The ins and outs of lncRNA structure: how, why and what comes next?

Amanda J Blythe\textsuperscript{a,1}, Archa Fox\textsuperscript{b}, Charles S Bond\textsuperscript{a,1}.

\textsuperscript{a}School of Chemistry and Biochemistry, University of Western Australia, Western Australia, 6009 Australia

\textsuperscript{b}Harry Perkins Institute of Biomedical Research, Western Australia, 6009 Australia

\textsuperscript{1}To whom correspondence should be addressed.

Charles Bond,
School of Chemistry and Biochemistry,
University of Western Australia,
35 Stirling Hwy,
Crawley, WA 6009
Australia
Tel: +61 8 6488 4406, Fax: +61 8 6488 1005
email: charles.bond@uwa.edu.au

Amanda Blythe,
School of Chemistry and Biochemistry,
University of Western Australia,
35 Stirling Hwy,
Crawley, WA 6009
Australia
Tel: +61 8 6488 4464, Fax: +61 8 6488 1005
email: amanda.blythe@uwa.edu.au
Abstract
The field of structural biology has the unique advantage of being able to provide a comprehensive picture of biological mechanisms at the molecular and atomic level. Long noncoding RNAs (lncRNAs) represent the new frontier in the molecular biology of complex organisms yet it remains the least characterised of all the classes of RNA. Thousands of new lncRNAs are being reported each year yet very little structural data exists for this rapidly expanding field. The length of lncRNAs range from 200 nt to over 100 kb in length and generally exhibit low cellular abundance. Therefore, obtaining sufficient quantities of lncRNA to use for structural analysis is challenging. However, as technologies develop structures of lncRNAs are starting to emerge providing important information regarding their mechanism of action. Here we review the current methods used to determine the structure of lncRNA and lncRNA:protein complexes and describe the significant contribution structural biology has and will make to the field of lncRNA research.

Keywords
lncRNA; structure; RNA structure; RNA-protein interaction; structural biology; non-coding

1. Introduction
Structural biology has long been at the forefront of being able to provide detailed mechanistic views of important cellular processes. The field of structural biology has made significant contributions to the understanding of protein function, with over 100,000 protein structures in the Protein Data Bank (PDB) to date. When this is compared to the
1,078 RNA and 1,522 RNA:protein structures, it is clear that structural biology is lagging behind in the field of RNA research. The ability of structural biology to contribute to RNA research has long been hindered by the difficulties in obtaining sufficient RNA, either by isolating the RNA target or producing it via \textit{in vitro} transcription. These difficulties are greatly exacerbated in the field of long non-coding RNA (lncRNA) research due to the length of the transcripts, which range from 200nt to in excess of 100kb, and their generally low cellular abundance. As such it is no surprise that, with the exception of ribosomal RNA (rRNA), lncRNA is the least characterised in terms of structure and function amongst the various classes of RNA.

LncRNAs represent a new frontier in the molecular biology of complex organisms as it is increasingly evident that they are involved in the regulation of almost every stage of gene expression, as well as being implicated in a variety of disease states. Unlike proteins, whose sequences are often highly conserved across species indicating functional importance, the primary structure of lncRNA generally displays extremely poor evolutionary conservation\textsuperscript{2-9}. This may be due to nucleotide substitutions in protein-coding sequences being more deleterious compared with non-coding sequence, and that conservation may be through functional or structural mechanisms, rather than being sequence based\textsuperscript{9}. To this effect, with the structural characterisation of some lncRNAs we are beginning to see evidence of secondary and tertiary structure conservation\textsuperscript{10-13}. The structure-function relationship has long since been proven for proteins, and it is becoming evident that the same can be said for RNA.

Certain functional properties of RNA are determined by their primary structure, with sequence specific processes like the base pairing of mature micro RNA and short interfering RNA molecules to their targets\textsuperscript{14}. However, RNA also has a unique ability to
adopt a variety of complex secondary and tertiary folds and, because of its long length, IncRNA has the potential for increased structural complexity. Indeed, recent studies have shown that IncRNAs exhibit greater structural complexity than messenger RNA with comparable stability profiles\textsuperscript{15, 16}. The structural versatility of RNA allows it to perform various functions ranging from precise protein recognition (ribonucleoproteins; RNPs), to catalysis (ribozymes\textsuperscript{17-19}) and metabolite sensing (riboswitches\textsuperscript{20-22}). Current research is demonstrating that many IncRNAs act in conjunction with proteins and other nucleic acids and that their secondary or tertiary structures are important for mediating these interactions.

Currently there are 111, 685 human annotated IncRNA transcripts (LNCipedia 3.1\textsuperscript{23, 24}); however, despite (or perhaps because of) their wide distribution across the genomes of complex organisms only a minute fraction of IncRNAs have been functionally and/or structurally characterised. Indeed, there still exists wide debate over the proportion of these IncRNAs that will eventually be proven to be functional\textsuperscript{25-32}. A contributing factor to slow rates of functional determination is certainly the excessive length and low cellular abundance of most IncRNAs, which presents a challenge to most \textit{in vitro} and \textit{in vivo} characterisation techniques. In this review we seek to outline the major advances structural biology has made in the understanding of the molecular mechanism behind IncRNA function, highlight the different methodological approaches the field uses to study IncRNA and to show how these methods can/will be applied in the future to IncRNA research.

