INTRON RETENTION AND RECOGNITION IN THE MICROSPORIDIAN *ENCEPHALITOZOON CUNICULI*

by

Renny Lee

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2008

ABSTRACT

Microsporidia are unicellular fungi that are intracellular parasites of animals, including humans. They are both complex and simple, armed with a sophisticated infection apparatus and possessing the smallest eukaryotic nuclear genomes. The microsporidian *Encephalitozoon cuniculi* has a genome size of 2.9 Mb, which is smaller than many bacterial genomes. Genome reduction and compaction in size, content, and form has been interpreted as an adaptation to parasitism. One of the effects of genome size reduction concerns intron evolution – *E. cuniculi* has retained only a few extremely short spliceosomal introns.

This thesis examines the splicing of introns in the spore stage. The introns were retained in spores, suggesting life-stage specific splicing and splicing inhibition. How the short introns are recognized was also examined. Unique splicing signal motifs were predicted, and were used to find additional introns. The intron density was doubled for this species, and I also obtained data that counter current views about intron evolution in compacted genomes with low intron densities. I also predict that *E. cuniculi* introns are recognized in a unique way by the spliceosome.

TABLE OF CONTENTS

List of Tables

List of Figures

Acknowledgments

I would like to thank the following people who had direct roles in assisting me through this research experience.

My supervisor, Dr. Naomi Fast. You're super-nice and super-bright. Thank you for taking me on as a student, and for so long, for your patience, and for training me. For being extremely approachable, and almost always in a good mood. I don't know where you get your energy, insane, really. I learned a lot while working with you. I will definitely miss working with you Naomi. May all your scientific dreams come true…

My committee member, Dr. Patrick Keeling. Thank you for your insights – especially the ones that materialized out of the ether, and honest criticisms, and keeping an open door.

My other committee member, Dr. Michael Murphy. You were an extremely under-utilized resource. Thank you for being on my committee.

When I first started, the Fast and Leander Labs shared a common space. Thanks to Dr. Mona Hoppenrath for showing me how to pour plates, and Susana Breglia for walking me through my first cloning reaction and showing me how to use the autoclave.

Erin, we started out at the same time, for showing me the gutter is tolerable. We were almost always by ourselves in the lab, and your presence and work ethic motivated me to work that much harder. May your drive and toughness bring you everything you'll ever want. Val, come back soon.

Past and present members of the Keeling Lab. A talented bunch, you were all invaluable. I hope to see you all again. Dr. Bryony Williams, for showing me the ropes during my warm-up project. And for allowing me to bug the hell out of you. Thank you for training me as well. Dr. Kevin Carpenter, for your wisdom through experience. Dr. Nico

Corradi, apologies for bugging the hell out of you. I wish you every success. Chitchai, for your wisdom at such a young age.

My world begins and ends with some. I want to thank them for their support.
My siblings.
Sunshine, "Whatever!", I miss you already…
Dan, back soon. *Oss*.
Spot OV, for our need for speed. For the Rush, and the Watch.
Big Lil' Bro A. For heaps. *Zanshin.*
Mara, for almost everything.

to my Dad

Van Iong Lee, *Amah*
Year of the Tiger.

Co-authorship Statement

Dr. Naomi Fast conceived of testing the splicing status of the annotated *E. cuniculi* introns as presented in Chapter 2, and I physically carried out this research on spores as specified in Chapter 2. Valerie Limpright and Erin Gill also contributed by testing for splicing status in spores and in meronts. I analyzed the data and drafted Chapter 2; Dr. Naomi Fast also analyzed data and contributed to the editing and writing of Chapter 2.

I conceived of the idea and method of searching for more introns in *E. cuniculi*, as described in Chapter 3. Erin Gill provided the raw material – *E. cuniculi* meront RNA – for this research, which I physically carried out. I analyzed the data and drafted Chapter 3; Dr. Naomi Fast also analyzed data and contributed to the editing and writing of Chapter 3.

CHAPTER 1. General Introduction

1.1 Microsporidia

Microsporidia are unicellular eukaryotes closely-related to fungi and comprise a large and diverse group of obligate intracellular parasites (Keeling and Fast, 2002). Infecting all animal phyla, several are also human parasites (Keeling and Fast, 2002). They possess eukaryotic features and structures that have diverged greatly in content, form, and function; a reflection of their parasitic lifestyles. One feature that has attracted the most attention is the mitosome, a mitochondrial remnant that has experienced severe reductive evolution (Tsaousis et al., 2008). Another highly-derived state concerns their genomes: they possess nuclear genomes that have undergone extreme size reduction (Keeling and Slamovits, 2005). As a result, their nuclear genomes, with a size range of 2.3 to 19.5 Mb, are the absolute smallest known amongst all eukaryotes (Keeling and Slamovits, 2005).

Sequence-level genome information is available for several microsporidia, but only one genome has been fully-sequenced. With a genome size of 2.9 Mb, *Encephalitozoon cuniculi* has one of the smaller genomes (Katinka et al., 2001). Its sequencing revealed extreme reduction, manifested as massive gene loss, and extreme compaction, by compressing what remains into a smaller space, with a high gene-density, achieved by shortening both genes and the intergenic spaces separating them. For example, *E. cuniculi* has retained only 1997 protein-coding genes, many of shortened length compared to yeast homologues, separated by an average intergenic distance of only 129 bases (Katinka et al., 2001). Sequencing also revealed massive intron loss and compaction: only twelve introns of reduced length have been retained in the *E. cuniculi* genome (Katinka et al., 2001).

A side effect of genome compaction in microsporidia is overlapping transcription: transcription for a gene can start within an upstream gene, and transcripts can contain multiple genes, hypothesized to be caused by loss or relocation of transcriptional control elements into adjacent genes (Williams et al., 2005). First described in *Antonospora*

locustae, this feature has been recently demonstrated for *E. cuniculi* transcripts as well (Corradi et al., 2008).

Besides their genomes, microsporidia are probably best known for their morphology and infection machinery. Outside of their hosts they exist as spores, encased by an environmentally-resistant coat, and are presumably dormant, but infectious (Keeling and Fast, 2002). The spore contains mainly structures specialized for infection. The prominent one is the polar tube, a tube coiled around the spore contents that, upon germination, everts, shoots out of the spore, and pierces a nearby host cell (Keeling and Fast, 2002). Spore contents are then extruded – also through this polar tube – into the host, an environment where microsporidia replicate (the 'meront' life-stage) before switching back to the infectious spore state (Keeling and Fast, 2002). *E. cuniculi*'s life-cycle is straightforward (as described), however many microsporidia have more elaborate life-cycles with multiple spore types and hosts (Becnel et al., 2005). The host-range is wide for *E. cuniculi* and includes humans, but the major host is rabbit (Mathis et al., 2005).

1.2 Introns

Introns, the intragenic (and intervening) regions in pre-mRNA, are sequences that are removed during mRNA maturation for the proper gene expression of exons, the expressed regions present in the mature mRNA transcript. Introns are also one of the hallmarks of eukaryotic genomes. Discovered over thirty years ago (Berget et al., 1977; Chow et al., 1977), their evolutionary origins and significance in genome evolution have been controversial to this day (Koonin, 2006). Intron removal, or splicing, is mediated by a macromolecular machine called the spliceosome. It has been suggested that the ancestor of all extant eukaryotes possessed a complex spliceosome (Collins and Penny, 2005), so it follows that introns likely arose and became widely distributed within nuclear genomes very early in eukaryotic evolution.

1.3 The intron splicing reaction is spliceosome-mediated

The splicing process is one of the fundamental steps in eukaryotic gene expression. It involves the excision of the intron from the pre-mRNA and subsequent ligation of the exons to yield translatable mRNA. These reactions are accomplished by the spliceosome, a large ribonucleoprotein complex consisting of five different subunits; each composed of a small nuclear RNA (snRNA) and a multitude of proteins unique to each subunit (reviewed in Jurica and Moore, 2003). The earliest studies on the splicing reaction focused on a variety of *in vivo* and *in vitro* models, particularly on the budding yeast *Saccharomyces cerevisiae* and mammalian cell lines (Jurica and Moore, 2003). Later, complementary studies in other organisms, particularly those of the genetic animal models, the fission yeast *Schizosaccharomyces pombe*, and the mustard weed *Arabidopsis thaliana*, showed the basic mechanism of splicing is conserved (Jurica and Moore, 2003).

Briefly and greatly simplified, splicing can be partitioned into a two-step process, both of which can be separated into a series of more steps (for a comprehensive review see Jurica and Moore, 2003):

(1) In the first step, the intron is recognized by the spliceosome. The U1 small nuclear ribonucleoprotein (snRNP) associates with the 5' splice site (SS) of the intron. The 5'SS consists of an invariant 5'-GU dinucleotide followed by a small stretch of ribonucleotides, forming a recognition motif usually six bases in length. This association between the U1 snRNP is accomplished by the complementary base-pairing between the 5'SS and the U1 snRNA. Another snRNP-intron association forms; the recognition of the intron branchpoint motif by the U2 snRNA. Complementary base-pairing between the U2 snRNA and the branchpoint motif results in an unpaired residue in the intron, called the branchpoint adenosine. Upon exposure of the bulged adenosine, a tri-snRNP complex consisting of the U4, U5, and U6 snRNPs associates with the intron. More specifically, the U6 snRNA associates with the 5'SS, and the U5 snRNP associates with both exons, bridging them. The U1 and U4 snRNPs are also released from the intron. Finally, the exposed branchpoint adenosine residue acts as a nucleophile, and attacks the

guanine residue of the 5'-GU dinucleotide and via a trans-esterification reaction, forms two products: the first exon and the intron-lariat-second exon.

(2) The second step completes the splicing reaction. The last residue of the excised exon acts as a nucleophile, and attacks the 3'SS. The 3'SS consists of an invariant dinucleotide at the intron boundary, an AG-3'. This trans-esterication reaction releases the intron-lariat and ligates the two exons.

The splicing reaction is a complex, ordered, and dynamic process comprised of a series of rearrangements involving various types of interactions: RNA-RNA, as demonstrated by intron-snRNA and snRNA-snRNA associations; RNA-protein, through intron-spliceosomal protein and snRNA-protein associations; and protein-protein interactions in the spliceosome.

1.4 Introns and microsporidia: the initial studies

The impetus for the initial spliceosome and intron studies in microsporidia was partly phylogenetic, since at that time, microsporidia were considered early-diverging eukaryotes. These studies strived to determine whether microsporidia have markers hypothesized to have been introduced by the endosymbiont that gave rise to mitochondria, these markers being the snRNAs and introns that allegedly evolved from fragmentation of self-splicing introns present in the alpha-proteobacterial mitochondrial ancestor (Cavalier-Smith, 1991). However, increasing numbers of phylogenetic studies soon began to suggest that microsporidia were not members of the Archezoa, a eukaryotic lineage that diverged prior to the acquisition of mitochondria. Rather, the various gene trees began to show that microsporidia was not early-diverging at all, and likely are either a sister group to fungi or are true fungi (reviewed in Keeling and Fast, 2002).

The first study involved cloning the U2 snRNA gene of the microsporidian *Vairamorpha necatrix*, and testing for its expression (Dimaria et al., 1996). The U2 snRNA, although highly-divergent and lacking many motifs, was nonetheless shown to be expressed. A similar and subsequent study involved cloning the U2 and U6 snRNAs of the

microsporidian *Antonospora locustae* (Fast et al., 1998). The study by Fast et al. (1998) showed that the most structurally conserved snRNAs, U2 and U6, were indeed present in the *A. locustae* genome. Further, the two snRNAs were shown to be expressed in spores, and RNA folding algorithms predicted folding profiles similar to that known for active orthologous snRNAs. These results strongly suggested the presence of splicing in microsporidia, despite the fact that, at the time, no introns were known since only a few microsporidian genes were sequenced. At the same time, a separate study by Fast et al. (1999) examined the phylogenetic placement of *A. locustae* based on the TATA Box Binding protein (TBP) gene. Most importantly, this study showed that introns present in the TBP gene of the ancestor of plants, fungi, and animals were all absent in *A. locustae*. This result suggested that microsporidia likely lost many spliceosomal introns throughout evolution. These three studies, combined with the genome size estimate of 2.9 Mb for the microsporidian *Encephalitozoon cuniculi*, predicted that microsporidian genomes were unlikely to contain intron densities comparable to other eukaryotes.

