THE EVOLUTION OF RNA PROCESSING IN REDUCED EUKARYOTES

by

Cameron James Grisdale

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

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2014

Abstract

RNA-processing encompasses several critical steps in the regulation of gene expression. Both transcription and pre-mRNA splicing are important for the formation of mature RNA. Most eukaryotic genes are interrupted by introns, the removal of which is catalyzed by the spliceosome. The spliceosome is a large molecular machine comprised of five small nuclear RNAs (snRNAs) and up to two hundred proteins. In addition to constitutive removal of introns, alternative splicing increases transcriptome complexity, as it allows for the formation of multiple transcript isoforms from a single pre-mRNA. Although these processes are well-studied in model systems, relatively little is known about their evolution in unicellular eukaryotes.

To investigate RNA-processing in reduced systems, I examined the transcriptomes of the microsporidian parasite *Encephalitozoon cuniculi*, and the red alga *Cyanidioschyzon merolae*. *E. cuniculi* and *C. merolae* harbour reduced genomes of 2.9Mbp and 16.5Mbp, respectively. Both genomes were annotated with fewer than 30 spliceosomal introns, and both have undergone reduction in spliceosomal components, including the loss of the U1 snRNA. Illumina RNAseq was used to sequence the transcriptomes of *E. cuniculi* at three time-points during its intracellular stage, and *C. merolae* under light and dark phases of its growth cycle. I found extremely low levels of pre-mRNA splicing for nearly all intron-containing genes in both organisms, under all conditions examined. These levels of splicing appear to be lower than in any other eukaryote examined, suggesting that reduction in unrelated spliceosomes reveals a common evolutionary trend: decreased splicing efficiency.

In addition to intron-retention, I found examples of other types of alternative splicing in these two reduced systems. *C. merolae* displayed all major types of alternative splicing, and

some events occurred at relatively high frequencies. The presence of few or no alternative splicing regulatory protein-coding genes in *C. merolae* and *E. cuniculi*, respectively, made this finding especially surprising. Also, I found high levels of antisense transcription in *C. merolae*, with the potential to play a regulatory role in gene expression.

Preface

A version of chapter 2 has been published. Grisdale, CJ, and Fast, NM. (2011) Patterns of 5′ untranslated region length distribution in *Encephalitozoon cuniculi*: implication for gene regulation and potential links between transcription and splicing. J Eukaryot Microbiol. 58:68-74. doi: 10.1111/j.1550-7408.2010.00523.x. The project was conceived by NMF and I. I conducted the RNA extractions, 5′ RACE PCR, Sanger sequencing, and base calling. NMF and I analyzed the data. I wrote the first draft. NMF and I edited the manuscript and wrote the final draft.

A version of chapter 3 has been published. Grisdale, CJ, Bowers, LC, Didier ES, and Fast, NM. (2013) Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of pre-mRNA splicing in a reduced eukaryote. BMC Genomics. 28;14:207. doi: 10.1186/1471-2164-14-207. The project was conceived by NMF and I. LCB and ESD grew the rabbit cell line, inoculated the cultured cells with *Encephalitozoon cuniculi* spores, and prepared the tissue to maintain RNA integrity. I conducted the RNA extractions, prepared the Illumina RNAseq cDNA library, analyzed the quality of Illumina reads, mapped Illumina reads to the reference genome, created custom Python scripts to analyze pre-mRNA splicing levels, and performed differential gene expression analysis. NMF and I analyzed the data. I wrote the first draft. NMF, LCB, ESD, and I edited the manuscript and wrote the final draft.

Chapter 4 is based on a manuscript in preparation. Grisdale, CJ, Tack, DC, and Fast, NM. (in preparation) High-throughput transcriptome sequencing of *Cyanidioschyzon merolae* reveals unexpected levels of constitutive and alternative splicing and antisense transcription. The project was conceived by NMF and I. I maintained *C. merolae* cultures, extracted RNA, prepared cDNA

libraries for Illumina sequencing, mapped reads to the reference genome, performed differential expression analysis, annotated new introns, assessed splicing levels, and quantified alternative splicing events. DCT and I developed Python scripts to analyze alternative splicing events from standard read alignment files. NMF and I interpreted the results. I wrote the first draft. NMF and I edited the manuscript.

Table of Contents

Abstract ..... **ii**

Preface ..... **iv**

Table of Contents ..... **vi**

List of Tables ..... **x**

List of Figures ..... **xi**

Acknowledgements ..... **xii**

Chapter 1: Introduction **1**

1.1 RNA processing 1

1.2 Genome reduction 6

1.3 Models of genome reduction and compaction 10

1.3.1 The microsporidian *Encephalitozoon cuniculi* 10

1.3.2 The red alga *Cyanidioschzyon merolae* 13

1.4 Research objectives 16

Chapter 2: Patterns of 5′ untranslated region length distribution in *Encephalitozoon cuniculi*: implications for gene regulation and potential links between transcription and splicing **20**

2.1 Introduction 20

2.2 Results 23

2.2.1 *Encephalitozoon cuniculi* 5′UTRs are very short 23

2.2.2 Exclusively short 5′UTRs for RPGs and intron-containing genes 24

2.3 Discussion 24

2.4 Conclusions 30
2.5 Materials and methods 30
2.5.1 RNA extraction and cDNA synthesis 30
2.5.2 Determining 5′UTR lengths 30
2.5.3 Gene selection 32
Chapter 3: Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of pre-mRNA splicing in a reduced eukaryote 41
3.1 Introduction 41
3.2 Results and discussion 43
3.2.1 Identification of novel transcribed regions 44
3.2.2 All coding regions are transcribed in intracellular *E. cuniculi* 45
3.2.3 High frequency of differentially expressed genes in the first 48 hrs 46
3.2.4 Analysis of pre-mRNA splicing 47
3.2.4.1 *E. cuniculi* has a reduced spliceosome 47
3.2.4.2 Discovery of introns and splice isoforms 48
3.2.4.3 Comparative analysis of intron-containing transcripts 50
3.3 Conclusions 54
3.4 Materials and methods 54
3.4.1 RNA preparation 54
3.4.2 RNA-seq library preparation 55
3.4.3 Illumina sequencing and data processing 55
3.4.4 Assessing splicing efficiency 56
3.4.5 Differential gene expression analysis 57

3.4.6 Search for novel transcribed regions (NTRs) 58

Chapter 4: High-throughput transcriptome sequencing of *Cyanidioschyzon merolae* reveals unexpected levels of constitutive and alternative pre-mRNA splicing and antisense transcription 62

4.1 Introduction 62

4.2 Results and discussion 67

4.2.1 Differential expression between light and dark phases 67

4.2.2 High prevalence of antisense transcription 70

4.2.3 Intron annotation 72

4.2.4 Pre-mRNA splicing levels 75

4.2.5 Alternative splicing 77

4.3 Conclusions 80

4.4 Materials and methods 81

4.4.1 Cell culture and RNA preparation 81

4.4.2 RNA-seq library preparation 81

4.4.3 Illumina sequencing and data processing 82

4.4.4 Assessing splicing efficiency 83

4.4.5 Differential gene expression analysis 84

4.4.6 Analysis of cis-NATs 84

Chapter 5: Conclusion 93

5.1 Summary 93

5.2 Future directions 96

Bibliography 99

Appendices.....120

Appendix A Supplementary tables and figures of chapter 2 120

Appendix B Supplementary tables and figures of chapter 3..... 122

List of Tables

Table 1.1 Protein content of spliceosomal complexes 18

Table 2.1: A list of the names and lengths of 5' untranslated regions (UTRs) of all 155 genes of *Encephalitozoon cuniculi* tested. 33

Table 3.1: Number of reads mapped to parasite and host genomes 59

Table 4.1: Number of reads mapping to *Cyanidioschyzon merolae* reference genome 86

Table 4.2: Frequency of alternative splicing events in *Cyanidioschyzon merolae* during light and dark growth 87

Table 4.3: Characteristics of 40 *Cyanidioschyzon merolae* introns in protein-coding genes 88

Table B.1: Gene expression levels in FPKM are shown for all 1985 *Encephalitozoon cuniculi* genes at three post-infection time-points 122

Table B.2: RNA decay genes in *Encephalitozoon cuniculi* 165

List of Figures

Figure 1.1: Mechanisms of alternative splicing 19

Figure 2.1: Bargraph of 5'UTR lengths in *Encephalitozoon cuniculi* 39

Figure 2.2: Distribution of 5'UTR lengths in *Encephalitozoon cuniculi* 40

Figure 3.1: Differential expression across three post-infection time-points 60

Figure 3.2: Splicing levels of all *E. cuniculi* intron-containing genes 61

Figure 4.1: Differential expression of 4494 *Cyanidioschyzon merolae* genes during light and dark conditions 89

Figure 4.2: Differential expression of *Cyanidioschyzon merolae* antisense transcripts during light and dark conditions 90

Figure 4.3: Distribution of antisense transcription levels for all *Cyanidioschyzon merolae* genes 91

Figure 4.4: Levels of pre-mRNA splicing for all 43 introns in *Cyanidioschyzon merolae* during light and dark 92

Figure A.1: The number of *Encephalitozoon cuniculi* genes examined at each 5' untranslated region (UTR) length between 0 and 20 bp 120

Figure A.2: A histogram of lengths of 5' untranslated regions (UTRs) for 155 genes of *Encephalitozoon cuniculi* 121

Figure B.1: Intron motifs of *Encephalitozoon cuniculi introns* 166

Figure B.2: Splicing levels in two fungal species 168

Acknowledgements

I would to thank my supervisor, Dr. Naomi Fast, for giving me the opportunity to work in her lab from undergraduate through PhD research, and for always being positive, encouraging, and understanding. I would like to thank my committee members, Dr. Keith Adams and Dr. Patrick Keeling, for their guidance and thoughtful suggestions regarding my research.

This research was funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Alexander Graham Bell Canada Graduate Scholarship (CGS-D3) awarded to me, as well as an NSERC discovery grant awarded to Naomi Fast.

I thank past and present lab members for their help with technical issues, and for being available to discuss research and non-research related topics: Alex Ardila, Dr. Erin Gill, Renny Lee, Donald Wong, and Thomas Whelan. I would also like to thank members of the Adams lab and Keeling lab for their help with various projects. Special thanks to David Tack for teaching me Python programming and for discussions of alternative splicing, as well as Dr. Jean-François Pombert for technical help with computing issues and teaching me the basics of Linux.

And finally, I would like to thank my family and friends for their support and understanding during a long and sometimes demanding process. Very special thanks to my fiancée Kim for supporting me throughout my time as a PhD student.

Chapter 1: Introduction

1.1 RNA processing

The control of transcription and pre-mRNA processing is fundamental in regulating gene expression. Regulation can be achieved via several distinct pathways, and involves both cis- and trans-acting elements. Transcription is a multi-step process involving a large number of proteins that interact with the DNA template, the RNA molecule being synthesized, and with one another. The dynamic interactions between transcription-associated proteins are responsible for modulating the level of transcription of all genes, both in time and space. Pre-mRNA splicing also involves a large number of components. Although self-splicing introns exist in prokaryotic and organellar genomes, pre-mRNA splicing involving cis- and trans-acting elements is unique to eukaryotes. The splicing process likely evolved as a result of the invasion of coding sequence by intron elements, and is thought to have played a major role in eukaryogenesis and in the expansion of eukaryotic diversity (Koonin 2006). Pre-mRNA splicing is also of medical importance as mis-splicing can play major roles in many human diseases (Nissim-Rafinia and Kerem 2005; Wang and Cooper 2007). Historically pre-mRNA splicing was thought to follow transcription and act on mature transcripts, however, we now know that these two processes are not independent. In fact, steps in RNA processing such as 5′ capping, 3′ polyadenylation, and pre-mRNA splicing have been shown to overlap, both in terms of timing and sharing of components, with transcription and the RNA polymerase II complex (Bentley 2002; Howe 2002; Kornblihtt *et al*. 2004; Maniatis and Reed 2002; Neugebauer 2002; Proudfoot 2003; Proudfoot, Furger, Dye 2002). This adds a further level of potential regulation by allowing for interactions

between complexes involved in the different steps of RNA processing, and makes for a highly complex system with many opportunities to regulate gene expression.

Spliceosomal introns present in pre-mRNAs must be removed in order for transcripts to become fully mature and be exported to the cytosol for translation. The excision of introns is catalyzed by the spliceosome, thought to be one of the largest molecular machines in the cell (Nilsen 2003). The spliceosome is composed of five small nuclear RNAs (snRNAs) and approximately twenty to more than two hundred protein components (Jurica and Moore 2003; Staley and Guthrie 1998). Each snRNA is associated with a group of core proteins as well as additional spliceosomal factors to form the five small nuclear ribonucleic particles (snRNPs). The snRNP components interact with conserved motifs within the intronic sequence, including the 5′ splice site, branch point motif, polypyrimidine tract, and 3′ splice site (Figure 1.1 A). These cis elements are essential for recruiting snRNP complexes, interacting with snRNAs during the splicing reactions, and also playing a role in the regulation of splicing.

The first step in pre-mRNA splicing involves the U1 snRNP identifying and binding to the 5′ splice site of the intron. The U2 snRNP then recognizes and binds the branch point motif. The U4/U6.U5 tri-snRNP is recruited to the pre-mRNA, bringing about major shifts in RNA-RNA and RNA-protein interactions within the spliceosome. These changes result in the dissociation of the U1 and U4 complexes and, followed by the action of the PRP2 RNA helicase, produce the catalytically active spliceosome. The first of two transesterification reactions takes place when the 2′ hydroxyl group of the intron's branch point adenosine performs a nucleophilic attack on the 5′ splice site, ligating the 5′ end of the intron with the branch point to form a lariat structure. The completion of the first catalytic step activates the spliceosome for the second transesterification reaction in which the 3′ hydroxyl of the 5′ exon performs a nucleophilic attack

on the 3′ splice site. The products of the second reaction are the free intron lariat and the ligated exons of the mRNA (reviewed in (Moore, Query, Sharp 1993). The mechanistic features of pre-mRNA splicing described above appear to be conserved in eukaryotes (Anantharaman, Koonin, Aravind 2002; Collins and Penny 2005; Davila Lopez, Rosenblad, Samuelsson 2008; Irimia, Penny, Roy 2007; Kaufer and Potashkin 2000; Schellenberg, Ritchie, MacMillan 2008). However, the lack of many essential splicing components in some reduced eukaryotes raises questions about the exact pathway used in these unusual systems.

Pre-mRNA splicing can be regulated to control gene expression in a number of cellular processes including cell cycle progression, as well as abiotic and biotic stress conditions (Dabeva and Warner 1993; Davis *et al*. 2000; Engebrecht, Voelkel-Meiman, Roeder 1991; Fewell and Woolford 1999; Juneau *et al*. 2007; Li, Vilardell, Warner 1996; Pleiss *et al*. 2007). There are many examples of autoregulatory splicing in yeast ribosomal genes, where the gene product can act on its own transcript to modulate the level of splicing (Dabeva and Warner 1993; Fewell and Woolford 1999; Li, Vilardell, Warner 1996). The *MER2* gene in yeast is an example of meiotically controlled splicing regulation (Engebrecht, Voelkel-Meiman, Roeder 1991). Although *MER2* is transcribed in both mitotic and meiotic phases of the yeast cell cycle, it is only spliced efficiently during meiosis as a result of the action of a related gene, *MER1* (Engebrecht, Voelkel-Meiman, Roeder 1991). Several other yeast genes have been shown to undergo regulatory splicing during meiosis, as well as during exponential growth (Davis *et al*. 2000; Engebrecht, Voelkel-Meiman, Roeder 1991; Juneau *et al*. 2007). Amino acid starvation in yeast results in both the upregulation and downregulation of groups of genes as a result of regulatory splicing (Pleiss *et al*. 2007).

