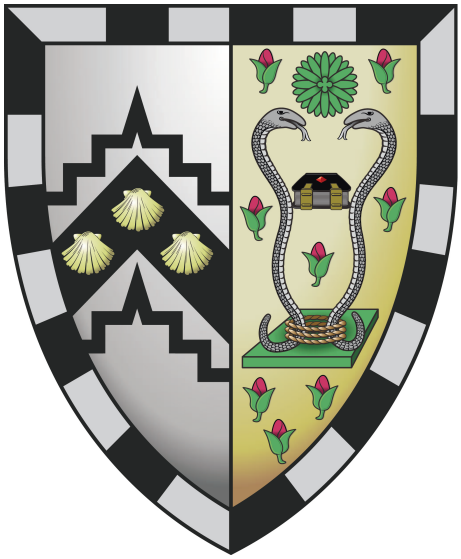
UNIVERSITY OF CAMBRIDGE

Investigation of snoRNA clusters in the imprinted *Dlk1-Dio3* and *Snrpn* regions and their role in splicing



Ujjawal Aditya Kumar

Group Leaders:

Dr. Russell S. Hamilton
Professor Anne C. Ferguson-Smith

Part II Natural Sciences Tripos — Genetics
Department of Genetics, University of Cambridge

Gonville & Caius College

2020-2021

Natural Sciences Tripos
Part II Genetics
Research Project Report — 2020/21

TITLE PAGE

Full Title:	Investigation of snoRNA clusters in the imprinted *Dlk1-Dio3* and *Snrpn* regions and their role in splicing
Group Leader:	Dr. Russell S. Hamilton, Professor Anne C. Ferguson-Smith
Word Count:	3496

Feedback Meeting

Name of the person who provided feedback on your draft project report:	Dr. Russell S. Hamilton
Date of this feedback meeting:	15th March 2021

Table of contents

List of figures

List of tables

Declaration

I declare that this Project Report is entirely my own work except where otherwise stated, either in the form of citation of published work, or acknowledgment of the source of any unpublished material.

_____

Ujjawal Aditya Kumar

28th March 2021

Acknowledgements

First and foremost, I would like to express my immense gratitude to Dr Russell Hamilton for all of his advice, help, and encouragement throughout this project. He was always incredibly helpful in answering my many questions, and I am extremely grateful for the time he gave to our discussions. I have learnt a tremendous amount about the vast field of structural bioinformatics(a field that I had minimal prior knowledge of), thus gaining a new perspective to approach my interest in the developmental mechanisms of disease. For this I am greatly indebted to him.

I would especially like to thank Professor Anne Ferguson-Smith for kindly hosting my project within her lab, and all of her guidance and insight on my project. I am most grateful to my fellow students: Aurily Constantino of the Hamilton Group for her practical advice on working with the snoRNAs and for providing their sequences; as well as Martin Limbäck-Stokin, Jemima Becker and Samuel Lloyd of the Ferguson-Smith Lab for all our discussions about our respective projects, science and much more. I would like to express my sincerest thanks to my advisor, Dr David Summers, for his advice and help throughout the year on

all matters relating to Part II Genetics. Thanks are also due to the other members of the Ferguson-Smith Group, for being so welcoming and friendly.

I would specifically like to thank my friends, Ruweena, Emilia, Jacques, Natasha, and especially Saffron, for all of their support; though we unfortunately couldn't spend time together in person, our conversations always put a smile on my face. Lastly, I am eternally grateful to my parents and my sister, Sneha, for always being so loving and supportive. Their advice and encouragement were, and always will be invaluable.

Abstract

The epigenetic phenomenon of genomic imprinting leads to specific genes being expressed in a parent-specific manner. Imprinted genes (IGs) have key roles in placental and embryonic development as well as an emerging role in brain development and function. The imprinted *Dlk1-Dio3* region contains two maternally expressed small nucleolar RNA (snoRNA) clusters, SNORD113 and SNORD114, containing nine and thirty-one copies respectively (Rocha et al., 2008). Human individuals with disruptions in this conserved region display growth retardation and facial dysmorphism (Kagami et al., 2008).

Deletion of the related SNORD116 cluster (paternally expressed) in the *Snrpn* imprinted region has been shown to be a direct cause of Prader-Willi syndrome (PWS) and postnatal growth retardation, (PNGR), (Sahoo et al., 2008; Skryabin et al., 2007). However, little is known about the mechanism of action of these snoRNAs.

Using open-source computational tools, I will model the snoRNA 2D and 3D structures of these regions (SimRNA – Boniecki et al., 2016; ViennaRNA Web Services – Gruber et al., 2008; Vienna RNA – Lorenz et al., 2011). Transcribed as a precursor long non-coding RNA, Rian/MEG8, the C/D box motif-containing snoRNAs of *Dlk1-Dio3* have been shown to be

directed by genomic imprinting and show brain-specific expression, unusually for snoRNAs. (Cavaillé et al., 2000; Rogelj, 2006).

Chapter 1

Introduction

1.1 Imprinting

In sexually reproducing organisms, each parent contributes chromosomes equally to the zygotic genome during fertilisation. For most genes, both parental copies are expressed. However, some genes in placental mammals show monoallelic expression through the epigenetic phenomenon of "genomic imprinting" (alleles are expressed/silenced depending on their parental origin). Notable examples include *Igf2* and *Igf2r*. As discussed in my Literature Review, imprinting is thought to have evolved as a means of regulating maternal-fetal resource balance via the placenta.

There are two main theories for its evolution:

- Kinship hypothesis – maternally inherited genes and paternally inherited genes have conflicting interests.
- Coadaptation hypothesis – imprinted genes act to optimise social interactions, specifically maternal-offspring interaction.

By demonstrating that androgenetic or parthenogenetic conceptuses were non-viable, Barton, Surani and colleagues (1984) showed the functional inequality of some parental chromosomes, suggesting the concept of mammalian imprinting. Studies of humans and other species (notably *Mus musculus*) show the influence of imprinting on developmental processes affecting growth and tissue differentiation (Rocha et al., 2008).

Disruptions to imprinted regions like Dlk1-Dio3 and Snrpn (found on Chr. 14 and 15 of the human genome) are associated with neurological disorders like PWS and AS, suggesting activity in the nervous system.

1.2 snoRNAs

As discussed in my Literature Review, *Dlk1-Dio3* and *Snrpn* encode snoRNAs within their intergenic regions (Figure 1.1), which are released via exonucleolytic trimming (Filipowicz and Pogačić, 2002).

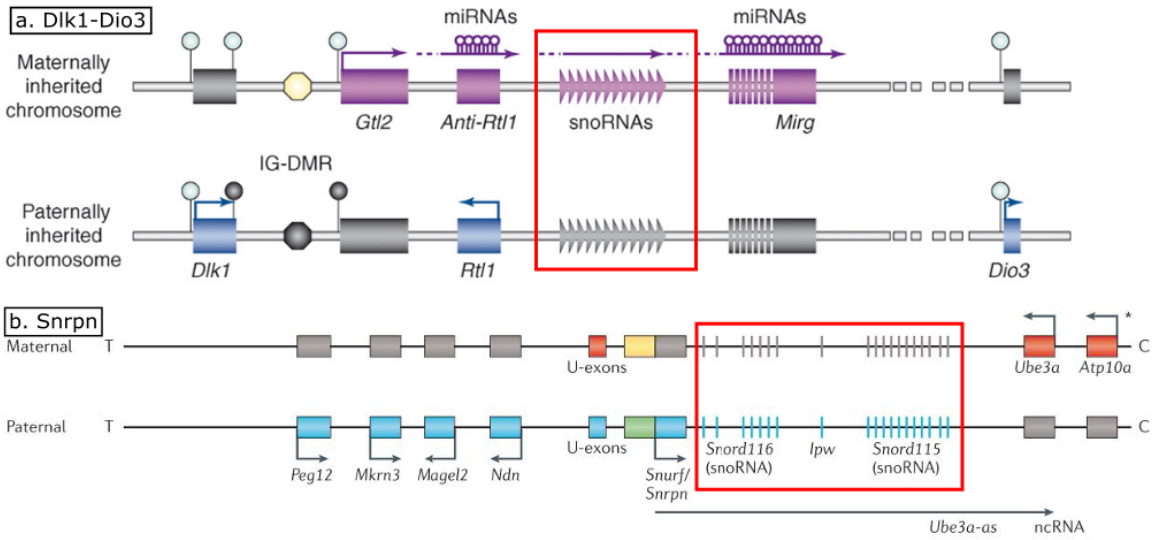


Fig. 1.1 Schematics of the 2 imprinted regions of interest: Dlk1-Dio3 (H. sapiens Chr14, M. musculus Chr12) and Snrpn (H. sapiens Chr15, M. musculus Chr7), with the snoRNA clusters shown specifically. **(a.)** In blue, genes expressed from the paternal chromosome and in purple, noncoding RNAs expressed from the maternally inherited chromosome, including the snoRNA clusters of interest SNORD113 and SNORD114. The imprinting control region for the domain is IG-DMR; found methylated in the paternally inherited chromosome (black circle) and unmethylated in the maternal chromosome (yellow circle). **(b.)** In blue, the paternally expressed genes of the Snrpn:Ube3a region, including the snoRNA clusters of interest SNORD116 and SNORD115, and in red, the maternally expressed genes Ube3a and Atp10a. Figures adapted from Peters, 2014 [Snrpn]; Rocha et al., 2008 [Dlk1-Dio3].

These snoRNAs are implicated in human diseases and disease phenotypes in mice. Particularly, aberrant Dlk1-Dio3 snoRNAs are linked to developmental defects like PNGR, a "bell-shaped thorax", and facial dysmorphisms (Kagami et al., 2008); disruption to the Snrpn snoRNAs (SNORD116 specifically) is known to be the causative factor in PWS pathogenesis (Bieth et al., 2015; Cavaillé, 2017; Sahoo et al., 2008), through an unknown mechanism.

There are four classes of snoRNAs (Figure 1.2), as discussed in Liang et al. (2019); these snoRNAs of interest fall into the "orphan" class.

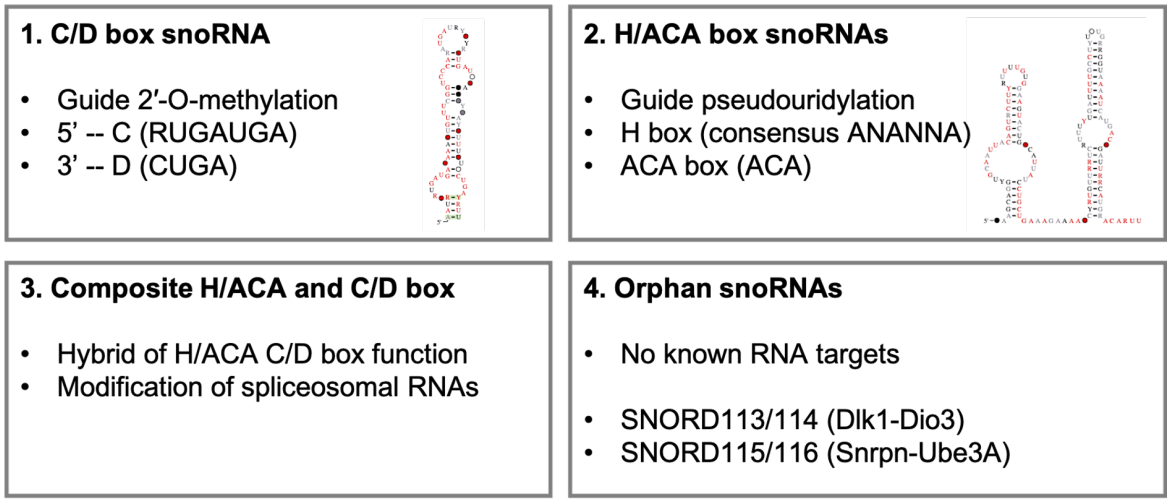


Fig. 1.2 The four classes of snoRNAs. The snoRNA clusters of interest are of the "orphan class" whose mechanism of action and structure is as yet unknown since they have no known RNA targets and do not show C/D boxes, or H/ACA box structural motifs. Thus, through this project I hope to explore their structures and further develop hypotheses for their mechanism of action.

There are a range of hypotheses for how these snoRNAs act. Most non-coding RNAs (ncRNAs) act via sequence complementarity to their targets. However, the only sequence complementarity found (computationally [Bazeley et al., 2008] using *snoTARGET* but not experimentally verified) for these snoRNAs is an 18-nucleotide motif in an alternative exon of the serotonin-2C-receptor [5-HT_{2C}R] (Cavaillé et al., 2000). This does not, however, explain the snoRNAs' roles in pathogenesis, since the 5-HT-R is in no way implicated. This link to alternative splicing is supported by the findings of Kishore et al., 2010, who found that over ten proteins' (e.g., DPM2, TAF1, RALGPS1, PBRM1, and CRHR1) alternative splicing was linked to snoRNAs in the cluster.

A site-specific base conversion from adenosine to inosine (A-to-I) is fundamental in the synthesis of a functional 5-HT-R. Vitali and colleagues showed that this ADAR2-Figsted conversion is inhibited by snoRNAs of SNORD115, through the formation of a snoRNP particle (Vitali et al., 2005). Thus, one can hypothesise that these snoRNAs are acting via a structural function.

This further supports the findings of Watkins and colleagues that there is structural similarity between some snoRNAs and the U4 spliceosomal snRNA (Watkins et al., 2000), a fundamental regulator of the major spliceosome (Yean and Lin, 1991). Specifically, the U4 spliceosomal snRNA and C/D box snoRNAs both show a Kink-Turn structural motif, which is likely to be important in brain-specific expression and is also seen in the snoRNAs of interest in this project. The presence of the KT motif in C/D box snoRNAs and the U4 snRNA has been shown by Moore et al., (2004) and Wozniak, (2005) respectively.

The KT motif is a short (around fifteen nucleotides), “two-stranded, helix-internal loop-helix” motif. Its first stem is canonical, ending at the internal loop with two Watson-Crick base pairs (typically C-G); the second, non-canonical stem follows the internal loop and starts with two non-Watson-Crick base pairs, typically G-A (Klein, 2001). KT motifs form in the presence of divalent cations such as Ca^{2+} (Matsumura et al., 2003), and therefore are indicative of brain-specific expression, which generally relies on Ca^{2+} sensitivity. When formed, these sharp bending structures force the nucleotides outwards, increasing accessibility for protein factors which can recognise and target RNA hairpins in a sequence-specific manner.

1.3 The spliceosome

The spliceosome is a large ribonucleoprotein complex found primarily in eukaryotic nuclei, and it is responsible for processing of the pre-mRNA transcript, removing introns and ligating exons to one another. It has a cyclic turnover pattern, shown in Figure 1.3.

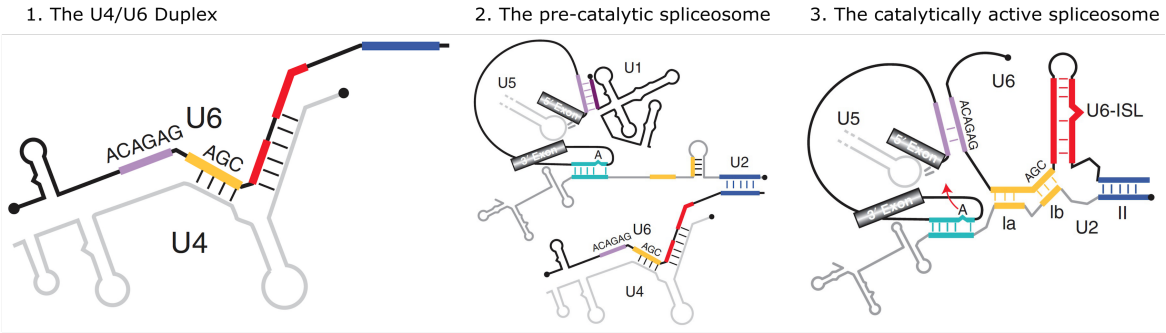


Fig. 1.3 Diagrams showing the 3 main stages of the major, U4-dependent spliceosome, adapted from Will and Lührmann, 2011. The structure at each stage is made up of different combinations of the five different spliceosomal snRNAs (U1, U2, U4, U5, U6). **1**. Complementary base pairing between the U4 spliceosomal snRNA and the U6 spliceosomal snRNA to form a duplex, that holds U6 inactive and thus unable to progress through the spliceosomal turnover. **2**. The pre-catalytic spliceosome is made up of U1, U2, U5 and U6. **3**. The catalytically active spliceosome is comprised of U2, U5 and U6.

Though the catalytically active spliceosome (Figure 1.4) is comprised of U2, U5 and U6, the U4 spliceosomal snRNA has been shown to have a fundamental role in the turnover of the major spliceosome, so much so that it is known as the U4-dependent spliceosome. Degradation of the U4 snRNA has been shown to inhibit spliceosomal function and therefore splicing (Black and Steitz, 1986). Owing to its duplex formation with U6, U4 is key to the correct timing and co-ordination of the different stages of the spliceosome's life cycle.

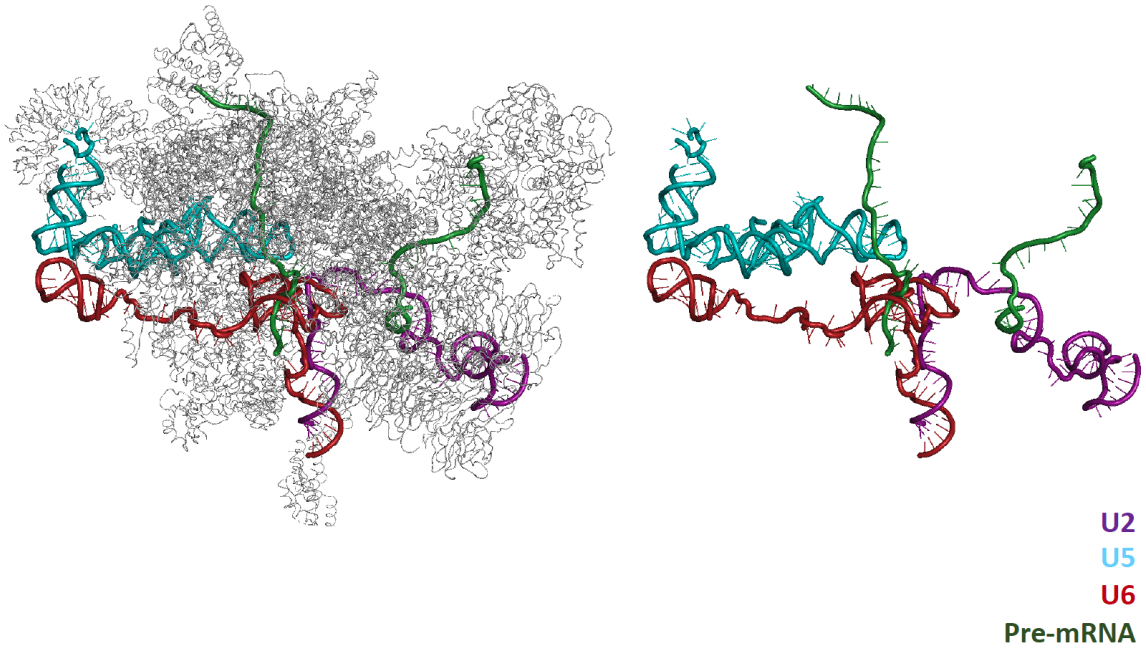


Fig. 1.4 The catalytically active spliceosome (Protein Data Bank #5LQW; Rauhut et al., 2016), showing its multitude of RNA:RNA interaction sites. The coloured structural elements represent RNAs, as shown in the key, with the grey structural elements representing the spliceosomal protein components. Visualised in *PyMol* (Schrodinger, 2015a).

The hypothesis that these snoRNAs might have structural similarity to the U4 spliceosomal snRNA allows us to propose that the snoRNAs interfere or hijack spliceosomal function through preventing U4's involvement in the spliceosomal life cycle, therefore disrupting spliceosome function. Through its similarity, it is possible that these snoRNAs bind to and act as a sponge for U4, thus reducing its availability for its normal function as part of the spliceosomal cycle in a concentration dependent manner. This equilibrium between U4 and the snoRNAs could explain the high number of snoRNA copies if they do hijack the spliceosome in an attempt to favour their use over the U4 spliceosomal snRNA. Alternatively, the snoRNAs may be binding to U6, preventing U4 from binding to U6 and forming the normal duplex (shown in Figure 1.5). Cornilescu and colleagues (2016) suggest that the

Snu13p protein is required to introduce the 60° kink seen, allowing sequential recruitment of the U4 and U6 ribonucleoproteins (Prp3, Prp4). It can thus be suggested that the snoRNAs may disrupt the U4 snRNA's formation of this KT motif, possibly by preventing Snu13p binding.

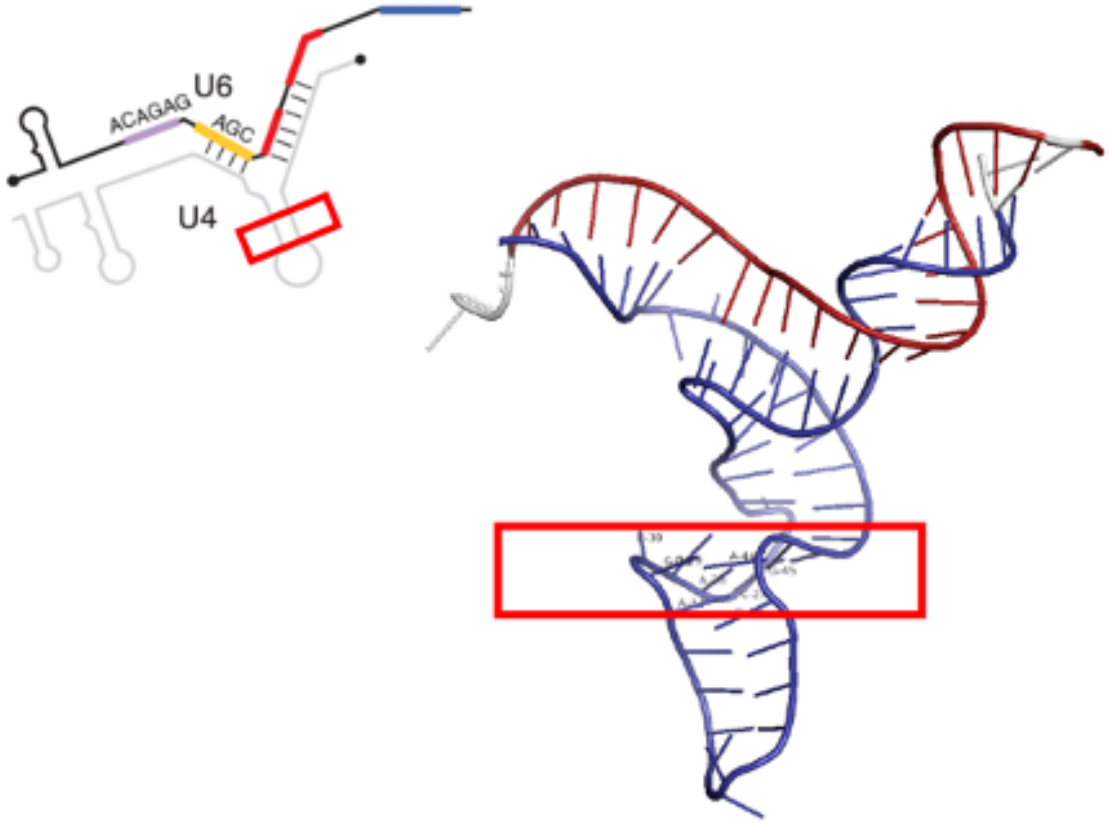


Fig. 1.5 The U4/U6 Duplex, with the Kink Turn motif highlighted within the red rectangle. NMR crystal structure (Protein Data Bank #2N7M; Cornilescu et al., 2016).