The very first microspordian intron, with canonical GU-AG intron boundaries, was predicted in a ribosomal protein-encoding gene - *L27a* of *E. cuniculi* (Biderre et al., 1998). This predicted intron was extremely short in length at 28 nt and was positioned immediately after the initiation codon ATG. Studies of the *L27a* homolog in model genetic systems have been limited to a single study in *S. cerevisiae* (DeLabre et al., 2002). That study demonstrated that L27a, while not essential for translation or viability, was required for the proper joining of the small and large ribosome subunits and for normal translation rates. Also, the protein contains a conserved N-terminal domain, which also happens to be present in the predicted coding region of the *E. cuniculi L27a* gene. In fact, this N-terminal domain is positionally-conserved across eukaryotes, as demonstrated by a protein sequence alignment (Biderrre et al., 1998). The intron was predicted based on its interruptive nature: left unspliced, a frameshift introduces a stop codon that results in premature translational termination eight amino-acyl residues into the transcript.

In conclusion these initial studies in microsporidia suggested that splicing may be active in microsporidian spores, and is probably less frequent. The prediction of the first intron also hinted at an ability to splice extremely short introns.

1.5 Information from the *Encephalitozoon cuniculi* genome sequence

Sequencing the *E. cuniculi* genome unveiled how little intron and splicing content is encoded by its genome. Of the annotated gene complement in *E. cuniculi* of approximately 1997 protein-coding genes, only 12 introns were originally annotated in the 2.9 Mb genome (Katinka et al., 2001). Ten of these introns are annotated in ten different ribosomal protein-encoding genes (RPGs), including previously described intron in *L27a* (Biderre et al., 1998). The other two introns are annotated in a single gene, annotated to code for a CDP-diacylglycerol serine phosphatidyltransferase. These annotated introns are remarkable for several reasons. Aside from their preponderance in ribosomal genes, the introns are also amongst the shortest known in length – with annotated lengths ranging from 23 to 52 bases, and occur very close to the translational start site, typically immediately after it. Also, like that of *L27a*, almost all the annotated introns cause frameshifts and introduce stop codons which would impede translational expression to protein products, if the introns are not spliced. In summary, the twelve *E. cuniculi* introns were annotated based on their capacity to introduce frameshifts, their possession of the canonical eukaryotic intron boundary splice sites 5'GU-AG3', and their disruptive nature in highly-conserved RPGs.

Besides the introns, sequencing the *E. cuniculi* genome also identified spliceosomal components (Katinka et al., 2001). For example, the genes for the two most highly-conserved snRNAs U2 and U6 are present, and approximately 30 spliceosomal proteins were also predicted, including the highly-conserved Prp8, a large spliceosome subunit with a potential role in splicing catalysis (Grainger and Beggs, 2005). The number of splicing protein components is alarmingly small, given that other eukaryotes have so many. For example, various estimates have suggested that the yeast *S. cerevisiae* spliceosome has at least 80 proteins, and as many as ~300, with these numbers being typical of spliceosomes from other organisms (Collins and Penny, 2005; Jurica and Moore, 2003). A subsequent

bioinformatics analyses of splicing components across a range of eukaryotes showed that *E. cuniculi* contains a few more splicing proteins than what was originally annotated (Collins and Penny, 2005). However, given that *E. cuniculi* (and microsporidia in general) possess such high levels of sequence divergence (Thomarat et al., 2004), it is likely these are underestimates.

In addition to the twelve annotated introns from the *E. cuniculi* genome, another three introns, in three different RPGs, were later annotated. Unlike the majority of the twelve introns, these three introns were annotated based not on their ability to cause frameshifts – they are in-frame with the coding sequences – but due to their GU-AG boundaries, and by the indels they create in otherwise conserved sequences. These three introns, like the originally-annotated twelve, are also extremely short and are near the translational start codons.

A sixteenth intron was later predicted in *E. cuniculi* when sequence comparisons between a known gene sequence in *A. locustae* and the annotated *E. cuniculi* homolog revealed extensive sequence similarity at the N-termini of the two encoded proteins (Wu et al., 2007). Based on this sequence conservation, the *E. cuniculi Sec61α* gene was re-annotated by extending the gene. This re-annotation predicted a short intron, which would need to be spliced for proper gene expression. The prediction of this intron suggests current annotations may be conservative, and highlights the problem of annotating introns in genes that are highly-divergent, particularly if introns are at the extreme 5' ends of genes.

Finally, in a proteomic analysis, peptide mass-fingerprinting from a *E. cuniculi* spore sample showed three of the RPGs annotated to harbor introns are expressed in the spore (Brosson et al., 2006). These intron-containing RPGs are *S17*, *S26*, and *L7a*. For *S17* and *S26*, protein expression is dependent on splicing, as frameshifts are introduced. For *L7a*, protein expression is not dependent on splicing since the intron neither introduces a frameshift, nor encodes a stop codon. These results provide additional, yet indirect, evidence to suggest splicing occurs in microsporidia.

To summarize, genomic and post-genomic analyses of *E. cuniculi* suggests the retention of only a handful of introns, all of reduced length. The short intron lengths and the retention of comparatively few spliceosomal proteins also suggest a paring-down of intron sequences and spliceosomal protein components to the bare essentials of what may be needed for splicing function. The positional and gene-category bias of the majority of the predicted introns to the beginnings of RPGs may reflect some sort of constraint, as will be described in the next section.

1.6 Intron evolution in *E. cuniculi*

That only a few introns, all of compacted length, are retained in the *E. cuniculi* genome may reflect several possibilities: they are kept for whatever functional qualities they have acquired, the introns are not selected for and their presence represents an intermediary stage in genome reduction, or a mixture of both.

If the remaining introns are simply remnants of a reducing genome, the introns would be expected to have no function, with their presence reflecting an inability or difficulty for their loss. For example, genomic deletion of the intron may be very inefficient and highly deleterious. Despite this, intron-shortening by deletion has no doubt occurred, given the compacted length of the retained introns. A different intron loss mechanism explains the 5' positional gene bias of the introns: introns would still be present because they could not be removed by a hypothetical mechanism involving recombination of a reverse-transcribed (RT) spliced cDNA product with the homologous genomic region (Boeke et al., 1985). This method preferentially removes more 3' positioned introns. This mechanism has been hypothesized to explain the 5' positional gene bias that is also seen for the introns of the related yeasts *S. cerevisiae* and *Ashyba gossyphi*, and also in the nucleomorph of *Guillardia theta* (Brachat et al., 2003; Douglas et al., 2001; Spingola et al., 1999). Nucleomorphs genomes are highly-reduced and compacted, and are found in only two unrelated algal lineages that have swallowed a photosynthetic alga and retained its chloroplast and nucleus (the nucleomorph). The *G.*

theta nucleomorph has retained only seventeen introns of relatively short length (Douglas et al., 2001). The *Bigellowiella natans* nucleomorph has retained many introns, including even shorter introns, including the shortest introns known (Gilson et al., 2006).

Another possibility is that the introns of *E. cuniculi* may be retained for a functional role. This hypothesis is popular because almost all sixteen annotated *E. cuniculi* introns are in RPGs. That we see this positional- and RPG-bias for the few introns present in other reduced genomes, like that of *S. cerevisiae* and the *G. theta* nucleomorph, also supports this view.

Studies have also shown that splicing is tightly-associated with other transcriptional and post-transcriptional events in the cell. Many of these processes are co-regulated with splicing, including transcription, mRNA capping, mRNA polyadenylation, and mRNA export out of the nucleus (Bentley, 2005). Further, since the majority of the *E. cuniculi* introns are in RPGs, it is possible they are involved in transcriptional regulation of their own mRNAs. In fact, studies of some *S. cerevisiae* mRNAs show that introns are indeed involved in post-transcriptional control. In one case, the intron in a transcript of a RPG forms a stable secondary structure through folding, in effect preventing further processing (Li et al., 1995). Protein levels modulated through negative feedback by intron-mediated autoregulation is not limited to ribosome components and their transcripts, however. As another example, an RNA export factor is also autoregulated by a similar process involving the intron forming a secondary structure that prevents further transcript processing (Preker and Guthrie, 2006). In another study, genes containing short introns are transcribed and exported out of the nucleus at greater rates than their counterparts without introns (Yu et al., 2002). Therefore, potential roles in ribosome synthesis and mRNA export could be envisioned for some of the *E. cuniculi* introns.

1.7 Introns in microsporidia: motives and thesis aims

Microsporidia offer a unique opportunity to study a process that is unwieldy to analyze in even the most tractable model systems. The splicing machinery and its intron targets are

considerably less complex in *E. cuniculi* than in organisms that have not undergone genome reduction. The annotated *E. cuniculi* introns are, for example, among the shortest known, and it is unknown how they might be recognized by the spliceosome. With just sixteen introns, *E. cuniculi* also has one of the lowest intron densities among eukaryotes. Along the same lines, the eukaryotic spliceosome may be the most complex macromolecular machine ever described (Nilsen, 2003). However, the *E. cuniculi* spliceosome, annotated to consist of just several dozen proteins (Katinka et al., 2001) may also be one of the least complex spliceosomes. Therefore, not only could divergent and potentially novel genome reduction-induced changes in the splicing process be discovered, but conserved mechanistic principles of spliceosomal function and intron recognition may also be revealed.

The knowledge obtained about microsporidian splicing may also give general information as to how genomes are reduced and what mechanisms intron loss may entail. As already mentioned, it is unknown if the annotated *E. cuniculi* introns are remnants of neutral evolution or products of selection. One strategy could involve biochemical assessments of the degree of spliced product for each of the predicted introns. Another complementary strategy may involve looking for introns and examining their splicing status in a range of microsporidia. In addition, these approaches may give clues as to whether the retained introns are selected for their ability to impart regulatory roles during gene expression - through regulated splicing, for instance.

The aims of this thesis are two-fold: to demonstrate that the annotated *E. cuniculi* introns are spliced, by examining a number of their transcript structures, and second, to assess how *E. cuniculi* introns are recognized.

References

Becnel, J.J., White, S.E., and Shapiro, A.M. (2005). Review of microsporidia-mosquito relationships: from the simple to the complex. *Folia Parasitol. (Praha).* **52**: 41-50.

Bentley, D.L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell. Biol.* **17**: 251-256.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U S A.* **74**: 3171-3175.

Biderre, C., Méténier, G., and Vivarès, C.P. (1998). A small spliceosomal-type intron occurs in a ribosomal protein gene of the microsporidian *Encephalitozoon cuniculi*. *Mol. Biochem. Parasitol.* **94**: 283-286.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. *Cell.* **40**: 491-500.

Brachat, S., Dietrich, F.S., Voegeli, S., Zhang, Z., Stuart, L., Lerch, A., Gates, K., Gaffney, T., and Philippsen, P. (2003). Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biol.* **4**: R45.

Brosson, D., Kuhn, L., Delbac, F., Garin, J., Vivarès, C.P., and Texier, C. (2006). Proteomic analysis of the eukaryotic parasite *Encephalitozoon cuniculi* (microsporidia): a reference map for proteins expressed in late sporogonial stages. *Proteomics.* **6**: 3625-3635.

Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. *Trends Genet.* **7**: 145-8.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell.* **12**: 1-8.

Collins, L., and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22**: 1053-1066.

Corradi, N., Gangaeva, A., and Keeling, P.J. (2008). Comparative profiling of overlapping transcription in the compacted genomes of microsporidia *Antonospora locustae* and *Encephalitozoon cuniculi. Genomics.* **91**: 388-393.

DeLabre, M.L., Kessl, J., Karamanou, S., and Trumpower, B.L. (2002). *RPL29* codes for a non-essential protein of the 60S ribosomal subunit in *Saccharomyces cerevisiae* and exhibits synthetic lethality with mutations in genes for proteins required for subunit coupling. *Biochim. Biophys. Acta.* **1574**: 255-261.

DiMaria, P., Palic, B., Debrunner-Vossbrinck, B.A., Lapp, J., Vossbrinck, and C.R. (1996). Characterization of the highly divergent U2 RNA homolog in the microsporidian *Vairimorpha necatrix. Nucleic Acids Res.* **24**: 515-522.

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.T., Wu, X., Reith, M., Cavalier-Smith, T., and Maier, U.G. (2001). The highly reduced genome of an enslaved algal nucleus. *Nature*. **410**: 1091-1096.

Fast, N.M., Logsdon, J.M. Jr., and Doolittle, W.F. (1999). Phylogenetic analysis of the TATA box binding protein (TBP) gene from *Nosema locustae*: evidence for a microsporidia-fungi relationship and spliceosomal intron loss. *Mol. Biol. Evol.* **16**: 1415-1419.

