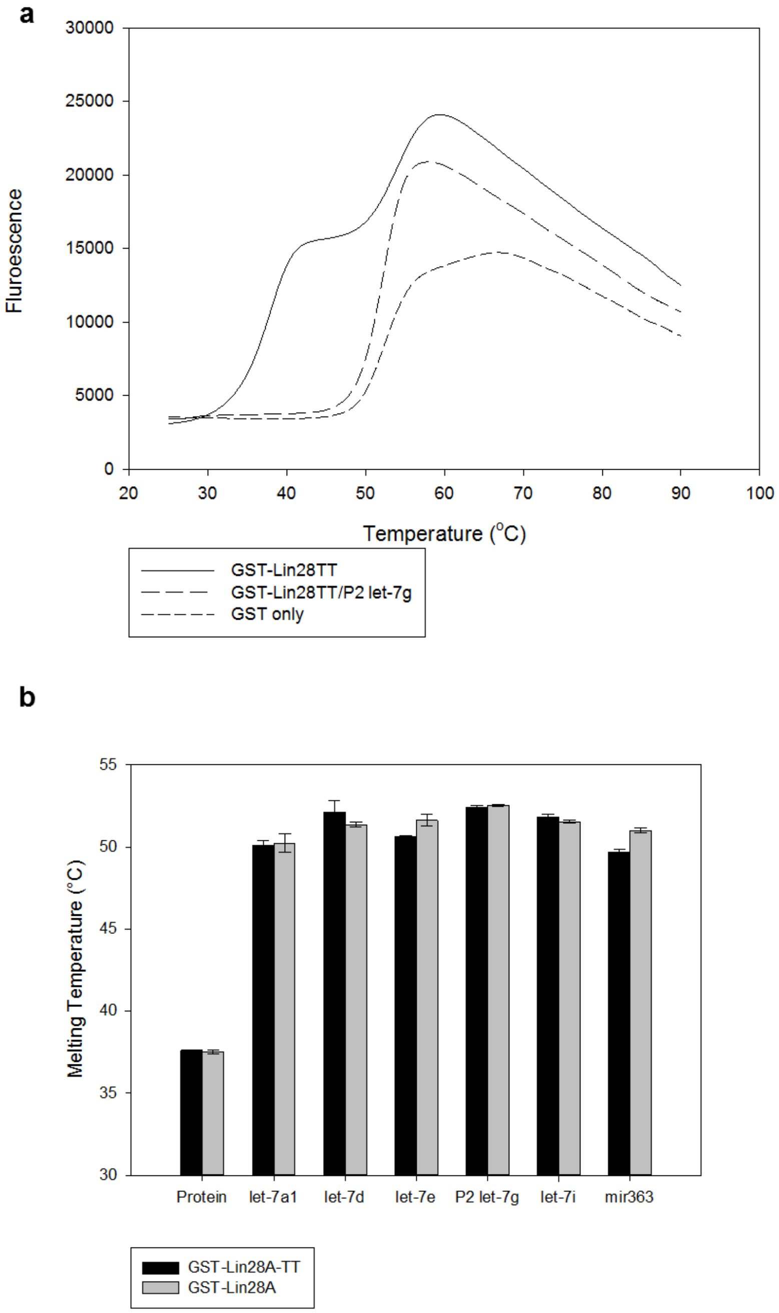


Figure 4.5: GST-Lin28TT is stabilised by miRNA segments *in vitro*.

(a) DSF melting curves show increased fluorescence intensity of the SYPRO® orange dye with temperature for the GST and GST-Lin28TT proteins, as well as the GST-Lin28TT/let-7g complex. The increased melting temperature of the Lin28/P2 let-7g mix is consistent with binding. (b) Melting temperatures of the fusion protein alone and when mixed with let-7 family miRNAs.

4.3.2 GST-Lin28A is stabilised by non-let-7 RNA and DNA oligonucleotides in a manner consistent with binding

DSF experiments were then repeated with a sequence derived from the terminal loop of the human mir-363 miRNA (mir363, see *Chapter 2*). mir-363 had been shown previously to be a potential binding partner of Lin28, as it contained a GGAG motif that was conserved throughout vertebrates, but in contrast to the let-7 miRNAs did not stimulate much uridylation activity when bound by Lin28 [92]. In addition its expression was found to be positively upregulated by Lin28 in *Xenopus tropicalis* by gene array experiments and binding to both native and recombinant *Xtr*Lin28 was observed by gel shift (Warrender et al. unpublished data). The results show that when mixed with mir363, the melting temperature of Lin28 shifts from 37°C to 52°C (**Fig. 4.5b**) similarly to the let-7 family RNA segments. This behaviour is consistent with binding.

The binding of Lin28 to other oligonucleotides was then probed by DSF. poly-T DNA sequences between 6 and 20 nucleotides in length were mixed with GST-Lin28TT. A step-wise increase in stability was seen with the increase in length of the oligonucleotide (**Fig. 4.6a**). The shift demonstrates that Lin28 is stabilised upon addition of poly-T DNA oligonucleotides in a manner consistent with binding, and correlates with the observed behaviour of the CSD of *Xtr*Lin28B [123].

To investigate differences between RNA and DNA sequences, poly-U RNA oligonucleotides were tested in the same way (**Fig. 4.6b**). Again, a step-wise increase in stability with the increasing length of RNA oligonucleotide was seen, although the T_m values calculated for each length were lower than for the poly-T sequences. No interaction could be observed between the dye and nucleic acids in the absence of protein (**Fig. 4.6c**). Therefore, Lin28 is stabilised by both non-let-7 RNAs and poly-pyrimidine oligonucleotides in a manner consistent with binding.

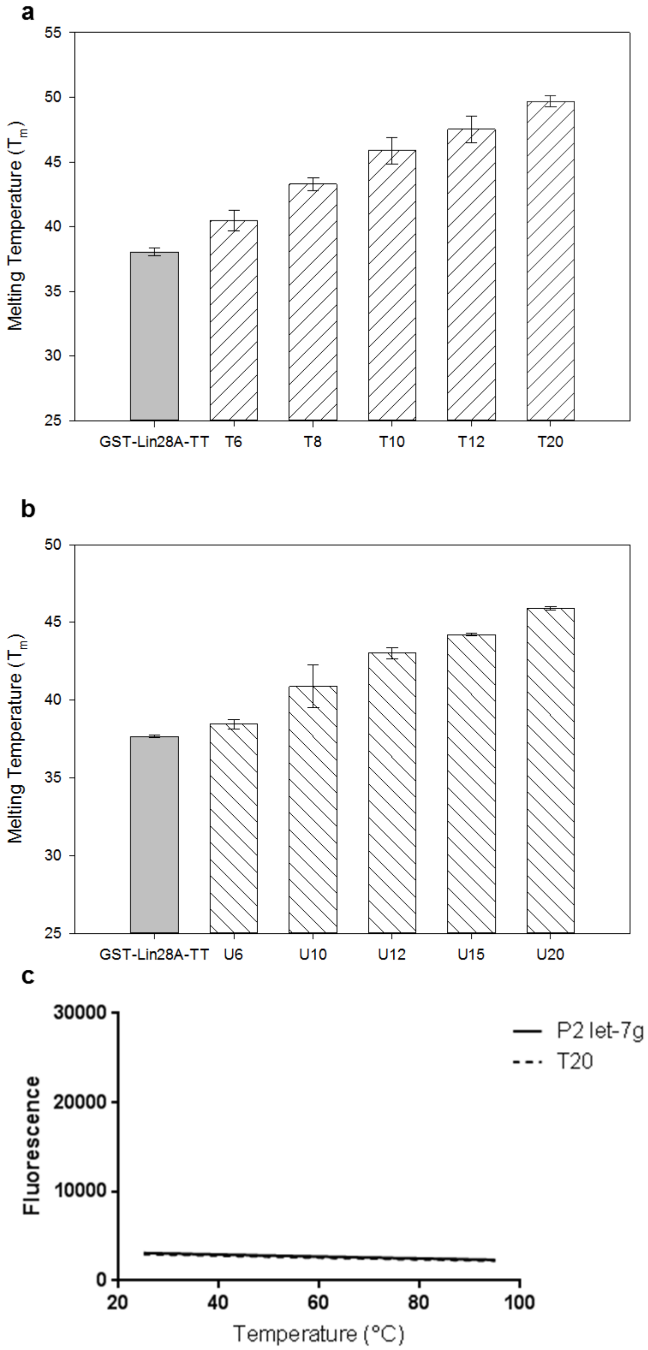


Figure 4.6: Lin28 is stabilised by nucleic acid segments of different length in a manner consistent with binding.

Melting temperatures obtained by DSF show Lin28 is stabilised by poly-pyrimidine DNA (a) and poly-U RNA (b) segments of different lengths. (c) DSF experiments conducted with nucleic acids in the absence of protein show no increase in fluorescence with temperature.

4.3.3 Affinity of GST-Lin28 towards nucleic acids

To further investigate the binding of nucleic acids by Lin28, fluorescently labelled preE-let-7g RNA was titrated with increasing concentrations of GST-Lin28ATT, and binding analysed by fluorescence anisotropy (**Fig. 4.7**). Anisotropy increased up to a plateau reached at a protein concentration of ~400nM. The data were fit with *Equation 14* resulting in a deduced K_d of 145.8 ± 25.2nM. The binding of GST-Lin28A was then tested with preE-let-7g and mir363 RNA oligonucleotides with data fit in each case by *Equation 14*, resulting in deduced K_d's of 156.8 ± 15.2nM and 21.4 ± 5.7nM respectively. Binding of GST-Lin28A was then tested with a variant of the mir363 oligonucleotide where the GGAG motif was mutated to AAAA (mir363(AAAA)). Data were fit with *Equation 14*, resulting in a deduced K_d of 68.7 ± 16.1nM. The ZnK domain is not able to bind to an AAAA motif [123] and so these results demonstrate that Lin28 will bind non-specific RNA sequences through its CSD. However, on examination, each binding curve is not an ideal fit, with R^2 values ≤ 0.98. This suggests that the single site binding equation may not fully describe the data.

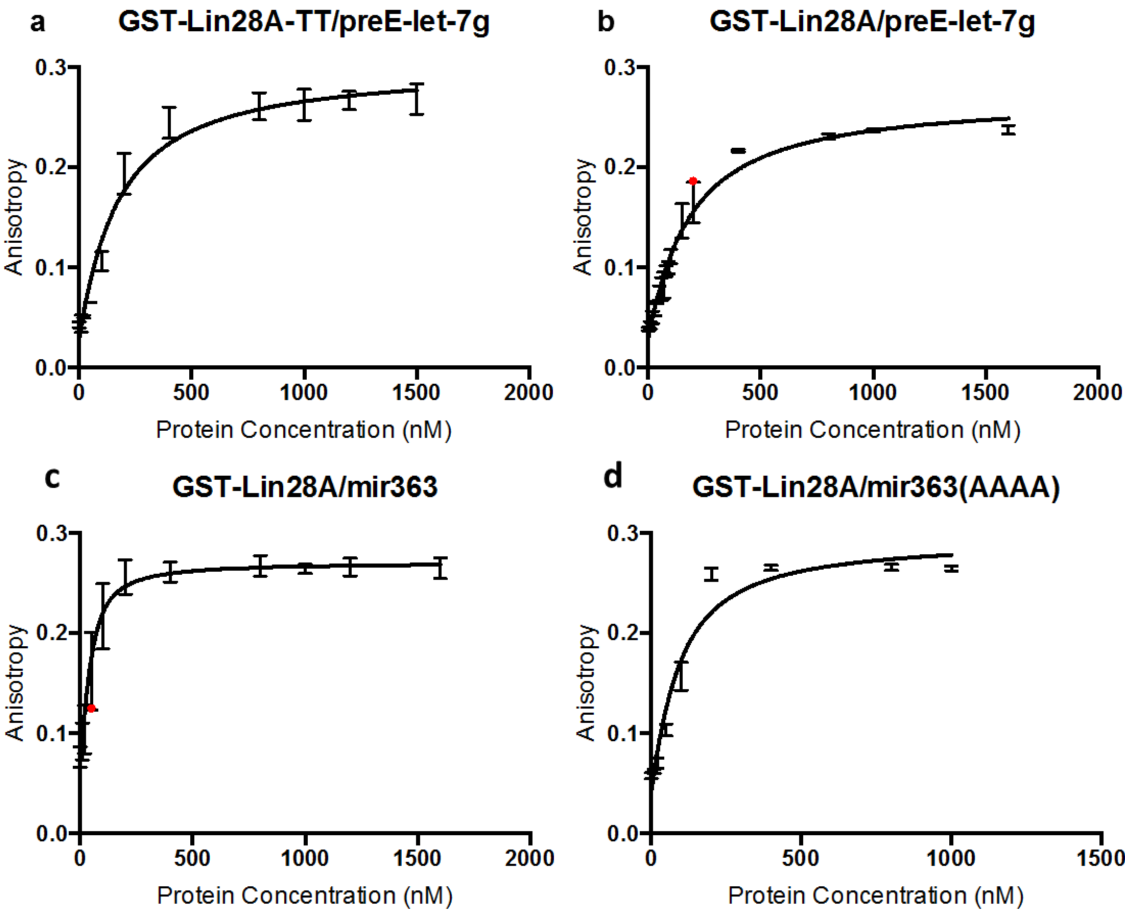


Figure 4.7: Fluorescence anisotropy analysis of GST fusion protein/RNA interactions.

Measurements were performed using preE-let-7g with GST-Lin28A (a) and GST-Lin28A-TT (b), as well as GST-Lin28A with mir363 (c) and mir363(AAAA) (d) RNA oligonucleotides. Outlying points omitted during fitting are highlighted in red.

4.3.4 Oligomeric state of the GST-Lin28A/let-7 complex

The oligomeric state the GST-Lin28/let-7 complexes was then determined by SEC-MALLS. Initially, the GST tag on its own was analysed (**Fig. 4.8a**). Although the molecular weight was variable across the UV peak, the elution profile was symmetrical. The average molecular weight of the peak was 49kDa, corresponding to the expected 45kDa of a dimeric GST proteins. GST-Lin28A was then applied to the column in the same conditions. Here, the average molecular weight was determined to be 94kDa, which compared with the expected 95kDa of a dimeric GST-Lin28A (**Fig. 4.8b**).

*t*preE-let-7g RNA (see *Chapter 2*) was then mixed in a 1:1 molar ratio with the protein and applied to the column. Again, though the molecular weight profile was variable, the average molecular weight was observed to be 113kDa (**Fig. 4.8c**). This corresponds to a single GST-Lin28A dimer, of 95kDa, plus two *t*preE- let-7g molecules, each of 9.5kDa, to make a total expected molecular weight of 114kDa.

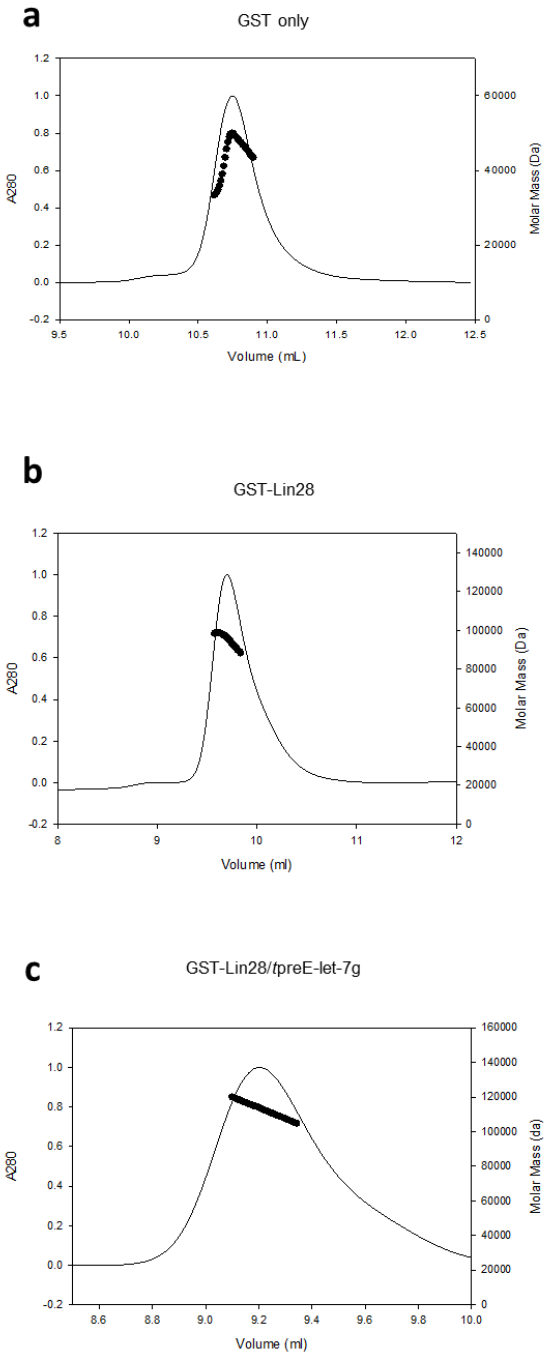


Figure 4.8: SEC-MALLS analysis of GST and GST fusion proteins.

Elution profiles of standalone GST (a) GST-Lin28 (b) and an equimolar mix of GST-Lin28 and *t*preE-let-7g (c).

Therefore, it was observed that GST and GST-Lin28A form dimers, and that for each GST-Lin28A dimer, two molecules of RNA were bound. This corresponds to a 1:1 Lin28:RNA binding stoichiometry. However, due to the complications involved in determining the stoichiometries and affinities of Lin28/RNA complexes with proteins dimerised through their GST tags, work with the GST-fusion proteins was suspended.

4.3.5 Determination of the oligomeric state of His-MBP-4A-Lin28ATT

To ascertain the suitability of the MBP fusion protein for binding studies, the oligomeric states of the unbound protein and its complexes with RNA were investigated. SEC-MALLS results demonstrated that the MBP tag on its own formed monomers, with the major elutant species of molecular weight 43.4kDa, comparing to an expected size of 42.2kDa for a monomeric state (**Fig. 4.9a**). The major species in the Lin28TT and CSD protein samples were also monomeric, with masses of 64.7kDa and 51.5kDa comparing to expected values of 59.4kDa and 52.7kDa respectively (**Fig 4.9 b,c**).

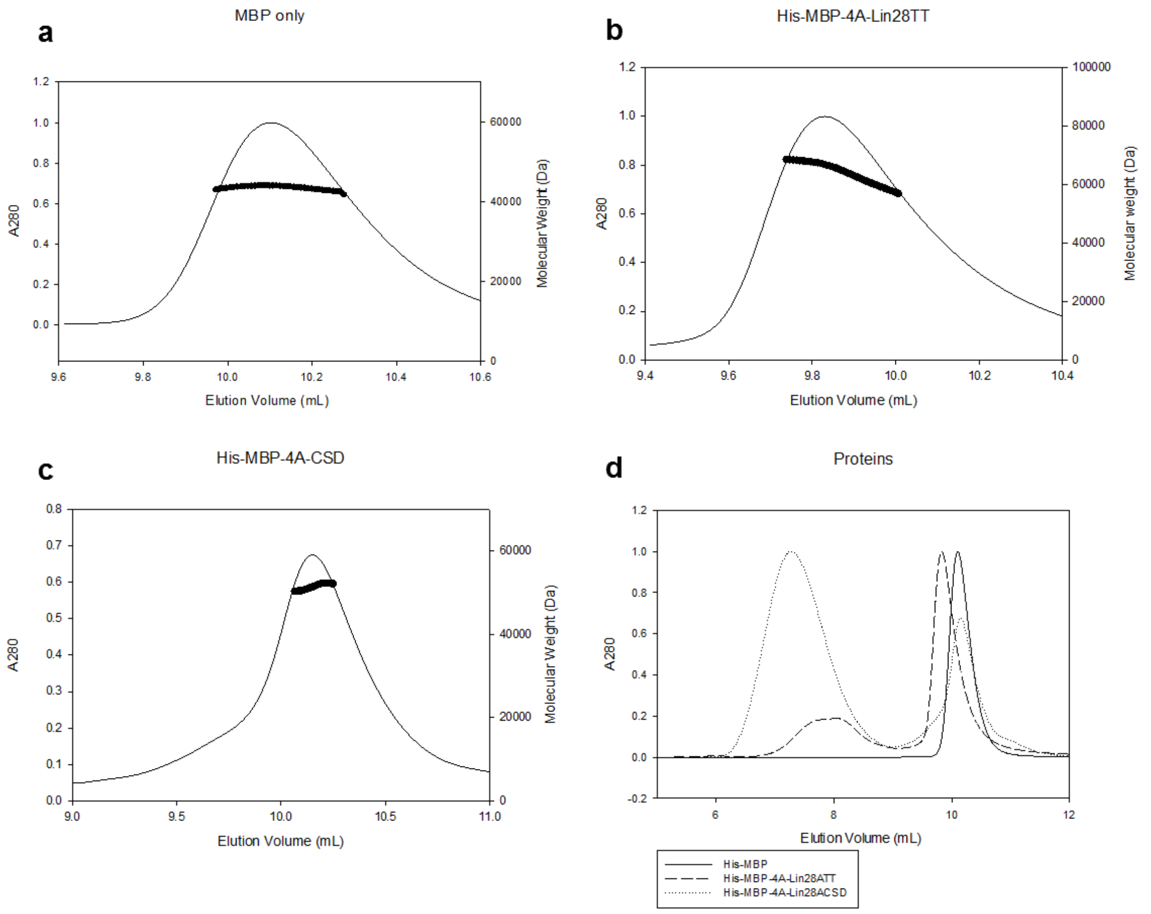


Figure 4.9: SEC-MALLS analysis of MBP and MBP fusion proteins.

The elution profiles of standalone MBP (a) His-MBP-4A-Lin28TT (b) and His-MBP-4A-CSD (c) fusion proteins are shown. (d) Overlay of the A_{280} absorbance of the elution peaks for the proteins used in this study.

The average molecular weights determined for *t*preE-let-7g and mir363 RNA, as well as dT29 DNA were similar to the expected values. All three species have molecular weights of ~10kDa and were observed to be 11.1kDa, 11.4kDa and 10.1kDa molecules respectively.

4.3.6 RNA-free Lin28 exists in equilibrium between active monomeric and aggregated states

All SEC-MALLS elution profiles of Lin28TT or CSD proteins contained an extended peak corresponding to species with large molecular weights. By overlaying the A_{280} absorbances of the complexes with their individual components (**Fig 4.9d, 4.10**), it can be seen that the size of this peak is dependent on the amount of nucleic acid initially added to the sample, where the unbound protein has the highest absorbance peak at this position.

The addition of nucleic acid to the protein sample causes the size of the high molecular weight peak to decrease. This implies that the protein that comprises this peak is active Lin28TT/CSD. This peak is of a greater molecular weight than the free protein, and is also present in the unbound protein sample, and so demonstrates that free Lin28 exists in equilibrium between monomeric and aggregated states. The disruption of these aggregates by the addition of nucleic acids shows that aggregation does not inhibit the activity of Lin28. In addition, the breadth of the peak implies that there is no fixed size of these aggregates and that their molecular weights are variable. This phenomenon was observed in both the truncated termini and CSD fusion proteins, suggesting the aggregation is due to the CSD, rather than the ZnK domain.

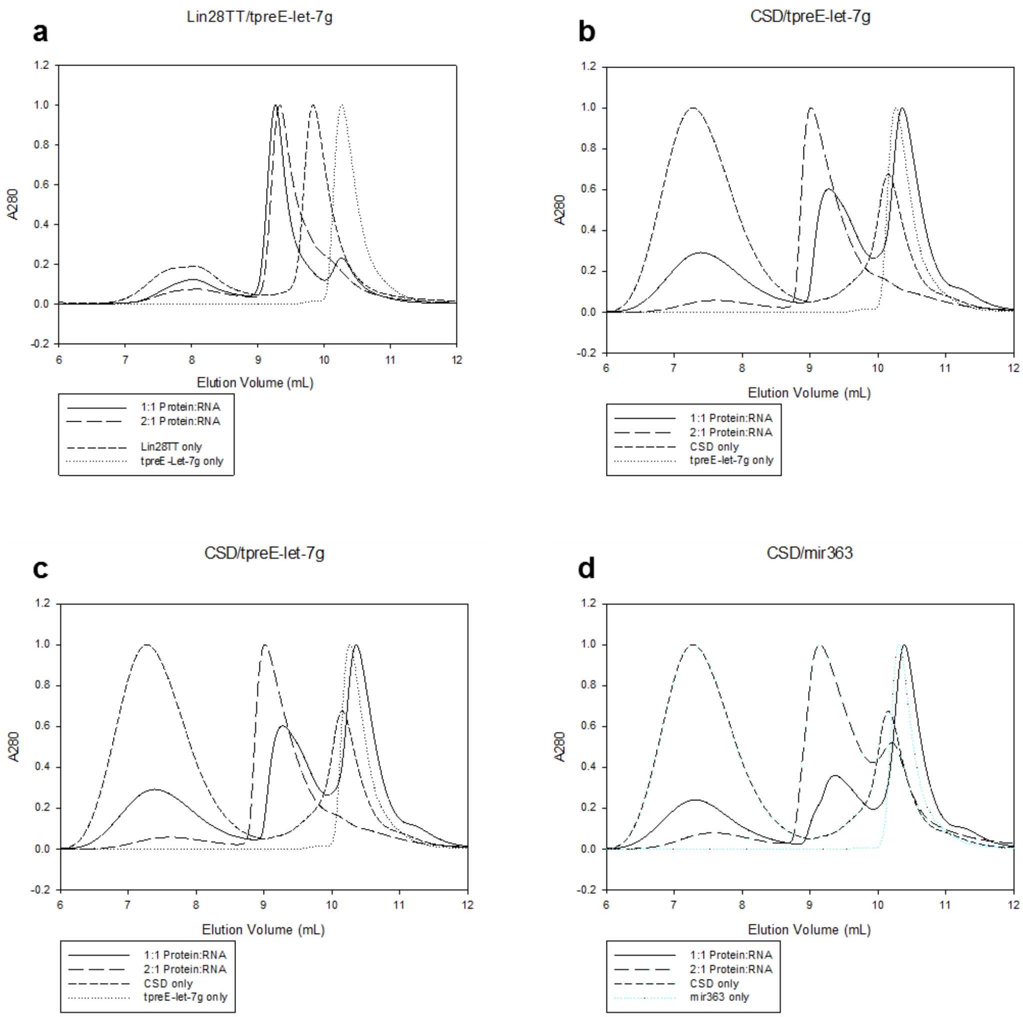


Figure 4.10: Elution profiles of Lin28TT/RNA complexes from SEC-MALLS.

A_{280} absorbance traces are overlaid from 1:1 and 2:1 protein:RNA molar ratio mixes, as well as the unbound Lin28TT or CSD proteins and the free nucleic acid. The profiles of complexes generated with *t*preE-let-7g and mir363 with Lin28TT (a, b) and CSD proteins (c, d) respectively are depicted.

4.3.7 Lin28 complexes with preE-let-7g in a 1:1 stoichiometry

The affinity of the MBP-Lin28 fusion proteins for preE-let-7g RNA was investigated using fluorescence anisotropy (**Fig. 4.11a**). *Equation 14* was fit to the data and a K_d of 38.8 ± 3.0nM was deduced. No binding was observed towards the His-MBP tag on its own, demonstrating that the Lin28 portion of the fusion protein is active (**Fig. 4.12**).

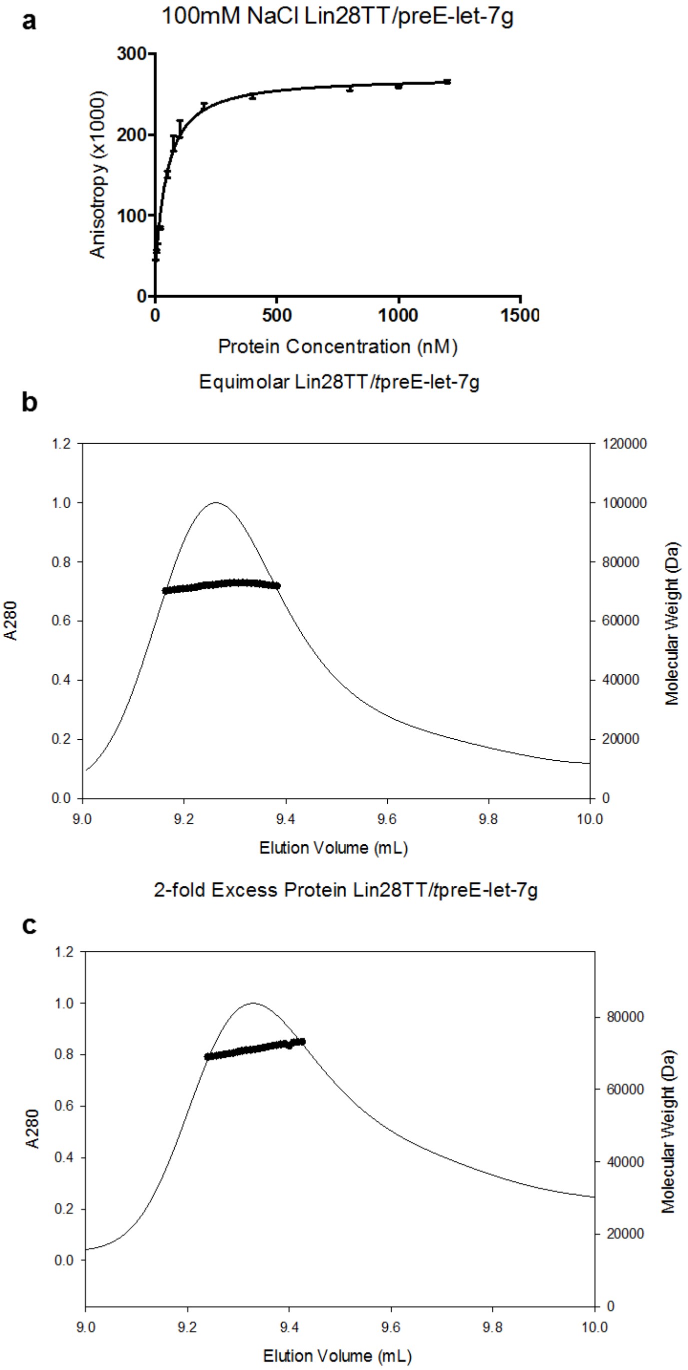


Figure 4.11: Affinity and Stoichiometry of Lin28/let-7g complexes.

(a) Fluorescence anisotropy change of preE-let7g when titrated with Lin28TT. (b) and (c) show SEC-MALLS elution profiles of the major complex peaks observed when Lin28TT is mixed with *t*preE-let-7g in 1 and 2 fold molar excesses respectively.

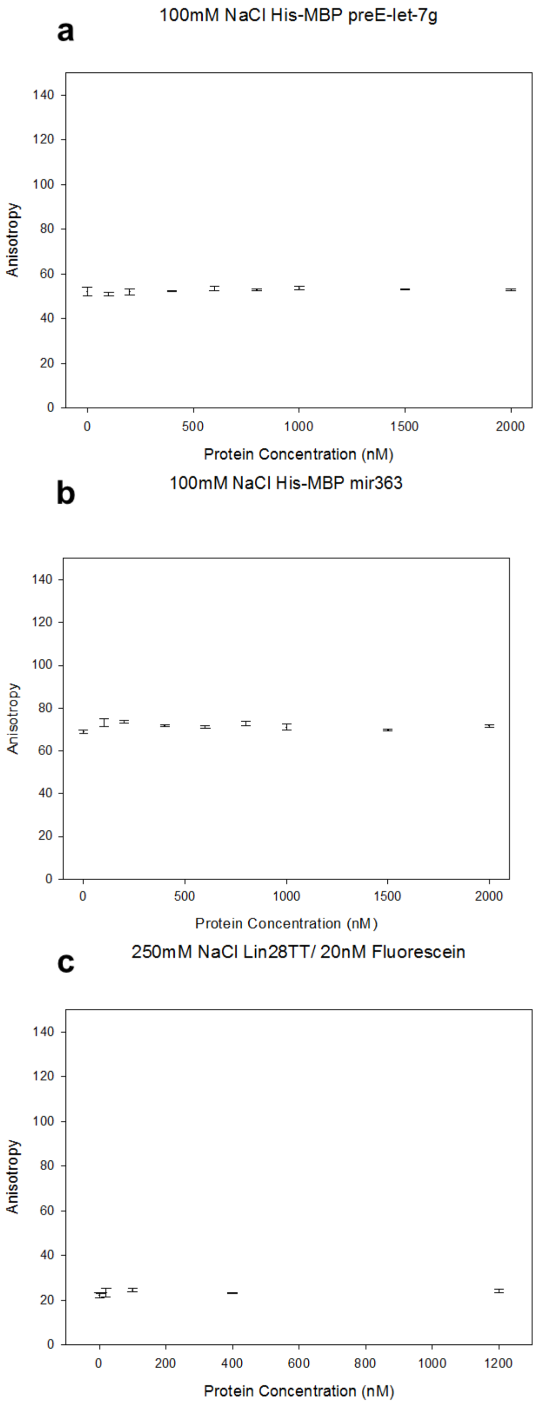


Figure 4.12: Fluorescence anisotropy of MBP/RNA and Lin28TT/Fluorescein mixes.

No major change in anisotropy could be seen when His-MBP protein was added to preE-let-7g (a) or mir363 (b) RNA at 100mM NaCl. (c) no major change in anisotropy was observed when Lin28TT was added to free fluorescein at 250mM NaCl.

The stoichiometry of the complex was then determined by SEC-MALLS. Lin28TT was mixed with *t*preE-let7g in equimolar amounts as well as with a 2-fold molar excess of protein. Molecular weights of 71.5kDa and 70.6kDa were observed (**Fig. 4.11 b,c**), corresponding to a 1:1 binding stoichiometry of Lin28 to let-7, which would have an expected molecular weight of 68.4kDa. Therefore, the Lin28TT fusion protein is active, and binds the let-7 terminal loop sequence with high affinity in a 1:1 stoichiometric ratio.

4.3.8 Lin28 binds the terminal loop of the pre-mir363 miRNA

mir363 RNA was found to interact with GST-Lin28 fusion proteins. The binding of this sequence by the MBP fusion proteins was determined by fluorescence anisotropy, with data fit by *Equation 14*. High affinity binding of this sequence (K_d = 16.6 ± 1.9nM) by the Lin28TT protein was observed (**Fig. 4.13a**).

The stoichiometry of the Lin28TT/mir363 complexes was then determined by SEC-MALLS. A 1:1 binding stoichiometry was observed, with molecular weights of 72.7kDa and 77.5kDa obtained (**Fig. 4.13 b,c**) for equimolar and 2-fold protein excess mixtures respectively, compared to an expected size of 68.1kDa for a 1:1 binding stoichiometry. The weight of the complex in the excess of protein was larger than expected due to the presence of a minor shoulder at a lower elution volume than the main complex peak, which resulted in a higher average reading.

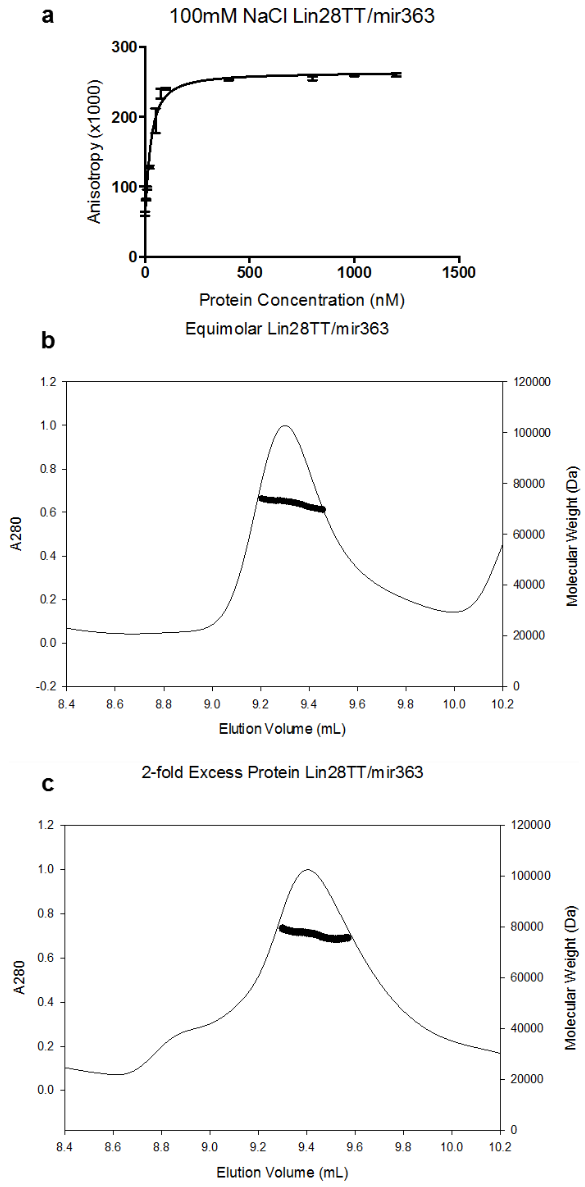


Figure 4.13: Affinity and Stoichiometry of Lin28TT/mir363 complexes.

(a) Anisotropy change of mir363 when mixed with different concentrations of Lin28TT. (b) and (c) show SEC-MALLS chromatograms of the major complex peaks observed when Lin28TT is mixed with mir363 in 1 and 2 fold molar excesses respectively.

4.3.9 The Lin28 ZnK domain is required for high affinity, 1:1 binding

The stoichiometries of the CSD protein with both *t*pre-let7g and mir363 were then investigated. Surprisingly, in both cases the CSD complexes formed had molecular weights intermediate between 1:1 and 2:1 stoichiometries, with weights of 76.6kDa and 80.3kDa for 1 and 2 fold molar excesses of protein to RNA respectively for *t*pre-let7g (**Fig. 4.14 a,b**), and 74.8kDa and 85.2kDa for mir363 (**Fig. 4.14 c,d**). These compared to expected 1:1 and 2:1 molecular weights of ~61kDa and ~114kDa. This suggests that the two stoichiometric states exist in equilibrium with each other.