2. \textbf{Methods to determine RNA structure}

Determining RNA secondary structure is an important first step towards understanding RNA function. RNA secondary structures comprise of single stranded regions that
connect helices, loops and bulges (Figure 1A). These secondary structures dictate the formation of RNA tertiary structure through the coaxial stacking of adjacent helices at junctions, pseudoknots and kissing loops; and, networks of triple helices, tetraloop-receptor interactions and other structures dictated by sequence-specific or backbone interactions (reviewed in \(^{33}\))(Figure 1B).

A variety of methods have been employed to predict and determine structures of RNA and these can be divided into three groups: bioinformatics, biochemical methods and 3-dimensional methods. We have summarised the applications, advantages and disadvantages of the individual methods from each group in Table 1 and briefly describe them here in the context of IncRNA; however, for a more detailed explanation see reviews \(^{34-37}\).

2.1. Bioinformatics

RNA structure prediction programs predict the secondary structure of RNA using the thermodynamics of base pairing, base stacking and hairpins. Whilst these programs have been proven to be accurate for short RNA sequences, the accuracy drops significantly as the length of the transcript increases, thus proving ineffective in predicting the structure of most IncRNAs. Furthermore, these programs are unable to cope with predicting non-Watson-Crick base-pairs and long range tertiary interactions\(^{38}\). Tertiary structure prediction programs have previously accurately predicted the 3D structures of RNA <50 nt in length\(^{39}\); however, by utilizing the experimental information from chemical probing, and as more atomic resolution structures are being determined, these programs are increasingly able to predict tertiary structures of RNAs up to a few hundred nucleotides\(^{40-43}\). In another novel approach to overcome computational limitations in RNA structure
prediction, “crowdsourcing” games have been developed to capitalize on the collective human intelligence and problem-solving capabilities to answer fundamental questions about RNA folding mechanisms. EteRNA (http://eterna.cmu.edu/web/, for 2D structure determination) has led to new RNA design rules\(^{44}\) and RNA-Puzzles\(^{45, 46}\) (for 3D structure determination) has identified the low accuracy of non-Watson-Crick interactions as the major bottleneck in RNA structure prediction\(^{46}\). Through the integration of both 2D and 3D experimental data, computational methods will increasingly be able to become more accurate for longer RNAs.

Phylogenetic analysis looks at the covariance between two nucleotides over a diverse range of species and is currently widely-used for RNA structure prediction\(^{47, 48}\). This technique has been successfully used in the case of ribosomal and riboswitch RNAs\(^{49-52}\); however, IncRNAs lack sequence diversity and exhibit low sequence conservation which significantly limits the number of available sequences for use in multiple sequence alignments. The development of methods to overcome these challenges and to apply these predictions to cover entire genomes is underway\(^{13}\). No matter how precise computational methodologies become, structure prediction techniques will always require experimental validation; however, when used in combination with biochemical structural determination methods, phylogenetic analysis can be a powerful technique in identifying functional regions of structure in IncRNA\(^{53, 54}\).

2.2. Biochemical

There have been many biochemical techniques designed to obtain RNA structural information, most commonly chemical or enzymatic probing. These approaches are able to differentiate between single stranded and helical regions with single nucleotide resolution that result in the generation of 2-D RNA structure maps. RNA cleavage with
nucleases or metal ions serves to map single or double stranded RNA regions depending on the nuclease specificity\textsuperscript{55-58}. Chemical probing relies on chemical base modification at specific single-stranded bases, or binding to mobile regions of the RNA backbone\textsuperscript{59-63}. These methods are followed by primer extension or end-labelling to detect the cleavage sites or modifications\textsuperscript{36, 64-66}. These techniques have been instrumental in determining the early ribosomal RNA secondary structures, revealing important structural elements in other RNAs and have since been applied to the structural determination of IncRNA as there are generally no length, quantity or heterogeneity restrictions\textsuperscript{49-51, 55, 57, 62, 67-69}. The 2-D information gained from these approaches directs mutagenesis and deletion studies and, in combination with foot-printing studies, can identify interaction motifs and provide important information regarding the mechanism of RNA function\textsuperscript{70}.

2.3. \textit{In vivo}

Given the difficulty in obtaining robust and plentiful samples for \textit{in vitro} structure determination, IncRNA structural biology is turning towards the structural interrogations of candidate RNAs \textit{in vivo}, generally from cultured cells. Several chemical probes (Pb\textsuperscript{2+} ions, dimethyl sulfate (DMS), selective 2'-hydroxyl acylation analysed by primer extension (SHAPE) reagents and hydroxyl radicals) have been successful in penetrating cellular membranes, acting on an RNA of interest, and thereby being used to determine RNA secondary structures\textsuperscript{55, 57, 62, 67, 68}. DMS and SHAPE reagents actively modify single-stranded or flexible regions of target RNAs, while Pb\textsuperscript{2+} and hydroxy radicals cleave the RNA at these regions. Pb\textsuperscript{2+} and hydroxyl radicals have been used to determine the secondary structure of small RNAs, while DMS and SHAPE have been used for the secondary structure determination of longer RNAs such as the HIV RNA genome and the SRA and HOTAIR IncRNAs\textsuperscript{10, 53, 54, 67, 71}. These methods provide us with the ability to assess RNA structure in cells; however, they target single molecules only and still rely on
RNA extraction and primer extension, which is problematic for many IncRNAs that are not present in sufficient abundance for structure probing in vivo.