1.4 Aims of the project

In this project, I will aim to answer the following:

- To what degree is the spliceosome conserved?
- Can these imprinted snoRNAs hijack/interfere with the spliceosome?
- Do they exert their effect via a structural similarity to the spliceosome components e.g., U4?
- Is this structural similarity evolutionarily conserved? (i.e., does it occur in mice and humans?)

Chapter 2

Materials and Methods

2.1 GitHub

All data and code was made available through my GitHub Repository (https://github.com/AFS-Part-II-Projects/Ujjawal_Kumar) to facilitate reproducibility. Hence, anyone may access my data, code and methodology to verify my results and/or conduct their own investigations. This kind of information sharing facilitates open research, improving the quality of and access to scientific research. I used various *R* and *bash* (run on a command line) scripts; examples of code are presented as figures throughout this section, with complete code provided as appendices.

2.2 Assessing conservation of spliceosomal snRNAs between *Mus musculus* and *Homo sapiens*

The first aim was to assess spliceosomal snRNAs conservation. The spliceosomal snRNA sequences (Appendix A) were downloaded from RNAcentral/Rfam, a database of ncRNAs (Kalvari et al., 2018, 2021; RNAcentral Consortium, 2021), with accession codes in Table 2.1.

Sequence alignment was used to investigate similarity at a one-dimensional (sequence) level. Local alignment (aligning a substring of the query sequence with a substring of the target sequence, using the Smith-Waterman algorithm [Smith and Waterman, 1981]) is more suitable than global alignment (aligning the whole query sequence with the whole target sequence, using the Needleman-Wunsch algorithm [Needleman and Wunsch, 1970]). This is because the sequences have different lengths, and we want to identify regions with high levels of similarity, rather than comparing whole sequences. A comparison of global and local alignment is shown in Figure 2.1.

The Smith-Waterman algorithm uses dynamic programming (Figure 2.2) to calculate alignment scores.

Spliceosomal snRNA	**ENA accession #**	**Rfam accession #**	**RNAcentral accession #**
H. sapiens U1 snRNA	CM000664.2	RF00003	URS000071A5D2_9606
H. sapiens U2 snRNA	CM000671.2	RF00004	URS000063164F_9606
H. sapiens U4 snRNA	CM000671.2	RF00015	URS0000715A86_9606
H. sapiens U5 snRNA	CM000663.2	RF00020	URS0000631BD4_9606
H. sapiens U6 snRNA	CM000675.2	RF00026	URS00006767A8_9606
M. musculus U1 snRNA	CM000996.2	RF00003	URS0000722349_10090
M. musculus U2 snRNA	CM000996.2	RF00004	URS0000726205_10090
M. musculus U4 snRNA	CM001007.2	RF00015	URS000064AD77_10090
M. musculus U5 snRNA	AC083892.19	RF00020	URS0001BC1F50_10090
M. musculus U6 snRNA	CM001010.2	RF00026	URS0000710FEE_10090

Table 2.1 Accession codes for the sequences of the spliceosomal snRNAs, downloaded as .fasta files from RNAcentral.

Fig. 2.1 Comparison of local and global alignment. Due to the differing lengths of the sequences to be compared, as well as the need to find sequence regions with a high degree of similarity, local alignment using the Smith-Waterman Algorithm was chosen. Figure from www.majordifferences.com/2016/05/differencebetween-global-and-local.html.

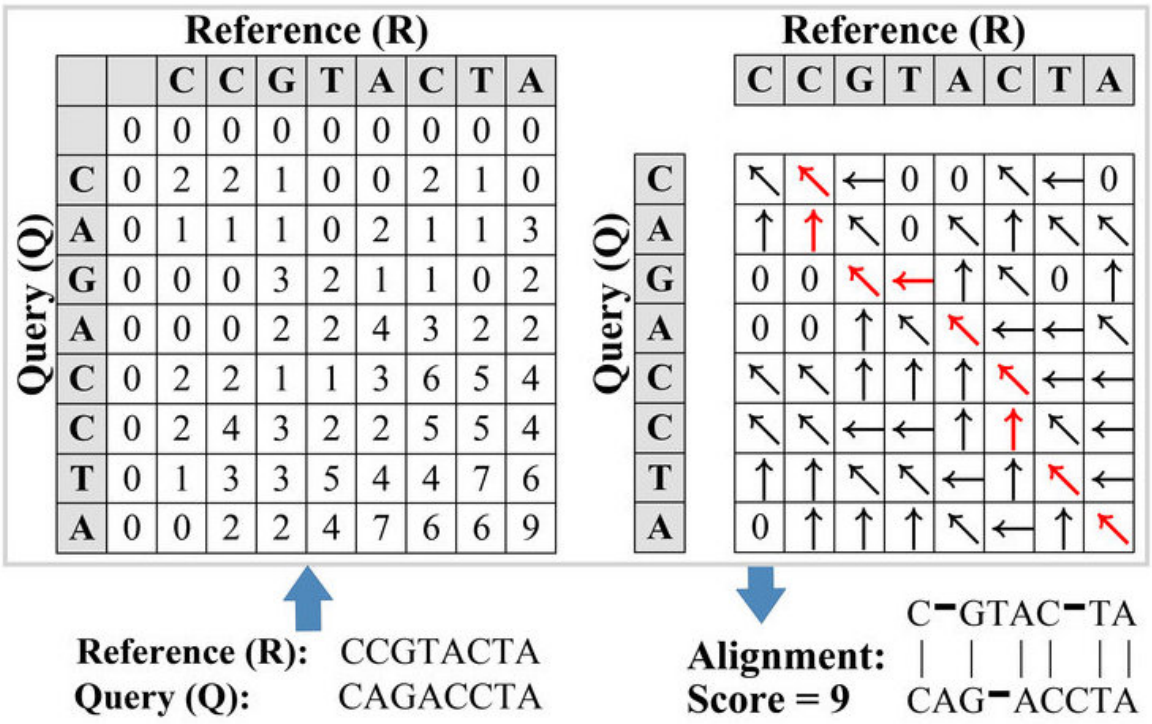


Fig. 2.2 Smith-Waterman algorithm uses dynamic programming to carry out local pairwise sequence alignment. Figure from Liao et al., 2018. The scoring parameters are match = +2, mismatch = -1, gap = 1. The vertical sequence is the input sequence, and the horizontal sequences is the reference. The left matrix is the corresponding (n + 1) by (m + 1) score matrix. The right matrix is the traceback matrix, with red arrows indicating the optimal alignment path. The null pointer is represented as 0.

Initially, EMBOSS Water (https://www.ebi.ac.uk/Tools/psa/emboss_water/; Madeira et al., 2019) was used to carry out local pairwise alignments. However, owing to the manual nature of loading sequences in EMBOSS Water, an R script Appendix D.1 was written to automate pairwise alignment. Identity values were tabulated and plotted in a heatmap, using the *ComplexHeatmap* R package (Gu et al., 2016; Zuguang Gu, 2017) and *circlize* (Gu et al., 2014) for the colour function.

To compare the two-dimensional similarity, the minimum free energy (MFE) secondary structures were first generated using *RNAfold*, initially using the web server (Gruber et al., 2008). However, command line *RNAfold* (part of the ViennaRNA package, Lorenz et al., [2011]; code in Appendix E.1) was used instead (quicker to run, requiring less manual input). *RNAfold* calculates the MFE using dynamic programming, maximising the number of base pairs and achieving as low a ΔG as possible, predicting a 2D structure. A more complete overview of the structures could be gained by investigating suboptimal structures other than just the single optimal structure. MFE structures were visualised in foRNA (Kerpedjiev et al., 2015) to obtain high quality 2D structural images for the spliceosomal snRNAs (Appendix B).

RNAdistance (also from ViennaRNA) was then run on the command line and used to calculate "distances" between spliceosomal snRNAs' secondary structures (Figure 2.3). These were tabulated and plotted as a heatmap (*ComplexHeatmap*, using code found in Appendix D.2).

```
uak20@ctr-web:~$ RNAfold -T37 < Spliceosomal\ snRNAs/Spliceosomal\ snRNAs\ Seqs/Spliceosomal\ snRNAs\ Seqs.fasta | RNAdistance -Xm
>hU1 snRNA
>hU2 snRNA
>hU4 snRNA
>hU5 snRNA
>hU6 snRNA
>mU1 snRNA
>mU2 snRNA
>mU4 snRNA
>mU5 snRNA
>mU6 snRNA
> f   10
154
172 156
154 164 124
119 181 159 135
180 160 148 128 157
157 187 195 161 164 189
173 171 129 111 168 133 182
154 168 122 44 127 134 143 105
118 154 148 106 99 154 143 135 94
```

Fig. 2.3 *RNAdistance* function used to obtain distances between pairs of spliceosomal snRNAs. N.B. this calculates the difference between the individual snRNAs, and therefore a lower value indicates greater similarity. *RNAdistance* was initially run using an R loop (see Appendix D.4), however the command line version (code shown here) was subsequently used instead, due to its greater speed.

2.3 snoRNA Sequence Retrieval

Sequences for the four snoRNA clusters were kindly provided by Aurily Constantino (*personal communication*) using *Infernal*, software that searches sequence databases for RNA sequence homologs, making RNA sequence alignments (Nawrocki and Eddy, 2013). Infernal uses multiple sequence alignments (gathered from Rfam entries: SNORD113/114 (RF00181), SNORD115 (RF00105), and SNORD116 (RF00108)) to build statistical (covariance) models of structurally annotated RNA, using the CMBUILD program. Once the models were calibrated using CMCALIBRATE in Infernal, CMSEARCH was used to find homologs in a sequence database. For instance, the coordinates for the forty snoRNAs found in Dlk1-Dio3 were retrieved from the entry corresponding to human chromosome 14 (AL132709). Hits found above the inclusion threshold were organised, taking into account the START and END coordinates, deleting those not belonging to the seeds available in Rfam. Using the coordinates obtained, snoRNA sequences were extracted using *getSequence* from the R package *Seqinr* (Charif and Lobry, 2007). I then converted these DNA sequences to RNA sequences using *DNA2RNA* (Figure 2.4), part of the *FastaUtils* R package (full code in Appendix D.3).

```
DNA2RNA(file="snoRNAs/snoRNA␣Seqs/SNORD113_hs.fasta", out="
    snoRNAs/snoRNA␣Seqs/SNORD113_hs.rna.fasta")
```

Fig. 2.4 Example of DNA2RNA code used to convert the sequences received for hSNORD113.

Subsequently, sequence (*pairwiseAlignment*) and 2D structural comparison (*RNAdistance*) was performed on combinations of spliceosomal snRNAs and individual snoRNAs of the 4 clusters of interest, plotting heatmaps for each (Figures 3.3 and 3.4)

2.4 Foldalign used to search for structurally similar regions between the snoRNAs and the U4 spliceosomal snRNA

Foldalign (Figure 2.5, Sundfeld et al., 2016) was used to find common elements between the U4 spliceosomal snRNA (the spliceosomal snRNA of interest), and the individual snoRNAs.

```
foldalign -plot_score hU4.fasta hSNORD113_1.fasta |
   locateHits > ~/U4\ x\ snoRNAs/Hit\ List\ hU4_hSNORD113/
   hU4_hSNORD113_01.txt
```

Fig. 2.5 Example of code used to run Foldalign, searching for common structural elements between the U4 spliceosomal snRNA and the individual snoRNAs of the clusters of interest. This will output a "hit list" as shown in Table C.1 (Appendix C).

A Foldalign "hit" suggests a region of structural similarity between the snoRNA and U4, possibly bound by a protein factor, supporting the hypothesis that the snoRNAs "soak up" the protein, preventing usual spliceosomal function. The mean p-score (where p <1.000) of the top five hits for each U4 vs snoRNA combination was calculated. The ten snoRNAs (mouse and human) with the lowest mean p-scores were chosen as the candidate snoRNAs to explore with 3D structural prediction and comparison.

2.5 *RNAcofold* used to predict duplex formation between candidate snoRNAs and spliceosomal snRNAs

RNAcofold (ViennaRNA, Figure 2.6) was used in order to predict duplex formation between the candidate snoRNAs and U4, U5 and U6, since they were shown to be involved in binding with U4. The number of base pairs within contact sites is an indicator of snoRNA binding to the spliceosomal snRNA, with a greater number of base pairs indicating greater likelihood of binding. Additionally, these base pairs should be contiguous to be an accurate predictor of a contact site.

```
RNAcofold -p --id-prefix=U4_SNORD114_13 < U4_SNORD114-13.seq
    > U4_SNORD114-13.RNAcofold.out
```

Fig. 2.6 Code used to predict secondary structure of duplexes formed between a candidate snoRNA and the U4 spliceosomal snRNA. Code for all candidate snoRNAs in Appendix E.3.

2.6 SimRNA used to predict the 3D structures of the twenty candidate snoRNAs and spliceosomal snRNAs

To predict 3D structures, SimRNA (Boniecki et al., (2016), a coarse-grained (Dawson et al., 2016) three-dimensional structure prediction package) was used (Figure 2.7). It represents each nucleotide as five pseudo-atoms, the constitutive phosphate and C4' sugar atoms in the backbone, and the base as a triangle of three atoms.

```
SimRNA -s ~/3D\ Structure/Candidates/Sequences/hSNORD114_06.
    fasta -c config.dat -S ~/3D\ Structure/Candidates/2D\
    Structures/hSNORD114_06\ MFE.rnafold.ss -r ~/3D\
    Structure/Candidates/Restraint\ Files/hSNORD114_06.pairs.
    con.cut
trafl_extract_lowestE_frame.py hSNORD114_06.fasta.trafl
SimRNA_trafl2pdbs hSNORD114_06.fasta-000001.pdb hSNORD114_06
    .fasta_minE.trafl 1 AA
```

Fig. 2.7 Example of code used to run SimRNA in order to carry out 3D structure prediction for one of my candidate snoRNAs. See Appendix G for full code and the Brackets2Restraints.pl code used to make the restraints files from the 2D structures predicted in *RNAfold*, as well as the config.dat file containing the parameters used to run the SimRNA 3D structure prediction. The AA at the end of the final line represents a command to replace the coarse-grained structures with All Atom versions in order to fully visualise the structures in *PyMol*.

Restraints may be assigned to any pair of atoms, however through specifying 2D structure (in dot/bracket, Vienna notation), base pairing restraints are created (using code as in Figure 2.8) between pairs of atoms that form the link between the triangular pseudo-bases. As in Taylor and Hamilton, (2017), the SimRNA parameters were fine-tuned (config.dat file, Appendix G).

```
RNAfold < ~/3D\ Structure/Candidates/Sequences/hU4\ snRNA.
    fasta | b2ct | perl ~/3D\ Structure/Brackets2Restraints.
    pl > ~/3D\ Structure/Candidates/Restraint\ Files/hU4\
    snRNA.pairs.con.cut
```

Fig. 2.8 An example of the code used to create the restraints files which SimRNA uses in its coarse-grain 3D structure prediction. This code makes use of a Brackets2Restraints.pl *Perl* script (see Appendix G) which was kindly provided by my supervisor, Dr Russell Hamilton.

Two-dimensional structures were generated using *RNAfold* as discussed before. Each three-dimensional prediction consisted of ten runs, selecting the MFE three-dimensional structure. Once 3D structure prediction was complete, the .pdb files (containing atomic coordinates for the candidate snoRNAs and the spliceosomal snRNAs) were visualised using *PyMol* (Schrodinger, 2015a) to view high quality three-dimensional structures.

Using the top Foldalign hit, the candidate snoRNAs were aligned to the U4 spliceosomal snRNA in *PyMol*, using *align* (Figure 2.9), obtaining RMSd values. RMSd analysis is a quantitative method of assessing the similarity between two biochemical structures and measures the square root of the mean of the squared distances between individual atoms in the aligned region. A smaller RMSd value indicates that the atoms are closer together and hence the structures of that region are more similar (Petitjean, 1999).

```
align hU4_snRNA.fasta_minE-000001_AA and resi 151-183,
    hSNORD114_06.fasta_minE-000001_AA and resi 22-54
```

Fig. 2.9 Code used to align the 3D structures of a candidate snoRNA to the U4 spliceosomal snRNA. This was done using the top hits found using a Foldalign search; the top hit is the predicted region of structural similarity with the lowest p score. A full list of top Foldalign hits for the Candidate snoRNAs against the U4 spliceosomal snRNA is shown in Table 3.

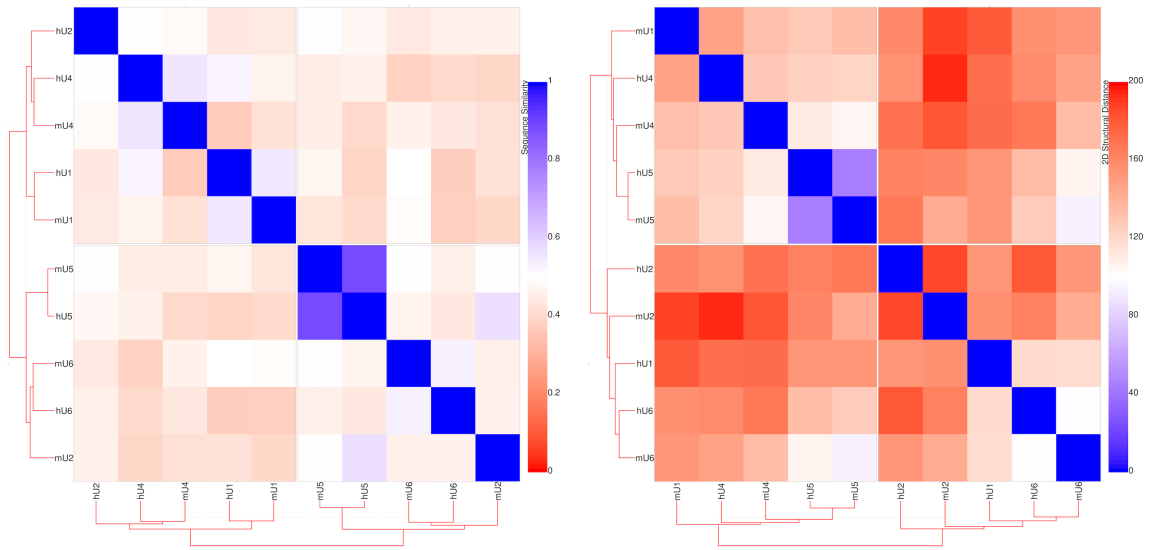
Chapter 3

Results

3.1 Key spliceosomal snRNAs show conservation between *Mus musculus* and *Homo sapiens*

To investigate conservation of the spliceosomal snRNAs between *Mm* and *Hs*, the degree of sequence similarity as well as 2D structural similarity was evaluated (results shown in Figure 3.1).

As illustrated in the heatmaps (Figure 3.1), there were particular spliceosomal snRNAs which showed greater similarity between the mouse and human homologues, specifically the U4 and U5 spliceosomal snRNAs, as well as U1 and U6. These are amongst the most important snRNAs of the spliceosome, acting as the key regulator (U4), and forming part of the catalytically active spliceosome (U5, U6). U4 and U5 were specifically investigated by comparing their 2D as well as 3D structures, by juxtaposing their structures. As can be seen in Figure 3.2, there are similar elements, but crucially not the whole structure, thus verifying the result from the sequence comparisons. Since there is notable similarity between key spliceosomal snRNAs across all three dimensions, I can therefore conclude that there is conservation of the U4-dependent, major spliceosome between *Mus musculus* and *Homo sapiens*.



a. Sequence Similarity b. 2D Structural Distance

Fig. 3.1 – Heatmaps representing the similarity analysis undertaken. **(a.)** Sequence similarity scores obtained through local pairwise sequence alignment, using the Smith-Waterman algorithm as described earlier. As shown in the legend, blue indicates greater similarity, with red indicating greater discrepancy **(b.)** 2D structural distance analysis. It is important to note that this depicts the **distance** and therefore the difference between a pair of RNAs. Again, as shown in the legend, blue indicates greater similarity, with red indicating greater discrepancy.