While the rate of splicing itself can be controlled by a variety of factors to influence the production of mature mRNA, introns can also be spliced alternatively to produce a multitude of functional or non-functional mRNAs from a single gene. The ability to create multiple distinct transcripts from a single gene increases the protein repertoire of an organism without the need for an increase in gene number, essentially providing an efficient means of increasing a genome's coding capacity (for review, see (Graveley and Nilsen 2010; Stamm et al. 2005). The variety of transcripts produced from a single gene can be staggering, such as in the case of the *Drosophila Dscam* gene, which has more than twice the number of alternative splice forms than there are genes in the genome (Adams *et al.* 2000; Schmucker *et al*. 2000). Alternative splicing also provides a system to negatively regulate gene expression by introducing premature stop codons, resulting in the degradation of mRNA via the nonsense mediated decay (NMD) pathway. Although not all pre-mRNA with premature stop codons are targeted for NMD, it has been shown experimentally that 13-18% of multi-exon genes are NMD sensitive in *A. thaliana* (Drechsel *et al*. 2013; Kalyna *et al*. 2011). Recent evidence suggests that alternative splicing is more prevalent than originally expected, with up to 100% of multi-exon human genes believed to produce multiple isoforms (Pan *et al*. 2008; Wang 2008). Plant genomes also encode a large number of alternatively spliced genes, with the most recent estimate showing approximately 61% of *Arabidopsis thaliana* genes producing multiple mRNA isoforms (Marquez *et al*. 2012). However, the frequency and types of alternative splicing events in unicellular eukaryotes, especially those with reduced or compacted genomes, have yet to be examined.

Alternative splicing comes in four basic forms: alternative 5′ splice site usage, alternative 3′ splice site usage, exon skipping, intron retention, or a combination of more than one of these (Figure 1.1 B-D). Some types occur more commonly in certain lineages, such as exon skipping

being the major form of alternative splicing in mammals, while intron retention is the most common form in plants and fungi (Kim, Magen, Ast 2007; McGuire et al. 2008; Ner-Gaon et al. 2004). While alternative splicing is fairly well characterized in metazoans, relatively little is known about the prevalence and importance of alternative splicing in unicellular eukaryotes. Current evidence from studies of unicellular eukaryotes suggests that alternative splicing occurs at a much lower frequency in these groups compared to metazoans (Grisdale *et al*. 2013; He *et al*. 1993; Kabran *et al*. 2012; Labadorf *et al*. 2010; Loftus *et al*. 2005; Marshall *et al*. 2013; Muhia *et al*. 2003; Sorber, Dimon, DeRisi 2011). Also, intron retention seems to be the most prevalent form of alternative splicing in unicellular eukaryotes, fungi, and plants, while exon skipping is among the rarest type of event (Ast 2004; Grisdale et al. 2013; Labadorf et al. 2010; McGuire et al. 2008; Wang and Brendel 2006). Some propose this is related to the phenotypic complexity of the organism, as metazoans can achieve much greater proteome complexity with high levels of exon skipping (Kim, Goren, Ast 2008). While this appears to hold true in most cases, a few notable exceptions have been found. The protist *Bigelowiella natans* is a unicellular eukaryote with high levels of alternative splicing, including frequent exon skipping events (Curtis *et al*. 2012). However, the evidence of a functional role for many splice forms in *Bigellowiella* was weak, leading us to suggest that the alternative splicing events observed in *B. natans* are likely a result of mis-splicing (Curtis *et al*. 2012). A protist closely related to metazoans, *Capsaspora owczarzaki*, shows regulated intron retention and exon skipping events that appear to be involved in life-cycle stage transitions (Sebé-Pedrós *et al.* 2013). In plants, it has been suggested that the high frequency of intron retention and rarity of exon skipping is a result of frequent whole genome duplication events (Kim, Goren, Ast 2008). When two copies of a gene are present, generally the selective pressure acting on one copy will be relaxed, resulting

in neofunctionalization, and potentially avoiding the need for multiple splice forms to achieve greater proteome diversity. Although alternative splicing has been documented in a small number of unicellular eukaryotes, it generally occurs at a lower frequency than in metazoans. Also, the high frequency of intron retention suggests that many alternative splicing events may be the result of mis-splicing. Therefore, assessing the prevalence and role of alternative splicing in unicellular eukaryotes will be important in understanding the origins and evolution of this important step in RNA processing.

1.2 Genome reduction

Nuclear genome size has a range of over 600,000 fold across eukaryote diversity (Gregory TR *et al*. 2005). From the tiny genomes of parasitic microsporidians (*E. intestinalis* 2.3Mb; (Corradi *et al*. 2010a)), to the enormous amoeba genomes (*Chaos chaos* 1,400,000Mb; (Friz 1968)), genome size can evolve by a variety of different mechanisms on both short and long timescales. If we were to include the miniaturized remnant nuclear genomes of algal endosymbionts (Douglas *et al*. 2001; Gilson *et al*. 2006), the genome size range would be closer to 2.8 million fold. Studying genomes at the extremes of this size range will help us to understand the mechanisms and drivers for genome evolution. However, studying large genomes has substantial technical challenges associated with it, such as difficulties in DNA sequencing due to repetitive sequences. On the other hand, it is much easier to examine small genomes since they require less sequence data to achieve a high depth of coverage over the entire genome. Studying reduced genomes can give us insight into the minimal requirements for molecular pathways, as well as provide a less complex (due to fewer components) background in which to identify interactions between molecules. Although genome evolution has been studied in several

models of eukaryotic genome reduction, there is still much to learn about the effects of reduction on processes related to gene expression.

Two major contributors to increases in genome size over evolutionary timescales are duplication events (both partial and whole genome) and the actions of selfish genetic elements such as transposons. However, certain pressures can lead to a long term decreasing trend in genome size. Prokaryotic and eukaryotic organisms involved in symbioses, such as endosymbionts and parasites, tend to show significant genome reduction and compaction in comparison to their free-living relatives (Corradi *et al*. 2010b; Martin and Herrmann 1998; McCutcheon and Moran 2011; Moore and Archibald 2009; Moran, McLaughlin, Sorek 2009; Sasaki *et al*. 2002). There are several mechanisms by which these organisms can become reduced over evolutionary timescales. The most severe form of reduction, in terms of the biology and metabolism of the organism, is the loss of genes.

The close association of an endosymbiont or parasite with its host can allow the symbiont to take advantage of readily available host metabolites. This relaxes evolutionary pressures to retain genes involved in the homologous metabolic pathways in the symbiont, which can result in gene loss. Microsporidian parasites provide a clear example of this as they have lost the genes required to produce ATP via the tricarboxylic acid pathway, but have several ATP/ADP transporters expressed on their outer membrane, allowing them to take up this key energy molecule from the host cytoplasm (Chen *et al*. 2013; Cuomo *et al*. 2012; Katinka *et al*. 2001; Tsaousis *et al*. 2008). The intracellular lifestyle of endosymbionts and some parasites can also allow for the loss of some genes that function in motility and cellular defense (Martin and Herrmann 1998; Martin *et al*. 2002; McCutcheon and Moran 2011). Another mechanism of gene loss from endosymbiont genomes, with less severe outcomes, is through gene transfer to the host

nucleus and co-translational or post-translational targeting to the endosymbiont (Bolte *et al*. 2009; Curtis *et al*. 2012; Martin *et al*. 2002; Timmis *et al*. 2004). Endosymbiont-to-host gene transfer is the mechanism by which mitochondria and plastids, which are of prokaryotic origin, have lost more than 90% of their predicted coding capacity compared to their free-living relatives (Gould, Waller, McFadden 2008; Gray, Lang, Burger 2004; Martin and Herrmann 1998; Reyes-Prieto, Weber, Bhattacharya 2007; Timmis *et al*. 2004). This phenomenon is also well known in eukaryotes with secondary and tertiary plastids that become obligate endosymbionts following prolific gene transfer and re-targeting (Bolte *et al*. 2009; Keeling and Palmer 2008; McFadden 1999; Patron, Waller, Keeling 2006).

In addition to gene loss, the compaction of genes and intergenic regions plays a role in reducing the overall size of a genome. Protein-coding genes in nucleomorph genomes have been found to be smaller than their homologs in other eukaryotes, displaying losses at their amino and carboxy termini, as well as internal deletions (Lane *et al*. 2007). Interestingly, this phenomenon is mirrored between nucleomorph genomes of slightly different size. The majority of genes in the genome of *Hemiselmis andersenii* (0.572Mbp) with homologs in *Guillardia theta* (0.551Mpb), are larger in *H. andersenii* (Lane *et al*. 2007). A similar trend appears when examining intergenic spacers, as the average intergenic region size is 92bp in *H. andersenii* and 50bp in *G. theta* (Lane *et al*. 2007). Reduction in gene size has also occurred in microsporidian parasites. Katinka *et al.* found that 85% of genes in *E. cuniculi* are shorter than their homologs in *Saccharomyces cerevisiae*, and it was hypothesized that this may be the result of a loss of protein interaction domains due to the reduction in the proteome (Katinka *et al*. 2001). A reduction in intergenic size and repetitive DNA accounts for perhaps the largest amount of change in microsporidian genome size relative to free living fungal relatives (Katinka *et al*. 2001; Keeling and Slamovits

2004). The compaction of genes and intergenic spaces appears to be a major contributor to the process of genome reduction.

The presence of abundant non-coding and repetitive DNA in large genomes suggests that these types of DNA strongly influence the size of the genomes in which they reside (Kidwell 2002; Taft, Pheasant, Mattick 2007). Although the correlation does not hold true in all lineages, the frequency and size of introns appears to be related to genome size (Deutsch and Long 1999; Vinogradov 1999). In the reduced genomes of hemiascomycetous yeasts, microsporidians, and algal endosymbionts, introns are rare and those present are shorter than the average intron size in larger eukaryotic genomes (Corradi *et al*. 2010a; Davis *et al*. 2000; Douglas *et al*. 2001; Katinka *et al*. 2001; Mitrovich *et al*. 2007). There are even known cases of total intron loss, typically accompanied by an absence of spliceosome-associated genes (Cuomo *et al*. 2012; Keeling *et al*. 2010; Lane *et al*. 2007). The apparent correlation between intron number and spliceosome complexity suggests a tight co-evolution between the components responsible for carrying out splicing, and their substrates. Some reduced genomes have shorter than average repeat units, such as in telomeres and tandem ribosomal subunits (Katinka *et al*. 2001; Matsuzaki *et al*. 2004). In fact, the ribosomal operons themselves can be packaged within the telomeric repeats, a phenomenon common to microsporidian and nucleomorph genomes (Douglas *et al*. 2001; Gilson and Mcfadden 1995; Slamovits, Williams, Keeling 2004; Zauner 2000). Thus, loss and compaction of non-coding and repetitive DNA are influential mechanisms of genome reduction in eukaryotes.

1.3 Models of genome reduction and compaction

1.3.1 The microsporidian *Encephalitozoon cuniculi*

Microsporidia are a large group of unicellular eukaryotes that are obligate intracellular parasites. The vast majority of these parasites infect animals, while a few species infect protists (Fokin *et al*. 2008; Larsson 2000; Scheid 2007; Wittner and Weiss 1999). Microsporidians were first discovered as the causative agent of the silkworm collapse during the 19th century, and were originally classified as fungi. Their phylogenetic classification has been debated and continually re-assessed since their discovery over 150 years ago: earlier hypotheses suggested microsporidians as primitive, early branching eukaryotes as part of the Archezoa hypothesis (Cavalier-Smith 1983; Cavaliersmith 1987), whereas recent molecular evidence points to their placement within the Fungi (Corradi and Keeling 2009; Gill and Fast 2006; James *et al*. 2006; Keeling, P.J.,and Fast,N.M. 2002; Keeling, Luker, Palmer 2000; Keeling 2003a; Lee *et al*. 2008). The current view is that Microsporidia belong within the fungi, however, the exact placement within this kingdom is not yet certain.

Microsporidia have a unique infection mechanism that involves the use of their characteristic coiled polar tube. Typical microsporidians have two life-stages, an extracellular spore stage in which the organism is thought to be largely dormant, and intracellular meront stages in which the parasite grows and divides asexually at a rapid rate. When a spore comes into proximity with a potential host cell, its posterior vacuole expands rapidly causing the polar tube to burst out of the spore at its apex (Wittner and Weiss 1999). When ejected, the polar tube inverts and will pierce any nearby host cell. The sporoplasm is then transferred through the polar tube and deposited into the host cell cytoplasm. The intracellular, or meront, stage now begins taking advantage of host energy supplies in order to grow and divide rapidly. Eventually the

population of meronts undergoes sporogony, producing walled spores that are ready to burst out of their host and infect new host cells (for review, see (Bigliardi and Sacchi 2001; Cali and Takvorian 1999; Dunn, A.M.,and Smith,J.E. 2001; Franzen 2004; Wittner and Weiss 1999)).

Although Microsporidia are well known for having some of the smallest nuclear genomes, those at the high end of the 10-fold size range are within the realm of typical unicellular eukaryotic genome sizes. At the small end of the range is the ultra-compact and reduced 2.3Mb genome of *Encephalitozoon intestinalis* (Corradi *et al*. 2010a), while the largest is estimated to be the 24Mb genome of *Octosporea bayeri*, which has low gene density and contains repetitive elements (Corradi *et al*. 2009). The first fully sequenced microsporidian genome was that of *Encephalitozoon cuniculi*, and its position as the model microsporidian likely contributed to the misconception that microsporidia all have tiny genomes. At just 2.9Mb and encoding less than two thousand genes, it was found to be extremely gene dense (1 gene/Kb), with short intergenic regions and a highly reduced gene-set (Katinka *et al*. 2001). This reduction in gene number has resulted in the loss of many genes involved in canonical eukaryotic metabolic pathways, including protein-coding genes that are essential in *Saccharomyces cerevisiae*. This makes *E. cuniculi* a useful model system in which to study essential cellular processes that involve large numbers of interacting factors in typical eukaryotes, such as transcription and pre-mRNA splicing.

Genome reduction and compaction has led to some unusual phenomena in *E. cuniculi*. Overlapping transcription has been characterized in the spore stage of more than one microsporidian species, though it is not typical among eukaryotes. The few other characterized examples of multi-gene transcription occur in the miniaturized nucleomorph genomes of remnant endosymbiont nuclei of cryptomonad and chlorarachniophyte algae (Williams *et al*. 2005). The

compaction in nucleomorph genomes has resulted in nearly all genes producing overlapping transcripts (Williams *et al*. 2005). The multi-gene transcripts produced are likely not functional in the way that bacterial operons are, as they generally contain only a single complete coding sequence. Interestingly, different types of multi-gene transcripts are present in different microsporidian species. In *E. cuniculi* transcripts tend to overlap with upstream ORFs, while those in the distantly related microsporidian *Antonospora locustae* typically overlap with downstream ORFs (Corradi, Gangaeva, Keeling 2008; Williams *et al*. 2005). A compelling function (if one exists) for overlapping transcription in microsporidian spores is not clear. However, we propose that overlapping transcripts are likely the result of transcriptional regulatory elements present in upstream or downstream ORFs due to extreme compaction, or perhaps just a result of a lack of regulation in the spore stage (Corradi, Gangaeva, Keeling 2008; Gill *et al*. 2010; Williams *et al*. 2005).

