Identification and Validation of the U2, U4, U5, and U6 Spliceosomal snRNAs in *Cyanidioschyzon merolae*

by

William St. Clair Dunn

B.Sc., University of Northern British Columbia, 2008

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTERS OF SCIENCE
IN
MATHEMATICAL, COMPUTER, AND PHYSICAL SCIENCES
(CHEMISTRY)

UNIVERSITY OF NORTHERN BRITISH COLUMBIA

April 2011

Your file Votre référence
ISBN: 978-0-494-75173-2
Our file Notre référence
ISBN: 978-0-494-75173-2

Abstract

Pre-mRNA splicing is an essential step in eukaryotic gene expression, and introns have been found in nearly all eukaryotic genomes sequenced to date. The red alga *Cyanidioschyzon merolae* is found in acidic thermal springs, and its recently sequenced genome revealed a surprising paucity of intron-containing genes, raising the question of whether the normal complement of splicing machinery is maintained to splice so few introns. To address this I searched for snRNAs computationally, successfully identifying *C. merolae* homologues for four of the five snRNAs. I experimentally confirmed their expression, found that their structural elements are similar to those known from other organisms, and demonstrated that U4 and U6 base pair to each other, as expected. My data support the proposed switch in U6-5' splice site base pairing between the two catalytic steps, as well as a recent model for free U6.

TABLE OF CONTENTS

List of Tables

List of Figures

Acknowledgements and Dedication

First and foremost I want to thank my brilliant wife Liz Dunn. She has been there for the entirety of my bullish charge into the world of biochemistry, patiently steering me around certain disaster, and with a steady hand taught me the techniques I required. She is the embodiment of what the science of biochemistry should be; constantly questioning, unafraid of new ideas, and in a continuous state of fascination.

I would also like to thank my supervisor Dr. Stephen Rader for the opportunity to study in this strange new discipline, the rest of the Rader Lab for embracing me as one of their own, my thesis committee members: Dr. Andrea Gorrell for delicious barbecued peppers and wackiness, and Dr. Alex Aravind for the extensive use of his computational resources, and finally the Misumi Lab at Rikkyo University Tokyo for our original *C. merolae* culture.

This work is dedicated to our little adventurer Link St. Clair Dunn, who was born during the course of my graduate studies, and was jointly funded through an NSERC Discovery Grant awarded to Dr. Stephen Rader and various UNBC administered research project awards awarded to William Dunn.

Introduction

The act of removing non-coding regions from precursor messenger RNA (pre-mRNA) transcripts to form the mature messenger RNA (mRNA) is a critical step in eukaryote gene expression and a steadfast component of nearly all eukaryote genomes. The cellular machinery that catalyses this process, called the spliceosome, consists of a dynamic complex of five small nuclear RNAs (snRNA) and over one hundred associated proteins (Jurica & Moore 2003). While the splicing reaction has been well studied in model organisms such as *Saccharomyces cerevisiae* (*S. cerevisiae*) and *Homo sapiens* (*H. sapiens*), very little is known about the function and three dimensional structure of the five snRNAs. Moreover, few snRNAs have been properly confirmed biochemically. It is my hope that my research into the snRNAs of the unique hot springs species, *Cyanidioschyzon merolae* (*C. merolae*), will be the first step in a new direction to finally answer our questions about the structure, function, and mechanism of the snRNAs.

Nuclear pre-mRNA Splicing

Pre-mRNA splicing removes non-coding regions (introns) from between two coding regions (exons) of the pre-mRNA that is transcribed from the DNA template in order to create mRNA, which is then ready to be translated into the encoded protein. Pre-mRNA splicing occurs by two transesterification reactions that require the assistance of the spliceosome (Fig. 1). In the first reaction the 2' hydroxyl of a bulged adenosine within the intron (branch point) attacks the 5' phosphoryl group at the 5' intron-exon junction (5' splice site), and concurrently forms a lariat intron-exon intermediate (lariat loop) and the free 5' exon (Padgett et al. 1984, Konarsak et al. 1985). In the second reaction the 3' hydroxyl of the 5' exon reacts with the 3' intron-exon contact (3' splice site), cleaving away the lariat loop

and ligating the 5' and 3' exons through a 3'-5' phosphodiester linkage (Padgett et al. 1984, Konarsak et al. 1985). The resulting exon-exon ligation product, assuming no additional introns need to be spliced, is then ready to be translated into the protein for which the mRNA encodes.

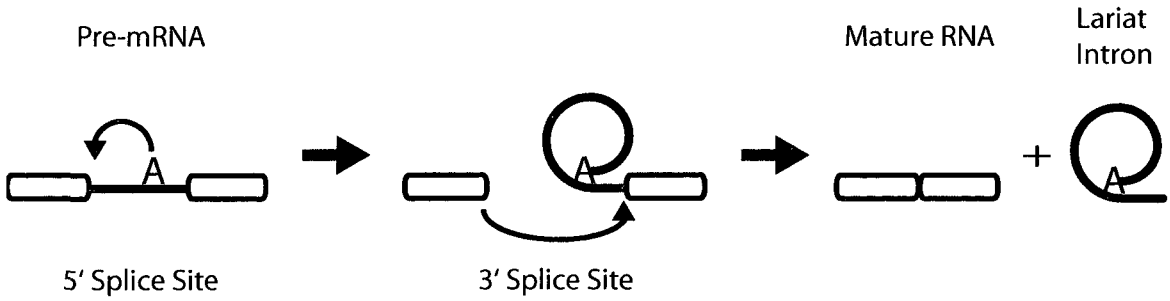

Figure 1: **Pre-mRNA Splicing**. Coding exons are shown as open boxes while the non-coding intron is indicated with a thick black line. The branch point adenosine is marked with an "A".

Spliceosome Assembly

The five major components of the spliceosome are thought to act on each new pre-mRNA transcript through the recognition and binding of three highly conserved sequences in the transcript: the 5' splice site, the branch point sequence, and the 3' splice site (Siliciano et al. 1987, Sawa & Abelson 1992, Lesser & Guthrie 1993, Parker et al. 1987, Umen & Guthrie 1995). It is still unknown whether the spliceosome assembles in a piecewise fashion on the transcript (Bindereif & Green 1987, Cheng & Abelson 1987, Konarska & Sharp 1987) or arrives as a pre-formed penta-snRNP (Stevens et al. 2002), but in either case, a number of intermolecular interactions between the snRNAs and the splicing transcript hold true. Many of these interactions occur through direct RNA/RNA base pairing between the transcript and the snRNAs; for example, both U1 and U6 snRNAs have been shown to base pair with the 5' splice site of the pre-mRNA transcript (Siliciano & Guthrie 1988, Seraphin et al. 1988), and

similarly, U2 snRNA has been demonstrated to base pair with the branch point (Parker et al. 1987).

Cyanidioschyzon merolae as a Model

Cyanidioschyzon merolae (*C. merolae*) is an acidophilic, unicellular red alga, whose genome was the first algal genome to be sequenced and the first 100% completed eukaryotic genome (Nozaki et al. 2007). At 16.5 million base pairs its genome is strikingly compact, the smallest of any photosynthetic organism (Matsuzaki et al. 2004), and indicative of the stripped-down metabolic machinery in this intriguing organism. Herein lies *C.* merolae's strength as a model organism. On the assumption that splicing takes place within *C. merolae*'s cells, as I suspect that it does based on computational identification of well conserved 5' and 3' splice sites and branch sequence (Matsuzaki et al. 2004), then ideally the rest of the cellular systems should be as simple as possible so as to introduce the minimum number of confounding factors into our experiments. *C. merolae*'s elegantly simplistic cells are well suited to study as they contain just a single nucleus, mitochondrion, and plastid, and do not even have rigid cell walls (Matsuzaki et al. 2004). Additionally, *C. merolae*'s division can be highly synchronized with light and dark cycles (Terui et al. 1995) which make it an excellent candidate for the study of the dividing apparatus of mitochondria and plastids (Kuroiwa 1998) and may offer a method of regulating splicing rates during experiments.

Most interesting of all is the fact that despite its small size, the *C. merolae* genome contains a comparable number of genes to the yeast *S. cerevisiae*, but only one tenth as many introns: 26 intron-containing genes (0.5% of the genome) (Matsuzaki et al. 2004) in *C. merolae* compared to 287 (~5%) (Juneau et al. 2007) in yeast. Its single plastid is small in size and contains no introns (Ohta et al. 2003). The small number of introns in *C. merolae* and the extreme environmental conditions that it exists in raises fascinating questions as to

whether or not the full complexity of the normal splicing machinery has been maintained, and if so, how these complexes continue to function under such extreme environmental pressures.

Project Motivation

Short Term

Ironically, with the advent of full genome sequencing satisfying our need for biological sequence data, we now have to contend with a much more difficult problem: too much data. Even the smallest and simplest genomes are well outside the capacity for manual screening and so we must turn to bioinformatics to assist us. Yet, the insights that bioinformatics provides are only as good as the wet-lab data on which the algorithm was trained.

It is important to remember that while bioinformatic techniques have already been used to identify thousands of potential snRNA homologues, very few of these snRNAs have been found and biochemically characterized in the wet-lab. With such a limited dataset it is likely that many of these potential snRNAs are not true homologues, but simply regions of DNA sequence that share some similar structural features, either because of now defunct pseudogenes or simply by chance.

As bioinformatics is the only tool we have for analyzing these enormous datasets, we must strive to improve it. The addition of new *biochemically confirmed* snRNAs to the training datasets is simply the most effective way to improve the accuracy and versatility of these algorithms in determining snRNA homologues in other species. *C. merolae*'s snRNAs would be of particular interest as *C. merolae* is highly divergent and may contain many features not yet present in the training dataset.

Long Term

The long term goal, of which this project is just a first step, arises from the questions as to how *C. merolae*'s snRNAs are able to splice under conditions that could begin to denature the snRNAs of other organisms. Under the assumption that spliceosomal splicing occurs within *C. merolae*, a necessary consequence is that it must have more robust snRNPs. This may be due to additional protein re-enforcement, more extensive base-pairing interactions, or some other factors within the snRNP itself that leads to a less flexible and more stable structure.

Attempts to crystallize any of the snRNPs in other model organisms have failed in all cases with the single exception of the *S. cerevisiae* U1 snRNP (Pomeranz Krummel et al. 2009), however *C.* merolae's robust and rigid snRNPs may be the solution to this problem. Crystallization requires the target molecules to align themselves in a regular repeating structural pattern. While this is relatively easy to facilitate with simple non-biological molecules, the difficulty increases dramatically when considering the complexity and flexibility of biological molecules. The *C. merolae* advantage lies in the fact that *C. merolae*'s more rigid snRNPs should more easily align themselves into the repeating patterns required for crystallization.

Chapter 1: Splicing within *Cyanidioschyzon merolae*

An important prerequisite of any study of splicing within *C. merolae* is the confirmation that splicing actually takes place *in vivo*. In the 2004 *C. merolae* genome sequencing project (Matsuzaki et al. 2004), 27 probable introns were found bioinformatically. These suspected introns each had a highly conserved 5' splice site, a branch point, and a 3' splice site, all of which are indicative of snRNA spliced introns.