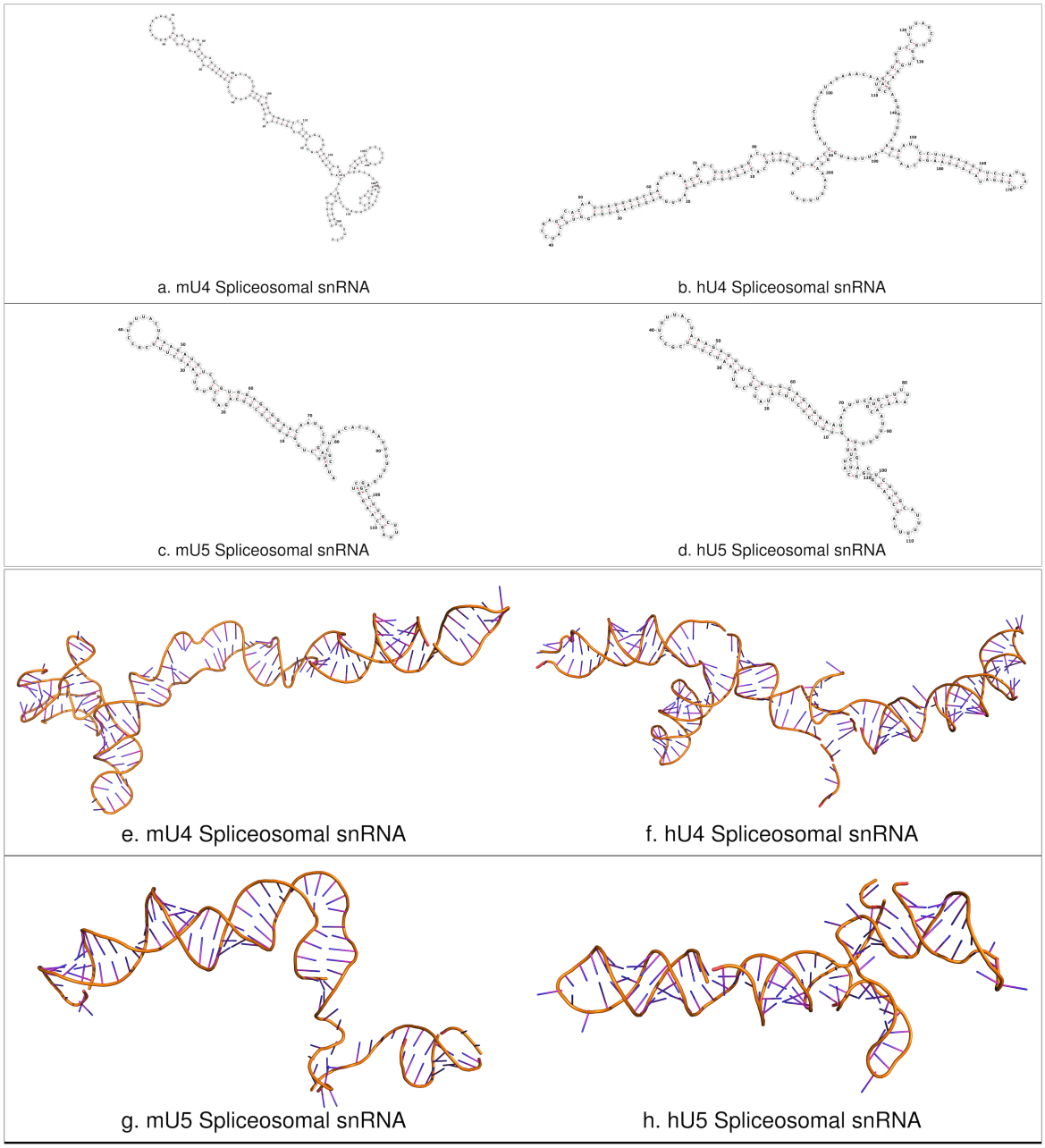


Fig. 3.2 – Juxtaposed structures of the mU4, hU4, mU5 and hU5 spliceosomal snRNAs. There is significant similarity between the 2D and 3D structures of the human and mouse homologues, with multiple common structural motifs. 2D MFE structures were calculated using RNAfold, visualised in foRNA. 3D structures were predicted using SimRNA, visualised in *PyMol*. See Appendix B for all 2D and 3D spliceosomal snRNA structures.

3.2 There is sequence and 2D structural similarity between a subset of snoRNAs and the spliceosomal snRNAs

To preliminarily investigate whether there is similarity between some of the snoRNAs of interest, pairwise sequence and 2D structural distance analysis was run for the *Mm* (Figure 3.3) and *Hs* (Figure 3.4) spliceosomal snRNAs and snoRNAs.

Through these heatmaps (Figures 3.3 and 3.4), we can see that there are certain groups of snoRNAs that show marked sequence similarity to the spliceosomal snRNAs, specifically U4. These snoRNAs are not necessarily grouped within one cluster and are quite varied. There seems to be a less notable structural similarity between the U4 spliceosomal snRNA and the snoRNAs, however this is to be expected. *RNAdistance* uses a kind of global alignment, and since the snoRNAs are much smaller than the spliceosomal RNAs (and will only have certain motifs in common) the lack of similarity is not contradictory to our structural similarity hypothesis.

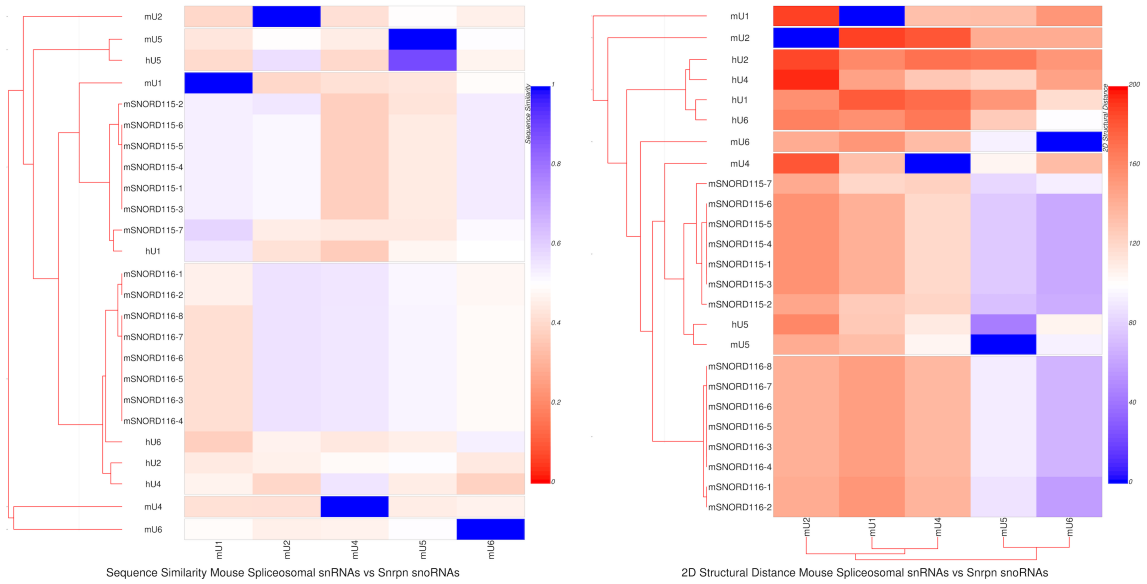


Fig. 3.3 – Heatmaps showing the sequence similarities and 2D structural distances of pairwise combinations of the murine spliceosomal snRNAs and snoRNAs (grouped by imprinted region within which they are found).

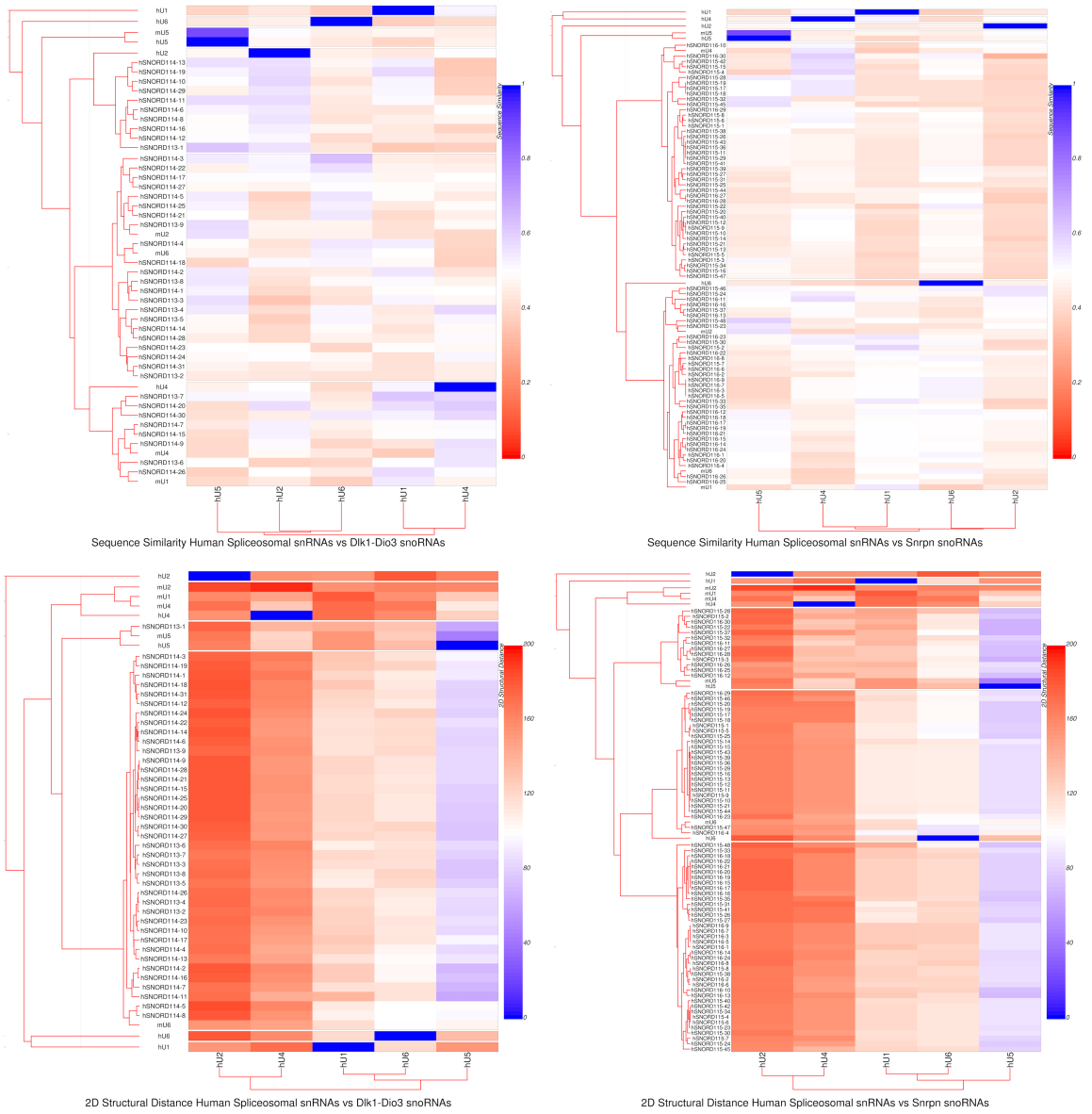


Fig. 3.4 – Heatmaps showing the sequence similarities and 2D structural distances of pairwise combinations of the human spliceosomal snRNAs and snoRNAs (grouped by imprinted region within which they are found).

3.3 snoRNAs show common structural elements to the U4 spliceosomal snRNA

Using Foldalign, I showed that there are common structural elements between individual snoRNAs and the U4 spliceosomal snRNA. A hit list was generated (see *Methods*), and the top five hits were identified, with the mean of their p-scores calculated. The ten snoRNAs for each of the two species with the lowest mean p-scores were identified (shown in Table 3.1), as these were the candidate snoRNAs going forward for 3D structural prediction and alignment.

Candidate snoRNA	**Mean p score of top 5 hits**
hSNORD114-13	0.4764
hSNORD114-6	0.486
hSNORD115-45	0.5216
hSNORD116-26	0.5736
hSNORD114-27	0.5878
hSNORD116-13	0.6036
hSNORD114-26	0.6044
hSNORD114-11	0.6084
hSNORD116-10	0.6242
hSNORD116-23	0.6252
mSNORD115-2	0.5202
mSNORD115-1	0.5344
mSNORD115-3	0.5344
mSNORD115-4	0.5344
mSNORD115-5	0.5344
mSNORD115-6	0.5344
mSNORD115-7	0.6748
mSNORD116-3	0.9384
mSNORD116-4	0.9384
mSNORD116-5	0.9384

Table 3.1 – The twenty candidate snoRNAs as identified through a Foldalign search for common structural elements (ten human, ten mouse).

These candidate snoRNAs' top two hits were then used to annotate the foRNA MFE structure of the U4 spliceosomal snRNA (Figures 3.5 and 3.6).

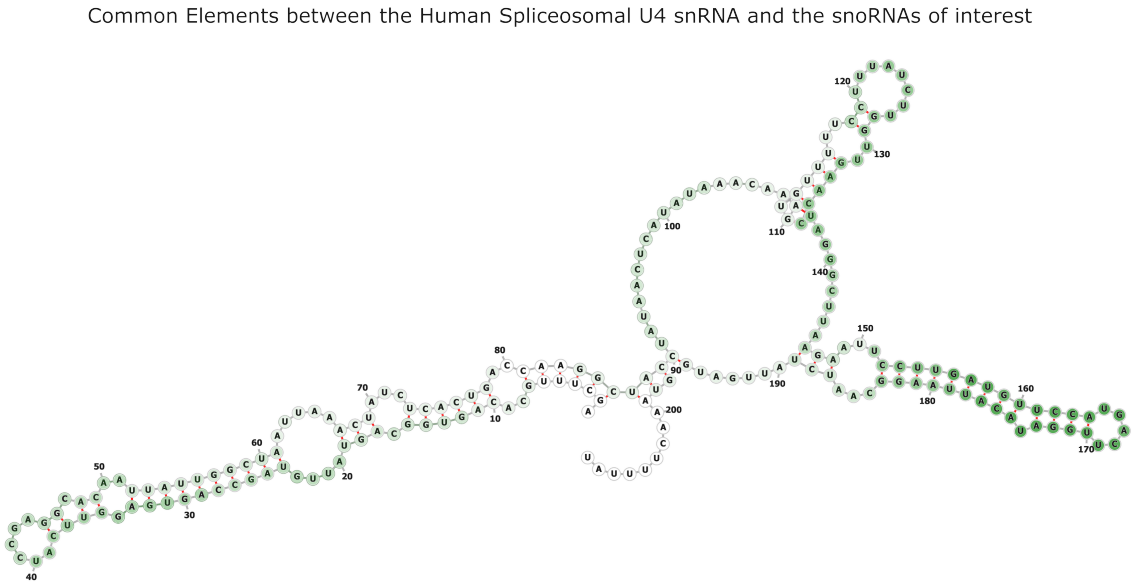


Fig. 3.5 – Annotated hU4 Spliceosomal snRNA structure to show the frequency of a site being part of a structural element shared with a candidate snoRNA. The opacity of colour represents the number of candidate snoRNAs that have a hit as part of a common structural element at that site. The similarity between the U4 spliceosomal snRNA and the candidate is found in discrete, well defined regions, with particular areas having a degree of similarity with multiple candidates, hence their dark opacity.

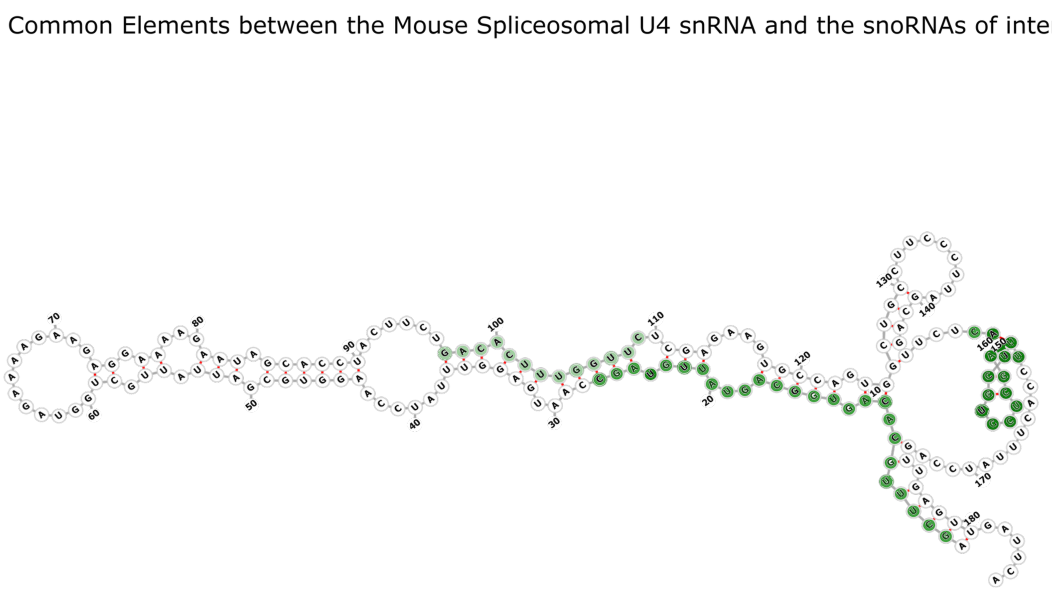


Fig. 3.6 – Annotated mU4 Spliceosomal snRNA structure to show the frequency of a site being part of a structural element shared with a candidate snoRNA. The opacity of colour represents the number of candidate snoRNAs that have a hit as part of a common structural element at that specific site. There are far fewer mouse snoRNAs, and they are extremely similar to each other, which explains the intense clustering of the common structural elements, and associated dark opacity of the relevant sites on the annotated MFE structure.

3.4 *RNAcofold* can be used to predict RNA/RNA annealing and duplex formation

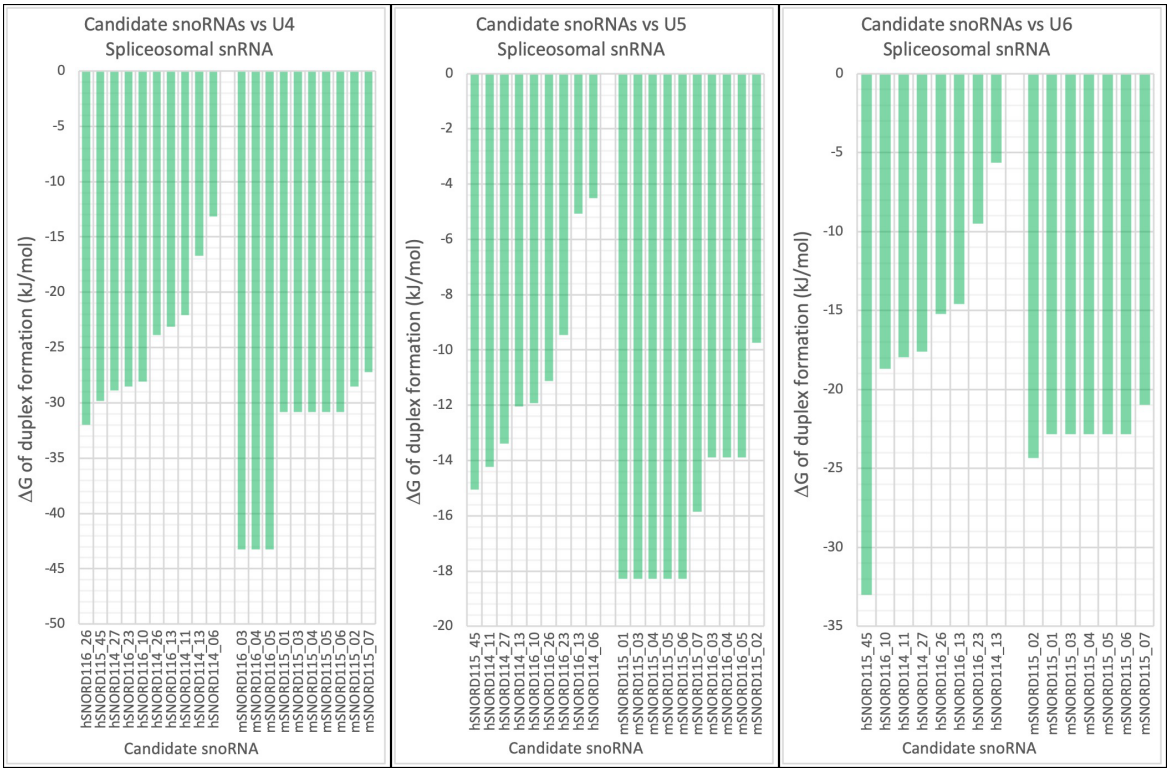


Fig. 3.7 – The ΔG of binding (and therefore duplex formation) of the U4, U5 and U6 Spliceosomal snRNAs to each candidate snoRNA. Positive ΔG values were excluded since they are endergonic and therefore will not occur spontaneously in a biological environment. More negative ΔG values indicate a greater decrease in the Gibbs free energy of the products of a biochemical reaction as compared to the reactants, and thus the more spontaneous a reaction is. The more spontaneous binding is, the more likely it is to take place, and thus more likely that duplex will form. Data for these graphs found in Table F.1 (Appendix F.2)

As shown in Figure 3.7, there are a range of ΔG values for the binding of candidate snoRNAs to the U4, U5 and U6 spliceosomal snRNAs. Those duplexes with the most negative ΔG value are thermodynamically the most likely to form. We can see that certain snoRNAs (Figure 3.8) have notably low ΔG values for binding to U4, U5 or U6, indicating the high feasibility of duplex formation.

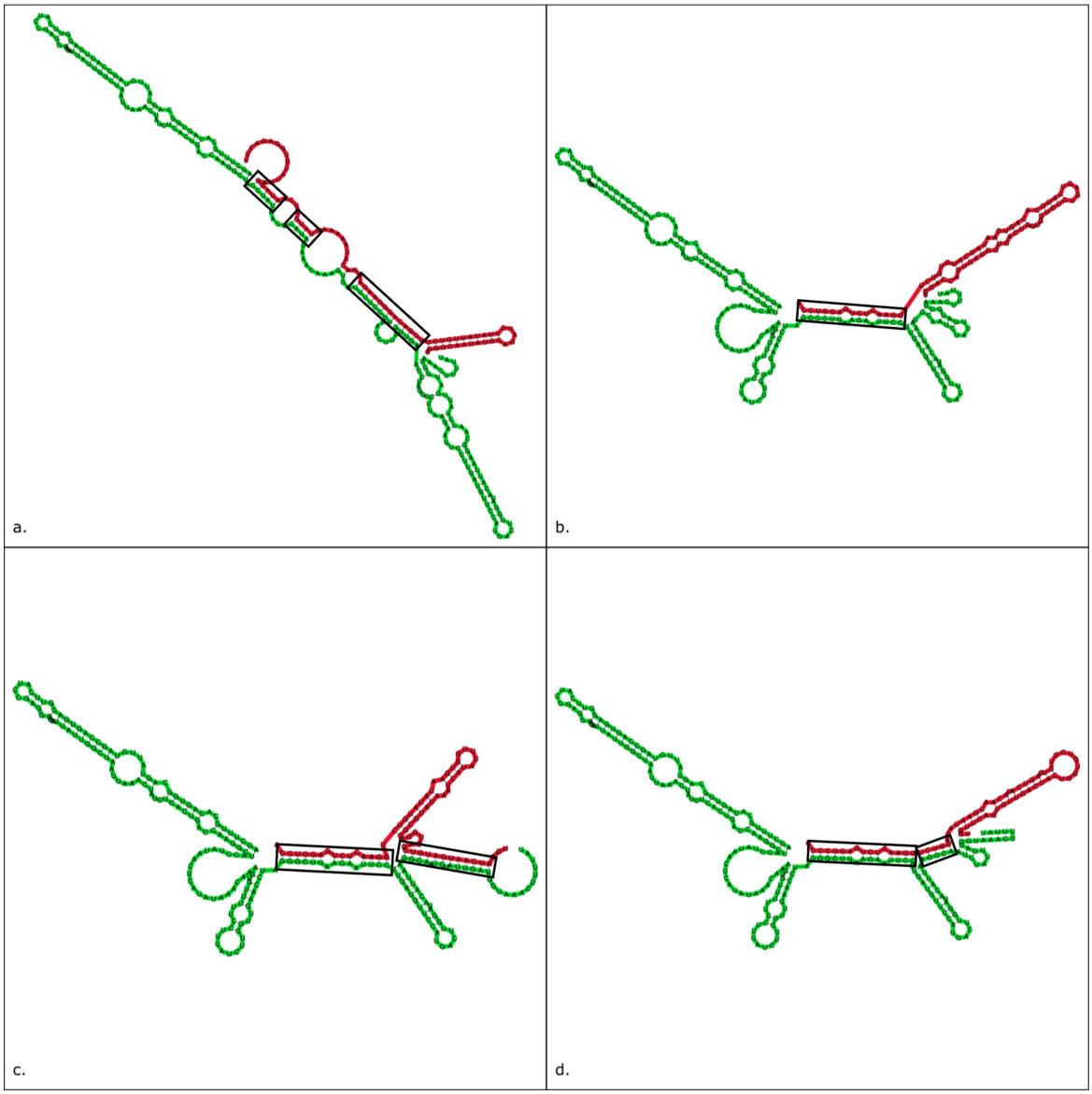


Fig. 3.8 – Predicted RNA duplexes formed between the human U4 spliceosomal snRNA and **(a.)** hSNORD115_45 **(b.)** hSNORD114_11 **(c.)** hSNORD114_06 **(d.)** hSNORD114_27. These are amongst the pairings with the lowest ΔG and therefore the most likely to form, thermodynamically speaking. The binding sites are outlined in the black rectangles, and crucially, these paired bases are contiguous with each other, thus forming binding sites that incorporate large numbers of bases.