In order to determine if overlapping transcription is unique to microsporidian spores or if it also occurs in the intracellular stage, we examined the transcription products of 31 genes from both life stages of *E. cuniculi* (Gill *et al*. 2010). We used the 5′ rapid amplification of cDNA ends (RACE) method in order to obtain the 5′ targeted region of the transcripts along with the full length 5′UTRs. We found that 5′UTR regions in the intracellular stage are much shorter than those in the spore stage, and as a result, multi-gene transcripts are much less common in the intracellular stage (Gill *et al*. 2010). Also, the number of transcription start sites observed was much higher in spore transcripts than in meront transcripts, indicating a loosening of transcriptional regulation (Gill *et al*. 2010). This lack of transcriptional regulatory control in the spore stage may be a result of a dormant physiological state of microsporidian spores. It is also possible that the long, multi-gene transcripts play a structural role in tethering ribosomes into

polyribosome structures, which has been observed in some microsporidians (Vavra, J. & Larsson, J.I.R. 1999). This could be an important energy saving mechanism employed by the spore, essentially stalling translation until it infects a viable host.

E. cuniculi was annotated with just 15 spliceosomal introns, suggesting severe intron loss took place during genome reduction. Since very few spliceosomal protein-coding genes and snRNAs were identified in *E. cuniculi,* and a lack of splicing would introduce premature stop codons in these 15 genes, it seemed pertinent to investigate pre-mRNA splicing in this reduced genome. In our study of spore and meront transcripts, we analyzed the status of all known *E. cuniculi* introns in pre-mRNA transcripts in both spore and meront stages (Gill *et al*. 2010). We found evidence of all introns being removed from meront transcripts, while spore transcripts always retained their introns (Gill *et al*. 2010). This suggested two very interesting implications. First, that the process of pre-mRNA splicing is proceeding successfully without the canonical, or perhaps identifiable, spliceosome components found in nearly all other eukaryotes. Second, that at least a small subset of spore transcripts is not functional. The latter solidified our belief that transcripts may have a translation-independent role in spores.

1.3.2 The red alga *Cyanidioschzyon merolae*

Cyanidioschyzon merolae belongs to a species-poor lineage of Rhodophyta comprised of organisms that live in extreme environments. *Cyanidioschyzon* is one of three recognized genera within the order Cyanidiales (Ciniglia *et al*. 2004). Phylogenetic evidence suggests that Cyanidiales are the basal lineage within the red algal tree, and that their common ancestor was thermo-acidic tolerant (Ciniglia *et al*. 2004; Yoon *et al*. 2002). Although the three genera within Cyanidiales: *Cyanidioschyzon*, *Galdieria*, and *Cyanidium*, are believed to be very low in species

number, recent surveys of biodiversity have found significantly more diversity than expected (Ciniglia *et al*. 2004; Gross *et al*. 2001). The sequence divergence levels between many environmentally sampled Cyanidiales lineages and/or species, are higher than the levels of between-order divergence of non-Cyanidiales red algae (Ciniglia *et al*. 2004). This suggests that either high sequence divergence rates are prevalent in Cyanidiales, likely as a result of their extreme habitats, or that perhaps there is more species richness than currently thought. It is also interesting to note the relatively high degree of genetic divergence between morphologically very similar organisms: putatively the product of long periods of isolation at geographically distant locations but similar habitats (Gross *et al*. 2001).

The *Cyanidioschzyon* type species was originally isolated from a thermal vent in the Campi Flegrei caldera in the west of Naples (De Luca, Taddei, Varano 1978). *C. merolae* is a tiny (~2μm), non-flagellated, non-walled, photoautotrophic unicell. Cell structure is very simple, with just a single nucleus, mitochondrion, plastid, Golgi, and endoplasmic reticulum. Little evidence exists in support of sexual reproduction in *C. merolae*, suggesting division proceeds only by binary fission. Also, cell division can be synchronized with appropriate light:dark cycles and CO_2 input growth conditions (Terui *et al*. 1995). These cellular features and the ability to synchronize cultures have brought much attention to *C. merolae* as a model system for the study of fundamental eukaryotic processes, such as organellar division (Kuroiwa 1998; Minoda *et al*. 2004; Misumi *et al*. 2005; Terui *et al*. 1995).

In 2004 the 16.5Mb haploid genome sequence of *C. merolae* was published (Matsuzaki *et al*. 2004), and in 2007 the reference genome was improved upon with the first 100% complete eukaryotic genome sequence (Nozaki *et al*. 2007). The genome encodes approximately five thousand genes on twenty short chromosomes. Chromosomes have typical eukaryotic properties,

with short telomeric repeats at their ends, and a single A/T rich centromeric region within each chromosomal core (Matsuzaki *et al*. 2004). Introns are unusually sparse, interrupting just 26 genes, and only a single gene contains more than one intron (Matsuzaki *et al*. 2004). The 5′ splice site motif of these introns appears to be very strict, a common phenomenon in intron poor species (Irimia, Penny, Roy 2007; Irimia and Roy 2008; Schwartz *et al*. 2008). *C. merolae* appears to have a minimal set of ribosomal RNAs, with just three copies of an 18S-5.8S-28S rDNA unit, and no tandem repeat units commonly found in eukaryotes (Matsuzaki *et al*. 2004). Overall, these features suggest that the genome of *C. merolae* is reduced and compacted relative to more typical eukaryotic genomes.

The relatively small coding capacity of *C. merolae* has given way to some molecular pathways being reduced relative to eukaryotes with larger, more gene-rich genomes. For example, several photosystem genes commonly found in phototrophic eukaryotes that are involved in dissipating light energy, are not found in *C. merolae*, suggesting a high sensitivity to light stress (Matsuzaki *et al*. 2004). As mentioned above, very few introns are annotated in *C. merolae*. In fact, just 0.5% of genes contain an intron. On the other hand, several genomes of a similar size encode a much larger fraction of intron-containing genes, such as *S. cerevisiae* (12Mb), *C. albicans* (16Mb), *Schizosaccharomyces pombe* (12.5Mb), and *Ostreococcus tauri* (12.5Mb), with 6%, 8%, 43%, and 39% intron-containing genes, respectively (Bruno *et al*. 2010; Derelle *et al*. 2006; Nagalakshmi *et al*. 2008; Wood *et al*. 2002). This paucity of introns is complemented by a reduced spliceosome, putatively lacking many conserved splicing factors. Just thirty ORFs with homology to spliceosomal protein-coding genes were identified in the *C. merolae* genome (Table 1.1), some with questionable levels of identity. Also, all five snRNAs were not identified, either by the original genome annotation, or a comprehensive computational

screen for snRNAs in all published genome sequences (Lopez, Rosenblad, Samuelsson 2008; Matsuzaki *et al*. 2004; Nozaki *et al*. 2007). This lack of splicing factors, many considered essential in yeast and other model systems, raises questions regarding the mechanism and outcomes of splicing in *C. merolae*.

1.4 Research objectives

The aim of my thesis research is to gain insight into the evolution of transcription, pre-mRNA splicing, and spliceosomal machinery, in reduced eukaryotic systems. The microsporidian *E. cuniculi* is a model of extreme genome reduction and compaction. Reductive pressures have led to many biological pathways being entirely absent or missing components that are considered essential in most other eukaryotes, while compaction has led to tiny intergenic spaces and frequent gene overlapping. *C. merolae* has a relatively small, intron poor, gene dense genome that lacks many spliceosomal components. Together, these provide an ideal comparative platform for studying the outcome of transcription and splicing in simplified systems with fewer interacting components. It also allows us to see how evolutionarily conserved processes can change at one extreme of genome size evolution. The conclusions drawn from such a comparison are strengthened by examining systems that have independently undergone genome reduction and that have been shaped by the evolutionary pressures of vastly different lifestyles. With little transcriptomic data available for microsporidians and unicellular red algae, I endeavored to examine the transcriptomes of these two reduced eukaryotic systems by completing the following projects:

1. Examine 5′UTR lengths in a subset of of *E. cuniculi* genes in order to elucidate any relationships between 5′UTR length and gene functional categories.

2. Perform a transcriptome-wide analysis of *E. cuniculi* at three post-infection time points to assess the levels of splicing and transcription during the intracellular stage of the life cycle.

3. Complete a transcriptome analysis of *C. merolae* during light and dark cycles to provide a comparison of splicing and transcription in an independently reduced eukaryote that is distantly related to *E. cuniculi*.

Table 1.1 Protein content of spliceosomal complexes

	H. sapiens[a]	*S. cerevisiae*[a]	*A. thaliana*[a]	*C. merolae*[b]	*E. cuniculi*[c]
U1 snRNP	4	11	8	0	2
U2 snRNP	16	12	17	9	5
U4/U6 snRNP	5	5	5	1	0
U5 snRNP	24	19	28	4	2
U4-5-6 tri-snRNP	4	3	5	1	1
Core Sm/Lsm	15	15	15	13	13
Splice site selection	7	2	9	2	0
hnRNP	12	0	29	0	0
EJC complex	8	1	9	0	0
SR proteins	13	0	10	2	1
Second step factors	6	5	6	3	0
SR protein Kinases	3	0	4	0	0
Other splicing factors	12	1	5	0	3
DEAD/H box helicases	8	3	6	0	3
Related to spliceosome	49	5	36	0	0
Total	186	82	192	35	30

[a] Data taken from the ASRG database (Wang and Brendel 2004).
[b] Data from (Matsuzaki *et al*. 2004).
[c] Data from (Katinka *et al*. 2001).

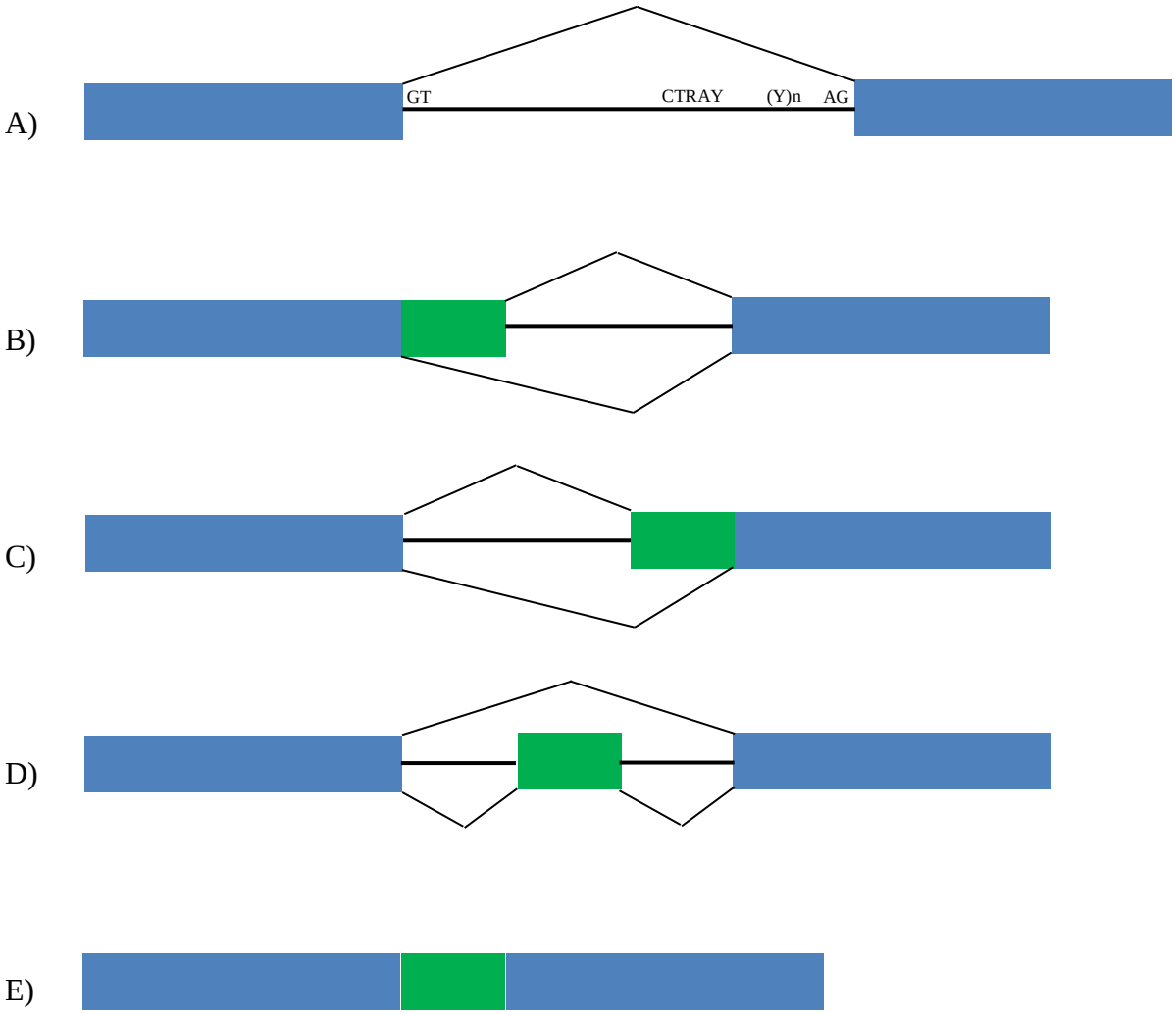


Figure 1.1: Mechanisms of alternative splicing

Four common mechanisms of alternative splicing in eukaryotes: alternative 5′ splice site usage (B), alternative 3′ splice site usage (C), exon skipping (D), and intron retention (E). The positions of the canonical 5′ splice site (GT), branch point (CTRAY), polypirymidine tract (Y), and 3′ splice site (AG) are shown on a constitutive intron (A).

Chapter 2: Patterns of 5′ untranslated region length distribution in *Encephalitozoon cuniculi*: implications for gene regulation and potential links between transcription and splicing

2.1 Introduction

Microsporidia are a group of unicellular eukaryotes that are intracellular parasites of many animals and several protist species. Although previously considered to be primitive eukaryotes lacking several key eukaryotic characteristics, microsporidia are now known to be specialized fungi (reviewed in (Corradi and Keeling 2009)). Their specific phylogenetic relationship with fungi is still a matter of debate; some evidence suggests that microsporidia could be sisters of fungi, whereas other data place microsporidia within the zygomycetes (Corradi and Keeling 2009; James *et al*. 2006; Keeling 2003b; Lee *et al*. 2008). A typical microsporidian life cycle consists of two stages, a proliferative intracellular stage and an infectious extracellular stage (Vavra, J. & Larsson, J.I.R. 1999). The extracellular spore accomplishes infection by injecting its cellular content into a host cell using its polar tube as a conduit (reviewed in (Franzen 2004; Vavra, J. & Larsson, J.I.R. 1999)).

By eukaryote standards, microsporidia have small genomes, ranging from 2.3 Mb in *Encephalitozoon intestinalis* to approximately 24 Mb in *Octosporea bayeria* (Corradi *et al*. 2009; Corradi *et al*. 2010a). The low end of this range represents the smallest genome of a free-living eukaryote—considerably smaller than many bacterial genomes. At just 2.9 Mb and encoding approximately 2,000 genes, *Encephalitozoon cuniculi* has a highly reduced and compacted genome (Katinka *et al*. 2001). A reduction in the number of genes and non-coding elements is

evident, as well as extreme compaction of intergenic spaces to an average of 129 bp (Katinka *et al*. 2001). Such compaction leaves very little room for cis-acting regulatory elements and the trans-acting factors that bind them.