To confirm the validity of a subset of these introns, while simultaneously testing for splicing within *C. merolae* cells, I used an RNA amplification technique called reverse transcription polymerase chain reaction (RT-PCR). I chose to amplify intron containing RNA regions of the expressed gene. The amplified regions were made to span not only the intron of interest, but also similar sized regions of exon on either side of the intron (see Fig. 2A). The amplified product and intron are both of known lengths, so when run on an ethidium bromide/agarose gel I expected to see a band corresponding to the full length pre-mRNA (Fig. 2B) and, if the intron of interest had been spliced out, a faster migrating band corresponding to the length of the two ligated exon regions or mRNA (Fig. 2C). It was also a possibility that in a case of extremely efficient splicing, no pre-mRNA band would be observed as it would all be found in the mRNA form. This chapter details my biochemical confirmation of splicing and the validation of a subset of putative introns within *C. merolae* cells.

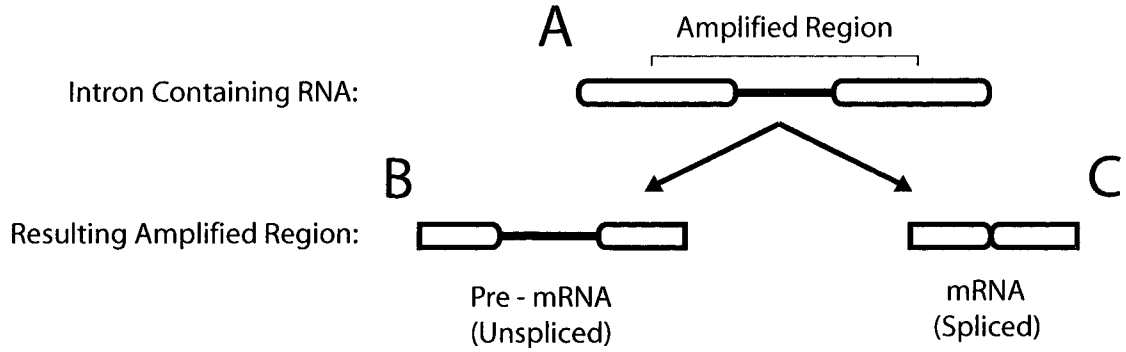

Figure 2: Schematic of the RNA of an expressed intron containing gene. Heavy black line represents regions of intron and rounded rectangle represents exon regions. **A** Full length unspliced intron containing RNA. **B** Unspliced pre-mRNA made up of two small sections of exon and the intron. **C** Spliced mRNA made up of just the ligated exon regions.

Materials and Methods

C. merolae Culturing

The 10D strain of *C. merolae* (NIES-1332), obtained from the Microbial Culture Collection at the National Institute for Environmental Studies in Tsukuba, Japan (http:// mcc.nies.go.jp/), was cultured in 50 mL of MA2 *C. merolae* media (Ohnuma et al. 2008) along with 200 μL each of trace element solution and Fe solution (Minoda et al. 2004). The cultures were grown under two 20 watt 60 Hz fluorescent aquarium lights (Marine-Glo) on a shaker at 45 C for 2-3 weeks.

Total RNA Preparation

The optical density (OD_{λ}) of the mature *C. merolae* culture was measured and the culture was divided into 15 mL conical tubes each with an $OD_{750} \approx 5$. These cultures were then spun down by spinning 6 minutes at 3000 g in a Beckman Coulter Allegra X-12R centrifuge with a SX4750 rotor, washed once with deionized water, and transferred to microcentrifuge tubes. The cells were spun down in an Eppendorf centrifuge. To prepare non-denatured total RNA, cell pellets were resuspended in 30 μL chilled RNA extraction buffer (50mM Tris-HCl pH 7.5, 100mM NaCl, 10mM EDTA). Two hundred μL 0.5mm Zirconia/

Silica beads (BioSpec Products Inc.) were added and tubes were vortexed for one minute on the maximum setting. Following a five minute incubation on ice, the tubes were vortexed for an additional minute before adding 300 µL chilled RNA extraction buffer, 60 µL 10% SDS, and 400 µL acid equilibrated phenol: chloroform (5:1, pH: 4) (Ambion). The samples were then vortexed for one minute on the highest setting and centrifuged in an Eppendorf Centrifuge 5415D at 13,200 rpm for five minutes at 4 C. The aqueous phase was transferred to a tube containing 500 µL cold acid equilibrated phenol: chloroform, and extracted as before. A third phenol: chloroform extraction was performed followed by an extraction with 500 µL chloroform (Sigma). The aqueous phase was transferred to a clean microcentrifuge tube and the RNA was precipitated with 40 µL 3M sodium acetate and 1 mL 100% cold ethanol. Samples were cooled at -80 C for at least 20 minutes. Precipitated RNA was pelleted by centrifugation in an Eppendorf Centrifuge 5415D at 13,200 rpm for 20 minutes at 4 C. The pellets were washed with 70% ethanol and allowed to air dry for 5-10 minutes prior to resuspension in 30 µL 10mM Tris-HCl, pH 7.5. Where appropriate, total RNA was denatured by heating for 3 minutes at 90 C.

***C. merolae* Splicing**

Two predicted intron-containing genes, CMS315C and CMS262C, were chosen to be screened for splicing through reverse transcription polymerase chain reaction (RT-PCR) analysis of their expressed RNA. Oligonucleotide pairs (Invitrogen) were designed upstream and downstream of the introns:

CMS315C:

oSDR734:CAGACAGGCCAACTGCTGGCTGGAA (17 nts upstream of 5' splice site)

oSDR735:GTGGTTTGTTCAGGCGCAAGTCGCA (114 nts downstream of 3' splice site)

CMS262C:

oSDR669:GGCGATATGGTCCTGGTTACG (106 nts upstream of 5' splice site)

oSDR668:GGCGATTGCTGAAGCCGCTGAGG (99 nts downstream of 3' splice site)

Heat-denatured total RNA was treated with 2 units of Turbo DNase (Ambion) and RT-PCR reactions were carried out using the appropriate primer pairs and AffinityScript reverse transcriptase (Stratagene). The 10μL reactions (2.0 μg *C. merolae* total RNA, 1 μL 10mM dNTPs, 0.5 μL 20 pmol / μL reverse primer) were incubated at 68C for 5 minutes and then moved onto ice for 1 minute. The AffinityScript solutions (1.5 μL 10x AffinityScript buffer, 1.0 μL 0.1M DTT, 1.0μL (20U) Superasin, 0.5 μL AffinityScript) were added to each reaction and then allowed to incubated for 1 hour at 45C. The reactions were then transferred to PCR tubes containing the PCR solution (5.0 μL 10x Standard Taq Buffer, 0.5 μL 10mM dNTP mix, 1.75 μL 20 pmol / μL reverse primer, 2.25 μL 20 pmol / μL forward primer, 2.0 μL 5U / μL Taq DNA Polymerase, 24.5μL dH_2O) and the PCR cycles were as follows: 95C for 2 minutes, 35 cycles of (95C for 1 minute, 57C for 1 minute, 72C for 1 minute 30 seconds), 72C for 10 minutes, and then hold at 4C.

The reaction products were run on a 1% agarose gel containing ethidium bromide. The gel was visualized on a Chemi Imager (Alpha Innotec) running AlphaEase FluorChem 5500. The resulting image was uniformly adjusted for contrast in a linear fashion.

Results and Discussion

Intron-containing Pre-mRNAs are Spliced in C. merolae

To confirm that some of the suspected introns within *C. merolae* transcripts were indeed spliced, I attempted to detect the presence of spliced (ie. intronless) transcripts in a preparation of total RNA using RT-PCR. I tested two genes, CMS315C and CMS262C,

predicted to have introns of 245 nts and 237 nts, respectively (Matsuzaki et al. 2004). I treated the samples with DNase prior to RT-PCR amplification to ensure that full-length products resulted from intron-containing transcripts and not DNA contamination. The presence of bands at both the predicted sizes for pre-mRNA and mRNA demonstrated that these *C. merolae* transcripts are spliced (Fig. 3A) while also validating the suspected introns, CMS315C and CMS262C, as true pre-mRNA spliced introns. I was unable to completely eliminate genomic DNA contamination in my reactions as demonstrated by a band in the control reaction lanes that lacked reverse transcriptase (Fig. 3B, lanes 2 and 5). These bands correspond to the PCR amplification of the associated region of genomic DNA from which our target RNA regions were transcribed. While ideally I would have no bands in the lanes lacking reverse transcriptase, these control lanes do provide some useful insight. In the lanes lacking reverse transcriptase the amplified genomic DNA region provides an effective size marker for our amplified pre-mRNA, and the lack of banding at the expected mRNA sizes indicates our RNA is not being amplified non-specifically. Additionally, in the corresponding RT-PCR reactions (Fig 3B, lanes 3 and 6), the bands on the level of the amplified genomic DNA region are intensified indicating that both our target pre-mRNA region and the genomic DNA region are being amplified in the presence of reverse transcriptase along with the mRNA product bands observable below. Having demonstrated that *C. merolae* does in fact splice, I sought to identify the five major components of the spliceosome: U1, U2, U4, U5, and U6 snRNAs (see Chapter 2).

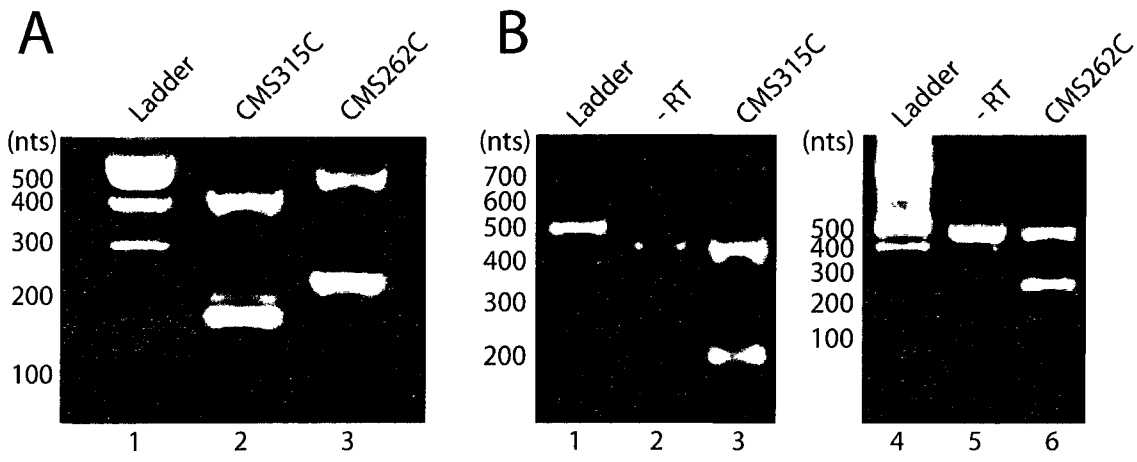

Figure 3: Intron containing genes in *C. merolae* are spliced. A. RT-PCR of two RNA regions, each spanning a *C. merolae* intron and a region of exon on either side, were amplified by RT-PCR. The products were run on a 1% agarose gel and visualized with ethidium bromide. **Lane 1:** 100 base-pair DNA ladder. **Lanes 2 and 3:** RT-PCR products of CMS315C and CMS262C genes respectively. The larger bands correspond to the expected unspliced amplicon size (CMS315C: 426 nts; CMS262C: 486 nts) and the smaller bands correspond to the expected spliced size (CMS315C: 181 nts; CMS262C: 249 nts). **B.** Control reactions for CMS315C and CMS262C. **Lanes 1 and 3:** 100 base-pair ladder. **Lanes 2 and 5:** RT-PCR reaction lacking reverse transcriptase. **Lane 3 and 6:** Standard RT-PCR reactions as in Part A.