Fast, N.M., Roger, A.J., Richardson, C.A., and Doolittle, W.F. (1998). U2 and U6 snRNA genes in the microsporidian *Nosema locustae*: evidence for a functional spliceosome. *Nucleic Acids Res.* **26**: 3202-3207.

Gilson, P.R., Su, V., Slamovits, C.H., Reith, M.E., Keeling, P.J., and McFadden, G.I. (2006). Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc. Natl. Acad. Sci. U S A.* **103**: 9566-9571.

Grainger, R.J., and Beggs, J.D. (2005). Prp8 protein: at the heart of the spliceosome. *RNA*. **11**: 533-557.

Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell.* **12**: 5-14.

Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., and Vivarès, C.P. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature.* **414**: 450-453.

Keeling, P.J., and Fast, N.M. (2002). Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu. Rev. Microbiol.* **56**: 93-116.

Keeling, P.J., and Slamovits, C.H. (2005). Causes and effects of nuclear genome reduction. *Curr. Opin. Genet. Dev.* **15**: 601-608.

Koonin, E.V. (2006). The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol. Direct.* **1**: 22.

Li, Z., Paulovich, A.G., and Woolford, J.L. Jr. (1995). Feedback inhibition of the yeast ribosomal protein gene *CRY2* is mediated by the nucleotide sequence and secondary structure of *CRY2* pre-mRNA. *Mol. Cell. Biol.* **15**: 6454-6464.

Mathis, A., Weber, R., and Deplazes, P. (2005). Zoonotic potential of the microsporidia. *Clin. Microbiol. Rev.* **18**: 423-45.

Nilsen, T.W. (2003). The spliceosome: the most complex macromolecular machine in the cell? *Bioessays.* **25**: 1147-1149.

Preker, P.J., and Guthrie, C. (2006). Autoregulation of the mRNA export factor Yra1p requires inefficient splicing of its pre-mRNA. *RNA.* **12**: 994-1006.

Spingola, M., Grate, L., Haussler, D., and Ares, M. Jr. (1999). Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA*. **5**: 221-234.

Thomarat, F., Vivarès, C.P., and Gouy, M. (2004). Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. *J. Mol. Evol.* **59**: 780-791.

Tsaousis, A.D., Kunji, E.R., Goldberg, A.V., Lucocq, J.M., Hirt, R.P., and Embley, T.M. (2008). A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi. Nature.* **453**: 553-556.

Williams, B.A., Slamovits, C.H., Patron, N.J., Fast, N.M., and Keeling, P.J. (2005). A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc. Natl. Acad. Sci. U S A.* **102**: 10936-10941.

Wu, Z., Li, Y., Pan, G., Li, C., Hu, J., Liu, H., Zhou, Z., and Xiang, Z. (2007). A complete Sec61 complex in *Nosema bombycis* and its comparative genomics analyses. *J. Eukaryot. Microbiol.* **54**: 379-380.

Yu, J., Yang, Z., Kibukawa, M., Paddock, M., Passey, D.A., and Wong, G.K. (2002). Minimal introns are not "junk". *Genome Res.* **12**: 1185-1189.

CHAPTER 2: Intron-retention in *Encephalitozoon cuniculi* transcripts[1]

2.1 Introduction

Splicing of the sixteen annotated *E. cuniculi* introns has never been demonstrated, but is a necessary first-step in validating that they are true introns and that splicing occurs. The introns possess characteristics that support the idea of their splicing: they all have the canonical GU-AG boundaries, and they introduce frameshifts in genes encoding key ribosomal components that are highly-conserved at the sequence level. Three of the sixteen intron-containing genes' protein products are also present in *E. cuniculi* spores, providing indirect evidence for splicing (Brosson et al., 2006). Lastly, the *E. cuniculi* genome annotation lists both protein and RNA components of the spliceosome, a highly-complex protein machine that is costly to maintain (Collins and Penny, 2005).

2.2 Materials and Methods

RNA extraction and 5'RACE cloning. *Encephalitozoon cuniculi* spores (Genotype II) were a generous gift from Dr. Elizabeth Didier (Tulane University, Louisiana). Glass bead-beating ruptured spores and total RNA extracted using an RNAqueous kit according to the manufacturer's (Ambion) instructions. Full-length mRNA was isolated with 5' RNA Ligase Mediated Rapid Amplification of cDNA Ends (5'RACE) according to the manufacturer's (Ambion) instructions, and using gene-specific primers. The kit-supplied 'TAP-minus' control was performed simultaneously to assess for DNA contamination, which confirmed the absence of DNA. 5' RACE products were visualized on agarose gels, and fragments were excised from gels and then purified using an Ultraclean15 MOBIO DNA purification kit (BIO/CAN Scientific, Mississauga, Ontario) according to the manufacturer's instructions. Amplicons were cloned into the pCR 2.1 vector using the TOPO TA Cloning kit (Invitrogen, Burlington, Ontario). A minimum of ten clones was

[1] A version of this chapter will be submitted for publication. Gill, E., Lee, R., Corradi, N., Limpright, V., Grisdale, C., Schmuland, R., Keeling, P., and Fast, N. (2008) Life-stage specific splicing and transcription patterns in microsporidia.

sequenced for each RACE product using ABI's Big Dye 3.1 chemistry. RACE products were assessed for splicing by comparisons to corresponding genomic sequences (*Encephalitozoon cuniculi*, Build 1.1, NCBI) using the sequence editing program Sequencher (Gene Codes, Ann Arbor, Michigan).

2.3 Results and Discussion

Of the sixteen annotated intron-containing genes, cDNA from transcripts of several were analyzed for splicing: *S26*, *L7a*, *S17*, *L37a*, *L27a*, and *Sec61α*. Because the introns are predicted to reside at the extreme 5' end of their genes, 5'RACE was performed to capture the sequence information of these regions – the first ATG codon would signify the start of the coding sequence, which would be followed by absence of the predicted introns if splicing occurs, or retention of the intron if splicing did not occur for that transcript.

I first describe the results of *S26* transcription, using it as an example for the others. The introns of *S26*, *L7a*, and *S17* were predicted to be most likely spliced, given protein expression data (Brosson et al., 2006). *L37a* transcripts were examined since *L37a* contains the longest annotated intron (Katinka et al., 2001). The intron of *L27a* was the first microsporidian intron described, so for historical reasons *L27a* transcripts were also tested. *Sec61a* is a housekeeping gene, so its transcripts were tested as well.

5'RACE revealed introns of *S26* transcripts are not spliced. Three *S26* RACE products were discernable on a 2% agarose gel (Figure 2.1), and these were cloned and sequenced. Figure 2.1 also shows the absence of similarly-sized RACE products in the 'minus-TAP control' lane, indicating the absence of DNA contamination. In fact, the gel profile of RACE products for *S26* is representative of the other genes tested with respect to its multiple bands in the experimental lane – for which we expect to see bands if transcription occurs for that gene, and for the control lanes (data not shown).

The predicted intron of *S26* is 42 nt, and is located immediately after the ATG codon. The *S26* intron is phase 0 and 42 nt, so this intron would not cause a frameshift in its transcripts. However, the intron does encode an in-frame stop codon, so it must be spliced for proper protein expression. Figure 2.2a is a schematic of the *S26* RACE products relative to its genomic locus. For *S26*, the three RACE products cloned represent three different transcripts. The two larger transcripts differ in 5'UTR lengths: one with a 611 nt UTR, and the other a 439 nt UTR. The smallest product represents a transcript initiating from within the *S26* gene, suggesting it is not a transcript that would lead to expression of the S26 protein. Sequencing of ten different clones from the two larger transcript types revealed that the 42 nt intron is retained in all, and that splicing has not removed the intron from these transcripts in the *E. cuniculi* spore.

The genomic context of *S26* is helpful in guiding interpretation that the two largest transcripts identified by RACE are indeed *S26* transcripts, and are not transcripts for adjacent genes. In fact, there are no annotated genes surrounding the *S26* locus for over 1 Kb on either the 5' or 3' side of the gene. Additional sequence examination I performed on the region surrounding the *S26* locus also did not indicate the presence of unannotated ORFs with a length greater than 100 amino acids. This is important because we know from studies that involved EST and RACE sequencing of *E. cuniculi* and *A. locustae* transcripts that a sizable proportion of transcripts contain multiple genes, that 5'UTRs can be quite large, and transcription for a target gene can start within an upstream gene – all due to loss of standard transcriptional regulatory signals caused by genome compaction (Corradi et al., 2008, Williams et al., 2003). The 5'UTR lengths observed coincide with previous reports (Corradi et al., 2008). The smallest transcript recovered may be a product for a small and unannotated downstream gene, or it can be a result of low-level spurious transcription – a sensible explanation, given its low abundance as indicated by its faint appearance (Figure 2.1).

To summarize, the *S26* transcripts are present in two forms within *E. cuniculi* spores, with the only difference between them being the length of their 5'UTRs. The genomic context for *S26* indicates the products from RACE arose from genuine *S26*

transcripts. The absence of alternative translational start sites, the conservation of S26 at the amino acid level, and ultimately, its presence in the spore proteome are all suggestive of splicing. All transcript forms examined are not spliced.

Other RPG transcripts are also not spliced in the spore. The other RPGs selected for study included *L7a* and *S17* (both of their encoded proteins were found to be present in the spore proteome (Brosson et al., 2006)), and *L37a* and *L27a.*

L7a transcripts are of a single size (Figure 2.2b). Based on the gene annotation, all have a 440 nt 5'UTR, followed by 15 protein-coding bases, and then the predicted 39 nt intron. These transcripts are very likely products of *L7a* transcription, as no other genes are present for over 5 Kb upstream and 2 Kb downstream. The intron does not introduce a frameshift, but does encode an in-frame stop. However, the level of amino acid conservation when comparing L7a sequences from a diverse set of organisms calls into question the predicted translational start, and the size of the gene: the annotated length of L7a is 193 amino acids, and the conserved portions occur after residue 50 (data not shown). A methionine occurs at residue 51, so this position can serve as an alternative translational start. If so, L7a length would be more in line with that from other organisms, and the actual 5'UTR would be longer than 440 nt. Consequently, the predicted intron may not be an intron.

Sequencing of cloned *L7a* RACE transcripts show none are spliced. It is uncertain, however, whether the apparent lack of splicing is due to control of splicing or if the gene has been mis-annotated, such that the "intron" is not actually an intron.

S17 is the remaining intron-containing RPG whose protein product was found to be present in *E. cuniculi* spores (Brosson et al., 2006). Its intron is the shortest with a predicted size of 23 nt (Katinka et al., 2001), is located immediately following the start codon, and introduces a frameshift. There are no alternative translational starts that would maintain its conserved amino acid nature. RACE showed three transcript types for *S17* (Figure 2.2c): one with a 109 nt 5'UTR, one with a 681 nt UTR, and one type

initiating within *S17*. The transcripts initiating within *S17* do so after the first 26 bp of *S17*, and likely represent transcripts for the downstream gene, located 101 bp following the *S17* stop codon. The authentic *S17* transcripts with the 109 and 681 nt UTRs transcriptionally initiate within the upstream gene of *S17* – only 49 bp separates it form *S17*. Introns are not spliced from either transcript type.

The largest predicted *E. cuniculi* intron is 52 nt (Katinka et al., 2001), and is found in *L37a*. This intron causes a frameshift and encodes stops, so its splicing is presumably necessary for proper expression. Also, there are no alternative start codons that would maintain amino acid conservation of L37a. Three transcript types were identified by 5'RACE (Figure 2.2d): one with a 42 nt 5'UTR, one with a 368 nt UTR, and one initiating after the first 195 bases of *L37a*. The transcripts that start within *L37a* are almost certainly transcripts for the downstream gene, since it resides only 100 bp downstream. The transcript type with the 368 nt UTR initiates within the upstream gene of *L37a*, which is located 345 bp away. This upstream intergenic region contains several short and unannotated ORFs, so these 368 nt UTR-containing transcripts may not be authentic *L37a* transcripts. Likely, only the 42 nt 5'UTR-containing transcripts are products of *L37a* transcription. Whatever the case, all sequenced RACE clones showed no splicing of the predicted intron.

The first microsporidian intron was predicted in *L27a* (Biderre et al., 1998), so its transcripts were also subjected to RACE, and three transcript types were recovered (Figure 2.2e). The transcriptional product of *L27a* has no 5'UTR, and is not spliced. The other two transcript types initiate within *L27a* and are likely transcriptional products of the downstream gene.