Presumably as a result of genome miniaturization in *E. cuniculi*, highly conserved processes, such as transcription and splicing, have been known to deviate from their canonical pathways. For example, overlapping transcription, a highly unusual phenomenon among eukaryotes, has been observed in some microsporidian species (Corradi, Gangaeva, Keeling 2008; Williams *et al*. 2005). Overlapping transcription has also been found in the tiny nucleomorph genomes of algal endosymbionts where it occurs in 80–100% of genes, and may represent a by-product of extreme genomic compaction (Williams *et al*. 2005). In the microsporidian *Antonospora locustae*, a locust parasite, and the mammalian parasite *E. cuniculi* there are high frequencies of multi-gene transcripts in the spore stage (Corradi, Burri, Keeling 2008; Williams *et al*. 2005). The types of multi-gene transcripts produced in these distantly related microsporidians differ in that *A. locustae* transcripts tend to overlap with downstream genes while *E. cuniculi* transcripts tend to start within upstream genes (Corradi, Gangaeva, Keeling 2008). However, these are unlikely to represent multiple functional transcripts like those produced from bacterial operons as there is often only one complete gene encoded on the transcript and sometimes the adjacent genes are encoded on opposite strands. This pattern of multi-gene transcription was discovered in transcripts isolated from the extracellular spore stage of the parasite. To determine whether this phenomenon is only associated with the spore stage or occurs in both microsporidian life stages, we analyzed a common set of gene transcripts from both spore and intracellular meront stages. We found evidence that overlapping transcription is much less common in meronts, where transcripts are significantly shorter (Gill *et al*. 2010).

The 5′ untranslated region (5′UTR) of a gene is the sequence between the transcription start site (TSS) and the translation start site. The 5′UTR lengths in eukaryotes are known to vary widely across species and even within the same genome. However, they do show a correlation with genome size: the average 5′UTR in *Saccharomyces cerevisiae* is 83 bp long, and any given 5′UTR length is positively correlated with its corresponding intergenic distance (Tuller, Ruppin, Kupiec 2009). This suggests that as genomes compact, the 5′UTRs should become shorter, thereby avoiding overlap with upstream genes, which could have negative consequences for gene expression.

In a recent study, we analyzed the transcripts of a set of 31 *E. cuniculi* genes isolated from spores and meronts to determine TSSs and 5′UTR sizes (Gill *et al*. 2010). We found a marked difference between spore and meront transcripts in terms of the number of TSSs per gene and the lengths of 5′UTRs. On average, spore transcripts had much longer 5′UTRs and had more TSSs than meront transcripts (Gill *et al*. 2010). The 15 intron-containing genes tested, mostly ribosomal protein-coding genes (RPGs), stood out as having very short 5′UTRs, ranging from 0 to 16 bp. However, we could not speculate about the relationship between short 5′UTRs and intron possession because only one intron-lacking RPG was examined raising the possibility that short 5′UTRs could be associated with RPGs in general. Recently, however, the number of annotated introns in *E. cuniculi* has increased to 34 (Lee *et al*. 2010), and 18 of these reside in non-RPGs. Therefore, in the current study the 5′UTR lengths of all *E. cuniculi* RPGs and all intron-containing genes are compared alongside an equivalent set of random genes (non-RPGs) to examine potential relationships between 5′UTR length and gene categories.

2.2 Results

2.2.1 *Encephalitozoon cuniculi* 5′UTRs are very short

Given that the *E. cuniculi* genome has undergone loss and compaction of coding and non-coding DNA, it is very likely that this compaction would also be evident in the transcriptome. In support of this, our analysis revealed that a high frequency of *E. cuniculi* genes have very short 5′UTRs (Table 2.1, Figure B.1 in Appendix A.1). The mean 5′UTR length for all 155 genes is 17 bp, with a standard deviation of 49. The median length is 3 bp, indicating that a considerable proportion of the 5′UTRs are extremely small.

It is possible that there is an over-representation of extremely short 5′UTRs in the *E. cuniculi* gene set examined because the intron-containing genes and RPGs were initially targeted for analysis based on previous results indicating their small 5′UTR sizes (see further discussion related to gene categories below). However, even in the set of intron-lacking, non-RPGs, 48 of 73 (65.8%) possess 5′UTRs shorter than 10 bp (Fig. 2.1). This indicates that a high frequency of *E. cuniculi* 5′UTRs are extremely short—quite possibly the shortest known.

The large difference between mean and median 5′UTR lengths reveals a biased distribution of lengths, in which there are many very short 5′UTRs and only a few that are much longer (Fig. 2.3 in Appendix A.1). This is indicative of a strongly right- or positively skewed distribution (Fig. 2.3 in Appendix A.1), which is also seen in the distribution of 5′UTRs of *S. cerevisiae* (Tuller, Ruppin, Kupiec 2009). Although rare, Table 1 indicates that some genes, especially those with longer 5′UTRs, have more than one 5′UTR length indicating multiple start sites. This is in line with previous results in *E. cuniculi* and *S. cerevisiae* where multiple start sites have been detected (Gill *et al*. 2010; Zhang and Dietrich 2005).

2.2.2 Exclusively short 5′UTRs for RPGs and intron-containing genes

Previous experiments in *E. cuniculi* have shown that intron-containing genes have very short 5′UTRs and that a set of randomly selected genes have much more variety in 5′UTR lengths (Gill *et al*. 2010). The current analysis expanded the dataset significantly and tested all intron-lacking RPGs in order to differentiate between the pattern of 5′UTR lengths of intron-containing RPGs and those lacking introns. The mean and median 5′UTR lengths of the intron-containing genes are 5 and 4 bp, those of the intron-lacking RPGs are 4 and 1 bp, respectively, and those of the intron-lacking non-RPGs are 32 and 3 bp, respectively. It is clear that both intron-containing genes and RPGs exclusively possess short 5′UTRs, while the intron-lacking, non-RPG category of genes includes several genes with much longer 5′UTRs, resulting in a considerable increase in the mean value (Fig. 2.2).

2.3 Discussion

The degree of 5′UTR reduction in *E. cuniculi* is evident when comparisons are made to other compact genomes: the average 5′UTR length in *S. cerevisiae* is 83 bp, with a standard deviation of 84 (Tuller, Ruppin, Kupiec 2009). The average 5′UTR lengths recorded for plants, animals, and other fungi are much longer, ranging from 143 to 246 bp (Pesole *et al*. 1997). Furthermore, a transcriptome-wide analysis of *S. cerevisiae* found that only 5% of genes have putative start codons <10 bp from the TSS (Nagalakshmi *et al*. 2008), whereas, in *E. cuniculi*, 119 of 155 (76.8%) genes tested have 5′UTRs of <10 bp, and 133 of 155 (85.8%) have 5′UTRs that are smaller than 20 bp.

There are several possible explanations for the abundance of short 5′UTRs. First, genomic compaction could have forced promoter elements and transcription machinery binding sites extremely close to the translation start codon, resulting in a shorter UTR. Compact

genomes, such as in Fugu, Cryptosporidium, and Giardia, tend to have short intergenic regions and this results in more tightly packed promoter regions in general (Abrahamsen *et al*. 2004; Franzén *et al*. 2009; Venkatesh, Gilligan, Brenner 2000). Although one analysis has found several potential transcription regulation motifs in various microsporidians (Peyretaillade *et al*. 2009), a more exhaustive investigation will be needed to determine all motifs used as promoters and their positions in the genome. Second, a loss of trans-acting regulatory protein-encoding genes could have allowed the intergenic spaces to shrink over time as fewer binding sites are being used. It has been suggested that during the reduction of the *E. cuniculi* genome interaction networks became simplified as genes were lost and shortened (Katinka *et al*. 2001). Following this reduction, intergenic regions that were no longer used to regulate expression would not be retained by selective pressure and could therefore be lost. It is also possible that reducing intergenic regions could have forced control elements into upstream genes, which may be one cause of the frequent overlapping transcription. Third, perhaps the need for complex regulation simply diminished as a result of an intracellular lifestyle and gave rise to a more constant level of transcription and translation that did not utilize extensive UTR regions. There may be two transcriptional states in microsporidia, an active state in the meront and a nearly inactive state in the spore. With only 2,000 genes in *E. cuniculi*, there could simply be continuous transcription of all genes in the meront, except those required to transition between spore and meront stages because these genes must have strict temporal regulation (Katinka *et al*. 2001). However, it seems hard to believe that nearly every gene encoded by an organism could be expressed at roughly the same rate. Therefore, some unknown factors must be involved in modulating expression at least to some degree. The *E. cuniculi* genome does encode a core set of approximately 50 transcriptional regulatory proteins (Katinka *et al*. 2001), however, this is much

reduced compared with the sets found in other fungal genomes. Further analysis of the cis- and trans-acting factors involved in gene regulation in microsporidia should help shed light on the evolution of these distinct types of 5′UTRs.

An analysis of human housekeeping genes shows that they have shorter exons, introns, and UTR regions than other genes (Eisenberg and Levanon 2003). Separate analyses have found that housekeeping genes are not special per se, but that highly expressed genes in general are more compact (Li, Feng, Niu 2007). Included in the gene ontology group of highly expressed housekeeping genes are the RNA-interacting genes (including RPGs), suggesting that RPGs could be compact in most or all eukaryotes (Anonymous; Eisenberg and Levanon 2003; Tuller, Ruppin, Kupiec 2009). The abundance of genes with short 5′UTRs in *E. cuniculi* suggests that more than just a small subset of genes may be highly expressed. Perhaps this makes sense because, as parasites, microsporidia likely require rapid cellular and genomic replication during infection and would therefore benefit from a "housekeeping" level of expression for most genes. The greatest impact of 5′UTR length could be seen at translation; the lack of a lengthy 5′UTR could allow for more rapid translation because regulatory proteins would not be able to bind upstream of the start codon and secondary structures that block translation initiation would not form. The primary sequence around the initiation codon of mRNAs is known to influence the efficiency of the 43S ribosomal complex recognizing the translation initiation codon (Baim and Sherman 1988; Day and Tuite 1998; Kozak 1986). However, with many of the transcripts analyzed here there are very few or even no nucleotides upstream of the start codon, leaving the possibility that a novel interaction takes place during translation initiation in *E. cuniculi*. Many of the core initiation factors are present, however, there are hundreds of genes whose proteins have unknown functions due to the divergent nature of *E. cuniculi* sequences (Katinka *et al*. 2001).

The ability to demonstrate such an interaction awaits the development of procedures for manipulating these parasites at the cellular and genetic levels.

The positively skewed distribution of 5′UTR lengths leads us to speculate on the evolution of the much longer and less common sequences, and their potential roles in gene expression. Given some of the hypotheses for the presence of short 5′UTRs provided above, it seems possible that some of the long 5′UTRs may simply be the result of mutational biases. Mutation rates are known to vary across genome regions based on characteristics such as base composition (Wolfe, Sharp, Li 1989) and different intergenic regions could undergo different rates of mutation. Another possibility is that the long 5′UTRs contain internal ribosome entry sites (IRES), which are used for cap-independent translation (Xia and Holcik 2009). Many viruses and some eukaryotic transcripts use IRES motifs to bypass the canonical 7mG cap-dependent translation mechanism, also referred to as the scanning mechanism, used by the vast majority of eukaryotic genes including those of *E. cuniculi* (Katinka *et al*. 2001; Van Eden *et al*. 2004). This could be useful for microsporidians as they enter a dormant state with inactive cap-dependent translation, and still require the translation of select genes. Although no IRES-specific motifs or proteins have been identified in the *E. cuniculi* genome, there are many genes with unknown functions, and it is possible to have IRES activity where the RNA structure promotes ribosome binding directly (Fitzgerald and Semler 2009). It has also been postulated that the long 5′UTRs could be playing a structural role rather than an informational one, and that perhaps they are involved in tethering polyribosomes for storage during the dormant phase (Gill *et al*. 2010). As mentioned above, the need for complex regulation could result in the retention of long 5′UTR regions. For example, the spore wall and polar tube proteins are flanked by much longer than average intergenic regions, providing space for regulatory factor binding sites and allowing for

longer UTR regions to be transcribed without overlapping expression (Corradi, Gangaeva, Keeling 2008). This supports the hypothesis above, as these two genes that are obvious candidates for complex regulation have long 5′UTRs (>100 bp), whereas the majority of other genes have much shorter 5′UTRs (Corradi, Gangaeva, Keeling 2008). The final and, perhaps, most likely scenario is that the long 5′UTRs are playing a role in down-regulating expression by reducing translation rates and increasing the time needed for transcription. In yeast, the folding of long 5′UTRs can regulate translation by affecting initiation through the modulation of ribosome accessibility (Ringner and Krogh 2005; Tuller, Ruppin, Kupiec 2009). In essence a long 5′UTR is more likely to fold into stable secondary structures that block access to the ribosome, thereby stalling translation initiation. Another consequence of the increased length is the time necessary for transcription. Because the average protein-coding gene length in *E. cuniculi* is only 1,080 bp, a difference in UTR length of a few hundred base pairs can account for up to ~30% of the transcript length and have a significant effect on gene expression (Katinka *et al*. 2001).

As seen in Fig. 2.2, intron-containing genes and RPGs have short 5′UTRs, while the intron-lacking, non-RPG category includes genes with much longer 5′UTRs. Could there be an explanation for the consistently short 5′UTR lengths observed in intron-containing genes and RPGs? In *S. cerevisiae*, several classes of genes have shorter than average 5′UTRs, and the gene category comprising rRNA processing genes and RPGs is the most significant among them (Hurowitz and Brown 2003; Tuller, Ruppin, Kupiec 2009). As mentioned above, short 5′UTRs may be involved in the increased levels of expression seen in housekeeping genes, such as those involved in transcription and splicing. Housekeeping genes are believed to require less post-transcriptional regulation than tissue-specific or environment-induced genes, and therefore can

dispense with long UTR regions (David *et al*. 2006; Tuller, Ruppin, Kupiec 2009). However, this does not explain why the set of intron-containing genes in *E. cuniculi* is entirely devoid of long 5′UTRs, because several of these genes function in a variety of pathways that would not be considered housekeeping. Perhaps the best explanation is a link between transcription and splicing.

The rate of successful splicing could be involved in regulating gene expression, and it may be modulated by associated factors that are also involved in transcription. This could allow several levels of regulation to exist using only proteins already present in the transcription and splicing machinery. For example, there is experimental evidence of introns increasing gene expression levels by improving mRNA stability (Luehrsen and Walbot 1991), as well as experiments showing regulatory feedback of splicing on RPG expression (Chung and Perry 1989; Dabeva, Post-Beittenmiller, Warner 1986; Dabeva and Warner 1993; Russo *et al*. 2010; Vilardell and Warner 1997; Warner *et al*. 1985). Further, physical links have been found between the transcription and splicing machinery, and pre-mRNA splicing has been shown to occur co-transcriptionally (reviewed in (Bentley 2002; Kornblihtt et al. 2004; Proudfoot, Furger, Dye 2002). In vitro and in vivo studies have found that the C-terminal domain of RNA polymerase II participates in splicing and at other stages of pre-mRNA maturation (Kornblihtt *et al*. 2004; McCracken *et al*. 1997; Neugebauer 2002). However, whether splicing occurs co- or post-transcriptionally depends on the position of the intron and the length of the transcript (Kornblihtt *et al*. 2004). The close proximity of a yeast intron to the 5′ end of a gene and the length of the second exon are major determinants of whether co-transcriptional splicing will occur (Kornblihtt *et al*. 2004). Given the positional biases of *E. cuniculi* introns, co-transcriptional splicing is likely occurring and could be playing a regulatory role in gene

expression. However, determining the potential mechanisms of regulation awaits further investigation.

2.4 Conclusions

We have shown that a high frequency of *E. cuniculi* transcripts have very short 5′UTRs. Highly expressed eukaryotic genes tend to have shorter than average 5′UTR regions, suggesting that *E. cuniculi* may have a large set of highly expressed genes. The longer *E. cuniculi* 5′UTRs might only be used to control regulation of a small subset of genes that require cell cycle or environment-induced regulation. Two categories of genes, the intron-containing genes and RPGs, have exclusively short 5′UTRs. This indicates that intron-containing genes, regardless of gene category, may be highly expressed. However, the presence of an intron may allow for regulation at multiple levels, suggesting links between transcription and splicing in *E. cuniculi*.