3.5 3D Structure Prediction allows superimposition of structures, RMSd calculation

Using SimRNA, the 3D structures of my twenty candidate snoRNAs were predicted, and subsequently visualised in *PyMol* (see Appendix H.1 for 3D structures of all twenty candidate snoRNAs). Each of the candidate snoRNAs was aligned (using the top Foldalign hit) to their respective U4 spliceosomal snRNA, and the RMSd values calculated (shown in Table 3.2).

Seq1	Seq1 Start	Seq1 End	Seq2	Seq2 Start	Seq2 End	RMSd (Å; Angstroms)
hU4	151	183	hSNORD114-06	22	54	4.429
hU4	12	40	hSNORD114-11	24	52	12.292
hU4	7	61	hSNORD114-13	3	58	20.874
hU4	151	183	hSNORD114-26	29	62	13.359
hU4	20	51	hSNORD114-27	7	38	17.761
hU4	156	177	hSNORD115-45	40	61	2.205
hU4	123	140	hSNORD116-10	82	99	13.534
hU4	123	140	hSNORD116-13	72	89	15.428
hU4	84	143	hSNORD116-23	34	92	25.066
hU4	151	189	hSNORD116-26	4	43	8.029
mU4	148	161	mSNORD115-01	2	15	7.807
mU4	148	161	mSNORD115-02	2	15	13.163
mU4	148	161	mSNORD115-03	2	15	10.814
mU4	148	161	mSNORD115-04	2	15	6.127
mU4	148	161	mSNORD115-05	2	15	9.575
mU4	148	161	mSNORD115-06	2	15	10.676
mU4	148	161	mSNORD115-07	2	15	14.045
mU4	97	109	mSNORD116-03	82	94	8.988
mU4	97	109	mSNORD116-04	82	94	9.145
mU4	97	109	mSNORD116-05	82	94	8.292

Table 3.2 – The alignment regions used, including the RMSd values obtained through *PyMol*. These regions were obtained as the top Foldalign hits for each candidate snoRNA when compared to U4 snRNA.

The lower an RMSd value, the lower the average distance between the atoms of each structure in the aligned region and therefore the structures of that region are more similar. These RMSd alignment values were plotted in Figure 3.9.

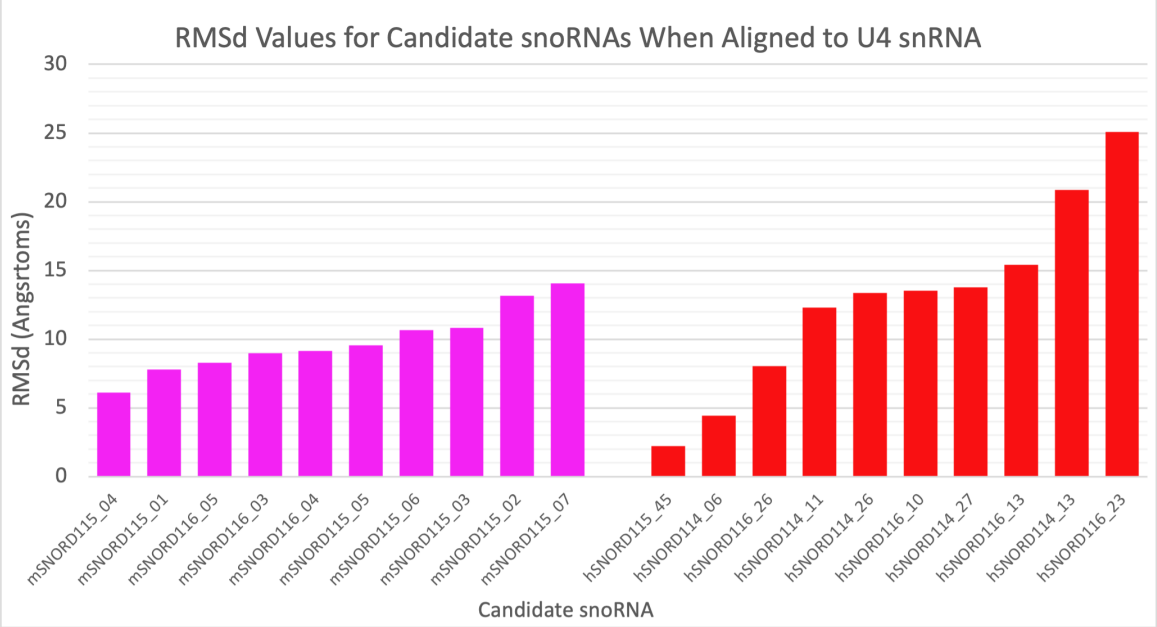


Fig. 3.9 – A graph showing the RMSd values for individual candidate snoRNAs when aligned using the top Foldalign hit to the U4 spliceosomal snRNA.

Thus, I identified the four snoRNAs for mouse and human with the lowest RMSd values and therefore conclude that these are the most similar to their respective U4 spliceosomal snRNA within this aligned region. Below are the 3D structures of each of these eight snoRNAs (Figure 3.10), as well as their alignments with the U4 snRNA (Figure 3.11).

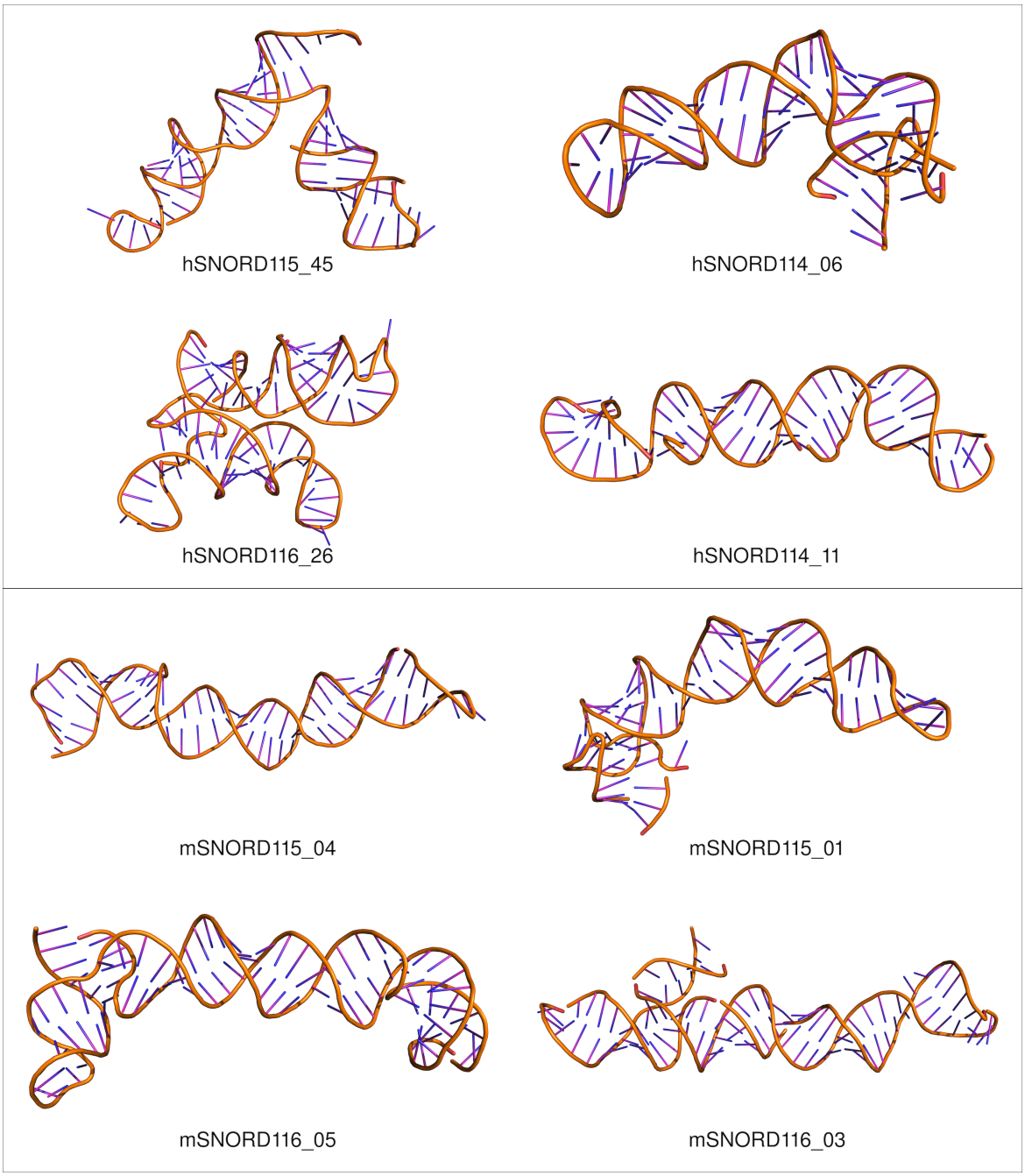


Fig. 3.10 – The four human and four mouse snoRNAs with the lowest RMSd values. Shown as 3D "cartoon structures".

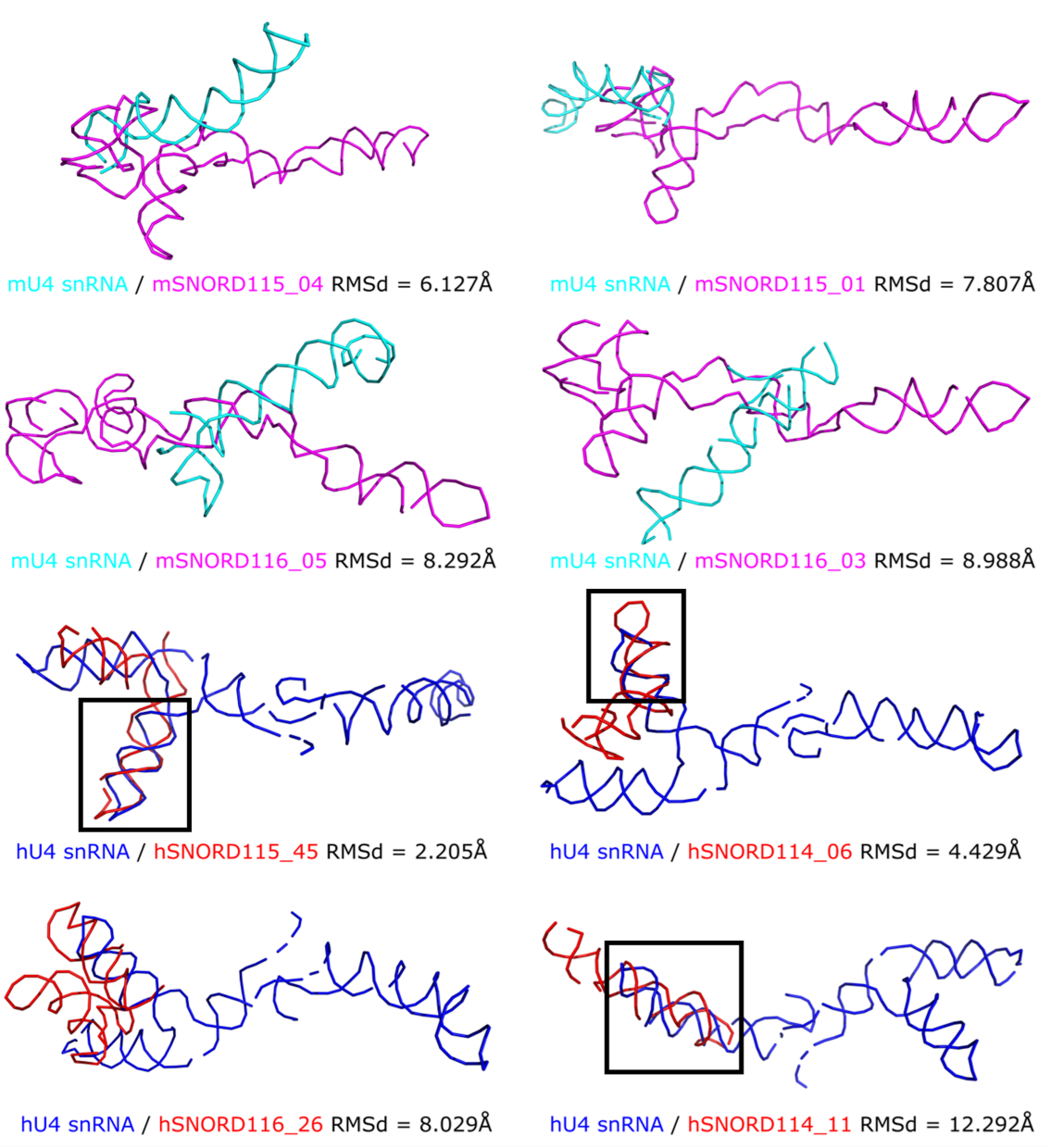


Fig. 3.11 – The top four mouse and top four human snoRNAs, aligned to their respective U4 snRNAs. Shown as 3D "ribbon structures".

As shown by the boxes (hSNORD115_45, hSNORD116_26, hSNORD114_11), the alignment region appears to contain the KT motif. As discussed earlier, this motif is fundamental in brain-specific expression with the U4 spliceosomal snRNA and these snoRNAs being shown to form stable KT motifs. The alignment encompassing this key motif (disease phenotypes associated with snoRNA disruption are linked to the nervous system) suggests that this is key to the normal functioning, and the mechanism of action of these snoRNAs.

Chapter 4

Discussion

4.1 Conclusions

In this project, I have used a variety of techniques in order to investigate the research questions I proposed in my "Aims" section. Through a combination of sequence and structural similarity analysis, I have shown that specific spliceosomal snRNAs (particularly U4 and U5) are conserved between *Homo sapiens* and *Mus musculus*. I have also shown that a subset of snoRNAs do show similarity to the spliceosomal snRNAs, including U4. This has been shown at a one-dimensional (sequence), two-dimensional (secondary structure) as well as at a three-dimensional level (tertiary structure). There seems to be structural similarity, as well as specific shared motifs such as the KT as previously discussed. My results throughout this project lend support to the hypothesis that these snoRNAs have a structure-based rather than sequence-dependent mechanism of action, with this structural similarity to the spliceosome allowing them to interfere with the spliceosomal machinery in some way. Lastly, I have also found this similarity in both species, indicating evolutionary conservation.

4.2 Future experiments and alternative hypotheses

However, these findings are computational predictions and need to be experimentally validated to be confident in this hypothesis and prove that there is binding between the U4 snRNA and these snoRNAs. This validation may be done using *in vitro* experiments to measure the binding affinity. Standard techniques used to do this include isothermal calorimetry (ITC), as well as others such as surface plasmon resonance and electrophoretic mobility shift assay (EMSA). However, these tend to have rather low throughputs, so perhaps more recent techniques such as "RNA Bind-n-Seq" (RBNS) as discussed in Lambert et al., (2015), might be more suited to this investigation. Binding reactions between tagged pairs of RNAs may be incubated, and subsequent high-throughput sequencing of RNA duplexes may be carried out. Though this technique was developed to quantify the binding affinity of RNA binding proteins and RNAs, there is potential for use in binding of pairs of nucleic acids such as this case.

However, it is not clear whether in fact U4 is the target of these snoRNAs, since they also seem to show notable similarity to other spliceosomal snRNAs. It is therefore possible that these snoRNAs interfere with another spliceosomal snRNA such as U5 or U6, since these are also fundamentally important, making up multiple stages of the spliceosomal life cycle and are key for the multitude of RNA:RNA interaction sites seen in both the pre-catalytic spliceosome (Figure 4.1), as well as the catalytically active spliceosome (Figure 1.4).

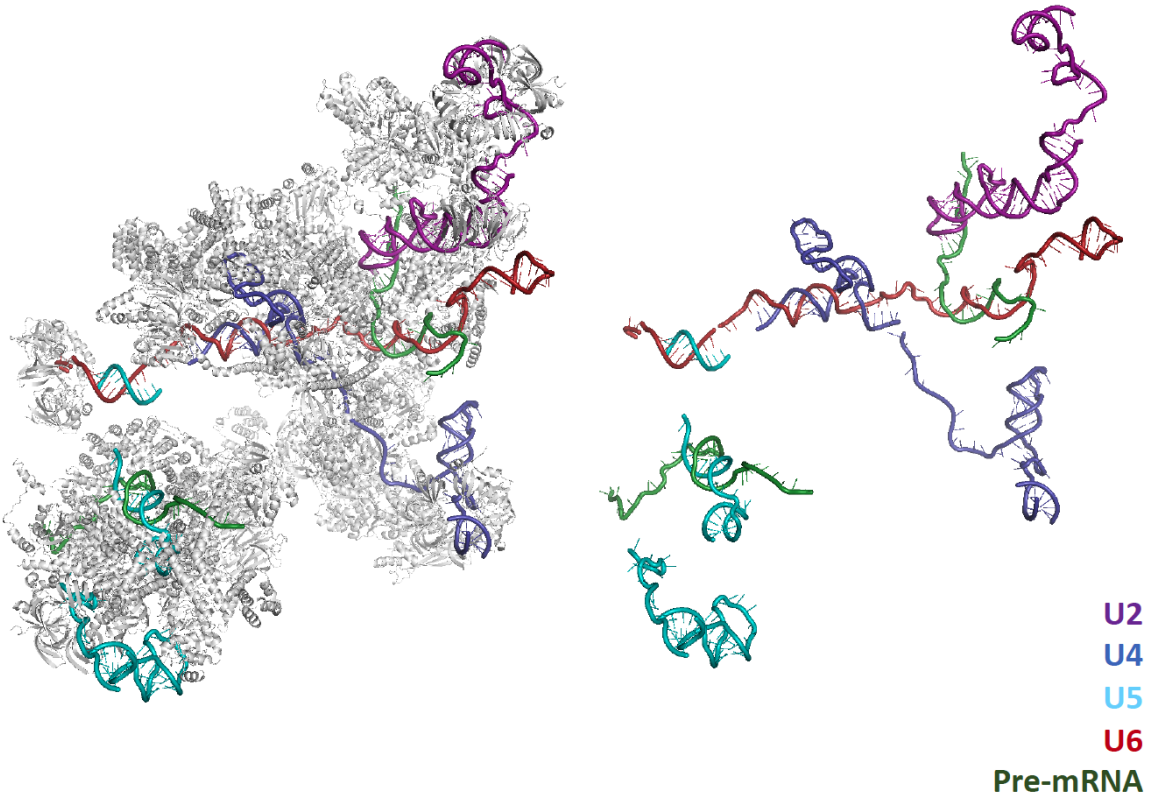


Fig. 4.1 – The pre-catalytically active spliceosome (Protein Data Bank #6AHD; Zhan et al., 2018), showing its multitude of RNA:RNA interaction sites. The coloured structural elements represent RNAs, as shown in the key, with the grey structural elements representing the spliceosomal protein components. Visualised in *PyMol* (Schrodinger, 2015a)

Another hypothesis suggested by Yin et al. (2012) is that these snoRNAs might act via an RNA:protein interaction and that the snoRNAs interfere with other splicing factors or the protein chaperones required to form the KT motif. They outline how during exonucleolytic trimming, some of the sequences between the snoRNAs are not in fact degraded, resulting in the formation of long non-coding RNAs (lncRNAs) that show snoRNA sequences on either end and lack PolyA tails or 5' caps. These sno-lncRNAs have been shown to associate strongly with the FOX family of proteins, notably FOX1, altering splicing patterns (Figure 4.2). Work to predict RNA/Protein interactions is currently being done by other members of the Hamilton Group in order to explore this hypothesis further.

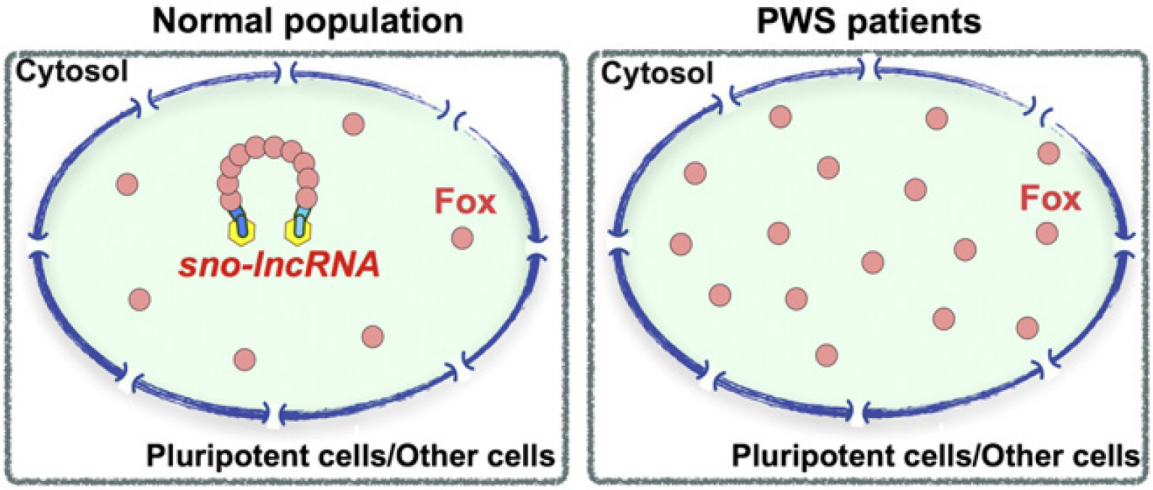


Fig. 4.2 – A suggested model for the action of sno-lncRNAs in normal cells, as well as in PWS patients. Figure from Yin et al., (2012). In normal cells, the PWS locus sno-lncRNAs are highly expressed, sequestering Fox family splicing factors in this model. This may result in reduced availability of Fox proteins throughout the nucleus, leading to global changes in alternative splicing regulation. Another possibility is an increased local concentration of Fox proteins in the vicinity of the sno-lncRNAs, leading to localised effects on alternative splicing. In PWS patients (PWS region sno-lncRNAs are deleted or not expressed), Fox splicing factors are more uniformly distributed throughout the nucleus, resulting in altered splicing patterns during early embryonic development and adulthood, leading to the disease phenotypes.