To calculate K_d values for the interactions of the CSD with RNA, *Equation 14* was fit to the data with the following assumptions. Each RNA has two single stranded regions of >7nt available, in the loop region and the 3' end. This is known to correspond to the binding site size of the CSD from structural data [122, 123]. It was therefore assumed that the maximum binding density of each nucleic acid would be two CSD molecules, and that each site would be recognized equally and independently by the CSD. In addition, it was assumed the equilibrium between the active aggregate and monomeric forms of Lin28 was not rate limiting, and would not interfere with RNA binding. While it is possible that there are differences between the binding sites, or that binding of one CSD affects the other, no distinction could be seen in the data and more complex models were over-parameterised, with co-dependent variables. Therefore, the data had to be fit with the simpler model defined by *Equation 14*, which is possibly inappropriate for describing this system. The K_d values calculated were 181.5 ± 12.3nM for the CSD/preE-let-7 interaction and 76.2 ± 7.3nM for the CSD/mir363 interaction (**Fig. 4.14 e,f**). This demonstrated that the CSD alone can bind these sequences, but that the ZnK domain is necessary for high affinity, 1:1 binding.

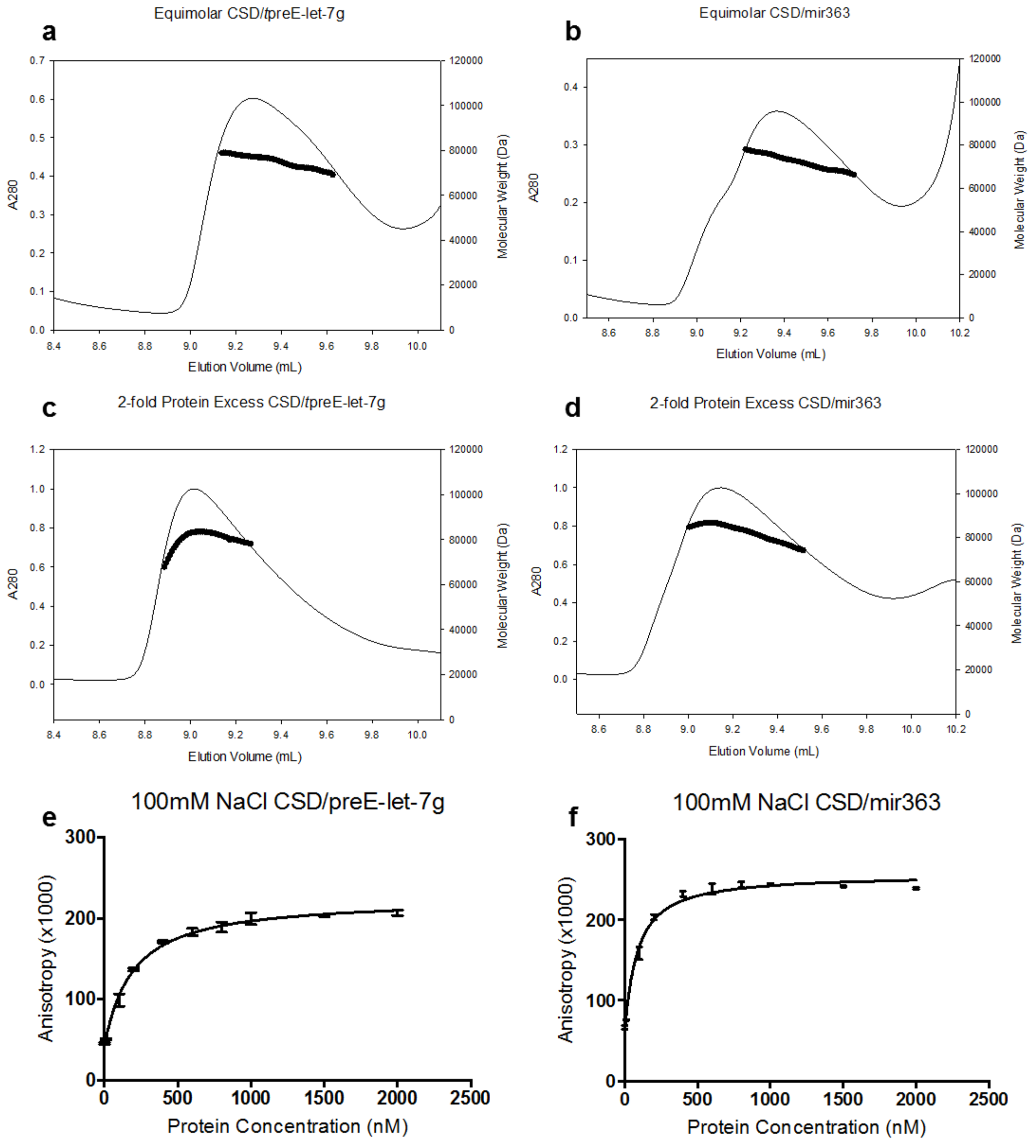


Figure 4.14: Affinities and stoichiometries of CSD/RNA complexes.

SEC-MALLS elution profiles of the major complex peak resulting from 1 and 2 fold molar excesses of CSD to *t*preE-let-7g are shown in (a) and (b), and with mir363 in (c) and (d). (e) and (f) represent the changes in anisotropy of preE-let-7g and mir363 RNAs when mixed with different concentrations of CSD protein.

4.3.10 Lin28 non-specifically binds GGAG mutant sequences

To further investigate the effect of the ZnK interaction with RNA, binding of Lin28TT to the mir363(AAAA) sequence was tested. Two peaks of average molecular weights 122.9kDa and 84.1kDa were observed when the RNA was mixed in an equimolar ratio with the Lin28TT protein (**Fig. 4.15a**), compared to expected sizes of ~128kDa and ~68kDa for 2:1 and 1:1 binding stoichiometries. The two peaks therefore likely correspond to 2:1 and an intermediate stoichiometry respectively. When mixed with a 2 fold molar excess of protein (**Fig. 4.15b**), only one peak of 120.8kDa could be seen, indicating that all of the RNA was bound in 2:1 protein:RNA complexes.

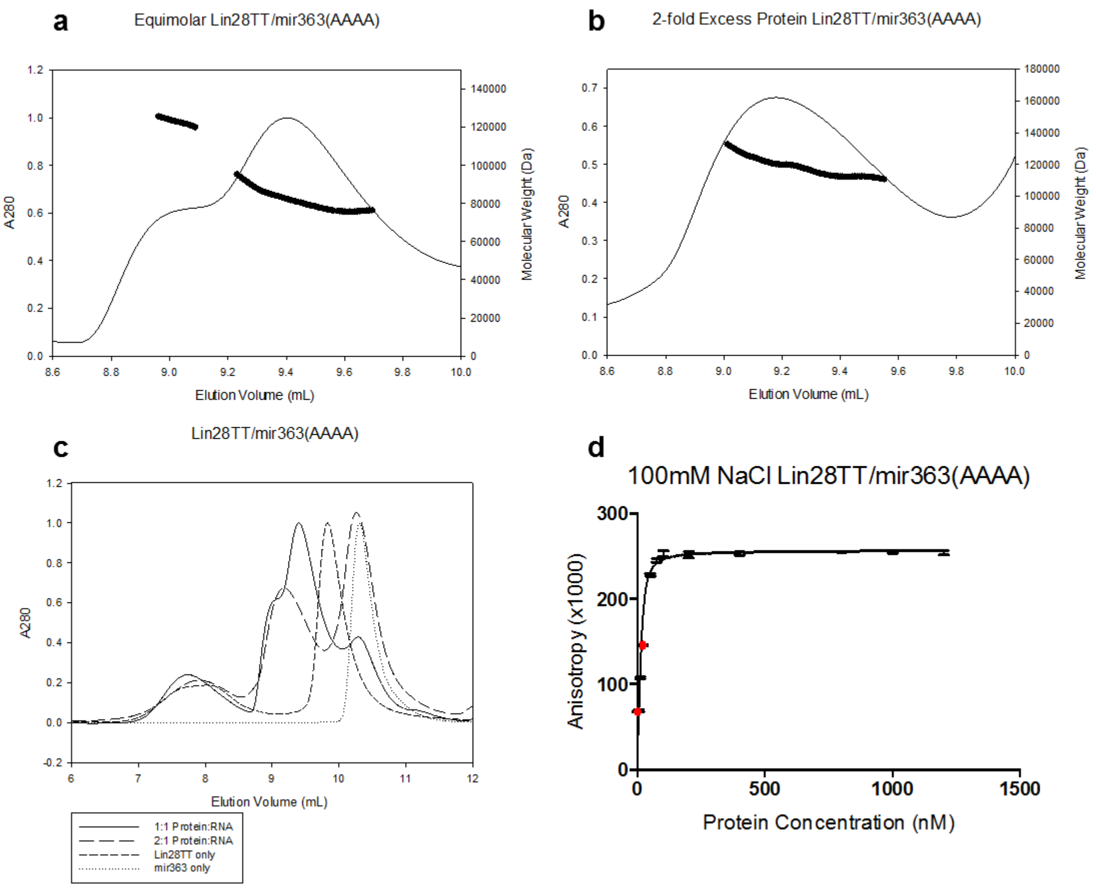


Figure 4.15: Stoichiometry and affinity of Lin28TT/mir363(AAAA) complexes.

(a) and (b) show SEC-MALLS elution profiles of the major complex peaks observed when Lin28TT is mixed with mir363(AAAA) in 1 and 2 fold molar excesses respectively. (c) shows the A_{280} absorbance traces from 1 and 2-fold protein:RNA molar ratio mixes, as well as the protein and mir363 RNA on their own. (d) shows the change in anisotropy of mir363(AAAA) RNA when mixed with different concentrations of Lin28TT protein. Outlying points excluded from data fitting are highlighted in red.

Equation 14 was then used to determine the affinity of the interaction, which had a K_d of 3.8 ± 5.4nM (**Fig. 4.15d**). The stoichiometry of the CSD alone with the mir363(AAAA) RNA was also determined to be intermediate between 2:1 and 1:1 (**Fig. 4.16 a,b**), with an dissociation constant of 54.7 ± 6.1nM (**Fig. 4.16d**), indicating that the CSD behaves similarly to all nucleic acids regardless of the presence of a GGAG sequence.

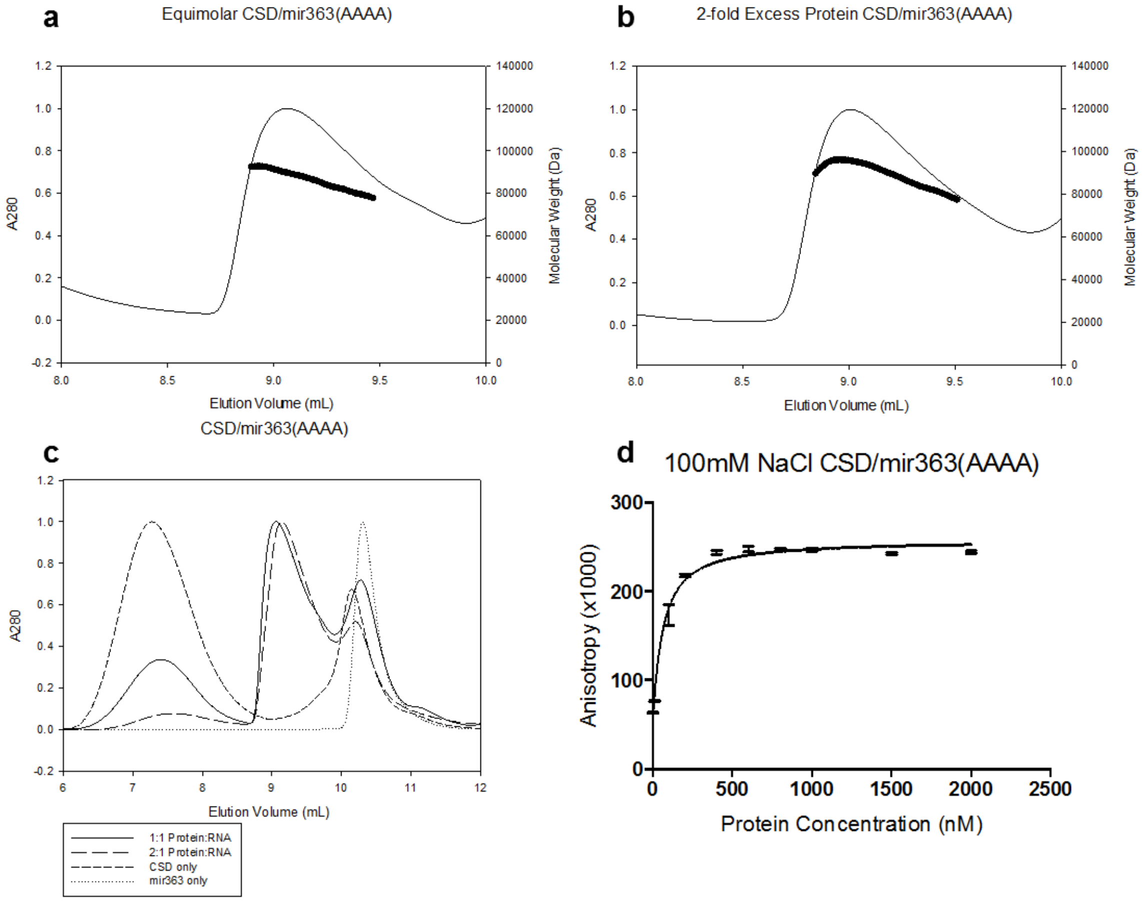
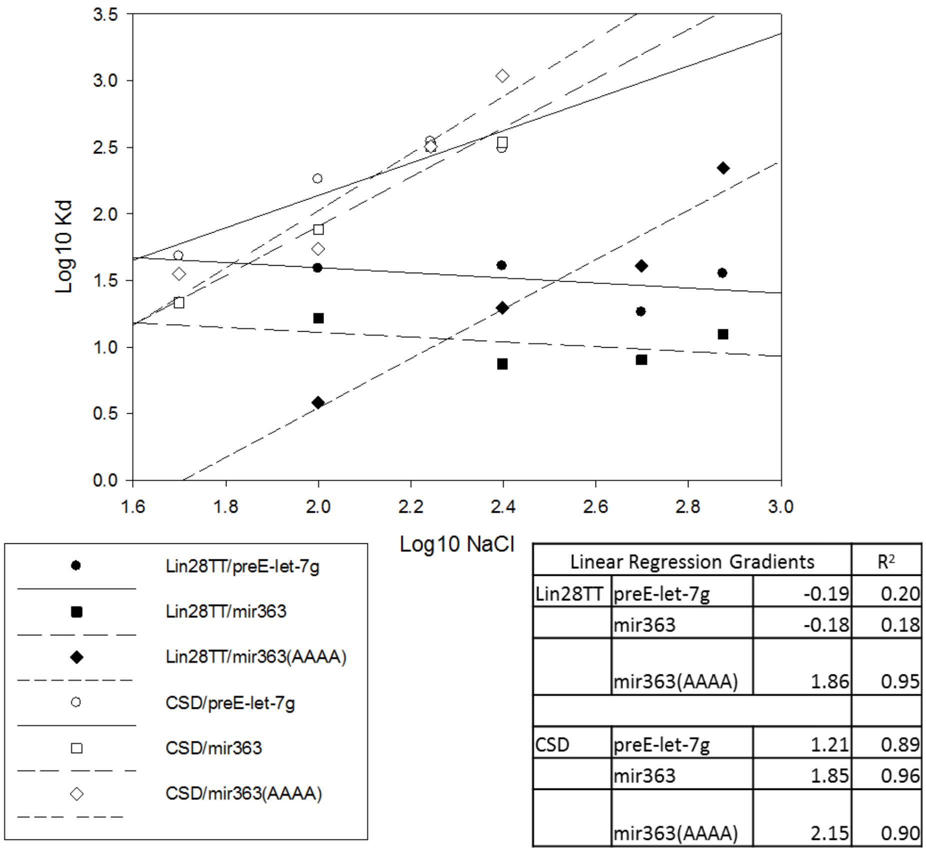


Figure 4.16: Stoichiometry and affinity of CSD/mir363(AAAA) complexes.

(a) and (b) show SEC-MALLS elution profiles of the major complex peaks observed when CSD is mixed with mir363(AAAA) in 1 and 2 fold molar excesses respectively. (c) shows the A_{280} absorbance traces from 1 and 2-fold protein:RNA molar ratio mixes, as well as the protein and mir363 RNA on their own. (d) shows the change in anisotropy of mir363(AAAA) RNA when mixed with different concentrations of CSD protein.

4.3.11 Non-specific Lin28/RNA complexes are highly sensitive to salt concentration

Non-specific protein-RNA complexes often form through electrostatic interactions. To investigate the specificity of the Lin28/RNA interactions, fluorescence anisotropy experiments were repeated with varying amounts of salt in the buffer. K_d's were determined using *Equation 14*. The Log_{10} of each K_d was plotted against the Log_{10} of the salt concentration in the buffer, to make the results easily comparable (**Fig. 4.17**).



Linear Regression Gradients			R²
Lin28TT	preE-let-7g	-0.19	0.20
	mir363	-0.18	0.18
	mir363(AAAA)	1.86	0.95
CSD	preE-let-7g	1.21	0.89
	mir363	1.85	0.96
	mir363(AAAA)	2.15	0.90

Figure 4.17: Influence of salt concentration on the affinity of Lin28 towards RNA.

Log-Log plot of K_d vs salt concentration for the interaction of Lin28TT or CSD with preE-let-7g, mir363 and mir363(AAAA).

The K_d of the interaction between Lin28TT and either preE-let-7g or mir363 did not vary greatly with increases in salt concentration, with linear regression gradients of -0.19 and -0.18 respectively. This implies that these interactions do not have a strong electrostatic component. In contrast, the interaction with the mir363(AAAA) was highly sensitive to the ionic strength of the buffer solution, with a linear regression gradient of 1.86, and K_d increasing steeply in solutions with the higher ionic strength.

Similarly, the dissociation constants of the CSD/RNA interactions were also sensitive to the ionic strength of the buffer solution, to the extent that very little binding could be seen above concentrations of 250mM NaCl. The affinity of the CSD/preE-let-7g interaction was less dependent on salt concentration than for the mir363 and mir363(AAAA) sequences, with regression gradients of 1.21, 1.85 and 2.15 for preE-let7g, mir363 and mir363(AAAA) respectively. However, the affinities of both the CSD/preE-let-7g and CSD/mir363 interactions were more sensitive to changes in the ionic strength of the buffer solution than the affinities of the Lin28TT/preE-let-7g and Lin28TT/mir363 interactions. This suggests that the interaction of the CSD domain alone with the RNA oligonucleotides has a greater electrostatic component than the interaction of the Lin28TT protein, which contains both CSD and ZnK domains, with the same RNA. Finally, the gradients of the linear regressions corresponding to the CSD interactions with the mir363 sequences were similar to that with the full-length protein with the mir363(AAAA). This means the variation of K_d with changes in the ionic strength of the buffer occurs similarly for the Lin28TT/mir363(AAAA), CSD/mir363 and CSD/mir363(AAAA) interactions. Each of these interactions therefore has a similar electrostatic component, and suggests that only the CSD is binding mir363(AAAA) in the case of the Lin28TT protein.

4.3.12 Lin28 binds dT29 in a maximum 2:1 stoichiometry

GST-Lin28 fusion proteins were bound to DNA sequences (see **Section 4.3.2**). To further investigate differences between specific and non-specific binding of Lin28, the stoichiometries of MBP-Lin28 fusion proteins in complex with dT29 were tested by SEC-MALLS. Lin28 will not bind specifically to any sequence elements in this oligonucleotide and so the stoichiometry of the complex will reflect how many Lin28 molecules can bind per oligonucleotide on average, and what the site size of the CSD is.

Equimolar and 2-fold protein excess mixtures of Lin28TT and dT29, as well as CSD and dT29, were analysed by SEC-MALLS (**Fig 4.18**). For the Lin28TT protein, three peaks were seen in the equimolar mix, of molecular weights 139.9kDa, 100.8kDa and 73.2kDa, which likely correspond to 2:1, intermediate equilibrium and 1:1 protein:RNA stoichiometric complexes, respectively. When the protein was added in excess, two peaks of molecular weight 138.4kDa and 95.9kDa could be observed, corresponding to a 2:1 protein:RNA stoichiometry and an intermediate equilibrium state respectively.

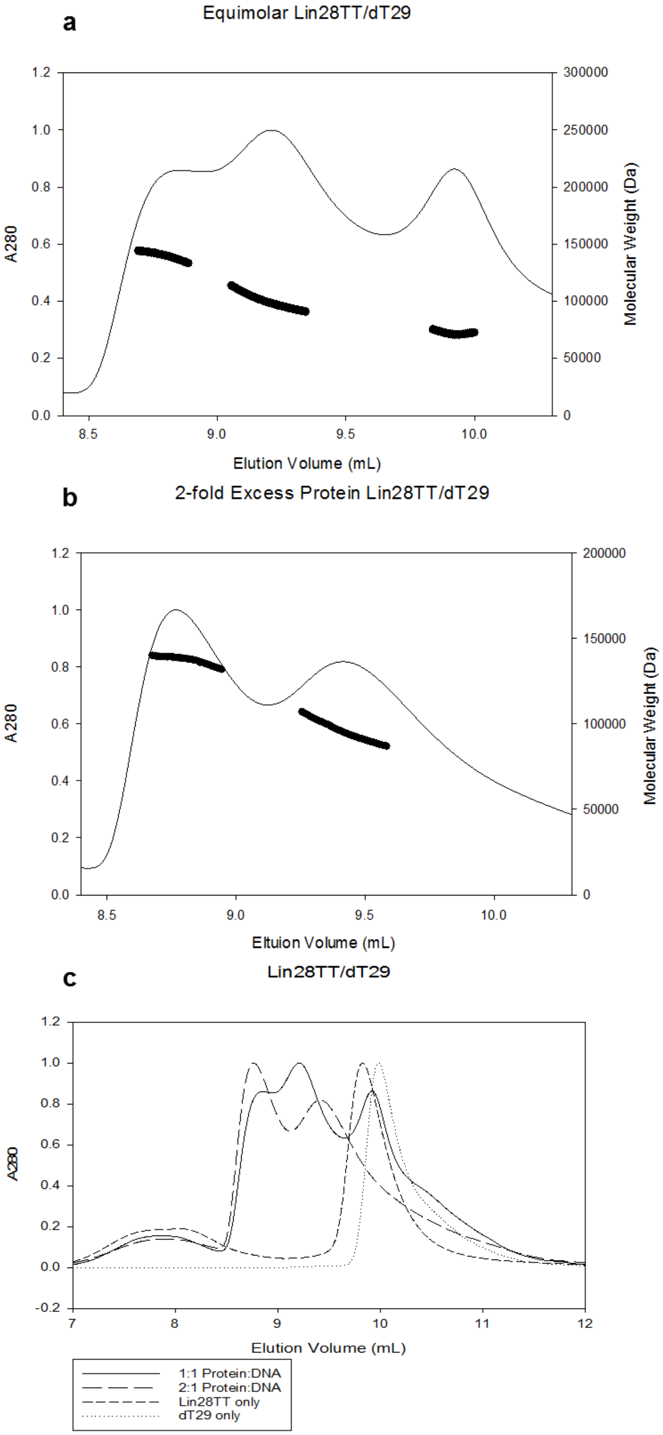


Figure 4.18: SEC-MALLS analysis of Lin28TT complexes with dT29 prepared in different molar rations.

(a) Protein-RNA complexes were mixed in 1:1 (a) and 2:1 (b) ratios. (c) Elution profiles (A_{280} absorbance) for each complex and for protein and DNA.

For the CSD protein, regardless of the molar ratios in the mixture, only one major peak corresponding to a 2:1 complex could be seen, with average molecular weights of 127.7kDa and 123.1kDa for the equimolar and 2-fold protein excess mixtures respectively (**Fig. 4.19**).

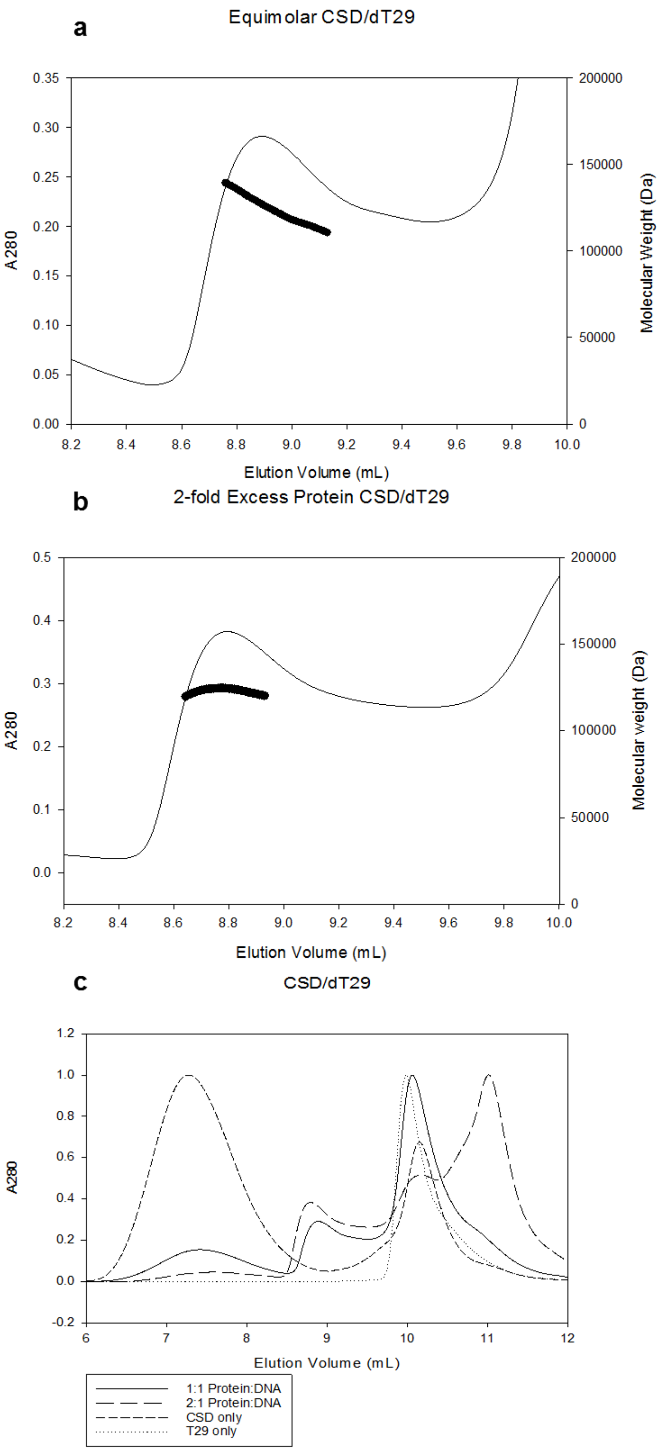


Figure 4.19: SEC-MALLS analysis of CSD complexes with dT29 prepared in different molar rations.

(a) Protein-RNA complexes were mixed in 1:1 (a) and 2:1 (b) ratios. (c) Elution profiles (A_{280} absorbance) for each complex and for protein and DNA.

For a sequence 29nt in length, therefore, the maximum binding density was two Lin28 proteins per oligonucleotide. The data imply that only the CSD is binding the DNA, as the binding density is equivalent for both the Lin28TT and CSD proteins. Site sizes for the CSD have been postulated to be between 7-11 [122, 123] nucleotides in length. For a 2:1 stoichiometric ratio, this would result in a combined binding site of 14-22nt, which correlates with the results seen here, as extra space between the proteins is needed to avoid steric hindrance. This could account for the extra seven nucleotides not directly bound by the CSD.

4.3.13 Lin28 binds let-7 mutant sequences

To determine the effect of the GGAG motif on the binding of preE-let-7 sequences by Lin28, the let-7mut sequence was used (see *Chapter 2*). Mutation of the GGAG to AAAA, as in the mir363(AAAA) sequence, would result in changes to the secondary structure of the miRNA, and so results would not be easily comparable. Instead, the 3' terminus of the *t*preE-let-7g sequence was truncated, in order to produce a shorter RNA consisting of only the let-7g loop, with no GGAG sequence present. SEC-MALLS analysis was conducted with equimolar and 2-fold protein excess mixtures of Lin28TT and let-7gmut and returned average molecular weight readings of 75.6kDa and 72.3kDa respectively (**Fig. 4.20 a,b**), corresponding to an expected 1:1 molecular weight of 67.9kDa. This implies the formation of a 1:1 stoichiometric complex.

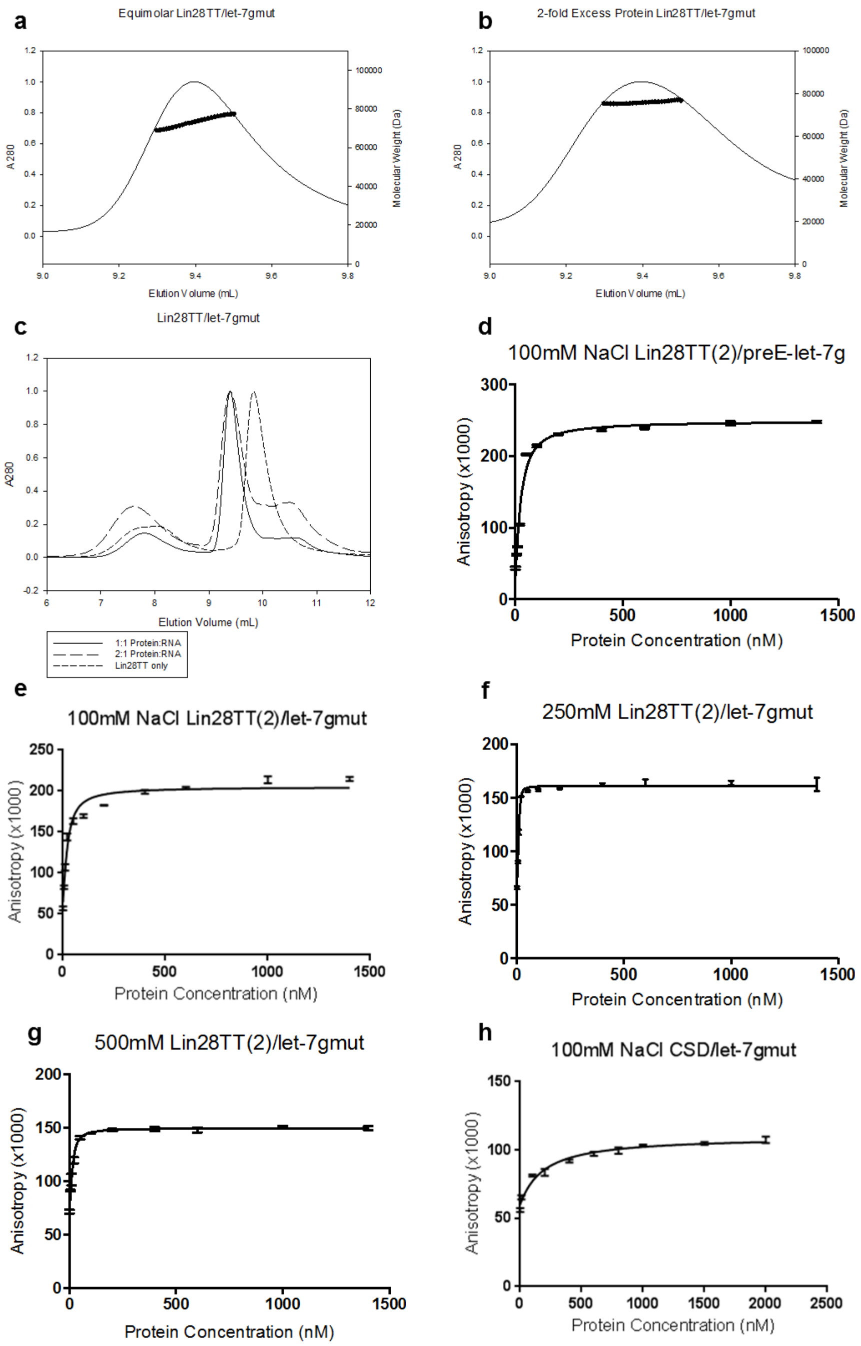


Figure 4.20: Interaction between Lin28TT and let-7gmut RNA.

SEC-MALLS elution profiles of Lin28TT mixed with let-7gmut in 1- (a) and 2-(b) fold molar ratios. (c) Elution profiles (A_{280} absorbance) for each complex and for standalone protein and RNA. (d) Anisotropy of preE-let-7g when mixed with different concentrations of Lin28TT(2) protein in 100mM NaCl. (e) Fluorescence anisotropy change of let-7gmut when mixed with different concentrations of Lin28TT(2) protein in 100mM NaCl. Anisotropy changes for the same interaction are shown in (e) and (f) at 250mM and 500mM NaCl concentrations respectively. (h) Fluorescence anisotropy change of let-7gmut when mixed with different concentrations of CSD protein in 100mM NaCl.

Affinity measurements were made using Lin28TT protein prepared by Dr.Vladimir Levdikov (Antson Group) using the protocol described in *Chapter 3, Section 3.3.2*. This protein will be referred to as Lin28TT(2). This protein preparation bound preE-let-7g ~2.3 times as strongly as the original Lin28TT sample (**Fig. 4.20d**) with a K_d of 16.9 ± 2.2nM compared to 38.8 ± 3.0nM .

The binding of the let-7gmut RNA by Lin28TT(2) was then investigated (**Fig. 4.20e**). Data were fit by *Equation 14,* resulting in a deduced K_d of 13.1 ± 2.8nM. However, the adjusted R^2 of this fit was relatively low (0.9614) suggesting that the model may not fully describe the data.

The experiments were repeated at 250mM and 500mM NaCl (**Fig. 4.20 f,g**). At 250mM and 500mM NaCl, the curves appeared sharper and had lower Amax values. *Equation 14* was fit to each set of data and K_d values of 0.69 ± 0.24nM and 4.77 ± 0.58nM were deduced respectively.

For the CSD protein, SEC-MALLS revealed that equimolar and 2-fold protein excess mixtures of CSD and let-7gmut had average molecular weights of 61.0kDa and 60.1kDa (**Fig. 4.21**), which also correspond to 1:1 stoichiometries, with an expected molecular weight of 61.2kDa. The value of the dissociation constant was then probed by fluorescence anisotropy. Data were fit by *Equation 14,* resulting in a deduced K_d of 178.7 ± 26.4nM (**Fig. 4.20h**). At higher salt concentrations very little binding activity could be seen, and no further conclusions could be made.

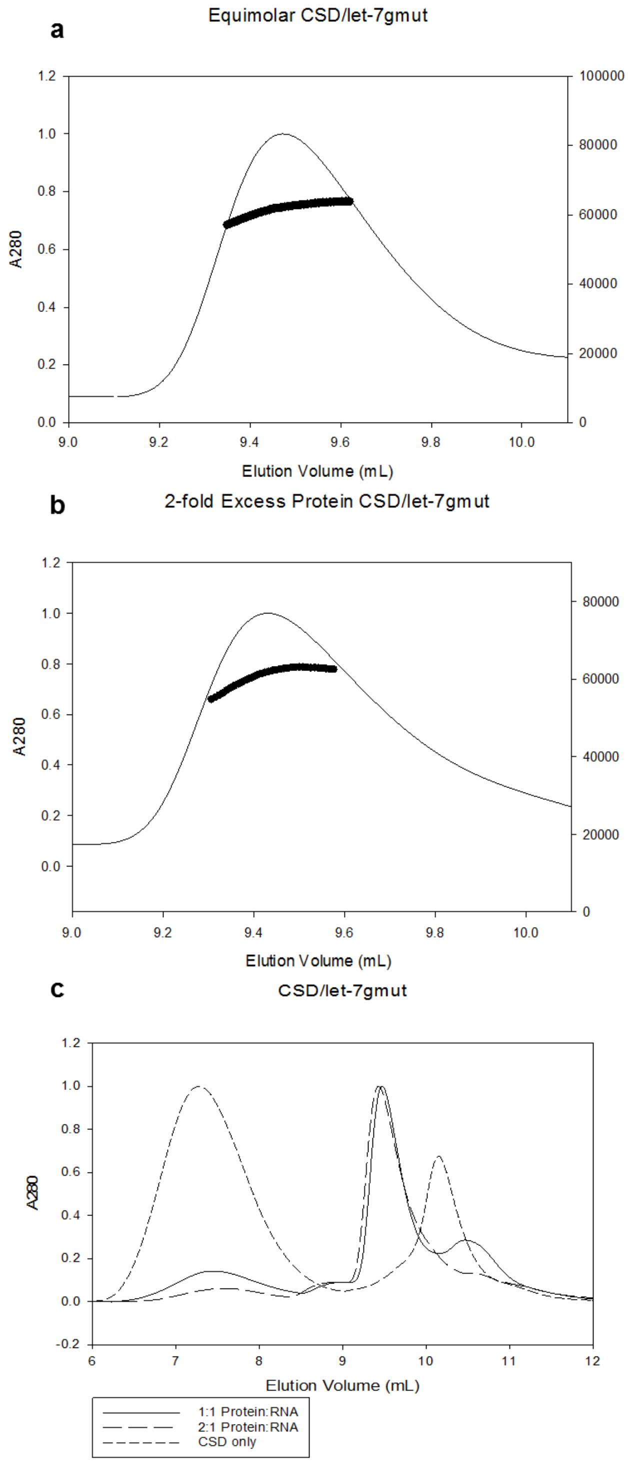


Figure 4.21: SEC-MALLS analysis of CSD complexes with let-7gmut prepared in different molar rations.

(a) Protein-RNA complexes were mixed in 1:1 (a) and 2:1 (b) ratios. (c) Elution profiles (A_{280} absorbance) for each complex and for protein and RNA.

4.3.14 Lin28 binds short let-7gΔ5 DNA oligonucleotides

To examine differences in binding between RNA and DNA sequences, a short DNA oligonucleotide with an equivalent sequence to that of let-7gΔ5 (see *Chapter 2*) which forms the same predicted hairpin secondary structure, was mixed with Lin28TT and binding measured by fluorescence anisotropy. The assay was conducted at two salt concentrations: 100mM and 500mM, with Lin28TT(2). The data obtained at 100mM NaCl (**Fig. 4.22a**) were fit by *Equation 14*. This resulted in a deduced K_d of 54.1 ± 7.5 nM. However, similarly to the let-7gmut results, the curve fit with a relatively low adjusted R^2 of 0.9648, suggesting that the model used may not fully describe the data. At 500mM NaCl (**Fig.4.22b**), the curve appeared sharper with a lower Amax, and data were fit by *Equation 14* to obtain a K_d of 2.1 ± 0.4nM. The binding of this short DNA sequence is very similar to that of the let-7gmut sequence, where the curve becomes sharper and flatter with increased salt concentration.