SHAPE is currently the gold standard of RNA secondary structure determination and in theory is applicable to the determination of all IncRNA structures, as it has no length restrictions. However, in practise, obtaining sufficient levels or quantities of RNA for this technique is still a major limitation for its use in IncRNA as it requires either in vitro transcription, which assumes the IncRNA will fold autonomously and correctly, or cellular extraction of the IncRNA requiring a high cellular abundance for sufficient detection.

The incorporation of next generation sequencing (NGS) has seen the advancement of secondary structure determination methods such as SHAPE-seq\textsuperscript{72} or DMS-seq\textsuperscript{73-75} allowing for lower initial RNA quantities (50 pmol of in-vitro transcribed RNA (SHAPE-seq\textsuperscript{76}) compared to 600 pmol (SHAPE\textsuperscript{67}) or 500 ng of total cellular input RNA (DMS-Seq\textsuperscript{74}) compared to 1-5 μg (DMS\textsuperscript{62})) with increased resolution. In the future, methods that address the ability to produce significantly longer RNA transcripts in vitro or couple in vivo RNA chemical modification with NGS to increase the detection sensitivity of low abundant RNA, would mean that SHAPE has the potential to be a highly effective secondary structure determination technique for IncRNA.

2.4. High-throughput / genome wide

The approaches described above are limited to characterising a single target RNA structure; however, with thousands of new IncRNAs being annotated each year, high-throughput structural characterisation techniques are being developed. Parallel analysis of RNA structure (PARS)\textsuperscript{77} and Fragmentation sequencing (FragSeq)\textsuperscript{78} use enzymatic hydrolysis coupled to NGS to analyse the RNA secondary structure content of entire
transcriptomes (Reviewed in 79). FragSeq has been successfully used to analyse the structural content of small nucleolar RNAs in mouse cells78; however, the technique isolates RNAs up to 100 bases long after nuclease digestion and so many larger RNAs may be under-represented. The use of PARS uncovered a number of general features for yeast mRNAs77 and also has the ability to probe structures of long RNAs. An extension of the PARS method incorporating temperature elevation (PARTE) provided a genome wide measurement of RNA folding energies. Diverse RNA structures and functionally important elements were distinguished and identified in the yeast transcriptome15. PARTE was able to discriminate between classes of RNAs through their melting temperatures and determined that lncRNAs are a stable species which appear to contain a level of structure in between the structural content of the ribosome and a typical mRNA.

2.5. 3-dimensional

3-dimensional (3-D) structural solution is the ultimate goal in macromolecular structure determination as it provides atomic resolution data at the molecular level. 3-D RNA structure determination techniques include small angle x-ray scattering (SAXS), nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography and electron microscopy. For small RNA molecules, 2-D mapping techniques have been almost completely superseded by these atomic resolution approaches; however, they are often labour intensive and typically rely on milligram quantities of robust and homogenous in vitro transcribed RNA that is often difficult to obtain for lncRNAs80. As such, with the exception of rRNA of which large quantities were readily available, there has been no atomic resolution 3-D structure solution of an entire lncRNA to date.

The various methods used to predict/determine structures of lncRNA all have the same major limitation: obtaining sufficient quantities of robust and homogeneous samples for
very long (over 3 kb) and low abundant lncRNAs. This is particularly problematic for 3-D structural characterisation methods. A limitation of the high-throughput techniques is that the RNAs are extracted from cells and renatured, thus the resulting structures may not be an accurate representation of the in vivo structures. Whilst there are now biochemical reagents that can probe RNA structure in living cells, such as SHAPE and DMS, these methods have not yet been adapted to deep sequencing platforms limiting their high-throughput or genome-wide applications. Nevertheless, various combinations of these techniques have been used successfully in characterising various structural elements and probing the entirety of some lncRNAs, which we have described below.