2.5 Materials and methods

2.5.1 RNA extraction and cDNA synthesis

Encephalitozoon cuniculi (Genotype II)-infected rabbit kidney (RK) tissue was a generous gift from Dr. Elizabeth Didier (Tulane University, Louisiana). Total RNA was extracted from RK13 cells 48 h post-infection using the Trizol reagent (Invitrogen, Carlsbad, CA) after grinding under liquid nitrogen. Meront cDNA of *E. cuniculi* was prepared using 5′ RNA ligase-mediated rapid amplification of cDNA ends (RLM-RACE; Ambion, Austin, TX), following the manufacturer's protocol exactly. The RLM-RACE protocol is designed specifically to amplify only full-length, capped mRNA.

2.5.2 Determining 5′UTR lengths

The 5′RACE cDNA fragments were amplified by nested-polymerase chain reaction (PCR) using forward primers specific to the 5′-adapter sequence and gene-specific reverse

primers. A list of primers for all genes analyzed in this study is provided in Table 2.2 in Appendix A.1. All PCR reactions used a 55 °C annealing temperature and 1.5 min extension time. The inner nested primer was labelled at the 5′-end with a 6-carboxyfluorescein (6-FAM) fluorescent molecule (Integrated DNA Technologies, San Diego, CA). Polymerase chain reaction products were diluted 1:20 and analyzed by capillary electrophoresis as described previously (Gill *et al*. 2010; Jorgenson and Lukacs 1983), using an ABI 3730S with GS600LIZ size standard (UBC NAPS, Vancouver, British Columbia, Canada). Product size was determined using Peak Scanner v1.0 software (Applied Biosystems, Carlsbad, CA). The 5′UTR lengths were calculated by subtracting the RACE-adapter sequence and the distance from the TSS (AUG) to the primer sequence, from the total product size. The UTR lengths determined by capillary electrophoresis were also corroborated by visualizing the products on 1.5% agarose gels. The method of detection post-amplification for intron-containing genes was done by cloning (TOPO TA kit, Invitrogen) and sequencing (BigDye3.1, ABI) as opposed to capillary electrophoresis. In multiple independent comparisons of these two detection methods, the fragment size results have been within 5 bp of one another, which is within the range of error of the capillary electrophoresis method (Corradi, Burri, Keeling 2008; Corradi, Gangaeva, Keeling 2008; Gill *et al*. 2010). This corroboration between sequencing data and capillary electrophoresis product size data and the fact that sequencing data did not show truncated mRNA, provides evidence that no degraded or shortened fragments are being amplified during the RLM-RACE protocol and that capillary electrophoresis is a valid alternative to sequencing (Corradi, Burri, Keeling 2008; Corradi, Gangaeva, Keeling 2008; Gill *et al*. 2010).

2.5.3 Gene selection

Transcripts for the following genes were compared in this study: all RPGs, a total of 67, where 16 contain introns; all intron-possessing genes, 16 RPGs, 15 non-RPGs; and a set of intron-lacking, non-RPGs with expression confirmed by proteomic analysis (Brosson *et al*. 2006). New 5′UTR data from this study include: 51 RPGs of which three could not be detected despite several attempts to amplify (L23A, L24, S5), and three were duplicates of other RPGs (L23A, L35, L35); and 43 intron-lacking, non-RPGs, which were grouped together with 31 genes of this category analyzed previously (Gill *et al*. 2010). The 5′UTRs from all new intron-containing genes were examined and added to existing data (Gill *et al*. 2010; Lee *et al*. 2010).

Table 2.1: A list of the names and lengths of 5' untranslated regions (UTRs) of all 155 genes of *Encephalitozoon cuniculi* tested.

Gene name	Gene ID	UTR1	2	3	4	5
RPGs (ribosomal protein-coding genes)						
RPL3	ECU03_1220	2				
RPL3	ECU09_1000	2				
RPL4	ECU08_0830	1	17			
RPL7	ECU03_0950	2				
RPL7A	ECU02_0750	0				
RPL8	ECU01_0310	0				
RPL9	ECU02_0800	16				
RPL10	ECU08_1570	0				
RPL10A	ECU05_0600	0				
RPL12	ECU08_2010	0				
RPL13	ECU03_0320	0				
RPL13A	ECU04_1380	0				
RPL15	ECU11_1380	0				
RPL17	ECU07_1410	0				
RPL18	ECU03_1490	5				
RPL18A	ECU08_0600i	41				
RPL21	ECU05_0900	1	3			
RPL22	ECU04_0740	8				
RPL23	ECU08_1160i	0				
RPL24	ECU02_0810	2				
RPL26	ECU08_0370	0	1			
RPL27	ECU04_0330	1				
RPL30	ECU05_1490	2				
RPL31	ECU03_0230	1				
RPL32	ECU04_1310	2	11			
RPL34	ECU03_0710	3	5			
RPL35	ECU11_2060	0				
RPL35A	ECU02_0900	0				

Gene name	Gene ID	UTR1	2	3	4	5
RPL36	ECU06_1120	0				
RPL44	ECU10_1300	4				
RPS2	ECU07_1700	0	3			
RPS3	ECU09_1250	3				
RPS4	ECU08_0870	0				
RPS7	ECU11_0780	14				
RPS9	ECU05_0920	0				
RPS11	ECU04_0640	0				
RPS12	ECU01_0920	0				
RPS13	ECU08_1060	0	7	10		
RPS14	ECU03_0650	1	4			
RPS15	ECU08_0350i	25				
RPS15A	ECU09_1350	0	4			
RPS16	ECU03_0310	0				
RPS18	ECU06_1110	0	2			
RPS19	ECU11_1620i	4				
RPS20	ECU11_0720	0				
RPS23	ECU10_0400	0				
RPS25	ECU08_1040	0				
RPS25	ECU08_1070	0				
RPS27	ECU04_1015	3				
RPS28	ECU09_1275	0	5			
RPS29	ECU04_0125i	0				
Intronless non-RPGs						
Gamma glutamyl transpeptidase	ECU05_1240	33	111	305	327	
Hypothetical protein	ECU04_1660	310				
PolyA binding protein 2	ECU10_1110	265				
Hypothetical protein	ECU07_0270	2	176			
Hypothetical protein	ECU09_1470	81	84	157		
Hypothetical protein YG22	ECU09_0180	6	30	100	108	152
Hypothetical protein	ECU03_0160	63	126			
Hypothetical protein	ECU04_1670	11	99			

Gene name	Gene ID	UTR1	2	3	4	5
Vacuolar ATP synthase-F	ECU03_0305	3	28	97		
Guanosine diphosphatase	ECU07_1260	13	15	74		
Glutaredoxin	ECU08_1380	68				
Hypothetical protein	ECU07_0120	3	30	54		
GTP-binding protein (RAB6)	ECU09_0170	47				
Similarity with WD-repeat proteins	ECU07_0260	5	44			
CDP-diacylglycerol synthase	ECU05_1250	9	29	42		
Hypothetical protein	ECU10_1070	39				
Chromobox protein	ECU03_0810	5	16	23		
Hypothetical protein	ECU08_1730	22				
Hypothetical protein YG22	ECU11_0670	21				
Hypothetical protein	ECU01_0250	0	20			
HSP-related 70 kDa protein	ECU03_0520	5				
Hypothetical protein	ECU03_0530	3				
Translation elongation factor 2	ECU11_1460	9				
Transport protein SEC13	ECU11_1450	3				
Hypothetical protein	ECU11_1390	12				
Ser/Thr protein phosphatase (PP11)	ECU11_0660	3	8	15		
ATP-dependent DNA-binding helicase	ECU02_1090	9				
Hypothetical protein	ECU07_0200	1				
Hypothetical protein	ECU07_0210	8				
Hypothetical protein	ECU07_1250	2				
Hypothetical protein	ECU08_1220	3				
Hypothetical protein	ECU08_1210	11				
Coatomer complex beta	ECU08_1100	8				
Guanine nucleotide binding protein beta	ECU08_1110	9				
BOS1-like vesicular	ECU07_1620	9				
Putative protein with Mut T domain	ECU07_1630	2				
Ubiquitin conjugating E2 (C7)	ECU01_1010	0				
GTP-binding protein (SAR1)	ECU05_0090	0				
Hypothetical protein	ECU07_0400	2	3			
Hypothetical protein	ECU09_1820	3				

Gene name	Gene ID	UTR1	2	3	4	5
Actin	ECU01_0460	0				
RNA polymerase II beta	ECU10_0250	0				
Glyceraldehype-3-P dehydrogenase	ECU07_0800	2				
Nucleoside diphosphate kinase A	ECU06_1530	1				
G6P 1-dehydrogenase	ECU08_1850	0				
UTP G1P uridyltransferase	ECU03_0280	0				
Septin	ECU11_1950	3				
TCP1 alpha subunit	ECU03_0220	0				
HSP90	ECU02_1100	0				
Peptidylpropyl *cis–trans* isomerase	ECU08_0470	0				
Phosphmannomutase	ECU05_0260	0				
Hypothetical protein	ECU11_1270	0				
Hypothetical protein	ECU09_1400	0				
Hypothetical protein	ECU08_1280	0				
Hypothetical protein	ECU05_0110	13				
Hypothetical protein	ECU02_0150	0				
Hypothetical protein	ECU01_0440	0				
Hypothetical protein	ECU01_0420	0				
Inorganic pyrophosphatase	ECU10_0340	0	1			
Glucosamine F-6-P aminotransferase	ECU07_1280	0				
Ser/Thr protein phosphatase (PP5)	ECU05_0440	2				
Arg/Ser-rich pre-mRNA splicing factor	ECU05_1440	0				
Cleavage stimulation factor	ECU08_0380	0				
Cysteinyl-tRNA synthetase	ECU08_0490	3				
TCP1 delta subunit	ECU02_0520	0				
Hypothetical protein	ECU09_1370	0				
Hypothetical protein	ECU11_1210	6				
Fructose bisP aldolase	ECU01_0240	6	11			
Trehalose phosphate synthetase	ECU01_0800	0				
Hypothetical protein	ECU09_1880	8				
TCP1 Eta subunit	ECU10_0630	2				
Proteosome alpha-type subunit (PRC5)	ECU07_1420	3				

Gene name	Gene ID	UTR1	2	3	4	5
Hypothetical protein	ECU02_0390	1				
Intron-containing genes						
RPL5	ECU06_0900i	3	6			
RPL6	ECU08_1780	3				
RPL11	ECU02_0610	3				
RPL19	ECU06_1080	4				
RPL27a	ECU10_0990	3				
RPL37	ECU07_1460	5				
RPL37a	ECU07_1005	3	4			
RPL39	ECU09_0395	2	4	5		
RPS3a	ECU05_0250	5				
RPS6	ECU05_0670	2	4			
RPS8	ECU02_0880	3				
RPS10	ECU04_1355	10				
RPS17	ECU02_0770	4	5			
RPS24	ECU10_1570	4	5			
RPS26	ECU06_1445	3	5			
RPS30	ECU10_1575	3	7			
Unknown7 (7p4)	Unannotated	2	3	4		
PolyA-binding protein	ECU09_1470	11	14			
Ubiquitin activator	ECU09_0860	0	2			
Unknown8-a/b (8m129)	Unannotated	0	4	7		
Unknown3a/b (8p118)	ECU08_1030	5	6			
Adenylate kinase (6p86)	ECU06_0650	1				
Sec61alpha	ECU09_0139	1				
Unknown1 (4p614)	Unannotated	1				
Unknown2 (8p140)	Unannotated	3	5			
Unknown4 (11p134)	ECU11_1060	4				
Unknown5 (10m152)	Unannotated	9	10			
Vacuolar sorter (7p204)	ECU07_1710	0	4			
Unknown6 (4p173)	Unannotated	1	13	16		
pgs-a/b (11p111)	ECU11_0850	1				

Gene name	Gene ID	UTR1	2	3	4	5
Unknown9 (6m617)	Unannotated	3				

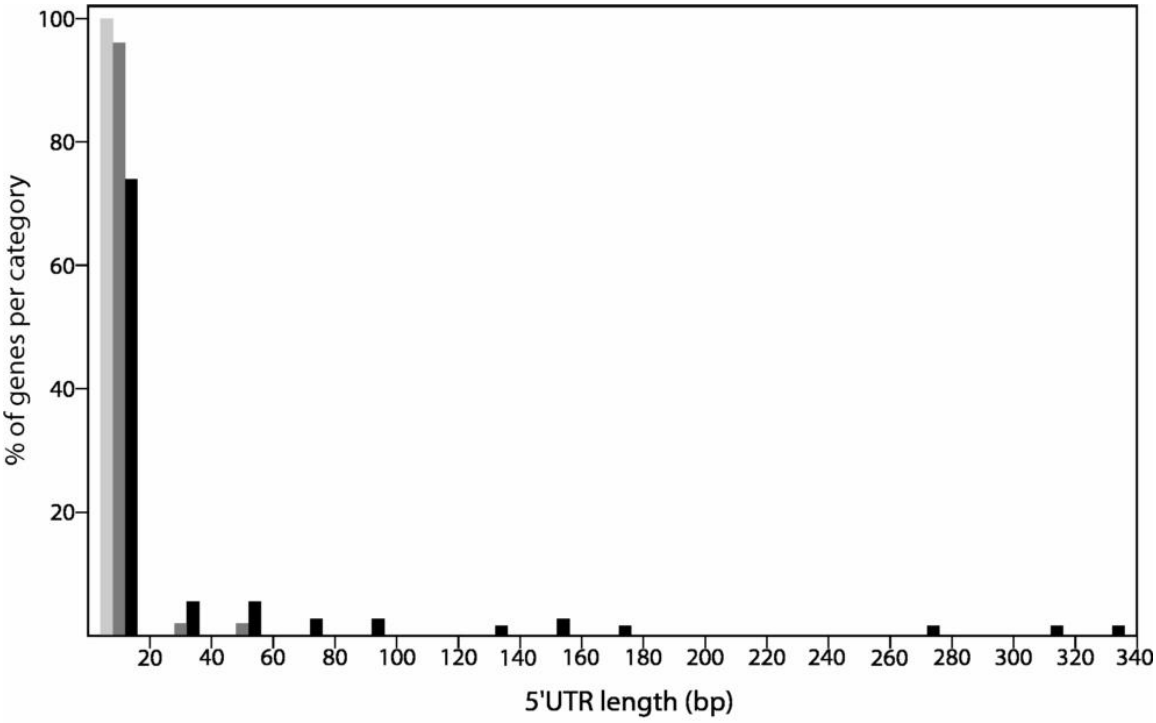


Figure 2.1: Bargraph of 5'UTR lengths in *Encephalitozoon cuniculi*

Shaded bars represent the percentage of genes of *Encephalitozoon cuniculi* from each category that fall within a particular range of 5' untranslated region (UTR) sizes. Light gray bars are intron-containing genes, dark gray bars are intron-lacking ribosomal protein genes, and black bars are intron-lacking non-ribosomal protein genes. See Figure A.1 in Appendix A for an expanded view of 5'UTR data of less than 20 bp.