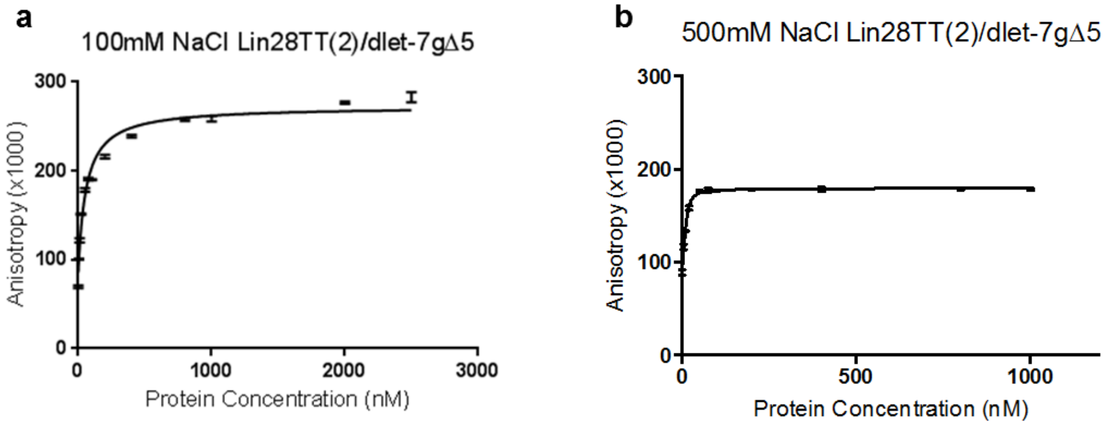


Figure 4.22: Fluorescence anisotropy analysis of Lin28TT(2) interaction with let-7gΔ5 DNA.

The analysis was performed in 100mM NaCl (a*)* and 500mM NaCl (b*)*.

4.4 Discussion

4.4.1 Specific Lin28 complexes interact with high affinity in a 1:1 binding stoichiometry

Multiple stoichiometries have been reported for Lin28/RNA complexes [122, 123, 137, 148]. The results presented within this chapter demonstrate that if both domains of Lin28 are present, as well as the GGAG motif, the terminal loop sequence of let-7g will be bound with nanomolar affinity, in a 1:1 ratio. The observation that Lin28B proteins bind pre-let-7g miRNAs in a 1:1 stoichiometry provides support for this conclusion [183]. Disrupting either of these factors will cause extra Lin28 molecules to associate with the RNA through their CSDs, and cause the complex to become more sensitive to the ionic strength of the buffer.

The mir-363 miRNA was identified as a potential binding partner of Lin28, as it contained a GGAG motif that was conserved throughout vertebrates and present on the 3' strand. Interestingly, it did promoted very little uridylation activity when bound by Lin28 relative to let-7a-1 [92]. In contrast to let-7 miRNAs, it was observed that in *Xenopus tropicalis*, knockdown of Lin28 was correlated with a decrease in mature mir-363 by gene array and was binding of this sequence by recombinant and endogenous *Xtr*Lin28 was observed by gel shift analysis (Warrender et al. unpublished data). The results here show that the mir363 sequence containing the terminal loop of this miRNA is specifically bound by Lin28. Similar to the let-7 sequence tested, mir363 was bound with nanomolar affinity and in a 1:1 stoichiometry. In addition the Lin28TT protein bound with stronger affinity than the CSD on its own, and was resistant to changes in the ionic strength of the buffer, implying that the ZnK domain specifically binds the GGAG motif through a hydrogen bonding network. This, in combination with the previous findings detailed above, suggests that although mir363 is a target of Lin28, the biological outcome of this binding may be different to that of the let-7 miRNAs.

4.4.2 The *in vitro* binding mechanism is dependent on the length of the nucleic acid

SEC-MALLS results revealed that complexes of Lin28 with the short let-7gmut and dlet-7gΔ5 formed at 1:1 stoichiometries, and can be fit with E*quation 14*, yet in low salt concentrations the curves fit relatively poorly. It is likely that this is due to a gradual increase in anisotropy seen at higher protein concentrations in both cases. Upon inspection of the data it can be seen that increasing the salt concentration prevents this increase in anisotropy, so the data instead show a sharp curve. It is possible that this curve corresponds to the interaction of the CSD with the loop of the sequence. If this is the case, then the RNA will not have a sufficiently long 3' end to accommodate a second Lin28/CSD molecule. Similarly, if the GGAG in the DNA sequence is inaccessible, then it will also be too short to fully accommodate another Lin28 molecule. Therefore, the anisotropy increase might correspond to a weak interaction between a second Lin28 molecule with a Lin28/nucleic acid complex. Alternatively, complexes of Lin28 with these shorter oligonucleotides may have a greater tendency to form aggregate species at higher protein concentrations in low ionic strength solutions, which would also explain the gradual increases in anisotropy observed in the low salt experiments.

For the let-7gmut RNA sequence, it is surprising Lin28 interacts more strongly with the shorter mutant sequence than the wild-type sequence. The shorter sequence can only form one of the two alternative structures that can be adopted by the wild-type sequence. Lin28 melts open the more highly structured RNA conformation, enabling it to bind [123]. It is possible, therefore, that the affinity of the CSD for this sequence might be increased as it could readily bind the loop without having to melt it. However, the observation that the CSD alone interacts with the let-7gmut RNA much more weakly than two domain protein suggests that this is unlikely to be the case, and that instead there is some interaction taking place between the ZnK domain and the mutant RNA. Further data will be required to fully explain these results and characterise the interaction of Lin28 with these shorter RNA oligonucleotides.

The DNA sequence used is identical to that used to solve the structure of the *m*Lin28/let-7g complex [122]. Lin28TT should therefore be able to contact the GGAG motif. If this motif is bound by the ZnK domain, then the binding would be expected to be similar to that of the Lin28TT/preE-let-7g interaction. The protein construct used for obtaining the crystal structure was truncated by removal of residues from the interdomain linker. It is therefore possible that the Lin28TT protein used in this study is too long to effectively interact with the GGAG motif of the shorter sequence, causing its binding to resemble that of the let-7gmut sequence. In addition, the sequence is composed of DNA rather than RNA, which may influence the binding in unknown ways.

The binding experiments with both the let-7gmut and dlet-7gΔ5 oligonucleotides were performed using the Lin28TT(2) protein. As this sample was from a different batch as that which was used for previous binding experiments it is, however, possible that the anomalies described above are due to differences in the preparation of each protein sample, rather than a more general feature of Lin28's binding activity towards shorter oligonucleotides. Support for this hypothesis can be found in the observation that the Lin28TT(2) protein bound more strongly to preE-let-7g than the original Lin28TT protein, which suggests differences between the two samples.

4.4.3 The binding of the Lin28 CSD has an electrostatic component

Lin28 interacts with miRNAs through its two domains [122]. Initially, the CSD of Lin28 binds to the single stranded region of the terminal loop of a miRNA. It has been suggested that a melting process could then facilitate the exposure and following interaction of the conserved GGAG motif with the ZnK domain [123, 124]. The structures of *m*Lin28A in complex with let-7 family miRNA sequences [122], and of the *Xtr*Lin28B CSD domain [123], show binding of the CSD to the extruded terminal through stacking interactions formed between the nucleobases of the RNA and aromatic amino acid side chains from a preformed hydrophobic binding platform present on the surface of the CSD. In tandem, the ZnK domain bind the GGAG motif through a hydrogen bonding network, with a stacking interaction formed between the side chain of Y140 and the bases of the final A and G of the motif.

The results highlight the importance of a previously unreported electrostatic component to the interaction. The sensitivity of the CSD to the ionic strength of the buffer implies that residues within this domain are forming electrostatic contacts with the negatively charged phosphate groups of the RNA in addition to the hydrophobic stacking interactions reported previously. The structure of the CSD homologue, CspB from *B.subtilis* [63, 64], contains a preformed hydrophobic binding platform surrounded by positively charged residues, which are thought to be important in attracting the negatively charged nucleic acid binding partners. In the CSD of Lin28A, many of these basic residues are conserved, with twice as many R,H and K residues found in the matching Lin28A sequence by alignment. Therefore, it is clear that there is a strong electrostatic component to the interaction between the CSD of Lin28A and nucleic acids, and explaining why the binding of the CSD only protein is sensitive to salt concentration.

Previous studies of Lin28's interaction with let-7g sequences were conducted in low ionic strength buffers (50-100mM NaCl) [80, 81, 122, 124, 148, 182]. The sensitivity of this interaction to salt concentration goes some way towards explaining the variation in the reported K_d values. Additionally, differences in the ways the proteins used in these studies were prepared could lead to differences in their affinity towards RNA. The MBP fusion proteins used here are highly suitable for the investigation into the interaction between Lin28 and its RNA binding partners, and the fluorescence anisotropy assay employed allows accurate, quantitative affinity measurements to be deduced.

4.4.4 The ZnK domain acts as an anchor in Lin28 binding

The absence of the ZnK/GGAG interaction, through mutation of either protein or RNA, results in higher protein:RNA stoichiometries as multiple CSDs bind the sequence.

CSD binding is non-specific and dominated by electrostatic forces, and takes place at one of two different sites on the RNA. The observation of intermediate stoichiometries suggests the CSD/RNA/CSD complexes exist in equilibrium and so each CSD/RNA complex is transient. When ZnK/GGAG binding occurs, a 1:1 complex forms, which is insensitive to increases in ionic strength. These results suggest that the ZnK domain acts as an anchor to keep the CSD bound in a particular position, whilst preventing the association of further CSDs with sequences of this length (**Fig. 4.23**).

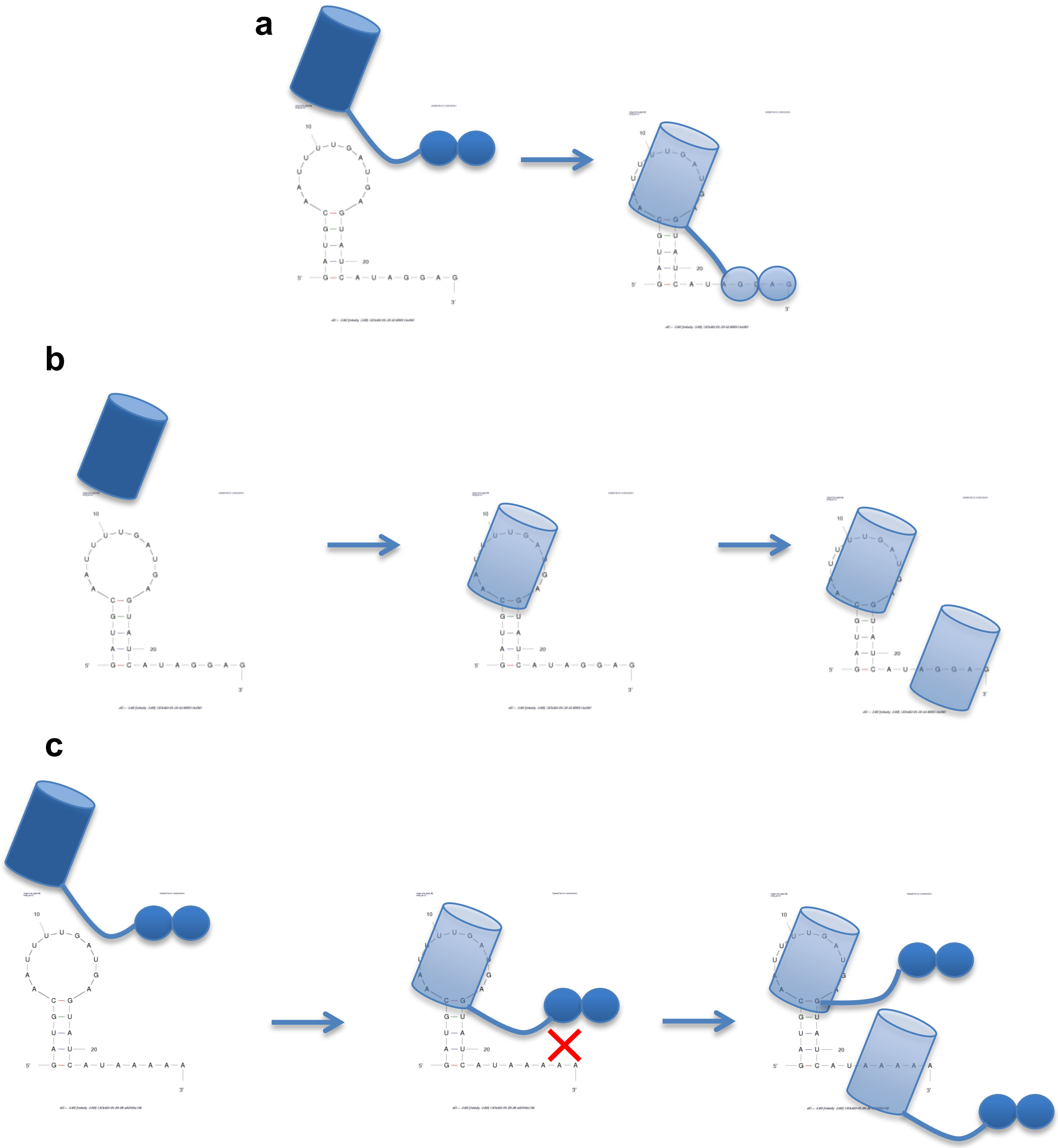


Figure 4.23: Specific and non-specific interactions of Lin28 with RNA.

Lin28 molecules are represented as cylinders, which correspond to CSDs, and circles, which represent zinc knuckles.(a) The Lin28 CSD is attracted to the mir363 RNA. The ZnK domain recognises the GGAG motif and binds, which is followed by binding of the CSD to the loop region of the RNA. (b) The CSD only construct is attracted to the RNA sequence and binds the loop region. The single stranded 3' region containing the GGAG motif is then bound by a second CSD molecule. This complex exists in equilibrium with its components. (c) The CSD of a Lin28 molecule is attracted to the mir363(AAAA) RNA. The ZnK domain cannot bind as no GGAG motif is present. The loop region is then bound by the CSD of one Lin28 molecule while the single stranded 3' region is bound by a second Lin28 molecule, giving rise to a 2:1 complex, which exists in equilibrium with its components.

The observation of sequential binding of Lin28 molecules to the let-7g terminal loop supports this model [182]. The *t*preE-let-7g sequence used here is shorter and contains only two of the three identified binding sites – the internal loop and GNRA tetraloop, and only one GGAG motif. This explains why a maximum stoichiometry of 2:1 is seen. It is possible that in longer sequences, higher protein:RNA stoichiometries could form, the specificity of such interactions would depend on whether extra GGAG binding sites were available. The use of SEC-MALLS in this study provides a quantitative measure of complex stoichiometry, which differentiates the results presented here from previous work [182].

These results suggest that, in the cell, Lin28 can use its CSD to effectively sample the transcriptome through transient electrostatic associations and increase the local concentration of ZnK domains around the RNA, increasing the likelihood of locating the short GGAG motif. The ZnK would bind this motif, and tether the CSD to the RNA, where it would interact with single stranded RNA elements and form a stable complex. Depending on the structure and sequence of the RNA, melting and the association of further Lin28 molecules could then occur. Future work will therefore need to concentrate on Lin28's interaction with the transcriptome in cells, in order to confirm this hypothesis.

Chapter 5 : Structure and Mechanism of Dihyrdouridine Synthase C from *E. coli*

5.1 Introduction

Specificity of the dihydrouridine synthase family of enzymes towards particular uridines in tRNA is important for normal cell functioning. As mentioned in *Chapter 1*, the dihydrouridine modification is commonly found at positions 16, 17, 20 and 20a of the D-loop of tRNA sequences as well as positions 20b and 47 in human and yeast [185]. The conservation of dihydrouridine at these positions suggests its importance in modifying the properties and structure of tRNA. One of the expected effects of the dihydrouridine modification is to locally increase the flexibility of the tRNA structure. This occurs through the prevention stacking interactions, as the base is no longer planar, and by converting the preferred ribose conformation from the C3’ endo to C2’ endo form, which is inherently more flexible [186]. It was postulated that the clustering of dihydrouridine nucleotides around the D loop allows the formation of the conserved G18-Ψ55 and G19-C56 tertiary base pairs.

Therefore, the exact positioning of dihydrouridine within the tRNA structure is important, but tRNA molecules differ in sequence, as well as in length [185] (**Fig. 5.1**). Dus enzymes have non-overlapping functions and are specific towards certain uracil positions in the tRNA [138]. Therefore, Dus enzymes must non-specifically bind all tRNA sequences containing uracil at the target position, and specifically modify a particular position within each molecule.

DusC substrate tRNAs

	D-loop																	T stem loop																
Position	10	11	12	13	14	15	16	17	18	19	20	20A	20B	21	22	23	24	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66
Asn_QUU	G	U	U	C	A	G	D	C	G	G	D	-	-	A	G	A	A	C	U	G	G	T	P	C	G	A	G	U	C	C	A	G	U	C
Asp_QUC	G	U	U	C	A	G	D	C	G	G	D	D	-	A	G	A	A	C	G	G	G	T	P	C	G	A	G	U	C	C	C	G	P	C
Gly_GCC	G	C	U	C	A	G	D	D	G	G	D	-	-	A	G	A	G	C	G	A	G	T	P	C	G	A	G	U	C	U	C	G	U	U
His_QUG	G	C	U	C	A	G	D	D	G	G	D	-	-	A	G	A	G	U	G	G	G	T	P	C	G	A	A	U	C	C	C	A	U	U
Leu_BAA	G	C	G	A	A	A	D	C	#	G	D	A	-	G	A	C	A	C	C	G	G	T	P	C	G	A	G	U	C	C	G	G	C	C
Leu_CAG	G	C	G	G	A	A	D	D	#	G	D	A	-	G	A	C	G	G	G	G	G	T	P	C	A	A	G	U	C	C	C	C	C	C
Leu_GAG	G	U	G	G	A	A	D	D	#	G	D	A	-	G	A	C	A	C	G	G	G	T	P	C	A	A	G	U	C	C	C	G	U	C
Leu_HAA	G	U	G	G	A	A	D	C	#	G	D	A	-	G	A	C	A	C	G	G	G	T	P	C	A	A	G	U	C	C	C	G	C	U
Lys_SUU	G	C	U	C	A	G	D	D	G	G	D	-	-	A	G	A	G	C	A	G	G	T	P	C	G	A	A	U	C	C	U	G	C	A
Met_MAU	G	C	U	C	A	G	D	D	#	G	D	D	-	A	G	A	G	C	A	G	G	T	P	C	G	A	A	U	C	C	C	G	U	C
Phe_GAA	G	C	U	C	A	G	D	C	G	G	D	-	-	A	G	A	G	U	U	G	G	T	P	C	G	A	U	U	C	C	G	A	G	U
Thr_GGU	G	C	U	C	A	G	D	D	G	G	D	-	-	A	G	A	G	G	C	A	G	T	P	C	G	A	A	U	C	U	G	C	C	U
Thr_GGU	G	C	U	C	A	G	D	D	G	G	D	-	-	A	G	A	G	C	C	A	G	T	P	C	G	A	C	U	C	U	G	G	G	U
Trp_CCA	G	U	U	C	A	A	D	D	G	G	D	-	-	A	G	A	G	G	G	A	G	T	P	C	G	A	G	U	C	U	C	U	C	C
Val_GAC	G	C	U	C	A	G	D	D	G	G	D	D	-	A	G	A	G	G	U	G	G	T	P	C	G	A	G	U	C	C	A	C	U	C
Val_GAC	G	C	U	C	A	G	D	D	G	G	D	D	-	A	G	A	G	U	U	G	G	T	P	C	G	A	G	U	C	C	A	A	U	U
A	0	0	0	1	16	5	0	0	0	0	0	4	0	12	4	12	5	0	2	4	0	0	0	0	3	16	4	0	0	0	3	3	0	1
C	0	11	0	12	0	0	0	5	0	0	0	0	0	0	0	4	0	9	3	0	0	0	0	16	0	0	1	0	16	12	8	3	6	8
G	16	0	4	3	0	11	0	0	11	16	0	0	0	4	12	0	11	4	7	12	16	0	0	0	13	0	10	0	0	0	4	9	2	0
U	0	5	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4	0	0	0	0	0	0	0	1	16	0	4	1	1	7	7
Modified	0	0	0	0	0	0	16	11	5	0	16	4	0	0	0	0	0	0	0	0	0	16	16	0	0	0	0	0	0	0	0	0	1	0
Consensus				C	A	G	D	D			D			A		A	G						P	C										

DusC non-substrate tRNAs

	D-loop																	T stem loop																
Position	10	11	12	13	14	15	16	17	18	19	20	20A	20B	21	22	23	24	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66
Ala_GGC	G	C	U	C	A	G	C	D	G	G	G	-	-	A	G	A	G	G	C	G	G	T	P	C	G	A	U	C	C	C	G	C	U	U
Ala_VGC	G	C	U	C	A	G	C	D	G	G	G	-	-	A	G	A	G	G	C	G	G	T	P	C	G	A	U	C	C	C	G	C	G	C
Ala_VGC	G	C	U	C	A	G	C	D	G	G	G	-	-	A	G	A	G	G	C	G	G	T	P	C	G	A	U	C	C	C	G	C	A	U
Arg_CCG	G	C	U	C	A	G	C	D	G	G	A	D	-	A	G	A	G	C	A	G	G	T	P	C	G	A	A	U	C	C	U	G	U	C
Arg_ICG	G	C	U	C	A	G	C	D	G	G	D	-	-	A	G	A	G	G	A	G	G	T	P	C	G	A	A	U	C	C	U	C	C	C
Arg_ICG	G	C	U	C	A	G	C	D	G	G	A	D	-	A	G	A	G	G	A	G	G	T	P	C	G	A	A	U	C	C	U	C	C	C
Arg_{CU	G	U	U	A	A	A	U	-	G	G	A	D	-	A	U	A	A	C	A	G	G	T	P	C	G	A	U	U	C	C	U	G	C	A
Arg_{CU	G	C	U	C	A	G	U	U	G	G	A	U	-	A	G	A	G	C	A	G	G	T	P	C	G	A	A	U	C	C	U	G	C	A
Cys_GCA	A	C	A	A	A	G	C	-	G	G	D	-	-	D	A	U	G	C	C	G	G	T	P	C	G	A	C	U	C	C	G	G	A	A
Gln_CUG	G	C	C	A	A	G	C	-	#	G	D	-	-	A	A	G	G	G	A	G	G	T	P	C	G	A	A	U	C	C	U	C	G	U
Gln_NUG	G	C	C	A	A	G	C	-	#	G	D	-	-	A	A	G	G	C	U	G	G	T	P	C	G	A	A	U	C	C	A	G	G	U
Glu_SUC	G	U	C	P	A	G	A	-	G	G	C	C	C	A	G	G	A	G	G	G	G	T	P	C	G	A	A	U	C	C	C	C	U	G
Glu_SUC	G	U	C	P	A	G	A	-	G	G	C	C	C	A	G	G	A	G	G	G	G	T	P	C	G	A	A	U	C	C	C	C	U	A
Glu_SUC	G	U	C	P	A	G	A	-	G	G	C	C	-	A	G	G	A	G	G	G	G	T	P	C	G	A	A	U	C	C	C	C	U	A
Gly_CCC	G	U	U	C	A	A	U	-	G	G	D	-	-	A	G	A	A	A	G	G	G	T	P	C	G	A	U	U	C	C	C	U	U	C
Gly_NCC	G	U	A	U	A	A	U	-	G	G	C	U	-	A	U	U	A	C	G	G	G	T	P	C	G	A	U	U	C	C	C	G	C	U
Ile_GAU	G	C	U	C	A	G	G	D	G	G	D	D	-	A	G	A	G	G	U	G	G	T	P	C	A	A	G	U	C	C	A	C	P	C
Ile_GAU	G	C	U	C	A	G	G	U	G	G	D	D	-	A	G	A	G	G	U	G	G	T	P	C	A	A	G	U	C	C	A	C	P	C
Ile_}AU	G	C	U	C	A	G	U	-	#	G	D	D	-	A	G	A	G	C	U	G	G	T	P	C	A	A	G	U	C	C	A	G	C	A
Sec_UCA	C	G	U	C	U	C	C	-	G	G	D	G	-	A	G	G	C	C	A	G	G	T	P	C	G	A	C	U	C	C	U	G	U	G
Ser_CGA	C	C	G	G	A	G	C	-	#	G	C	D	G	A	A	C	G	G	G	G	G	T	P	C	A	A	A	U	C	C	C	C	C	U
Ser_GCU	G	C	C	G	A	G	A	-	G	G	C	D	G	A	A	G	G	G	G	G	G	T	P	C	G	A	A	U	C	C	C	C	G	C
Ser_GGA	U	C	C	G	A	G	U	-	#	G	C	D	G	A	A	G	G	G	G	G	G	T	P	C	G	A	A	U	C	C	C	C	C	C
Ser_GGA	U	C	C	G	A	G	U	-	#	G	D	D	G	A	A	G	G	G	G	G	G	T	P	C	G	A	A	U	C	C	C	C	C	C
Ser_VGA	G	C	C	G	A	G	C	-	#	G	D	D	G	A	A	G	G	A	G	A	G	T	P	C	G	A	A	U	C	U	C	U	G	C
Tyr_QUA	C	C	C	G	A	G	C	-	#	G	C	C	A	A	A	G	G	A	A	G	G	T	P	C	G	A	A	U	C	C	U	U	C	C
Tyr_QUA	C	C	C	G	A	G	C	-	#	G	C	C	A	A	A	G	G	A	A	G	G	T	P	C	G	A	A	U	C	C	U	U	C	C
Val_VAC	G	C	U	C	A	G	C	D	G	G	G	-	-	A	G	A	G	G	C	G	G	T	P	C	G	A	U	C	C	C	G	U	C	A
A	1	0	2	4	27	3	4	0	0	0	4	0	2	27	10	13	6	4	9	1	0	0	0	0	4	28	16	0	0	0	4	0	2	7
C	4	21	11	13	0	1	15	0	0	0	9	5	2	0	0	1	1	8	5	0	0	0	0	28	0	0	2	4	28	27	10	15	12	13
G	21	1	1	7	0	24	2	0	19	28	4	1	5	0	16	12	21	16	10	27	28	0	0	0	24	0	3	0	0	0	5	8	5	2
U	2	6	14	1	1	0	7	2	0	0	0	2	0	0	2	2	0	0	4	0	0	0	0	0	0	0	7	24	0	1	9	5	7	6
Modified	0	0	0	3	0	0	0	8	9	0	11	11	0	1	0	0	0	0	0	0	0	28	28	0	0	0	0	0	0	0	0	0	2	0
Consensus				C	A	G	C	-			D			A		A	G						P	C										

Figure 5.1: Sequence alignment of EcDusC tRNA substrates and non-substrates.

Mature *E. coli* tRNA sequences were separated into two groups by the presence or absence of D at position 16 and aligned. The consensus row shows the most common nucleoside at each position. Figure courtesy of Dr.Rob Byrne (Antson group).

Although the structure of the *Thermus thermophilus* Dus is available in tRNA bound and unbound states [187], it was not clear which factors define the enzymes specificity towards particular uridines of tRNA,, as protein:RNA complex structures with alternate specificities had not yet been determined. In addition, the position specificities of the *E. coli* enzymes were unknown, except that DusA is able to modify position 20 [138]. It was difficult to draw solid conclusions from the structure of the *Tt*Dus:tRNA^{phe} complex [187], due to the covalent bond formed between the active site Cys residue and D20 of the tRNA, questioning whether the *Tt*Dus was active and whether the tRNA molecule was trapped in a natural way.

5.1.1 Factors defining position specificity in ArcTGT

In archaea, position G15 of the D loop in many tRNA sequences is modified to 7-formamidino-7-deazaguanosine, known as Archaeosine, through a 7-cyano-7-deazaguanine intermediate, called preQ$_0$. G15 is modified to preQ$_0$ through the ArcTGT (Archaeosine tRNA-transglycosylase) enzyme, which cleaves the N-glycosidic bond of G15, then reforms the bond with free preQ$_0$ base [188]. Structurally, this reaction is challenging, as G15 is buried within the core of the tRNA and is involved in a tertiary base pair with C48 of the variable loop. ArcTGT is highly specific for this base, but does not recognise any other part of the tRNA molecule [189].

The 3.3Å structure of *P. horikoshii* ArcTGT in complex with tRNA^{val} (**Fig. 5.2a**) demonstrates how this reaction takes place, and which factors define specificity towards this single position [190]. ArcTGT forms a homodimer [191]. The dimer binds two tRNA^{val} molecules, with each tRNA bound by both subunits [190]. The structure of the ArcTGT did not change considerably upon binding tRNA, whereas the tRNA underwent a significant change in conformation, adopting the λ conformation instead of the classic L-shape. In the λ conformation, nucleotides U8-22 of the D arm form an extended structure, while the remainder of the D-stem forms a helix with the variable loop, called the DV helix (**Fig. 5.2b**) [190].

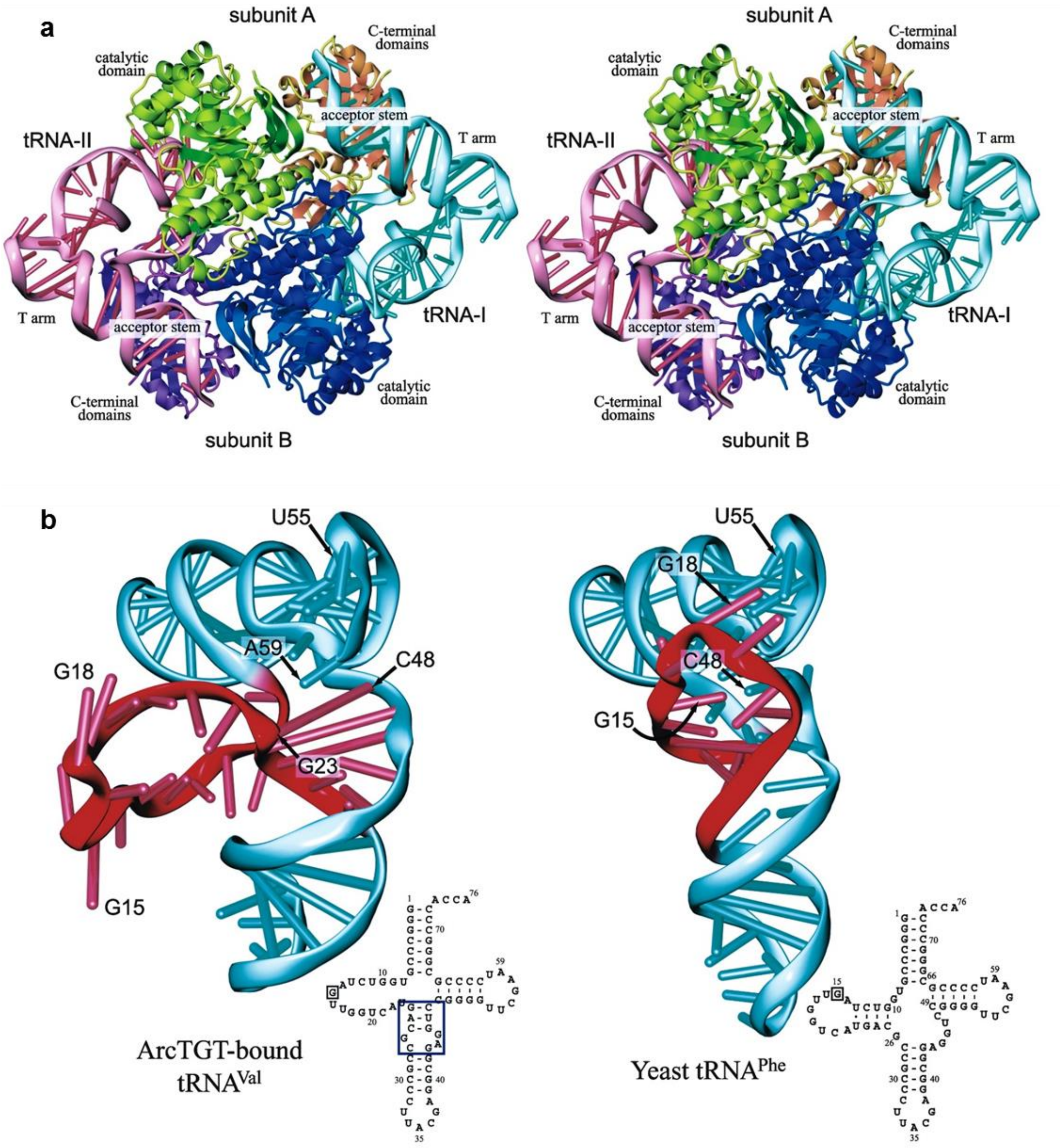


Figure 5.2: ArcTGT binds the λ conformation of tRNA.

(a) Stereo view of the crystal Structure of ArcTGT with tRNA^{val}. Two tRNA molecules are bound, with each contacted by both subunits of the ArcTGT dimer. Relevant regions are labelled. (b) Comparison of the λ form of tRNA^{val} with the L-form of yeast tRNA^{phe}. Nucleotides of the D arm and the DV helix are coloured in red. Secondary structures are shown alongside 3D models with the DV helix surrounded by a box. Adapted from Ishitani et al. (2003).

Subunit A interacts with the acceptor stem through electrostatic contacts formed between the phosphates of the tRNA and basic residues found in the C3 domain of ArcTGT, known as the PUA (**P**seudo**U**ridine synthase and **A**rchaeosine TGT) domain

[190], which is widely found in tRNA modification enzymes [192]. These interactions allow the PUA domain to bind to the 5’ end of the tRNA acceptor stem, while the 3’ end is recognised by electrostatic interactions between tRNA phosphates and basic residues in a β sheet of the C2 domain. This allows ArcTGT to accurately position itself onto the tRNA [190].

Nucleotides 8-13 of the extruded loop of the λ form tRNA are bound in the cleft between the two ArcTGT subunits through interactions between the tRNA phosphates and polar residues from both subunits. Subunit B forms hydrogen bonds to the 2-amino groups of G9 and G10 but does not interact with the bases of other nucleotides in this region. A14 and U16 are then bound in hydrophobic pockets of the catalytic domain, and the substrate G15 buried deep within [190], held in place by the backbones of S98 and F99, as well as the S98 sidechain [191].

The PUA and C2 domains of subunit A bind the acceptor stem of the tRNA through non base specific contacts, which positions the catalytic domain of subunit B so that it binds to the extruded region of the λ form tRNA. Non-specific hydrogen bonds are formed to the backbone atoms of this region, allowing the enzyme to count the number of nucleotides up to U13, where the next three nucleotides are buried in hydrophobic pockets and G15 is inserted into the active site of the enzyme, where it is modified. Position specificity of the ArcTGT enzyme is therefore maintained through a base counting mechanism, facilitated by a rearrangement of the tRNA structure, and the PUA domain of the protein.

5.1.2 Factors determining position specificity in Pseudouridine Synthases

Pseudouridine (Ψ) is a common nucleotide modification of uracil that stabilises RNA sequences by introducing a new hydrogen bond donor group in place of C5, as well as enhancing stacking interactions in single and double stranded RNAs [193]. The modification occurs through the isomerisation of uracil nucleotides present in RNA sequences and is catalysed by five families of enzymes: RluA, RsuA, TruA, TruB and TruD [194]. These enzymes have a common core consisting of an eight stranded mixed β sheet intersected by a catalytic cleft adjacent to several helices and loops. One of these loops contains the universally conserved aspartate residue responsible for catalysis, and it is likely that all pseudouridine synthases have the same catalytic mechanism [194].

The range of substrate nucleotide positions modified by pseudouridine synthases is diverse, and includes both tRNA and rRNA sequences [194]. As the pseudouridine synthases all have a common core, substrate specificity must be generated by other means. As many tRNAs contain Ψ at several positions, the structural problem faced by pseudouridine synthases is similar to that faced by the Dus enzymes, where a specific position must be recognised in a non-specific context.

5.1.2.1 TruB

TruB is the pseudouridine synthase that modifies position U55 to Ψ in tRNA; a modification found in the majority of tRNA molecules. Recognition of this base is challenging, as it is present within the folded core of the tRNA due to a long range base pair between U55 and a nucleotide of the D-loop, so the tRNA must be opened to allow modification of this residue [195].

The structure of TruB in complex with a synthetic RNA corresponding to the T stem loop (TSL), revealed the mechanism of substrate access and recognition. A base flipping mechanism is used to access U55 as the bases of nucleotides 55, 56 and 57 are extruded from the helical stack they usually form in tRNA (**Fig. 5.3a**) by the conserved H43 and its neighbouring amino acids. The “thumb” domain of the protein closes in on the TSL during binding and forces H43 to stack underneath the universal A58:U54 reverse Hoogsteen basepair in a conformation stabilised by a hydrogen bonding network (**Fig. 5.3b**). This insertion prevents the stacking of U55 under this pair (**Fig. 5.3 c,d**), and so the base flips out of the TSL helix into the active site cavity, which is inaccessible to solvent, thus sequestering the substrate in close proximity to the active site aspartate. Modelling U55 into the structure, which uses 5-fluorouracil in place of U55, showed that an active site tyrosine forms a stacking interaction with U55. This prevents the intermediate detached nucleobase from rotating in the active site, ensuring efficient catalysis [195]. TruB catalyses Ψ modification equally in both tRNA^{phe} and T-arm segments [196], and superposition of tRNA^{phe} with the TSL in the crystal structure revealed few steric clashes or contacts [195], and so it is likely that the recognition mechanism observed in this structure is also true for full tRNA molecules.