Transcripts of the housekeeping gene *Sec61α* are unspliced. The other intron-containing *E. cuniculi* genes that are not ribosomal-encoding are *Pgs1*, which was incorrectly annotated as a CDP diacylglycerol serine o-phosphatidyltransferase (Katinka et al., 2001), and *Sec61α*. BLAST results I obtained for the incorrectly annotated ORF, which was annotated to contain two introns (Katinka et al., 2001), suggested similarity to

Pgs1, the first enzyme in the cardiolipin synthesis pathway in mitochondria (Li et al., 2007). RACE showed many transcript types for *Pgs1*, which were not spliced (data not shown). However, the role of Pgs1 in *E. cuniculi* is unknown, and its splicing may not be important. On the other hand, Sec61α is a component of the Sec61 complex, which forms the core of an ER protein translocation channel (Hartmann et al., 1994) that functions in the general protein secretory pathway. The *Sec61α* intron is predicted to be 33 nt in length, and is not spliced in the two transcript types obtained: *Sec61α* has a transcript with a 24 nt 5'UTR, and another transcript initiating within, likely for the downstream gene, located just 62 bp downstream (Figure 2.2f).

Introns are not spliced from *E. cuniculi* spore transcripts. From the number of transcripts sequenced, and the number of clones sequenced per transcript, the most simple and logical explanation is that splicing does not occur in *E. cuniculi* spores. It is possible that splicing of some of the transcripts is rare, and that RACE may not have detected this low-level of splicing since spliced transcripts are not abundant, but this is unlikely. Many of the intron-containing genes are essential for viability in other organisms; for instance, according to the Saccharomyces Genome Database (www.yeastgenome.org), homozygous deletions of *S17*, *Pgs1*, or *Sec61α* result in lethality. And while deletions of *S26*, *L37a*, or *L27a* are by themselves viable in *S. cerevisiae* it is unlikely all of them are expendable.

The observation that all of these transcripts encoding ribosomal components are not spliced, and yet are present at the protein level in the spore – at least for a subset of those analyzed – leads to the conclusion that splicing of RPG transcripts is inhibited in the *E. cuniculi* spore. In addition, splicing inhibition does not appear to be gene-category specific as transcripts of non-RPGs are also not spliced. This leads me to predict that splicing in general is inhibited in the *E. cuniculi* spore. Future studies can focus on the mechanism for splicing inhibition as it likely exerts its effects at least at the level of the spliceosome, if not more globally.

Significance of splicing regulation in microsporidia. Since microsporidia alternate between the extracellular spore and the intracellular meront stages, developmentally-regulated splicing may reflect a control point for this transition. Inhibition of splicing in the spore could be a mechanism to prevent germination of spores or to maintain the dormant state. What the effects could be of repressing ribosomal protein and Sec61α expression in the spore is puzzling. Perhaps threshold titers of ribosomal components and Sec61α are needed for progression into the meront stage. Alternatively, and perhaps more likely, splicing control may have nothing directly to do with the spore-meront transition, and any potential differential-splicing may merely reflect differential cellular environments – one conducive for splicing, and another not. Perhaps, spores are more dormant ("inactive") than currently thought, and splicing, translation, and protein secretion – among all other processes, are all globally suppressed in the spore. In this case, the lack of splicing observed for intron-containing transcripts in the spore may not be as important – splicing is not specifically inhibited in the spore; rather spores are dormant and everything is inhibited, repressed, or deregulated in some way. Regardless of whether splicing is a cause or an effect of the microsporidian life-cycle, it is clear that: (1) splice inhibition prevents unnecessary upkeep in the spore, contributing to its inactive, dormant, and environmentally-resistant state, and (2) splicing is life-stage specific: it does not occur in the spore, so it occurs elsewhere, and probably in the meront stage, when replicating inside host cells.

That splicing is life-stage specific in *E. cuniculi* is encouraged from examples witnessed in *S. cerevisiae*: some genes are spliced only during meiosis (Juneau et al., 2007), and some genes, mainly RPGs, splice only when conditions are ripe for growth, and not under conditions of amino acid starvation (Pleiss et al., 2007). Both examples closely parallel the spore-meront transition – different cell types and different environments. However, these examples from yeast are correlated with a subset of genes with distinct or proposed suboptimal splicing motifs within the affected introns. The splicing inhibition observed for microsporidia seems to involve not a subset of the intron-containing genes, but rather, all of them, and it remains to be seen whether splicing

motifs are present within *E. cuniculi* introns (Chapter 3), and if so, whether distinct classes of splicing signals can be observed.

Addendum

Since this work was done, others have continued it: Valerie Limpright and Erin Gill performed RACE for the remainder of the *E. cuniculi* intron-containing RPGs, and showed their transcripts were also not spliced in the spore. Erin Gill also performed RACE on all the intron-containing transcripts from *E. cuniculi* meronts, showing they are irrefutably spliced. These results confirm my hypothesis of life-stage specific splicing. However, meront *L7a* transcripts were, like those in the spore, unspliced; the next chapter provides a molecular explanation as to why some introns are spliced and others are not.

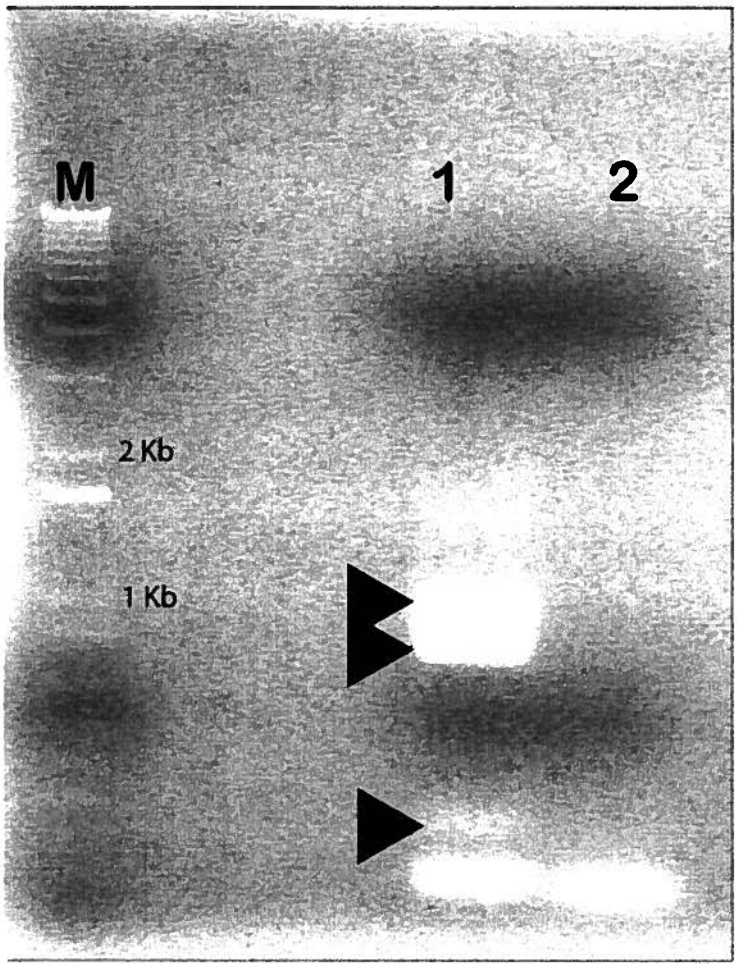

Figure 2.1. 5'RACE profile of E. cuniculi S26 transcripts. Lane 1 is the experimental, with arrows denoting RACE products. Lane 2 is the 'TAP-minus' control.

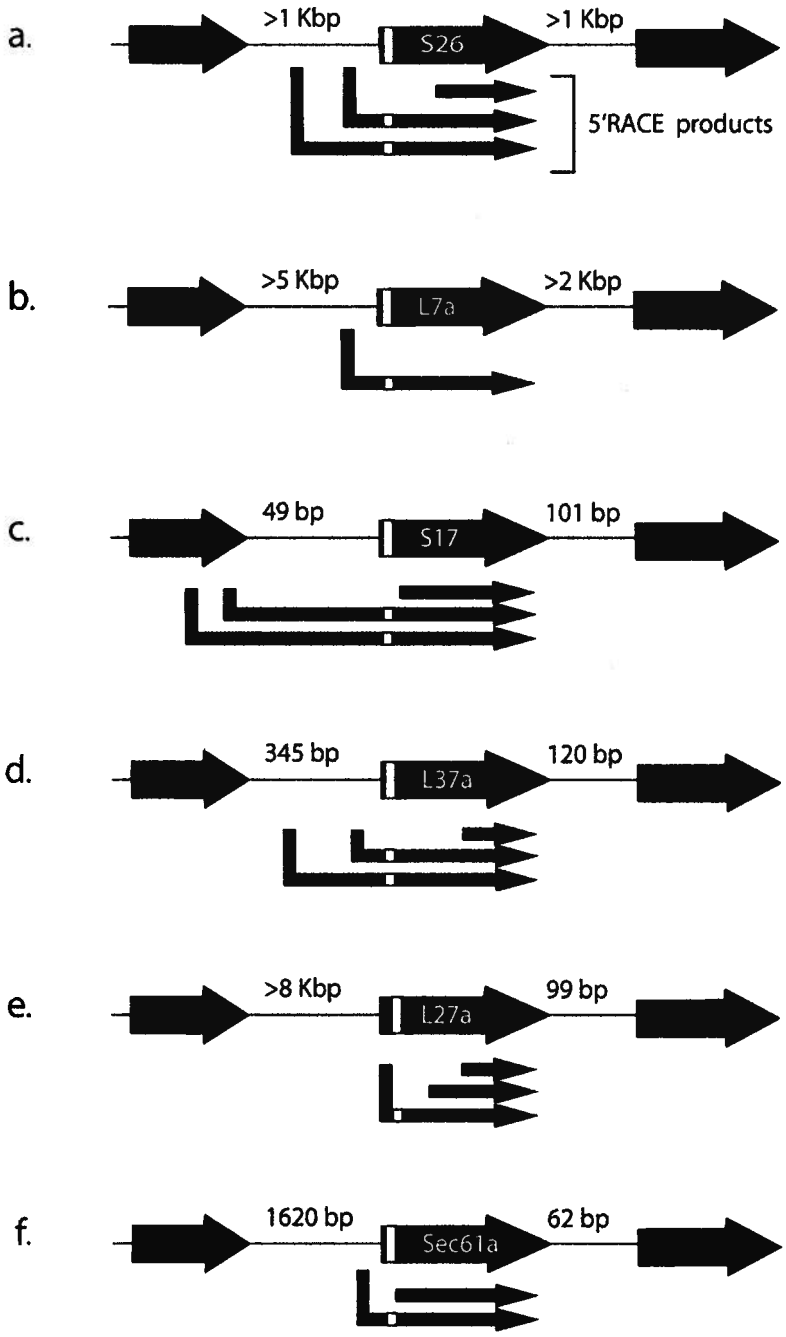

Figure 2.2. Schematic of 5'RACE profiles for the examined intron-containing E. cuniculi genes. For each of a-f, the genomic locus is shown, and corresponding transcripts are shown below it. The white bars denote introns. Numbers are intergenic distances, and is not to scale. See text for details.

References

Biderre, C., Méténier, G., and Vivarès, C.P. (1998). A small spliceosomal-type intron occurs in a ribosomal protein gene of the microsporidian *Encephalitozoon cuniculi*. *Mol. Biochem. Parasitol.* **94**: 283-286.

Brosson, D., Kuhn, L., Delbac, F., Garin, J., Vivarès, C.P., and Texier, C. (2006). Proteomic analysis of the eukaryotic parasite *Encephalitozoon cuniculi* (microsporidia): a reference map for proteins expressed in late sporogonial stages. *Proteomics*. **6**: 3625-3635.

Collins, L., and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22**: 1053-1066.

Corradi, N., Gangaeva, A., and Keeling, P.J. (2008). Comparative profiling of overlapping transcription in the compacted genomes of microsporidia *Antonospora locustae* and *Encephalitozoon cuniculi*. *Genomics*. **91**: 388-393.

Hartmann, E., Sommer, T., Prehn, S., Görlich, D., Jentsch, S., and Rapoport, T.A. (1994). Evolutionary conservation of components of the protein translocation complex. *Nature*. **367**: 654-657.

Juneau, K., Palm, C., Miranda, M., and Davis, R.W. (2007). High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proc. Natl. Acad. Sci. U S A.* **104**: 1522-1527.

Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., and Vivarès, C.P. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature.* **414**: 450-453.