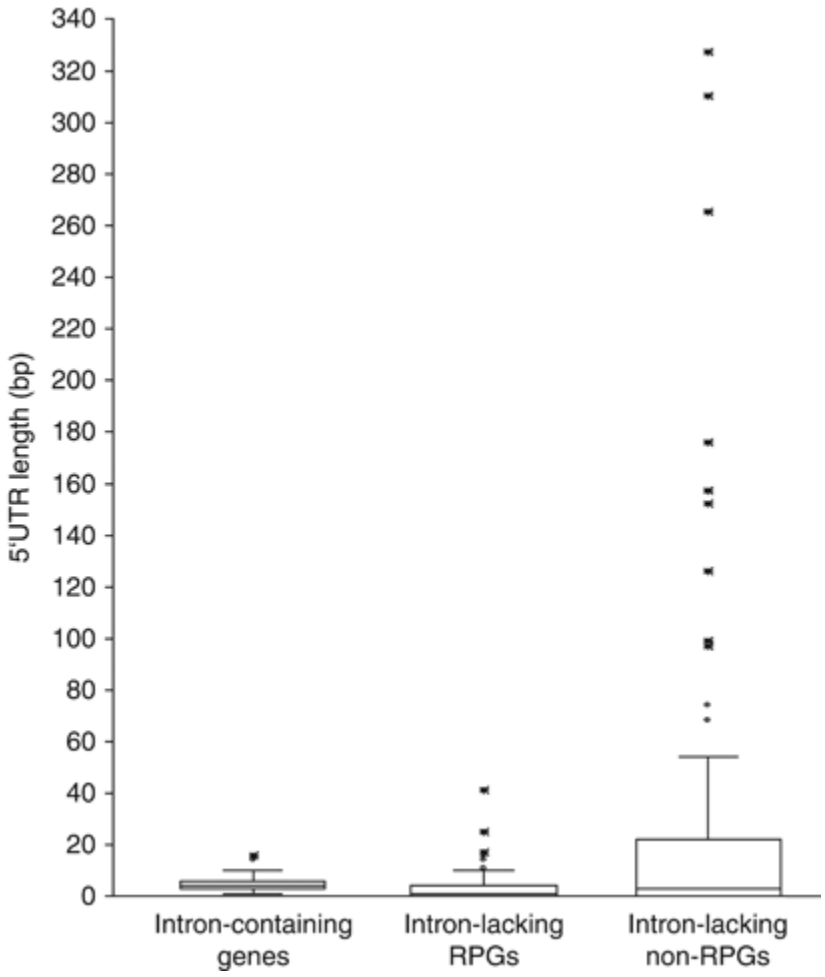


Figure 2.2: Distribution of 5'UTR lengths in *Encephalitozoon cuniculi*

A box and whisker plot showing the distribution of lengths of 5' untranslated regions (UTRs) of three categories of genes of *Encephalitozoon cuniculi*. The box represents the distribution of 50% of the data points (i.e. 25-75 percentile), with 25% above and below the median value (line within the box). Whiskers represent data points up to 1.5 times the box height from the top and bottom of the box. Values up to 3 times the box height from the box are depicted as circles, while those further away are depicted as stars.

Chapter 3: Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of pre-mRNA splicing in a reduced eukaryote

3.1 Introduction

Microsporidia possess among the smallest, most compact eukaryotic genomes known (Corradi *et al*. 2010a). Microsporidia are intracellular parasites that alternate between a thick-walled, extracellular stage (spore) and intracellular stages (meronts, sporonts, and sporoblasts). When triggered, a specialized structure called the polar tube shoots out of the spore and, upon contacting a host cell, creates a passageway into the host (Delbac and Polonais 2008). If a host cell is infected, meronts will proliferate, then undergo sporogony before being released from the host cell. The mammalian parasite *Encephalitozoon cuniculi* typically infects humans with compromised immunity due to HIV-infection or immune-suppressive therapy (Katinka et al. 2001; Vavra, J. & Larsson, J.I.R. 1999). *E. cuniculi* was the first microsporidian to have its genome completely sequenced, and at 2.9 Mb this highly reduced genome possesses many unusual features. It has a reduced coding capacity, encoding less than two thousand protein-coding genes, most of which are shorter than their homologs in yeast (Katinka *et al*. 2001). It lacks genes for several biosynthetic pathways and components of the energy-producing tricarboxylic acid cycle. This stripped down genome provides an opportunity to study cellular processes that generally require large, complex sets of components, yet in microsporidia such complexity is reduced, while retaining function.

The spliceosome is a large macromolecular machine that is responsible for removing nuclear spliceosomal introns from pre-mRNA via two transesterification reactions (Jurica and

Moore 2003; Wahl, Will, Luhrmann 2009). In humans, this complex rivals the size of the bacterial ribosome and contains hundreds of protein components and five small nuclear RNAs (snRNAs). Conversely, *E. cuniculi* is only predicted to possess 30 spliceosomal proteins (Katinka *et al*. 2001). Such reduced eukaryotes could hold important information about intron and spliceosome evolution as they harbor so few spliceosomal introns (fewer than 40), and some microsporidia are completely devoid of introns and splicing machinery (Cuomo *et al*. 2012; Keeling *et al*. 2010).

In a previous study we assessed differences in *E. cuniculi* transcription and spliced transcript levels between intracellular and extracellular life stages (Gill *et al*. 2010). We found that transcripts have much longer untranslated regions (UTRs) and more transcription start sites in the spore stage compared to the intracellular stage. Splicing appears to take place exclusively in the intracellular stage leaving long, unspliced transcripts in the spore, that may play a structural rather than an informational role (Gill *et al*. 2010). Although pre-mRNA splicing occurs, we found no evidence for alternative splicing or mis-splicing (Lee *et al*. 2010). We also found that *E. cuniculi* intron-containing genes have exclusively short 5'UTRs and that, on average, intracellular stage 5'UTR lengths are among the shortest known (Grisdale and Fast 2011). Another unusual feature of microsporidian transcription is the presence of overlapping transcripts in the extracellular stage. *E. cuniculi* and the distantly related microsporidian *Antonospora locustae* were both found to have overlapping transcripts (Corradi, Gangaeva, Keeling 2008; Williams *et al*. 2005). However, transcripts in the former often initiate in upstream genes, while those in the latter often terminate in downstream genes (Corradi, Gangaeva, Keeling 2008; Williams *et al*. 2005). These peculiarities of microsporidian molecular biology and the differences in transcripts between extracellular and intracellular life stages led us

to conduct a comprehensive investigation of the parasite's transcriptome during intracellular stages.

Using Illumina HiSeq technology we performed deep RNA sequencing of the *E. cuniculi* transcriptome 24 hr, 48 hr, and 72 hr post-infection. This allowed us to assess spliced transcript and gene expression levels at multiple time points, find novel transcribed regions (NTRs), and improve gene annotations. RNA-seq is an ideal method for examining transcriptomes, as it is relatively unbiased, has greater sensitivity than hybridization methods such as microarrays, and produces high coverage of transcripts (Agarwal *et al*. 2010; Marioni *et al*. 2008; Wang, Gerstein, Snyder 2009; Wilhelm *et al*. 2010). We analyzed splicing at all 37 splice junctions to assess the role of these few remaining *E. cuniculi* introns, determined gene expression levels of all annotated genes, found several novel ORFs and, in general, increased our understanding of the dynamic transcriptomes of these unusual parasites.

3.2 Results and discussion

Genomic analyses of microsporidian species have revealed a number of unusual features that are distinct from other eukaryotes. To date, the microsporidia examined have either done away with introns and splicing machinery entirely, or retain very few of each. How these introns are spliced with greatly reduced machinery, and why so few are retained are questions that pertain both specifically to the evolution of these parasites and, more generally, to intron splicing in eukaryotes. In this study, we present the first transcriptomic analysis of *E. cuniculi*. Intracellular stage genotype 2 *E. cuniculi* was examined at three time-points: 24 hr, 48 hr, and 72 hr after infection of RK13 cells (rabbit kidney fibroblast cell line). A total of 525.9 million reads were produced (Table 1), 40.6 million (7.7%) of which aligned to the *E. cuniculi* genotype 2

reference genome (GenBank accession AEWQ01000000) (Pombert *et al*. 2013), and 273.5 million (52.0%) of which aligned to the rabbit host (Oryctolagus cuniculus reference genome: GenBank accession AAGW02000000). We saw no evidence of cross mapping between host and parasite genomes, as expected, owing to the availability of reference genomes for both organisms and the high level of divergence between them (data not shown). The number of reads mapping to *E. cuniculi* at 24 hr, 48 hr, and 72 hr post-infections were 13.9 million, 17.5 million, and 9.3 million, respectively. This was sufficient coverage to assess splicing and examine gene expression levels at each time-point in order to address questions regarding intron function and evolution, as well as the expression of pathogenesis-related and microsporidia-specific genes.

3.2.1 Identification of novel transcribed regions

We annotated eleven previously unidentified, transcribed ORFs, three of which have the potential to play a role in pathogenesis. These eleven ORFs are distributed over eight chromosomes. *E. cuniculi* chromosomes were annotated using GLIMMER to find putative ORFs with a minimum length cut-off of either 300 or 150nt (Katinka *et al*. 2001). ORFs were used for BLAST searches followed by protein domain identification. This type of annotation leaves open the possibility of ORFs not being annotated due to their small size or lack of known, conserved domains. In order to find novel ORFs that may have been overlooked by the automated annotation software, we examined each chromosome visually using Integrated Genomics Viewer (Robinson *et al*. 2011) (see Methods for details).

The novel ORF on chromosome 3 (ECU03_0255) is a potential candidate for a pathogenesis-related gene involved in cell entry. Although no clear function for this ORF could be predicted from similarity searches, weak (30%) similarity to a viral entry protein could suggest that the product of this ORF functions in host invasion. We discovered two additional

ORFs that are so far unique to microsporidia, and therefore may play a role relating to their parasitic lifestyle. Novel ORF ECU03_0715 has a clear homolog in *E. hellem*, sharing 72% identity over all 116 amino acids. Although not present in all known microsporidian genomes, this ORF shares similarity with genes of unknown function in *Antonospora locustae* and *Nematocida parisii*, two distantly related microsporidia. A second ORF that appears to be microsporidia-specific is ECU06_0735, which shares 41% identity over 133 of its 146 amino acids with homeobox domain-containing transcription factors in other *Encephalitozoon* species. The products of these ORFs will require functional analysis to ascertain the cellular roles of their microsporidia-specific protein products.

An additional ORF (ECU08_1555) we discovered has no predicted connection to pathogenesis, but may play an important cellular role as it shows similarity to the nucleolar protein NOP10. NOP10 is associated with snoRNAs in ribonucleoprotein complexes that are involved in 18S rRNA production, rRNA pseudouridylation, and are components of the telomerase complex (Henras *et al*. 1998). Additional novel ORFs had very weak similarity to known proteins, and were identified based on transcription signal alone (data not shown). Also, several predicted intergenic regions were transcribed with distinct boundaries but no ORF could be assigned on either strand. These may represent important non-coding RNAs or possibly even unknown selfish genetic elements.

3.2.2 All coding regions are transcribed in intracellular *E. cuniculi*

The expression data revealed that nearly all 1981 genes had detectable levels of expression in all three time-points (see Appendix Table B.1): all 1981 genes were expressed 24 hr post-infection, 1980 genes were expressed 48 hr post-infection, and 1979 genes were expressed 72 hr post-infection. The twenty genes with highest average expression, in descending

order, include spore wall protein 1, RNA-binding domain-containing protein (discussed below), translation elongation factor 1 alpha, actin, histones H2B/H3/H2A/H4, heat shock protein 70, and ribosomal protein L9. The remaining ten genes encode hypothetical proteins with unknown functions. As expected, many highly-expressed genes have housekeeping functions; however, the most highly expressed gene, excluding hypotheticals, is a spore wall protein-encoding gene. This highlights the priority of preparing to form the infective stage, even as early as the first 72 hrs following infection. In summary, essentially all *E. cuniculi* protein-coding genes are expressed during the first three days post-infection in tissue culture.

3.2.3 High frequency of differentially expressed genes in the first 48 hrs

Although nearly all genes are expressed at all time-points, we found an abundance of genes with considerable differences in expression levels between time-points. There were 746 (37.7%) genes differentially expressed between 24 hr and 48 hr post-infection and 867 (43.8%) genes differentially expressed between 24 hr and 72 hr (Figure 3.1A,B). However, between 48 hr and 72 hr there were only 4 genes differentially expressed (Figure 3.1C), all with fairly weak fold changes of less than 0.5. This pattern, where many genes are differentially expressed within the first 48 hrs but not after, has implications for the life-cycle of this parasite, such as the possibility that spore formation begins by 48 hr post-infection.

Evidence from expression data suggests that *E. cuniculi* meronts undergo a shift towards producing spore-related genes by 48 hr post-infection. The ten genes with largest positive and negative fold change between 24 hr and the two subsequent time-points include mostly housekeeping genes and genes encoding hypothetical proteins. An exception to this is polar tube protein 2, whose gene had some of the strongest positive fold changes, both from 24 hr to 48 hr (2.25) and from 24 hr to 72 hr (2.45). Also, the gene encoding polar tube protein 1 showed a

similar pattern, with a fold change of 2.05, while the spore wall protein-encoding gene had a fold change of 1.55. This suggests that expression of spore-related genes increases by 48 hr and spore formation could be taking place, however we did not see evidence of spore-specific transcripts with extended 5′UTRs (Grisdale and Fast 2011), even by 72 hr post-infection. This is in line with previous experiments, which have found a spore related gene to have increased expression between 24 hr and 72 hr post-infection (Taupin *et al*. 2006), and evidence of spore-containing vacuoles beginning at 120 hr post-infection (Fischer *et al*. 2008).

Housekeeping genes are down-regulated after 24 hr, providing evidence that proliferation is taking place very soon after spore germination and likely for a very brief time. Among the ten most strongly down-regulated genes are several ribosomal protein genes, ubiquitin, an RNA polymerase, two novel ORFs, and several hypothetical protein encoding genes. Down-regulation of housekeeping genes after the 24 hr time-point likely occurs because their expression is high upon germination. We also found that, while many ribosomal protein genes have relatively weak changes in expression ratio, they are all negative, further evidence that housekeeping genes as a whole are being down-regulated after 24 hrs. In summary, it seems that spore-specific proteins are produced early in the intracellular life-stage, although spores are likely not formed until after 72 hr post-infection, and housekeeping genes are being down-regulated after 24 hr, possibly as a result of slowing intracellular stage replication rates.

3.2.4 Analysis of pre-mRNA splicing

3.2.4.1 *E. cuniculi* has a reduced spliceosome

Gene annotation in *E. cuniculi* identified just 30 ORFs with similarity to spliceosomal components (Katinka *et al*. 2001), predicting one of the smallest functional spliceosomes known. Several components that are required for viability in yeast are absent in *E. cuniculi*, raising

questions about the necessity of these components, the redundancy built into this pathway, and the flexibility of the spliceosome. Also, one of the five RNA components, the U1 snRNA, has not been identified (Lopez, Rosenblad, Samuelsson 2008). This suggests that splicing may be occurring without a complete U1 complex, which is involved in the key first step of splicing when the intron is recognized and bound at the 5' splice site (Wahl, Will, Luhrmann 2009). The reduction in *E. cuniculi* spliceosome machinery is severe and is likely to have an effect on the splicing reaction, potentially reducing splicing efficiency.