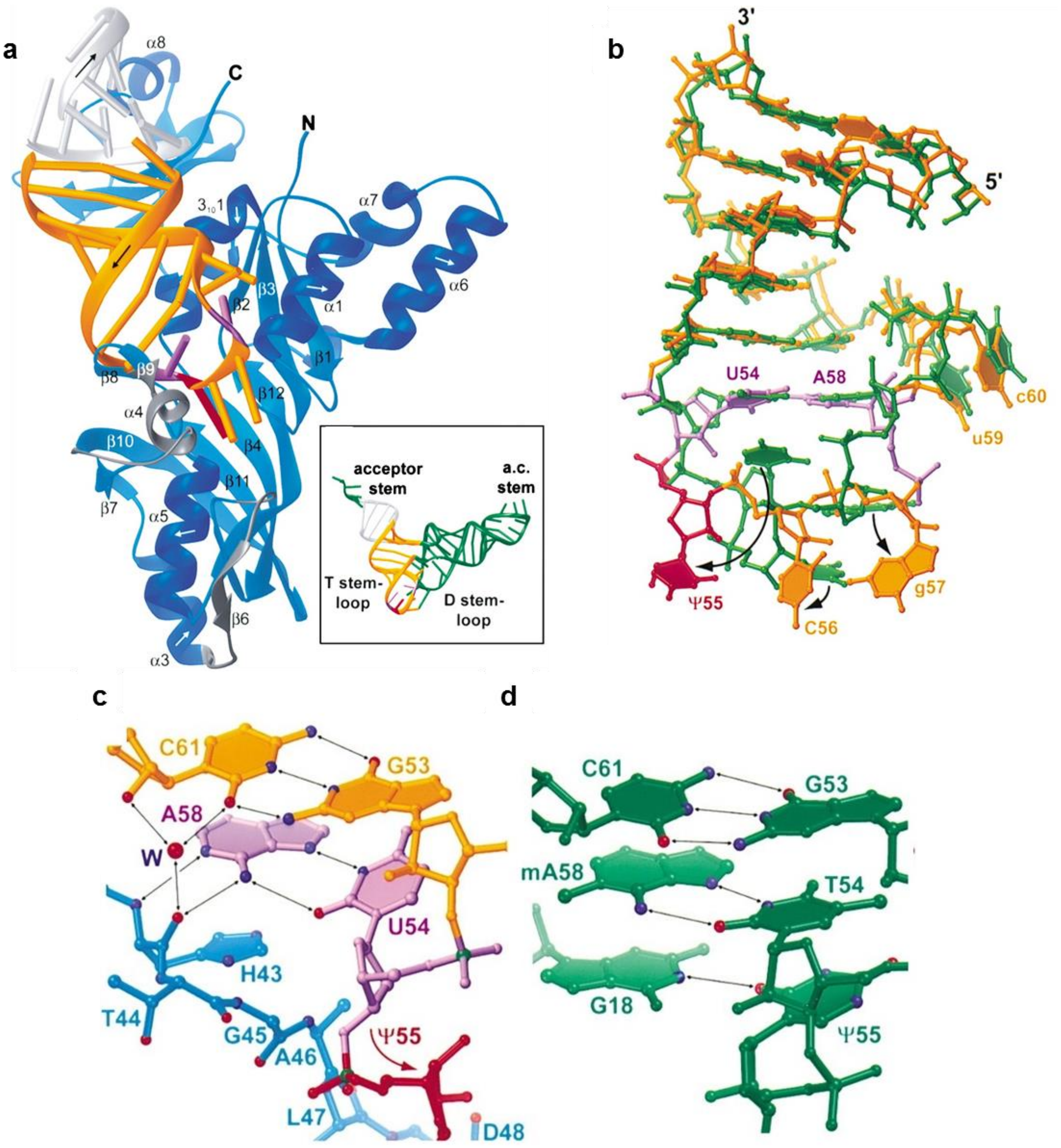


Figure 5.3: Structure of TruB in complex with TSL segment.

(a) Ribbon diagram of the complex. Protein is shown in blue, with grey segments highlighting characteristic TruB family elements. The TSL is coloured in orange, apart from the U54:A58 reverse Hoogsteen basepair (magenta) and the substrate nucleotide 55 (red). The position of the TSL sequence within yeast tRNA^{phe} is shown alongside the structure. (b) Superposition of the TruB bound TSL with the corresponding RNA segment from tRNAphe (green), aligned by C1' atoms excluding nucleotides 55, 56 and 57, which are flipped out upon TruB binding. (c) H43 (blue) displaces nucleotide 55 from its original position under the reverse Hoogsteen basepair. (d) In tRNAphe, G18 makes contact with nucleotide 55. (c) and (d) show the differences in the TSL structure upon TruB binding. Adapted from Hoang and Ferré-D'Amaré (2001).

5.1.2.2 TruD

The recognition mechanism used by TruD has not been established [194]. TruD modifies position U13 of the tRNA D loop, which is present in the core structure of the tRNA, and so is inaccessible without rearrangement of the tRNA. The tRNA-free structure of TruD reveals two domains: a catalytic domain, and a αβ insertion domain which is not homologous to any known fold. The catalytic domain displays homology to the catalytic segments of two other members of the pseudouridine synthase family of enzymes, TruA and TruB. The conserved catalytic aspartate lies in the active site cleft of TruD, which is formed between the two domains. An active site phenylalanine, which is conserved in the TruD subfamily enzymes, corresponds to the conserved active site tyrosine residue found in TruB, suggesting conservation of the base flipping mechanism of nucleotide recognition. However, docking of tRNA into the active site cleft between the TruD domains implies that it is more likely that TruD binds the λ form of tRNA [197], as seen in the ArcTGT structure [190]. In this form, the extended D-loop would lie along the active site cleft and the insertion domain would stabilise the conformation. It is possible that L-form tRNA could be accommodated by the active site cleft of the enzyme, as its structure is flexible, but U13 would still be inaccessible in this case [197].

5.1.2.3 TruA

TruA catalyses the modification of U38-40 located in the anticodon stem loop (ASL). It differs from TruB and TruD as it forms a homodimer. tRNA^{leu} was bound by both subunits, with the ASL positioned in the active site cleft that forms at the dimerization interface between the N terminal domain of one subunit, and the C terminal domain of the second subunit. The substrate nucleotide is therefore positioned close to the active site aspartate, which resides in the active site cleft. The conformations of both protein and tRNA in the complex and in their standalone states do not differ significantly. The tRNA is bound to the protein through hydrophobic interactions with the bases, and hydrogen bonding contacts with the sugar-phosphate backbone of the elbow connecting the D and T stem loops. Molecular dynamics simulations, in combination with the structural data, then revealed that the substrate nucleotide is flipped into the active site with the assistance of an active site arginine residue [198].

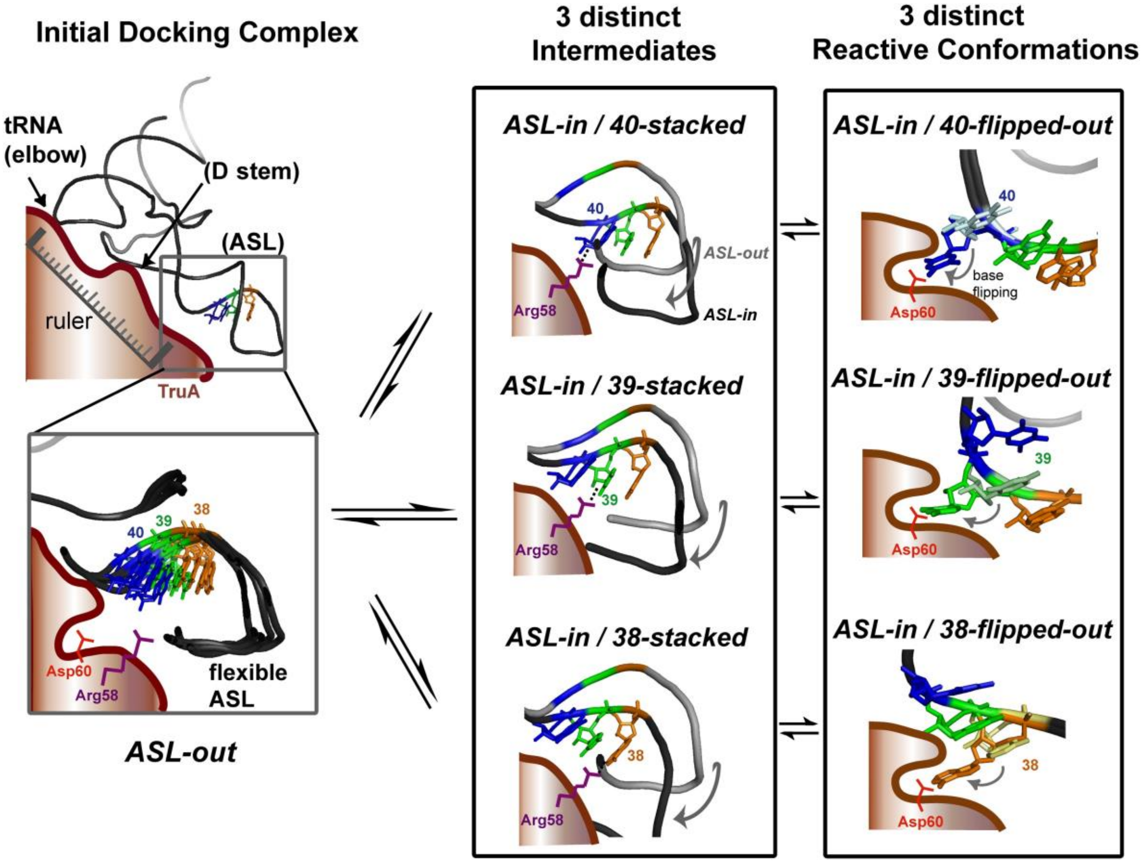


Figure 5.4: Model of nucleotide recognition by TruA.

TruA initially binds the elbow of the tRNA in a non-sequence specific manner, which maintains the flexibility of the ASL but positions it close to the active site. Arg58 then helps to bend the ASL through an intermediate stage, before the flipped out conformation is formed, resulting in nucleotide modification. Each position, 38-40, can be sampled through the intermediate and, as the ASl is flexible, any U in these positions can be modified by TruA. Adapted from Hur and Stroud, (2007).

Any tRNA, regardless of sequence, can therefore be recognised by the TruA dimer, as the overall positioning of the tRNA in the complex is determined by non-sequence specific contacts to a conserved fold adopted by all tRNAs. The ASL is thus positioned in the active site cleft in such a manner that it remains flexible in the enzyme/tRNA complex. It can then adopt one of two conformations, with bases 38-40 either stacked or flipped out, and these bases will be sampled by TruA during one or more binding events. Due to the size of the active site, any base can be accommodated, but only uracil bases have the correct chemistry to be modified by the enzyme. In such a way, three different positions can be recognised and specifically modified by TruA in all tRNA sequences (**Fig. 5.4**).

5.1.3 *E. coli* DusC is specific for position U16

Recent work by Dr.Rob Byrne and Dr.Fiona Whelan (Antson group) identified which particular uridine of tRNA^{phe} is modified by *E. coli* DusC. A reverse transcriptase assay, which extends a primer across unmodified tRNA^{phe} until terminated by the presence of dihyrouridine, showed the enhancement of termination at position 17 in tRNA that had been incubated with wild-type DusC, but not with mutant DusC^{C98A}, where the active site cysteine was mutated to alanine. This indicated that DusC was modifying position U16, and so terminating the chain at position 17.

This result was confirmed by analytical size exclusion chromatography (SEC) and gel shift assays where DusC^{C98A} was seen to form a stable complex with tRNA^{phe}, but not with tRNA^{cys}, which does not contain D at position 16 in any of its mature sequences. tRNA^{trp}, which contains D at positions 16 and 17 in its mature sequence, also formed a complex with DusC^{C98A}, but was not seen to bind by gel shift analysis. These results indicate that DusC specifically modifies position 16, and possibly 17, of the D loop of tRNA.

5.1.4 DusC uses an alternate mechanism of substrate selectivity

However, it was still not known how DusC recognised tRNA, and how specificity towards particular uridine positions within tRNA was generated. The mechanisms used by ArcTGT and the pseudouridine synthases to modify certain nucleotide positions suggested specificity towards particular nucleotides is generated through either rearrangement of the tRNA structure, or a nucleotide flipping mechanism. In each case, recognition involves an auxiliary domain, e.g. the PUA domain of ArcTGT, the thumb domain of TruB, or the insertion domain of TruD. In some cases, like TruA, recognition is assisted through dimerization. The structures of DusC [147] and *Tt*Dus [187] showed no such structural features, and instead a mechanism involving an unknown co-factor was proposed [147]. Therefore, structural data on the complex of DusC with tRNA was needed in order to fully elucidate the mechanism by which position specificity is generated in Dus family enzymes.

5.2 Materials and Methods

5.2.1 Plasmids and RNA

Plasmids encoding *Ec*DusC C98A mutant protein and *E. coli* tRNA^{phe} were produced by Dr.Fiona Whelan (Antson group) and Dr.Andrey Konevega (Max Planck Institute for Physical Biochemistry, Göttingen) respectively.

5.2.2 Protein Purification

DusC^{C98A} protein was produced by James Stowell (Antson group) (**Fig. 5.5**). Briefly, the protein was overexpressed in Rosetta2 *E. coli* cells at 37°C after induction with 1mM IPTG at 0.6 OD_{600}. Cells were resuspended in 20mM TRIS pH 7.5, 500mM NaCl, 20mM Imidazole, 2mM DTT buffer supplemented with an EDTA-free protease cocktail, and lysed by sonication. The insoluble fraction was removed by centrifugation and the soluble fraction purified by Ni^{2+} IMAC. The elution fractions were concentrated and buffer exchanged into 20mM TRIS pH 7.5, 50mM NaCl, 5mM DTT and purified by anion exchange chromatography. The His tag was then removed by digestion overnight with thrombin, and the cleaved DusC^{C98A} protein purified by a second Nickel affinity column. The protein was then further purified by Size-Exclusion chromatography to be stored at -80°C in 20mM TRIS pH 7.5, 100mM NaCl, 5mM DTT.

tRNAphe was purified by Dr.Andrey Konevega (Max Planck Institute for Physical Biochemistry, Göttingen) .

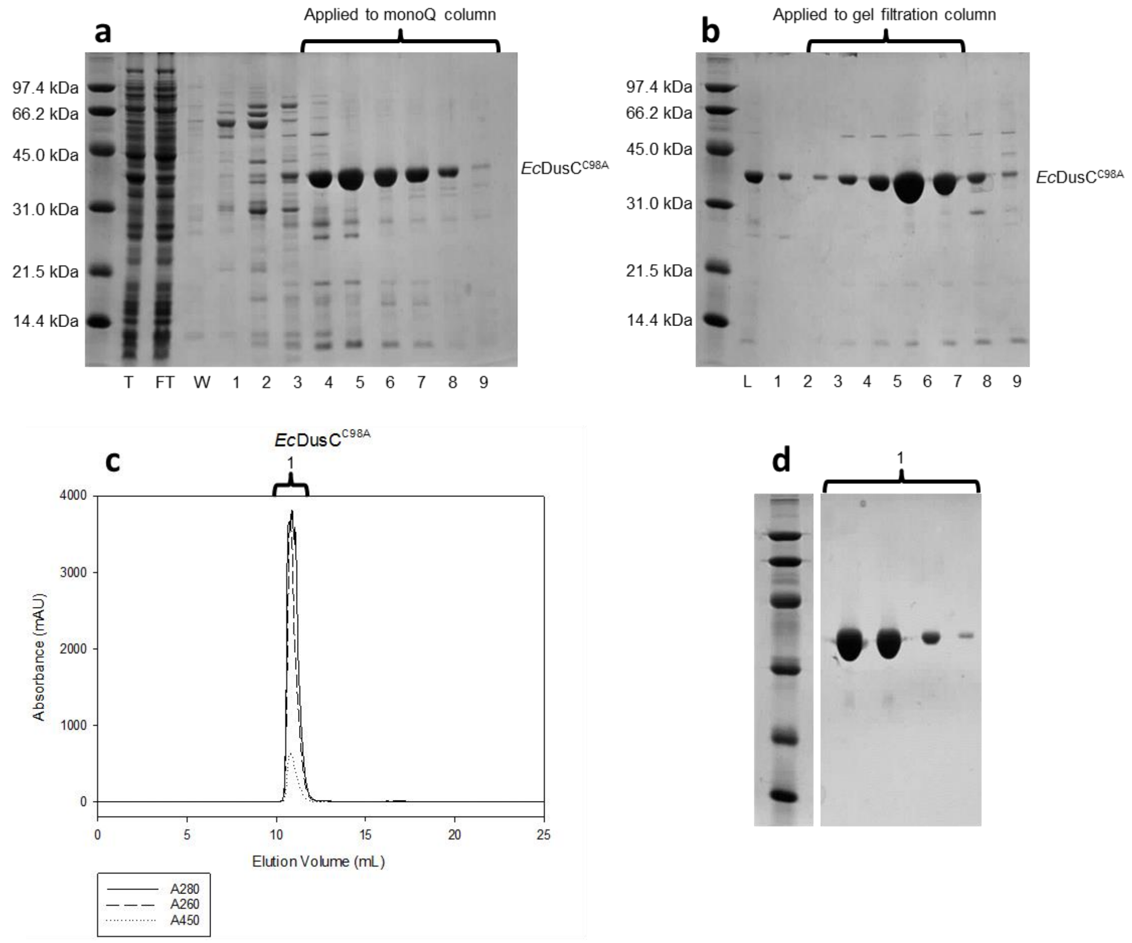


Figure 5.5: Purification of *Ec*DusC^{C98A}

(a) SDS-PAGE analysis of the fractions eluting from the Ni^{2+} IMAC column. T = Total soluble protein, FT = column flow-through (unbound protein), W = Wash fraction. Fractions from lanes 7-12 were pooled and applied to the monoQ anion exchange column shown in (b). (b) SDS-PAGE analysis of the fractions eluting from the anion exchange column. L = Sample loaded onto column. Fraction from lanes 6-11 were pooled and applied to the gel filtration column shown in (c). (c) Elution profile from gel filtration column. Elution fractions were analysed by SDS-PAGE, shown in (d). Experiments performed by James Stowell (Antson group).

5.2.3 Crystallization

The DusC^{C98A}-tRNA^{phe} complex was formed by mixing DusC^{C98A} and tRNA^{phe} in a 1:1 molar ratio, following buffer exchange of the protein into a solution containing 100 mM $MgCl_2$ and 10 mM HEPES-NaOH pH 7.0 using a 0.5 mL Vivaspin 10 kDa MWCO concentrator. The resulting complex was at a final protein concentration of 7.5 mg mL^{-1} (approximately 200 µM complex). Crystals were produced by hanging drop vapour diffusion at 20°C and grew over the course of three days. 1 µL of the complex was mixed with 1 µL of the reservoir solution containing 100 mM HEPES pH 7.0, 200 mM $MgCl_2$, 10 mM $MnCl_2$ and 11% w/v PEG 6K and equilibrated over 1 mL of reservoir. Before flash-cooling in liquid nitrogen, crystals were transferred into a solution containing 100 mM HEPES pH 7.0, 150 mM $MgCl_2$, 10 mM $MnCl_2$, 18% w/v PEG 6K and 15% v/v glycerol.

5.2.4 Structure Determination

X-ray data were collected at Diamond Light Source (Didcot, United Kingdom). Two data sets were collected from a single crystal of the DusC^{C98A}-tRNA^{phe} complex at two different wavelengths: 0.9795 Å data for structure refinement and longer-wavelength data (1.300 Å) for identification of Mn^{2+} using anomalous differences. After integration with *XDS [199]* , the data were imported into the CCP4 suite [200] for subsequent tasks. The data were analysed with *Pointless*, scaled with *Aimless [201]* and the resulting intensities were converted to amplitudes with *cTruncate* (**Table 1**).

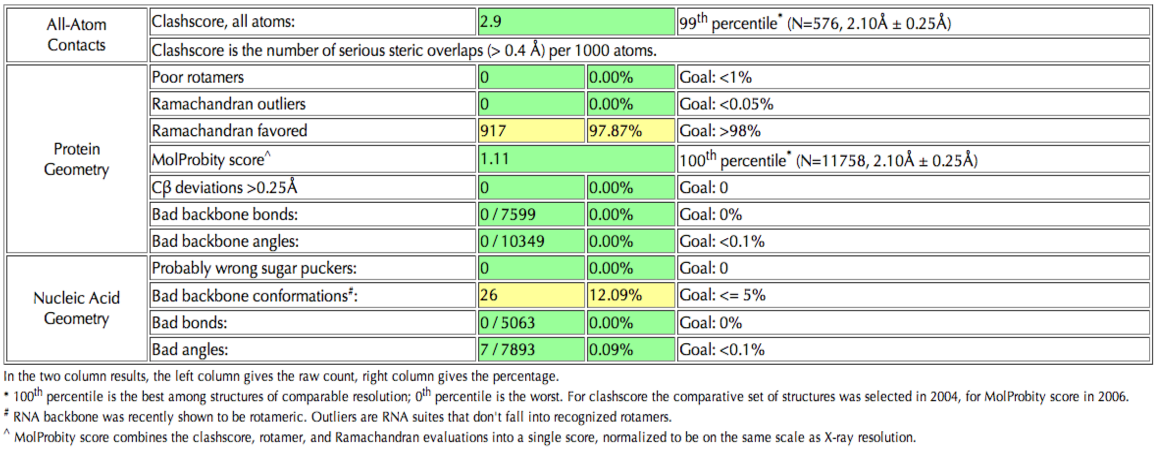
The structure of the complex was determined by molecular replacement with *Phaser* using the structures of DusC (PDB code 4BFA) and unmodified tRNA^{phe} (PDB code 3L0U). Three molecules each of DusC^{C98A} and tRNA^{phe} were found in the asymmetric unit and these were rebuilt and refined using isotropic atomic *B* factors and 21 TLS groups identified by *TLSMD* [202]. Toward the end of model building and refinement, water molecules were added using *ARP/WARP*. [203] Mg^{2+} ions were identified as peaks above 6 σ in the mF_o-DF_c difference electron density maps where coordinating distances were in agreement with those expected for Mg^{2+} ions. Similarly, FMN molecules were built into positive density in the mF_o-DF_c electron density maps that corresponded to the FMN cofactor in the structure of DusC. To identify Mn^{2+} ions within the structure, the data collected at a wavelength of 1.300 Å were analysed by *ANODE* [204] and inspection of the resulting anomalous difference map revealed the locations of three Mn^{2+} ions, one per molecule of tRNA, that were subsequently included in the refinement. The final model contains three DusC^{C98A}-tRNA^{phe} complexes, three molecules of FMN, 24 Mg^{2+} ions, 3 Mn^{2+} ions and 608 water molecules. The DusC and tRNA^{phe} models are complete except for amino acids 100-105 (chains A-C), 313-315 (chain B) and 315 (chain C) and nucleotides 75-76 (chain D), 1-2, 70-76 (chain E) and 72-76 (chain F).

Molecular interfaces were analyzed with *PISA*. [205] The geometry of all models was assessed by *MolProbity [206]*. RNA backbone geometry was improved using RCrane [207]. Figures were created with *ccp4mg* [208]. Additional superpositions were performed in *ProSMART [209]* by Dr.Rob Byrne (Antson group). Multiple sequence alignments were created with *Clustal Omega* [210] and *ESPript* [211].

Table Error! No text of specified style in document.**1: Crystallography statistics for DusC^{C98A}/tRNA^{phe} structure**

Data Collection		
	Wavelength (Å)	0.9795
	Space Group	C 2 2 21
	a, b, c (Å)	100.585, 176.8950, 238.4130
	Resolution (Å)	49.26-2.10 (2.14-2.10)
	Number of Reflections	
	Total	556847 (28446)
	Unique	123558 (6073)
	Completeness (%)	99.9 (100)
	Multiplicity	4.5 (4.7)
	<I/σ(I)>	15.7 (1.7)
	R$_{merge}$	0.061 (0.999)
	R$_{pim}$	0.033 (0.522)
	Wilson B factor (Å^2)	37.2
Refinement		
	Resolution (Å)	49.26-2.10
	Number of Reflections	
	Working	122260
	Free	1242
	R$_{work}$	18.8%
	R$_{free}$	22.1%
	Mean B factor (Å^2)	25.8
	Geometry	
	RMSD Bond Lengths	0.008
	RMSD Bond Angles	1.22
	Ramachandran plot (%)	
	Favoured	97.87
	Alowed	2.13

Table 2: Analysis of DusC^{C98A}/tRNA^{phe} geometry by *MolProbity* webserver [206].



All-Atom Contacts	Clashscore, all atoms:	2.9		99th percentile* (N=576, 2.10Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	0	0.00%	Goal: <1%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	917	97.87%	Goal: >98%
	MolProbity score^	1.11		100th percentile* (N=11758, 2.10Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad backbone bonds:	0 / 7599	0.00%	Goal: 0%
	Bad backbone angles:	0 / 10349	0.00%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	0	0.00%	Goal: 0
	Bad backbone conformations#:	26	12.09%	Goal: <= 5%
	Bad bonds:	0 / 5063	0.00%	Goal: 0%
	Bad angles:	7 / 7893	0.09%	Goal: <0.1%

In the two column results, the left column gives the raw count, right column gives the percentage.

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst. For clashscore the comparative set of structures was selected in 2004, for MolProbity score in 2006.

RNA backbone was recently shown to be rotameric. Outliers are RNA suites that don't fall into recognized rotamers.

^ MolProbity score combines the clashscore, rotamer, and Ramachandran evaluations into a single score, normalized to be on the same scale as X-ray resolution.

5.3 Results

5.3.1 Structure of DusC-tRNA^{phe} complex

In order to elucidate the mechanism of DusC binding and catalysis, a stable complex of C98A mutant enzyme with *E. coli* tRNA^{phe} was prepared and diffraction quality crystals of this complex were obtained. The best crystals belonged to the orthorhombic space group C222$_1$ and diffracted to a resolution of 2.1Å. The structure was solved by molecular replacement and refined to a R$_{work}$/R$_{free}$ of 18.8/22.1%. Three copies of the complex were found in the asymmetric unit (**Fig. 5.6**). Protein chains A,B and C were highly similar to each other, superimposing with a maximum RMSD of 0.3Å over 1164 main chain atoms (residues 5-300), as were tRNA chains D,E and F, with a maximum RMSD of 0.6Å over 792 backbone atoms (nucleotides 4-69).

Several magnesium ions were also present in the structure: tRNA chain D containing five ions clustered around the D and TΨC loops; chain E containing five ions found mainly throughout the anticodon stem; and chain F containing six spread out throughout the tRNA structure (**Fig. 5.7a**). In addition, three manganese ions were identified using the anomalous difference maps, with one ion found per tRNA chain, always in the same position adjacent to the D loop (**Fig. 5.7b**). As $MnCl_2$ was present in the crystallization buffer, it is likely that the manganese ions came from there. In addition, although crystals could be grown in the absence of manganese, the diffraction quality was poorer, indicating that the presence of manganese was important for high quality crystal packing.

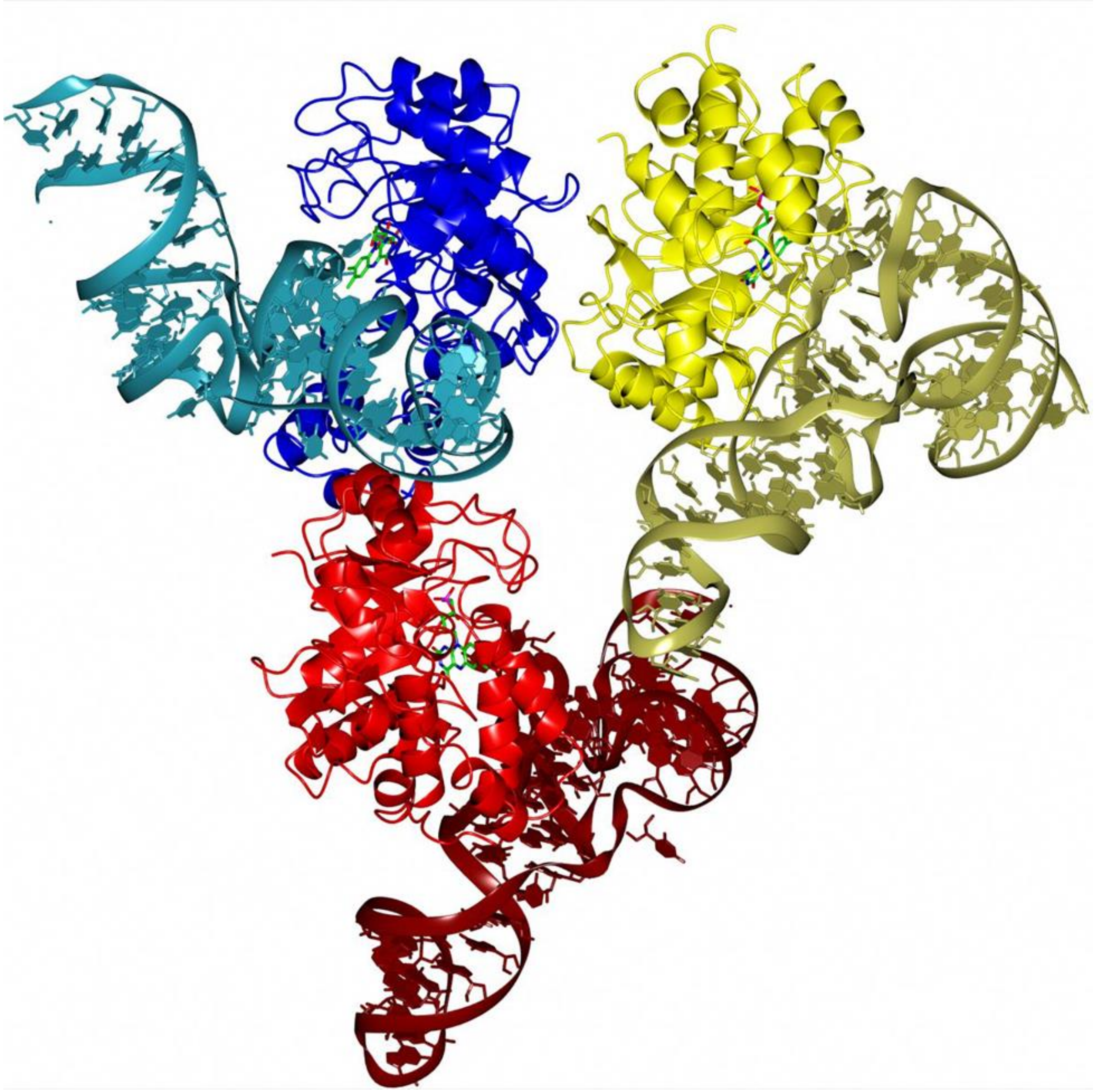


Figure 5.6 Asymmetric unit of the DusC^{C98A}:tRNA^{phe} complex.

Protein/RNA chains A/D, B/E and C/F are depicted in blue, red and yellow respectively.

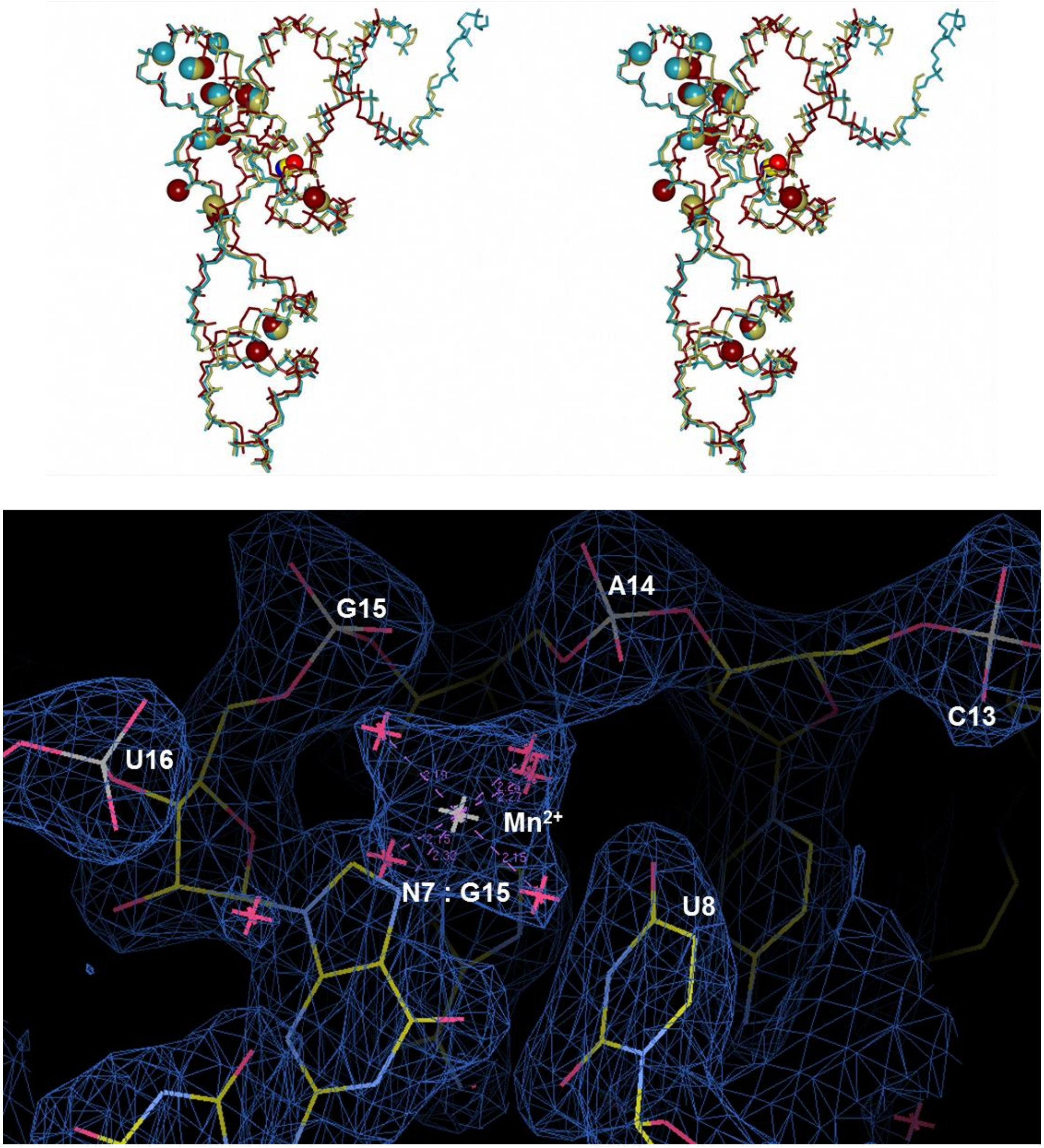


Figure 5.7: Metal ion binding in the *Ec*DusC^{C98A}:tRNA^{phe} complex.

(a) Superposition of the three tRNA chains present in the unit cell of the complex (cyan, gold and burgundy). Bound magnesium ions are shown as spheres coloured according to chain, and one manganese ion is shown for each molecule (blue, yellow and red). (b) Uncommon co-ordination of magnesium by the N7 atom of a nucleotide base. The co-ordination distance is very similar to water (pink crosses) coordination. Electron density from the 2*m*F$_o$-*D*F$_c$ map is shown contoured at 1σ.

In the structure, the N-terminal TIM barrel catalytic domain contacts the D-stem loop, whilst the C-terminal recognition domain mainly interacts with the D stem loop and the

TΨC stem loop. In contrast, the anticodon and acceptor stem loops point away from the enzyme and are not contacted. The result of this interaction is to insert uracil 16 into the active site of the enzyme, to place it in a stacking interaction with the plane of the FMN (**Fig. 5.8a**).

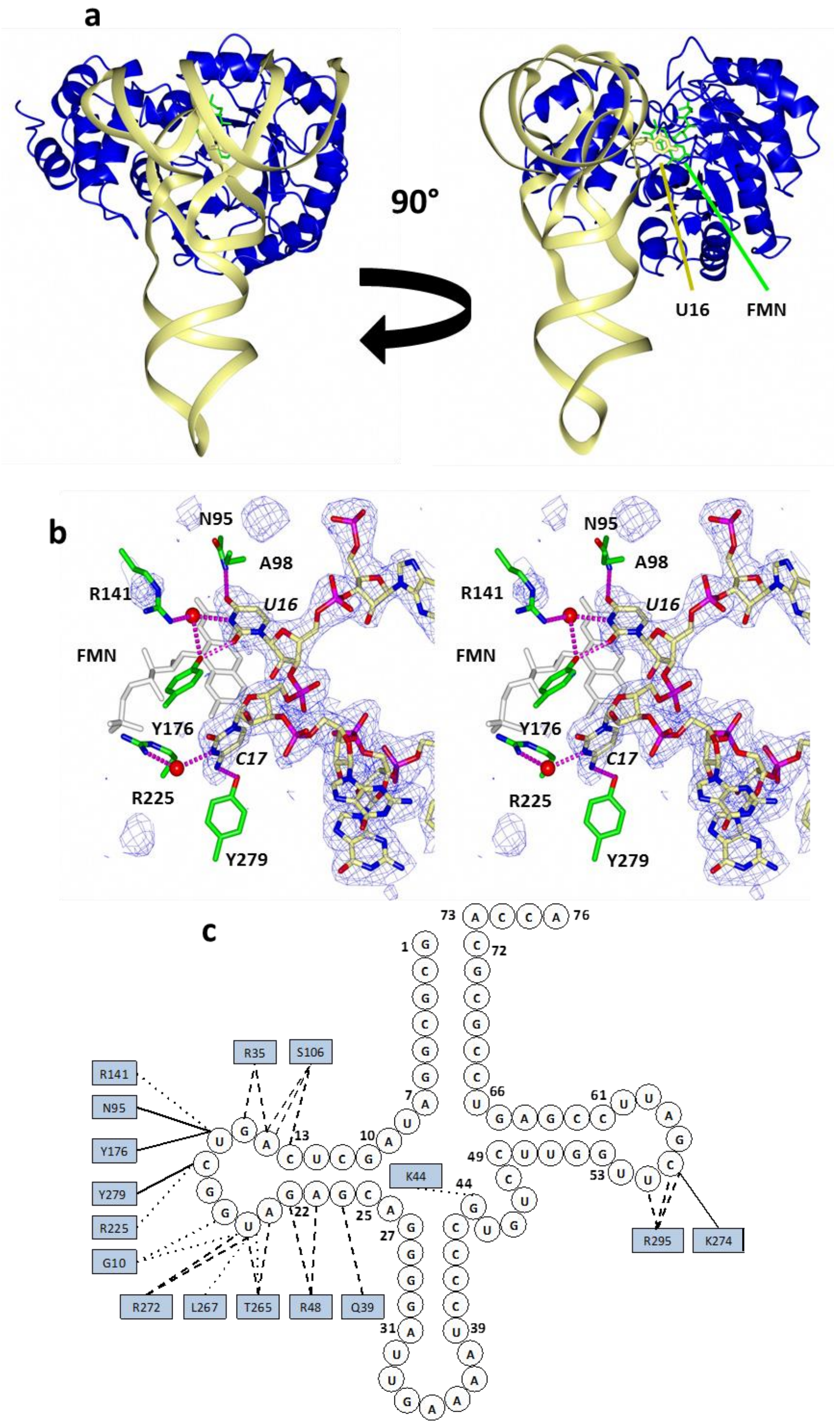


Figure 5.8: Structure of the DusC^{C98A}:tRNA^{phe} complex.