C. merolae's Intron Evolution

Having shown that pre-mRNA splicing occurs in *C. merolae*, one stubborn question remained: would the extremely genomically minimalist organism *C. merolae* maintain all of the cellular machinery required for pre-mRNA splicing, including the five snRNAs and their hundreds of associated proteins, all for the sake of (potentially) as few as 27 introns? *C. merolae*'s inherent simplicity and paucity of introns lead me to initially suspect that its genome was that of a basal 'living fossil'. However recent studies of the evolution of spliceosomal introns proved my intuition wrong.

Intron gain is, with few exceptions, rare (< 0.0002 gain per gene per 10^6 years) while intron loss is much more common and variable (0 to 10% per 10^8 years) (Irimia & Roy

2008). Additionally, massive scale intron loss events have been shown to occur regularly in evolutionary history, while large scale intron gains are almost non-existent (Roy & Gilbert 2006). A consequence of this realization is the discounting of the long-held belief that introns appeared late in the evolution of modern organisms. The majority of their introns can be traced back to early eukaryote evolution and even the intron rich, complex modern organisms such as humans and mice have seen no intron gain in the last 75 million years (Waterson et al. 2002). The current model of intron evolution is one of variable speed intron loss stemming from extremely intron rich ancestors, wherein a reduced number of introns in a genome corresponds to both greater sequence change over time and longer phylogenetic branch length (Irimia & Roy, 2008).

Considering the near intronless *C. merolae*, this is indicative of huge amounts of sequence variability in its evolution. *C. merolae* is a hot environment acidophile, has a short generation time and a large population; given this and considering that *C. merolae*'s genome is tightly constrained by the additional selective pressures of its environment, it seems likely that *C. merolae* has lost nearly all of its original introns over the course of its evolution. Additionally I noted that *C. merolae*'s strong 5' splice site conservation was in keeping with the observations that wide-spread intron loss is commonly associated with strengthening of the 5' splice site consensus of remaining introns (Irimia et al. 2007).

In conclusion, I submit that *C. merolae* does in fact splice at least some of its 27 pre-mRNA introns, that it has done so since very early in its evolutionary history, and that through intensive genomic alteration and streamlining, presumably due to its harsh environmental conditions, nearly all of its ancestral introns have been eliminated.

Chapter 2: Bioinformatic Candidate Determination

No snRNA homologues had previously been found in *C. merolae* for any of the five snRNAs, as the small size of snRNA genes makes their identification by traditional sequence searches challenging. In order to search for the five snRNA homologues in *C. merolae* I turned to bioinformatic techniques trained on snRNA data from other organisms. At 16.5 million base-pairs, the *C. merolae* genome is considered to be quite concise, the smallest of all known photosynthetic eukaryotes (Matsuzaki et al. 2004), but is well out of the reach of effective manual screening. I chose the Infernal program as it was well suited to this study and offered superior sensitivity to that of more traditional homology searching methods.

The Infernal Advantage

The Infernal program (Nawrocki et al. 2009) is superior to many other homology searching methods in that it examines the sequencing data for not only primary structure homology, but also secondary structure homology. Infernal uses a training dataset of Stockholm aligned sequences, which contain a consensus secondary structure, to search sequence data for potential RNA homology. The Infernal program initially builds RNA secondary structure profiles called covariance models. These models allow for primary and secondary structure screening but are extremely computationally expensive and so Hidden Markov Models (HMMs) are used initially to 'prune' out highly unlikely sequences. The remaining sequences are searched with the covariance models. Sequences are scored by the Bit Score, a measure of whether the sequence is a better fit for the profile model (> 0) or the null model of non-homogeneous sequences (< 0). From the Bit Score a criterion for significance is calculated, the E-Value. The E-value gives the number of false positives

expected at or above this Bit Score. These two values allow for a quick and efficient screening of how well the candidate sequences match the constructed covariance model.

Between two highly divergent organisms, in this case the stripped down *C. merolae* and the splicing model organism *S. cerevisiae*, wherein the primary structure of a true homology may be quite divergent, I still expected a similar secondary structure as the functional nature of the secondary structure is more likely to be preserved. Infernal secondary structure sensitivity was extremely beneficial as the *C. merolae* genome contains an elevated G+C content (55%) relative to other organisms studied thus far (Matsuzaki et al. 2004), which I expected to result in less primary sequence conservation between the potential *C. merolae* candidates and the existing consensus. In the case of the *C. merolae* genome, the Infernal program provided a decisive advantage over more traditional primary sequence homology searching methods.

It should be noted that a previous computational study using the same toolset did not find any snRNA homologues in the *C. merolae* genome (Davila López et al. 2008). However, this study had screened an impressive 149 eukaryotic genomes, discovering potential snRNA homologues in every genome save for *C. merolae* and the deep-branching *G. lamblia* and had done so with a fairly strict criterion for homology while examining only sections of each genome. I was confident that with the luxury of focusing on just a single genome, I could search the entire genome while manipulating searching thresholds and sequence clustering in such a way as to increase search sensitivity and still keeping the number of potential candidates feasibly small for additional manual screening.

Materials and Methods

The 99.98% complete *C. merolae* genome was downloaded from the *C. merolae* genome project website (http://merolae.biol.s.u-tokyo.ac.jp/), and this formed my search area (Matsuzaki et al. 2004). I downloaded 'seed' training data sets for the five snRNAs U1, U2, U4, U5, and U6 from the Rfam database version 9.1 (Griffiths-Jones et al. 2005). The Infernal program version 1.0 was used with each seed dataset to search for the corresponding snRNA in the *C. merolae* genome on a Sun Microsystems unix machine running Solaris Express Community Edition (snr_105 SPARC) with 2 gigabytes of available RAM. The program was initially run using the default settings of a single covariance model based upon the entire seed dataset. I then instructed the program to divide the seed dataset into clusters of 60% or greater sequence identity and then re-run using multiple covariance models wherein each model was constructed from a single cluster to increase search sensitivity. In the case of the elusive U1 snRNA, the cluster threshold was further increased to 88%.

The set of sequences returned by Infernal for each snRNA was refined by only considering those sequences with an E-value of less than 0.5 and a Bit Score of greater than 15. The set of possible candidate sequences was further reduced by excluding sequences that were at odds with regions of high or invariant conservation among well characterized snRNAs. The candidate *C. merolae* snRNAs were chosen from their remaining respective sequence sets through individual examination of each candidate's ability to form snRNA secondary structures, as well as for their ability to form the extensive intermolecular base pairing interactions known to exist between snRNAs. Candidate sequences with a strong possibility of homology were found for four of the five snRNAs in *C. merolae*.

Sequence Alignments were prepared using the ClustalX program version 2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html) (Thompson et al. 1997, Larkin et al. 2007). In the

case of U4 and U5, where I was able to identify the Sm-binding site by manual inspection, I first aligned the sequences at the Sm region with low gap opening and extension penalties and then aligned the remaining regions with the default settings. The U2 and U6 sequence sets were aligned using the default settings.

Results and Discussion

Identification and Characterization of the candidate C. merolae snRNAs

To investigate the *C. merolae* splicing machinery I looked for snRNA sequences within the *C. merolae* genome that could potentially be snRNA homologues. The top *C. merolae* snRNA candidates for the U2 and U4 snRNAs were found using an Infernal single covariance model search while the U5 and U6 snRNA were found using multiple covariance models (see methods).

The increased clustering of multiple covariance model implemented for the U1 snRNA seed *did* increase the search sensitivity and provided twenty-five U1 snRNA candidates. Unfortunately none of these candidates possessed both a 5' splice site binding region and an Sm binding site, and none of the top five strongest candidates seem to be expressed. I then sought to further refine the *C. merolae* U2, U4, U5, and U6 snRNA candidates bioinformatically before moving on to biochemical validation (see Chapter 3).

While I was confident about the genomic location of each snRNA candidate within the *C. merolae* genome, I was less confident about the precise 5' and 3' boundaries of each candidate. Taking the U2 snRNA sequence as an example, the U2 snRNA in most organisms is approximately 160 nts, however in *S. cerevisiae* U2 is 1,175 nts but shares a strong complimentarity with the U2 snRNA of other organisms through its 5' region. In essence, the homologous functional core is shared through all U2 snRNAs but there is some variability in the number of nucleotides preceding (5') and following (3') this core region. While I was

confident in the candidates 'core' homology, I was well aware that the variable 5' and 3' sequence length of the Rfam seed dataset could cause Infernal reported boundaries to be close but not exact.

In order to refine each of the candidate snRNAs' 5' and 3' ends, I aligned each candidate with five biochemically confirmed snRNAs of the same type from different organisms. Noting where the core homology commenced and terminated in the other organisms I was able to bound each of the candidate *C. merolae* snRNAs and establish an overall length. The location of the candidate snRNA genes, and their Infernal scores, are shown in Table 1, while the sequences are shown in Figure 5.

Table 1: *C. merolae* candidate snRNA characteristics.

C. merolae snRNA	Chromosome	snRNA Accession Number	Strand	Range (nts)	GC Content (%)	Infernal Bit Score	Infernal E-Value
U2	11	AP006493	Plus	762863 - 762997	39	19.59	0.01083
U4	5	AP006487	Plus	222390 - 222571	50	21.29	0.02654
U5	17	BK008013	Plus	771503 - 771672	59	27.48	0.01949
U6	19	AP006501	Minus	483364 - 483492	52	17.8	0.1409

Mapping the snRNA Candidate's 5' and 3' Ends

***C. merolae* U2 snRNA** shares strong 5' end conservation with other well characterized U2 snRNAs, having 45% identity in the first 67 nucleotides (Fig. 5). This sequence conservation, along with conservation of secondary structure elements (see below), allowed us to anchor our tentative 5' end in the *C. merolae* genome. Sequence conservation drops off dramatically through the central region and 3' end, an observation that was not

unexpected as there is little 3' consensus among the biochemically characterized U2 snRNAs (Fig. 5). Consequently the 3' end of the *C. merolae* U2 snRNA was determined principally from the size estimate provided by our Northern blot (See Chapter 3). With an overall length of only 135 nucleotides, *C. merolae* U2 snRNA is by far the smallest characterized U2 snRNA.