3. Examples of lncRNA structures
Recently, Watts et al, used SHAPE to interrogate the entire 9173 nt HIV-1 RNA genome identifying new functional elements and long-range RNA interactions\textsuperscript{54}. Structures in single-stranded RNA viral genomes are critical to viral replication and host defence evasion. Analysis of the HIV-1 genome structure identified that unstructured protein loops are derived from highly structured RNA elements, suggesting the existence of another level of the genetic code. Additionally, the HIV-1 genome structure revealed that the regions which have high sequence variation that are a critical component for viral host evasion are fundamentally unstructured, while the stable structure regions are evolutionarily conserved. This provides insight into the mechanism of genetic diversity in ssRNA viruses. The HIV-1 genome structure showed the potential for an astonishing amount of information being encoded into the structures of large RNA molecules and makes it apparent that there is much to be gained from the global structural analyses of RNA.
The first experimental characterisation of the secondary structure of an entire human IncRNA, was for the steroid receptor RNA activator (SRA)\textsuperscript{10}. The authors used a combination of chemical and enzymatic approaches to map the entire secondary structure of the 870 nt IncRNA. The structure determination revealed that the SRA consisted of four major domains organised into a complex secondary structure and, whilst being short enough for mechanistic studies it was long enough to encompass many of the structural characteristics commonly found in rRNA\textsuperscript{10}. Structure determination of SRA in combination with previous deletion and mutational analysis provided evidence for IncRNA function requiring the complex structural organisation of the entire molecule, rather than being limited to one particular substructure\textsuperscript{10,81}.

Very recently, the secondary structure of the entire 2.2 kb IncRNA HOTAIR (HOX transcript antisense intergenic RNA) was experimentally determined\textsuperscript{53}. The structural map obtained by using a combination of chemical probes shows that HOTAIR, like SRA, folds into four independent domains. Two of the four sub-domains correspond to predicted protein-binding domains with high conservation displayed in the secondary structure elements surrounding protein-binding motifs. Validation of the map with phylogenetic analysis identified potential key functional regions through highly conserved secondary structure elements.

The secondary structure determination of SRA and HOTAIR provides exciting possibilities for the structural determination of other human IncRNAs of similar length; however, the RNAs for these studies were produced via \textit{in vitro} transcription: a method that is generally not applicable for IncRNAs of lengths over 3 kb. With the many challenges of determining the entire structure of IncRNA meaning that very few IncRNAs have been characterised,
Researchers often settle for the high-resolution structure determination of smaller fragments.

For example, the NMR structure of a 32-nucleotide (nt) fragment from the 0.4 kb human telomerase RNA (hTR; also called telomerase RNA component (TERC)) showed a stable hairpin containing an asymmetric internal loop in a domain that was important for telomerase activity\(^\text{82}\) (Figure 2A). The twist in the RNA structure caused by the loop provides a key functional feature by generating a docking site for the enzymatic protein component of the telomerase, hTERT as well as identifying possible sites for drug targeting. Other examples include the NMR structural solution of a 14 nt hairpin from the 17 kb Xist IncRNA involved in X-chromosome inactivation, identifying a novel AUCG tetraloop. The tetraloop is unusual in that there is no base-pair between the first and last nucleotide, a turn is introduced between the first two nucleotides rather than the second and third nucleotide, and the final nucleotide of the tetraloop is solvent exposed which is a rare and unusual observation (Figure 2B). Further experiments showed that this novel tetraloop was important for X-chromosome inactivation\(^\text{83}\). A combined crystallography and NMR approach elucidated the molecular mechanism for the binding of a 20 nt stem loop region within the 600 nt long intergenic non-coding RNA (lincRNA) growth arrest-specific 5 (Gas5) to steroid receptors (SRs)\(^\text{84}\) (Figure 2C). The double stranded stem region of Gas5 competes directly with DNA for the DNA binding domain of the SRs to inhibit their transcriptional activity; and a crystal structure characterising a triple helix from the 8 kb metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) IncRNA that is critical for the prevention of nuclear decay identifies it as a potential target for reducing MALAT1 levels in cancer cells\(^\text{85}\) (Figure 2D).
With the exception of the ribosome, as yet there are no atomic resolution structures of any complete lncRNAs thus the intricacies of tertiary interactions within lncRNA remain a mystery. However, the atomic resolution structures of large RNA molecules illustrate that RNA molecules can display a level of versatility and complexity comparable to proteins. For example, the structure of the 155 nt HIV-1 RNA packaging signal, solved by NMR, adopts a tandem three-way junction. This single RNA conformation is able to inhibit translation and selectively promote the packaging of its unspliced genome during the viral replication of HIV. Formation of the junction results in the sequestration of splice and translation initiation sites and exposes weakly paired or unpaired guanosines enabling their binding to the HIV proteins directing genome packaging (Figure 3A)\(^66\).

The crystal structure of a bacterial ribonuclease (RNase) P, a 346 nt enzyme that processes the 5’ leader sequence of precursor tRNA molecules, is an example of an RNA catalyst which constitutes multiple domains - one for substrate recognition and the other forming the scaffold for the active site that requires the coordination of divalent metal ions for activity\(^67\) (Figure 3B). The structure identified the location of the active site, the requirement of the protein cofactor in tRNA catalysis and revealed that the enzyme uses shape selective recognition and specific RNA-RNA contacts such as base stacking, A-minor interactions and canonical base-pairing to recognise its tRNA substrate\(^67\).