(a) tRNA^{phe} (yellow ribbon) bound to DusC^{C98A} (blue ribbon) (left) rotated 90° about the y-axis (right) with substrate target residue U16 (yellow sticks) and cofactor FMN (green sticks) indicated. (b) Stereo view of the active site with electron density from the 2*m*F$_o$-*D*F$_c$ nucleotide omit map contoured at 3σ (blue); tRNA (yellow, numbering in italics); side-chains (green); FMN (white), waters (red spheres) and enzyme side-chains (green sticks). (c) Schematic of residues hydrogen bonded to tRNA through side-chains (solid lines), protein backbone (dashed lines) and water mediated (dotted lines).

Binding of the tRNA by the protein results in a buried surface area of ~1057Å^2. In total, 26 hydrogen bonds are formed between the DusC and tRNA^{phe}, out of which only four are made to RNA bases (**Fig. 5.8 b,c**). Two of the four hydrogen bonds are to the active site uracil, U16, one to C17 and another to the fully conserved C56. The remaining 22 hydrogen bonding interactions are comprised of 7 water mediated interactions and hydrogen bonds to the sugar-phosphate backbone atoms in the tRNA. The majority of these hydrogen bonds form between the protein and the D loop, but several also form between the recognition domain of the protein and the TΨC loop. This region also contains one of the four specific hydrogen bonds, which forms between the side chain of K274 and the O2 carbonyl group of C56, which is located in the elbow of the tRNA.

There are no significant conformational changes that take place in the DusC structure upon binding, which can be seen by comparison with the structure of tRNA-free protein (PDB code 4BFA), which superimposes with an RMSD of 0.5 Å (1152 main chain atoms, residues 5-300). The most prominent difference is in the active site loop (residues 98-107), which becomes disordered upon binding tRNA. This loop is disordered in one of the two protein molecules in the tRNA-free structures and may be inherently flexible. This could be advantageous for the enzyme, as the flexibility of the loop may allow tRNA molecules with different length D-loops to dock without hindrance.

Similarly, the structure of the tRNA does not change significantly upon binding (**Fig. 5.9a**), with the RMSD between the bound and unbound (3L0U) tRNAs being 2.3 Å (792 backbone atoms, nucleotides 4-69). Here, differences are mainly in the D loop with the largest differences being found for the active site nucleotides U16 and C17 (**Fig. 5.9b**).

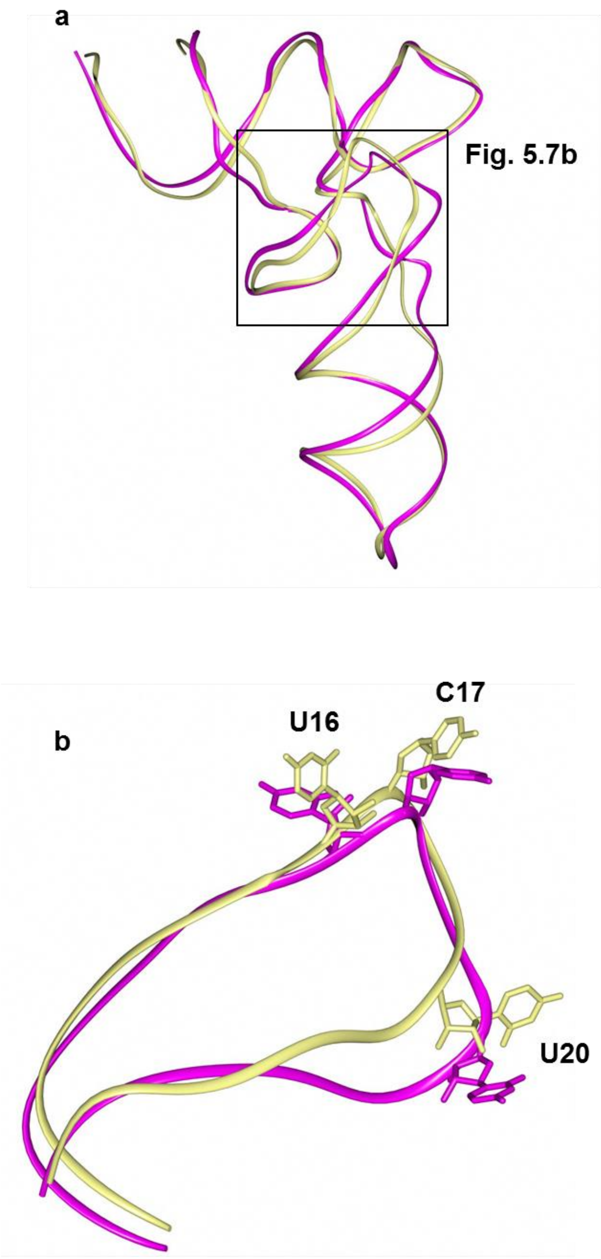


Figure 5.9: Comparison of bound and standalone tRNA^{phe}.

(a) Superposition of *Ec*DusC bound (blue) and unbound tRNA^{phe} (purple, 3l0u). (b)A close-up image of the D-loop (defined by a box in (a)) shows remodeling of residues neighbouring the target residue U16.

These comparisons therefore suggest that DusC-tRNA^{phe} complexes form through rigid-body docking of the two molecules followed by induced fit, where small conformational adjustments take place around the active site of the protein and D loop of the tRNA.

5.3.2 Recognition and modification of U16 by DusC

Examination of the active site reveals that the U16 nucleotide has been inserted to stack above the FMN co-factor. The proposed reaction mechanism for the Dus enzymes involves a transfer of hydride from the N5 atom of FMN onto the C6 atom of the substrate uridine followed by a transfer of a proton from Sγ of C98 onto the C5 atom of the substrate uridine [142]. Due to the position of the U16 relative to the rest of the active site, it is likely that the modification process follows the same mechanism for DusC, and that U16 is the base that would be modified in the wild-type enzyme. This correlates well with the results showing U16 as the base that is specifically recognised by DusC.

U16 is bound in the active site through direct hydrogen bonds between the side chains of Y176 and N95, which are both highly conserved throughout the DusC subfamily, and the O2 and O4 oxygen atoms of carbonyl groups of the base. In addition, a water-mediated hydrogen bond also forms between the N5 endocyclic amine group of the base, and the highly conserved residues R141 and H168, which were previously found to be essential for DusA activity [212] (**Fig. 5.10**). The water molecules that take part in this interaction occupy a region of electron density that was previously unassignable, and assumed to be an unidentified co-factor [187] (3B0U and 3B0V). Although the difference density maps indicate that there could potentially be another two water molecules present, further refinement suggests that these sites have only partial occupancy, and that there is no evidence for an additional co-factor.

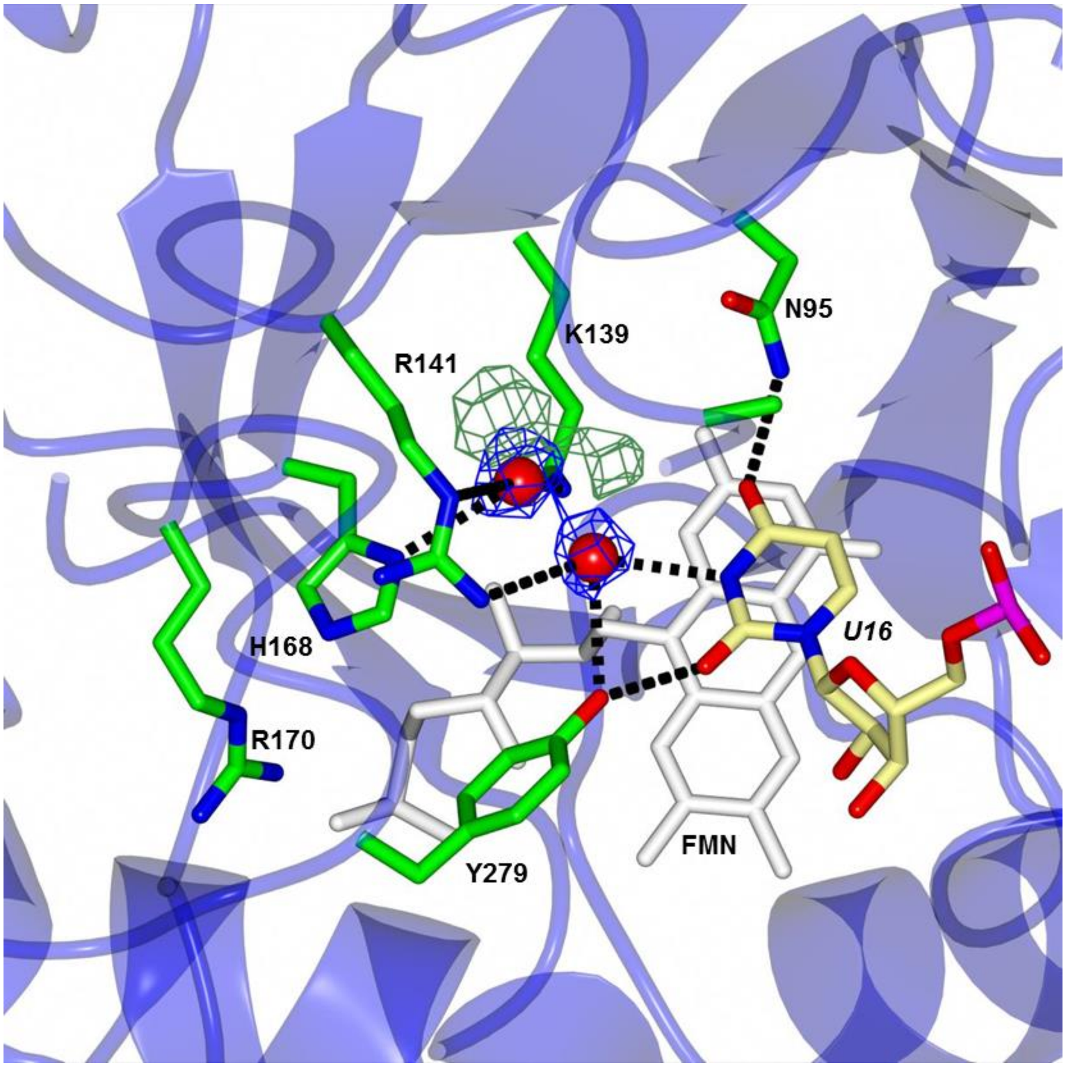


Figure 5.10: Difference density in the active site of *Ec*DusC.

Previously unidentified positive difference density in the active site of other Dus enzymes is present in the active site of *Ec*DusC $2F_o$-F_c map contoured at 1σ (blue) and F_o-F_c map contoured at 4σ (green and red). The density and refinement suggest that it is due to the presence of two water molecules (red spheres). Also shown are FMN (green sticks), target uracil U16 (cyan sticks), and active site side-chains (green sticks).

The neighbouring nucleotide, C17, is also bound in the active site forming a direct hydrogen bonding interactions between its N4 exocyclic amine group and the side chain of Y279, as well as through water mediated hydrogen bond between the N3 imine group and the side chain of R225. In all *E. coli* tRNA sequences with D at position 16, position

17 is either a C or U, and it is clear from the hydrogen bonding network that U could be bound in the same position as C if it was present instead.

5.4 Discussion

5.4.1 Comparison with the *Tt*Dus structures

Comparison of the position-16 specific DusC structures with the position-20 specific *Tt*Dus structures reveals how the enzymes are able to maintain specificity for different nucleotide positions. Examination of the unbound DusC structure (PDB code 4BFA) with the unbound *Tt*Dus structure shows that the proteins adopt a highly similar fold with RMSD's of 2.0 Å (212 Cα atoms) and 2.5 Å (42 Cα atoms) for the catalytic and recognition domains, respectively (**Fig. 5.11a**). In the structure of the *Tt*Dus-tRNA^{phe} complex, however, the tRNA substrate is rotated by ~160° relative to the *Ec*DusC^{C98A}-tRNA^{phe} complex structure, allowing U20 to be inserted into the active site instead of U16 (**Fig. 5.11b**).

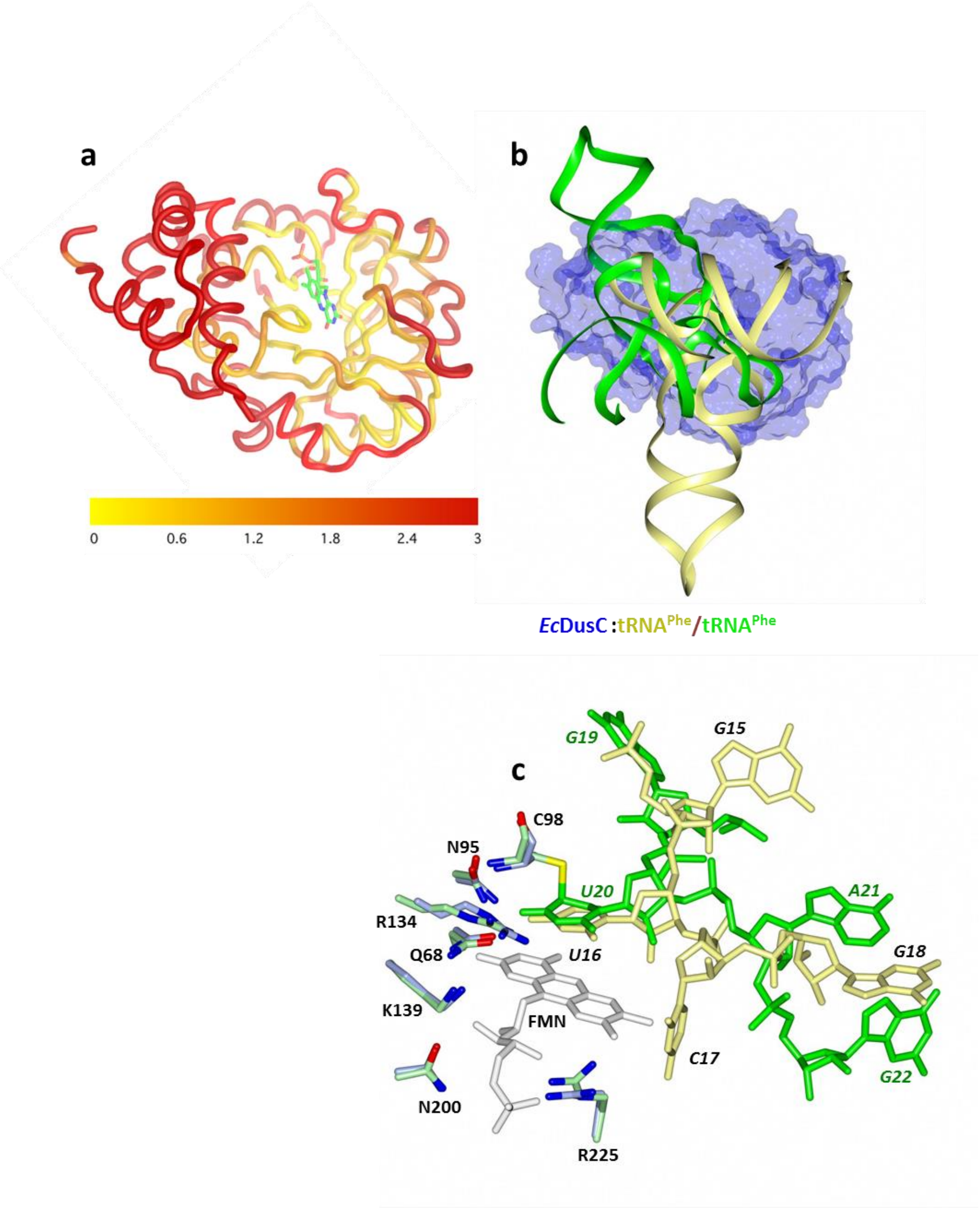


Figure 5.11: Comparison of *Ec*DusC and *Tt*Dus and their complexes with tRNA^{phe}.

(a) The structure of *Ec*DusC coloured according to the distance between atom pairs after superposition with *Tt*Dus by *ProSMART* alignment. Figure courtesy of Dr.Rob Byrne, Antson group. (b) Superposition of the enzymes within their respective tRNA complexes reveal that tRNA (yellow ribbon) is bound with a ~160° rotation by the *E. coli* enzyme (blue surface) relative to *Tt*Dus:tRNA^{phe} (green ribbon). (c) Superposition of FMN (white) shows that the position of enzyme substrate target residues U16 (yellow)/U20 (green) is invariant in the active site of both complexes *Ec*DusC (light blue) and *Tt*Dus (pale green), respectively.

The cause of this rearrangement is due to differences in the distribution of charges on the surface of each enzyme. The DusC enzyme surface contains a positively charged, “L-shaped” groove into which the tRNA slots (**Fig. 5.12a**), contrasting with the surface of the *Tt*Dus enzyme which has a cylindrically shaped groove (**Fig. 5.12b**), causing the tRNA to dock in the opposite orientation. tRNA molecules are thus prevented from binding in the alternative orientation due to mismatching charges between the phosphates of the tRNA backbone and the protein side-chains.

Next, sequence alignments of position-16 specific DusC subfamily members (**Fig. 5.13**) and position-20 specific DusA subfamily members (**Fig. 5.14**) (sequences selected by Dr.Rob Byrne, Antson group) were generated to investigate whether the pattern of charges was conserved throughout subfamily members (**Fig. 5.15**). Several putative, subfamily specific “hotspots” were identified, which define the orientation of the tRNA on the surface of the enzyme (**Fig. 5.12 c,d**).

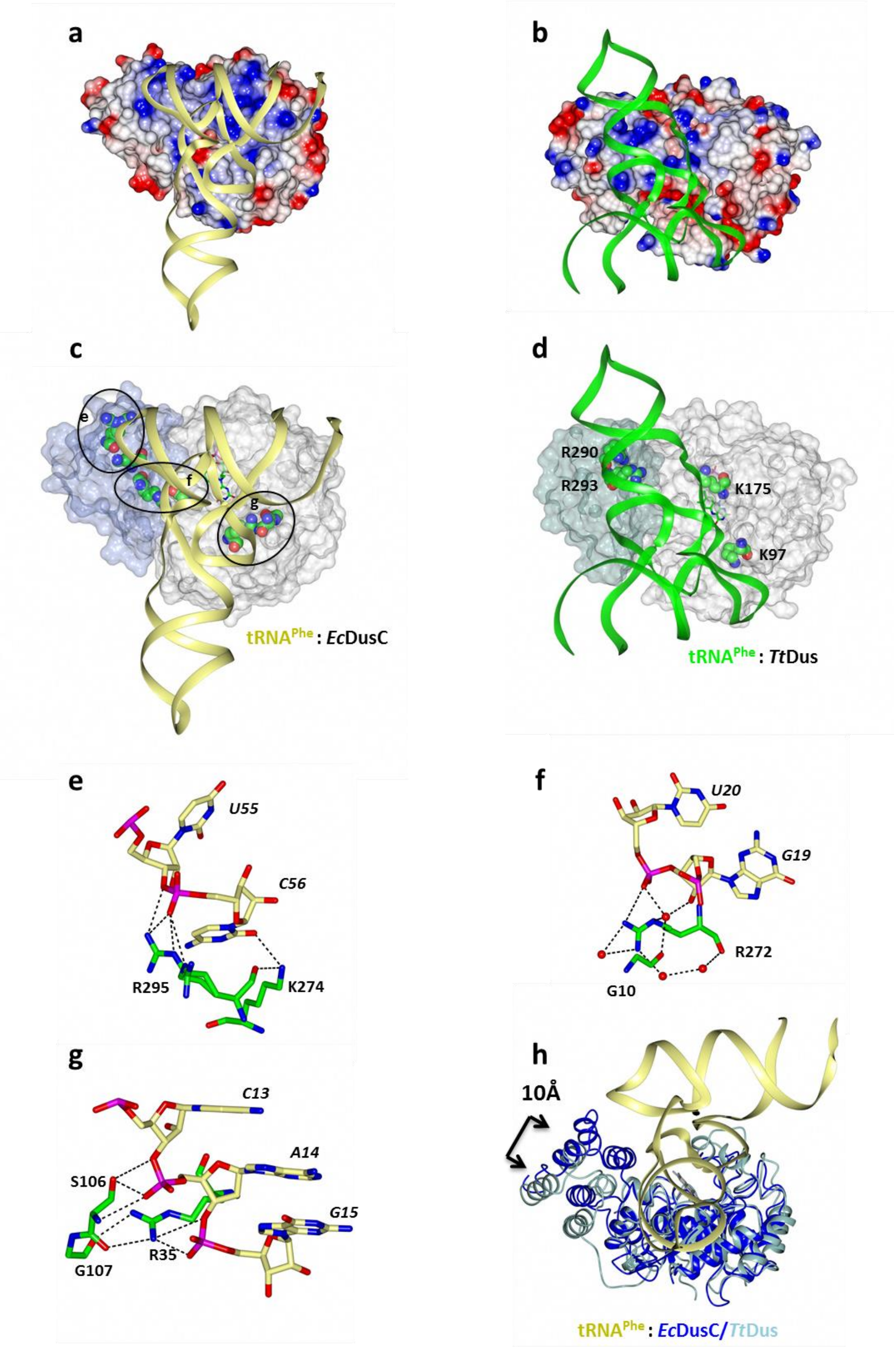


Figure 5.12: Structural features that determine specificity in Dus enzymes.

The electrostatic surface potential was calculated using unbound enzyme and rendered for the tRNA^{phe} complex (blue, positive; red, negative) for (a) *Ec*DusC and (b) *Tt*Dus. Specific residues proposed to determine tRNA binding orientation (spheres) by (c) *Ec*DusC (recognition domain, blue; enzyme core, white) and (d) *Tt*Dus(recognition domain pale green; enzyme core, white). *Ec*DusC 'Hot-spots' are circled in black; lettering corresponds to subsequent figures (e, f and g). (e), (f) and (g) Illustrate the molecular interactions between *Ec*DusC and tRNA^{phe} highlighted in (a). (h) Superposition of the catalytic domains of *Ec*DusC (blue) and *Tt*Dus (pale green) reveals a high level of structural conservation, with a change in relative disposition of the recognition domain.

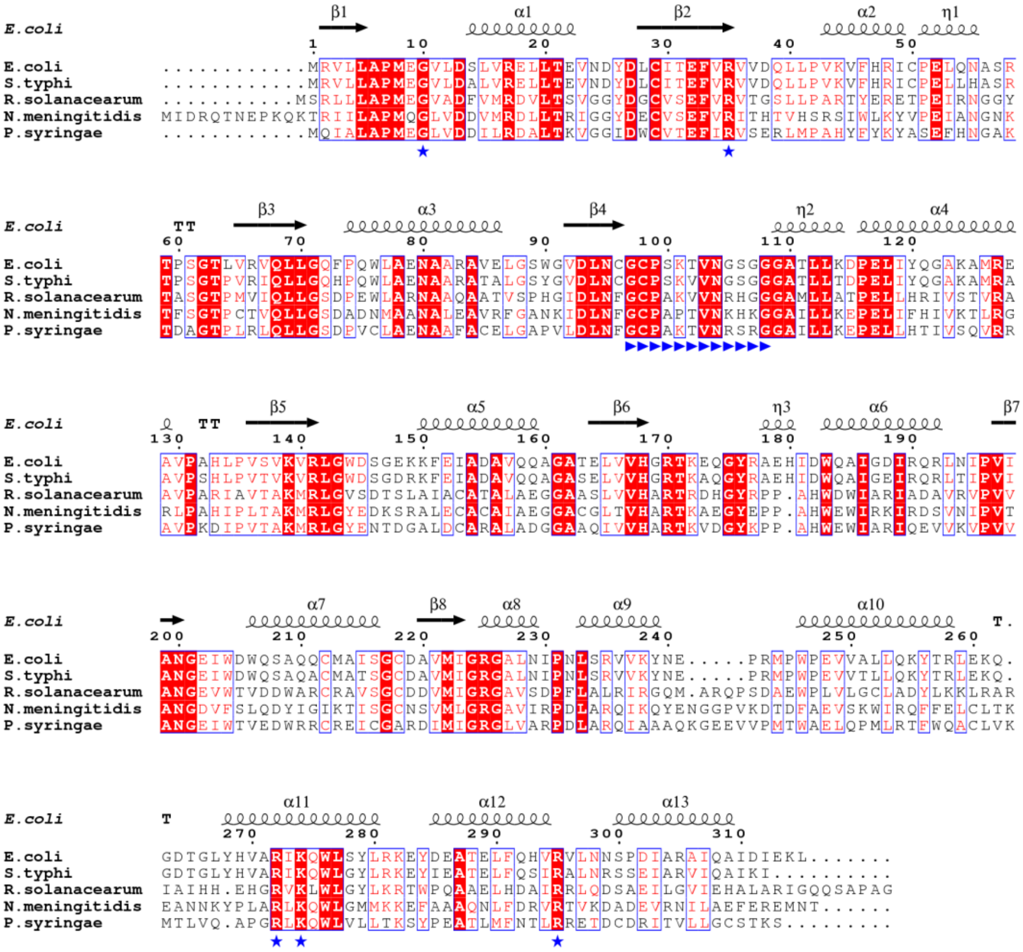

E.coli
β1
α1
β2
α2
η1
1
10
20
30
40
50
E.coli ............MRVLLAPMEGVLDSLVRELLTEVNDYDLCITEFVRVVDQLLPVKVFHRICPELQNASR
S.typhi ............MRVLLAPMEGVLDALVRELLTEVNDYDLCITEFVRVVDQLLPVKVFHRICPELLHASR
R.solanacearum ...........MSRLLLAPMEGVADFVMRDVLTSVGGYDGCVSEFVRVTGSLLPARTYERETPEIRNGGY
N.meningitidis MIDRQTNEPKQKTRIILAPMQGLVDDVMRDLLTRIGGYDECVSEFVRITHTVHSRSIWLKYVPEIANGNK
P.syringae ............MQIALAPMEGLVDDILRDALTKVGGIDWCVTEFIRVSERLMPAHYFYKYASEFHNGAK
E.coli
TT
β3
α3
β4
η2
α4
60
70
80
90
100
110
120
E.coli TPSGTLVRVQLLGQFPQWLAENAARAVELGSWGVDLNCGCPSKTVNGSGGGATLLKDPELIYQGAKAMRE
S.typhi TPSGTPVRIQLLGQHPQWLAENAARATALGSYGVDLNCGCPSKVVNGSGGGATLLTDPELIYQGAKAMRA
R.solanacearum TASGTPMVIQLLGSDPEWLARNAAQAATVSPHGIDLNFGCPAKVVNRHGGGAMLLATPELLHRIVSTVRA
N.meningitidis TFSGTPCTVQLLGSDADNMAANALEAVRFGANKIDLNFGCPAPTVNKHKGGAILLKEPELIFHIVKTLRG
P.syringae TDAGTPLRLQLLGSDPVCLAENAAFACELGAPVLDLNFGCPAKTVNRSRGGAILLKEPELLHTIVSQVRR
E.coli
TT
β5
α5
β6
η3
α6
β7
130
140
150
160
170
180
190
E.coli AVPAHLPVSVKVRLGWDSGEKKFEIADAVQQAGATELVVHGRTKEQGYRAEHIDWQAIGDIRQRLNIPVI
S.typhi AVPSHLPVTVKVRLGWDSGDRKFEIADAVQQAGASELVVHGRTKAQGYRAEHIDWQAIGEIRQRLTIPVI
R.solanacearum AVPARIAVTAKMRLGVSDTSLAIACATALAEGGAASLVVHARTRDHGYRPP.AHWDWIARIADAVRVPVV
N.meningitidis RLPAHIPLTAKMRLGYEDKSRALECACAIAEGGACGLTVHARTKAEGYEPP.AHWEWIRKIRDSVNIPVT
P.syringae AVPKDIPVTAKMRLGYENTDGALDCARALADGGAAQIVVHARTKVDGYKPP.AHWEWIARIQEVVKVPVV
E.coli
α7
β8
α8
α9
α10
T.
200
210
220
230
240
250
260
E.coli ANGEIWDWQSAQQCMAISGCDAVMIGRGALNIPNLSRVVKYNE.....PRMPWPEVVALLQKYTRLEKQ.
S.typhi ANGEIWDWQSAQACMATSGCDAVMIGRGALNIPNLSRVVKYNE.....PRMPWPEVVTLLQKYTRLEKQ.
R.solanacearum ANGEVWTVDDWARCRAVSGCDDVMIGRGAVSDPFLALRIRGQM.ARQPSDAEWPLVLGCLADYLKKLRAR
N.meningitidis ANGDVFSLQDYIGIKTISGCNSVMLGRGAVIRPDLARQIKQYENGGPVKDTDFAEVSKWIRQFFELCLTK
P.syringae ANGEIWTVEDWRRCREICGARDIMIGRGLVARPDLARQIAAAQKGEEVVPMTWAELQPMLRTFWQACLVK
E.coli
T
α11
α12
α13
270
280
290
300
310
E.coli GDTGLYHVARIKQWLSYLRKEYDEATELFQHVRVLNNSPDIARAIQAIDIEKL.......
S.typhi GDTGLYHVARIKQWLGYLRKEYIEATELFQSIRALNRSSEIARVIQAIKI..........
R.solanacearum IAIHH.EHGRVKLWLGYLKRTWPQAAELHDAIRRLQDSAEILGVIEHALARIGQQSAPAG
N.meningitidis EANNKYPLARLKQWLGMMKKEFAAAQNLFDRVRTVKDADEVRNILAEFEREMNT......
P.syringae MTLVQ.APGRLKQWLVLLTKSYPEATLMFNTLRRETDCDRITVLLGCSTKS.........
★
★
★
★
★

Figure 5.13: Sequence alignment of DusC type enzymes from homologues with 30-50% sequence identity.

Sequences selected by Dr.Rob Byrne, Antson group. The secondary structure of *Ec*DusC is shown, and DusC putative 'hotspot' residues are highlighted (blue stars). Blue triangles denote the position of the active site loop.

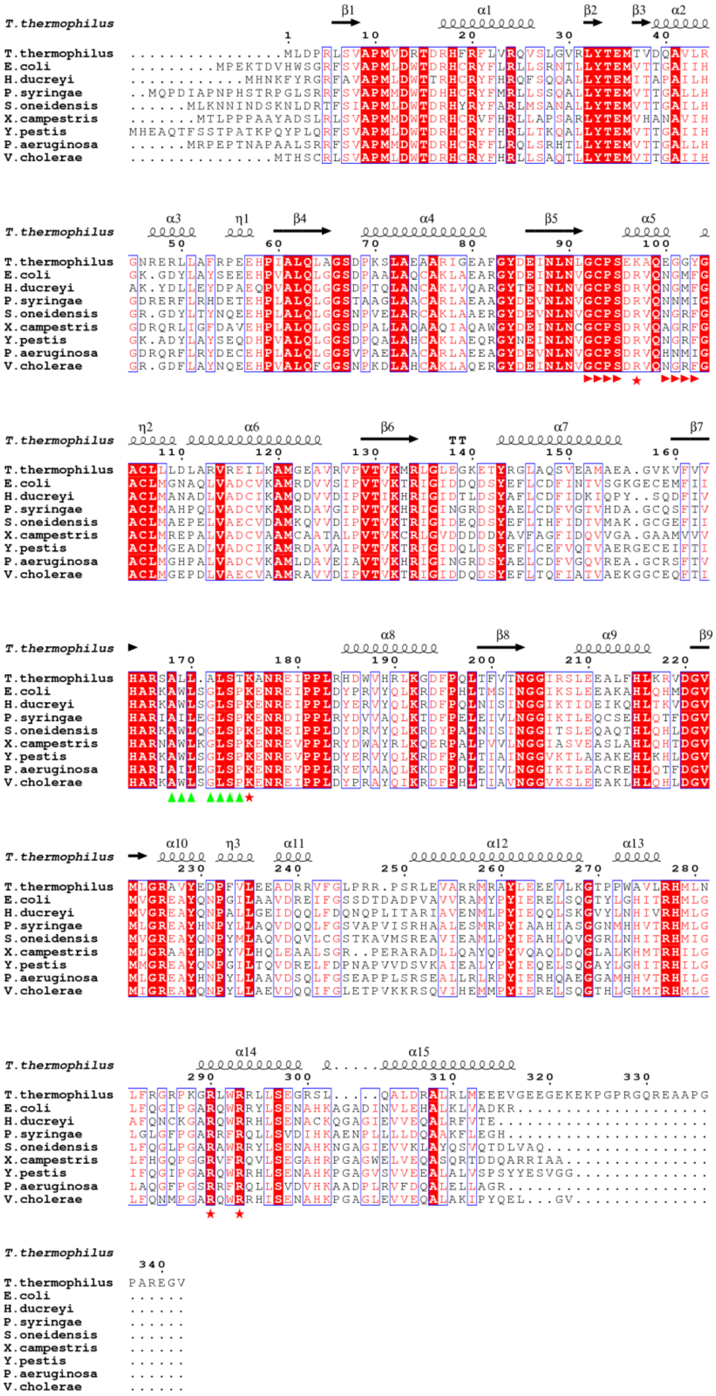
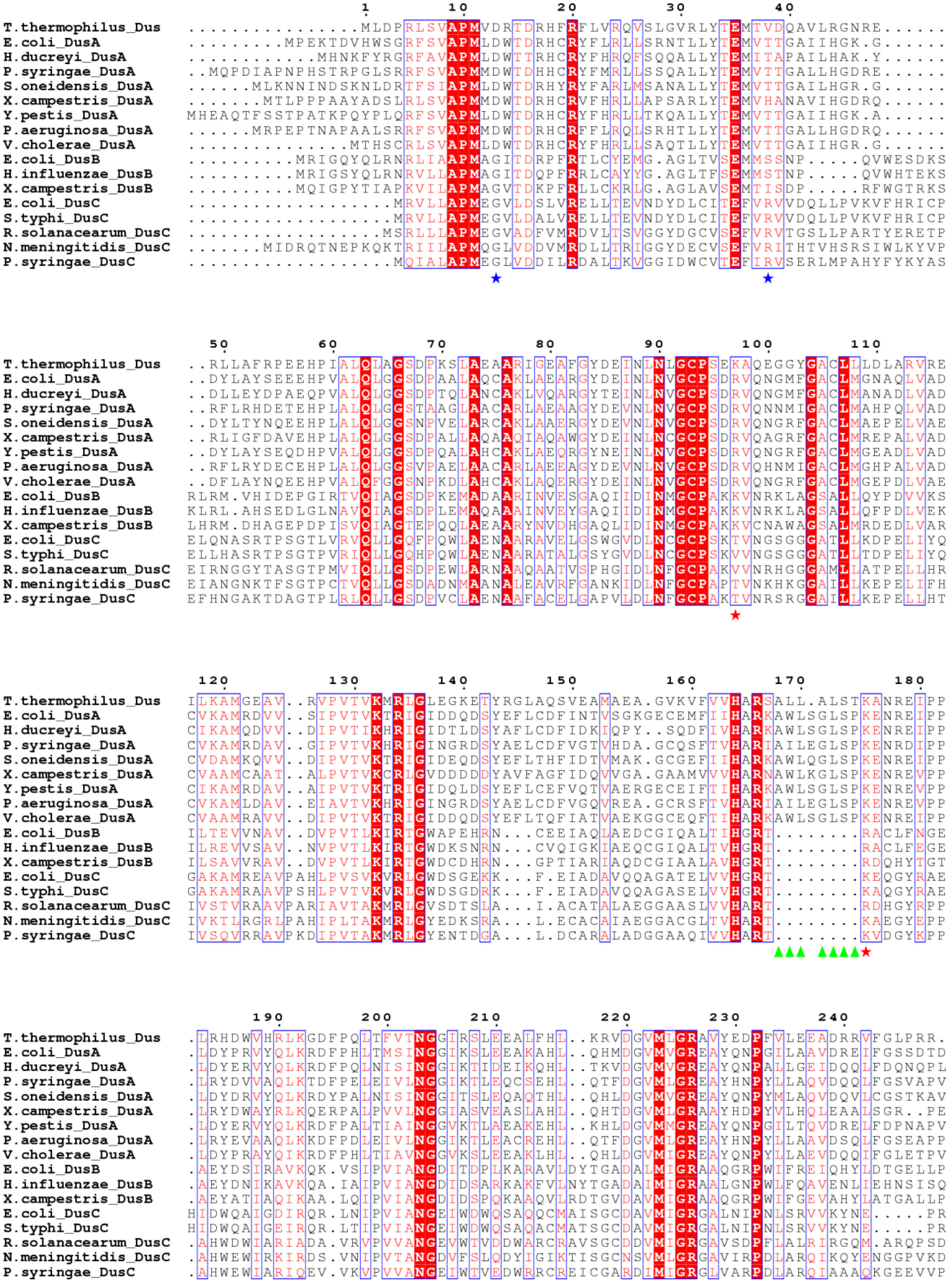


Figure 5.14: Sequence alignment of DusA type enzymes from homologues with 30-50% sequence identity.