Li, G., Chen, S., Thompson, M.N., and Greenberg, M.L. (2007). New insights into the regulation of cardiolipin biosynthesis in yeast: implications for Barth syndrome. *Biochim. Biophys. Acta.* **1771**: 432-441.

Pleiss, J.A., Whitworth, G.B., Bergkessel, M., and Guthrie, C. (2007). Rapid, transcript-specific changes in splicing in response to environmental stress. *Mol. Cell.* **27**: 928-937.

Williams, B.A., Slamovits, C.H., Patron, N.J., Fast, N.M., and Keeling, P.J. (2005). A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc. Natl. Acad. Sci. U S A.* **102**: 10936-10941.

CHAPTER 3: Intron recognition in *Encephalitozooon cuniculi*[1]

3.1 Introduction

Aside from the 5'-GU-AG-3' intron boundaries, the sixteen annotated *E. cuniculi* introns have never been examined for any sequence motifs necessary for their recognition by the spliceosome. The annotated sizes of the introns range from 23 to 52 bases (Katinka et al., 2001). Being so short, there are two possibilities: either the introns are enriched for motifs, or there are no motifs – or at least, no recognizable sequence motifs beyond the GU-AG boundaries. Possession of motifs increases the likelihood that the annotated introns are introns, particularly if the motifs resemble those from other organisms, and would also hint at some sort of selection for a consensus to ensure or maintain efficient splicing. The introns of the ciliate *Paramicium tetraurelia*, for example, are also extremely short (20 to 33 bases), contain some motifs, and are spliced (Russell et al., 1994). On the other hand, the absence of obvious motifs does not refute the possibility of splicing, as splicing may operate on cryptic intron sequence motifs or on none at all. As an example, the introns of the chlorarachniophyte *Bigelowiella natans* nucleomorph are the absolutely shortest known with a size range of 18 to 21 bases (Gilson et al., 2006). Comparisons between genomic and EST data, in addition, show that these short introns are spliced at discernable levels (Gilson et al., 2006). Yet even through comparisons of introns of the same length, where motifs are likely to be the most striking, there are no apparent motifs, aside from the standard GU-AG boundaries (Gilson et al., 2006). While the absence of motifs may not indicate a lack splicing for a particular intron, possession of such features is strongly suggestive of it.

The compacted nature of these microsporidian introns poses several questions. How are they recognized and spliced by the spliceosome? Have these introns retained characteristic splicing motifs, or have the splice signals degenerated to a cryptic state, or

---

[1] A version of this chapter will be submitted for publication. Lee, R., Gill, E., and Fast, N. (2008) Motif-searching and experimental validation reveal double the intron number in the microsporidian *Encephalitozoon cuniculi.*

have they been lost altogether? Knowing answers to these questions would reveal properties of splicing in *E. cuniculi*, and may shed light on the splicing mechanism more generally. Here I describe a bioinformatic assessment of what the annotated *E. cuniculi* introns look like: what motifs may be present, and if present, their variety. If motifs are indeed found to be present in the annotated introns, these data would augment and strengthen the argument for their intron status. Another goal involved using the predicted motifs to search the *E. cuniculi* genome for additional introns. Additional introns were predicted, and their splicing was validated experimentally. The set of these newly-predicted introns possess characteristics that counter current notions of what *E. cuniculi* introns look like, their numbers, the types of genes they reside in, and their positions within those genes. In addition, this new data provide an explanation for how *E. cuniculi* introns may be recognized by the spliceosome.

3.2 Materials and Methods

Intron-finding. The sequence-editing program Sequencher (Gene Codes, Ann Arbor, Michigan) was used to search for introns in the *Encephalitozoon cuniculi* genome (Build 1.1, NCBI). Briefly, the total available sequence length (2,497,413 bases distributed across eleven chromosomes) was imported into Sequencher. All variations of the 5' splice site (5'SS) motif and the branchpoint (bpA) motif were then highlighted and color-coded using the 'motif definition' option in Sequencher: 5'SS motifs were color-coded 'blue' and bpA motifs were color-coded 'pink'. Variations of the 5'SS included: GUAAGU, GUAGGU, GUAAGG, and GUGACU. Variations of the bpA motif included: CUAACUU, UUAACUU, CUAAUUU, UUAAUUU, AUAAUUU, and GUAAUUU. The intron search profile was 5'SS...bpA motif...AG-3' and the maximum intron size was arbitrarily set at 100 bases. Introns were then manually searched for, on both strands, by perusing the genome for this 'blue-pink' combination contained in a 100 base window. All previously annotated *E. cuniculi* introns were recovered using this search method.

Sequences were subjected to some criteria before they were labeled as putative introns. Some of these conditions were: removing the sequence revealed an unannotated ORF, or

removing the sequence extended an existing ORF, or removing the sequence generated an ORF that gave a higher BLAST score. Sequences were not introns if a clear ORF was present in the opposite strand, or if removing the sequence resulted in a shorter ORF, for example.

Validation of predicted and unannotated introns. *E. cuniculi* (Genotype II) meronts were a generous gift from Dr. Elizabeth Didier (Tulane University, Louisiana). RNA prepared from meronts was supplied by Erin Gill. Full-length mRNA was isolated with a 5' RNA Ligase Mediated Rapid Amplification of cDNA Ends (5'RACE) kit according to the manufacturer's (Ambion) instructions, and using sequence-specific primers. 5'RACE products were visualized on agarose gels, and fragments were excised from gels and then purified using a Ultraclean15 MOBIO DNA purification kit (BIO/CAN Scientific, Mississauga, Ontario) according to the manufacturer's instructions. Amplicons were cloned into the pCR 2.1 vector using the TOPO TA Cloning kit (Invitrogen, Burlington, Ontario). Several independent clones were sequenced using ABI's Big Dye 3.1 chemistry. RACE products were assessed for splicing by comparisons to corresponding genomic sequences using the sequence editing program Sequencher.

Miscellaneous data presentation. Sequence logos were generated using WebLogo (weblogo.berkeley.edu). Sequence alignments were generated using MacClade and subsequent editing done by eye.

3.3 Results and Discussion

***E. cuniculi* introns have the essential types of splicing motifs.** The sixteen annotated introns have the standard 5'-GU and AG-3' splice site boundaries, but unknown was the extent of additional sequence conservation into the introns, and whether motifs are present. For the prediction of the 5' splice site (5'SS) and recognition motif, the first eight annotated intron bases were used to create a sequence logo. Figure 3.1a shows the 5' SS motif for *E. cuniculi*, which is 5'-GUAAGU. This 5'SS motif is the most common amongst eukaryotes (Irimia et al., 2007), and its sequence conservation is likely maintained

for RNA-RNA interactions between the intron and the snRNA components of the spliceosome, most notably, the U1 and U6 snRNAs (Valadkhan, 2005). Similarly, for the prediction of the 3' SS motif, the last eight annotated intron bases were aligned and a sequence logo was made. Figure 3.1b shows the 3'SS motif as AG-3', and also shows that the adjacent intron bases do not contribute to the 3' SS motif. This is not surprising, as the majority of diversely-sampled eukaryotes also have a 3'SS motif that is no more than an AG-3' (Collins and Penny, 2006). This is a weak and non-specific signal for splicing when compared to the 5'SS.

Another motif present in virtually all introns is the branchpoint adenosine (bpA) motif, with the adenosine residue being a necessary participant in the splicing mechanism – particularly in the initial steps of splicing, when the intron is spliced at the 5'SS boundary and a intron lariat intermediate is formed (Jurica and Moore, 2003). A multiple sequence alignment encompassing the entire lengths of all sixteen annotated *E .cuniculi* introns was generated. From this, a bpA motif is evident (Figure 3.2). The bpA consensus motif is YUAAYUU, which resembles the eukaryotic consensus of CURAY (Spingola et al., 1999). The predicted *E. cuniculi* bpA motif, however, is longer in length than those from other organisms, with the exception of the *S. cerevisiae* bpA consensus of UACUAAC (Spingola et al., 1999), which happens to be the same size as the *E. cuniculi* bpA motif (Figure 3.2). When compared with introns in well-studied animal and plant splicing models, introns in *S. cerevisiae* contain fewer numbers of motifs, possibly reflecting fewer intron-spliceosome interactions (Collins and Penny, 2006). It has been argued that having a more stringent bpA motif compensates for the lack of additional sequence signals (Spingola et al., 1999), and this may be true for *E. cuniculi* as well.

Another common intron motif is the poly-pyrimidine (polyY) tract, which is prevalent in animal and plant introns, and in most yeast introns (Spingola et al., 1999). Almost exclusively positioned 3' to the bpA motif, they can also be found upstream of the bpA motif, as seen in *S. cerevisiae* introns (Spingola et al., 1999). From the *E. cuniculi* intron alignment, a polyY tract is not discernable in all the introns (Figure 3.2). However, a small stretch (three bases) of uridines are present in some of the introns that are 31 or 32

bases in length – in the introns of *L19*, *L37*, *s8*, and *L39* (Figure 3.2). The significance of the triple-U tract in these introns is unknown, and it is debatable if they are bona fide or even remnant polyY tracts, given the lack of obvious polyY tract-binding protein homologues listed in the *E. cuniculi* genome annotation (Katinka et al., 2001). On the other hand, the last two uridine bases of the predicted bpA motif may be the polyY tract. The significance of these last two uridines is a mystery, as having them in the *E. cuniculi* bpA motif makes the *E. cuniculi* bpA motif the most 3' extended compared to all other eukaryotes. Another possibility could be that the extra two uridines participate in pairing with the U2 snRNA, but this has been discounted given the sequence of the U2 snRNA (data not shown).

Despite being so compacted in length, *E. cuniculi* introns have retained all three dominant splicing signals: the 5'SS, the bpA motif, and the 3'SS. This indicates that even with extraordinary pressures to reduce the genome, *E. cuniculi* has maintained the most essential elements of intron splicing, and that the splicing mechanism in this organism likely resembles the core of most splicing pathways. The lack of an obvious polyY tract, however, is not detrimental to this view, as short introns having the dominant signals have also been shown to lack polyY tracts (i.e. Paramecium). Moreover, increasing amounts of data suggests the polyY tract as being necessary only when considering longer introns (Coolidge et al., 1997), to aid the spliceosome in selecting the correct 3'SS, a general problem in splicing.

The *E. cuniculi* intron splicing motifs were used to predict intron status and to formulate an intron model. Based on the 5'SS of 5'-GUAAGU and the bpA motif of YUAAYUU, two of the sixteen annotated introns, found in *L7a* and *s29*, were predicted not to be introns, as both lack a 5'SS and bpA motif resembling either consensus (Figure 3.2). In addition, unlike the other fourteen annotated introns, those of *L7a* and *s29* do not induce frameshifts or encode stops. 5'RACE was performed on meront material for the sixteen annotated introns, and while splicing removed the majority of the sixteen introns, splicing did not occur for the *L7a* and *s29* annotated introns (Erin Gill, personal

communication), indicating these are not introns and highlighting the importance of both motifs when defining *E. cuniculi* introns.

An initial 5'RACE result showed that one of the annotated introns was actually shorter than annotated: s26 had an annotated intron length of 42 bases, but 5'RACE showed it is only 33 bases (Erin Gill, personal communication), a difference of a few codons in length. Using this information, and considering the intron alignment – the majority of the annotated introns have only a few bases separating the bpA motif and the 3'SS (Figure 3.2), a second set of predictions was generated. Based on the predicted bpA motif and its downstream sequence content, several additional introns were predicted to be shorter: *L5* intron is 26 and not 38 bases, *pgs1*#1 intron is 25 and not 43 bases, *s24* intron is 29 and not 44 bases, and *L37a* intron is 49 and not 52 bases. Subsequent meront 5'RACE experiments confirmed my predictions (Erin Gill, personal communication).

The two sets of confirmed predictions indicate two necessary conditions: first, an *E. cuniculi* sequence is an intron only if it has motifs similar to the 5'SS and the bpA motif, and second, *E. cuniculi* introns have a bpA motif-3'SS distance with a definable length range. Knowing this, I wondered if there were more than just fourteen (sixteen minus *L7a* minus *s29*) introns in the *E. cuniculi* genome.

Using the *E. cuniculi* intron model to search for additional and unannotated introns. With a defined intron model on hand, it became possible to find unannotated introns more effectively. If the original annotation of *E. cuniculi* introns was conservative, which it likely was (introns were not specifically searched for), then finding additional introns was a strong possibility. In addition, we know from the *E. cuniculi* genome annotation that roughly 60% of genes have homologs, and that 40% are either hypothetical, highly-divergent, or microsporidian-specific genes (Katinka et al., 2001). With this gene-category ratio, and considering that all annotated introns fall into the category of genes with identifiable homologs, it was expected that roughly ten unannotated introns in the hypothetical gene-category could be found. However, due to the extreme compaction of intergenic regions in the *E. cuniculi* genome, it was also possible that additional introns are

not present. To distinguish between these possibilities, introns were searched for (see Materials and Methods).