***C. merolae* U4 snRNA** is similar to other well characterized U4 snRNAs throughout its length, with 72 of 182 (40%) nucleotide sequence identity between the *S. cerevisiae* and *C. merolae* sequences. Both the 5' and 3' ends of *C. merolae* U4 were easily mapped as a result of the high sequence conservation at the 5' end and through the Sm binding site at the 3' end, giving a total length of 182 nucleotides. This is one of the longest characterized U4 snRNAs, with a 31 nt insertion that can form a stem loop (nts 77-107) that is not present in other well characterized U4 snRNAs (Fig. 5).

***C. merolae* U5 snRNA** shares a strongly conserved core region of 20% sequence identity in the midsection, which is centered around a continuous 9 nucleotide sequence called Loop 1 that is completely conserved across all five species. This sequence is an important Prp8 binding site and is thought to align the 5' and 3' exons for ligation in the second step of splicing (Kershaw et al. 2009). As with the other U5 snRNAs, the *C. merolae* U5 shares little other sequence conservation save for a loosely conserved uridine rich Sm binding region near the 3' end (position 157 - 170 in Figure 5). As there was little 5' and 3' end conservation of sequence or length, the length of the *C. merolae* U5 snRNA was taken from the Infernal determined length of 171nts and this result was later confirmed through Northern blots (See chapter 3). While the *C. merolae* U5 snRNA shares little primary sequence conservation, its secondary structure is highly conserved among the U5 snRNAs.

The *C. merolae* U5 snRNA shares all of the secondary structure features established in yeast (Kershaw et al. 2009) (see Figure 4). These features include Loop 1 and Internal Loop 1 which are very nearly identical, in both primary and secondary structure, to the corresponding structures in *S. cerevisiae* and are well conserved across all five U5 snRNAs. Also present and structurally similar are the Variable Stem Loop and the 3' stem loop. Notably, Internal Loop 2, which is a stem bulge in *S. cerevisiae*, has the ability to be completely base paired in *C. merolae* and may offer additional structural stabilization to the *C. merolae* U5.

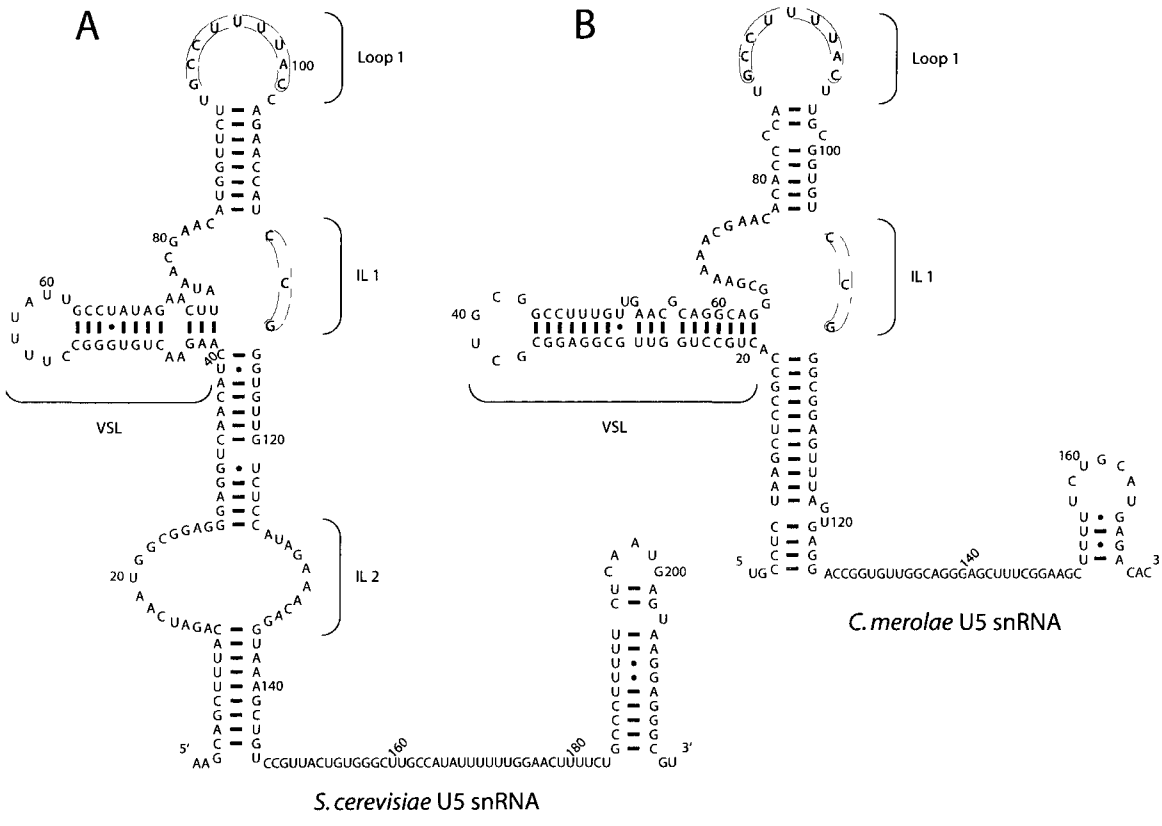

Figure 4: Free form *S. cerevisiae* and *C. merolae U5* snRNAs. The Loop 1 structure (Loop 1), internal loops (IL) and the variable stem loops (VSL) are indicated. The dark grey regions denote sequence conservation across all five U5 snRNAs aligned in Figure 5. Light grey regions denote the region of internal loop 1 that is conserved across both *S. cerevisiae* and *C. merolae*. The *S. cerevisiae* shown is the long form (214 nts).

***C. merolae* U6 snRNA** shares sequence elements with other biochemically characterized U6 snRNAs, with more than 60% sequence identity through the middle third of the molecule (Fig. 5). The 3' end of U6 was easily mapped due to the presence of the highly conserved, uridine-rich Lsm-binding site. To map the 5' end, mfold (Zuker 2003) was used, in conjunction with our size estimate from Northern blotting (See Chapter 3), to examine our tentative 5' end for its ability to form the phylogenetically conserved 5' stem loop (Fig. 7). The *C. merolae* stem loop is large compared to other characterized metazoan 5' stem loops, however the melting temperature is very similar to that of the *S. cerevisiae* 5' stem loop, with estimated melting temperatures of 96.3 C and 94.5 C respectively (Owczarzy et al. 2008).

Table 2: Accession numbers for sequence alignment snRNAs.

Species	**Accession Numbers**			
	U2	**U4**	**U5**	**U6**
C. reinhardtii	X71483	X71485	X67000	X71486
H. sapiens	M19204	M15956	K03167	M14486
S. pombe	X55772	X15491	X15310	X14196
S. cerevisiae	M14625	M17238	NC_001139 :939675 - 939497	X12565
C. merolae	AP006493	AP006487	AP006499: 771503 - 771673	AP006501

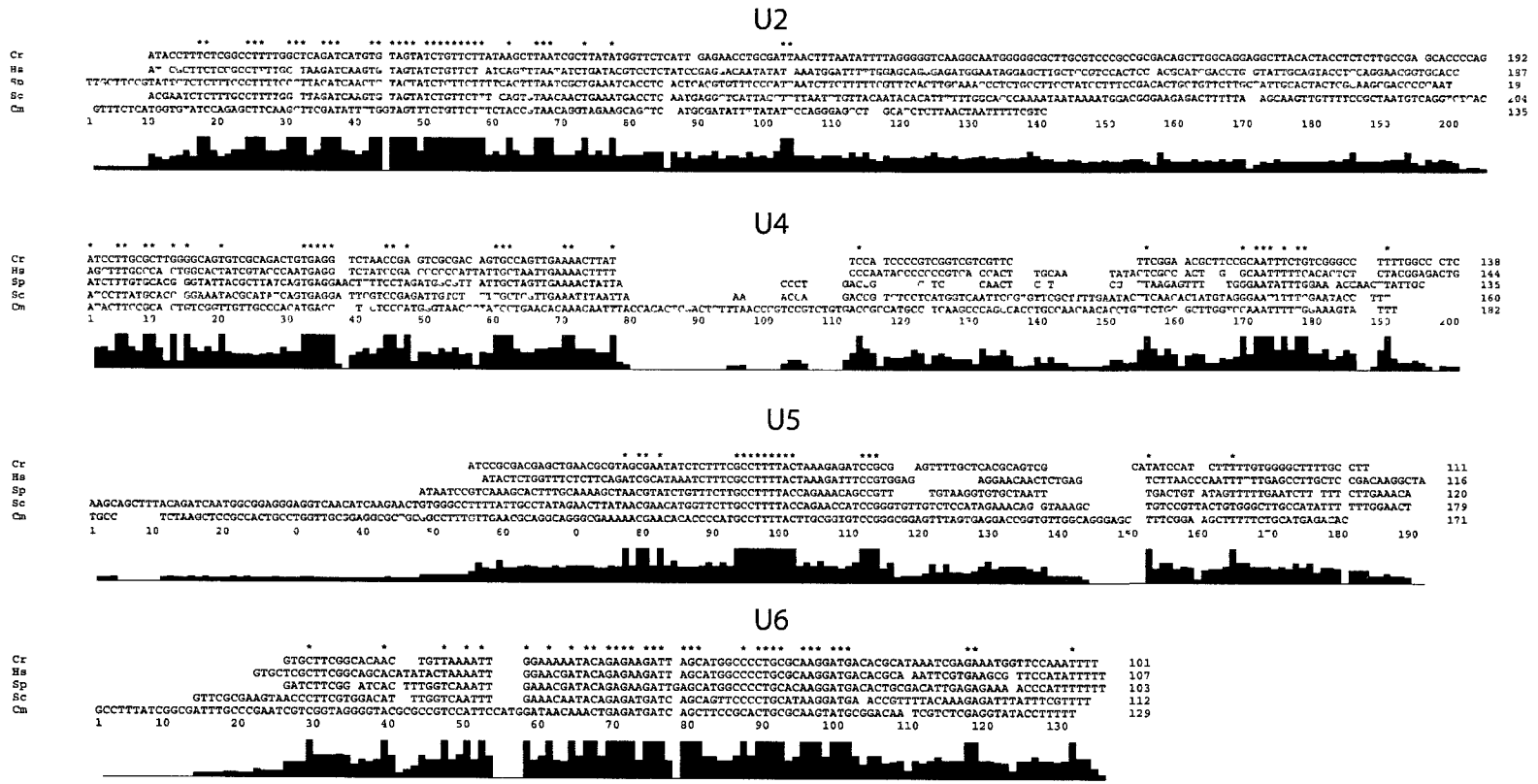

Figure 5: Sequence Alignments of known snRNAs with *C. merolae* candidates.
C merolae snRNA candidates aligned against biochemically verified snRNAs from *C reinhardtii* (*Cr*), *Homo sapiens* (*Hs*), *S pombe* (*Sp*), *and S cerevisiae* (*Sc*). The *S cerevisiae* U2 snRNA has been truncated at nucleotide 204 of 1175 for clarity The *S cerevisiae U5 snRNA* shown is the shorter version and the *H sapiens* U5 snRNA is HUMU5B1

Chapter 3: Candidate snRNA Experimental Validation

Having isolated and refined my top *C. merolae* snRNA candidates I was ready to begin biochemical verification. In order to validate my bioinformatically determined snRNA candidates, I needed to demonstrate the following within *C. merolae*: RNA in the Infernal predicted genes is being transcribed, the length of the transcribed RNA agrees with bioinformatically predicted lengths, and the transcribed RNAs behave as true snRNAs forming the intermolecular interactions that have been shown to be required for splicing reactions.