Finally the crystal structure of the ribosome, the longest 3-D structure of RNA to be solved to date, illustrates the ability of long RNA molecules to fold into multiple and distinct functional domains. The highly organised RNA of the ribosome is largely an arrangement of helical elements held together by the coaxial stacking between helices, stabilising long-range interactions of single stranded regions, and protein binding (Figure 3C)\(^68\).
4. IncRNA:protein interactions

Interacting molecules can have an impact on RNA structure and a major recurring theme in IncRNA biology is its ability to function through the recruitment of protein factors. LncRNA:protein complexes function to alter the chromatin state, organize nuclear substructures, and regulate gene expression; however, the mechanisms of these important cellular processes are still largely unknown. In comparison to other classes of RNA, a feature of IncRNA is its propensity to contain a huge number of protein-binding sites. Thus IncRNA can provide a platform for the multimerisation of proteins, such as the Xist RNA for Rev proteins\textsuperscript{83}, or can provide a scaffold for the assembly of multiple proteins into larger functional units, such as NEAT1 in the paraspeckle\textsuperscript{89-91}. Critical toward understanding the function of IncRNA is the understanding of how they interact with proteins.

The recognition of small RNAs by RNA-binding proteins (RBPs) is through distinct structural elements, RNA-binding domains (RBDs) that are responsible for the specific recruitment of RNA to the target protein\textsuperscript{92-95}. The RNA recognition motif (RRM) constitutes the most abundant RBD and is hence the most extensively studied. Other common RBDs include the heterogeneous ribonucleoprotein particle homology (KH) domain, zinc-fingers (ZnF), arginine rich motifs (ARMs), and the RGG box. The comprehensive structural analyses of RBDs indicate that they primarily recognise short ssRNA sequences; however, other than occurring in single stranded stretches, these regions can occur in RNA secondary structures such as hairpin loops or in bulges. There are examples of proteins binding to double-stranded (ds) RNA stems and while these proteins seem to recognise the dsRNA shape, they also bind to specific dsRNA targets indicating some element of sequence specificity\textsuperscript{96-101}. There has been a multitude of RNA:protein structures contributing to the characterisation of how these common RNA binding domains recognise
specific RNA molecules, providing us with a detailed understanding of their mechanism of interaction (Figure 4).

Unlike ribosomal proteins, which interact with the rRNA through electrostatic interactions with the sugar phosphate backbone\(^1\)\(^2\), there is evidence that many IncRNAs bind their proteins partners in a canonical manner. Characterised examples include LARP7 (La-related proteins group 7) that uses its La domain and two RRM domains to bind a stretch of U’s at the 3’ end and the apical loop of the 3’ hairpin of the 331 nt 7SK RNA\(^3\) (Figure 5A); and, the ARMs of the viral Rev protein homodimer bind with high specificity to the major grooves along the two legs of the 233 nt viral Rev-response element (RRE) IncRNA\(^4\)-\(^7\) (Figure 5B). Further to these characterised structures, domain analysis shows that many IncRNA-binding proteins (IncRBPs) contain these common RBDs and in multiple copies. For example, the Drosophila behaviour human splicing (DBHS) proteins known to interact with NEAT1 contain two RRMs\(^8\),\(^9\), heterogeneous nuclear ribonuclear protein K (HNRNPK) which interacts with both lincRNA-p21 and NEAT1 contains three KH domains and an RGG box\(^10\),\(^11\), and the myriad of proteins shown to co-localise to the paraspeckle contain multiple copies and combinations of canonical RBD’s\(^11\).

In contrast to canonical recognition, other IncRBPs such as the polycomb repressor group complex (PRC) proteins do not contain characterised canonical RBD’s but are known to interact with multiple IncRNAs\(^12\)-\(^16\). This suggests that they interact with IncRNA through non-canonical backbone interactions, in a similar manner to ribosomal proteins, or they could display unique methods of interactions, such as Gas5 IncRNA binding to the DNA-binding domain of steroid receptors\(^8\). Further characterization of these IncRNA:protein
assemblies is needed in order to elucidate the molecular mechanisms behind these non-canoncal protein:RNA interactions.

5. IncRNA as scaffolds for many proteins.

Many IncRNAs interact with multiple different proteins suggesting that they may act as scaffolds to amalgamate protein complexes into larger functional units. The most well characterised example of this is the ribosome, of which two-thirds is made up of non-coding RNA. The RNA forms two subunits, which in eukaryotes consist of approximately 86 proteins interspersed throughout a 1.9 kb and a 4.7 kb IncRNA and two shorter RNAs (120 nt and 160 nt). Early structural mapping of rRNA using cross-linking and chemical foot-printing methods provided initial physical constraints for building models of the ribosome\textsuperscript{117-120} and the initial two subunit morphology of the ribosome was determined by negative-stain electron microscopy. Early cryoEM studies began to provide some details about the mechanism of action\textsuperscript{121, 122} including evidence of the structural dynamics\textsuperscript{123}, however, it was the high resolution crystal structure of the ribosome that provided fundamental insight into its mechanism, paving the way for more complex biochemical and genetic experiments to further evolve our understanding of mechanisms and regulation of ribosome function.