The genome of *E. cuniculi* has retained at least twice the number of introns than previously thought. Fifteen additional and unannotated introns were predicted (Table 3.1), more than doubling the current *E. cuniculi* intron number. Of the fifteen newly-predicted introns, four are in RPGs (including one in a 5'UTR), one intron is in a gene encoding a polyadenylate-binding protein, one intron is in a gene encoding an ubiquitin-activating protein, two introns are in unannotated genes encoding proteins of conserved but unknown functions, three introns are in annotated *E. cuniculi* hypothetical genes, and four introns are in unannotated *E. cuniculi* hypothetical genes (Table 3.1). Previously, of the fourteen annotated and verified introns, eleven introns were in RPGs, and three introns were in two non-RPGs with putative functions. This preponderance of introns in RPGs is also seen in the highly-reduced and compacted nucleomorph genome of *Guilliardia theta*, which has also retained only a few – seventeen – introns (Douglas et al., 2001). Considering these newly-predicted *E. cuniculi* introns, the gene-category of intron distribution now shifts substantially, with approximately half of the introns being retained in RPGs. This pattern resembles what is seen in the yeast *S. cerevisiae*, where of its several hundred introns, over a third are in RPGs (Spingola et al., 1999). That low intron density in organisms with hyper-compacted genomes is correlated with an almost *exclusive* preponderance of introns in RPGs is therefore no longer true, assuming that the intron annotation in the *G. theta* nucleomorph was also exhaustive.

The size range for the originally annotated and verified fourteen *E. cuniculi* introns is 23 to 49 bases. The addition of the newly-predicted introns does not change the lower size limit, but it does change the upper limit: the largest predicted intron is 76 bases, more than half in length larger than the next largest intron (49 bases of the annotated *L37a*). This 76 base intron is in a gene encoding for a polyA-binding protein, and is of interest precisely because it is such an extreme size outlier relative to the other intron lengths (Figure 3.3). One wonders whether this 76 base intron contains unique splicing signals that the other *E. cuniculi* introns have lost, which may have constrained its compaction, or whether its larger

size is a result of a more recent insertion, that arose subsequent to genome reduction. Considering that almost all the variability in *E. cuniculi* intron length is due to length variation between the 5'SS and the bpA motif (Figure 3.3), either scenario is possible. An insertion could have occurred in the region between the 5'SS to bpA motif, but this would seem at odds with a genome experiencing extreme pressures to reduce and compact in size. However, given what is now known of microsporidian genome size evolution – that genomes are not static and may even be re-expanding following reduction (Williams et al., 2008) – this idea cannot be overlooked. If on the other hand, this 76 base intron has retained unique splice signals, this would hint at an added layer of complexity in microsporidian splicing, beyond what can be envisioned as an "essential core".

Experimental verification of intron predictions demonstrate these intron-containing genes are expressed and that their introns are spliced. The skew of the majority of the newly-predicted introns towards the lower end of the observed intron length spectrum (Figure 3.3) suggested that, even in the absence of experimental validation, most are genuine introns. This interpretation is based on the statistical argument that for the motif signals to occur together within the specified search window is rare; it is even rarer for them to occur in a decreasing size window. Surprisingly, all of the intron-containing genes are expressed, as indicated by 5'RACE bands on agarose gels (data not shown), and splicing has thus far been confirmed for thirteen of the fifteen newly-predicted introns, including the 76 base longest intron and the intron in the 5'UTR of the RPG *s10*. Of the fifteen newly-predicted introns, thirteen splice at boundaries exactly as predicted. Two of the predicted introns have yet to be confirmed via splicing: one intron is the second intron (denoted 'Intron 12' in Table 3.1 and Figure 3.3) in the annotated hypothetical gene of Ecu08_1030 (the first predicted intron is spliced) – this intron is too far away from the 5'end of the gene to be recovered by 5'RACE so 3'RACE will later be attempted to show its splicing.

The other intron not yet verified is in an unannotated gene encoding a highly-conserved protein of unknown function (denoted as 'Intron 7' in Table 3.1 and Figure 3.3). This predicted intron looks like an intron in every way (Figure 3.3), and its encoded

product can only be fully expressed if splicing occurred. The fact that its splicing is not demonstrated indicates that either it splices at a lower level compared to all other *E. cuniculi* introns, or its splicing is actively inhibited at the level of the gene, post-transcriptionally – which would be the first such case in microsporidia, and is something that will be further explored. That it might splice at a very low level, which is undetectable by RACE, can be confirmed by other methods, such as capillary electrophoresis.

The predicted and verified *E. cuniculi* introns exhibit an unusual positional distribution. It is generally accepted that for organisms with reduced genomes and few introns, introns are positionally-biased to the extreme 5' ends of the genes in which they reside. This correlation is evident in the reduced *G. theta* nucleomorph genome and also in *S. cerevisiae* (Douglas et al., 2001; Spingola et al., 1999). To account for this positional skew, the mechanism of reverse transcriptase-mediated intron-loss is usually invoked. Briefly, during this process, the spliced mRNA is reverse-transcribed into cDNA, and subsequent recombination with the genomic locus would result in a gene missing an intron (Roy and Gilbert, 2006). This mechanism for intron loss is preferred over genomic deletion as it results in complete removal of the intron, as opposed to likely only partial intron loss via genomic intron deletion, which more often than not, is deleterious. Further, since RT initiates from the 3' ends of transcripts and is of low processivity, introns closer to the 3' ends of genes are lost at a greater rate – or, introns closest to the start of genes are the most difficult to lose (Roy and Gilbert, 2006). If this is true, then we can infer that the few introns left in these genomes likely have no functional basis that hinders their removal, and that they could be "on their way out". Alternatively, the 5' positional bias may reflect some sort of function, perhaps some role associated with gene expression processes such as transcription, capping, or export (Bentley, 2005). This hypothesis is popular since almost all the introns in the *G. theta* nucleomorph, and a significant proportion of the few introns in *S. cerevisiae*, happen to be retained at the 5' ends of RPGs (Douglas et al., 2001; Spingola et al., 1999).

The originally-annotated *E. cuniculi* introns also show this distribution (Figure 3.4, top). And with the exception of two introns, which both reside in a non-RPG, the introns

are situated at the most 5', very often after the first codon. However, inclusion of the fifteen newly-predicted introns suggests that not all introns are preferentially biased to the start of genes, and that introns are generally more dispersed within their genes (Figure 3.4, bottom).

It is obvious now that only RPG introns show this 5' bias – of the fifteen predicted introns, four are in RPGs, and three of the RPG introns are at the 5' ends. The other RPG intron is in the 5'UTR of its transcripts, also close to the 5' translational start (Figure 3.4). That the *E. cuniculi* genome has retained an intron in a 5'UTR is puzzling with respect to microsporidian genome evolution, but its occurrence may illuminate mechanistic aspects of intron loss and retention. UTR introns do not, unlike introns within protein-encoding regions, have to be spliced for gene expression. Also, in contrast with introns in protein-coding regions, UTR introns are relatively immune to the deleterious effects of mutation on introns via incomplete mutational removal – through genomic deletion, for instance – so the fact that it is present in *E. cuniculi*, which has ratcheted down its genome, is perplexing. The evolutionary significance of UTR introns is unknown (Roy et al., 2007), but studies have suggested that some UTR introns play a regulatory role (Hong et al., 2006). Considering this information, and that a UTR intron is present in *E. cuniculi* despite the apparent ease at which it can be removed relative to non-UTR introns, it could indicate that this intron likely imparts a functional role. Moreover, the fact that this UTR intron is in a RPG strengthens the argument that the other RPG introns may also be regulatory, and that their retention is a result of positive selection.

Unlike the RPG introns, the non-RPG introns do not show a positional bias and are dispersed across their gene lengths (Figure 3.4, bottom). Whether these introns may also be functional is unknown. Recall that one of the fifteen newly-predicted introns has not yet been shown to splice, possibly reflecting active splicing inhibition of that intron. If they are not selected for, and if RT-mediated intron loss was the major mechanism that lead to the purging of most *E. cuniculi* introns, then it is expected that introns would never be observed to be as diffuse across their genes, as they are here. However, if intron loss in *E. cuniculi* has been predominantly RT-mediated, then perhaps it is a slower process then

currently thought. Alternatively, RT-activity may have been lost and all retained introns are "stuck".

Another possibility is that intron loss may never have been RT-mediated in *E. cuniculi*. Given the small size of RPGs (personal observation), compared to other genes, it is difficult to imagine that intron removal at their transcripts' 5'ends would be prohibited by this mechanism. That intron loss may not be RT-mediated also explains the diffuse distribution of introns in non-RPGs.

3' Splice site selection is constrained by *E. cuniculi* intron structure. A general splicing problem is the accurate recognition of the correct AG-3' splice site (3'SS) in the midst of multiple possible splice sites (Luukkonen and Seraphin, 1997). 3' splice site selection in well-studied splicing models indicates roles for additional sequence signals – motifs that are not present in *E. cuniculi*. For example in *S. cerevisiae*, the polypyrimidine tract located after the branchpoint aids in selection of the correct 3' splice site (Chen et al., 2000), as does the binding of many auxillary spliceosomal proteins to this intron region (McPheeters and Muhlenkamp, 2003). In multicellular organisms, splice site selection is regulated by signals within adjacent exons (Collins and Penny, 2006) – a presumably logical way to cope with the long intron length in these organisms. How the correct 3'SS of the short introns of *E. cuniculi* (which lack these additional splicing motifs) are recognized, was uncovered in the present study.

The sum of all obtained *E. cuniculi* 5'RACE sequence data show only one splice product per intron-containing gene, suggesting intron recognition is highly accurate. How the spliceosome is able to operate with such fidelity is best illustrated by what it does not do: a handful of the *E. cuniculi* introns have an adjacent AG dinucleotide immediately following the observed 3' splice site (*L5* intron, *s24* intron, intron 9, and intron 2 from Figure 3.5), yet splicing at these possible cryptic sites never occurs. The alignment of all the *E. cuniculi* introns (Figure 3.3) offers a clue to this problem. It shows that the spacing between the bpA motif to the splice site ranges from zero to three bases, suggesting that this spacing of up to a few bases may be the factor in determining accurate 3' splice site

recognition. Further, that this threshold of up to three bases results in accurate 3'SS recognition and splicing explains why the introns with an adjacent 3'SS dinucleotide are never observed to be mis-spliced: the distance from the bpA motif to the adjacent dinucleotide are up to four bases, exceeding the proposed threshold of three bases (Figure 3.5).

Additional support for this claim comes from considering sequences that look like introns, in that they contain the necessary *E. cuniculi* intron motifs, but that we know are not spliced. For example, sequences in *Prp8* and in Ecu05_1440 contain the motifs, yet are not truly introns. *Prp8* encodes the sixth largest protein in *E. cuniculi* (personal observation), which is highly-conserved both size and in sequence composition. If splicing were to occur for this intron-like sequence, this protein would be rendered nonfunctional. Examining the intron-like sequence of *Prp8* (Figure 3.5) shows that mis-splicing is prevented by the spacing between the bpA motif and the 3'SS: the spacing between these motifs is four bases, larger than the threshold for *E. cuniculi* 3'SS selection and splicing; therefore, the spliceosome does not recognize it. The same logic applies to the intron-like sequence within Ecu05_1440 (Figure 3.5). All things considered, possession of motifs is important for accurate splicing, but their spacing is equally important: as a way to prevent mis-splicing.

Although it seems clear that the bpA-3'SS directs correct splicing, the possibility that mis-spliced products are shuttled out of the nucleus very quickly and then degraded must be considered. Nonsense-mediated mRNA decay (NMD) is the pathway by which mis-spliced mRNAs are processed (Jaillon et al., 2008). The apparent absence of NMD components in *E. cuniculi* (Lynch, 2006) decreases the likelihood of this scenario. However the possibility of another degradation pathway cannot formally be ruled out. Coupling the lack of a recognized degradation pathway in *E. cuniculi* with the fact that no mis-spliced products were recovered supports the idea that *E. cuniculi* splicing is highly-accurate and specific.