This chapter details the experiments used to confirm that snRNA candidates were being transcribed and to confirm our bioinformatic estimations of overall RNA length using denatured Northern blotting, as it provides information about transcription and overall RNA length, and the use of non-denatured solution hybridization analysis (Li & Brow 1993) which allows for the examination of native properties of the RNAs and facilitates comparisons to established snRNA properties.

Materials and Methods

Denaturing Northern Analysis

To determine candidate expression and the overall length of each snRNA, Denatured Northern blotting reactions were performed. Denatured *C. merolae* and *S. cerevisiae* total RNA was run on an 8% polyacrylamide 7M urea gel at 400 volts for 1.5 hours, then transferred onto a Hybond N+ membrane (GE Healthcare) using a Panther Semidry Electroblotter HEP-3 (Owl) for 15 min at 450 mAmps. RNAs were crosslinked to the membrane in a UV Stratalinker 1800 (Stratagene) with 120000 joules of ultraviolet radiation before being probed for the *C. merolae* snRNA of interest as well as the *S. cerevisiae*

snRNAs U4, U5, and U6 for size comparison. Oligonucleotide probes were kinased using gamma ^{32}P ATP 3000 Ci/mmol (PerkinElmer) and T4 polynucleotide kinase (NEB) according to the manufacturer's instructions. Blots were prehybridized in 10 mL of Rapid-Hyb Buffer (GE Healthcare) at 65 C for 30 min, then allowed to cool to room temperature over 30 min. Probe was added and hybridized at room temperature, followed by three 3 minute washes in Northern blot wash solution (0.2% Sodium dodecyl sulfate, 900mM NaCl and 90mM sodium citrate). Probed blots were imaged on a phosphor imager screen overnight and visualized with a Cyclone Phosphor Imager and OptiQuant software© (Packard Instruments). The blot was then stripped by bringing the Northern strip (0.1% Sodium dodecyl sulfate, 15mM NaCl and 1.5mM sodium citrate, 40mM Tris-HCl pH: 7.5) to a boil, suspending the blot in 10 mL for 10 min. This process was then repeated three times and the blot was imaged overnight as above. The blot was then re-probed for each of the five snRNAs. Each of the resulting images was uniformly adjusted for contrast.

The following oligos were used to probe the blot for *C. merolae* U2, U4, U5, and U6 snRNAs:

U2: CAGAAACTACCAAAATATCGAAGCTTGAAGCTC (oSDR745)

U4: AAATTGTTTGTGTTCAGCATACCGTT (oSDR597)

U5: GGACACCGCAAGTAAAAGGCATGG (oSDR768)

U6: AAAAAGGTATACCTCGAGACGATTGTC (oSDR598)

The following oligos were used to probe the blot for *S. cerevisiae* U4, U5S, U5L and U6:

U4: AGGTATTCCAAAAATTCCCTAC (14b)

U5S and U5L: AAGTTCCAAAAAATATGGCAAGC (U5-75mWTNR)

U6: TTGTTTCAAATTGACC (oSDR467)

Solution Hybridization analysis

Non-denaturing solution hybridization analysis was performed as described previously (Li & Brow 1993). ^{32}P-labeled oligonucleotides for *C. merolae* (U4: oSDR597, U6: oSDR598) and *S. cerevisiae* (U4: 14b, U6: oSDR467) were added to non-denatured *C. merolae* and *S. cerevisiae* total RNA and the reaction was split in half. One half was incubated under non-denaturing conditions (on ice for 20 min), while the other was incubated under denaturing conditions (75 C for 5 min and then 15 min at RT). The RNA-oligo complexes were run on a 9% non-denaturing polyacrylamide gel at 250 volts at 4 C for 1 hour and 45 minutes, imaged on a phosphor screen overnight at -80 C, and the resulting autoradiogram was visualized with a Cyclone Phosphor Imager and OptiQuant software© (Packard Instruments). The resulting image was then uniformly adjusted for contrast.

Results and Discussion

When probed with ^{32}P end labeled DNA oligonucleotides complementary to each of the five snRNAs, a single band was observed in each case, confirming that the genes identified by Infernal were in fact transcribed into the corresponding RNAs. In addition to confirming RNA expression, the Northern blot was used to obtain a size estimate for each RNA. These size estimates were in close agreement with the bioinformatic candidate lengths determined by primary sequence comparison to other well characterized snRNAs (Fig. 5). The lengths of the snRNAs were found to be 135, 182, 171, and 129 nucleotides for U2, U4, U5, and U6 snRNA respectively.

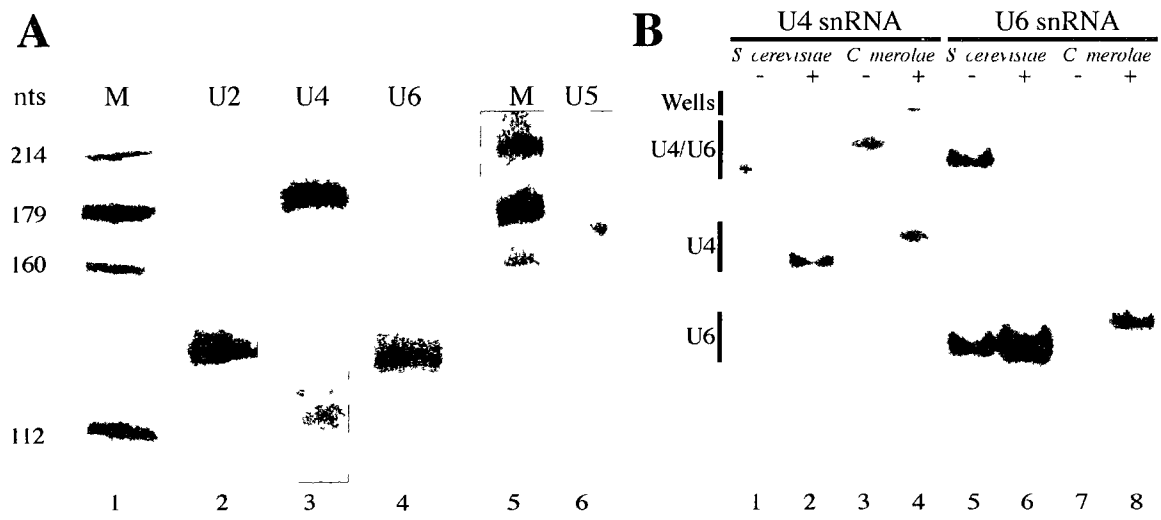

Figure 6: Expression of snRNAs in *C. merolae*. **A** Two denaturing northern blots reveals the expression of U2 (lane 2), U4 (lane 3), U5 (lane 6), and U6 (lane 4). *S. cerevisiae* U4, U5-S, U5-L, and U6 are used as size markers (lanes 1 and 5). Sizes, in nucleotides, are indicated. **B** A non-denaturing solution hybridization gel demonstrates that *C. merolae* U4 and U6 are base paired. Cold phenol extracted total RNA from *S. cerevisiae* or *C. merolae* was run on a non-denaturing acrylamide gel with ^{32}P-labeled oligos complementary to U4 or U6, as indicated at the top. The -/+ above the lanes indicates whether the RNA was heat-denatured prior to loading. Positions of free U4 and U6 and base-paired U4/U6 are shown on the left.

The most discriminating test of the U4 and U6 snRNA candidates was to examine if they would form the extensively base paired U4/U6 complex known to be essential for splicing in other species. To do this, I performed a solution hybridization experiment in which a ^{32}P-labeled probe against U4 (Fig. 6B, lanes 1-4) or U6 (lanes 5-8) was incubated with total RNA and separated on a non-denaturing gel. The *C. merolae* U4 (lane 3) and U6 (lane 7) co-migrate, indicating that they are in a base paired complex that dissociates upon heat treatment (lanes 4 and 8). *S. cerevisiae* RNA was used as a control to show heating-induced dissociation of U4 (compare lanes 1 and 2) and U6 (lanes 5 and 6), as well as to provide size markers. Essentially all of the U4 snRNA in *C. merolae* was found base paired to U6 snRNA (96%), leaving very little free U4 (lane 3); this is consistent with observations in *S. cerevisiae* (95%). In contrast, 28% of the total U6 snRNA present was found in a free

species, with the remainder base paired to U4 snRNA (lane 7) whereas I observed 57% free in *S. cerevisiae*. When compared with the corresponding heat denatured lane, no less than 95% of all species were found in the free form. The similarity of U4 and U6 snRNA properties between *S. cerevisiae* and *C. merolae* confirms the identification of our *C. merolae* candidates as snRNAs.

I have demonstrated in this chapter that all of the bioinformatically determined snRNA candidates are expressed and that the U4 and U6 candidates mirror the unusual properties of other known snRNAs. I submit that I demonstrated that the U4 and U6 candidates are in fact snRNA homologues and that the remaining candidates are extremely likely to be snRNA homologues, a claim I further support through a more detailed analysis of the snRNAs in Chapter 4.

Chapter 4: The *C. merolae* snRNAs

Conservation of Base Pairing Interactions Between Spliceosomal RNAs

It is known from other organisms that genetically determined interactions within the spliceosome - between U2 and U6, U2 and the intron, and U6 and the intron - are not stable enough to detect electrophoretically, with the exception of U4/U6. I therefore modeled these interactions manually, along with those between U4 and U6, to determine whether they are similar to what has been observed in other organisms. Consider my secondary structure model for the *C. merolae* U4/U6 complex (Fig. 7), with the *S. cerevisiae* complex inset for comparison. The best-characterized interactions, in stems I and II, are highly conserved in *C. merolae*, as is the phylogenetically conserved stem III (Brow & Vidaver 1995, Jakab et al. 1997). Although there is no experimental evidence for U4/U6 stem III, it remains possible that it exists transiently during some stage in the splicing cycle that has so far eluded detection.