The ribosome represents a highly structured and rigid scaffold that could be used for structural studies due to its high cellular abundance and robustness allowing it to be purified intact from cells. Other IncRNAs are suggested to form flexible scaffolds that are not so abundant or robust, making them significantly harder to characterise. The yeast telomerase RNA, Xist and HOTAIR have all been suggested to be IncRNA scaffolds that tether a multitude of proteins together\textsuperscript{124-127}. In the case of HOTAIR it is proposed that the
IncRNA may serve as a scaffold by providing a binding surface to assemble and tether two separate histone modification complexes\textsuperscript{127}. For the yeast telomerase, large stretches of the RNA could be deleted without significant loss in function suggesting that the RNA serves primarily to tether the protein binding subunits together\textsuperscript{128}, and a similar phenomena, whereby significant segments of the RNA are dispensable for function, has been seen for Xist\textsuperscript{129}. The mechanism for how these IncRNA:protein complexes are assembled is currently unknown; however, in light of the recent secondary structure determination of HOTAIR, combined with more targeted biochemical and structural experiments, a detailed molecular understanding of its assembly mechanism may soon be revealed.

Flexible IncRNA scaffolds can also play a role in the assembly of nuclear bodies. There are several types of non-coding RNAs that are physiologically enriched within various types of nuclear bodies such as Cajal bodies, nuclear speckles, paraspeckles and nuclear stress bodies (reviewed in \textsuperscript{130-133}). In several cases, formation of these bodies occurs through a nucleating RNA, resulting in the recruitment and accumulation of associating proteins. The best-studied IncRNA involved in the formation of nuclear bodies is NEAT1’s nucleation of paraspeckles.

Paraspeckles are a dynamic, ribonucleoprotein body found in mammalian cell nuclei consisting of the 23 kb NEAT1 transcript (NEAT1\_v2), and a shorter 3.7 kb isoform (NEAT1\_v1)\textsuperscript{111,134}. These two isoforms provide a scaffold for the recruitment of in excess of 40 nuclear RNA binding proteins\textsuperscript{89-91}. Unlike the IncRNAs Gomafu or MALAT1, which are both enriched in, but not required for the formation of gomafu speckles and nuclear speckles respectively\textsuperscript{90,135}, paraspeckle biogenesis is dependant on the transcription of
NEAT1 and requires seven essential paraspeckle proteins HNRNPK, NONO, RBM14, SFPQ, DAZAP, FUS and HNRNPH3$^{30, 91, 111, 136}$. The macromolecular structure of the paraspeckle is yet to be elucidated; further, it is unclear as to the stoichiometry of paraspeckle proteins and NEAT1 RNA within a single paraspeckle.

Electron microscopy studies suggest that NEAT1 accumulates in paraspeckles to form a regular network of RNA fibres with NEAT1_v1 and the 5’ and 3’ ends of NEAT1_v2 localising to the periphery of the paraspeckle, and the central sequence of NEAT1_v2 localising to the inner core$^{137}$. The macromolecular arrangement of NEAT1 and the paraspeckle proteins is currently unknown. The scaffold of the paraspeckle could be formed via RNA-RNA interactions between NEAT1 transcripts; alternatively, multiple NEAT1_v2 transcripts could be connected by means of interaction with the paraspeckle proteins. This is highly likely given that essential DBHS paraspeckle proteins, PSPC1, NONO and SFPQ have the propensity to form dimers and oligomers$^{108, 138}$, and that many paraspeckle proteins are required for paraspeckle formation without being required to maintain NEAT1 levels$^{139}$. In this latter case, protein-protein interactions may be bridging RNA-protein interactions to form a paraspeckle (Figure 6).

The oligomerisation of the essential DBHS paraspeckle proteins could provide key insight into the mechanism of macromolecular assembly of paraspeckles. Recent structural characterisation of SFPQ revealed a remarkably extended coiled-coil through which extensive interactions mediated infinite polymerization of SFPQ homodimers$^{108}$. The extended coiled-coil is a feature of the DBHS proteins and their ability to polymerise through this domain was proven to be critical for paraspeckle formation and localization of SFPQ and PSPC1 to paraspeckles. The oligomerisation of DBHS proteins through the coiled-coil brings together their globular functional domains, such as the RRM which are
the proposed sites of NEAT1 interaction. Thus, one model for the assembly of paraspeckle complexes is through the cooperative oligomerisation of DBHS proteins on the NEAT1 scaffold utilizing both protein-protein and protein-RNA interactions\textsuperscript{108} (Figure 5). A similar mechanism of protein oligomerisation on lncRNA is seen for the formation of the Rev-RRE complex formation\textsuperscript{140}. It is unknown what defines the composition and architecture of paraspeckles but interestingly, the polymerisation domains of DBHS homo- and heterodimers were shown to reversibly associate and dissociate. Comparable to the network of coiled-coil interactions Lee et al. observed for the DBHS proteins is the reversible and functional fibrillar-like aggregates of the low complexity domains found in many nucleic acid binding proteins, including several paraspeckle proteins\textsuperscript{141}. These data suggest a mechanism for the dynamic formation and dissociation of the paraspeckle through an equilibrium in and out of the functional oligomerised or aggregated state of paraspeckle proteins.