Therefore, accurate splicing appears to be constrained by the bases following the bpA motif, but what limits this dimension of intron length to just three nucleotides? Since the distance from the bpA motif to the 3'SS never exceeds three nucleotides, it strongly suggests that the *E. cuniculi* spliceosome can distinguish this difference. We can hypothesize that recognition is constrained by the physical interaction between the spliceosome and the intron, so that a bpA-3'SS distance of four or more bases is sterically excluded. Spliceosomal protein factors are numerous even in systems where splicing is regarded to be relatively less complex, as in *S. cerevisiae*, but knowledge of all the spliceosomal components and their interactions with one another and with intron elements remains incomplete (McPheeters and Muhlenkamp, 2003). Further dissection of this sub-system of splicing in *E. cuniculi*, therefore, offers an opportunity to understand it completely. This is the first case where limiting the spacing between the bpA motif and the 3'SS is required for accurate splicing. There has never been another example where intron length has an upper bound, and it is easy to visualize why this might be so: introns can form secondary structures (Li et al., 1995), and looping of the content between the bpA motif and 3'SS is one way to cope with large introns. Perhaps without a recognizable way to deal with mis-spliced transcripts, 3'SS recognition is more tightly controlled, concomitant with splicing reduction.

The lower limit spacing between the bpA motif and the 3'SS is zero bases (Figure 3.3). However, the possession of the three bases (YUU) immediately following the branchpoint adenosine also appears necessary and could prevent further length reduction. A minimal length from the bpA motif to 3'SS is seen in other systems. In *S. cerevisiae*, there is a minimum (but unknown) 3' intron end length, as studies show that the first and closest possible 3'SS are not always used (Luukkonen and Seraphin, 1997). In other cases a single motif may be responsible for both bpA and 3'SS recognition. In both *Trichomonas vaginalis* and *Giardia intestinalis*, there is exact sequence and size conservation at the 3' ends of their introns (five bases bounded by the branchpoint adenosine and the 3'SS), and their bpA motifs and 3'SS appear fused, suggesting simultaneous motif recognition (Vanacova et al., 2005). Such lower limits for motif separation could be viewed as stretching a piece of string (where the bpA motif is at one end of the string and the 3'SS is

at the other). The length of the string is restricted by the "length" of bound spliceosomal protein(s). The *E. cuniculi* triplet is part of the bpA motif (Figure 3.3), and its sequence conservation suggests that it participates in how the *E. cuniculi* introns are recognized. But whether it is solely involved in recognition of the branchpoint adenosine base or also involved in 3'SS recognition is unknown. There are likely several points for recognition beyond recognition of the motifs, such as the variable spacing between them.

Lastly, consideration of one of the newly-predicted (but not yet verified) introns suggest that not only is 3'SS selection in *E. cuniculi* based solely on the distance following the bpA motif, which may be an unique case if additional sequence motifs are not involved, but also the possibility that 3'SS recognition is based on a scanning mechanism following bpA motif recognition. Examination of this sequence shows the distance from the bpA motif to its observed 3'SS is just a single base (intron 12 of Figure 3.5). That it also has a possible AG-3' dinucleotide directly adjacent to the intron is telling for an additional reason: the distance from the bpA motif to this adjacent AG dinucleotide is three bases, the threshold for splice site selection. If splicing is observed for the true 3'SS, it would suggest the first possible 3'SS is selected, and probably through a scanning mechanism subsequent to bpA motif recognition.

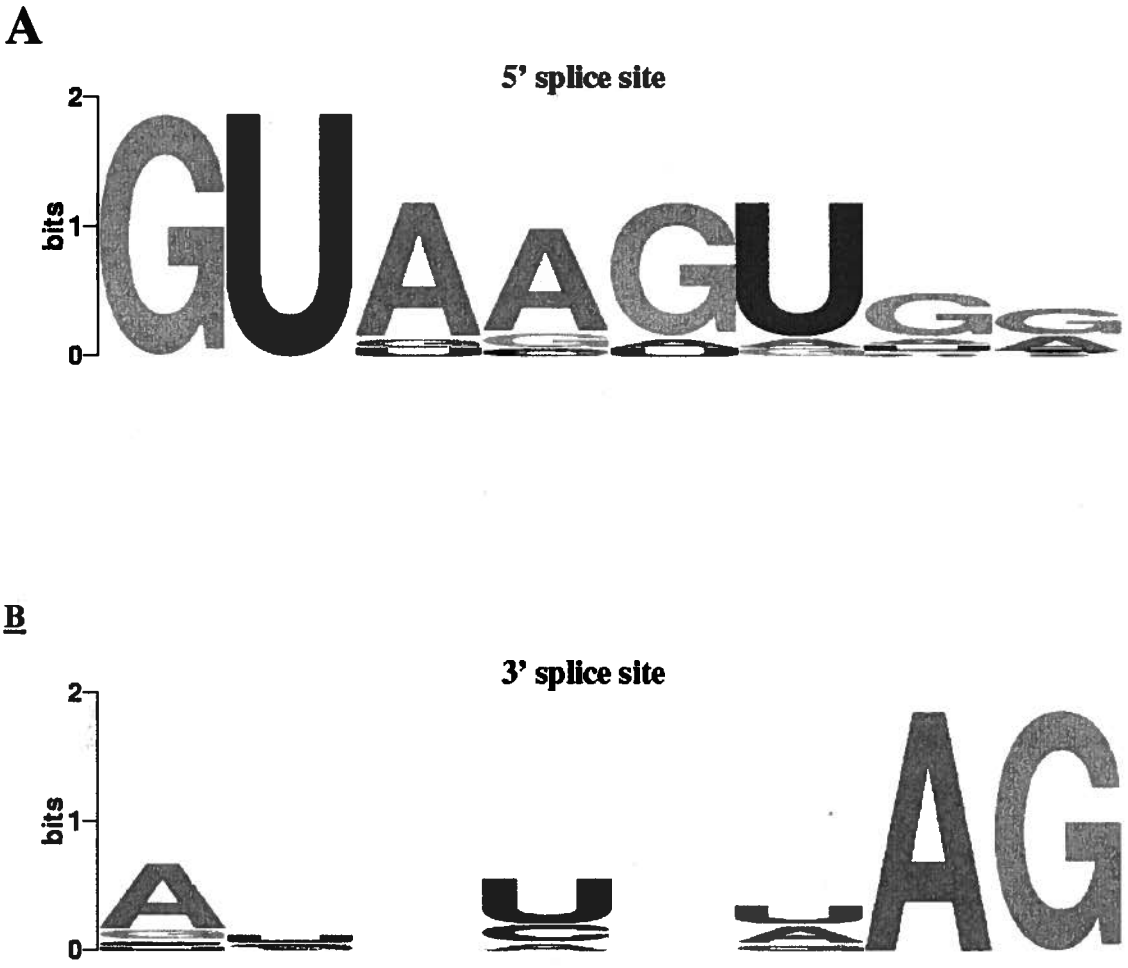

Figure 3.1. Splice site motifs illustrated as sequence logos (generated using WebLogo: http://weblogo.berkeley.edu). Intron bases from all previously annotated *E. cuniculi* introns were used.

```
s17      GUAGGUGG-------------------------GAGCU-UUAAUUU-U-------------------AG  23 nt
pgs#2    GUAAGUAG-----------------------ACAACA-CUAAUUU-CU-----------------AG  25 nt
L27a     GUAAGUGG-------------------UUGCUUGAG-CUAAUUU-GA-----------------AG  28 nt
L19      GUAAGUGA-----------------GACGUGCUGUUU-CUAAUUU-GA-----------------AG  31 nt
L37      GUAAGUCG-----------------GCUUUAGACUCA-AUAAUUU-GU-----------------AG  31 nt
s8       GUAAGUGG-----------------GCUUUGAAGAAGA-CUAAUUU-U-----------------AG  31 nt
L39      GUAAGGUG-----------------AUUUGGCUCCAGG-CUAACUU-CU-----------------AG  32 nt
s29      GUUCUUUU--------CGACGUACCAGUGAAGAAGAAUG----------------------------AG  33 nt
Sec61a GUAAGUGA---------------AGAUAAACAUUGAAA-CUAAUUU-U-----------------AG  33 nt
L5       GUAAGUGG----------------------UAGCGAC-GUAAUUU-UAAGAGCAAUGCAA------AG  38 nt
L7a      GUGAAAAA--GAUUCCGGUUGAGGAACCUAGGCUGACAA----------------------------AG  39 nt
s26      GUAAGUGG----------------UAGGCAAAGUGCU-CUAAUUU-UGUAGACUGUUA--------AG  42 nt
pgs#1    GUAAGUGA----------------------AGCAUU-UUAAUUU-CAAGAUAUCUUGAGGUCCAGAG  43 nt
s24      GUAAGUGC-------------------GAAGACGAAG-CUAACUU-GUAGAGUUUGGAACUGG---AG  44 nt
s30      GUAGGUGG---AAUCUUGCAGUCUCUCGGAGACCUGUU-CUAACUU-U-----------------AG  45 nt
L37a     GUAAGUGA-CUUGGGAUUGUAGUAGCUCCUGGAGCAGA-CUAACUU-UGUAGGGAAA----------AG  52 nt
E.cuniculi branchpoint consensus              YUAAYUU
S.cerevisiae branchpoint                     UACUAAC
eukaryotic branchpoint consensus              CURAY
                                                *
```

Figure 3.2. Alignment of all sixteen previoously *E. cuniculi* introns optimised to highlight the branchpoint motif. The asterisk denotes the branchpoint adenosine involved in the first major step of the splicing reaction.

Intron	Genome coordinates	Intron size (nt)	Gene
1	Chromosome 2, minus strand, 123178-123201	24	Ribosomal L11
2	Chromosome 4, minus strand, 44622-44658	37	5'UTR of ribosomal s10
3	Chromosome 5, plus strand, 36830-36865	36	Ribosomal s3a
4	Chromosome 5, plus strand, 95407-95437	31	Ribosomal s6
5	Chromosome 9, minus strand, 83634-83709	76	Ecu09_1470 polyA-binding
6	Chromosome 9, minus strand, 144043-144067	25	Ecu09_0860, Ubq-activator
7	Chromosome 4, plus strand, 61467-61489	23	Unannotated, conserved
8	Chromosome 4, plus strand, 173862-173885	24	Unannotated, hypothetical
9	Chromosome 6, minus strand, 61797-61826	30	Unannotated, hypothetical
10	Chromosome 7, plus strand, 48247-48270	24	Unannotated, conserved
11	Chromosome 8, plus strand, 118293-118316	24 (1^{st})	Ecu08_1030, hypothetical
12	Chromosome 8, plus strand, 118609-118631	23 (2^{nd})	Ecu08_1030, hypothetical
13	Chromosome 8, plus strand, 140176-140198	23	Unannotated, hypothetical
14	Chromosome 8, minus strand, 130005-130028	24	Unannotated, hypothetical
15	Chromosome 11, plus strand, 134220-134242	23	Ecu11_1060, hypothetical

Table 3.1. Listed are all fifteen, unannotated introns, all predicted in this study. Six introns were found in unannotated genes. Two introns are predicted in one gene. Genome coordinates refer to NCBI's *E. cuniculi* genome (Build 1.1).