3.2.4.2 Discovery of introns and splice isoforms

The original genome annotation of *E. cuniculi* predicted 16 introns, almost all of which were in ribosomal protein genes (Katinka *et al*. 2001). The number of introns was increased to 34 after a thorough search was performed with a combination of visual and string-search algorithm methods (Lee *et al*. 2010). Many of these new introns were found in non-ribosomal protein-coding genes, which has implications for our understanding of intron retention and evolution in Microsporidia (discussed in (Lee *et al*. 2010)). Ranging in size from 22–76 nt, *E. cuniculi* introns are among the smallest spliceosomal introns found in nature, surpassed only by the miniature introns of *Paramecium tetraurelia* (Russell, Fraga, Hinrichsen 1994) and the Chlorarachniophyte nucleomorph genomes (Gilson and McFadden 1996). All *E. cuniculi* introns have standard GT-AG boundaries, and relatively strict 5' splice site and branch point motifs (see Appendix Figure B.1). This is in line with phylogenetically broad genomic analyses, which have shown that strict splicing motifs are common in intron-poor genomes (Irimia, Penny, Roy 2007; Irimia and Roy 2008). Utilizing the RNA-seq dataset we confirmed that all previously annotated introns are indeed spliced and are bona fide introns. Also, we found one new intron that creates a novel ORF (ECU09_1255), and confirmed splicing of two others that were recently discovered in a

comparison of four *Encephalitozoon* species (Pombert *et al*. 2012). These three recently detected introns were each confirmed with more than a hundred spliced transcripts, as well as having motifs that are characteristic of *E. cuniculi* introns (Appendix Figure B.1).

We have found the first evidence of alternative splicing in a microsporidian parasite. A small proportion of transcripts for three intron-containing genes utilize alternative downstream acceptor sites. Although unexpected to find alternative splicing in such a reduced, streamlined system, the alternative transcripts are so rare that they may represent erroneous splice events. In all cases observed, the alternative isoform represents less than 5% of the reads at the corresponding junction. Despite their low abundance, it is possible that the alternative forms could be utilized as another post-transcriptional regulatory mechanism by inducing rapid decay, as has been hypothesized in *P. falciparum* (Sorber, Dimon, DeRisi 2011). If these transcripts were to induce decay this would help explain the rarity at which we observe them. Unfortunately, we lack the tools needed to manipulate decay rates in microsporidia, and therefore cannot test this hypothesis directly. We also see evidence of alternative intron retention, most notably in ECU11_0850 (Figure 3.2). In this case the upstream intron is spliced at higher levels than the downstream intron, which would result in some transcripts being truncated at the 3' end, but potentially still functional. Since no genes that function in alternative splicing regulation, such as SR protein family genes, have been found in *E. cuniculi*, we suggest that variation in intron motif features are responsible for differing levels of intron retention within a gene. It has been shown previously that modification to intron motifs can affect splicing efficiency (Skelly *et al*. 2009). Therefore, alternative splicing could be playing a minor role in *E. cuniculi* gene expression.

3.2.4.3 Comparative analysis of intron-containing transcripts

We quantified transcript abundance of intron-containing genes to assess levels of intron-retention versus intron removal in order to get a better understanding of the roles of pre-mRNA splicing and RNA decay in *E. cuniculi*. There are several possible scenarios with regards to levels of pre-mRNA splicing and RNA decay. One scenario would be that decay rates are low, and the levels of intron retention or removal are dictated by splicing levels. Another option would be that decay is efficient, creating high levels of spliced transcripts whether or not splicing is efficient, as well. We found that, on average, levels of spliced transcripts in *E. cuniculi* were very low (Figure 3.2). A staggering 30 of 37 introns (81.1%) had less than 50% of transcripts with introns removed, and 22 (59.5%) of these had below 20% spliced transcripts. Levels of intron-lacking, or spliced, transcripts ranged from less than 5% to over 85%, with one particularly interesting outlier at the high end of the range. The gene ECU09_1470, an RNA binding domain-containing protein-coding gene, had previously been noted as unusual for containing the longest *E. cuniculi* intron. In this study we found further reason to examine this gene closely as it had the highest levels of splicing and it was one of the few introns with significant differences in splicing levels between time-points. On the other hand, since all *E. cuniculi* introns contain stop codons or cause frameshifts if not properly removed, it is surprising that the majority of them appear to be spliced at such low levels. For example, over half of the transcripts of thirty of these genes appear to be non-functional because they retain introns. This suggests that decay rates are low (discussed below) and pre-mRNA splicing has a strong influence on levels of transcripts with introns retained or removed.

To assess whether these transcripts were unique to microsporidia or common to parasites and organisms with compact genomes, we performed a similar examination of splicing levels in

a free-living and a parasitic fungus. The transcriptomes of *Saccharomyces cerevisiae* and *Candida albicans* encode 306 and 540 introns, respectively (Bruno *et al*. 2010; Nagalakshmi *et al*. 2008). The introns of both are similar in size, generally in the 50-1000nt range (Mitrovich *et al*. 2007; Spingola *et al*. 1999). Although these fungi possess similarly sized spliceosomes that lack over twenty components found in mammals, they encode more than twice as many components as *E. cuniculi*, and therefore, *E. cuniculi* still represents a model of extreme reduction.

Levels of splicing in both *S. cerevisiae* and *C. albicans* were distinctly different from those observed in *E. cuniculi*, with averages of 80% in *S. cerevisiae* and 95% in *C. albicans* (Appendix Figure B.2). We found that 32 of 46 (69.6%) *S. cerevisiae* introns were spliced at levels above 80%, while 39 of 46 (84.8%) were spliced at levels above 50%. Splicing levels in *C. albicans* were comparable with 39 of 48 (81.3%) spliced at levels above 80%, and 43 of 48 (89.6%) spliced at levels above 50%. Also, a similar analysis of splicing levels has been performed in the relatively reduced parasitic protist *P. falciparum*, the causative agent of Malaria (Sorber, Dimon, DeRisi 2011). The authors found that in this unicellular parasite splicing levels were quite high on average, with a median of five times more spliced reads than intron-retained reads observed (Sorber, Dimon, DeRisi 2011). They also note that only 5.6% of introns were spliced at levels below 50% (Sorber, Dimon, DeRisi 2011). Therefore, spliced transcript levels in *E. cuniculi* are drastically lower than those in both a fungal and a very distantly related protistan parasite, as well as a free-living fungus. This result, along with the fact that the *E. cuniculi* spliceosome is much more reduced than *P. falciparum* and both fungal species, indicates that it may not be the life-style of the organism that is having such an effect on splicing, but the severe reduction of the spliceosomal machinery. If, over evolutionary timescales, the loss of

spliceosomal components resulted in decreases in splicing levels, the reduction of the spliceosome could not have reached its current point unless the levels of intron-containing gene expression were acceptable for cell viability and decay rates increased to compensate for increased intron-containing transcripts. Therefore, the spliceosomal core is likely much smaller than we expect, since mutations in introns and increases in gene expression levels can compensate for decreased splicing levels.

One possible reason for the abundance of transcripts with introns retained is that they could be playing a functional role in gene regulation. For example, several ribosomal protein-coding genes in yeast are known to perform autoregulatory splicing: where the product of the splicing reaction inhibits further splicing by specifically binding to newly made transcripts (Dabeva and Warner 1993; Fewell and Woolford 1999; Li, Vilardell, Warner 1996). Other yeast genes have their splicing regulated by environmental stress, such as amino acid starvation (Pleiss *et al*. 2007), or in conjunction with the meiotic cycle (Engebrecht, Voelkel-Meiman, Roeder 1991). We found 11 of 37 junctions with significant differences in splicing levels between time-points, most with relatively modest changes (Figure 3.2). Interestingly, one of the few genes with two introns had significant changes in splicing in both introns, including the largest change (30%), and high variability in levels between introns (Figure 3.2). This provides evidence that splicing may be playing a regulatory role. However, even with nearly a third of intron-containing genes showing differences in splicing levels over the course of infection, nearly all changes are too small to warrant strong evidence of regulatory splicing. Also, we failed to find any strong compensatory role of splicing to moderate expression levels of ribosomal genes, in order to balance their relatively high levels of variability. The low levels of splicing observed do not

seem to be the result of regulation at the level of splicing in most cases, however, the splicing patterns of a few genes are indicative of regulation and will require further examination.

Another plausible explanation for the elevated levels of intron-retained transcripts is that RNA decay may not be functioning efficiently in *E. cuniculi*. Since all *E. cuniculi* introns either contain stop codons or induce frameshifts that result in downstream pre-mature stop codons, intron retention should induce transcript degradation by an RNA decay pathway. Metabolic pathways are generally reduced in *E. cuniculi* (Katinka *et al*. 2001), so complete RNA decay pathways would not be expected. However, *E. cuniculi* appears to have retained a small number of decay proteins, encoding ORFs with similarity to key players including Upf1, Dcp2, Dis3, Dhh1, Ccr4, and Nmd5 (Appendix Table B.2). It is likely that these few decay proteins have evolved to function in the absence of their canonical reaction partners, similar to the spliceosome and DNA repair system (Gill and Fast 2007), as the cell would presumably not be able to function properly without RNA degradation. However, as we predict with spliceosomal functioning, there may be a significant reduction in decay efficiency that could play a part in increasing the proportion of unspliced transcripts present. Yet, to invoke reduced RNA decay as the sole source of these results, decay would have to be very inefficient indeed - a situation that seems unlikely given that no other obvious abnormalities are observed in the transcriptome. Although a formal possibility, it seems unlikely that decay alone is the cause of the high levels of unspliced transcripts. Therefore, the loss of spliceosome components is likely the cause of reduced splicing activity, and in combination with low decay rates, results in a large proportion of unspliced transcripts.

3.3 Conclusions

Assessing the transcriptome of *E. cuniculi* allowed us to improve the genome annotation, uncover novel transcribed regions that could play a role in pathogenesis, discover new introns, and assess levels of intron splicing. We found spliced transcript levels to be surprisingly low on average, most likely as a result of spliceosomal reduction, but with the potential for decreased decay rates to be playing a role. Gene expression levels vary over the course of infection; tremendous numbers of genes are differentially expressed in the first 48 hrs post-infection, suggesting a major genetic change that likely precedes a life-stage change. The reduction of spliceosome and RNA decay pathway components appears to be the cause of decreased splicing efficiency and an accumulation of unspliced, non-functional transcripts. This suggests that a balance is maintained between inefficiency resulting from gene loss, and continued pressure of genome reduction.

3.4 Materials and methods

3.4.1 RNA preparation

E. cuniculi (Genotype II) was cultured in the rabbit kidney fibroblast cell line (CCL-37, American Type Culture Collection, Manassas, VA USA). Intracellular meront stages of *E. cuniculi* appear to bind to the parasitophorous vacuole membrane and thus cannot be physically separated from host cells. Total RNA therefore, was extracted from two biological replicates of RK13 cells in 25cm2 tissue culture flasks 24 hr, 48 hr, and 72 hr post-infection using the Ambion RNAqueous kit (Ambion, Austin, TX). Extracted RNA was treated with TURBO DNase (Ambion, Austin, TX) to eliminate any contaminating DNA. RNA quality was assessed on an Agilent Bioanalyzer 2100 (Agilent, Santa Clara, CA) and RNA quantity was measured on a Qubit 2.0 fluorometer (Life Technologies Corp., Carlbad, CA).

3.4.2 RNA-seq library preparation

A total of six Illumina cDNA libraries were prepared according to the TruSeq library preparation protocol (Illumina, Hayward, CA). A total of 4ug of RNA from each of the six DNase-treated samples was used as starting material. Library quality control and pooling were performed by the Biodiversity Research Centre (BRC) sequencing facility (UBC, Vancouver, BC).

3.4.3 Illumina sequencing and data processing

Paired-end sequencing was performed on an Illumina HiSeq 2000 at the BRC sequencing facility. The six libraries were multiplexed and sequenced in two lanes in order to give two technical replicates for each of the biological replicates, and to help avoid bias associated with a particular flow cell or lane therein (Auer and Doerge 2010). Although paired-end RNA-seq does not account for possible antisense and overlapping transcription, our previous work has indicated that such transcripts are limited to the extracellular spore stage of the parasite (Corradi, Gangaeva, Keeling 2008; Gill *et al*. 2010; Grisdale and Fast 2011; Williams *et al*. 2005).

Raw sequence data was processed and converted to fastq format. Since RNA was obtained from *E. cuniculi* genotype 2 infected RK13 cells, reads were mapped to the genotype 2 reference genome (GenBank accession AEWQ01000000) (Pombert *et al*. 2013), as opposed to strain GB-M1 (Katinka *et al*. 2001). Reads mapping to *E. cuniculi* are available at the NCBI Sequence Read Archive under study accession SRP017112. The short-read aligner Bowtie version 0.12.7 (Langmead *et al*. 2009) was used for read mapping, using default mismatch parameters, and allowing only a single alignment for each read. Biological and technical replicates showed extremely high levels of correlation (see Additional file 1), as has been seen in previous RNA-seq experiments (Bruno *et al*. 2010; Mortazavi *et al*. 2008; Nagalakshmi *et al*.

2008; Wilhelm *et al*. 2008). SAMtools version 0.1.18 (Li *et al*. 2009) was used to process SAM and BAM alignment files. Alignments were visualized using the Integrative Genomics Viewer version 2.0.7 (Robinson *et al*. 2011). Expression levels were measured in the standard fragments per kilobase per million mapped reads (FPKM) format (Mortazavi *et al*. 2008). We found 45 genes with less than twenty reads of coverage in at least one time-point, suggesting that their expression may be the result of background or antisense transcription, and therefore not of biological significance. However, 42 of these encode tRNAs, 5S rRNA, or U2 snRNA, which were not expected to have read coverage following polyA-selected library preparation.

3.4.4 Assessing splicing efficiency

Attempts were made to use Tophat (Trapnell, Pachter, Salzberg 2009) as a splice junction mapper, however it was not able to detect introns in *E. cuniculi*. Therefore, a custom Bowtie reference was made in order to automate splicing level counts. The sequences of all *E. cuniculi* introns and one hundred flanking nucleotides were obtained from the genome reference (Pombert *et al*. 2013). Two reference sequences were created for each intron locus, one containing the intron sequence and one with the intron sequence removed. The flanking sequence was also reduced to 96nt at each end of the splice junction. Therefore, in order for a read to map to one of the reference sequences, it must overlap the splice junction (without the intron) or the intron itself by a minimum of 5nt. The data set was mapped to this reference using Bowtie, producing a Sequence Alignment/Map (SAM) format output file containing sequence alignment information. SAMtools was used to obtain mapping statistics for the reference sequences, producing counts of the number of reads that map to the spliced and unspliced reference sequences. The number of spliced reads was then divided by the total number of reads covering each splice junction, in order to produce a measure of the splicing levels. Pairwise comparisons of splicing levels for

each intron-containing gene were performed with corrected Pearsons's chi-squared tests in R (R Development 2011). Pairs of splicing level values were considered to be significantly different if the chi-squared p-value was less than 0.01. As described above, splicing levels observed are unlikely to result from antisense transcription in this stage of the parasite. Indeed, the presence of significantly different splicing levels across time points for several introns, different splicing levels for two introns in the same gene, and several genes showing high levels of splicing, further supports previous observations that antisense transcription is not widespread in intracellular *E. cuniculi*, and is therefore unlikely to be responsible for the splicing levels observed.

Custom Bowtie reference sequences were prepared for 80 randomly selected introns from *Saccharomyces cerevisiae* and Candida albicans, as described above. All RNA-seq reads from the publicly available datasets for *S. cerevisiae* (SRX000559-SRX000564) (Nagalakshmi *et al*. 2008) and *C. albicans* (SRP002852) (Bruno *et al*. 2010) were mapped against the respective custom reference sequences, allowing spliced and unspliced reads to be counted (as above). Forty-six *S. cerevisiae* and forty-eight *C. albicans* junctions remained after filtering for those with at least 50X coverage.