Potential base pairing interactions between U2, U6, and the intron are similarly conserved in *C. merolae* (Fig. 8). These interactions are different depending on whether they occur in the four helix junction form, thought to correspond to the first chemical step of splicing (Fig. 8A) (Sashital et al. 2004) or the three helix junction form, corresponding to the second step of splicing (Fig. 8B) (Hilliker & Staley 2004). In figure 8A, interactions between U2 and the branch point, U6 and the 5' splice site, and U2 and U6 are similar to those modeled in *S. cerevisiae* (inset). Similarly, in figure 8B, the potential interactions, now including U2/U6 helices Ia, Ib, and II, comprise the same regions in *C. merolae* as in *S. cerevisiae*. In summary, the potential secondary structures and base pairing interactions in *C. merolae* are similar to those known in other organisms.

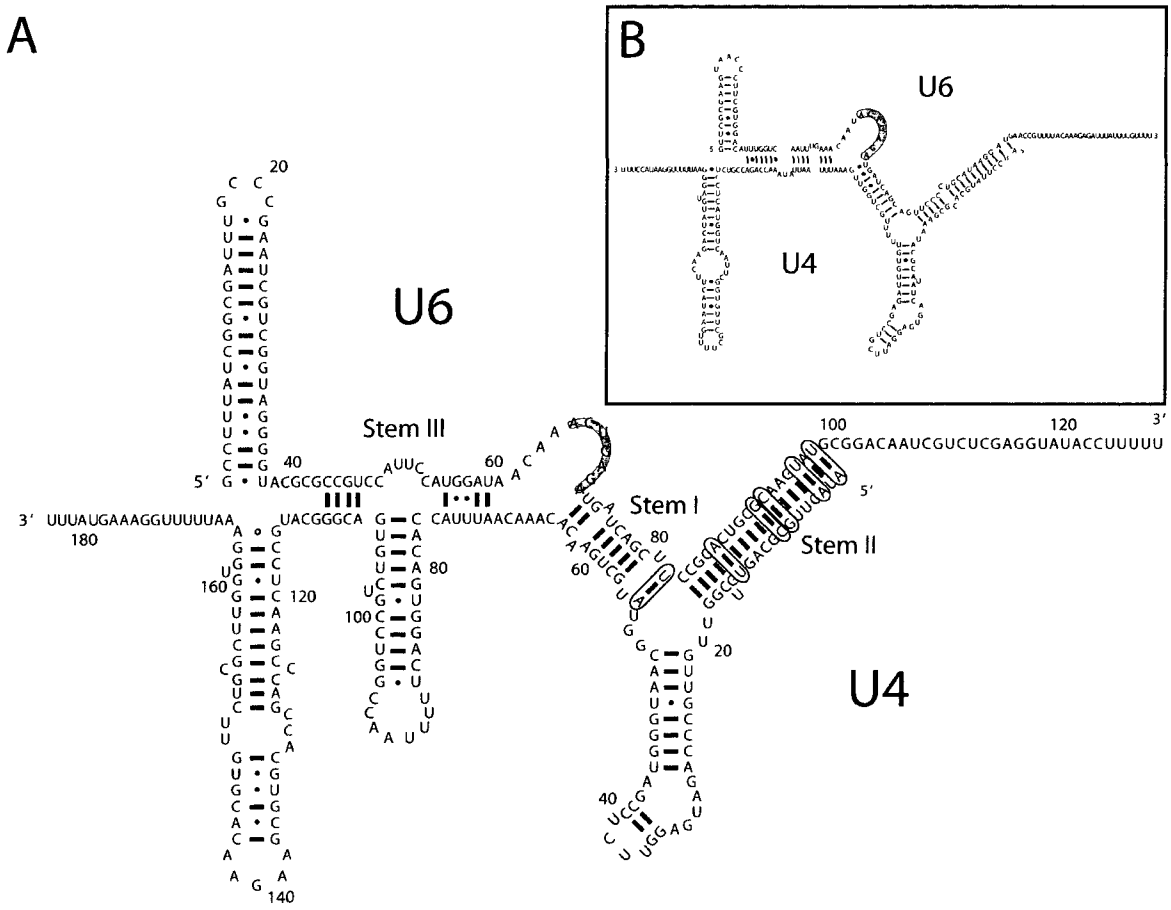

Figure 7: Predicted secondary structure of *C. merolae* snRNA candidates U4 and U6 in their base paired form. A. Model of *C. merolae* U4/U6 interactions including stem I, stem II, the phylogenetically conserved stem III, and a central insertion unique to *C. merolae* U4. The characteristic U6 sequence AC(A/U)GAGA is highlighted. Base pairs that co-vary between *S. cerevisiae* and *C. merolae* are circled in the *C. merolae* structure. B. Model of the *S. cerevisiae* U4/U6 complex for comparison.

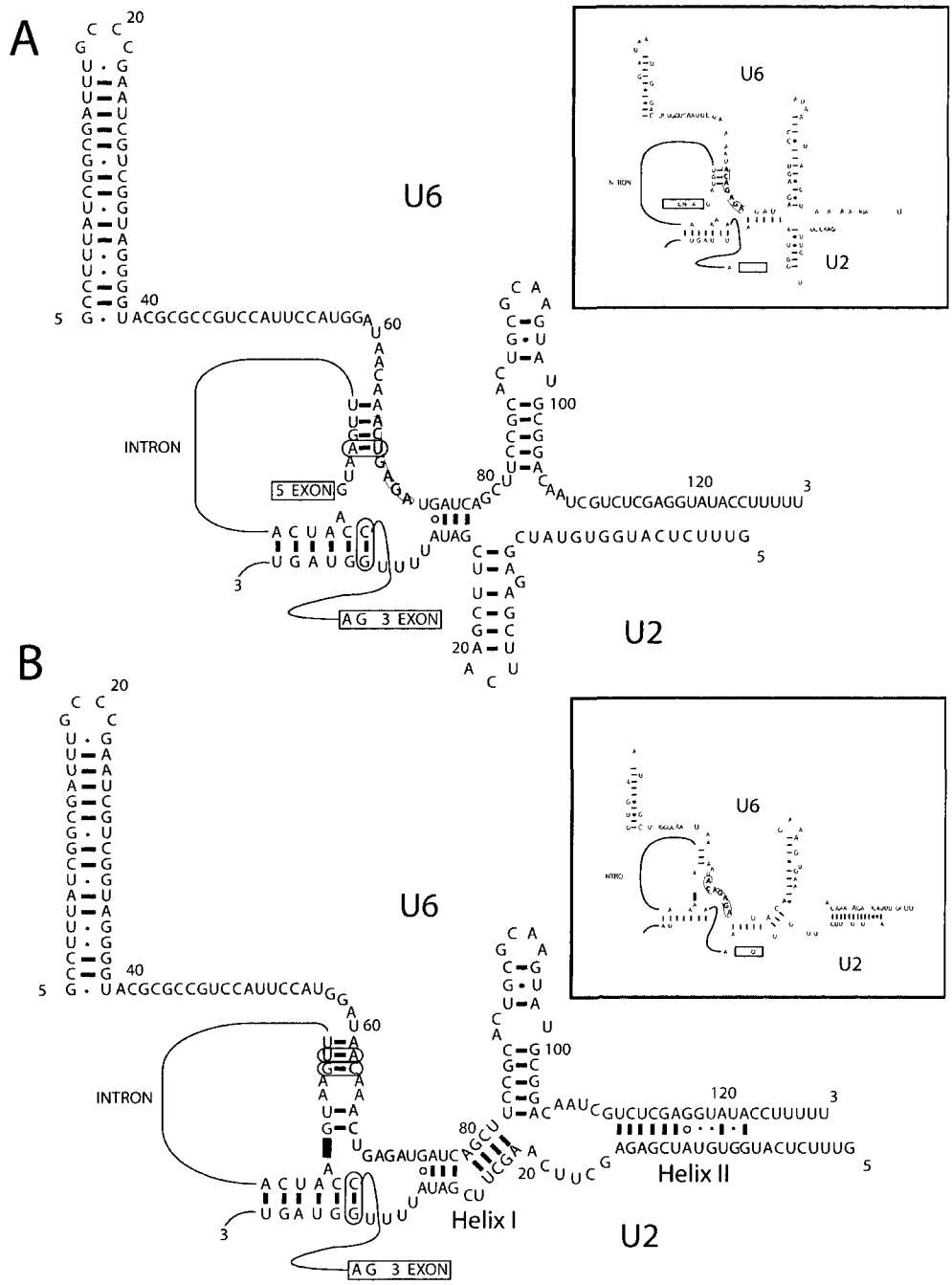

Figure 8: Secondary structure model of U2 and U6 interactions. A Four helix junction (step 1) model of *C. merolae* U2, U6, and intron interactions. The U6 AC(A/U)GAGA sequence is highlighted, and covariant base pairs between U6 and the intron 5' splice site and U2 and the branch point are circled. The *S. cerevisiae* complex is shown for reference (inset). **B** Three helix junction (step 2) model of *C. merolae* U2, U6, and intron interactions. U6 sequences and covariant base pairs are indicated as in A. The heavy black line denotes the covalent bond formed between the 5' end of the intron and the branch point of the transcript. The *S. cerevisiae* complex is shown for reference (inset).

Phylogenetic Co-variation in Spliceosomal Secondary Structures

Analysis of phylogenetic co-variation in RNA is a powerful tool for RNA structure prediction; the existence of a proposed base pair is strongly supported by variations in sequence in which the identity of the paired bases changes while the ability to form a base pair is maintained (Noller et al. 1981, Woese et al. 1983). This strategy has been used extensively to predict RNA secondary and tertiary structures ranging from small RNAs to large macromolecular complexes such as the ribosomal subunits (Noller et al. 1981, Woese et al. 1983). In fact, the 16S and 23S ribosomal RNA secondary structures predicted by comparative sequence analysis were later shown to be more than 97% accurate by X-ray crystallography (Wimberly et al. 2000, Ban et al. 2000, Gutell et al. 2002).

The *C. merolae* snRNAs described here were identified in part through the conservation of known secondary structure elements, such as the 5' stem loop in U6. I have observed 10 examples of intermolecular co-variation, one in U4/U6 stem I (Fig. 7), five in U4/U6 stem II (Fig. 7), one in the U2/branch point interaction (Fig. 8), and three between U6 and the 5' splice site (Fig. 8). The most notable of these is the interaction between U6 and the 5' splice site, which involves a mutation in the phylogenetically nearly invariant U6 AC**A**GAGA sequence, which in *C. merolae* has mutated to AC**U**GAGA. Compensation for this mutation occurs by a co-variation in the *C. merolae* 5' splice site consensus sequence, which changes from GUA**U**GU in *S. cerevisiae* to GUA**A**GU in *C. merolae* (Fig. 8A).

U6/Intron Co-variation Supports the 5' Splice Site Interaction

Given the high conservation of U6, the presence of a U at position 68 in the ACAGAGA sequence raised questions about whether this was the correct U6 gene. The only proposed intermolecular base pairing partner for this position is in the 5' splice site. Strikingly, 24 out of 27 introns in *C. merolae* have GUA**A** (bold nucleotide is complementary

to U6 U68) in their 5' splice site (compared to GUAU in *S. cerevisiae*), while two of the remaining three introns have GUA**G**, which would also be able to base pair with U6 (Fig. 8). This provides the first support for a U6/5' splice site base pair based on co-variation.