Using a technique such as individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP), developed by the Ule Lab (Huppertz et al., 2014), it may be possible to quantify these reactions, through crosslinking individual RNAs and proteins (process shown in Figure 4.3). This might allow us to see that the proposed interaction of these snoRNAs with the splicing factor proteins (as discussed above) is also seen with the spliceosomal snRNAs, and thus support the RNA/Protein interaction model.

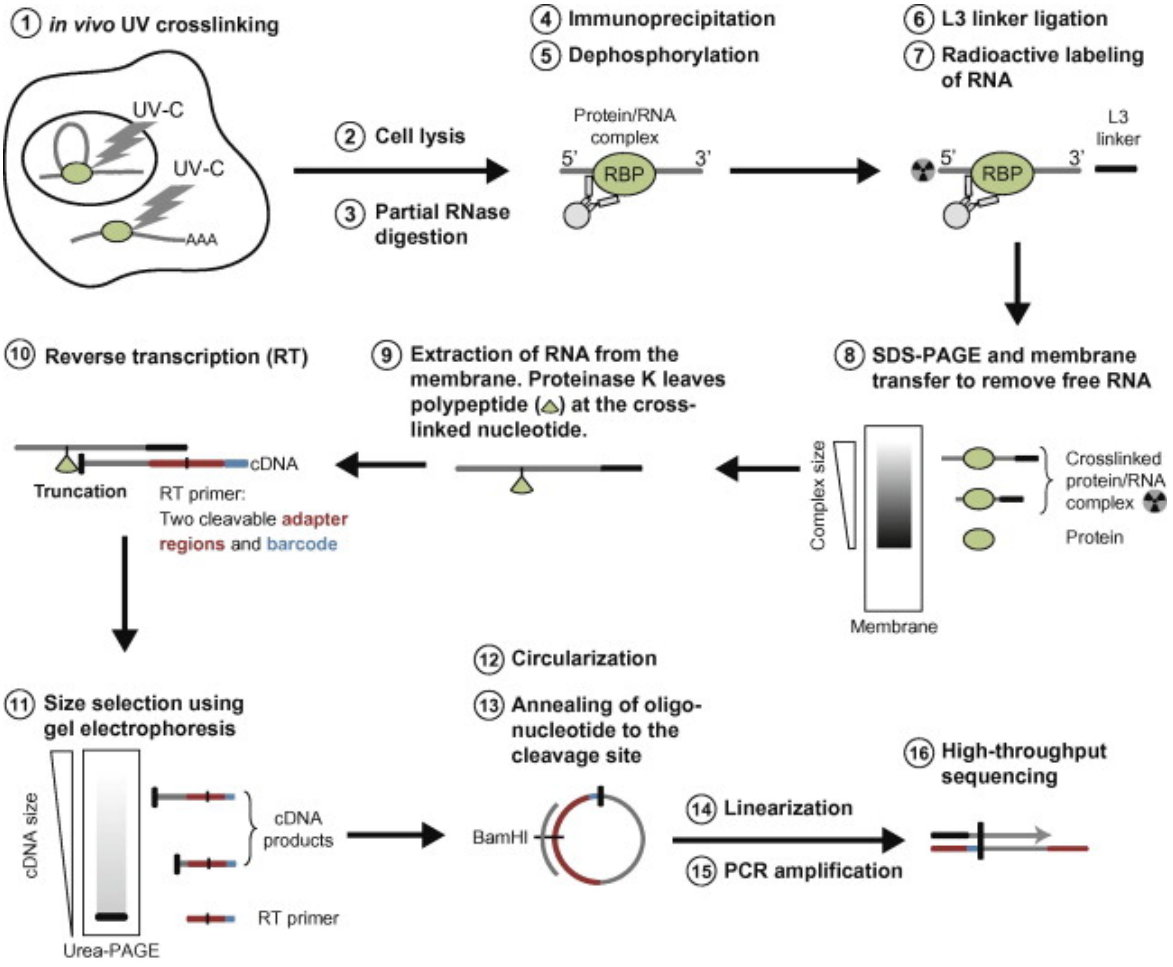


Fig. 4.3 – the iCLIP Process. Figure from Huppertz et al., (2014). The extremely high resolution (single nucleotide) and specificity of this process is thanks to the intramolecular cDNA circularization step, enabling analysis of cDNAs that truncate at the protein–RNA crosslink sites. iCLIP is an improvement over the already established CLIP methods discussed in Lee and Ule, (2018); Sugimoto et al., (2012).

Another possibility is the use of Oxford Nanopore sequencing in order to resolve these highly complex regions of the reference genomes. This will allow us to specifically sequence these snoRNA clusters, rather than relying on the reference genomes, which often have poorly resolved sequences for these complex regions.

Lastly, validation of these results in the presence of ribonucleoproteins will be necessary, since all predictions have been made for just the RNAs themselves, however we know that within the spliceosome, there are also ribonucleoproteins which form a core part of spliceosomal structure and influence the RNAs' structures.

Chapter 5

References

Barton, S.C., Surani, M. a. H., and Norris, M.L. (1984). Role of paternal and maternal genomes in mouse development. Nature 311, 374–376.

Bazeley, P.S., Shepelev, V., Talebizadeh, Z., Butler, M.G., Fedorova, L., Filatov, V., and Fedorov, A. (2008). snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. Gene 408, 172–179.

Bieth, E., Eddiry, S., Gaston, V., Lorenzini, F., Buffet, A., Conte Auriol, F., Molinas, C., Cailley, D., Rooryck, C., Arveiler, B., et al. (2015). Highly restricted deletion of the SNORD116 region is implicated in Prader-Willi Syndrome. Eur. J. Hum. Genet. EJHG 23, 252–255.

Black, D., and Steitz, J.A. (1986). Pre-mRNA splicing in vitro requires intact U4/U6 small nuclear ribonucleoprotein. Cell 46, 697–704.

Boniecki, M.J., Lach, G., Dawson, W.K., Tomala, K., Lukasz, P., Soltysinski, T., Rother, K.M., and Bujnicki, J.M. (2016). SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic Acids Res. 44, e63–e63.

Cavaillé, J. (2017). Box C/D small nucleolar RNA genes and the Prader-Willi syndrome: a complex interplay. Wiley Interdiscip. Rev. RNA 8, e1417.

Cavaillé, J., Buiting, K., Kiefmann, M., Lalande, M., Brannan, C.I., Horsthemke, B., Bachellerie, J.P., Brosius, J., and Hüttenhofer, A. (2000). Identification of brain-specific and

imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc. Natl. Acad. Sci. U. S. A. 97, 14311–14316.

Charif, D., and Lobry, J.R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 207–232.

Cornilescu, G., Didychuk, A.L., Rodgers, M.L., Michael, L.A., Burke, J.E., Montemayor, E.J., Hoskins, A.A., and Butcher, S.E. (2016). Structural Analysis of Multi-Helical RNAs by NMR–SAXS/WAXS: Application to the U4/U6 di-snRNA. J. Mol. Biol. 428, 777–789.

Coulson, R.L., Powell, W.T., Yasui, D.H., Dileep, G., Resnick, J., and Lasalle, J.M. (2018). Prader–Willi locus Snord116 RNA processing requires an active endogenous allele and neuron-specific splicing by Rbfox3/NeuN. Hum. Mol. Genet.

Dawson, W.K., Maciejczyk, M., Jankowska, E.J., and Bujnicki, J.M. (2016). Coarse-grained modeling of RNA 3D structure. Methods 103, 138–156.

Filipowicz, W., and Pogačić, V. (2002). Biogenesis of small nucleolar ribonucleoproteins. Curr. Opin. Cell Biol. 14, 319–327.

Flouri, T. xflouris GitHub Repository .

Grant, B.J., Rodrigues, A.P.C., Elsawy, K.M., Mccammon, J.A., and Caves, L.S.D. (2006). Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22, 2695–2696.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R., and Hofacker, I.L. (2008). The Vienna RNA Websuite. Nucleic Acids Res. 36, W70–W74.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in R. Bioinformatics 30, 2811–2812.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinforma. Oxf. Engl. 32, 2847–2849.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: Protein–RNA interactions at nucleotide resolution. Methods 65, 274–287.

Kagami, M., Sekita, Y., Nishimura, G., Irie, M., Kato, F., Okada, M., Yamamori, S., Kishimoto, H., Nakayama, M., Tanaka, Y., et al. (2008). Deletions and epimutations affecting the human 14q32.2 imprinted region in individuals with paternal and maternal upd(14)-like phenotypes. Nat. Genet. 40, 237–242.

Kalvari, I., Nawrocki, E.P., Argasinska, J., Quinones-Olvera, N., Finn, R.D., Bateman, A., and Petrov, A.I. (2018). Non-coding RNA analysis using the Rfam database. Curr. Protoc. Bioinforma. 62, e51.

Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 49, D192–D200.

Kaur, H., van der Feltz, C., Sun, Y., and Hoskins, A.A. (2021). Network Theory Reveals Principles of Spliceosome Structure and Dynamics. BioRxiv 2021.03.03.433650.

Kerpedjiev, P., Hammer, S., and Hofacker, I.L. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 31, 3377–3379.

Kishore, S., Khanna, A., Zhang, Z., Hui, J., Balwierz, P.J., Stefan, M., Beach, C., Nicholls, R.D., Zavolan, M., and Stamm, S. (2010). The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. Hum. Mol. Genet. 19, 1153–1164.

Klein, D.J. (2001). The kink-turn: a new RNA secondary structure motif. EMBO J. 20, 4214–4221.

Kumar, U. Ujjawal Kumar GitHub Repository. (github.com/AFS-Part-II-Projects/Ujjawal_Kumar)

Lambert, N.J., Robertson, A.D., and Burge, C.B. (2015). RNA Bind-n-Seq: Measuring the Binding Affinity Landscape of RNA Binding Proteins. Methods Enzymol. 558, 465–493.

Lee, F.C.Y., and Ule, J. (2018). Advances in CLIP Technologies for Studies of Protein-RNA Interactions. Mol. Cell 69, 354–369.

Liang, J., Wen, J., Huang, Z., Chen, X., Zhang, B., and Chu, L. (2019). Small Nucleolar RNAs: Insight Into Their Function in Cancer. Front. Oncol. 9.

Liao, Y.-L., Li, Y.-C., Chen, N.-C., and Lu, Y.-C. (2018). Adaptively Banded Smith-Waterman Algorithm for Long Reads and Its Hardware Accelerator. In 2018 IEEE 29th International Conference on Application-Specific Systems, Architectures and Processors (ASAP), (Milan: IEEE), pp. 1–9.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. 6, 26.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47, W636–W641.

Matsumura, S., Ikawa, Y., and Inoue, T. (2003). Biochemical characterization of the kink-turn RNA motif. Nucleic Acids Res. 31, 5544–5551.

Moore, T., Zhang, Y., Fenley, M.O., and Li, H. (2004). Molecular Basis of Box C/D RNA-Protein Interactions. Structure 12, 807–818.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933–2935.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2020). Biostrings (Bioconductor).

Peters, J. (2014). The role of genomic imprinting in biology and disease: an expanding view. Nat. Rev. Genet. 15, 517–530.

Petitjean, M. (1999). On the root mean square quantitative chirality and quantitative symmetry measures. J. Math. Phys. 40, 4587–4595.

Rauhut, R., Fabrizio, P., Dybkov, O., Hartmuth, K., Pena, V., Chari, A., Kumar, V., Lee, C.-T., Urlaub, H., Kastner, B., et al. (2016). Molecular architecture of the Saccharomyces cerevisiae activated spliceosome. Science 353, 1399–1405.

RCSB Protein Data Bank 2N7M: Structure of the core of the U4/U6 di-snRNA.

RCSB Protein Data Bank 5LQW: yeast activated spliceosome.

RCSB Protein Data Bank 6AHD: The Cryo-EM Structure of Human Pre-catalytic Spliceosome (B complex) at 3.8 angstrom resolution.

RNAcentral Consortium (2021). RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res. 49, D212–D220.

Rocha, S.T.D., Edwards, C.A., Ito, M., Ogata, T., and Ferguson-Smith, A.C. (2008). Genomic imprinting at the mammalian Dlk1-Dio3 domain. Trends Genet. 24, 306–316.

Rogelj, B. (2006). Brain-Specific Small Nucleolar RNAs. J. Mol. Neurosci. 28, 103–110.

Sahoo, T., Del Gaudio, D., German, J.R., Shinawi, M., Peters, S.U., Person, R.E., Garnica, A., Cheung, S.W., and Beaudet, A.L. (2008). Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat. Genet. 40, 719–721.

Salazar, G. FastaUtils (https://github.com/GuillemSalazar/FastaUtils).

Schrodinger (2015a). The PyMOL Molecular Graphics System, Version 1.8.

Schrodinger (2015b). The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8.

Schrodinger (2015c). The JyMOL Molecular Graphics Development Component, Version 1.8.

Skryabin, B.V., Gubar, L.V., Seeger, B., Pfeiffer, J., Handel, S., Robeck, T., Karpova, E., Rozhdestvensky, T.S., and Brosius, J. (2007). Deletion of the MBII-85 snoRNA Gene Cluster in Mice Results in Postnatal Growth Retardation. 3, e235.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol. 13, 1–13.

Sundfeld, D., Havgaard, J.H., de Melo, A.C.M.A., and Gorodkin, J. (2016). Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. Bioinforma. Oxf. Engl. 32, 1238–1240.

Surani, M.A., Barton, S.C., and Norris, M.L. (1984). Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature 308, 548–550.

Taylor, W.R., and Hamilton, R.S. (2017). Exploring RNA conformational space under sparse distance restraints. Sci. Rep. 7, 44074.

Vitali, P., Basyuk, E., Le Meur, E., Bertrand, E., Muscatelli, F., Cavaillé, J., and Huttenhofer, A. (2005). ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. J. Cell Biol. 169, 745–753.

Watkins, N.J., Ségault, V., Charpentier, B., Nottrott, S., Fabrizio, P., Bachi, A., Wilm, M., Rosbash, M., Branlant, C., and Lührmann, R. (2000). A Common Core RNP Structure Shared between the Small Nucleoar Box C/D RNPs and the Spliceosomal U4 snRNP. 103, 457–466.

Will, C.L., and Lührmann, R. (2011). Spliceosome Structure and Function. Cold Spring Harb. Perspect. Biol. 3, a003707.

Wozniak, A.K. (2005). Detecting protein-induced folding of the U4 snRNA kink-turn by single-molecule multiparameter FRET measurements. RNA 11, 1545–1554.

Yean, S., and Lin, R.J. (1991). U4 small nuclear RNA dissociates from a yeast spliceosome and does not participate in the subsequent splicing reaction. Mol Cell Biol 11, 5571–5577.

Yin, Q.-F., Yang, L., Zhang, Y., Xiang, J.-F., Wu, Y.-W., Carmichael, G.G., and Chen, L.-L. (2012). Long Noncoding RNAs with snoRNA Ends. Mol. Cell 48, 219–230.

Zhan, X., Yan, C., Zhang, X., Lei, J., and Shi, Y. (2018). Structures of the human pre-catalytic spliceosome and its precursor spliceosome. Cell Res. 28, 1129–1140.

Zuguang Gu (2017). ComplexHeatmap (Bioconductor).

Comparison of Global and Local Alignment. https://www.majordifferences.com/2016/05/difference-between-global-and-local.html#.YF46nmSmP0o.

Appendix A

Spliceosomal snRNA sequences

Human Spliceosomal snRNA Sequences

>URS000071A5D2 Homo sapiens U1 snRNA

ACUCUCACCUGGCAGAGGAGAUGCCAUGAUCACAAAGGUGGUUUUCUCAGGGUGAGACUUAUGGAUUG
CAUUCCGGGUCUGAUGACCCGCACGGUUUUCCCAGAUAUGAGAAACUCAUAGAUGCAGAAACUGCAUA
AUUUAUGGUUGAGAGGGACUGCAUUUCUACUUUACCCUGU

>URS000063164F Homo sapiens U2 snRNA

AUACAAGUUAGGCCUUUUGGCUAAGAUCAAGUGUAGUAUCUGUUCUUAUCAGUUUAAAUACUAGUUAA
CACAUUGGUUUCAGUUUGGCUGUGGGGUCCUUCACUGAACAAUUCUGAUGAAGUCAAAUGAACAAUUC
CUUACAUUUUGUUUGUGCUUUUGGUAACUUGCUUAAGAGUUCUUUCCCAUACCAGUGUCAUAACCUAU
AGUUUCUUCUGUAGCUUUAUAGUUUUGCUAUUCA

>URS0000715A86 Homo sapiens U4 snRNA

AGCUUUGCACAGUGGCAGUAUUGUAGCCAGUGAGGUUCAUCCGAGGCACAAUUAUUGGCUAAUUAAAC
UAUCUCACUGACCAAGGCUACCUAUAACUCAUAUAAACAAAGUGUUUUUCCUUUAUCUUGGUUGAACC
UAGGGCUUAAGAAUUCCUUGAUGUUCCAUGACUUGGAUACAUUAAGGCAAUCUAUUGAUGGUAAACUU
UUAU

>URS0000631BD4 Homo sapiens U5 snRNA

UUACUCUAGUUUCUCUUCAUAUCGCAUAAAUCUUUCGCCUUUUACUAAAGAUUUCCGUGGAGAGGAAU
AAUUCUGAGUUUUAAACCAAUUUUUUAGAGGUCUUGCAUUUUUUAGCAAGGC

>URS00006767A8 Homo sapiens U6 snRNA

GUUCUGCACUUGUAUCCCAGAACUUGAAAUAAAUUUUUUAAAAAAAGAAUAUUAAAAAAAAAAAAGUAC

UCACUUGGGCAGGUCACAUACUGAAAUUGGAGUUAGCACGGCCCCUGUGCAAGGAUGCCAAGCAAAUC

UGUGAAGCGUUUCGUAUUUUU

Murine Spliceosomal snRNA Sequences

>URS0000722349 Mus musculus U1 snRNA

UUCUUAACCUGGCAGGGGAGAUAUCAUGAUCACAAAAGUGUUUUGUUGUUAUUGCUGUUGUUGUUUUG

UGUGUGUUUUAUUUUGUUUUUUGUACAAGACUUAGCCAUUGCACUCCUGAUGUGCUGACCCCUGCCAU

UUUCACAAAUGUGGGAAACUAGACUGCAUAACUUGUGGUAGCAAGACACUCC

>URS0000726205 Mus musculus U2 snRNA

AUCGCUUCUCAGCCUUUUGGCUAAGAUCAAGUGUAGAAAUCCGCCUGCCUCUGCCUCCCGAGUGCUGA

GAUUAAAGGCGUGCACCACCACGCCCGGCAUACGUUGCUUUCUUCUUAUACAUACAUUUGUAUGAAUU

UUUUAAUCAUCUCAACUUCUGAGAAAGAUAACUUUGAAUAAUUGAAAGAAAAUUUUGAUACUGUUUCA

AAUUCUUUUAAAUUUGA

>URS000064AD77 Mus musculus U4 snRNA

AGCUUUGCACAGUGGCAGUAUUGUAGCCAAUGAGGUUUAUCCAAGGUGCGAUUAUUGCUGGUAGAAAA

GAAGAGGAAAAGAAUAGCACCUACUUCUGACACUUUGGUUCUCGAGAAGUGCCAGUGGCUGCCUUCCC

UUAGCAGUUCUCAAGCUCCUGCUUUCCACUUUAUCCAGUUGAGUUGAUUCA

>URS0001BC1F50 Mus musculus U5 snRNA

AUACUCUGGUUUCUCUUCAGAUCGUAUAAAUCUUUCGCCUUUUACUAAAGAUUUCCGUGGAGAGGAAC

AAUUCUGAGUCUUACACUAAUUUUUUUGAGGCCUUGCUUUAGCAAGGCU

>URS0000710FEE Mus musculus U6 snRNA

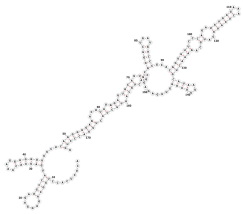
GUGCUUACCUCAGCAGCACAUACACCGACAUUGGAACAAUACAUGGCCACUGCACACAAUACAGAGAU

CAUCAUGGCCACUGCACAAGGAUGCCAUGCAAAUUCAUCAACGGUCCCAUAUUUUU

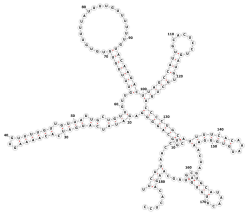
Appendix B

Spliceosomal snRNA structures

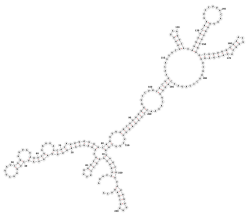
2D MFE Structures



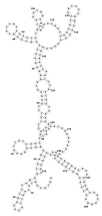
Human U1 snRNA



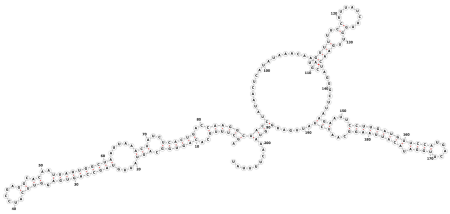
Mouse U1 snRNA



Human U2 snRNA

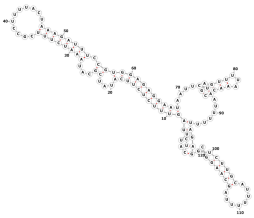


Mouse U2 snRNA



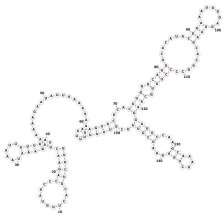
Human U4 snRNA

Mouse U4 snRNA

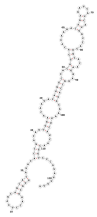


Human U5 snRNA

Mouse U5 snRNA

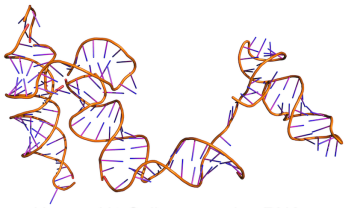


Human U6 snRNA

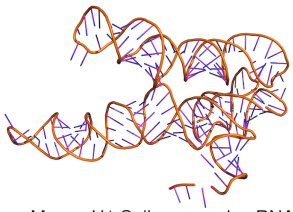


Mouse U6 snRNA

3D Predicted Structures

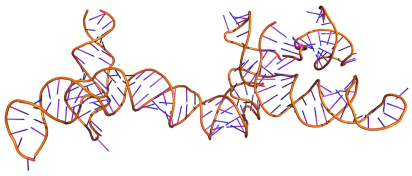


Human U1 Spliceosomal snRNA

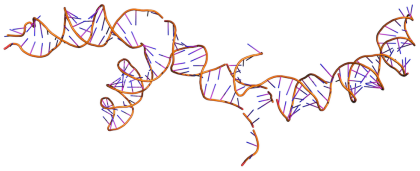


Mouse U1 Spliceosomal snRNA

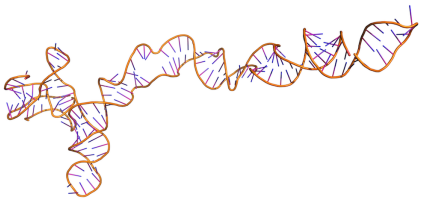
Human U2 Spliceosomal snRNA



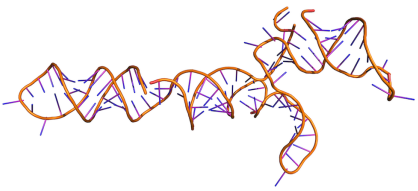
Mouse U2 Spliceosomal snRNA



Human U4 Spliceosomal snRNA

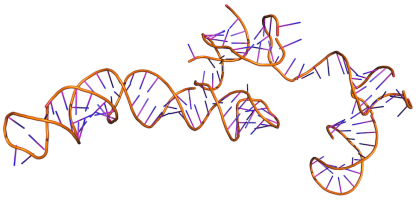


Mouse U4 Spliceosomal snRNA

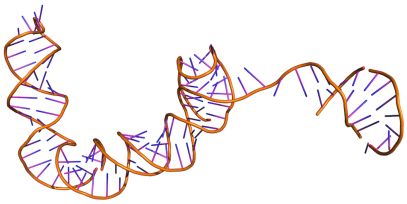


Human U5 Spliceosomal snRNA

Mouse U5 Spliceosomal snRNA



Human U6 Spliceosomal snRNA



Mouse U6 Spliceosomal snRNA

Appendix C

Example Foldalign “hit list”