The paraspeckle represents a unique and intriguing lncRNA:protein assembly of which there are many unanswered questions surrounding the details of its macromolecular structure. Techniques such as CryoEM or cryoelectron tomography are ideal to study the structure of these micron sized nuclear bodies, yet the extreme length and low cellular abundance of NEAT1 makes \textit{in vitro} transcription or cellular extraction difficult and the dynamic nature of paraspeckles renders them challenging to isolate. The difficulties in the structural determination of paraspeckles are representative of the challenges associated with many large lncRNA:protein complexes, and highlight the requirement for technological advancements in current structural biology techniques.

The mechanism for lncRNA:protein complex assembly is unknown, and the task of interrogating these mechanisms requires detailed knowledge of the lncRNA structure and
of the individual RNA:protein interactions. Interacting proteins need to be identified, the RNA binding specificity of the individual proteins characterised, protein binding sites on the RNA identified, biochemical and structural characterisation of the individual protein:RNA complexes, and the global structure of RNA determined. A task such as this is daunting given our current technological limitations, yet necessary if we want to ascertain the molecular mechanisms behind lncRNA function.

6. Discussion

LncRNAs have emerged as essential regulators in almost all aspects of biology and are being found to play a role in a variety of disease states. LncRNAs display greater tissue specificity compared to other classes of RNA making them desirable candidates in the quest for diagnostic markers; therefore, an improved understanding of the role of lncRNA would aid pharmaceutical design and novel therapeutic strategies. Structural biology is of critical importance for future studies on lncRNAs through building the necessary groundwork to interpret their functions, study the effect of mutations on lncRNA function and elucidating their direct connection to disease.

The overall architecture of lncRNA remains unknown; however, similar to the domain configuration seen in the ribosome, determining the secondary structure of the SRA and HOTAIR lncRNA revealed the presence of distinct domains within the sequences. With the possibility that all lncRNAs consist of multiple structural or functional domains, it follows that each domain could be studied independently, comparable to the approach used for multidomain proteins or large protein complexes. By focusing on individual domains, lncRNAs could be separated into more manageable lengths for atomic resolution studies. Thus perhaps a more realistic way to interrogate lncRNA structures in the future
would be in first characterizing the secondary structure of IncRNA in order to identify the functional and structural regions to target for tertiary structure determination.

Tertiary nucleic acid structures are publically available through the PDB and the Nucleic Acid Database (NDB)\textsuperscript{143, 144}, and their secondary structure information is able to be extracted for visualization and manipulation through tools such as RnaMLView\textsuperscript{145}. However, there is no similarly available database for the sharing of RNA secondary structure information, such as the structural annotations of the HOTAIR or SRA IncRNA maps. As more secondary structures of IncRNAs are experimentally determined, the generation of a centralized depository of RNA secondary structure annotations will be important for the dissemination of the data. An obvious choice would be to incorporate RNA secondary structure maps into the NDB, which is already a hub for RNA tertiary structures, as a sequence and secondary structure annotation file. Another candidate system for the distribution of RNA secondary structure data could be the inclusion of experimentally determined RNA secondary structure tracks into the UCSC genome browser\textsuperscript{146}; analogous perhaps to the work of Smith et al.\textsuperscript{13} whose evolutionary conserved RNA secondary structure predictions can be visualized on the human genome browser via a public track hub (www.martinalexandersmith.com/ecs/). The RNA secondary structures annotated at nucleotide resolution are then easily extracted as input into RNA folding software.

While there are many available 2-D and 3-D techniques to study the structures of proteins and small RNAs, they have not yet been sufficiently adapted for the IncRNA field. Size restrictions are still the major limitation of most available techniques. Out of the 100,000+ annotated IncRNAs, only two have experimentally determined secondary structures. Thus, it is obvious that new technological advances are required for the ability to produce
long RNA transcripts through either *in vitro* or *in vivo* methods in order to obtain high-resolution structural characterisation of IncRNAs and their binding partners. CryoEM is revolutionizing the structural biology of macromolecular protein assemblies. It negates the need for crystals, requires minimal sample and can differentiate biochemical and conformational heterogeneities of the samples. Indeed it has been predicted that cryoEM will largely supersede crystallography\textsuperscript{147}. The future of IncRNA structural biology will most likely require a combined approach of high-resolution global imaging techniques such as cryoEM in combination with 2D structural determination methods such as SHAPE. Structural maps can then be used to guide detailed biophysical explorations in order to obtain a complete overview of the molecular functions and mechanisms of the new frontier of molecular biology.

**Acknowledgements**

This work was supported by the National Health and Medical Research Council (NHMRC) grant numbers 1048659 and 1050585.