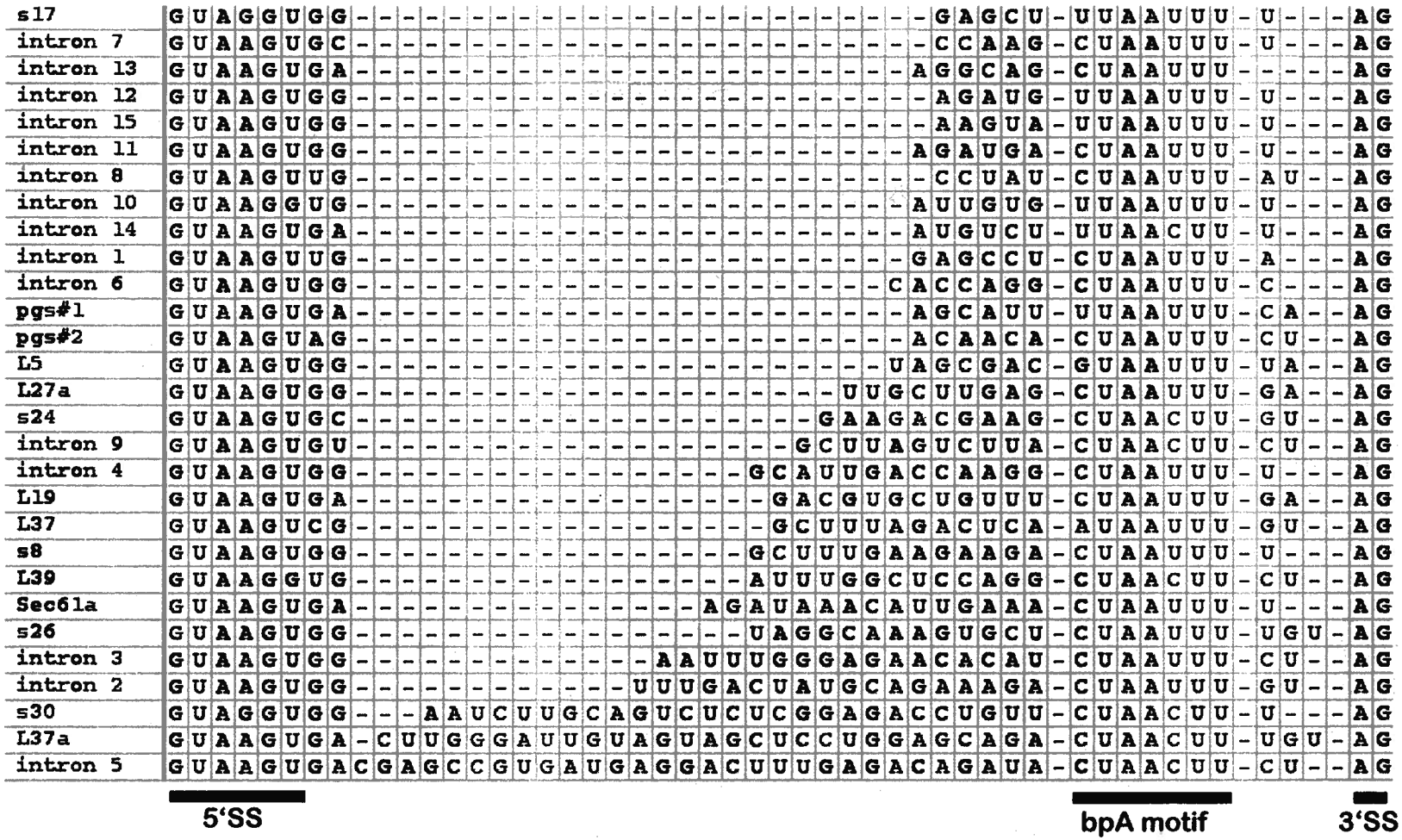

Figure 3.3. An alignment of all known E.cuniculi introns, consisting of previously annotated introns which have been verified experimentally, with revised boundaries and also all newly-predicted and unannotated introns from this study. The alignment was drawn in such a way as to highlight the predicted splice motifs. The newly-predicted introns are labelled from 1-15 (refer to Table 1 for gene correspondance and other info). The shortest intron is at the top. The longest intron is at the bottom (this 76 base intron was truncated to make the alignment).

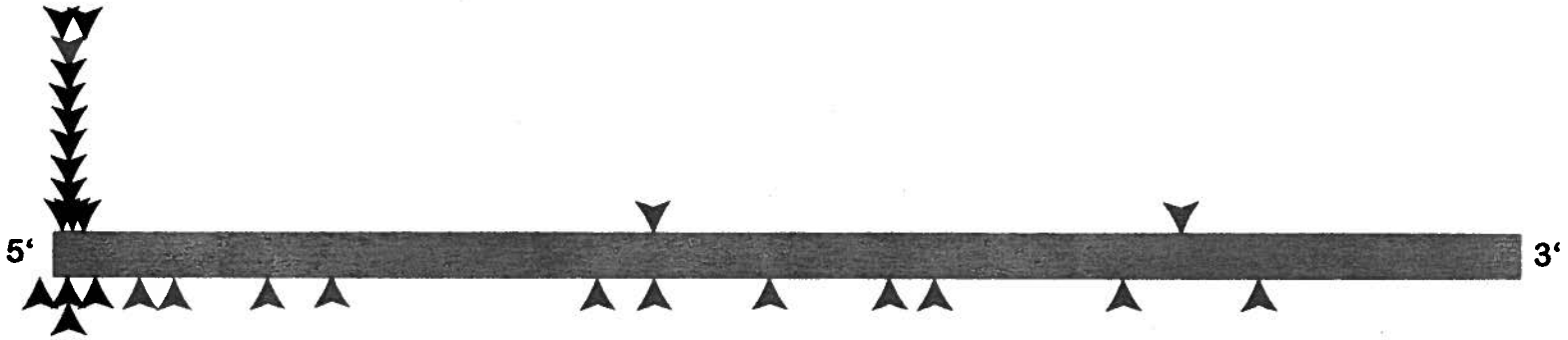

Figure 3.4. New introns give an unexpected intron positional distribution for an intron-sparse reduced genome. Arrows point to intron positions. Black arrows show RPG-introns, grey arrows show non-RPG intron positions. Previously annotated introns that were verified in this study are shown at the top. Introns predicted from this study are shown at the bottom.

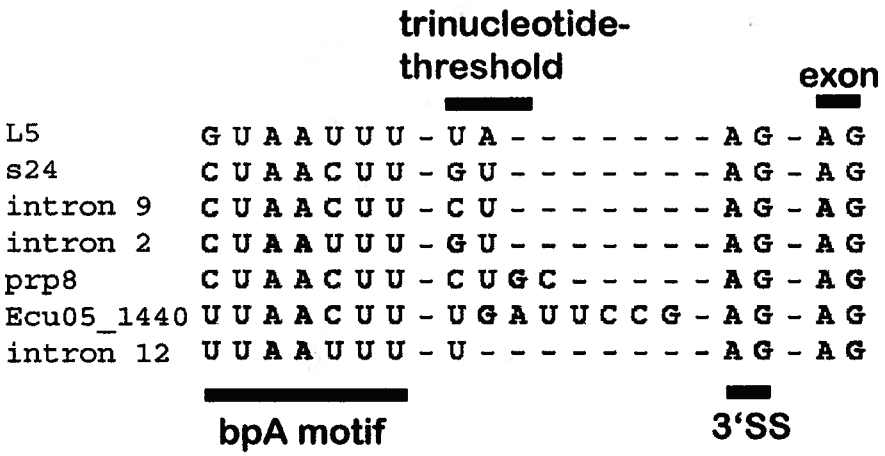

Figure 3.5. Schematic showing how the distance immediately following the bpA motif constrains 3'SS selection to allow for accurate splicing and the prevention of mis-splicing. To highlight the features of this model, the intron bases upstream of the bpA motif are omitted. The nucleotide content between the bpA motif and the 3'SS has a size threshold of three bases, preventing mis-splicing at the adjacent AG exonic dinucleotide for the introns shown. See text for further details.

References

Bentley, D.L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell. Biol.* **17**: 251-256.

Chen, S., Anderson, K., and Moore, M.J. (2000). Evidence for a linear search in bimolecular 3' splice site AG selection. *Proc. Natl. Acad. Sci. U S A.* **97**: 593-598.

Collins, L., and Penny, D. (2006). SMBE Tri-National Young Investigators. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Investigating the intron recognition mechanism in eukaryotes. *Mol. Biol. Evol.* **23**: 901-910.

Coolidge, C.J., Seely, R.J., and Patton, J.G. (1997). Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.* **25**: 888-896.

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.T., Wu, X., Reith, M., Cavalier-Smith, T., and Maier, U.G. (2001). The highly reduced genome of an enslaved algal nucleus. *Nature*. **410**: 1091-1096.

Gilson, P.R., Su, V., Slamovits, C.H., Reith, M.E., Keeling, P.J., and McFadden, G.I. (2006). Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc. Natl. Acad. Sci. U S A.* **103**: 9566-9571.

Hong, X., Scofield, D.G., and Lynch, M. (2006). Intron size, abundance, and distribution within untranslated regions of genes. *Mol. Biol. Evol.* **23**: 2392-2404.

Irimia, M., Penny, D., and Roy, S.W. (2007). Coevolution of genomic intron number and splice sites. *Trends Genet.* **23**: 321-325.

Jaillon, O., Bouhouche, K., Gout, J.F., Aury, J.M., Noel, B., Saudemont, B., Nowacki, M., Serrano, V., Porcel, B.M., Ségurens, B., Le Mouël, A., Lepère, G., Schächter, V., Bétermier, M., Cohen, J., Wincker, P., Sperling, L., Duret, L., and Meyer, E. (2008). Translational control of intron splicing in eukaryotes. *Nature.* **451**: 359-362.

Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell.* **12**: 5-14.

Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., and Vivarès, C.P. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature.* **414**: 450-453.

McPheeters, D.S., and Muhlenkamp, P. (2003). Spatial organization of protein-RNA interactions in the branch site-3' splice site region during pre-mRNA splicing in yeast. *Mol. Cell. Biol.* **23**: 4174-4186.

Li, Z., Paulovich, A.G., and Woolford, J.L. Jr. (1995). Feedback inhibition of the yeast ribosomal protein gene *CRY2* is mediated by the nucleotide sequence and secondary structure of *CRY2* pre-mRNA. *Mol. Cell. Biol.* **15**: 6454-6464.

Luukkonen, B.G., and Séraphin, B. (1997). The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in *Saccharomyces cerevisiae*. *EMBO J.* **16**: 779-792.

Lynch, M (2006). The origins of eukaryotic gene structure. *Mol. Biol. Evol.* **23**: 450-468.

Roy, S.W., and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.* **7**: 211-221.

Roy, S.W., Penny, D., and Neafsey, D.E. (2007). Evolutionary conservation of UTR intron boundaries in Cryptococcus. *Mol. Biol. Evol.* **24**: 1140-1148.

Russell, C.B., Fraga, D., Hinrichsen, and R.D. (1994). Extremely short 20-33 nucleotide introns are the standard length in Paramecium tetraurelia. *Nucleic Acids Res.* **22**: 1221-1225.

Spingola, M., Grate, L., Haussler, D., and Ares, M. Jr. (1999). Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA*. **5**: 221-234.

Valadkhan, S. (2005). snRNAs as the catalysts of pre-mRNA splicing. *Curr. Opin. Chem. Biol.* **9**: 603-608.

Vanácová, S., Yan, W., Carlton, J.M., and Johnson, P.J. (2005). Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. *Proc. Natl. Acad. Sci. U S A.* **102**: 4430-4435.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*. **14**: 802-813.

Williams, B.A., Lee, R.C., Becnel, J.J., Weiss, L.M., Fast, N.M., and Keeling, P.J. (2008). Genome sequence surveys of *Brachiola algerae* and *Edhazardia aedis* reveal microsporidia with low gene densities. *BMC Genomics*. **9**: 200.

Chapter 4: Conclusions and proposed research possibilities

In this thesis I have examined issues relating to intron evolution in a microsporidian with a highly reduced genome. Splicing is inhibited in spores and is life-stage specific in the microsporidian *Encephalitozoon cuniculi*. This aspect of splicing regulation is proposed to reflect the "dormant" cellular conditions within the spore. Few introns are retained in the *E. cuniculi* genome, and this retention is carried to the transcriptome through splicing inhibition.

The other major conclusions, as described in Chapter 3, concern *E. cuniculi* intron recognition. There is at least twice the current inventory of introns, and the introns all have the hallmark splicing motif signals, suggesting that general splicing mechanisms are conserved in this organism. My predictions and validation of these additional introns also give unexpected gene-category and gene-positional intron distributions (Katinka et al., 2001); these data have implications for intron evolution within Microsporidia. Most important, splicing specificity is observed for all *E. cuniculi* introns, and this is achieved by maintenance of a finite sequence length between splice signals – thus solving a general splicing problem for *E. cuniculi*, via a unique mechanism.

In addition to the immediate plans already proposed, future studies will strive to resolve whether the spore intron-containing transcripts are the same transcripts that are spliced in the meronts, which would tell us if spore transcripts are processed to exhaustion, and would inform us more about spore biology and post-transcriptional regulation. Another subject for future research is the mechanism for inhibiting splicing.

I plan to continue to look for *E. cuniculi* introns that are slightly more variable than the ones I have found so far. I have already found a potentially "decayed" intron; such findings would help us better picture how introns are lost or compacted. I am also looking for introns in other microsporidian species.

I am particularly interested in the structural basis (of the spliceosome) for 3'SS selection, but the lack of methods given the intracellular lifestyle of Microsporidia makes this type of research difficult. I am also interested in the functions of the non-RPG intron-containing genes.

References

Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., and Vivarès, C.P. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature.* **414**: 450-453.