Seq1	Seq1 Start	Seq1 End	Seq2	Seq2 Start	Seq2 End	Foldalign Score	p-Score
hU4	161	174	hSNORD113-1	35	48	132	0.654
hU4	25	64	hSNORD113-1	3	43	131	0.663
hU4	120	136	hSNORD113-1	38	54	106	0.857
hU4	194	208	hSNORD113-1	34	48	105	0.863
hU4	98	116	hSNORD113-1	27	45	89	0.944
hU4	155	179	hSNORD113-2	47	71	279	0.035
hU4	5	17	hSNORD113-2	57	69	101	0.891
hU4	103	117	hSNORD113-2	34	48	100	0.896
hU4	101	115	hSNORD113-2	51	65	99	0.901
hU4	20	32	hSNORD113-2	12	24	96	0.917
hU4	161	174	hSNORD113-3	34	47	236	0.106
hU4	17	60	hSNORD113-3	14	55	110	0.844
hU4	123	136	hSNORD113-3	40	53	105	0.874
hU4	31	39	hSNORD113-3	63	71	95	0.925
hU4	102	116	hSNORD113-3	30	44	82	0.969
hU4	20	40	hSNORD113-4	11	30	122	0.794
hU4	162	199	hSNORD113-4	21	57	109	0.874
hU4	43	75	hSNORD113-4	23	55	101	0.914
hU4	31	39	hSNORD113-4	66	74	95	0.938
hU4	124	136	hSNORD113-4	41	53	88	0.96

Table C.1 – Table containing the top 5 hits for a Foldalign comparison between the Human U4 spliceosomal snRNA and 4 snoRNAs of the SNORD113 cluster.

Appendix D

R code used in the project

D.1 *R* code for local pairwise sequence alignment on the Spliceosomal snRNA sequences

```
# Matrix from https://github.com/xflouris/gapmis/blob/master/EDNAFULL.h (note T changed to
    U for RNA)
EDNAFULL\_matrix <- matrix( c(
5, -4, -4, -4, -4, 1, 1, -4, -4, 1, -4, -1, -1, -1, -2,
 -4, 5, -4, -4, -4, 1, -4, 1, 1, -4, -1, -4, -1, -1, -2,
 -4, -4, 5, -4, 1, -4, 1, -4, 1, -4, -1, -1, -4, -1, -2,
 -4, -4, -4, 5, 1, -4, -4, 1, -4, 1, -1, -1, -1, -4, -2,
 -4, -4, 1, 1, -1, -4, -2, -2, -2, -2, -1, -1, -3, -3, -1,
 1, 1, -4, -4, -4, -1, -2, -2, -2, -2, -3, -3, -1, -1, -1,
 1, -4, 1, -4, -2, -2, -1, -4, -2, -2, -3, -1, -3, -1, -1,
 -4, 1, -4, 1, -2, -2, -4, -1, -2, -2, -1, -3, -1, -3, 1,
 -4, 1, 1, -4, -2, -2, -2, -2, -1, -4, -1, -3, -3, -1, -1,
 1, -4, -4, 1, -2, -2, -2, -2, -4, -1, -3, -1, -1, -3, -1,
 -4, -1, -1, -1, -1, -3, -3, -1, -1, -3, -1, -2, -2, -2, -1,
 -1, -4, -1, -1, -1, -3, -1, -3, -3, -1, -2, -1, -2, -2, -1,
 -1, -1, -4, -1, -3, -1, -3, -1, -3, -1, -2, -2, -1, -2, -1,
 -1, -1, -1, -4, -3, -1, -1, -3, -1, -3, -2, -2, -2, -1, -1,
 -2, -2, -2, -2, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1 ),
nrow=15,ncol=15,byrow=TRUE)
colnames(EDNAFULL\_matrix) <- c("A","U","G","C","S","W","R","Y","K","M","B","V","H","D","N"
    )
rownames(EDNAFULL\_matrix) <- c("A","U","G","C","S","W","R","Y","K","M","B","V","H","D","N"
    )
# Read .fasta file
fastaFile <- readRNAStringSet("Spliceosomal snRNAs/Spliceosomal snRNAs Seqs/Spliceosomal
    snRNAs
Seqs.fasta", format="fasta")
sequences <- data.frame(Headers=names(fastaFile), Sequences=paste(fastaFile))
str(sequences)
# Run pairwise alignment using the .fasta file with sequences
identities <- matrix(nrow=nrow(sequences),ncol=nrow(sequences))
colnames(identities) <- sequences$Headers
rownames(identities) <- sequences$Headers
for (iseq in c(1:nrow(sequences)))
{for (jseq in c(1:nrow(sequences)))
 {aln <- pairwiseAlignment(pattern = sequences$Sequences[iseq], subject = sequences$
     Sequences[jseq], type="local", gapOpening=10, gapExtension=0.5, substitutionMatrix=
     EDNAFULL_matrix)
 identities[iseq,jseq] <- pid(aln)
 message(paste0(sequences$Headers[iseq], " vs ", sequences$Headers[jseq], "ID=", identities
     [iseq,jseq]))}}
```

Fig. D.1 – R code using the pairwiseAlignment function of Biostrings (Pagès et al., 2020) and a scoring matrix available from https://github.com/xflouris/gapmis/blob/master/EDNAFULL.h. This code is used to to carry out local pairwise sequence alignment on combinations of the spliceosomal snRNA sequences using the Smith-Waterman Algorithm (Figure 2.2. Sequences were read in from the relevant .fasta files using the FastaUtils package.

D.2 *R* code for plotting heatmaps

```
library(circlize), library(ComplexHeatmap), library(viridis), library(bio3d)

png(filename = "Spliceosomal snRNA Sequence Similarity.png", width = 5000, height = 5000)
col_fun = colorRamp2(c(0, 0.5, 1), c("red", "white", "blue"))
Heatmap(as.matrix(Sequence_Similarity_Matrix), "Spliceosomal snRNA Sequence Similarity",
border = TRUE,
width = unit(4000, "points"),
height = unit(4250, "points"),
heatmap_legend_param = list(
at = c(0, 20, 40, 60, 80, 100),
title = "% Sequence Similarity",
title_position = "lefttop-rot",
title_gp = gpar(fontsize = 75),
legend_height = unit(3500, "points"),
grid_width = unit(175, "points"),
labels_gp = gpar(fontsize = 80)),
row_names_gp = gpar(fontsize = 80),
row_names_centered = TRUE,
row_names_max_width = unit(400, "points"),
row_names_side = c("left"),
column_names_gp = gpar(fontsize = 80),
column_names_centered = TRUE,
column_names_max_height = unit(400, "points"),
cluster_rows=T, cluster_columns=T,
column_km = 2,
column_gap = unit(15, "points"),
row_km = 2,
row_gap = unit(15, "points"),
row_dend_gp = gpar(col="red", lwd = 7.5),
row_dend_width = unit(400, "points"),
row_dend_side = c("left"),
column_dend_gp = gpar(col="red", lwd = 7.5),
column_dend_side = c("bottom"),
column_dend_height = unit(400, "points"),
col=col_fun )
dev.off()
```

Fig. D.2 – R code for plotting heatmaps. The data was read in from a .csv matrix (using read.csv); specific parameters of this code were adjusted for individual heatmaps.

D.3 *R* code utilising *DNA2RNA* to convert DNA sequences to RNA sequences

```
DNA2RNA(file="snoRNAs/snoRNA Seqs/SNORD113_hs.fasta", out="snoRNAs/snoRNA Seqs/SNORD113_hs.
    rna.fasta")
DNA2RNA(file="snoRNAs/snoRNA Seqs/SNORD114_hs.fasta", out="snoRNAs/snoRNA Seqs/SNORD114_hs.
    rna.fasta")
DNA2RNA(file="snoRNAs/snoRNA Seqs/SNORD115_hs.fasta", out="snoRNAs/snoRNA Seqs/SNORD115_hs.
    rna.fasta")
DNA2RNA(file="snoRNAs/snoRNA Seqs/SNORD116_hs.fasta", out="snoRNAs/snoRNA Seqs/SNORD116_hs.
    rna.fasta")
DNA2RNA(file="snoRNAs/snoRNA Seqs/SNORD115_mm.fasta", out="snoRNAs/snoRNA Seqs/SNORD115_mm.
    rna.fasta")
DNA2RNA(file="snoRNAs/snoRNA Seqs/SNORD116_mm.fasta", out="snoRNAs/snoRNA Seqs/SNORD116_mm.
    rna.fasta")
```

D.4 *R* code to run pairwise 2D structural distance analysis

```
library("Biostrings"), library("FastaUtils"), library("seqinr")

# Matrix from https://github.com/xflouris/gapmis/blob/master/EDNAFULL.h (note T changed to
    U for RNA)
EDNAFULL_matrix <- matrix( c(
 5, -4, -4, -4, -4, 1, 1, -4, -4, 1, -4, -1, -1, -1, -2,
 -4, 5, -4, -4, -4, 1, -4, 1, 1, -4, -1, -4, -1, -1, -2,
 -4, -4, 5, -4, 1, -4, 1, -4, 1, -4, -1, -1, -4, -1, -2,
 -4, -4, -4, 5, 1, -4, -4, 1, -4, 1, -1, -1, -1, -4, -2,
 -4, -4, 1, 1, -1, -4, -2, -2, -2, -2, -1, -1, -3, -3, -1,
 1, 1, -4, -4, -4, -1, -2, -2, -2, -2, -3, -3, -1, -1, -1,
 1, -4, 1, -4, -2, -2, -1, -4, -2, -2, -3, -1, -3, -1, -1,
 -4, 1, -4, 1, -2, -2, -4, -1, -2, -2, -1, -3, -1, -3, 1,
 -4, 1, 1, -4, -2, -2, -2, -2, -1, -4, -1, -3, -3, -1, -1,
 1, -4, -4, 1, -2, -2, -2, -2, -4, -1, -3, -1, -1, -3, -1,
 -4, -1, -1, -1, -1, -3, -3, -1, -1, -3, -1, -2, -2, -2, -1,
 -1, -4, -1, -1, -1, -3, -1, -3, -3, -1, -2, -1, -2, -2, -1,
 -1, -1, -4, -1, -3, -1, -3, -1, -3, -1, -2, -2, -1, -2, -1,
 -1, -1, -1, -4, -3, -1, -1, -3, -1, -3, -2, -2, -2, -1, -1,
 -2, -2, -2, -2, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1 ),
 nrow=15,ncol=15,byrow=TRUE)
colnames(EDNAFULL_matrix) <- c("A","U","G","C","S","W","R","Y","K","M","B","V","H","D","N")
rownames(EDNAFULL_matrix) <- c("A","U","G","C","S","W","R","Y","K","M","B","V","H","D","N")
# Read .fasta file
fastaFile <- readRNAStringSet("Spliceosomal snRNAs/Spliceosomal snRNAs Seqs/Spliceosomal
    snRNAs Seqs.fasta", format="fasta")
sequences <- data.frame(Headers=names(fastaFile), Sequences=paste(fastaFile))
str(sequences)
identities <- matrix(nrow=nrow(sequences),ncol=nrow(sequences))
colnames(identities) <- sequences$Headers
rownames(identities) <- sequences$Headers
# Run 2D for whole fasta file to get distances
distances <- matrix(nrow=nrow(sequences),ncol=nrow(sequences))
colnames(distances) <- sequences$Headers
rownames(distances) <- sequences$Headers
for (iseq in c(1:nrow(sequences)))
{write.fasta(sequences = sequences$Sequences[iseq], names=sequences$Headers[iseq], file.out
    ="temp_i_seq.fasta")
 system("RNAfold < temp_i_seq.fasta > temp_i_seq.ss", intern=T)
 for (jseq in c(1:nrow(sequences)))
 {aln <- pairwiseAlignment(pattern = sequences$Sequences[iseq], subject = sequences$
     Sequences[jseq], type="local", gapOpening=10, gapExtension=0.5, substitutionMatrix=
     EDNAFULL_matrix)
 identities[iseq,jseq] <- pid(aln)
 message(paste0(sequences$Headers[iseq], " vs ", sequences$Headers[jseq], " ID=",
     identities[iseq,jseq]))
 write.fasta(sequences = sequences$Sequences[jseq], names=sequences$Headers[jseq], file.out
     ="temp_j_seq.fasta")
 system("RNAfold < temp_j_seq.fasta > temp_j_seq.ss", intern=T)
 system("cat temp_i_seq.ss temp_j_seq.ss > temp_ij_seq.ss")
 system("RNAdistance -Xp < temp_ij_seq.ss > temp_ij_seq.ss.mat 2>&1", intern=T)
 distmatrix <- read.table("temp_ij_seq.ss.mat", comment.char = ">")
 distances[iseq,jseq] <- distmatrix[1,c("V2")]}}
```

Fig. D.4 – *R* code to carry out pairwise 2D structural distance analysis using *RNAdistance*. The values were tabulated in a distance matrix and heatmaps plotted.

Appendix E

bash code used in the project

E.1 *bash* code to calculate 2D MFE structures of the spliceosomal snRNAs

```
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/hU1\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/hU1\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/hU2\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/hU2\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/hU4\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/hU4\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/hU5\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/hU5\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/hU6\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/hU6\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/mU1\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/mU1\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/mU2\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/mU2\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/mU4\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/mU4\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/mU5\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/mU5\ snRNA\ MFE.rnafold.ss
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/mU6\ snRNA.fasta >
    ~/3D\ Structure/Candidates/2D\ Structures/mU6\ snRNA\ MFE.rnafold.ss
```

Fig. E.1 – *bash* code utilising *RNAfold* to calculate 2D MFE structures of the spliceosomal snRNAs.

E.2 *bash* code to calculate 2D structural distances between the spliceosomal snRNAs

```
RNAfold -T37 < Pairwise\ Combinations/hU1\ x\ hSNORD113.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU2\ x\ hSNORD113.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU4\ x\ hSNORD113.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU5\ x\ hSNORD113.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU6\ x\ hSNORD113.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU1\ x\ hSNORD114.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU2\ x\ hSNORD114.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU4\ x\ hSNORD114.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU5\ x\ hSNORD114.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU6\ x\ hSNORD114.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU1\ x\ hSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU2\ x\ hSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU4\ x\ hSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU5\ x\ hSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU6\ x\ hSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU1\ x\ hSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU2\ x\ hSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU4\ x\ hSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU5\ x\ hSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/hU6\ x\ hSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU1\ x\ mSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU2\ x\ mSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU4\ x\ mSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU5\ x\ mSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU6\ x\ mSNORD115.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU1\ x\ mSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU2\ x\ mSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU4\ x\ mSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU5\ x\ mSNORD116.fasta | RNAdistance -Xm
RNAfold -T37 < Pairwise\ Combinations/mU6\ x\ mSNORD116.fasta | RNAdistance -Xm
```

Fig. E.2 – *bash* code utilising *RNAdistance* to calculate 2D structural distances.

E.3 *bash* code to predict duplexes formed between the U4 spliceosomal snRNA and the candidate snoRNAs

```
RNAcofold -p --id-prefix=hU4_hSNORD114_06 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD114-06.seq
RNAcofold -p --id-prefix=hU4_hSNORD114_11 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD114-11.seq
RNAcofold -p --id-prefix=hU4_hSNORD114_13 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD114-13.seq
RNAcofold -p --id-prefix=hU4_hSNORD114_26 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD114-26.seq
RNAcofold -p --id-prefix=hU4_hSNORD114_27 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD114-27.seq
RNAcofold -p --id-prefix=hU4_hSNORD115_45 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD115-45.seq
RNAcofold -p --id-prefix=hU4_hSNORD116_10 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD116-10.seq
RNAcofold -p --id-prefix=hU4_hSNORD116_13 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD116-13.seq
RNAcofold -p --id-prefix=hU4_hSNORD116_23 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD116-23.seq
RNAcofold -p --id-prefix=hU4_hSNORD116_26 < ~/3D\ Structure/Annealing/Sequences/
    hU4_hSNORD116-26.seq
RNAcofold -p --id-prefix=mU4_mSNORD115_01 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD115-01.seq
RNAcofold -p --id-prefix=mU4_mSNORD115_02 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD115-02.seq
RNAcofold -p --id-prefix=mU4_mSNORD115_03 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD115-03.seq
RNAcofold -p --id-prefix=mU4_mSNORD115_04 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD115-04.seq
RNAcofold -p --id-prefix=mU4_mSNORD115_05 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD115-05.seq
RNAcofold -p --id-prefix=mU4_mSNORD115_06 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD115-06.seq
RNAcofold -p --id-prefix=mU4_mSNORD115_07 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD115-07.seq
RNAcofold -p --id-prefix=mU4_mSNORD116_03 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD116-03.seq
RNAcofold -p --id-prefix=mU4_mSNORD116_04 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD116-04.seq
RNAcofold -p --id-prefix=mU4_mSNORD116_05 < ~/3D\ Structure/Annealing/Sequences/
    mU4_mSNORD116-05.seq
```

Fig. E.3 – *bash* code utilising *RNAcofold* to predict duplexes. The predicted duplexes are shown in Appendix F.