Sequences selected by Dr.Rob Byrne, Antson group. The secondary structure of *Tt*Dus is shown, and DusA specific residues are highlighted (red stars). The conserved 7-reside $\beta\alpha_6$ insert is highlighted by green triangles.



```
                                      1       10        20        30        40
T.thermophilus_Dus    ................MLDPRLSVAPMVDRTDRHFRFLVRQVSLGVRLYTEMTVDQAVLRGNRE......
E.coli_DusA           .........MPEKTDVHWSGRFSVAPMLDWTDRHCRYFLRLLSRNTLLYTEMVTTGAIIHGK.G......
H.ducreyi_DusA        ............MHNKFYRGRFAVAPMLDWTTRHCRYFHRQFSQQALLYTEMITAPAILHAK.Y......
P.syringae_DusA       ..MQPDIAPNPHSTRPGLSRRFSVAPMMDWTDHHCRYFMRLLSSQALLYTEMVTTGALLHGDRE......
S.oneidensis_DusA     ......MLKNNINDSKNLDRTFSIAPMLDWTDRHYRYFARLMSANALLYTEMVTTGAILHGR.G......
X.campestris_DusA     .......MTLPPPAAYADSLRLSVAPMMDWTDRHCRVFHRLLAPSARLYTEMVHANAVIHGDRQ......
Y.pestis_DusA         MHEAQTFSSTPATKPQYPLQRFSVAPMLDWTDRHCRYFHRLLTKQALLYTEMVTTGAIIHGK.A......
P.aeruginosa_DusA     ......MRPEPTNAPAALSRRFSVAPMMDWTDRHCRFFLRQLSRHTLLYTEMVTTGALLHGDRQ......
V.cholerae_DusA       ...............MTHSCRLSVAPMLDWTDRHCRYFHRLLSAQTLLYTEMVTTGAIIHGR.G......
E.coli_DusB           ..........MRIGQYQLRNRLIAAPMAGITDRPFRTLCYEMG.AGLTVSEMMSSNP.....QVWESDKS
H.influenzae_DusB     ..........MRIGSYQLRNRVLLAPMAGITDQPFRRLCAYYG.AGLTFSEMMSTNP.....QVWHTEKS
X.campestris_DusB     ..........MQIGPYTIAPKVILAPMAGVTDKPFRLLCKRLG.AGLAVSEMTISDP.....RFWGTRKS
E.coli_DusC           ...................MRVLLAPMEGVLDSLVRELLTEVNDYDLCITEFVRVVDQLLPVKVFHRICP
S.typhi_DusC          ...................MRVLLAPMEGVLDALVRELLTEVNDYDLCITEFVRVVDQLLPVKVFHRICP
R.solanacearum_DusC   ..................MSRLLLAPMEGVADFVMRDVLTSVGGYDGCVSEFVRVTGSLLPARTYERETP
N.meningitidis_DusC   .......MIDRQTNEPKQKTRIILAPMQGLVDDVMRDLLTRIGGYDECVSEFVRITHTVHSRSIWLKYVP
P.syringae_DusC       ...................MQIALAPMEGLVDDILRDALTKVGGIDWCVTEFIRVSERLMPAHYFYKYAS
                                                  ★                        ★

                        50        60        70        80        90        100       110
T.thermophilus_Dus    ..RLLAFRPEEHPIALQLAGSDPKSLAEAARIGEAFGYDEINLNLGCPSEKAQEGGYGACLLLDLARVRE
E.coli_DusA           ..DYLAYSEEEHPVALQLGGSDPAALAQCAKLAEARGYDEINLNVGCPSDRVQNGMFGACLMGNAQLVAD
H.ducreyi_DusA        ..DLLEYDPAEQPVALQLGGSDPTQLANCAKLVQARGYTEINLNVGCPSDRVQNGMFGACLMANADLVAD
P.syringae_DusA       ..RFLRHDETEHPLALQLGGSTAAGLAACARLAEAAGYDEVNLNVGCPSDRVQNNMIGACLMAHPQLVAD
S.oneidensis_DusA     ..DYLTYNQEEHPLALQLGGSNPVELARCAKLAAERGYDEVNLNVGCPSDRVQNGRFGACLMAEPELVAE
X.campestris_DusA     ..RLIGFDAVEHPLALQLGGSDPALLAQAAQIAQAWGYDEINLNCGCPSDRVQAGRFGACLMREPALVAD
Y.pestis_DusA         ..DYLAYSEQDHPVALQLGGSDPQALAHCAKLAEQRGYNEINLNVGCPSDRVQNGRFGACLMGEADLVAD
P.aeruginosa_DusA     ..RFLRYDECEHPLALQLGGSVPAELAACARLAEEAGYDEVNLNVGCPSDRVQHNMIGACLMGHPALVAD
V.cholerae_DusA       ..DFLAYNQEEHPVALQFGGSNPKDLAHCAKLAQERGYDEINLNVGCPSDRVQNGRFGACLMGEPDLVAE
E.coli_DusB           RLRM.VHIDEPGIRTVQIAGSDPKEMADAARINVESGAQIIDINMGCPAKKVNRKLAGSALLQYPDVVKS
H.influenzae_DusB     KLRL.AHSEDLGLNAVQIAGSDPLEMAQAAAINVEYGAQIIDINMGCPAKKVNRKLAGSALLQFPDLVEK
X.campestris_DusB     LHRM.DHAGEPDPISVQIAGTEPQQLAEAARYNVDHGAQLIDINMGCPAKKVCNAWAGSALMRDEDLVAR
E.coli_DusC           ELQNASRTPSGTLVRVQLLGQFPQWLAENAARAVELGSWGVDLNCGCPSKTVNGSGGATLLKDPELIYQ
S.typhi_DusC          ELLHASRTPSGTPVRIQLLGQHPQWLAENAARATALGSYGVDLNCGCPSKVVNGSGGATLLTDPELIYQ
R.solanacearum_DusC   EIRNGGYTASGTPMVIQLLGSDPEWLARNAAQAATVSPHGIDLNFGCPAKVVNRHGGAMLLATPELLHR
N.meningitidis_DusC   EIANGNKTFSGTPCTVQLLGSDADNMAANALEAVRFGANKIDLNFGCPAPTVNKHKGAILLKEPELIFH
P.syringae_DusC       EFHNGAKTDAGTPLRLQLLGSDPVCLAENAAFACELGAPVLDLNFGCPAKTVNRSRGAILLKEPELLHT
                                                                        ★

                      120         130       140       150        160       170        180
T.thermophilus_Dus    ILKAMGEAV..RVPVTVKMRLGLEGKETYRGLAQSVEAMAEA.GVKVFVVHARSALL.ALSTKANREIPP
E.coli_DusA           CVKAMRDVV..SIPVTVKTRIGIDDQDSYEFLCDFINTVSGKGECEMFIIHARKAWLSGLSPKENREIPP
H.ducreyi_DusA        CIKAMQDVV..DIPVTIKHRIGIDTLDSYAFLCDFIDKIQPY..SQDFIVHARKAWLSGLSPKENREIPP
P.syringae_DusA       CVKAMRDAV..GIPVTVKHRIGINGRDSYAELCDFVGTVHDA.GCQSFTVHARIAILEGLSPKENRDIPP
S.oneidensis_DusA     CVDAMKQVV..DIPVTVKTRIGIDEQDSYEFLTHFIDTVMAK.GCGEFIIHARKAWLQGLSPKENREIPP
X.campestris_DusA     CVAAMCAAT..ALPVTVKCRLGVDDDDDYAVFAGFIDQVVGA.GAAMVVVHARNAWLKGLSPKENREVPP
Y.pestis_DusA         CIKAMRDAV..AIPVTVKTRIGIDQLDSYEFLCEFVQTVAERGECEIFTIHARKAWLSGLSPKENREVPP
P.aeruginosa_DusA     CVKAMLDAV..EIAVTVKHRIGINGRDSYAELCDFVGQVREA.GCRSFTVHARIAILEGLSPKENREVPP
V.cholerae_DusA       CVAAMRAVV..DIPVTVKTRIGIDDQDSYEFLTQFIATVAEKGGCEQFTIHARKAWLSGLSPKENREIPP
E.coli_DusB           ILTEVVNAV..DVPVTLKIRTGWAPEHRN...CEEIAQLAEDCGIQALTIHGRT........RACLFNGE
H.influenzae_DusB     ILREVVSAV..NVPVTLKIRTGWDKSNRN...CVQIGKIAEQCGIQALTVHGRT........RACLFEGE
X.campestris_DusB     ILSAVVRAV..DVPVTLKIRTGWDCDHRN...GPTIARIAQDCGIAALAVHGRT........RDQHYTGT
E.coli_DusC           GAKAMREAVPAHLPVSVKVRLGWDSGEKK...F.EIADAVQQAGATELVVHGRT........KEQGYRAE
S.typhi_DusC          GAKAMRAAVPSHLPVTVKVRLGWDSGDRK...F.EIADAVQQAGASELVVHGRT........KAQGYRAE
R.solanacearum_DusC   IVSTVRAAVPARIAVTAKMRLGVSDTSLA...I.ACATALAEGGAASLVVHART........RDHGYRPP
N.meningitidis_DusC   IVKTLRGRLPAHIPLTAKMRLGYEDKSRA...L.ECACAIAEGGACGLTVHART........KAEGYEPP
P.syringae_DusC       IVSQVRRAVPKDIPVTAKMRLGYENTDGA...L.DCARALADGGAAQIVVHART........KVDGYKPP
                                                                          ▲▲▲ ▲▲▲▲★

                           190       200       210         220       230       240
T.thermophilus_Dus    .LRHDWVHRLKGDFPQLTFVTNGGIRSLEEALFHL..KRVDGVMLGRAVYEDPFVLEEADRRVFGLPRR.
E.coli_DusA           .LDYPRVYQLKRDFPHLTMSINGGIKSLEEAKAHL..QHMDGVMVGREAYQNPGILAAVDREIFGSSDTD
H.ducreyi_DusA        .LDYERVYQLKRDFPQLNISINGGIKTIDEIKQHL..TKVDGVMVGREAYQNPALLGEIDQQLFDQNQPL
P.syringae_DusA       .LRYDVVAQLKTDFPELEIVLNGGIKTLEQCSEHL..QTFDGVMLGREAYHNPYLLAQVDQQLFGSVAPV
S.oneidensis_DusA     .LDYDRVYQLKRDYPALNISINGGITSLEQAQTHL..QHLDGVMVGREAYQNPYMLAQVDQVLCGSTKAV
X.campestris_DusA     .LRYDWAYRLKQERPALPVVLNGGIASVEASLAHL..QHTDGVMLGRAAYHDPYVLHQLEAALSGR..PE
Y.pestis_DusA         .LDYERVYQLKRDFPALTIAINGGVKTLAEAKEHL..KHLDGVMMGREAYQNPGILTQVDRELFDPNAPV
P.aeruginosa_DusA     .LRYEVAAQLKKDFPDLEIVLNGGIKTLEACREHL..QTFDGVMLGREAYHNPYLLAAVDSQLFGSEAPP
V.cholerae_DusA       .LDYPRAYQIKRDFPHLTIAVNGGVKSLEEAKLHL..QHLDGVMIGREAYQNPYLLAEVDQQIFGLETPV
E.coli_DusB           .AEYDSIRAVKQK.VSIPVIANGDITDPLKARAVLDYTGADALMIGRAAQGRPWIFREIQHYLDTGELLP
H.influenzae_DusB     .AEYDNIKAVKQA.IAIPVIANGDIDSARKAKFVLNYTGADAIMIGRAALGNPWLFQAVENLIEHNSISQ
X.campestris_DusB     .AEYATIAQIKAA.LQIPVIANGDIDSPQKAAQVLRDTGVDAVMIGRAAQGRPWIFGEVAHYLATGALLP
E.coli_DusC           HIDWQAIGDIRQR.LNIPVIANGEIWDWQSAQQCMAISGCDAVMIGRGALNIPNLSRVVKYNE.....PR
S.typhi_DusC          HIDWQAIGEIRQR.LTIPVIANGEIWDWQSAQACMATSGCDAVMIGRGALNIPNLSRVVKYNE.....PR
R.solanacearum_DusC   .AHWDWIARIADA.VRVPVVANGEVWTVDDWARCRAVSGCDDVMIGRGAVSDPFLALRIRGQM.ARQPSD
N.meningitidis_DusC   .AHWEWIRKIRDS.VNIPVTANGDVFSLQDYIGIKTISGCNSVMLGRGAVIRPDLARQIKQYENGGPVKD
P.syringae_DusC       .AHWEWIARIQEV.VKVPVVANGEIWTVEDWRRCREICGARDIMIGRGLVARPDLARQIAAAQKGEEVVP
```

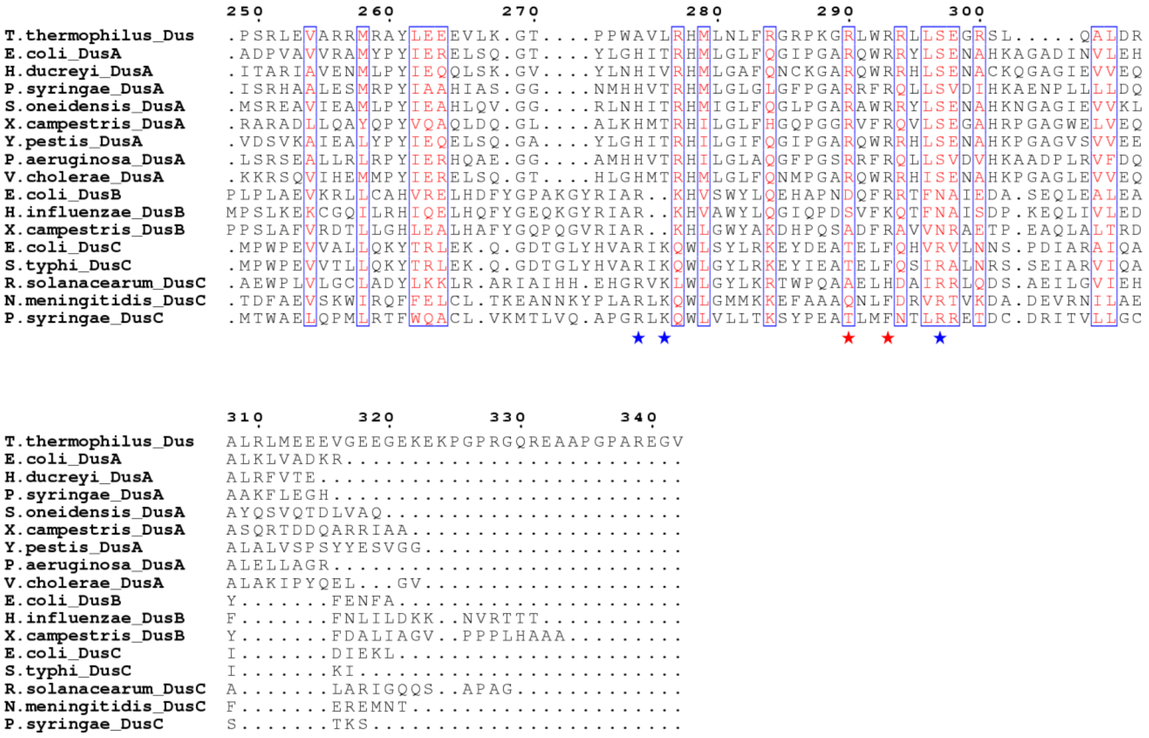


Figure 5.15: Sequence alignment of Dus enzymes from subfamilies DusA, DusB and DusC.

Regional conservation (red background, white text) and similarity (red text) are shown. DusC putative hotspots are shown as blue stars and DusA putative hotspots as red stars. The conserved 7-reside βα$_6$ insert is highlighted by green triangles.

In the position-16 specific enzymes, three putative hotspots were found: (i) K274/R295 (**Fig. 5.12e**), (ii) G10/R272 (**Fig. 5.12f**) and (iii) R35/S106/G107 (**Fig. 5.12g**). In comparison, the hotspots identified in the position-20 specific enzymes were K97, K175 and R290/R293 corresponding to residues T102, E173 and K282/D285 respectively in the DusC protein. In the first set, residues K274/R295 of the recognition helix interact with the TΨC loop of the substrate tRNA. These residues form hydrogen bonds with the phosphate group of C56 as well as the previously mentioned hydrogen bond with the pyrimidine ring of C56. The second set, consisting of the G10/R272 pair, binds the centre of the D loop, which directly interacts with the phosphate group of U20 and makes further contacts with G19 and U20 through water-mediated hydrogen bonds. In the final set, the 5' end of the D loop is bound via hydrogen bonding interactions between the R35/S106/G107 triplet, which contact the phosphate of G15 and the O3′ ribose atom of G14. It is important to note that in this group, however, S106 and G107 are not fully conserved throughout the DusC subfamily.

In conjunction with these putative hotspots, the relative displacement of the recognition domains of the two proteins also acts as a determinant of specificity. Compared to the *Tt*Dus enzyme, the recognition domain of DusC is raised by ~10Å (**Fig. 5.12h**), closing the cylindrical groove between the two domains, and so preventing the binding of the substrate tRNA in the *Tt*Dus orientation. Additionally, small deletions and insertions in the Dus sequences, notably insertion of residues 168-175 (*Tt*Dus numbering) in position 20-specific enzymes, could further contribute to defining the position specificity of Dus enzymes. The determinants of specificity in Dus enzymes are therefore amino acid hotspots on the surface of the enzyme, and variation in relative displacement of the recognition domain, which together define the orientation of bound tRNA and hence which uracil is positioned to access the active site for modification.

5.4.2 Alternate tRNA sequences can be accommodated by induced fit

The crystal structure also clarifies how DusC is able to modify U16 in tRNAs that lack a strong consensus sequence in the regions contacted by the enzyme. First, plasticity in the active site, owing to flexibility of the active site loop, could serve to accommodate variations in the D loop lengths, sequences and overall tertiary structures of tRNA substrates [185, 213]. Multiple non-sequence specific interactions, involving hydrogen bonds to sugar-phosphate backbone, serve to stabilise complex formation. Consequently, it is expected that DusC modifies a wide range of tRNA substrates, in agreement with the observation that mature *E. coli* tRNA sequences containing dihydrouridine at position 16 lack a strong consensus sequence, especially in the regions of the D and TΨC stem-loops contacted by DusC [185]. Second, this binding mode is insensitive to the length of the variable loop, which is exposed and does not interact with the enzyme. This would therefore allow DusC to modify tRNAs with elongated variable loops, such as tRNA^{Leu} in *E. coli* [185]. Third, DusC selects for fully-folded tRNA through both the interactions described above, and by specific recognition of C56 in the elbow region. Although previous studies demonstrated that Dus enzymes have significantly lower activity against unmodified tRNA than partially modified tRNA [142, 187], the structures of DusC and *Tt*Dus do not implicate specific recognition of any residues typically modified in mature tRNA. Instead,the preference for partly modified tRNAs is more likely due to the fact that these modifications stabilise the L-shaped conformation of tRNA, and hence may increase the affinity of enzyme binding [214]. It is interesting to note that specific recognition of the G19-C56 base pair has also been reported for TruA, which is only able to act on a fully-folded tRNA and consequently acts as a ‘tertiary structure checkpoint’ [198]. As a result, the final stages of tRNA maturation are delayed until the tRNA has adopted the L-shaped conformation.

5.4.3 Evolution of Dus Specificity

Dihydrouridine synthases that modify uridines at different positions belong to paralogous groups of sequences that appeared as a result of gene duplication [144]. It is now clear that the same protein fold can provide the basis for a duplicated gene to become position 16- or 20-specific, without addition of auxiliary domains [138, 145]. Specificity is generated by the introduction of three ‘hot spots’ containing polar and charged residues that define docking of tRNA in a specific orientation. Binding is assisted by positional adjustments of the C-terminal recognition domain. Additional variations in the length of exposed protein loops serve to further increase affinity towards particular orientations of substrate tRNAs.

It is possible that an ancestral dihydrouridine synthase was able to bind tRNA in several alternative orientations. Differentiation of Dus subfamilies may have proceeded by mutations at putative “hotspots” which favored docking of tRNA in a specific orientation, facilitating modification of evolving sequences of tRNA that derived evolutionary benefit from modification at specific position(s).

This evolutionary pathway would give rise to the observations here – that in addition to the active site residues necessary for catalysis, the DusA and DusC subfamilies contain specifically conserved charged residues that determine the orientation of the tRNA on the enzyme surface, and hence constrain substrate specificity to the uracil that is positioned for insertion into the active site.

5.4.4 Comparisons with other tRNA modification enzymes

Other tRNA modification enzymes use different strategies for generating nucleotide specificity than DusC by including extra structural features. TruB [195], TruD [197], and ArcTGT [190] contain extra domains which confer position specificity whereas in TruA, specificity is generated by the formation of a dimeric state [198].

Structural data on position 16- (these results) and position 20-specific [187] Dus subfamilies reveal a different mechanism by which positional specificity can be conferred. In these enzymes, clusters of residues conserved in each family (‘hotspots’) guide docking of the substrate tRNA in a specific (and completely different) orientation. Interaction with the tRNA is facilitated by positional adjustments of the C-terminal ‘recognition’ domain. In the case of DusC, tRNA recognition is further assisted by induced fit adjustments in both the conformation of the D loop of tRNA and an opposing flexible segment (residues 100-108) in DusC. The relative simplicity of the Dus mechanism therefore represents a novel addition to the recognition strategies used by tRNA modifying enzymes.

While the currently identified Dus enzymes of archaea and bacteria consist of a ‘core region’, comprising just the TIM-barrel and recognition domains, a number of eukaryotic Dus enzymes have additional RNA domains linked to the core. In humans, for example, putative Zn-fingers are present in the C-terminus of Dus1 and the N-terminus of Dus3, while Dus2 contains a C-terminal double stranded RNA binding motif [144]. The tRNA recognition roles played by these domains have not been investigated to date, yet their presence is intriguing, given that adaptation of the core region appears to be sufficient in archaea and bacteria to target all positions of dihydrouridine incorporation within the D loop. It is therefore possible that eukaryotic Dus enzymes determine modification position specificity through a combination of the mechanism presented here, and the mechanisms used by other nucleotide modification enzymes which involve auxiliary RNA binding domains.

Chapter 6 : Discussion

The primary focus of this study was to examine the strategies used by RNA binding proteins for generating specificity towards their target sequences, and how the specific and non-specific recognition of RNA by protein influences their function. In this regard, the RNA binding proteins Lin28 and DusC were investigated. In both cases, the specificity of the protein:RNA interactions was known to be important for biological function, but relatively little was known about how specificity was generated in each case. To elucidate the mechanisms used by these proteins to recognize their targets, both structural biology and biochemical techniques were employed.

Being composed of ribonucleotides and having a negative charge, an RNA oligonucleotide is capable of making multiple specific and non-specific interactions with protein molecules. Electrostatic contacts from charged amino acid side chains in the protein can help to guide the phosphate backbone around the surface of the protein by either attraction or repulsion. Hydrogen bonding contacts are often crucial in complex formation, and could be either specific or non-specific depending on whether they are formed with RNA bases or sugar-phosphate backbone. Hydrophobic stacking interactions between the RNA bases and the aromatic side chains of several amino acids can also form, which can again be non-specific, or semi-specific if the protein utilizes specifically sized hydrophobic binding pockets which for example, might only accommodate pyrimidines.

6.1.1 Specific and non-specific interactions are used in the Lin28 and DusC mechanisms of RNA binding.

Both the Lin28 and DusC proteins are likely to use a combination of both specific and non-specific interactions at different points in their binding mechanisms. Within the cell, Lin28 must recognize a multitude of different RNA targets of varying structure and sequence, united only by the small GGAG motif. As demonstrated in *Chapter 4,* Lin28 first employs non-specific electrostatic contacts in order to sample the available nucleic acid sequences until it finds one containing the GGAG motif, which is recognized by specific hydrogen bonding interactions between the RNA and the ZnK domain of the protein. The ZnK domain then acts as an anchor and facilitates the binding of the CSD through a mixture of non-specific stacking interactions, and base specific hydrogen bonds. By utilizing such a mechanism, it can be speculated that Lin28 is able to increase its efficiency in recognizing RNA target sequences.

For DusC, one specific uracil must be recognized at the same position on all tRNA molecules that contain it, regardless of differences in sequence and D-loop lenth. It is not surprising then, that, as shown in *Chapter 5*, the majority of the contacts formed between the protein and RNA are hydrogen bonding interactions between protein side chains and the phosphate or ribose moieties of the RNA. In this system, tRNAs must again be first recognized non-specifically by electrostatic forces in order to dock the tRNA in specific orientation at the positively charged surface. Following this, U16 can be positioned in the active site using a specific hydrogen bonding network. In addition, the flexibility of the active site loop allows tRNAs to bind regardless of the different D-loop lengths. It can therefore be seen, once again, that the protein (DusC) employs a mixture of non-specific and sequence specific interactions to fulfil its biological role.

6.1.2 Lin28 and DusC use modularity to create an extended RNA binding surface

In *Chapter 1*, the idea that RNA binding proteins exploit modularity in order to create their RNA binding surfaces was discussed. This strategy can be seen in the way both the Lin28 and DusC proteins bind their targets.

The two domains of Lin28 are connected by a flexible linker region, and it has been speculated that this could be important for the recognition of targets of varying lengths [122]. Both domains are variants of classical RNA binding domains, but arranged together in a unique combination. These domains recognize target miRNA species through two different motifs of the miRNA; a structural motif in the case of the CSD/loop interaction, and a sequence motif in the case of the ZnK/GGAG interaction. By retaining flexibility between the two domains, different RNA binding surfaces can be created that can recognize RNAs with different sizes and shape. This is likely to facilitate Lin28's biological function by allowing it to recognize a wide variety of different RNAs.

DusC is not composed of classical RNA binding domains, and there is no evidence to suggest the linker between the N-terminal catalytic domain and the C-terminal recognition domain is flexible. However, the tethering of these two domains defines a distinctive surface. The shape of this surface is different depending on the relative displacement of the domains, as can be seen from the comparison of DusC and *Tt*Dus structures (**Fig. 5.12h**). The evolution of the divergent positively charged hotspots then defines specificity for binding tRNA in a particular orientation. As the mechanisms of both the DusC and *Tt*Dus are thought to be the same, it is the subtle changes in the structure of the RNA binding proteins that confers the completely different specificities of the two enzymes. Thus, by altering their surface properties (but not folds), these RNA binding proteins are able to bind the same tRNA substrate in completely different orientations, resulting in alternate specificities.

In conclusion, the structural and biophysical data presented here show that the control of substrate specificity is determined by the structural properties of proteins, which utilize the chemical properties of RNA in order to alternately recognize RNA through specific and non-specific means. This careful control of specificity towards RNA targets allows both Lin28 and DusC to fulfil their biological functions efficiently.

6.1.3 Future work: Lin28

Further structural work is needed to fully understand Lin28's interaction with nucleic acids. The structure of a Lin28/mir363 complex would be informative as it would demonstrate whether Lin28 binds this RNA in the same manner as let-7 miRNAs. In addition, it would be useful to obtain structures of Lin28 in complex with slightly longer RNA sequences and longer constructs of Lin28 in order to visualize a more complete structure. However, such structural information could be challenging to produce due to the flexibility of both the RNA and the protein, especially if the CSD/RNA interaction is transient.

Following the results presented here, it would be informative to perform single molecule experiments to investigate how Lin28 interacts with RNA in time. Recent studies of the bacterial EcoRV restriction enzyme and human oxoguanine DNA glycosylase 1 (hOgg1) have used total internal reflection fluorescence microscopy (TIRFM) to analyse systems where a DNA binding protein interacts with an oligonucleotide both specifically and non-specifically [215, 216]. This technique involves illuminating a sample with incident light at such an angle as to achieve total internal reflection. This generates a small evanescent field that can illuminate fluorophores in a very small area, allowing single molecules to be visualized [217]. In both of the above reports, a length of DNA was immobilized, and single protein molecules were observed translocating across the DNA by "sliding" and "jumping" mechanisms. These mechanisms occur through non-specific interactions between the protein and DNA as the protein searches for a specific binding site. In each case, the experiments were repeated at different salt concentrations [215, 216]. For hOgg1, increasing the salt concentration did not greatly alter the diffusion coefficient, implying a sliding mechanism [215]. In contrast, EcoRV uses a combination of sliding and jumping mechanisms. At higher salt concentrations the amount of time spent associated with the DNA decreased, but the distribution of jump distances remained the same. These data imply that, for this protein, the jumping mechanism is favoured over the sliding mechanism in higher ionic strength solutions [216].

Similar approaches could be used for measuring non-specific, salt dependent interaction of Lin28 with RNA. In *Chapter 4*, it was discussed that Lin28 may use it's CSD to scan the transcriptome for binding sites, and use the ZnK domain to bind to specific RNA sequences. A long RNA sequence (mRNA or pre-miRNA) could be immobilized on a slide. Lin28 with a conjugated fluorophore could then be added and examined by TIRFM. The amount of time that Lin28 spends interacting with RNA sequences that either contain or do not contain the GGAG sequence could then be measured. These data could be obtained at different ionic strengths. The results would reveal how multiple single molecules of Lin28 interact with long RNA sequences, the kinetics of the association, and whether Lin28 can use "jumping" or "sliding" mechanisms to scan RNA target sequences.

6.1.4 Further work: Dihydrouridine synthases

The mechanism used by DusC to modify uridine at position-16 of tRNA is now known. However, data obtained previously in YSBL by Dr. Rob Byrne and Dr. Fiona Whelan did not identify whether position 17 could also be modified by DusC. It would therefore be useful to obtain two mutated variants of tRNA^{trp}, which was shown to be bound by DusC. This tRNA has U at position 16 and 17. tRNA^{trp} with U16/C17 and C16/U17 would allow biochemical investigation into whether DusC could modify both positions. If DusC can modify position 17, then the corresponding protein-tRNA complex could be structurally characterised, possibly using the same crystallization conditions as described in *Chapter 5* for the DusC^{C98A}/tRNA^{phe} complex.

It is now known that DusA modifies position 20, and that DusC modifies position 16 (and possibly also position 17). The position specificity of DusB, however, remains unknown. It would be useful to perform electromobility shift assays, followed by the reverse transcriptase assay, mentioned in *Chapter 5* (**5.1.3**). This would inform which tRNAs could be bound by DusB and which positions are modified by it. Structural investigations could then begin. The comparison of tRNA-bound DusC, DusB and *Tt*Dus (homologous to DusA), would provide a complete picture of how position specificity is generated by bacterial Dus enzymes.

Finally, a structural investigation of eukaryotic Dus enzymes would be highly informative. The hotspots described in *Chapter 5* are not conserved in the eukaryotic enzymes, and in addition they contain extra domains. The mechanism for generating position specificity in bacterial enzymes is therefore unlikely to be conserved in the eukaryotic enzymes. The human hDus2 enzyme has been implicated in non-small cell lung cancer. Investigation of its structure and functional mechanism would therefore be of medical importance.

6.1.5 Conclusion

The study of protein/RNA interactions is challenging. Proteins can interact both specifically and non-specifically with nucleic acids and many modular proteins composed of multiple domains are intrinsically flexible. This makes them capable of adjusting or even significantly changing the positions of their domains in response to the environment and/or interaction with nucleic acids. In addition, RNAs are flexible and often unstable, and predicting their secondary and tertiary structures could be a non-trivial task. The factors can make producing appropriate binding models for fitting biophysical data difficult. Furthermore, production of recombinant protein constructs composed of multiple domains can also be complicated, slowing down progress in obtaining diffracting crystals for structural analysis. Future work will likely take advantage of modern techniques that are actively being developed, such as single molecule methods, and improvements in electron microscopy. These approaches allow visualization of single molecules and are therefore very informative and useful for the development of more accurate models. Electron microscopy is not limited to fully ordered species, and so as technology improves the resolution of this technique, it could become a more appropriate tool for the structural investigation of these interactions. What is clear, however, is that it will be the combination of both structural and functional studies that will help to elucidate protein/RNA interactions and the associated mechanisms that underpin many fundamental processes in biology.

Appendix: Non-linear regression fits and statistics

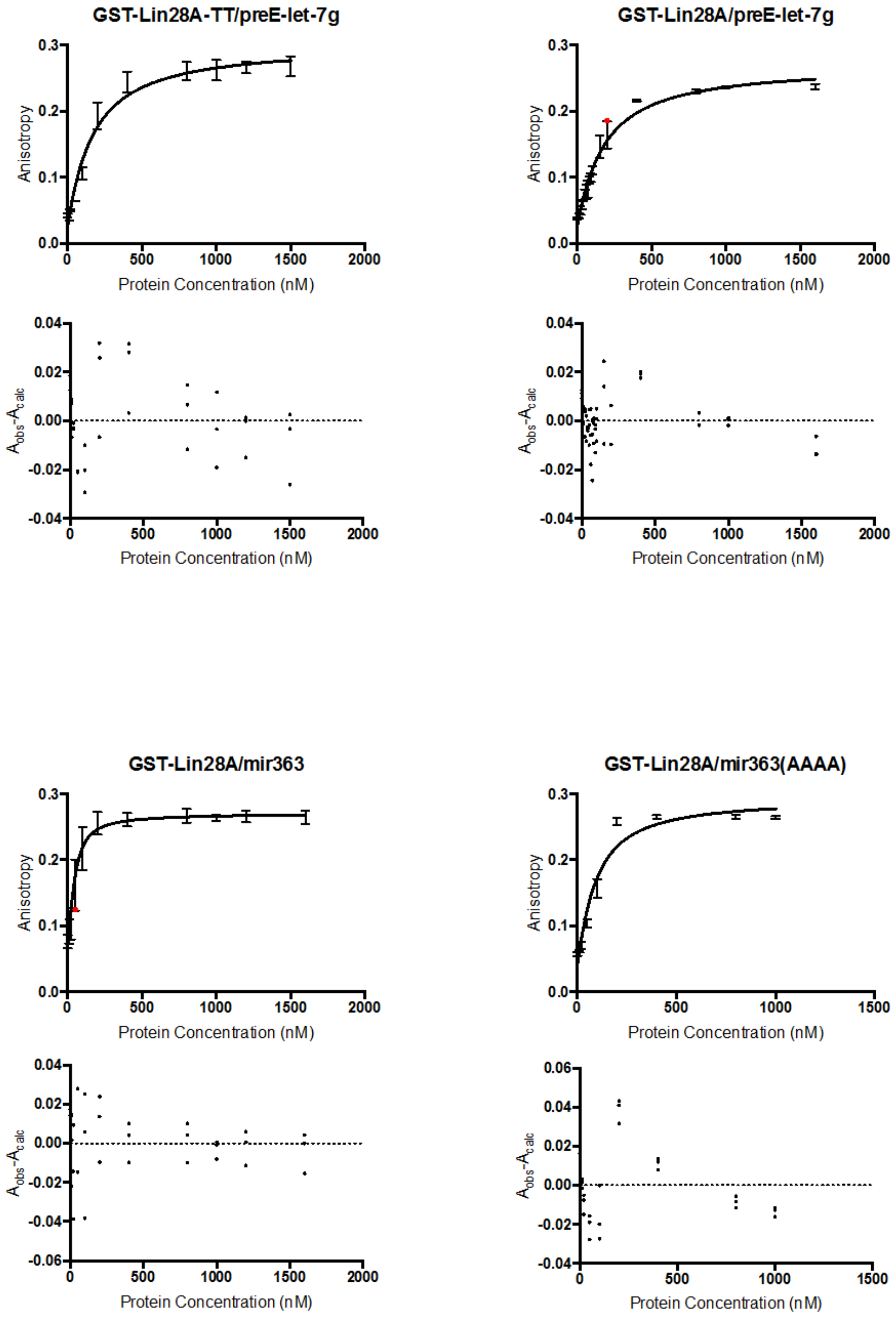


Figure A1: Analysis of interactions between GST tagged proteins and RNA by fluorescence anisotropy.

Changes in anisotropy when GST-Lin28A-TT and GST-Lin28A are added to preE-let7g, mir363 and mir363(AAAA) RNAs are shown, fit by *Equation 14*. Residuals from each fit are shown beneath each curve. Outlying points excluded from data fitting are shown in red.

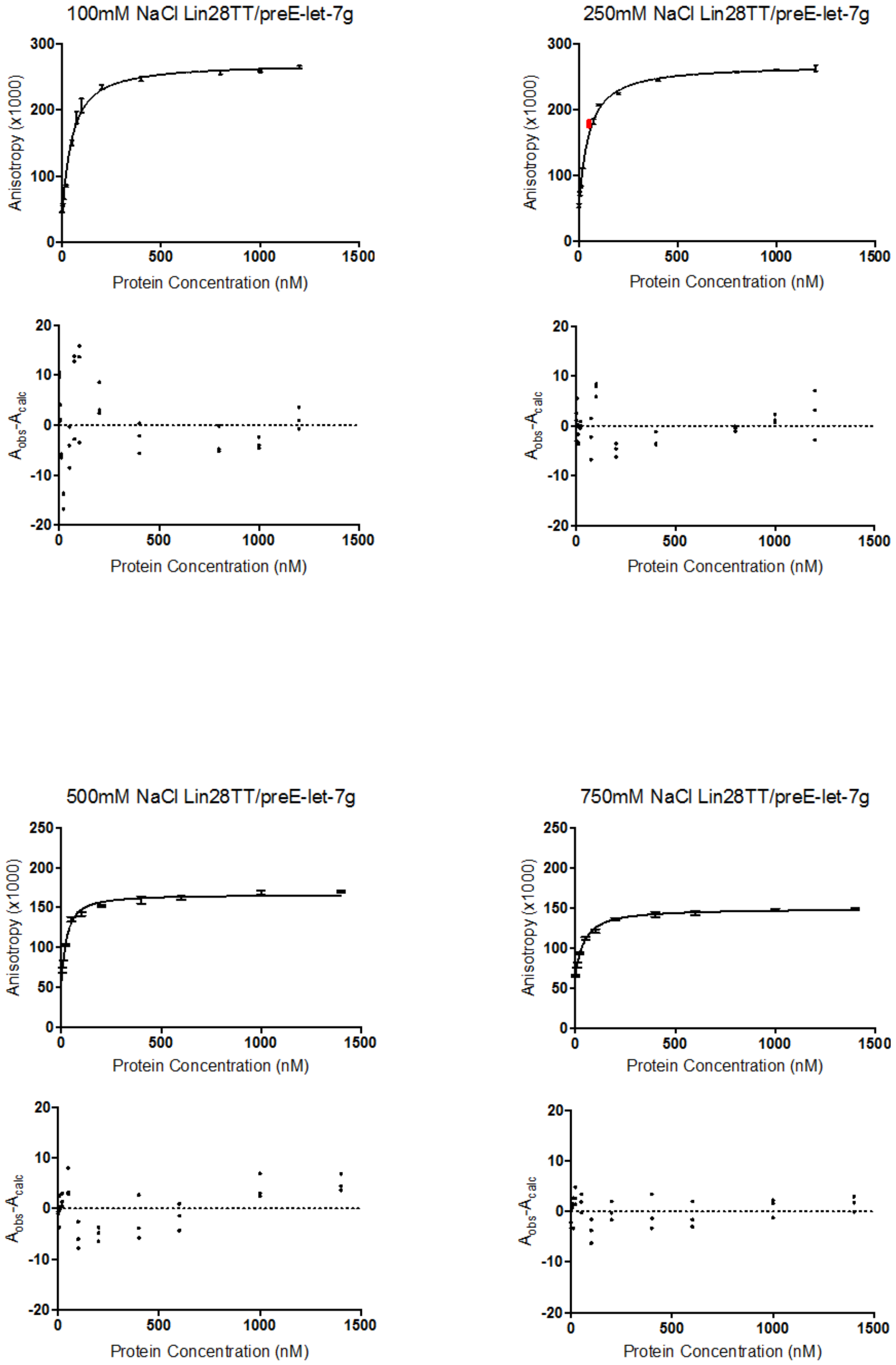


Figure A2: Analysis of the Lin28TT/preE-let-7g interaction at different ionic strengths by fluorescence anisotropy.

Changes in anisotropy when Lin28TT is added to preE-let-7g RNA in increasing ionic strength conditions are shown fit by *Equation 14*. Residuals from each fit are shown beneath each curve. Outlying points excluded from data fitting are shown in red.