It has recently been shown that not only is the GUAAGU sequence of the 5' splice site a common feature of intron-poor genomes, but that intron-poor species show a strict adherence to this consensus (Irimia et al. 2007). Notably, the predominant 5' splice site sequence of various microsporidia species, which also have an ACUGAGA sequence in U6 snRNA, is GUAA as well (Fast et al. 1998). Coupled with our results presented here, this raises the possibility that the U6/5' splice site co-variation might be a common feature in organisms that possess stripped down splicing machinery.

This phylogenetic support for the U6/5' splice site interaction adds to a growing body of data that suggests that the interaction takes place early in spliceosome assembly, prior to the first catalytic step. The strongest evidence for this was reported in the 4-thio-uridine cross-links observed between the *S. cerevisiae* ACAGAGA sequence and the 5' splice site of full-length pre-mRNA, which formed in spliceosomes stalled prior to the first catalytic step. These cross-linked species could be then chased through the first splicing reaction (Kim & Abelson 1996).

Interestingly, two additional co-variations have been identified 5 nucleotides upstream of the ACUGAGA sequence in *C. merolae* (Fig. 8). This AAC sequence could base pair to the GUU sequence in the *C. merolae* intron at positions +5, +6 and +7 (Fig. 8). The corresponding residues in *S. cerevisiae* U6 snRNA, ACA, base pair to the UGU sequence at positions +5, +6 and +7 of the *S. cerevisiae* intron (Fig. 8 inset). In *S. cerevisiae*, this interaction is supported by a cross-link between the ACA of U6 and the 5' splice site of the lariat intron/exon2 splicing intermediate, suggesting that this interaction takes place

following the first catalytic reaction (Sawa & Abelson 1992). Furthermore, when this interaction was hyperstabilized by genetic mutation, aberrant cleavage was increased (Lesser & Guthrie 1993), again suggesting that the interaction is important following the first reaction. Taken together with other genetic and biochemical data, our co-variation data support Sawa and Abelson's (1992) proposal that base pairing between U6 snRNA and the pre-mRNA transcript undergoes a conformational rearrangement following the first splicing reaction.

U6 Covariation Supports a Recent Model for Free U6

The most striking divergence of *C. merolae* U6 snRNA is its non-conformity to the nearly invariant ACAGAGA region, of which there are only a few other biochemically characterized examples (Xu et al. 1994, Fast et al. 1998). The *C. merolae* U6 snRNA possesses an AC**U**GAGA sequence (i.e. U6 A68U, *C. merolae* numbering), which is notable since point mutations in this sequence in *S. cerevisiae* result in lethality *in vivo* and dramatically reduced levels of splicing *in vitro* (Madhani et al. 1990, McPheeters 1996, Fabrizio & Abelson 1990). Intriguingly, this A to U mutation is coupled in *C. merolae* with a complementary change in U6 that supports a recently proposed model of free U6 (Dunn & Rader 2010).

Considerable work over the years has focused on determining the secondary structure of free U6, ie. U6 prior to base pairing with U4, but the field has failed to settle on one model (Jandrositz & Guthrie 1995, Dunn & Rader 2010, Fortner et al. 1994, Vidaver et al. 1999, Karaduman et al. 2006, McManus et al. 2007). A recent reassessment of U6's intramolecular base pairing potential led to the suggestion that free U6 contains a three-helix junction, rather than the 3' internal stem loop previously proposed, as shown in Figure 9 (Dunn & Rader 2010). Unfortunately, structure probing data do not distinguish between the models, and the

high sequence conservation of U6 limits the availability of co-variation data that might differentiate between the models. The A68U sequence reported here is paired with a complementary U to A change at position 87 in the Dunn model, but not in other models of free U6. This provides the first co-variation support for the existence of Dunn's stem loop A.

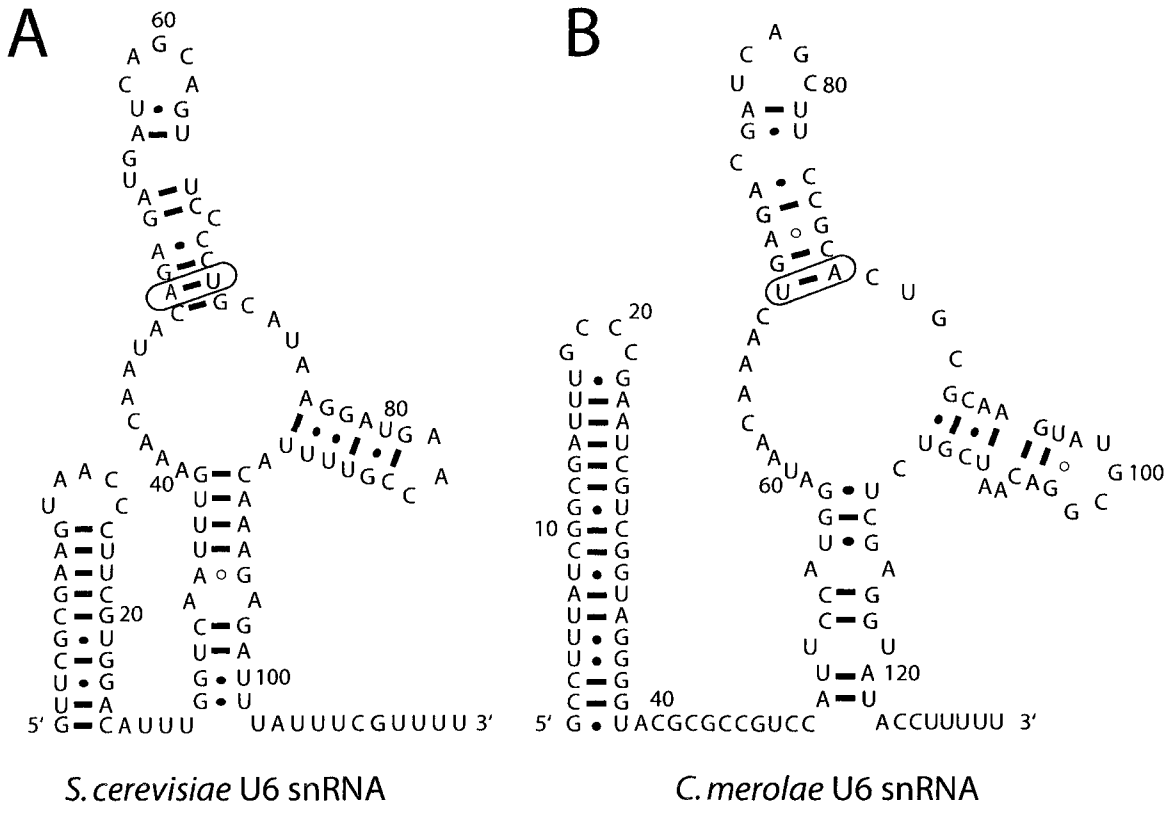

Figure 9: Intramolecular co-variation in U6 snRNA. The co-varying position is circled. A. Secondary structure model of U6 from the free U6 snRNP in *C. merolae*, showing the recently proposed stem loops A and B (Dunn & Rader 2010). B. Secondary structure model of free U6 snRNP in *S. cerevisiae*.

U4 has a Large Insertion

The presence of a 31 nucleotide stem loop located in the 3' half of U4, dividing U4/U6 stem III into two sections, is unique to *C. merolae* (Fig. 7). This stem loop is situated in a similar location to a three nucleotide bulge in the U4 side of the proposed *S. cerevisiae* U4/

U6 stem III (Brow & Vidaver 1995, Jakab et al. 1997). Given *C. merolae*'s harsh growing conditions in hot springs (pH 1.5 and 45 C), the additional stem loop might be required for increased structural stability, either of the RNA itself, or potential RNA/protein interactions. Alternatively, it is possible that this stem loop obviates the need for a protein to catalyze base pair formation between U4 and U6, as neither I nor others have found a PRP24 homologue (Misumi et al. 2005).

Chapter 5: Concluding Remarks and Future Directions

The identification of non-coding RNAs in newly sequenced genomes has become largely a matter of routine algorithmic exploration (Griffiths-Jones et al. 2005). It is important to remember, however, that computational results are only as good as the training dataset of known sequences on which our search model is based. This necessarily introduces an unfortunate bias into our view of RNA sequence space, relative to sequences that are found experimentally, and impoverishes knowledge of sequence diversity. My work provides an example of the importance of complementing computational approaches with experimental validation.

This first report of snRNAs in *C. merolae* provides compelling evidence that splicing in this extremophile organism proceeds via the normal spliceosomal reaction pathway, in spite of the small number of intron-containing substrates and apparent absence of numerous splicing factors. The splicing RNAs in *C. merolae* are notable for sequence changes within highly conserved elements, including in the branch point, its complementary region in U2, the 5' splice site, and its interacting region in U6: the nearly invariant ACAGAGA sequence. These compensatory changes support a U6/5' splice site interaction, as well as a switch in this interaction between the two chemical steps of splicing. It will be important to continue computational and biochemical characterization of *C. merolae* splicing factors to determine whether it contains the normal repertoire, and whether they differ substantially from those in other species. My comparative sequence analysis has identified the first true co-variations in the 5' splice site base pairs with U6 snRNA, providing phylogenetic support for these proposed interactions.

Future Directions

The discovery and identification of the four of the five snRNAs in *C. merolae* is but the first step in a new direction to determine the three dimensional structure of the snRNPs through the exploitation of *C. merolae*'s more robust and rigid snRNPs. In order to get snRNP crystals we need to bioinformatically determine the sequence of the snRNAs so we can design tagged primers complementary to each snRNA and use these oligos to pull the snRNP out of cell extract and begin crystallography trials. Once crystallized we can interpret the data and build a three dimensional structure for each snRNP and begin to speculate how these molecules function *in vivo*.

The trimethylguanosine caps present in all but the U6 snRNA, offer a useful target for possible antibody pulldown experiments. Antibodies targeting this cap could be used to pull down the snRNAs and their associated proteins and offer an opportunity to further confirm the snRNA results presented here, as well as providing a solution of snRNA associated proteins suitable for mass spectrometry. The protein fragments determined through mass spectrometry would be quite a useful companion to a *C. merolae* bioinformatic protein investigation.

Another interesting direction would be completing a genome wide confirmation of the introns proposed in the 2004 Matsuzaki paper, extending my splicing confirmation experiment to include all of *C. merolae*'s proposed introns. Once the true introns were confirmed, they could be used to form a more accurate training dataset for Infernal or Blast. A search with this more uniquely calibrated training set should find all remaining introns and once these were confirmed as above, we would have definitively determined *C. merolae*'s introns.

Works Cited

Ban N, Nissen P, Hansen J, Moore PB, and Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science.* **289**(5481):905-920.

Bindereif A and Green MR (1987) An ordered pathway of snRNP binding during mammalian pre-mRNA splicing complex assembly. *EMBO J.* **6**(8):2415-2424.