Appendix F

Duplex formation prediction

F.1 Structural figures of predicted duplexes

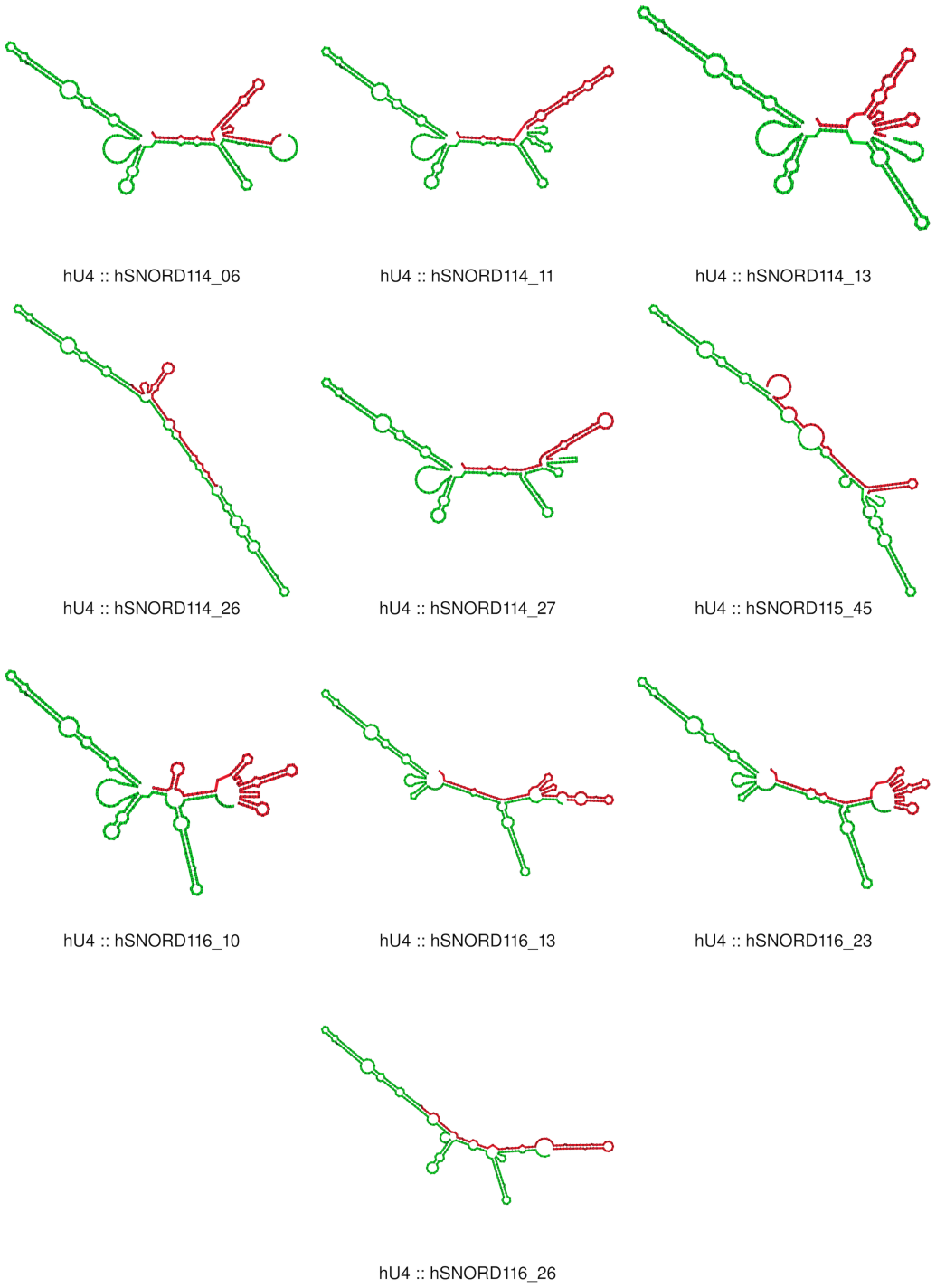


Fig. F.1 – Human Candidate snoRNAs with Human U4 Spliceosomal snRNA

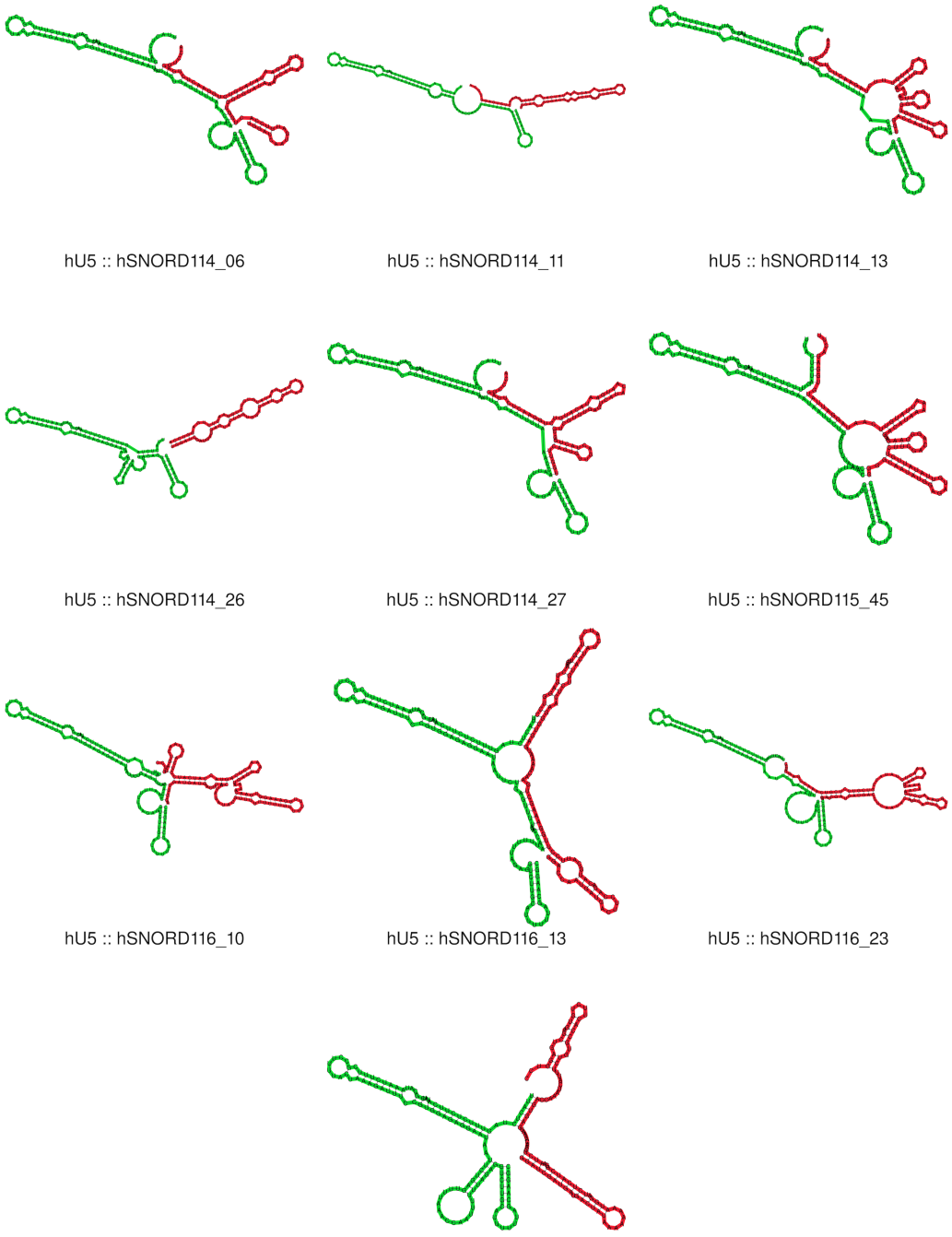


Fig. F.2 – Human Candidate snoRNAs with Human U5 Spliceosomal snRNA

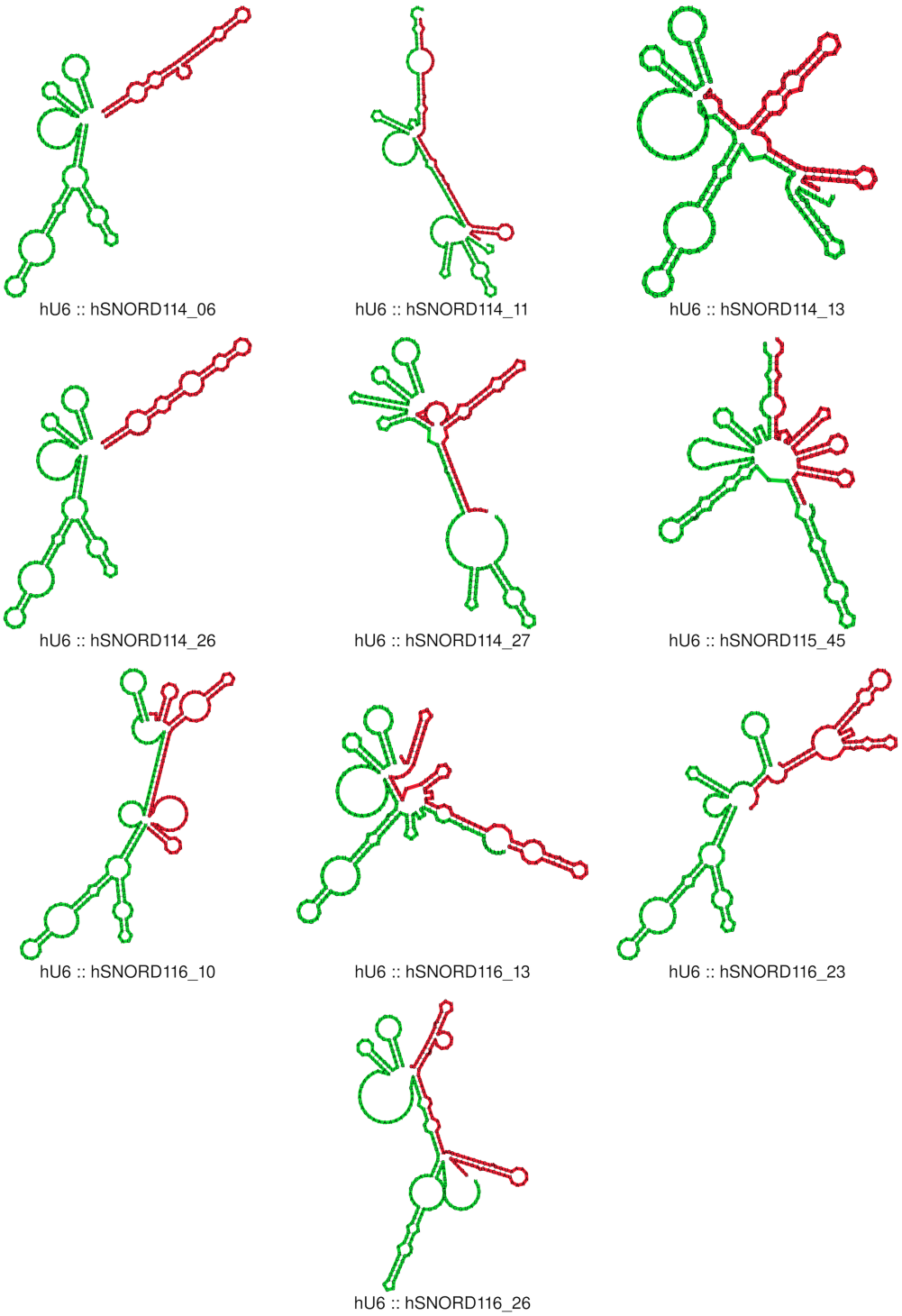


Fig. F.3 – Human Candidate snoRNAs with Human U6 Spliceosomal snRNA

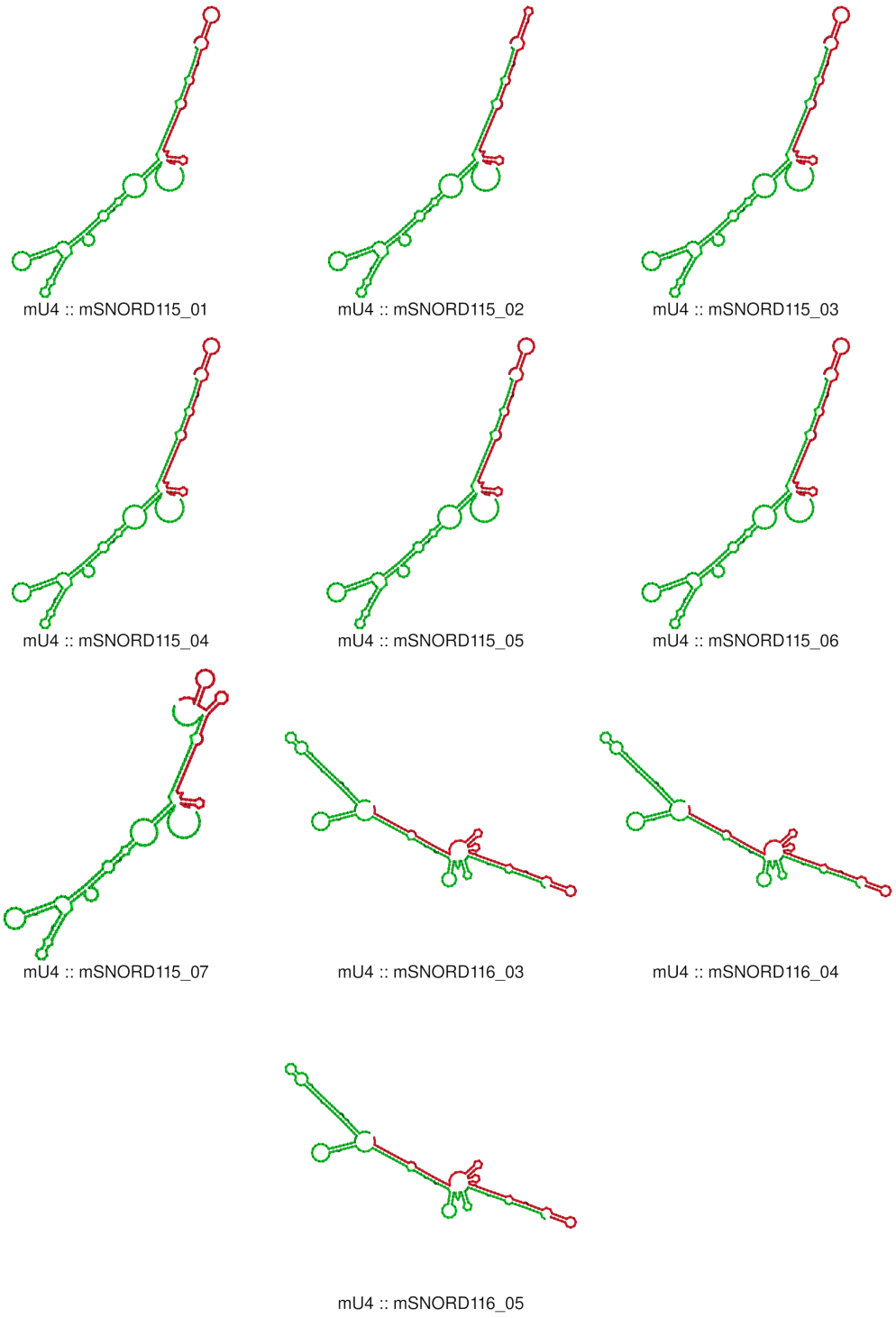


Fig. F.4 – Murine Candidate snoRNAs with Murine U4 Spliceosomal snRNA

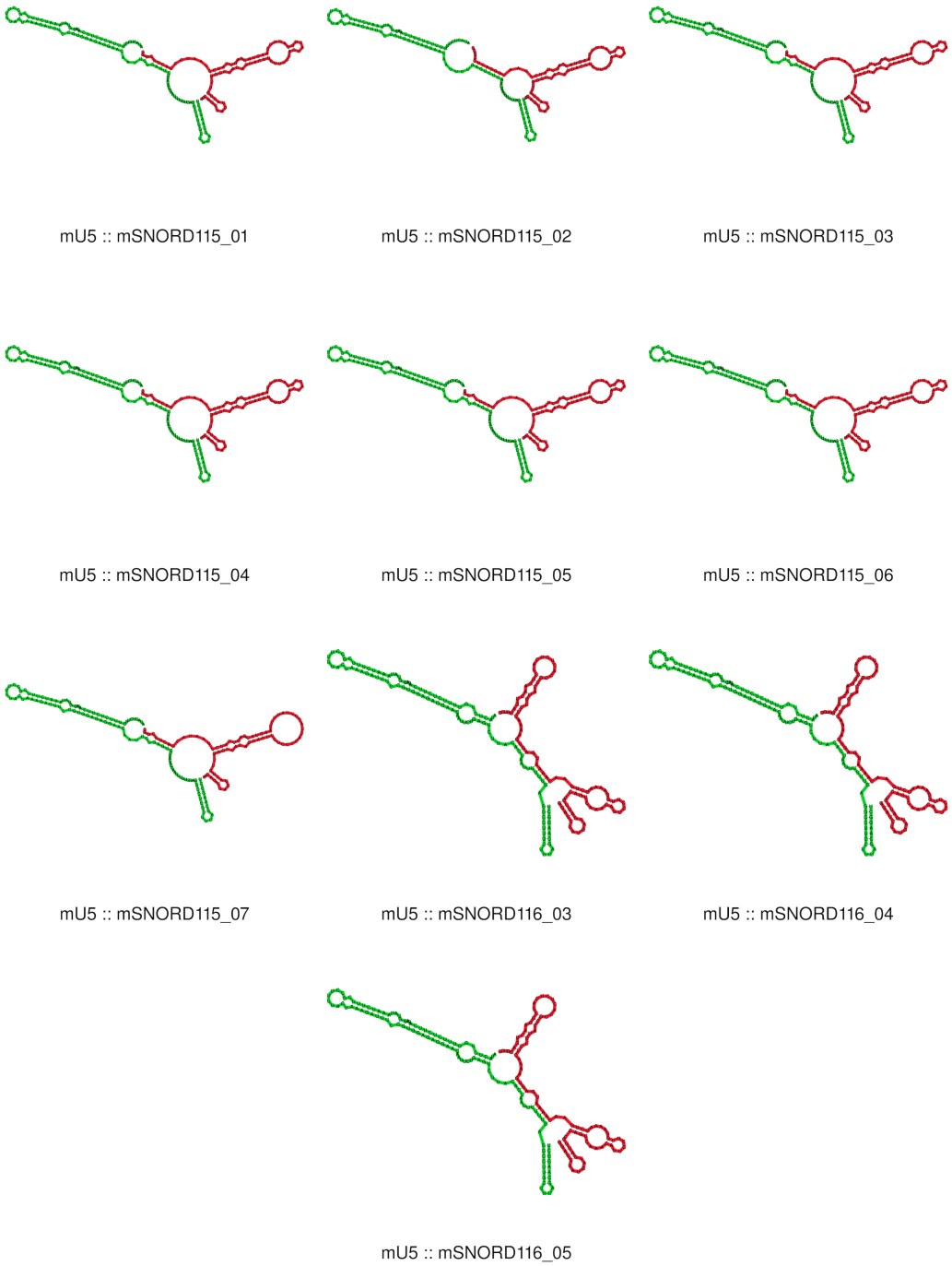


Fig. F.5 – Murine Candidate snoRNAs with Murine U5 Spliceosomal snRNA

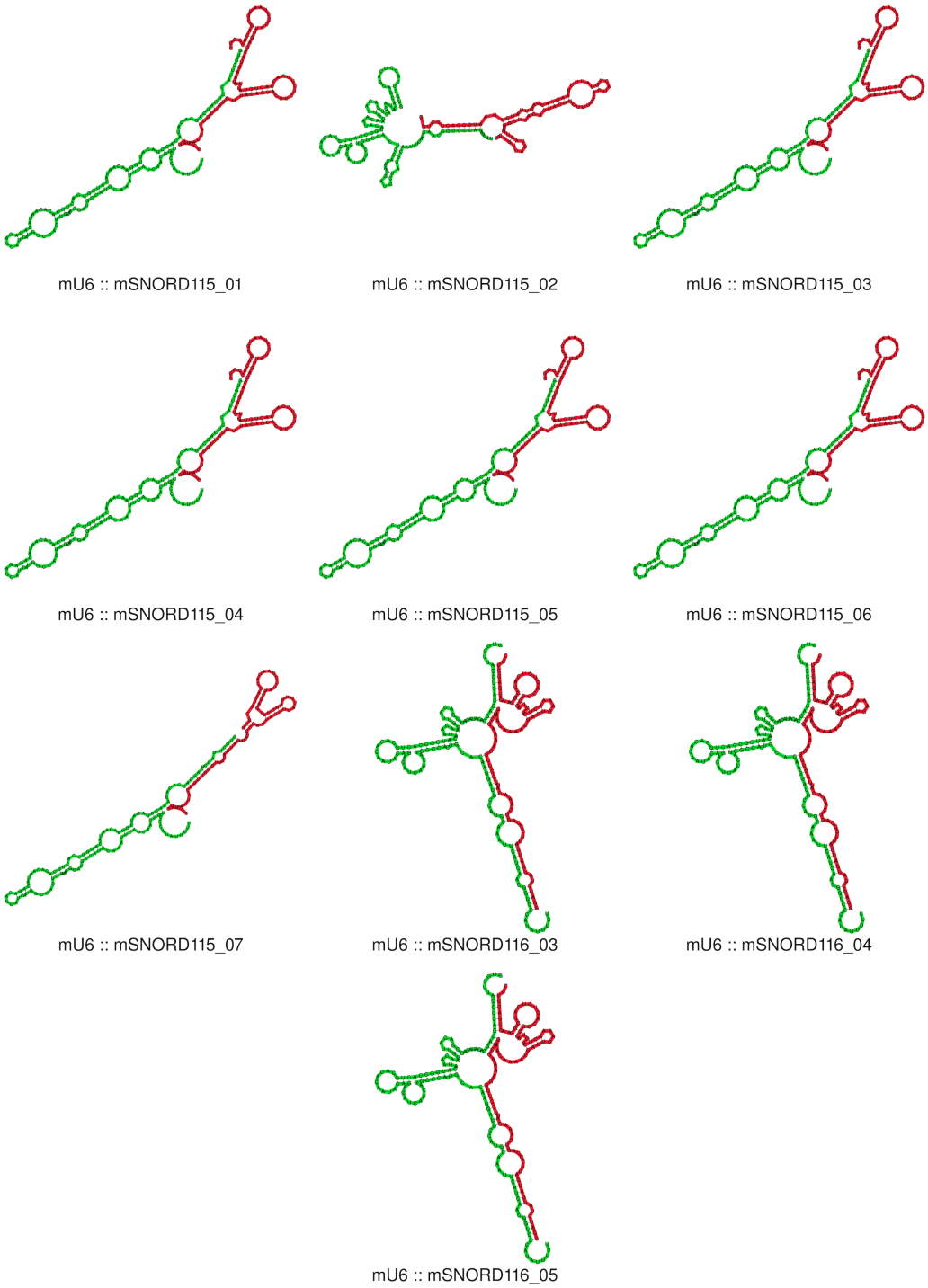


Fig. F.6 – Murine Candidate snoRNAs with Murine U6 Spliceosomal snRNA

F.2 ΔG of binding values

Candidate snoRNA	ΔG of binding (kJ/mol)		
	U4 spliceosomal snRNA	*U5 spliceosomal snRNA*	*U6 spliceosomal snRNA*
hSNORD114_06	-13.13776	-4.51872	6.52704
hSNORD114_11	-22.04968	-14.2256	-17.94936
hSNORD114_13	-16.69416	-12.04992	-5.6484
hSNORD114_26	-23.8488	6.73624	0.16736
hSNORD114_27	-28.8696	-13.3888	-17.61464
hSNORD115_45	-29.83192	-15.0624	-33.01176
hSNORD116_10	-28.07464	-11.9244	-18.70248
hSNORD116_13	-23.09568	-5.06264	-14.60216
hSNORD116_23	-28.53488	-9.45584	-9.49768
hSNORD116_26	-31.96576	-11.12944	-15.22976
mSNORD115_01	-30.83608	-18.28408	-22.84464
mSNORD115_02	-28.53488	-9.74872	-24.35088
mSNORD115_03	-30.83608	-18.28408	-22.84464
mSNORD115_04	-30.83608	-18.28408	-22.84464
mSNORD115_05	-30.83608	-18.28408	-22.84464
mSNORD115_06	-30.83608	-18.28408	-22.84464
mSNORD115_07	-27.23784	-15.85736	-20.96184
mSNORD116_03	-43.26256	-13.89088	47.61392
mSNORD116_04	-43.26256	-13.89088	47.61392
mSNORD116_05	-43.26256	-13.89088	47.61392

Table F.1 ΔG of binding values obtained using *RNAcofold*, part of the ViennaRNA package (Lorenz et al., 2011). Positive values indicate an endergonic reaction, with the products (snoRNA-spliceosomal snRNA duplex) having a greater Gibbs free energy than the reactants (individual RNAs), and thus the reaction is not spontaneous and will not occur in a physiological environment. Negative values indicate an exergonic reaction (products have less Gibbs free energy than reactants) and thus a spontaneous reaction/duplex formation.