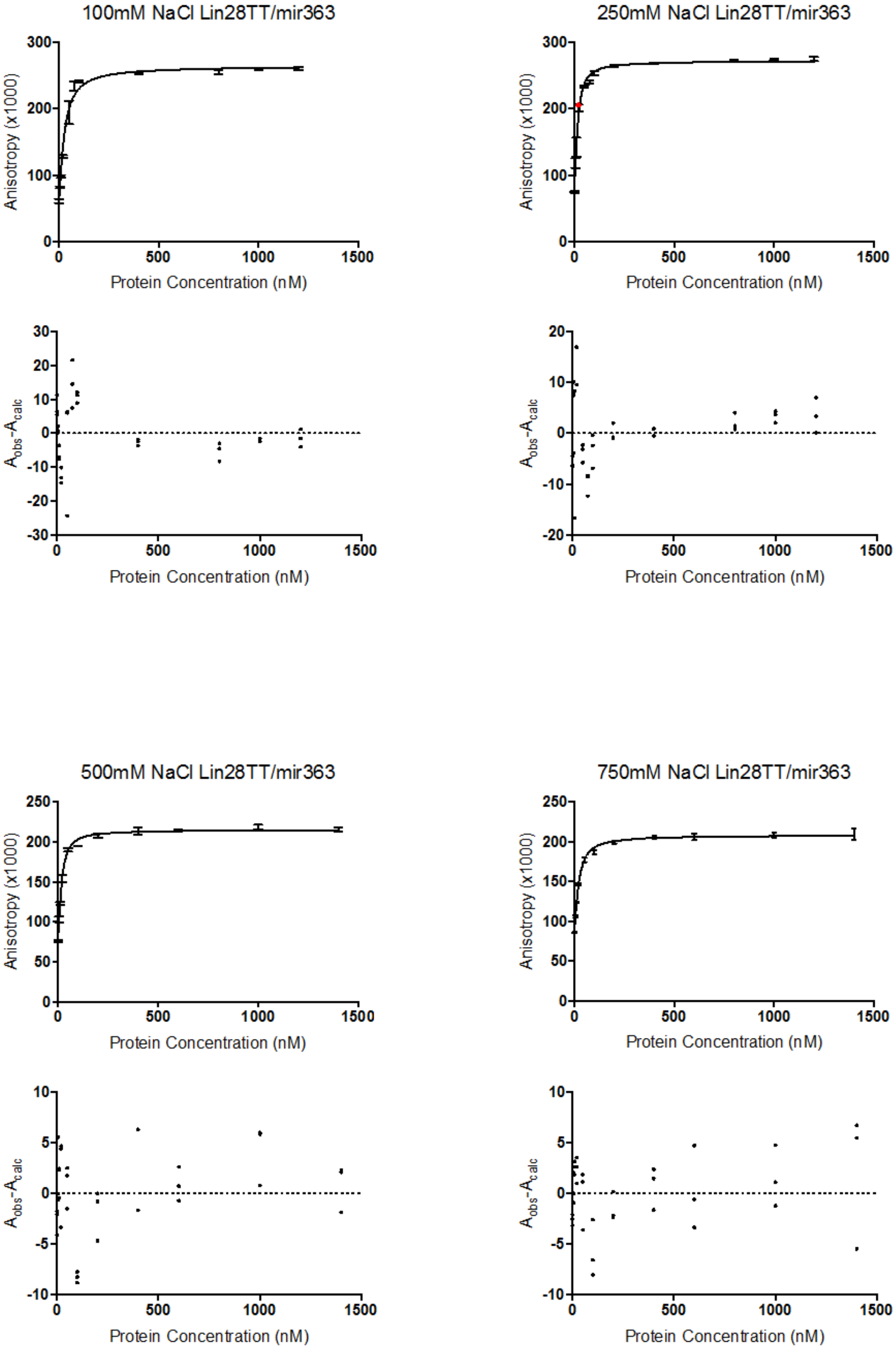


Figure A3: Analysis of the Lin28TT/mir363 interaction at different ionic strengths by fluorescence anisotropy.

Changes in anisotropy when Lin28TT is added to mir363 RNA in increasing ionic strength conditions are shown fit by *Equation 14*. Residuals from each fit are shown beneath each curve. Outlying points excluded from data fitting are shown in red.

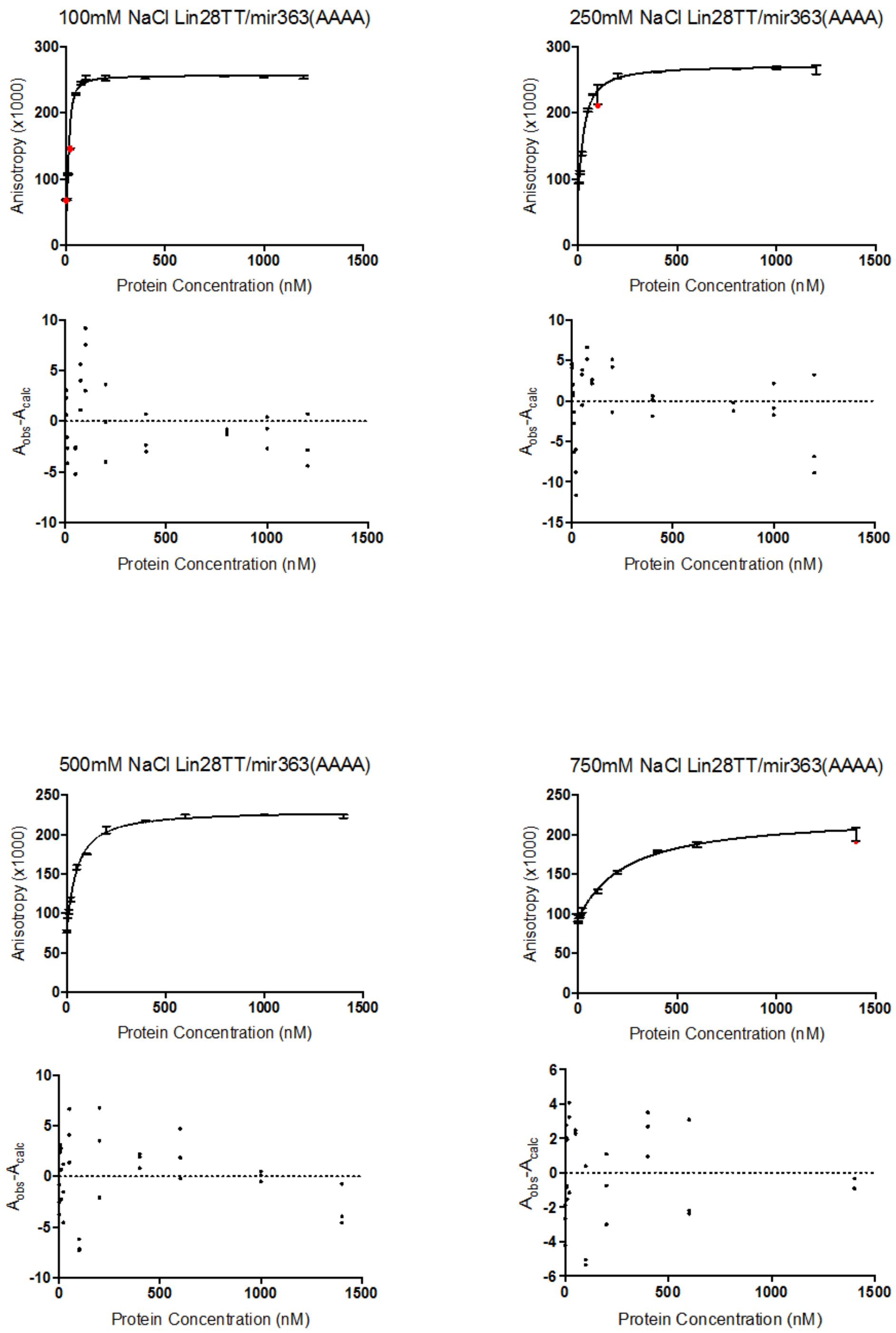


Figure A4: Analysis of the Lin28TT/mir363(AAAA) interaction at different ionic strengths by fluorescence anisotropy.

Changes in anisotropy when Lin28TT is added to mir363(AAAA) RNA in increasing ionic strength conditions are shown fit by *Equation 14*. Residuals from each fit are shown beneath each curve. Outlying points excluded from data fitting are shown in red.

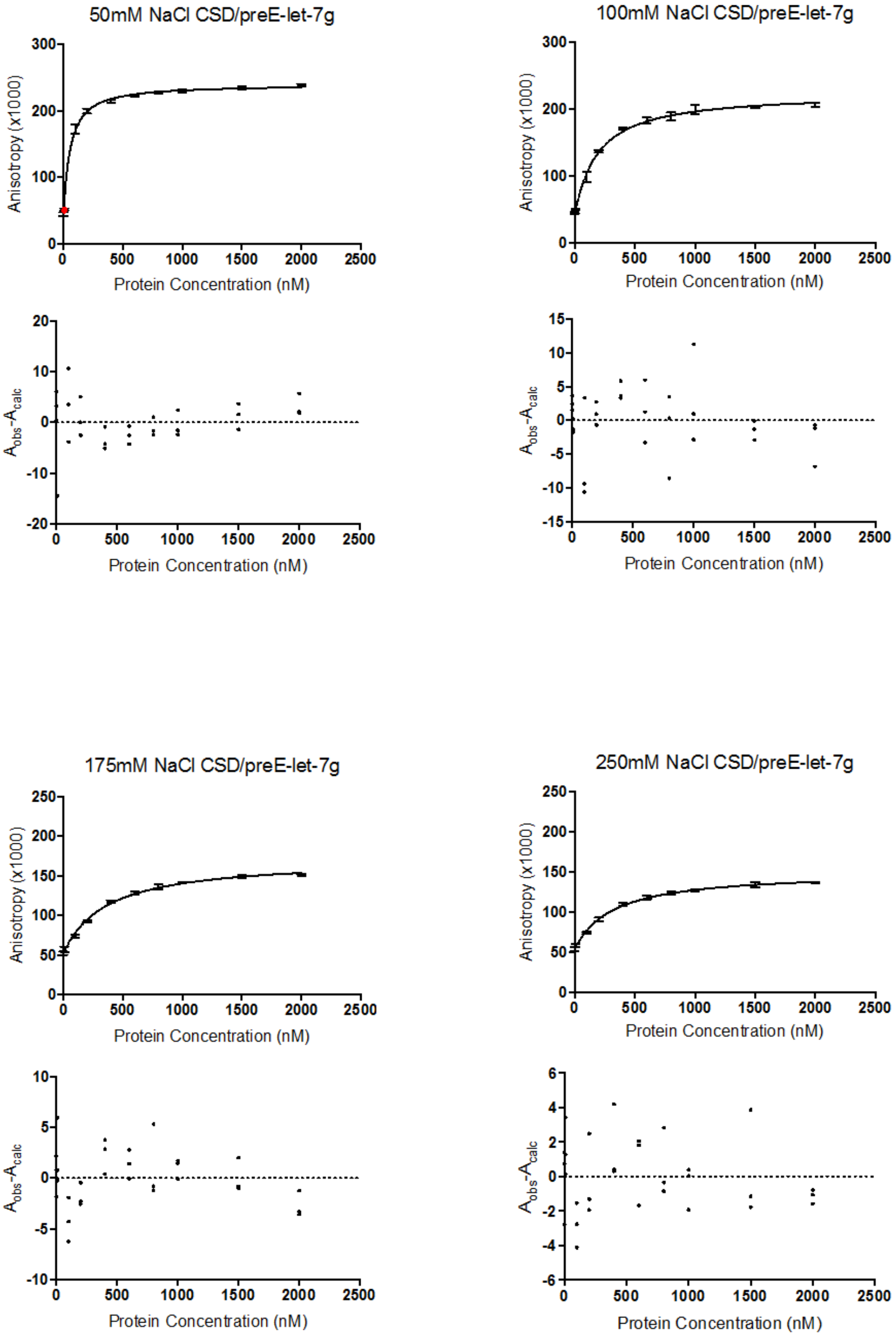


Figure A5: Analysis of the CSD/preE-let-7g interaction at different ionic strengths by fluorescence anisotropy.

Changes in anisotropy when CSD is added to preE-let-7g RNA in increasing ionic strength conditions are shown fit by *Equation 14*. Residuals from each fit are shown beneath each curve. Outlying points excluded from data fitting are shown in red.

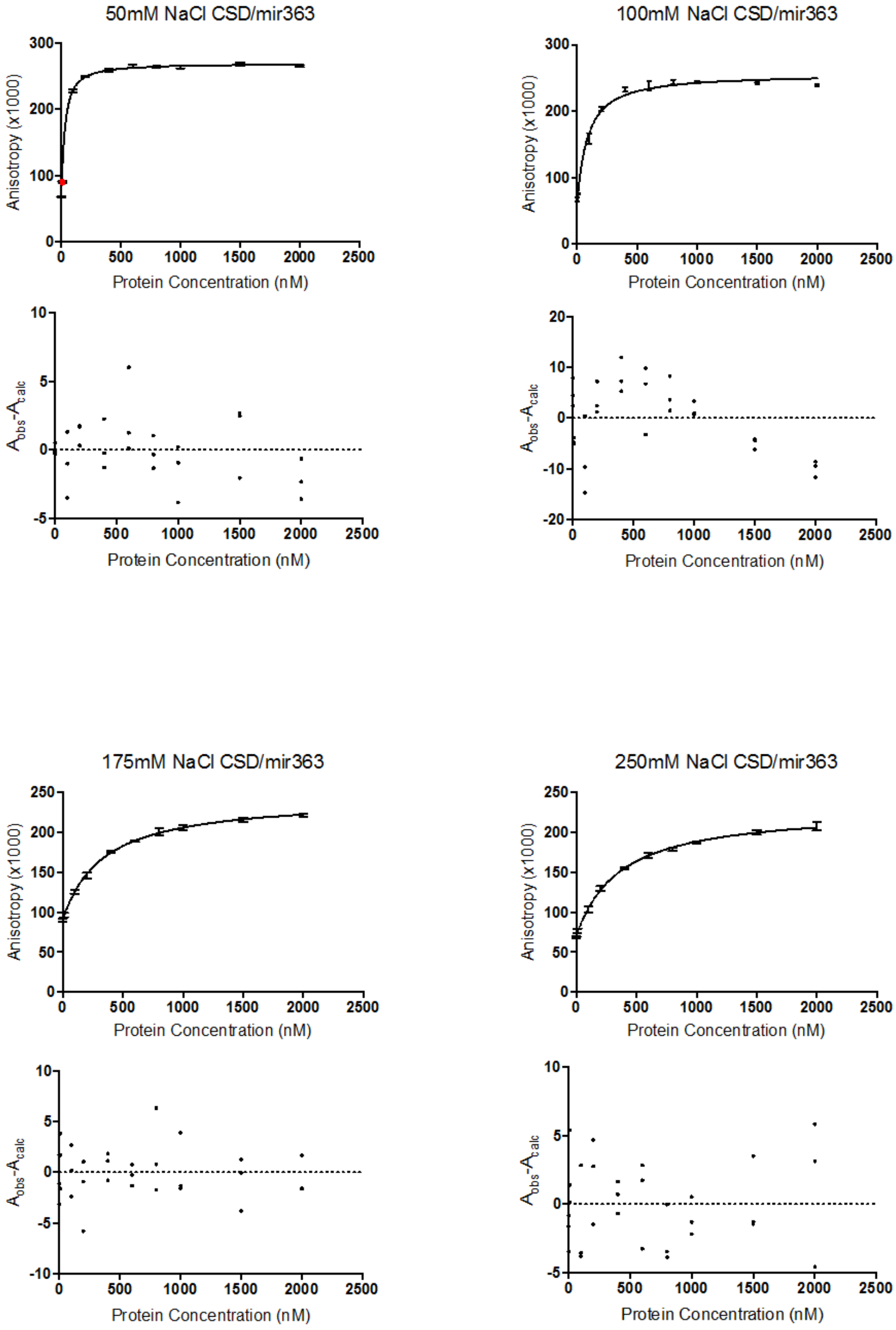


Figure A6: Analysis of the CSD/mir363 interaction at different ionic strengths by fluorescence anisotropy.

Changes in anisotropy when CSD is added to mir363 RNA in increasing ionic strength conditions are shown fit by *Equation 14*. Residuals from each fit are shown beneath each curve. Outlying points excluded from data fitting are shown in red.

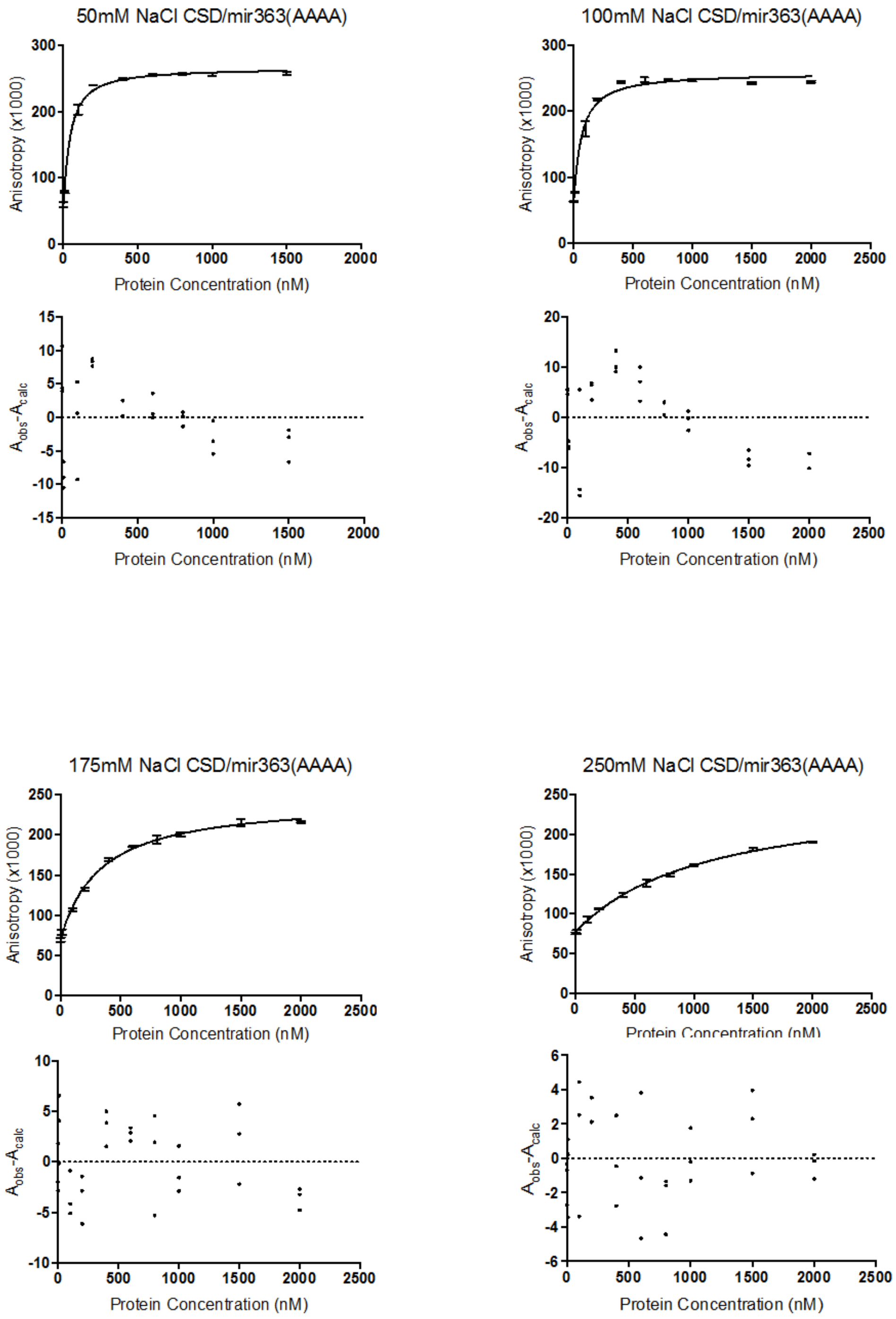


Figure A7: Analysis of the CSD/mir363(AAAA) interaction at different ionic strengths by fluorescence anisotropy.

Changes in anisotropy when CSD is added to mir363(AAAA) RNA in increasing ionic strength conditions are shown fit by *Equation 14*. Residuals from each fit are shown beneath each curve.

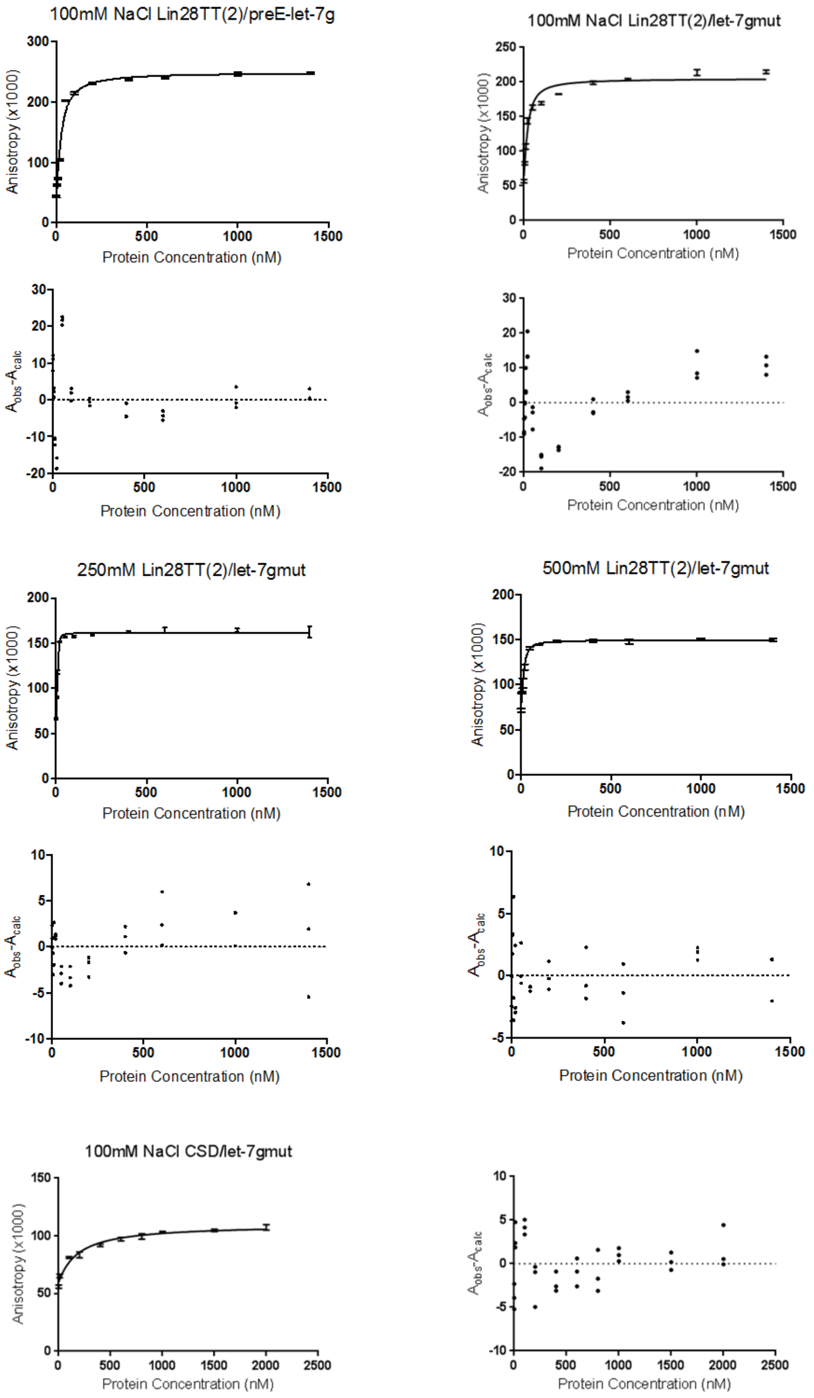


Figure A8: Analysis of the interactions of preE-let-7g amd let-7gmut RNA with Lin28TT(2) and CSD at different ionic strengths by fluorescence anisotropy.

The change in anisotropy when Lin28TT(2) is added to preE-let-7g is shown, followed by the changes in anisotropy when added to let-7gmut in increasing ionic strength conditions, with data fit by *Equation 14*. Finally, the change in anisotropy when CSD is added to let-7gmut RNA is shown, with data fit by E*quation 14*. Residuals from each fit are shown beneath each curve, except for the CSD/let-7gmut interaction, where they are shown to the side.

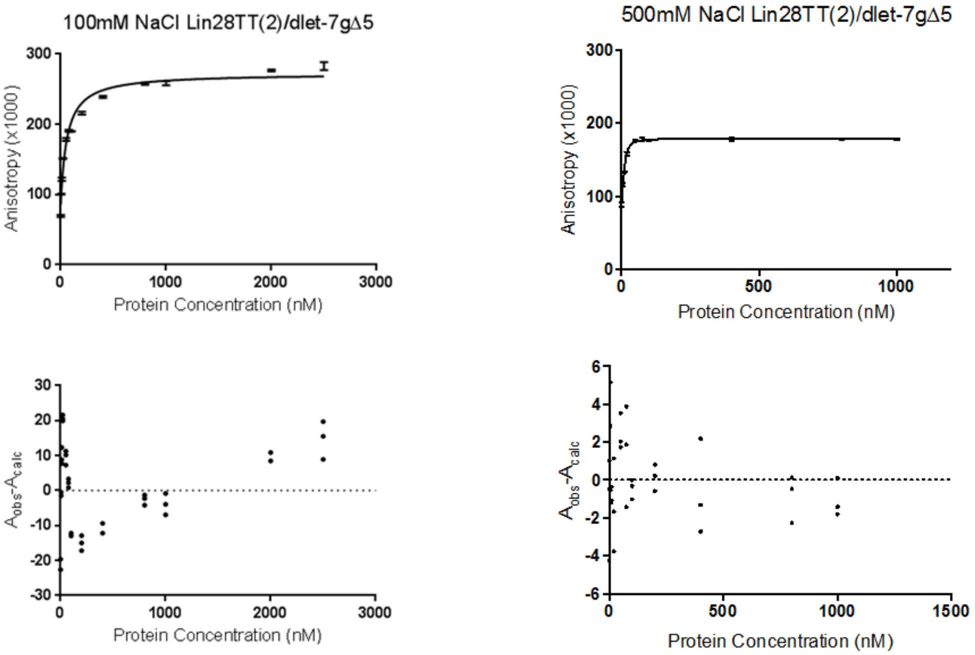


Figure A9: Analysis of the Lin28TT(2)/dlet-7gΔ5 interaction at different ionic strengths by fluorescence anisotropy.

Changes in anisotropy when Lin28TT(2) is added to dlet-7gΔ5 DNA in conditions containing 100mM and 500mM NaCl are shown, with data fit by *Equation 14*. Residuals from each fit are shown beneath each curve.

Protein	RNA	Buffer Salt Concentration	Equation	Amin	Std. error	Dependancy	Amax	Std. error	Dependancy	c	Std. error	Dependancy	K_d	Std. error	Dependancy
GST-Lin28A-TT	preE-let-7g	100mM NaCl	14	0.027	0.006	0.422	0.302	0.009	0.71	50.0	-	-	145.8	25.2	0.79
GST-Lin28A	preE-let-7g	100mM NaCl	14	0.027	0.003	0.638	0.271	0.006	0.71	50.0	-	-	156.8	15.2	0.84
	mir363	100mM NaCl	14	0.066	0.007	0.266	0.271	0.005	0.43	50.0	-	-	21.4	5.7	0.55
	mir363(AAAA)	100mM NaCl	14	0.040	0.008	0.394	0.295	0.011	0.66	50.0	-	-	68.7	16.1	0.75

Protein	RNA	Buffer Salt Concentration	Equation	Adj R2	Sum of Squares	Normality Test P-value D'Agostino-Pearson	Normality Test P-value Shapiro-Wilks	Number of Points used for fit	Outliers
GST-Lin28A-TT	preE-let-7g	100mM NaCl	14	0.9693	0.008949	0.5667	0.4761	33	0
GST-Lin28A	preE-let-7g	100mM NaCl	14	0.9803	0.004639	0.4913	0.6624	50	1
	mir363	100mM NaCl	14	0.9548	0.007914	0.2399	0.2842	32	1
	mir363(AAAA)	100mM NaCl	14	0.954	0.009403	0.2169	0.1206	27	0

Table A1: Non-linear regression fit values and statistics for GST fusion proteins.

Normality tests for residuals are passed if P > 0.05.

Protein	RNA	Buffer Salt Concentration	Equation	Amin	Std. error	Dependancy	Amax	Std. error	Dependancy	c	Std. error	Dependancy	K_d	Std. error	Dependancy
Lin28TT	let-7g	100mM	14	35.1	3.2	0.41	271.6	2.9	0.52	20.0	-	-	38.8	3.0	0.67
		250mM	14	54.1	1.6	0.37	268.4	1.5	0.52	20.0	-	-	40.6	1.9	0.66
		500mM	14	58.2	1.8	0.37	166.6	1.3	0.43	20.0	-	-	18.2	1.9	0.60
		750mM	14	66.8	1.1	0.40	149.2	1.0	0.49	20.0	-	-	35.5	2.9	0.65
	mir363	100mM	14	53.6	4.1	0.36	264.8	3.0	0.44	20.0	-	-	16.6	1.9	0.60
		250mM	14	80.5	3.1	0.28	272.2	2.0	0.42	20.0	-	-	7.5	0.9	0.55
		500mM	14	78.3	1.8	0.32	215.2	1.1	0.37	20.0	-	-	8.0	0.8	0.53
		750mM	14	88.6	1.6	0.35	208.2	1.1	0.40	20.0	-	-	12.4	1.1	0.57
	mir363(AAAA)	100mM	14	58.4	-	-	262.5	3.12	0.42	20.0	-	-	10.13	1.358	0.42
		250mM	14	68.6	2.0	0.37	272.8	1.5	0.46	20.0	-	-	19.7	1.1	0.62
		500mM	14	79.6	1.5	0.40	230.6	1.4	0.51	20.0	-	-	40.8	2.5	0.66
		750mM	14	92.2	0.9	0.41	224.2	2.8	0.81	20.0	-	-	221.9	14.8	0.85

Protein	RNA	Buffer Salt Concentration	Equation	Adj R2	Sum of Squares	Normality Test P-value D'Agostino-Pearson	Normality Test P-value Shapiro-Wilks	Number of Points used for fit	Outliers
Lin28TT	preE-let-7g	100mM	14	0.9906	2225	0.87	0.27	36	0
		250mM	14	0.9975	475	0.38	0.20	33	3
		500mM	14	0.9893	520	0.79	0.67	33	0
		750mM	14	0.9923	220	0.51	0.27	33	0
	mir363	100mM	14	0.9847	2725	0.55	0.82	33	0
		250mM	14	0.9894	1584	0.70	0.99	35	1
		500mM	14	0.9933	497	0.48	0.20	33	0
		750mM	14	0.9934	384	0.79	0.92	33	0
	mir363(AAAA)	100mM	14	0.9792	4178	0.50	0.09	36	0
		250mM	14	0.9960	695	0.09	0.02	35	1
		500mM	14	0.9955	426	0.75	0.69	33	0
		750mM	14	0.9951	200	0.31	0.18	29	1

Table A2: Non-linear regression fit values and statistics for Lin28TT interactions.

Normality tests for residuals are passed if P > 0.05.

Protein	RNA	Buffer Salt Concentration	Equation	Amin	Std. error	Dependancy	Amax	Std. error	Dependancy	c	Std. error	Dependancy	K_d	Std. error	Dependancy
CSD	let-7g	50mM	14	41.1	2.6	0.19	240.0	1.6	0.53	20.0	-	-	48.0	3.4	0.59
		100mM	14	42.5	2.1	0.37	225.0	2.6	0.73	20.0	-	-	181.5	12.3	0.79
		175mM	14	52.2	1.2	0.45	171.5	2.2	0.83	20.0	-	-	348.3	23.7	0.87
		250mM	14	54.1	0.9	0.43	150.8	1.6	0.81	20.0	-	-	307.0	19.1	0.86
	mir363	50mM	14	68.1	1.3	0.05	269.8	0.7	0.47	20.0	-	-	21.5	1.0	0.49
		100mM	14	61.6	3.3	0.27	256.4	2.7	0.59	20.0	-	-	76.2	7.3	0.66
		175mM	14	90.7	1.1	0.44	243.4	1.9	0.82	20.0	-	-	321.2	15.2	0.86
		250mM	14	70.2	1.3	0.44	229.5	2.4	0.83	20.0	-	-	343.9	18.9	0.87
	mir363(AAAA)	50mM	14	52.8	2.9	0.24	266.5	2.1	0.53	20.0	-	-	35.6	3.2	0.61
		100mM	14	58.0	3.7	0.25	258.2	2.7	0.54	20.0	-	-	54.7	6.1	0.62
		175mM	14	70.2	1.6	0.44	244.0	2.8	0.82	20.0	-	-	318.9	19.4	0.86
		250mM	14	76.6	1.0	0.55	253.6	5.5	0.95	20.0	-	-	1082.8	75.3	0.96

Protein	RNA	Buffer Salt Concentration	Equation	Adj R2	Sum of Squares	Normality Test P-value D'Agostino-Pearson	Normality Test P-value Shapiro-Wilks	Number of Points used for fit	Outliers
CSD	preE-let-7g	50mM	14	0.9943	582	0.03	0.14	28	2
		100mM	14	0.9936	634	0.36	0.23	30	0
		175mM	14	0.9940	218	0.86	1.00	30	0
		250mM	14	0.9949	127	0.63	0.59	30	0
	mir363	50mM	14	0.9987	121	0.24	0.46	27	3
		100mM	14	0.9889	1375	0.49	0.57	30	0
		175mM	14	0.9971	178	0.47	0.78	30	0
		250mM	14	0.9961	254	0.31	0.25	30	0
	mir363(AAAA)	50mM	14	0.9942	823	0.88	0.67	27	0
		100mM	14	0.9871	1711	0.31	0.25	30	0
		175mM	14	0.9951	386	0.07	0.22	30	0
		250mM	14	0.9957	184	0.53	0.49	30	0

Table A3: Non-linear regression fit values and statistics for CSD protein interactions.

Normality tests for residuals are passed if P > 0.05.

Protein	RNA	Buffer Salt Concentration	Equation	Amin	Std. error	Dependancy	Amax	Std. error	Dependancy	c	Std. error	Dependancy	K_d	Std. error	Dependancy
Lin28TT(2)	preE-let-7g	100mM NaCl	14	33.2	4.5	0.37	249.6	3.2	0.42	20.0	-	-	16.9	2.2	0.59
	let-7gmut	100mM NaCl	14	64.0	4.6	0.35	205.3	3.1	0.40	20.0	-	-	13.1	2.8	0.57
		250mM NaCl	14	65.0	1.6	0.34	161.7	0.7	0.26	16.9	1.0	0.70	0.7	0.2	0.67
		500mM NaCl	14	73.5	1.1	0.27	149.9	0.7	0.33	20.0	-	-	4.8	0.6	0.48
CSD	let-7gmut	100mM NaCl	14	59.9	1.3	0.37	109.9	1.5	0.73	20.0	-	-	178.7	26.4	0.79
Lin28TT(2)	dlet-7gΔ5	100mM NaCl	14	90.0	4.8	0.42	272.1	4.1	0.46	20.0	-	-	54.1	7.5	0.64
		500mM NaCl	14	90.7	1.2	0.40	179.6	0.6	0.43	16.2	1.2	0.81	2.1	0.4	0.82

Protein	RNA	Buffer Salt Concentration	Equation	Adj R2	Sum of Squares	Normality Test P-value D'Agostino-Pearson	Normality Test P-value Shapiro-Wilks	Number of Points used for fit	Outliers
Lin28TT(2)	preE-let-7g	100mM NaCl	14	0.9835	3206	0.36	0.06	33	0
	let-7gmut	100mM NaCl	14	0.9614	3221	0.59	0.73	33	0
		250mM NaCl	14	0.9916	270	0.70	0.79	33	0
		500mM NaCl	14	0.9715	217	0.78	0.72	30	0
CSD	let-7gmut	100mM NaCl	14	0.9703	226	0.78	0.74	30	0
Lin28TT(2)	dlet-7gΔ5	100mM NaCl	14	0.9648	5705	0.20	0.28	39	0
		500mM NaCl	14	0.9947	142.1	0.5231	0.7192	33	0

Table A4: Non-linear regression fit values and statistics for interactions between MBP fusion proteins and preE-let-7g/let-7gmut RNA and dlet-7gΔ5 DNA.

Normality tests for residuals are passed if P > 0.05.