Brow DA and Vidaver RM (1995) An element in human U6 RNA destabilizes the U4/U6 spliceosomal RNA complex. *RNA.* **1**:122-131.

Cheng S and Abelson J (1987) Spliceosome assembly in yeast. *Genes & Development.* **1**:1014-1027.

Davila Lopez M, Rosenblad MA, and Samuelsson T (2008) Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res.* **36**(9):3001-3010.

Dunn EA and Rader SD (2010) Secondary structure of U6 small nuclear RNA: implications for spliceosome assembly. *Biochem Soc Trans.* **38**(4):1099-1104.

Fabrizio P and Abelson J (1990) Two domains of yeast U6 small nuclear RNA required for both steps of nuclear precursor messenger RNA splicing. *Science.* **250**(4979):404-409.

Fast NM, Roger AJ, Richardson CA, and Doolittle WF (1998) U2 and U6 snRNA genes in the microsporidian Nosema locustae: evidence for a functional spliceosome. *Nucleic Acids Research.* **26**(13):3202-3207.

Fortner DM, Troy RG, and Brow DA (1994) A stem/loop in U6 RNA defines a conformational switch required for pre-mRNA splicing. *Genes & Development.* **8**(2): 221-233.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, and Bateman A (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**:121-124.

Gutell RR, Lee JC, and Cannone JJ (2002) The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol.* **12**(3):301-310.

Hilliker AK and Staley JP (2004) Multiple functions for the invariant AGC triad of U6 snRNA. *RNA.* **10**:921-928.

Irimia M and Roy SW (2008) Spliceosomal introns as tools for genomic and evolutionary analysis. *Nucleic Acids Res.* **36**(5):1703-12.

Irimia M, Penny D, and Roy SW (2007) Coevolution of genomic intron number and splice sites. *Trends in Genetics.* **23**(7):318-325.

Jakab G, Mougin A, Kis M, Pollak T, Antal M, Branlant C, and Solymosy F (1997) Chlamydomonas U2, U4 and U6 snRNAs. An evolutionary conserved putative third interaction between U4 and U6 snRNAs which has a counterpart in the U4atac-U6atac snRNA duplex. *Biochemie.* **79**(7):387-395.

Jandrositz A and Guthrie C (1995) Evidence for a Prp24 binding site in U6 snRNA and in a putative intermediate in the annealing of U6 and U4 snRNAs. *EMBO J.* **14**(4):820-832.

Juneau K, Palm C, Miranda M, and Davis RW (2007) High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *PNAS.* **104**(5):1522-1527.

Jurica MS and Moore MJ (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell.* **12**(1):5-14.

Karaduman R, Fabrizio P, Hartmuth K, Urlaub H, and Lührmann R (2006) RNA structure and RNA-protein interactions in purified yeast U6 snRNPs. *J Mol Biol.* **356**(5):1248-1262.

Kershaw CJ, Barrass D, Beggs JD, and O'Keefe RT (2009) Mutations in the U5 snRNA result in altered splicing of subsets of pre-mRNAs and reduced stability of Prp8. *RNA.* **15**:1292-1304.

Kim CH and Abelson J (1996) Site-specific crosslinks of yeast U6 snRNA to the pre-mRNA near the 5' splice site. *RNA.* **2**(10):995-1010.

Konarska MM and Sharp PA (1987) Interactions between small nuclear ribonucleoprotein particles in formation of spliceosomes. *Cell.* **49**:763-774.

Konarska MM, Grabowski PJ, Padgett RA, and Sharp PA (1985) Characterization of the branch site in lariat RNAs produced by splicing of mRNA precursors. *Nature.* **313**:552-557.

Kuroiwa T (1998) The primitive red algae Cyanidium caldariumand Cyanidioschyzon merolaeas model system for investigating the dividing apparatus of mitochondria and plastids. *BioEssays.* **20**:344-354.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, and Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics.* **23**(21):2947-2948.

Lesser CF and Guthrie C (1993) Mutations in U6 snRNA that alter splice site specificity: implications for the active site. *Science.* **262**:1982-1988.

Li Z and Brow DA (1993) A rapid assay for quantitative detection of specific RNAs. *Nucleic Acids Research.* **21**(19):4645-4646.

Madhani HD, Bordonné R, and Guthrie C (1990) Multiple roles for U6 snRNA in the splicing pathway. *Genes & Development.* **4**:2264-2277.

Matsuzaki M, Misumi O, Shin-i T, Maruyama S, Takahara M, Miyagishima S, Mori T, Nishida K, Yagisawa F, Nishida K, Yoshida Y, Nishimura Y, Nakao S, Kobayashi T, Momoyama Y, Higashiyama T, Minoda A, Sano M, Nomoto H, Oishi K, Hayashi H, Ohta F, Nishizaka S, Haga S, Miura S, Morishita T, Kabeya Y, Terasawa K, Suzuki Y, Ishii Y, Asakawa S, Takano H, Ohta N, Kuroiwa H, Tanaka K, Shimizu N, Sugano S, Sato N, Nozaki H, Ogasawara N, Kohara Y, and Kuroiwa T (2004) Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature.* **428**:653-657.

McManus CJ, Schwartz ML, Butcher SE, and Brow DA (2007) A dynamic bulge in the U6 RNA internal stem-loop functions in spliceosome assembly and activation. *RNA.* **13**(12): 2252-2265.

McPheeters DS (1996) Interactions of the yeast U6 RNA with the pre-mRNA branch site. *RNA.* **2**(11):1110-1123.

Minoda A, Sakagami R, Yagisawa F, Tsuneyoshi, and Tanaka K (2004) Improvement of Culture Conditions and Evidence for Nuclear Transformation by Homologous Recombination in a Red Alga, Cyanidioschyzon merolae 10D. *Plant Cell Physiol.* **45**(6): 667-671.

Misumi O, Matsuzaki M, Nozaki H, Miyagishima S, Mori T, Nishida K, Yagisawa F, Yoshida Y, Kuroiwa H, and Kuroiwa T (2005) Cyanidioschyzon merolae genome. A tool for facilitating comparable studies on organelle biogenesis in photosynthetic eukaryotes. *Plant Physiol.* **137**(2):567-85.

Nawrocki EP, Kolbe DL, and Eddy SR (2009) Infernal 1.0: Inference of RNA alignments. *Bioinformatics.* **25**(10):1335-1337.

Noller HF, Kop J, Wheaton V, Brosius J, Gutell RR, Kopylov AM, Dohme F, and Herr W (1981) Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Research.* **9**(22): 6167-6189.

Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, Maruyama S, Nishida K, Yagisawa F, Yoshida Y, Fujiwara T, Takio S, Tamura K, Chung SJ, Nakamura S, Kuroiwa H, Tanaka K, Sato N, and Kuroiwa T (2007) A 100% complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae. *BMC Biol.* **5**:28.

Ohnuma M, Yokoyama T, Inouye T, Sekine Y, and Tanaka K (2008) Polyethylene Glycol (PEG)-Mediated Transient Gene Expression in a Red Alga, Cyanidioschyzon merolae 10D. *Plant Cell Physiol.* **49**(1):117-120.

Ohta N, Matsuzaki M, Misumi O, Miyagishima S, Nozaki H, Tanaka K, Shin-i T, Kohara Y, and Kuroiwa T (2003) Complete Sequence and Analysis of the Plastid Genome of the Unicellular Red Alga Cyanidioschyzon merolae. *DNA Research.* **10**:67-77.

Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA, Almabrazi HG, Pedersen KF, Lin Y, Garretson J, McEntaggart NO, Sailor CA, Dawson RB, and Peek AS (2008) IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Res.* **36** (Web Server issue):163-9.

Padgett RA, Konarska MM, Grabowski PJ, Hardy SF, and Sharp PA (1984) Lariat RNAs as intermediates and products in the splicing of messenger RNA precursors. *Science.* **225**:898-903.

Parker R, Siliciano PG, and Guthrie C (1987) Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell.* **49**:229-239.

Pomeranz Krummel DA, Oubridge C, Leung AKW, Li J, and Nagai K (2009) Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature.* **458**(7237):475-480.

Roy SW and Gilbert W (2005) Rates of intron loss and gain: Implications for early eukaryotic evolution. *PNAS.* **102**(16):5773-5778.

Roy SW and Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet.* **7**(3):211-21.

Sashital DG, Cornilescu G, McManus CJ, Brow DA, and Butcher SE (2004) U2-U6 RNA folding reveals a group II intron-like domain and a four-helix junction. *Nat Struct Mol Biol.* **11**(12):1237-1242.

Sawa H and Abelson J (1992) Evidence for a base-pairing interaction between U6 small nuclear RNA and the 5' splice site during the splicing reaction in yeast. *Proc. Nati. Acad. Sci. USA.* **89**:11269-11273.

Seraphin B, Kretzner L, and Rosbash M (1988) A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. *EMBO J.* **7**(8):2533-2538.

Siliciano PG and Guthrie C (1988) 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. *Genes Dev.* **2**:1258-1267.

Siliciano PG, Brow DA, Roiha H, and Guthrie C (1987) An essential snRNA from S. cerevisiae has properties predicted for U4, including interaction with a U6-like snRNA. *Cell.* **50**:585-592.

Stevens SW, Ryan DE, Ge HY, Moore RE, Young MK, Lee TD, and Abelson J (2002) Composition and Functional Characterization of the Yeast Spliceosomal Penta-snRNP. *Molecular Cell.* **9**:31-44.

Terui S, Suzuki K, Takahashi H, Itoh R, and Kuroiwa T (1995) Syncronization of choloroplast division in the ultramicroalga cyanidioschyzon merolae (rhodophyta) by treatment with light and aphidicolin. *Journal of Phycology.* **31**(6):958-961.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, and Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research.* **25**(24):4876-4882.

Umen JG and Guthrie C (1995) The second catalytic step of pre-mRNA splicing. *RNA.* **1**:869-885.

Vidaver RM, Fortner DM, Loos-Austin LS, and Brow DA (1999) Multiple functions of Saccharomyces cerevisiae splicing protein Prp24 in U6 RNA structural rearrangements. *Genetics.* **153**(3):1205-1218.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, and et al PA (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature.* **420**:520-562.

Wimberly BT, Brodersen DE, Jr WMC, Carter RJMAP, Vonrhein C, Hartsch T, and Ramakrishnan V (2000) Structure of the 30S ribosomal subunit. *Nature.* **407**(6802):306-307.

Woese CR, Gutell R, Gupta R, and Noller HF (1983) Detailed Analysis of the Higher-Order Structure of 16S-Like Ribosomal Ribonucleic Acids. *Microbiology Reviews.* **47**(4):621-669.

Xu G, Wieland B, and Bindereif A (1994) trans-Spliceosomal U6 RNAs of Crithidia fasciculata and Leptomonas seymouri: Deviation from the Conserved ACAGAG Sequence and Potential Base Pairing with Spliced Leader RNA. *Molecular and Cellular Biology.* **14**(7): 4565-4570.

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