Appendix G

Three-dimensional structure prediction and comparison

```
use strict;
my @Restraints;
my $cnt = 0;
while (<>){
    my $line = $_;
    chomp($line);
    if($line !~ /ENERGY/){
        $line =~ s/^\s+//g;
        my @data = split(/\s+/, $line);
        if($data[4] != 0){
            if($data[4] > $data[5]){
                $Restraints[$data[5]] = $data[4];}
                else {
                $Restraints[$data[4]] = $data[5];}
            #print "$cnt ::: $line\n";}
        }$cnt++;}
my $i=0;
for($i=0; $i<=$#Restraints; $i++){
    if($Restraints[$i] > 0){
        print "SLOPE A/", $i, "/MB A/", $Restraints[$i], "/MB 5.0 10.0 1.0\n";
        print "WELL A/", $i, "/MB A/", $Restraints[$i], "/MB 5.0 10.0 5.0\n";}}
```

Fig. G.1 – *Perl* script used to convert 2D MFE structures in dot/bracket notation to restraints files for *SimRNA* 3D structure prediction.

```
NUMBER_OF_ITERATIONS 1600000
TRA_WRITE_IN_EVERY_N_ITERATIONS 16000
INIT_TEMP 1.35
FINAL_TEMP 0.90
BONDS_WEIGHT 1.0
ANGLES_WEIGHT 1.0
TORS_ANGLES_WEIGHT 0.0
ETA_THETA_WEIGHT 0.40
```

Fig. G.2 – config.dat file for *SimRNA* runs. This sets the parameters for the 3D structural prediction runs, e.g. a starting temperature of 1.35 and a final temperature of 0.90

```
RNAfold --noLP ~/3D\ Structure/Candidates/Sequences/_____.fasta > ~/3D\ Structure/
    Candidates/2D\ Structures/_____\ MFE.rnafold.ss
```

Fig. G.3 – *bash* code utilising *RNAfold* to calculate 2D MFE structures of the candidate snoRNAs. _____ represents the name of the individual candidate snoRNA being investigated

```
RNAfold < ~/3D\ Structure/Candidates/Sequences/hU4\ snRNA.fasta | b2ct | perl ~/3D\
    Structure/Brackets2Restraints.pl > ~/3D\ Structure/Candidates/Restraint\ Files/hU4\
    snRNA.pairs.con.cut
```

Fig. G.4 – *bash* code utilising *RNAfold* and the *Perl* script to obtain a restraints file for use in 3D structure prediction using *SimRNA*

```
SimRNA -s ~/3D\ Structure/Candidates/Sequences/_____.fasta -c config.dat -S ~/3D\ Structure
    /Candidates/2D\ Structures/_____\ MFE.rnafold.ss -r ~/3D\ Structure/Candidates/
    Restraint\ Files/_____.pairs.con.cut

trafl_extract_lowestE_frame.py _____.fasta.trafl

SimRNA_trafl2pdbs _____.fasta-000001.pdb _____.fasta_minE.trafl 1 AA
```

Fig. G.5 – *bash* code utilising SimRNA to predict 3D structures of the candidate snoRNAs. _____ represents the name of the individual candidate snoRNA being investigated

```
SimRNA -s ~/3D\ Structure/Candidates/Sequences/___\ snRNA.fasta -c config.dat -S ~/3D\
    Structure/Candidates/2D\ Structures/___\ snRNA\ MFE.rnafold.ss -r ~/3D\ Structure/
    Candidates/Restraint\ Files/___\ snRNA.pairs.con.cut

trafl_extract_lowestE_frame.py ___\ snRNA.fasta.trafl

SimRNA_trafl2pdbs ___\ snRNA.fasta-000001.pdb ___\ snRNA.fasta_minE.trafl 1 AA
```

Fig. G.6 – *bash* code utilising SimRNA to predict 3D structures of the spliceosomal snRNAs. _____ represents the name of the individual snRNA being investigated

```
align hU4_snRNA.fasta_minE-000001_AA and resi 151-183, hSNORD114_06.fasta_minE-000001_AA
    and resi 22-54
align hU4_snRNA.fasta_minE-000001_AA and resi 12-40, hSNORD114_11.fasta_minE-000001_AA and
    resi 24-52
align hU4_snRNA.fasta_minE-000001_AA and resi 7-61, hSNORD114_13.fasta_minE-000001_AA and
    resi 3-58
align hU4_snRNA.fasta_minE-000001_AA and resi 151-183, hSNORD114_26.fasta_minE-000001_AA
    and resi 29-62
align hU4_snRNA.fasta_minE-000001_AA and resi 20-51, hSNORD114_27.fasta_minE-000001_AA and
    resi 7-38
align hU4_snRNA.fasta_minE-000001_AA and resi 156-177, hSNORD115_45.fasta_minE-000001_AA
    and resi 40-61
align hU4_snRNA.fasta_minE-000001_AA and resi 123-140, hSNORD116_10.fasta_minE-000001_AA
    and resi 82-99
align hU4_snRNA.fasta_minE-000001_AA and resi 123-140, hSNORD116_13.fasta_minE-000001_AA
    and resi 72-89
align hU4_snRNA.fasta_minE-000001_AA and resi 84-143, hSNORD116_23.fasta_minE-000001_AA and
     resi 34-92
align hU4_snRNA.fasta_minE-000001_AA and resi 156-177, hSNORD116_26.fasta_minE-000001_AA
    and resi 4-43
align mU4_snRNA.fasta_minE-000001_AA and resi 148-161, mSNORD115_01.fasta_minE-000001_AA
    and resi 2-15
align mU4_snRNA.fasta_minE-000001_AA and resi 148-161, mSNORD115_02.fasta_minE-000001_AA
    and resi 2-15
align mU4_snRNA.fasta_minE-000001_AA and resi 148-161, mSNORD115_03.fasta_minE-000001_AA
    and resi 2-15
align mU4_snRNA.fasta_minE-000001_AA and resi 148-161, mSNORD115_04.fasta_minE-000001_AA
    and resi 2-15
align mU4_snRNA.fasta_minE-000001_AA and resi 148-161, mSNORD115_05.fasta_minE-000001_AA
    and resi 2-15
align mU4_snRNA.fasta_minE-000001_AA and resi 148-161, mSNORD115_06.fasta_minE-000001_AA
    and resi 2-15
align mU4_snRNA.fasta_minE-000001_AA and resi 148-161, mSNORD115_07.fasta_minE-000001_AA
    and resi 2-15
align mU4_snRNA.fasta_minE-000001_AA and resi 97-109, mSNORD116_03.fasta_minE-000001_AA and
     resi 82-94
align mU4_snRNA.fasta_minE-000001_AA and resi 97-109, mSNORD116_04.fasta_minE-000001_AA and
     resi 82-94
align mU4_snRNA.fasta_minE-000001_AA and resi 97-109, mSNORD116_05.fasta_minE-000001_AA and
     resi 82-94
```

Fig. G.7 – Code in *Pymol* aligning the 3D predicted structures of each individual candidate snoRNA to their respective U4 spliceosomal snRNA. This was done using the alignment co-ordinates in Table 3.2

Appendix H

snoRNA structures

H.1 3D structures of candidate snoRNAs

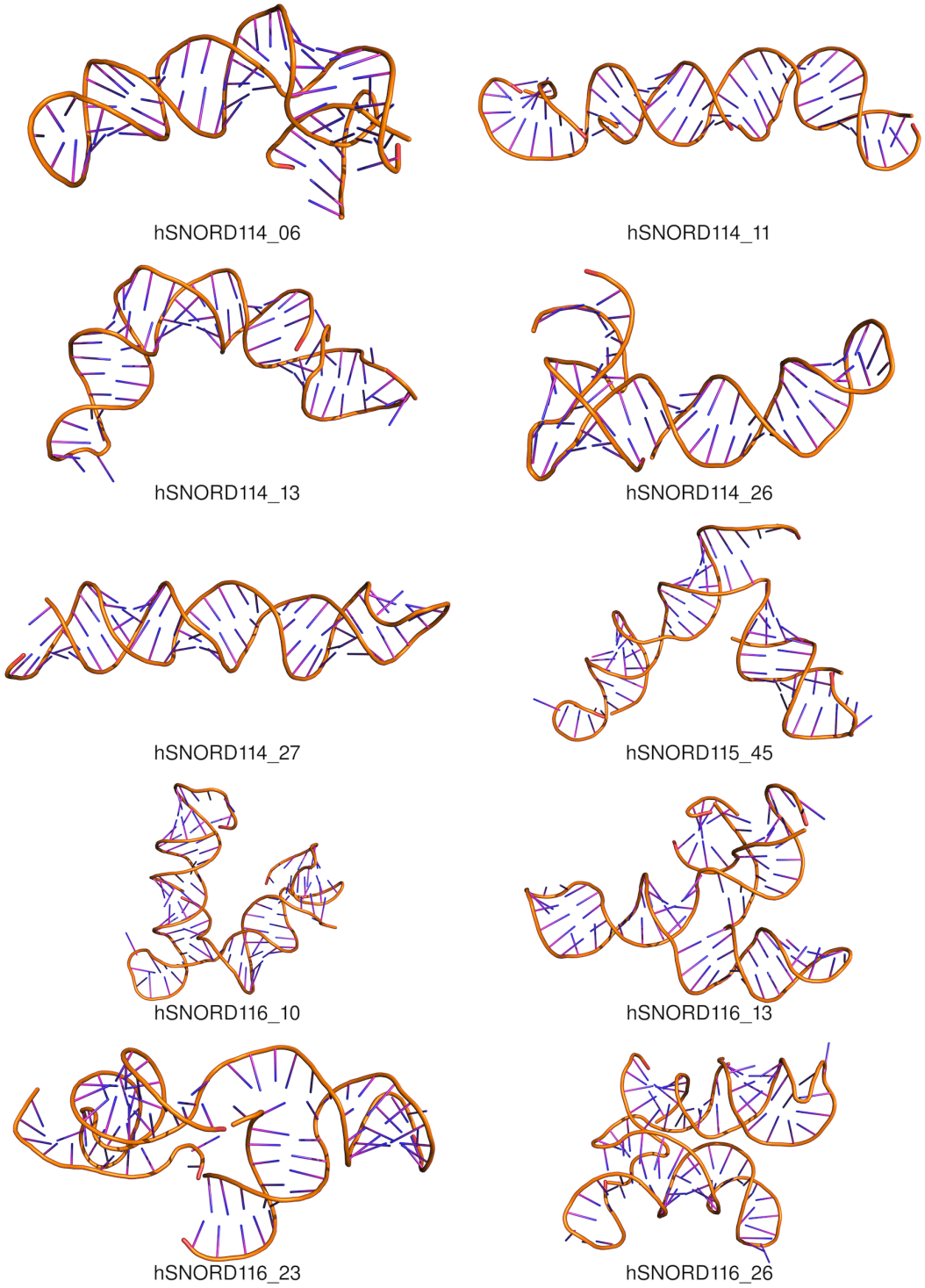


Fig. H.1 – 3D structures of human candidate snoRNAs

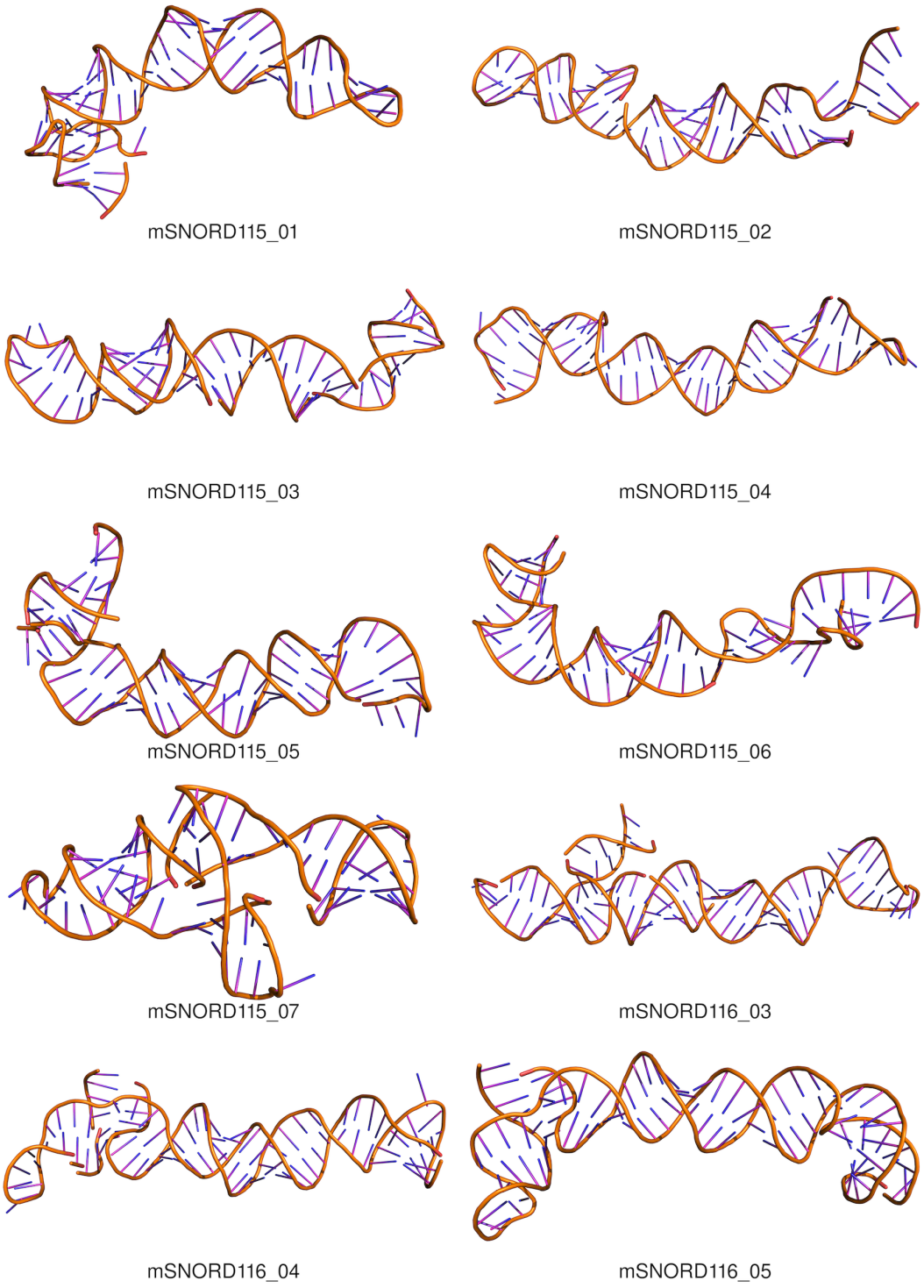


Fig. H.2 – 3D structures of murine candidate snoRNAs

H.2 Candidate snoRNAs aligned to U4 snRNA

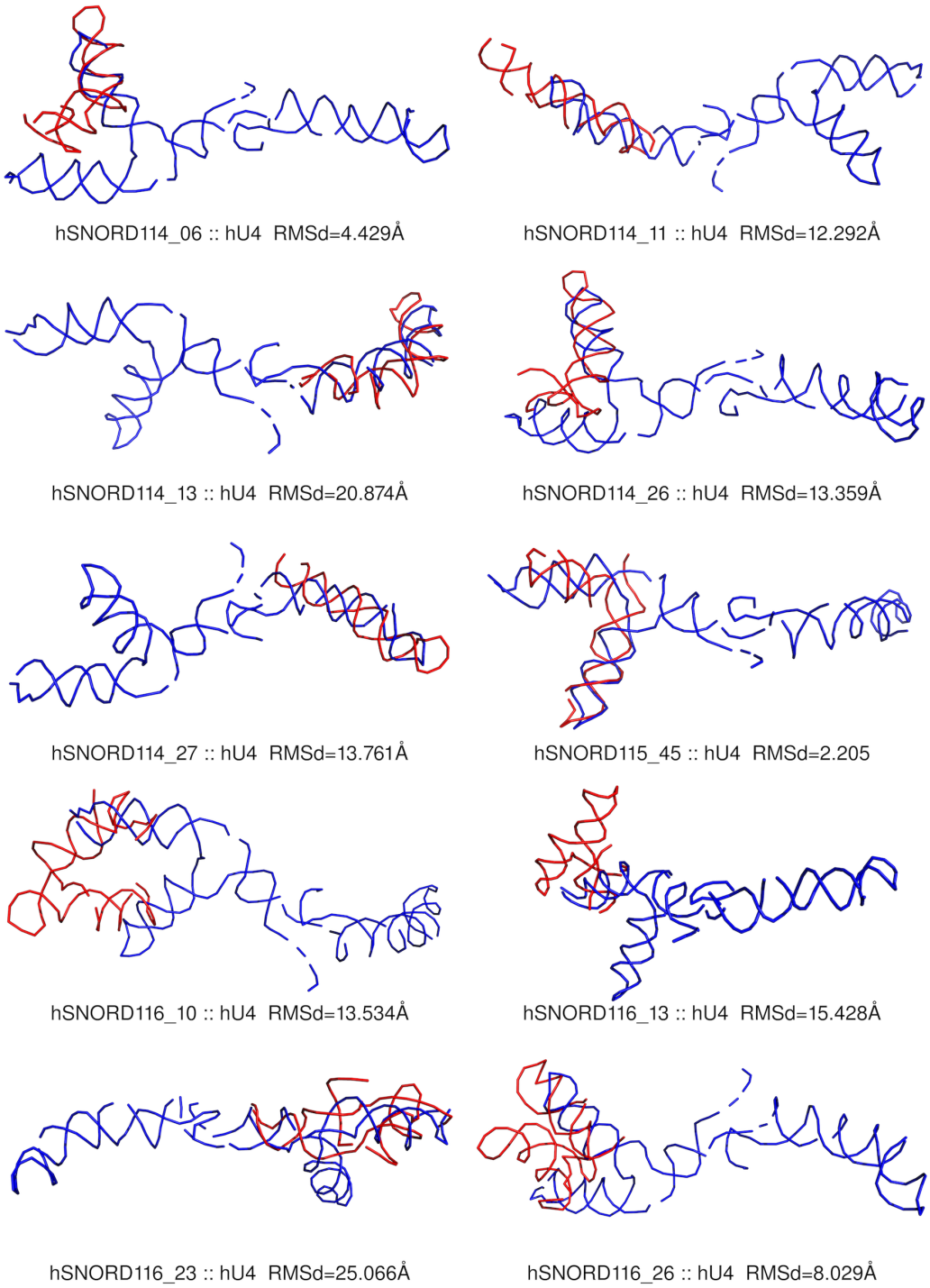


Fig. H.3 – 3D structures of human candidate snoRNAs aligned to the 3D structure of the human U4 spliceosomal snRNA. Alignment carried out using the co-ordinates in Table 3.2.

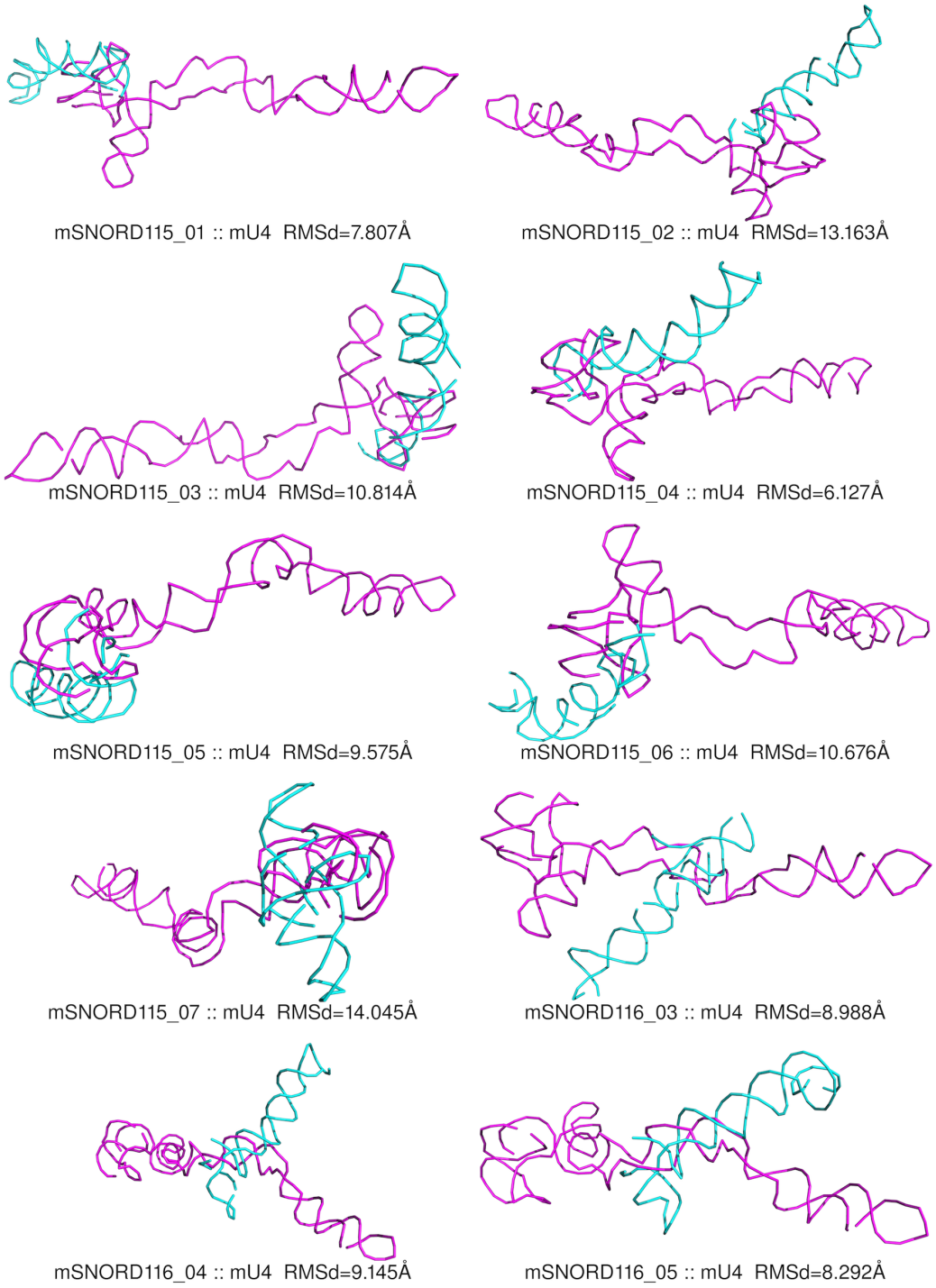


Fig. H.4 – 3D structures of murine candidate snoRNAs aligned to the 3D structure of the murine U4 spliceosomal snRNA. Alignment carried out using the co-ordinates in Table 3.2.