List of Abbreviations

AUC	Analytical Ultracentrifugation
CSD	Cold Shock Domain
CV	Column Volumes
DMSO	Dimethyl Sulfoxide
DSF	Differential Scanning Fluorimetry
dsRBD	Double-stranded RNA Binding Domain
DTT	Dithiothreitol
Dus	Dihydrouridine Synthase
FMN	Flavin Mononucleotide
GSH	Glutathione (reduced)
GST	Glutathione-S-transferase
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
IMAC	Immobilized Metal Affinity Chromatography
IPTG	Isopropyl-β-D-1-thiogalactopyranoside
LB	Luria-Bertani growth medium
lncRNA	Long non-coding RNA
LRE	Lin28 Response Element
MBP	Maltose Binding Protein
miRNA	microRNA
mRNA	Messenger RNA
MWCO	Molecular Weight Cut Off
ncRNA	Non-coding RNA
PAGE	Polyacrylamide Gel Electrophoresis

PCR	Polymerase Chain Reaction
PEI	Polyethyleneimine
piRNA	Piwi-interacting RNA
PolII	RNA polymerase II
PreE	The pre-element, or terminal loop of a microRNA
Pre-miRNA	Precursor microRNA
Pri-miRNA	Primary microRNA
RBP	RNA Binding Protein
RISC	RNA Induced Silencing Complex
RRM	RNA Recognition Motif
rRNA	Ribosomal RNA
SDS	Sodium Dodecyl Sulphate
SEC	Size Exclusion Chromatography
SEC-MALLS	Size Exclusion Chromatography-Multiple Angle Laser Light Scattering
snoRNA	Small Nucleolar RNA
snRNA	Small Nuclear RNA
snRNP	Small Nuclear Ribonuclear particle
TEMED	Tetramethylethylenediamine
TRIS	tris(hydroxymethyl)aminomethane
tRNA	Transfer RNA
ZnK	Zinc Knuckle

References

1. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
2. DP, B., *MicroRNAs: Genomics,Biogenesis, Mechanism, and Function.* Cell, 2004. **116**.
3. Lee, Y., et al., *MicroRNA genes are transcribed by RNA polymerase II.* EMBO J, 2004. **23**(20): p. 4051-4060.
4. Cai, X., C.H. Hagedorn, and B.R. Cullen, *Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs.* Rna, 2004. **10**(12): p. 1957-66.
5. Ameres, S.L. and P.D. Zamore, *Diversifying microRNA sequence and function.* Nat Rev Mol Cell Biol, 2013. **14**(8): p. 475-488.
6. Han, J., et al., *Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex.* Cell, 2006. **125**(5): p. 887-901.
7. MacRae, I.J., et al., *Structural Basis for Double-Stranded RNA Processing by Dicer.* Science, 2006. **311**(5758): p. 195-198.
8. Hutvágner, G. and P.D. Zamore, *A microRNA in a Multiple-Turnover RNAi Enzyme Complex.* Science, 2002. **297**(5589): p. 2056-2060.
9. Hutvagner, G. and M.J. Simard, *Argonaute proteins: key players in RNA silencing.* Nat Rev Mol Cell Biol, 2008. **9**(1): p. 22-32.
10. Filipowicz, W., S.N. Bhattacharyya, and N. Sonenberg, *Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?* Nat Rev Genet, 2008. **9**(2): p. 102-114.
11. Jones-Rhoades, M.W., D.P. Bartel, and B. Bartel, *MicroRNAs AND THEIR REGULATORY ROLES IN PLANTS.* Annual Review of Plant Biology, 2006. **57**(1): p. 19-53.
12. Bartel, D.P., *MicroRNAs: Target Recognition and Regulatory Functions.* Cell, 2009. **136**(2): p. 215-233.
13. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Research, 2009. **19**(1): p. 92-105.
14. Baek, D., et al., *The impact of microRNAs on protein output.* Nature, 2008. **455**(7209): p. 64-71.
15. Sabin, Leah R., M.J. Delás, and Gregory J. Hannon, *Dogma Derailed: The Many Influences of RNA on the Genome.* Molecular Cell, 2013. **49**(5): p. 783-794.
16. Ulitsky, I. and David P. Bartel, *lincRNAs: Genomics, Evolution, and Mechanisms.* Cell, 2013. **154**(1): p. 26-46.
17. Kung, J.T.Y., D. Colognori, and J.T. Lee, *Long Noncoding RNAs: Past, Present, and Future.* Genetics, 2013. **193**(3): p. 651-669.
18. Brockdorff, N., et al., *The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus.* Cell, 1992. **71**(3): p. 515-526.
19. Wutz, A., T.P. Rasmussen, and R. Jaenisch, *Chromosomal silencing and localization are mediated by different domains of Xist RNA.* Nat Genet, 2002. **30**(2): p. 167-174.

20. Brown, C.J., et al., *The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus.* Cell, 1992. **71**(3): p. 527-542.
21. Lyon, M.F., *Gene Action in the X-chromosome of the Mouse (Mus musculus L.).* Nature, 1961. **190**(4773): p. 372-373.
22. Lee, Jeannie T. and Marisa S. Bartolomei, *X-Inactivation, Imprinting, and Long Noncoding RNAs in Health and Disease.* Cell, 2013. **152**(6): p. 1308-1323.
23. Brown, C.J., et al., *Localization of the X inactivation centre on the human X chromosome in Xq13.* Nature, 1991. **349**(6304): p. 82-84.
24. Chureau, C., et al., *Comparative Sequence Analysis of the X-Inactivation Center Region in Mouse, Human, and Bovine.* Genome Research, 2002. **12**(6): p. 894-908.
25. Jeon, Y. and Jeannie T. Lee, *YY1 Tethers Xist RNA to the Inactive X Nucleation Center*. Cell, 2011. **146**(1): p. 119-133.
26. Lee, J.T., *Disruption of Imprinted X Inactivation by Parent-of-Origin Effects at Tsix.* Cell, 2000. **103**(1): p. 17-27.
27. Sado, T., et al., *Regulation of imprinted X-chromosome inactivation in mice by Tsix.* Development, 2001. **128**(8): p. 1275-1286.
28. Lee, J., L.S. Davidow, and D. Warshawsky, *Tsix, a gene antisense to Xist at the X-inactivation centre.* Nat Genet, 1999. **21**(4): p. 400-404.
29. Sun, B.K., A.M. Deaton, and J.T. Lee, *A Transient Heterochromatic State in Xist Preempts X Inactivation Choice without RNA Stabilization.* Molecular Cell, 2006. **21**(5): p. 617-628.
30. Zhao, J., et al., *Polycomb Proteins Targeted by a Short Repeat RNA to the Mouse X Chromosome.* Science, 2008. **322**(5902): p. 750-756.
31. Tian, D., S. Sun, and J.T. Lee, *The Long Noncoding RNA, Jpx, Is a Molecular Switch for X Chromosome Inactivation.* Cell, 2010. **143**(3): p. 390-403.
32. Malone, C.D. and G.J. Hannon, *Small RNAs as Guardians of the Genome.* Cell, 2009. **136**(4): p. 656-668.
33. Brennecke, J., et al., *Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila.* Cell, 2007. **128**(6): p. 1089-1103.
34. Aravin, A.A., et al., *A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice.* Molecular Cell, 2008. **31**(6): p. 785-799.
35. Aravin, A.A., et al., *Developmentally Regulated piRNA Clusters Implicate MILI in Transposon Control.* Science, 2007. **316**(5825): p. 744-747.
36. Wiedenheft, B., S.H. Sternberg, and J.A. Doudna, *RNA-guided genetic silencing systems in bacteria and archaea.* Nature, 2012. **482**(7385): p. 331-338.
37. Fineran, P.C. and E. Charpentier, *Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information.* Virology, 2012. **434**(2): p. 202-209.
38. Semenova, E., et al., *Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence.* Proceedings of the National Academy of Sciences, 2011. **108**(25): p. 10098-10103.
39. Marraffini, L.A. and E.J. Sontheimer, *Self versus non-self discrimination during CRISPR RNA-directed immunity.* Nature, 2010. **463**(7280): p. 568-571.

40. Moore, M.J. and N.J. Proudfoot, *Pre-mRNA Processing Reaches Back toTranscription and Ahead to Translation.* Cell, 2009. **136**(4): p. 688-700.

41. Wahl, M.C., C.L. Will, and R. Lührmann, *The Spliceosome: Design Principles of a Dynamic RNP Machine.* Cell, 2009. **136**(4): p. 701-718.

42. Will, C.L. and R. Lührmann, *Spliceosome Structure and Function.* Cold Spring Harbor Perspectives in Biology, 2011. **3**(7).

43. Moore, P.B. and T.A. Steitz, *The Roles of RNA in the Synthesis of Protein.* Cold Spring Harbor Perspectives in Biology, 2011. **3**(11).

44. Jühling, F., et al., *tRNAdb 2009: compilation of tRNA sequences and tRNA genes.* Nucleic Acids Research, 2009. **37**(suppl 1): p. D159-D162.

45. Ibba, M. and D. Söll, *AMINOACYL-tRNA SYNTHESIS.* Annual Review of Biochemistry, 2000. **69**(1): p. 617-650.

46. Byrne, R.T., et al., *The crystal structure of unmodified tRNAPhe from Escherichia coli.* Nucleic Acids Research, 2010. **38**(12): p. 4154-4162.

47. Robertus, J.D., et al., *Structure of yeast phenylalanine tRNA at 3 [angst] resolution.* Nature, 1974. **250**(5467): p. 546-551.

48. Motorin, Y. and M. Helm, *tRNA Stabilization by Modified Nucleotides.* Biochemistry, 2010. **49**(24): p. 4934-4944.

49. Sprinzl, M., et al., *Compilation of tRNA sequences and sequences of tRNA genes.* Nucleic Acids Res, 1998. **26**(1): p. 148 - 153.

50. Dunin-Horkawicz, S., et al., *MODOMICS: a database of RNA modification pathways.* Nucleic Acids Research, 2006. **34**(suppl 1): p. D145-D149.

51. Limbach, P.A., P.F. Crain, and J.A. McCloskey, *Summary: the modified nucleosides of RNA.* Nucleic Acids Research, 1994. **22**(12): p. 2183-2196.

52. Lunde, B.M., C. Moore, and G. Varani, *RNA-binding proteins: modular design for efficient function.* Nat Rev Mol Cell Biol, 2007. **8**(6): p. 479-490.

53. Mackereth, C.D. and M. Sattler, *Dynamics in multi-domain protein recognition of RNA.* Current Opinion in Structural Biology, 2012. **22**(3): p. 287-296.

54. Shamoo, Y., N. Abdul-Manan, and K.R. Williams, *Multiple RNA binding domains (RBDs) just don't add up.* Nucleic Acids Research, 1995. **23**(5): p. 725-728.

55. Cléry, A., M. Blatter, and F.H.T. Allain, *RNA recognition motifs: boring? Not quite.* Current Opinion in Structural Biology, 2008. **18**(3): p. 290-298.

56. Muto, Y. and S. Yokoyama, *Structural insight into RNA recognition motifs: versatile molecular Lego building blocks for biological systems.* Wiley Interdiscip Rev RNA, 2012. **3**(2): p. 229-46.

57. Dreyfuss, G., V.N. Kim, and N. Kataoka, *Messenger-RNA-binding proteins and the messages they carry.* Nat Rev Mol Cell Biol, 2002. **3**(3): p. 195-205.

58. Jiang, W., Y. Hou, and M. Inouye, *CspA, the Major Cold-shock Protein of Escherichia coli, Is an RNA Chaperone.* Journal of Biological Chemistry, 1997. **272**(1): p. 196-202.

59. Graumann, P.L. and M.A. Marahiel, *A superfamily of proteins that contain the cold-shock domain.* Trends in Biochemical Sciences, 1998. **23**(8): p. 286-290.

60. Graumann P, et al., *A family of cold shock proteins in Bacillus subtilis is essential for cellular growth and for efficient protein synthesis at optimal and low temperatures.* Mol Microbiol., 1997. **25**(4): p. 741-56.

61. Mihailovich, M., et al., *Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression.* BioEssays, 2010. **32**(2): p. 109-118.

62. Arcus, V., *OB-fold domains: a snapshot of the evolution of sequence, structure and function.* Current Opinion in Structural Biology, 2002. **12**(6): p. 794-801.

63. Max, K.E.A., et al., *T-rich DNA Single Strands Bind to a Preformed Site on the Bacterial Cold Shock Protein Bs-CspB.* Journal of Molecular Biology, 2006. **360**(3): p. 702-714.

64. Schindelin, H., M.A. Marahiel, and U. Heinemann, *Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein.* Nature, 1993. **364**(6433): p. 164-168.

65. Hall, T.M.T., *Multiple modes of RNA recognition by zinc finger proteins.* Current Opinion in Structural Biology, 2005. **15**(3): p. 367-373.

66. Brown, R.S., *Zinc finger proteins: getting a grip on RNA.* Current Opinion in Structural Biology, 2005. **15**(1): p. 94-98.

67. Krishna, S.S., I. Majumdar, and N.V. Grishin, *Structural classification of zinc fingers.* Nucleic Acids Research, 2003. **31**(2): p. 532-550.

68. Razin, S.V., et al., *Cys2His2 zinc finger protein family: Classification, functions, and major members.* Biochemistry (Moscow), 2012. **77**(3): p. 217-226.

69. De Guzman, R.N., et al., *Structure of the HIV-1 Nucleocapsid Protein Bound to the SL3 Ψ-RNA Recognition Element.* Science, 1998. **279**(5349): p. 384-388.

70. Pavletich, N. and C. Pabo, *Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A.* Science, 1991. **252**(5007): p. 809-817.

71. Nolte RT, et al., *Differing roles for zinc fingers in DNA recognition: Structure of a six-finger transcription factor IIIA complex*

Proc Natl Acad Sci U S A, 1998. **95**(6): p. 2938-2943.

72. Masliah, G., P. Barraud, and F.T. Allain, *RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence.* Cellular and Molecular Life Sciences, 2013. **70**(11): p. 1875-1895.

73. Ryter, J.M. and S.C. Schultz, *Molecular basis of double‐stranded RNA‐protein interactions: structure of a dsRNA‐binding domain complexed with dsRNA.* The EMBO Journal, 1998. **17**(24): p. 7505-7513.

74. Doyle, M. and M.F. Jantsch, *New and old roles of the double-stranded RNA-binding domain.* Journal of Structural Biology, 2002. **140**(1–3): p. 147-153.

75. Kato, T., et al., *A Novel Human tRNA-Dihydrouridine Synthase Involved in Pulmonary Carcinogenesis.* Cancer Research, 2005. **65**(13): p. 5638-5646.

76. Griffiths, S., et al., *Crystallization and preliminary X-ray crystallographic analysis of the catalytic domain of human dihydrouridine synthase.* Acta Crystallographica Section F, 2012. **68**(3): p. 333-336.

77. Moss, E.G., R.C. Lee, and V. Ambros, *The Cold Shock Domain Protein LIN-28 Controls Developmental Timing in C. elegans and Is Regulated by the lin-4 RNA.* Cell, 1997. **88**(5): p. 637-646.

78. Huang, Y., *A mirror of two faces: Lin28 as a master regulator of both miRNA and mRNA.* Wiley Interdisciplinary Reviews: RNA, 2012. **3**(4): p. 483-494.

79. Viswanathan, S.R., G.Q. Daley, and R.I. Gregory, *Selective Blockade of MicroRNA Processing by Lin28.* Science, 2008. **320**(5872): p. 97-100.

80. Piskounova, E., et al., *Determinants of MicroRNA Processing Inhibition by the Developmentally Regulated RNA-binding Protein Lin28.* Journal of Biological Chemistry, 2008. **283**(31): p. 21310-21314.

81. Piskounova, E., et al., *Lin28A and Lin28B Inhibit let-7 MicroRNA Biogenesis by Distinct Mechanisms.* Cell, 2011. **147**(5): p. 1066-1079.

82. Thornton, J.E. and R.I. Gregory, *How does Lin28 let-7 control development and disease?* Trends in Cell Biology, 2012. **22**(9): p. 474-482.

83. Büssing, I., F.J. Slack, and H. Großhans, *let-7 microRNAs in development, stem cells and cancer.* Trends in Molecular Medicine, 2008. **14**(9): p. 400-409.

84. Roush, S. and F.J. Slack, *The let-7 family of microRNAs.* Trends in Cell Biology, 2008. **18**(10): p. 505-516.

85. Guo, Y., et al., *Identification and characterization of lin-28 homolog B (LIN28B) in human hepatocellular carcinoma.* Gene, 2006. **384**(0): p. 51-61.

86. Yu, J., et al., *Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells.* Science, 2007. **318**(5858): p. 1917-1920.

87. Newman, M.A., J.M. Thomson, and S.M. Hammond, *Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing.* RNA, 2008. **14**(8): p. 1539-1549.

88. Rybak, A., et al., *A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment.* Nat Cell Biol, 2008. **10**(8): p. 987-993.

89. Viswanathan, S.R. and G.Q. Daley, *Lin28: A MicroRNA Regulator with a Macro Role.* Cell, 2010. **140**(4): p. 445-449.

90. Thornton, J.E., et al., *Lin28-mediated control of let-7 microRNA expression by alternative TUTases Zcchc11 (TUT4) and Zcchc6 (TUT7).* RNA, 2012. **18**(10): p. 1875-1885.

91. Hagan, J.P., E. Piskounova, and R.I. Gregory, *Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells.* Nat Struct Mol Biol, 2009. **16**(10): p. 1021-1025.

92. Heo, I., et al., *TUT4 in Concert with Lin28 Suppresses MicroRNA Biogenesis through Pre-MicroRNA Uridylation.* Cell, 2009. **138**(4): p. 696-708.

93. Heo, I., et al., *Lin28 Mediates the Terminal Uridylation of let-7 Precursor MicroRNA.* Molecular Cell, 2008. **32**(2): p. 276-284.

94. Ustianenko, D., et al., *Mammalian DIS3L2 exoribonuclease targets the uridylated precursors of let-7 miRNAs.* RNA, 2013.

95. Chang, H.-M., et al., *A role for the Perlman syndrome exonuclease Dis3l2 in the Lin28-let-7 pathway.* Nature, 2013. **497**(7448): p. 244-248.

96. Murray, M.J., et al., *LIN28 Expression in malignant germ cell tumors downregulates let-7 and increases oncogene levels.* Cancer Res, 2013. **73**(15): p. 4872-84.

97. Mayr, F. and U. Heinemann, *Mechanisms of Lin28-Mediated miRNA and mRNA Regulation—A Structural and Functional Perspective.* International Journal of Molecular Sciences, 2013. **14**(8): p. 16532-16553.

98. Hanna, J., et al., *Direct cell reprogramming is a stochastic process amenable to acceleration.* Nature, 2009. **462**(7273): p. 595-601.

99. Viswanathan, S.R., et al., *Lin28 promotes transformation and is associated with advanced human malignancies.* Nat Genet, 2009. **41**(7): p. 843-848.

100. Yang, X., et al., *Double-Negative Feedback Loop between Reprogramming Factor LIN28 and microRNA let-7 Regulates Aldehyde Dehydrogenase 1–Positive Cancer Stem Cells.* Cancer Research, 2010. **70**(22): p. 9463-9472.

101. Zhang, Wen C., et al., *Glycine Decarboxylase Activity Drives Non-Small Cell Lung Cancer Tumor-Initiating Cells and Tumorigenesis.* Cell, 2012. **148**(1–2): p. 259-272.

102. Jiang, X., et al., *Blockade of miR-150 Maturation by MLL-Fusion/MYC/LIN-28 Is Required for MLL-Associated Leukemia.* Cancer Cell, 2012. **22**(4): p. 524-535.

103. Chang, T.C., et al., *Lin-28B transactivation is necessary for Myc-mediated let-7 repression and proliferation.* Proc Natl Acad Sci U S A, 2009. **106**(9): p. 3384-9.

104. Iliopoulos, D., H.A. Hirsch, and K. Struhl, *An Epigenetic Switch Involving NF-κB, Lin28, Let-7 MicroRNA, and IL6 Links Inflammation to Cell Transformation.* Cell, 2009. **139**(4): p. 693-706.

105. Rosen, J.M. and C.T. Jordan, *The Increasing Complexity of the Cancer Stem Cell Paradigm.* Science, 2009. **324**(5935): p. 1670-1673.

106. Zhu, H., et al., *Lin28a transgenic mice manifest size and puberty phenotypes identified in human genetic association studies.* Nat Genet, 2010. **42**(7): p. 626-630.

107. Ong, K.K., et al., *Genetic variation in LIN28B is associated with the timing of puberty.* Nat Genet, 2009. **41**(6): p. 729-733.

108. Lettre, G., et al., *Identification of ten loci associated with height highlights new biological pathways in human growth.* Nat Genet, 2008. **40**(5): p. 584-591.

109. Yokoyama, S., et al., *Dynamic gene expression of Lin-28 during embryonic development in mouse and chicken.* Gene Expression Patterns, 2008. **8**(3): p. 155-160.

110. Faas, L., et al., *Lin28 proteins are required for germ layer specification in Xenopus.* Development, 2013. **140**(5): p. 976-986.

111. West, J.A., et al., *A role for Lin28 in primordial germ-cell development and germ-cell malignancy.* Nature, 2009. **460**(7257): p. 909-913.

112. Papaioannou, G., et al., *let-7 and miR-140 microRNAs coordinately regulate skeletal development.* Proceedings of the National Academy of Sciences, 2013. **110**(35): p. E3291-E3300.

113. Polesskaya, A., et al., *Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency.* Genes & Development, 2007. **21**(9): p. 1125-1138.

114. Cimadamore, F., et al., *SOX2–LIN28/let-7 pathway regulates proliferation and neurogenesis in neural precursors.* Proceedings of the National Academy of Sciences, 2013. **110**(32): p. E3017-E3026.

115. Balzer, E., et al., *LIN28 alters cell fate succession and acts independently of the let-7 microRNA during neurogliogenesis in vitro.* Development, 2010. **137**(6): p. 891-900.

116. Zhu, H., et al., *The Lin28/let-7 Axis Regulates Glucose Metabolism.* Cell, 2011. **147**(1): p. 81-94.

117. Zhang, J., et al., *The polymorphism in the let-7 targeted region of the Lin28 gene is associated with increased risk of type 2 diabetes mellitus.* Molecular and Cellular Endocrinology, 2013. **375**(1–2): p. 53-57.

118. Shyh-Chang, N., et al., *Lin28 Enhances Tissue Repair by Reprogramming Cellular Metabolism.* Cell, 2013. **155**(4): p. 778-792.

119. Ramachandran, R., B.V. Fausett, and D. Goldman, *Ascl1a regulates Muller glia dedifferentiation and retinal regeneration through a Lin-28-dependent, let-7 microRNA signalling pathway.* Nat Cell Biol, 2010. **12**(11): p. 1101-1107.

120. McCarty, M.F., *Metformin may antagonize Lin28 and/or Lin28B activity, thereby boosting let-7 levels and antagonizing cancer progression.* Medical Hypotheses, 2012. **78**(2): p. 262-269.

121. Oliveras-Ferraros, C., et al., *Micro(mi)RNA expression profile of breast cancer epithelial cells treated with the anti-diabetic drug metformin: Induction of the tumor suppressor miRNA let-7a and suppression of the TGFβ-induced oncomiR miRNA-181a.* Cell Cycle, 2011. **10**(7): p. 1144-1151.

122. Nam, Y., et al., *Molecular Basis for Interaction of let-7 MicroRNAs with Lin28.* Cell, 2011. **147**(5): p. 11.

123. Mayr, F., et al., *The Lin28 cold-shock domain remodels pre-let-7 microRNA.* Nucleic Acids Research, 2012.

124. Loughlin, F.E., et al., *Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28.* Nat Struct Mol Biol, 2012. **19**(1): p. 84-89.

125. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Research, 2003. **31**(13): p. 3406-3415.

126. Balzer, E. and E. Moss, *Localization of the developmental timing regulator Lin28 to mRNP complexes, P-bodies and stress granules.* RNA Biology, 2007. **4**

(1): p. 16-25.

127. Qiu, C., et al., *Lin28-mediated post-transcriptional regulation of Oct4 expression in human embryonic stem cells.* Nucleic Acids Research, 2010. **38**(4): p. 1240-1248.

128. Xu, B. and Y. Huang, *Histone H2a mRNA interacts with Lin28 and contains a Lin28-dependent posttranscriptional regulatory element.* Nucleic Acids Research, 2009. **37**(13): p. 4256-4263.

129. Xu, B., K. Zhang, and Y. Huang, *Lin28 modulates cell growth and associates with a subset of cell cycle regulator mRNAs in mouse embryonic stem cells.* RNA, 2009. **15**(3): p. 357-361.

130. Lei, X.-X., et al., *Determinants of mRNA recognition and translation regulation by Lin28.* Nucleic Acids Research, 2012. **40**(8): p. 3574-3584.

131. Peng, S., et al., *Genome-Wide Studies Reveal That Lin28 Enhances the Translation of Genes Important for Growth and Survival of Human Embryonic Stem Cells.* STEM CELLS, 2011. **29**(3): p. 496-504.

132. Jin, J., et al., *Evidence that Lin28 stimulates translation by recruiting RNA helicase A to polysomes.* Nucleic Acids Research, 2011. **39**(9): p. 3724-3734.

133. Kallen, A.N., J. Ma, and Y. Huang, *Does Lin28 antagonize miRNA-mediated repression by displacing miRISC from target mRNAs?* Frontiers in Genetics, 2012. **3**.

134. Qiao, C., et al., *Drosha mediates destabilization of Lin28 mRNA targets.* Cell Cycle, 2012. **11**(19): p. 3590-3598.

135. Wilbert, Melissa L., et al., *LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance.* Molecular Cell, 2012. **48**(2): p. 195-206.

136. Cho, J., et al., *LIN28A Is a Suppressor of ER-Associated Translation in Embryonic Stem Cells.* Cell, 2012. **151**(4): p. 765-777.

137. Hafner, M., et al., *Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition.* RNA, 2013. **19**(5): p. 613-626.

138. Bishop, A., et al., *Identification of the tRNA-dihydrouridine synthase family.* J Biol Chem, 2002. **277**(28): p. 25090 - 25095.

139. Dalluge, J.J., et al., *Conformational Flexibility in RNA: The Role of Dihydrouridine.* Nucleic Acids Research, 1996. **24**(6): p. 1073-1079.

140. Kowalak, J.A., et al., *The Role of Posttranscriptional Modification in Stabilization of Transfer RNA from Hyperthermophiles.* Biochemistry, 1994. **33**(25): p. 7869-7876.

141. Dalluge, J., et al., *Posttranscriptional modification of tRNA in psychrophilic bacteria.* J Bacteriol, 1997. **179**(6): p. 1918 - 1923.

142. Rider, L.W., et al., *Mechanism of Dihydrouridine Synthase 2 from Yeast and the Importance of Modifications for Efficient tRNA Reduction.* Journal of Biological Chemistry, 2009. **284**(16): p. 10324-10333.

143. Yu, F., et al., *Molecular basis of dihydrouridine formation on tRNA.* Proc Natl Acad Sci U S A, 2011. **108**(49): p. 19593 - 19598.

144. Kasprzak, J., A. Czerwoniec, and J. Bujnicki, *Molecular evolution of dihydrouridine synthases.* BMC Bioinformatics, 2012. **13**(1): p. 153.

145. Xing, F., et al., *The specificities of four yeast dihydrouridine synthases for cytoplasmic tRNAs.* J Biol Chem, 2004. **279**(17): p. 17850 - 17860.

146. Park, F., et al., *The 1.59 A resolution crystal structure of TM0096, a flavin mononucleotide binding protein from Thermotoga maritima.* Proteins, 2004. **55**(3): p. 772 - 774.

147. Chen, M., et al., *Structure of dihydrouridine synthase C (DusC) from Escherichia coli.* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2013. **69**(Pt 8): p. 834-8.

148. Desjardins, A., et al., *Importance of the NCp7-like domain in the recognition of pre-let-7g by the pluripotency factor Lin28.* Nucleic Acids Research, 2012. **40**(4): p. 1767-1777.

149. Sambrook, J. and D. Russell, *Molecular Cloning: A Laboratory Manual*. 3rd ed. Vol. 2. 2001, Cold Spring Harbour: Cold Spring Harbour Press.

150. Joyce, C.M. and S.J. Benkovic, *DNA Polymerase Fidelity: Kinetics, Structure, and Checkpoints†.* Biochemistry, 2004. **43**(45): p. 14317-14324.

151. Sambrook, J. and D. Russell, *Molecular Cloning: A Laboratory Manual*, in *Molecular Cloning: A Laboratory Manual*. 2001, Cold Spring Harbour Press: Cold Spring Harbour. p. 1.84-1.87.

152. Sambrook, J. and D. Russell, *Molecular Cloning: A Laboratory Manual*, in *Molecular Cloning: A Laboratory Manual*. 2001, Cold Spring Harbour Press: Cold Spring Harbour. p. 5.4-5.8.

153. Sambrook, J. and D. Russell, *Molecular Cloning: A Laboratory Manual*, in *Molecular Cloning: A Laboratory Manual*. 2001, Cold Spring Harbour Press: Cold Spring Harbour. p. 15.14-15.21.

154. Sambrook, J. and D. Russell, *Molecular Cloning: A Laboratory Manual*, in *Molecular Cloning: A Laboratory Manual*. 2001, Cold Spring Harbour Press: Cold Spring Harbour. p. A8.40-A8.42.

155. Niesen, F.H., H. Berglund, and M. Vedadi, *The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability.* Nat. Protocols, 2007. **2**(9): p. 2212-2221.

156. Matulis, D., et al., *Thermodynamic Stability of Carbonic Anhydrase: Measurements of Binding Affinity and Stoichiometry Using ThermoFluor.* Biochemistry, 2005. **44**(13): p. 5258-5266.

157. Schulz, M.N., J. Landström, and R.E. Hubbard, *MTSA—A Matlab program to fit thermal shift data.* Analytical Biochemistry, 2013. **433**(1): p. 43-47.

158. Lakowicz, J., *Fluorescence Anisotropy*, in *Principles of Fluorescence Spectroscopy*. 2006, Springer. p. 353-380.

159. Hulme, E.C. and M.A. Trevethick, *Ligand binding assays at equilibrium: validation and interpretation.* British Journal of Pharmacology, 2010. **161**(6): p. 1219-1237.

160. Wyatt, P.J., *Light scattering and the absolute characterization of macromolecules.* Analytica Chimica Acta, 1993. **272**(1): p. 1-40.

161. Wyatt. *SEC-MALLS Theory*. 2012 [cited 2013 091213]; Available from: http://www.wyatt.eu/index.php?id=molar_mass.

162. Wen, J., T. Arakawa, and J.S. Philo, *Size-Exclusion Chromatography with On-Line Light-Scattering, Absorbance, and Refractive Index Detectors for Studying Proteins and Their Interactions.* Analytical Biochemistry, 1996. **240**(2): p. 155-166.

163. Wyatt. *Mass determination by SEC-MALLS*. 2012 [cited 2013 091213]; Available from: http://www.wyatt.eu/index.php?id=absolute-technique&L=0%2522%20onfocus%253D%2522blurLink%2528this%2529%253B.

164. Blow, D., *Outline of Crystallography for Biologists*. 2004, Oxford: Oxford University Press.

165. Read, R. *Basic diffraction theory: waves, interference, reciprocal space*. 1999 [cited 2014 24/02/2014]; Available from: http://www-structmed.cimr.cam.ac.uk/Course/Basic_diffraction/Diffraction.html.

166. Read, R. *Advanced diffraction: waves, interference and complex numbers*. 1999 [cited 2014; Available from: http://www-structmed.cimr.cam.ac.uk/Course/Adv_diff1/Diffraction.html.

167. IUCr. *Dictionary of Crystallography: Atomic Scattering Factor*. [cited 2014; Available from: http://reference.iucr.org/dictionary/Atomic_scattering_factor.

168. Smyth, D.R., et al., *Crystal structures of fusion proteins with large-affinity tags.* Protein Science, 2003. **12**(7): p. 1313-1322.

169. Baneyx, F., *Recombinant protein expression in Escherichia coli.* Current Opinion in Biotechnology, 1999. **10**(5): p. 411-421.

170. Shih, Y.-P., et al., *High-throughput screening of soluble recombinant proteins.* Protein Science, 2002. **11**(7): p. 1714-1719.

171. Braun, P., et al., *Proteome-scale purification of human proteins from bacteria.* Proceedings of the National Academy of Sciences, 2002. **99**(5): p. 2654-2659.

172. McTigue, M.A., D.R. Williams, and J.A. Tainer, *Crystal Structures of a Schistosomal Drug and Vaccine Target: Glutathione S-Transferase fromSchistosoma japonicaand its Complex with the Leading Antischistomal Drug Praziquantel.* Journal of Molecular Biology, 1995. **246**(1): p. 21-27.

173. Zhan, Y., X. Song, and G.W. Zhou, *Structural analysis of regulatory protein domains using GST-fusion proteins.* Gene, 2001. **281**(1–2): p. 1-9.

174. Carter, A.P., et al., *Crystal Structure of the Dynein Motor Domain.* Science, 2011. **331**(6021): p. 1159-1165.

175. Kapust, R.B. and D.S. Waugh, *Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused.* Protein Science, 1999. **8**(8): p. 1668-1674.

176. Spurlino, J.C., G.Y. Lu, and F.A. Quiocho, *The 2.3-A resolution structure of the maltose- or maltodextrin-binding protein, a primary receptor of bacterial active transport and chemotaxis.* J Biol Chem, 1991. **266**(8): p. 5202-19.

177. Kobe, B., et al., *Crystal structure of human T cell leukemia virus type 1 gp21 ectodomain crystallized as a maltose-binding protein chimera reveals structural evolution of retroviral transmembrane proteins.* Proceedings of the National Academy of Sciences, 1999. **96**(8): p. 4319-4324.

178. Center, R.J., et al., *Crystallization of a trimeric human T cell leukemia virus type 1 gp21 ectodomain fragment as a chimera with maltose-binding protein.* Protein Science, 1998. **7**(7): p. 1612-1619.

179. Liu, Y., et al., *Crystal structure of the SarR protein from Staphylococcus aureus.* Proceedings of the National Academy of Sciences, 2001. **98**(12): p. 6877-6882.

180. Ke, A. and C. Wolberger, *Insights into binding cooperativity of MATa1/MATα2 from the crystal structure of a MATa1 homeodomain-maltose binding protein chimera.* Protein Science, 2003. **12**(2): p. 306-312.

181. Cornell, N.W. and K.E. Crivaro, *Stability constant for the zinc-dithiothreitol complex.* Analytical Biochemistry, 1972. **47**(1): p. 203-208.

182. Desjardins, A., J. Bouvette, and P. Legault, *Stepwise assembly of multiple Lin28 proteins on the terminal loop of let-7 miRNA precursors.* Nucleic Acids Research, 2014.

183. Shaik Syed Ali, P., et al., *Recognition of the let-7g miRNA precursor by human Lin28B.* FEBS Letters, 2012. **586**(22): p. 3986-3990.

184. Motulsky, H. and R. Brown, *Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate.* BMC Bioinformatics, 2006. **7**(1): p. 123.

185. Juhling, F., et al., *tRNAdb 2009: compilation of tRNA sequences and tRNA genes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D159-62.

186. Dalluge, J., et al., *Conformational flexibility in RNA: the role of dihydrouridine.* Nucleic Acids Res, 1996. **24**(6): p. 1073 - 1079.

187. Yu, F., et al., *Molecular basis of dihydrouridine formation on tRNA.* Proceedings of the National Academy of Sciences, 2011. **108**(49): p. 19593-19598.

188. Watanabe, M., et al., *Biosynthesis of Archaeosine, a Novel Derivative of 7-Deazaguanosine Specific to Archaeal tRNA, Proceeds via a Pathway Involving Base Replacement on the tRNA Polynucleotide Chain.* Journal of Biological Chemistry, 1997. **272**(32): p. 20146-20151.

189. Watanabe, M., et al., *tRNA Recognition of tRNA-guanine Transglycosylase from a Hyperthermophilic Archaeon, Pyrococcus horikoshii.* Journal of Biological Chemistry, 2001. **276**(4): p. 2387-2394.

190. Ishitani, R., et al., *Alternative Tertiary Structure of tRNA for Recognition by a Posttranscriptional Modification Enzyme.* Cell, 2003. **113**(3): p. 383-394.

191. Ishitani, R., et al., *Crystal Structure of Archaeosine tRNA-guanine Transglycosylase.* Journal of Molecular Biology, 2002. **318**(3): p. 665-677.

192. Aravind, L. and E.V. Koonin, *Novel Predicted RNA-Binding Domains Associated with the Translation Machinery.* Journal of Molecular Evolution, 1999. **48**(3): p. 291-302.

193. Davis, D.R., *Stabilization of RNA stacking by pseudouridine.* Nucleic Acids Research, 1995. **23**(24): p. 5020-5026.

194. Hamma, T. and A.R. Ferré-D'Amaré, *Pseudouridine Synthases.* Chemistry & Biology, 2006. **13**(11): p. 1125-1135.

195. Hoang, C. and A.R. Ferré-D'Amaré, *Cocrystal Structure of a tRNA Ψ55 Pseudouridine Synthase: Nucleotide Flipping by an RNA-Modifying Enzyme.* Cell, 2001. **107**(7): p. 929-939.

196. Gu, X., et al., *Molecular Recognition of tRNA by tRNA Pseudouridine 55 Synthase†.* Biochemistry, 1998. **37**(1): p. 339-343.

197. Ericsson, U.B., P. Nordlund, and B.M. Hallberg, *X-ray structure of tRNA pseudouridine synthase TruD reveals an inserted domain with a novel fold.* FEBS Letters, 2004. **565**(1–3): p. 59-64.

198. Hur, S. and R.M. Stroud, *How U38, 39, and 40 of Many tRNAs Become the Targets for Pseudouridylation by TruA.* Molecular Cell, 2007. **26**(2): p. 189-203.

199. Kabsch, W., *Xds.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 125-32.

200. Winn, M.D., et al., *Overview of the CCP4 suite and current developments.* Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 4): p. 235-42.

201. Evans, P.R. and G.N. Murshudov, *How good are my data and what is the resolution?* Acta Crystallogr D Biol Crystallogr, 2013. **69**(Pt 7): p. 1204-14.

202. Painter, J. and E.A. Merritt, *Optimal description of a protein structure in terms of multiple groups undergoing TLS motion.* Acta Crystallogr D Biol Crystallogr, 2006. **62**(Pt 4): p. 439-50.

203. Langer, G.G., et al., *Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7.* Nature Protocols, 2008. **3**: p. 1171-1179. .

204. Thorn, A. and G.M. Sheldrick, *ANODE: anomalous and heavy-atom density calculation.* Journal of Applied Crystallography, 2011. **44**(6): p. 1285-1287.

205. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* J. Mol. Biol., 2007. **372**: p. 774--797.

206. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 1): p. 12-21.

207. Keating, K.S. and A.M. Pyle, *RCrane: semi-automated RNA model building.* Acta Crystallogr D Biol Crystallogr, 2012. **68**(Pt 8): p. 985-95.

208. Potterton, L., et al., *Developments in the CCP4 molecular-graphics project.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 12 Pt 1): p. 2288-94.

209. Nicholls, R.A., F. Long, and G.N. Murshudov, *Low-resolution refinement tools in REFMAC5.* Acta Crystallographica Section D-Biological Crystallography, 2012. **68**: p. 404-417.

210. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Mol Syst Biol, 2011. **7**(539): p. 539.

211. Gouet, P., et al., *ESPript: analysis of multiple sequence alignments in PostScript.* Bioinformatics, 1999. **15**(4): p. 305-8.

212. Savage, D., V. de Crecy-Lagard, and A. Bishop, *Molecular determinants of dihydrouridine synthase activity.* FEBS Lett, 2006. **580**(22): p. 5198 - 5202.

213. Giegé, R., et al., *Structure of transfer RNAs: similarity and variability.* Wiley Interdisciplinary Reviews: RNA, 2012. **3**(1): p. 37-61.

214. Nobles, K.N., et al., *Highly conserved modified nucleosides influence Mg2+-dependent tRNA folding.* Nucleic Acids Res, 2002. **30**(21): p. 4751-60.

215. Blainey, P.C., et al., *A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA.* Proceedings of the National Academy of Sciences, 2006. **103**(15): p. 5752-5757.

216. Bonnet, I., et al., *Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA.* Nucleic Acids Research, 2008. **36**(12): p. 4118-4127.

217. Deniz, A.A., S. Mukhopadhyay, and E.A. Lemke, *Single-molecule biophysics: at the interface of biology, physics and chemistry.* Journal of The Royal Society Interface, 2008. **5**(18): p. 15-45.