**References**


33


Figure legends

Figure 1. Common secondary and tertiary structures of RNA. A. The dot-bracket notation and 2-D map of a small region from the 5’ end of B2-SINE IncRNA showing common secondary structure elements. B. 2-D and 3-D representations of RNA tertiary structures. A double-stranded region of the expression and nuclear retention element (ENE) sequesters the RNA poly (A) tail (magenta) from Kaposi’s sarcoma-associated herpesvirus using Watson-Crick and Hoogsteen interactions to form a stabilising triple helix (PDB entry 3P22\textsuperscript{148}). A kissing loop between the stem-loop substrate and stem-loop catalytic domain by the Neurospora Varkud satellite ribozyme (PDB entry 2MI0\textsuperscript{149}). The three Watson-Crick base-pairs responsible for the interaction between the two loops are indicated with a dotted line. A classical pseudoknot from the turnip yellow mosaic virus showing the reciprocal interaction between two stemloops where the loop of each stemloop forms the stem of the other (PDB entry 1A60\textsuperscript{150}). All tertiary structures are shown as cartoon representations coloured by element. Wobble and non-Watson-Crick base-pairs are shown in the 2-D maps as unfilled and filled circles respectively.

Figure 2. Atomic resolution structures of IncRNA fragments. A. The NMR structure of a 32-nucleotide (nt) hairpin from the human telomerase IncRNA adopts a canonical A-form RNA structure within double helical stem region capped by a typical UUCG tetraloop (PDB entry 1Z31\textsuperscript{82}). The hairpin contains an asymmetric internal loop introducing a twist in the RNA structure. The loop provides a key functional feature through the generation of a docking site for the enzymatic protein component of the telomerase, hTERT. B. A novel AUCG tetraloop in a hairpin from the IncNRA Xist is an important functional element (PDB entry 2Y95\textsuperscript{83}). C. A double stranded helix from the IncRNA Gas5 (PDB entry 4MCE\textsuperscript{84}).
contains a widened major groove width of 15 Å in comparison to a typical A-form RNA helix that has a major groove width of 3.4 Å (PDB entry 1RNA\textsuperscript{151}). The widened major groove of Gas5 mimics DNA to compete for the binding of steroid receptors. D. The crystal structure of a triple helix element at the 3’ end of the IncRNA MALAT1 is critical for the prevention of nuclear decay (PDB entry 4PLX\textsuperscript{85}). All tertiary structures are shown as cartoon representations coloured by element. Wobble and non-Watson-Crick base-pairs are shown in the 2-D maps as unfilled and filled circles respectively.

**Figure 3.** Atomic resolution structures of large RNAs. A. The secondary and NMR structure of the HIV-1 RNA packaging signal (PDB entry 2N1Q\textsuperscript{86}) forming a tandem three-way junction. Brackets indicate each junction and guanosine residues exposed by the junction are coloured red. B. The crystal structure of a bacterial RNase P (PDB entry 3Q1R\textsuperscript{87}). The metal ions are indicated and shown as grey spheres and the tRNA substrate is coloured magenta. C. The crystal structure of the 30S subunit of the ribosome from *Thermus thermophilus* (PDB entry 1J5E\textsuperscript{88}). RNA tertiary structures are shown as cartoon representations coloured by element, and the surface representations of protein structures shown in light blue. Wobble and non-Watson-Crick base-pairs are shown in the 2-D map as unfilled and filled circles respectively.

**Figure 4.** Structures of RNA binding motifs bound to their cognate RNA. Nucleic acids are coloured by element and shown in the cartoon representation. Both the surface and cartoon representation of the proteins are shown (light blue) with the nucleic acid interacting residues coloured purple and shown as sticks. A. The RRM of Human Fox-1 binds to UGCAUGU in a canonical way through three conserved aromatic residues on the four-stranded B-sheet surface. (PDB entry 2ERR\textsuperscript{152}) B. The dsRBM of ADAR2
recognises double stranded RNA in the form of a stem-loop (PDB entry 2L3C\textsuperscript{155}).  C. The KH domain of the Nova protein recognises an RNA stem loop (PDB entry 1EC6\textsuperscript{154}).  D. The zinc finger of ZRANB2 binds to an AGGUAA motif through side-chain hydrogen bonds and the formation of a guanine-tryptophan-guanine “ladder” (PDB entry 3G9Y\textsuperscript{153}).

**Figure 5.** Structures of IncRNA-protein interactions. Nucleic acids are coloured by element and shown in the cartoon representation. Both the surface and cartoon representation of the proteins are shown (light blue) with the nucleic acid interacting residues coloured purple and shown as sticks. A. The La-domain and RRM1 of LARP interaction with a U-rich hairpin from 7SK IncRNA (PDB entry 4WKR\textsuperscript{103}). Amino acids from the RRM that are involved in RNA interactions are shown as sticks. B. The ARMs in the viral Rev protein homo-dimer bind to the Rev-response element (RRE) IncRNA (PDB entry 4PMI\textsuperscript{156}).

**Figure 6.** Model for the assembly of a paraspeckle. NEAT1\_v1 and the 5’ and 3’ ends of NEAT1\_v2 localise to the periphery of the paraspeckle with the central sequence localising to the inner core. DBHS proteins oligomerise along NEAT1 through protein-protein and protein-RNA interactions while the other paraspeckle proteins may be bridging protein-protein or protein-RNA interactions to form a paraspeckle.