3.4.5 Differential gene expression analysis

After mapping with Bowtie, read counts were obtained for all *E. cuniculi* ORFs using HTseq (http://www-huber.embl.de/users/anders/HTSeq/ webcite). The read counts were then analyzed for differential expression (DE) using DESeq (Anders and Huber 2010), an R/Bioconductor package (Gentleman *et al*. 2004; R Development 2011). A p-value cut-off of 0.01 was used for the DE analysis. The custom *E. cuniculi* gene annotation file used with DESeq was created from the ecotype II genome assembly files (Pombert *et al*. 2013) using a custom Python (Python 2.6.2) script (available upon request).

3.4.6 Search for novel transcribed regions (NTRs)

The extreme gene-dense nature of the *E. cuniculi* genome made it unreliable to use a custom script to search for NTRs. Therefore, the read alignment files were searched visually for NTRs using IGV. The search parameters used were: a minimum of 10X coverage, no overlap with previously annotated ORFs, and distinguishable borders with regards to the reads mapping to adjacent ORFs, in order to avoid counting untranslated regions.

Table 3.1: Number of reads mapped to parasite and host genomes

The number of reads mapping to *Encephalitozoon cuniculi* and *Oryctolagus cuniculus* at three post-infection time-points are shown.

Time-point	***Encephalitozoon cuniculi***	***Oryctolagus cuniculus***	**Total reads mapped**
T1	13895384	92511829	106407213
T2	17454072	84792245	102246317
T3	9286586	96176009	105462595
Total	40636042	273480083	314116125

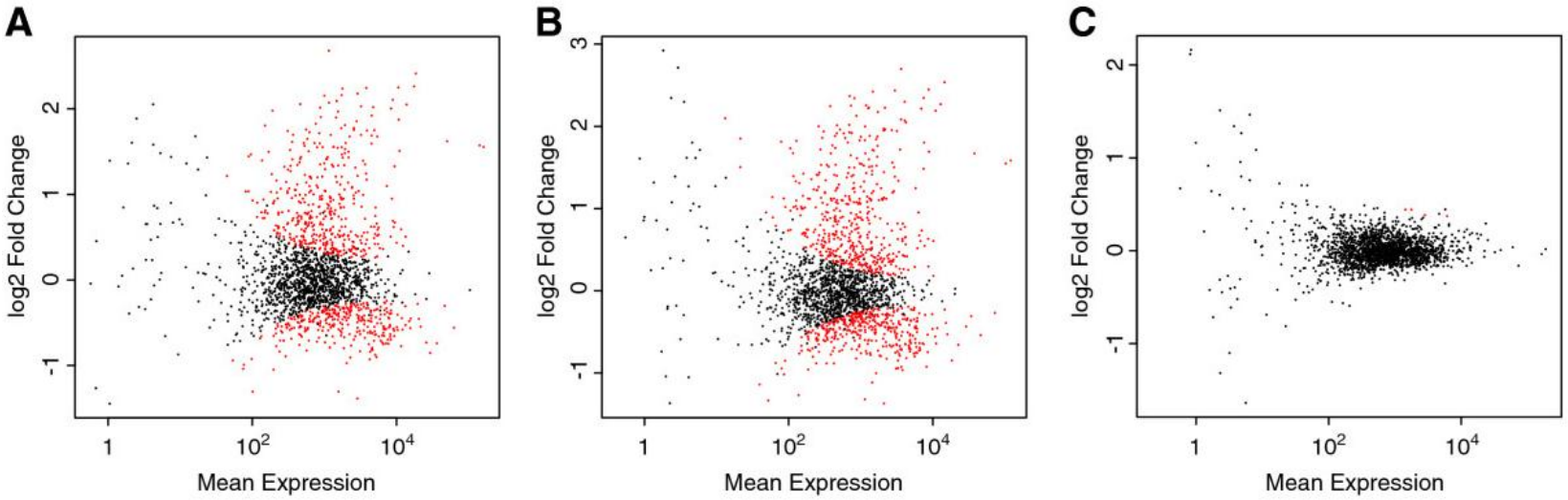


Figure 3.1: Differential expression across three post-infection time-points

Plot of log2 fold change versus mean expression level for all *E. cuniculi* genes. Red dots indicate those genes that are differentially expressed and black dots indicate those that are not. (A) Differential expression between 24 hr and 48 hr, (B) 24 hr and 72 hr, and (C) 48 hr and 72 hr post-infection.

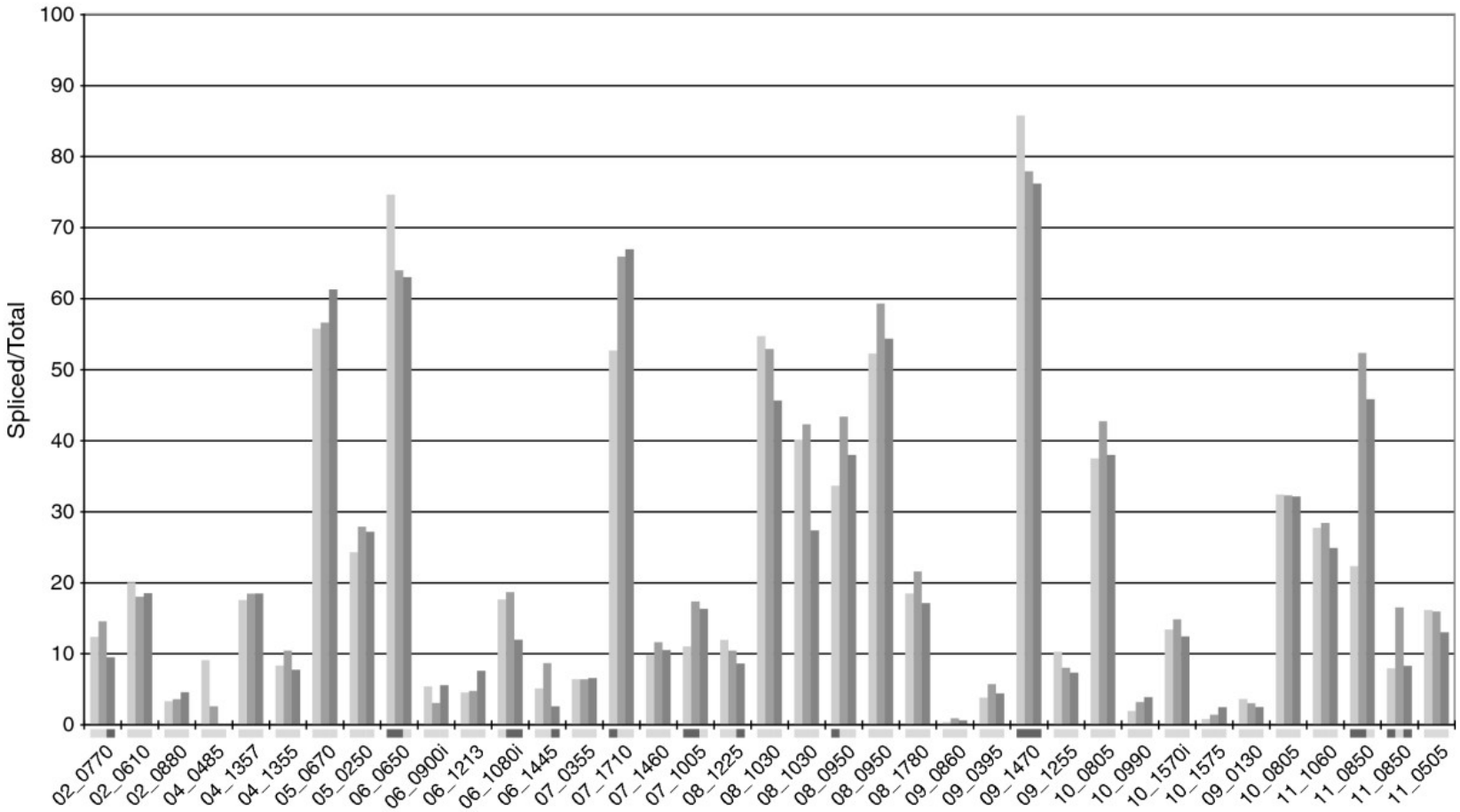


Figure 3.2: Splicing levels of all *E. cuniculi* intron-containing genes

Levels of splicing were determined by measuring the number of spliced and unspliced transcripts and dividing spliced by total transcripts to produce a percentage of splicing. From left to right, splicing levels at 24 hr, 48 hr, and 72 hr are indicated by grey bars. *E. cuniculi* gene names are on the x-axis. Significant differences in splicing levels between time-points are shown by darkened boxes along the x-axis. From left to right, darkened boxes indicate significant differences between 24 hr and 48 hr, 24 hr and 72 hr, and 48 hr and 72 hr.

Chapter 4: High-throughput transcriptome sequencing of *Cyanidioschyzon merolae* reveals unexpected levels of constitutive and alternative pre-mRNA splicing and antisense transcription

4.1 Introduction

Cyanidioschyzon merolae is a tiny, unicellular red alga that thrives in acidic hot springs (De Luca, Taddei, Varano 1978). It belongs to one of three species-poor genera of thermo-acidophilic algae within the Cyanidiales. Although these lineages were originally believed to contain just 1-2 species each, recent biodiversity surveys of Cyanidiales suggest that this group contains much higher diversity, with many geographically isolated and highly diverged strains (Ciniglia *et al*. 2004; Gross *et al*. 2001). *C. merolae* has very simple cellular architecture with a single nucleus, mitochondrion, and plastid: all of which can be made to divide synchronously under 12h:12h light:dark cycles (Terui *et al*. 1995). The genome of *C. merolae* has been completely sequenced, and was found to be reduced relative to most other photosynthetic eukaryotes (Matsuzaki *et al*. 2004; Nozaki *et al*. 2007). A total of 16.5Mb of DNA spread over 20 chromosomes encodes approximately five thousand genes (Matsuzaki *et al*. 2004; Nozaki *et al*. 2007). Reduction has shaped the *C. merolae* genome in a number of ways that make it unique compared to other eukaryotes. It has three non-repeat rRNA (18S-5.8S-28S) units, instead of the typical tandem repeat rRNA units of most eukaryotes (Matsuzaki *et al*. 2004). Along with three nearly identical copies of the 5S rRNA, this makes the rRNA complement of *C. merolae* the smallest known (Matsuzaki *et al*. 2004). Widespread intron loss has occurred, leaving just 27 introns in 26 genes (Matsuzaki *et al*. 2004). Why *C. merolae* has retained so few introns, requiring dozens of spliceosome components for their removal, remains a mystery.

The process of spliceosome-mediated intron removal involves a large complex of proteins and RNAs that interact in a dynamic process to perform two transesterification reactions. Spliceosomal components are conserved across most eukaryotic lineages, with over one hundred spliceosomal protein-coding genes present in yeast and more than two hundred in mammals (Jurica and Moore 2003; Wahl, Will, Luhrmann 2009). However, several exceptions exist regarding the highly conserved nature of spliceosome components. For example, the model microsporidian *E. cuniculi* retains just 30 ORFs with homology to splicing-associated genes and putatively lacks one of five snRNAs (Katinka *et al*. 2001; Lopez, Rosenblad, Samuelsson 2008). Transcriptome data from two distantly related microsporidians, *Edhazardia aedis* and *Nematocida parisii,* suggest that splicing is not active in these two species that have lost most canonical splicing machinery and introns (Akiyoshi *et al*. 2009; Cuomo *et al*. 2012). Barring the potential nuclear-encoded, nucleomorph-targeted proteins, the nucleomorph genomes of the Cryptophyte *Guillardia theta* and the Chlorarachniophyte *Bigellowiella natans* lack many conserved splicing proteins, most notably in *G. theta* (Gilson *et al*. 2006). These two nucleomorph genomes are predicted to retain fewer than five snRNAs each, suggesting a loss of these essential splicing components (Lopez, Rosenblad, Samuelsson 2008). The genome sequence of *C. merolae* revealed that it also has a small number of introns and a highly reduced spliceosome (Matsuzaki *et al*. 2004; Nozaki *et al*. 2007). Given the large phylogenetic distance between red algae and reduced systems in other major eukaryotic lineages, examining splicing in *C. merolae* will provide a powerful comparison in an independently reduced eukaryote.

Alternative splicing is a process that allows for the creation of multiple isoforms from a single transcript. This can lead to an increase in proteome diversity without a need to encode more genes, thereby increasing the phenotypic complexity of an organism without an increase in

genome size (for reviews, see (Graveley and Nilsen 2010; Stamm *et al*. 2005)). Alternative splicing can also provide a means for regulating gene expression, as the introduction of premature stop codons in unspliced transcripts can lead to efficient decay via the nonsense mediated decay (NMD) pathway (Chang, Imam, Wilkinson 2007; Lareau *et al*. 2007a; Palusa and Reddy 2010). In *Arabidopsis thaliana*, where approximately 60% of genes produce multiple transcript isoforms, experimental work has shown that nearly 20% of multi-exon genes are NMD sensitive (Drechsel *et al*. 2013; Kalyna *et al*. 2011; Marquez *et al*. 2012). Although alternative splicing is widespread and well documented in multicellular eukaryotes, very little is known of the presence and implications of alternative splicing in eukaryotes with reduced genomes. Individual cases of alternative splicing have been documented in several unicellular eukaryotes, while just a few of these species have had transcriptome-wide analyses performed (Curtis *et al*. 2012; Grisdale *et al*. 2013; Iriko *et al*. 2009; Jaillon *et al*. 2008a; Kabran *et al*. 2012; Labadorf *et al*. 2010; Loftus *et al*. 2005; Mekouar *et al*. 2010; Muhia *et al*. 2003; Sorber, Dimon, DeRisi 2011). An RNA-seq analysis of the highly reduced microsporidian *E. cuniculi* showed just a few cases of alternative 3′ splice site usage occurring at extremely low frequencies (Grisdale *et al*. 2013). However, levels of intron retention were very high in this species with a reduced spliceosome, suggesting that inefficient splicing (or mis-splicing) could be the result of spliceosomal component losses (Grisdale *et al*. 2013). Understanding the role of alternative splicing in unicellular eukaryotes will shed light on the origin and importance of this process throughout eukaryotic evolution. Also, examining the retention of introns in organisms with reduced genomes should provide insight into the evolution of intron function and gene architecture.

The recent availability of strand-specific library preparation for high-throughput sequencing allows read mapping software to detect which strand a read originated from, providing a new layer of transcriptomic information. Strand information can be used to discover where antisense transcription is occurring, that is, transcription on the strand opposite to annotated (sense) loci. Two types of regulatory antisense transcripts have been described: cis-natural antisense transcripts (cis-NATs) that overlap only with the locus from which they originated, and trans-NATs that typically have non-perfect complementarity to more than one genomic locus allowing them to bind multiple loci. There are three possible cis-NAT orientations: tail-to-tail, head-to-head, and one transcript contained completely within another. By far the most prevalent form found in genome-wide studies to date is the tail-to-tail oriented cis-NATs pairs (Lapidot and Pilpel 2006). The frequency of cis-NATs has been estimated in several model systems, with a range of 5-29% of transcripts forming cis-NATs in animals (Chen *et al*. 2004; Katayama *et al*. 2005; Misra *et al*. 2002; Sun *et al*. 2006; Yelin *et al*. 2003), 7-26% in plants (Jen *et al*. 2005; Osato *et al*. 2003; Poole *et al*. 2008; Wang, Gaasterland, Chua 2005), and 24% in the malaria parasite, *Plasmodium* (Siegel *et al*. 2014). A recent study has even suggested that 90% of *S. cerevisiae* genes have detectable levels of antisense transcription (Goodman, Daugharthy, Kim 2013).

Cis-NATs have been implicated in gene regulation through four mechanisms, and can affect changes in processes such as transcription, mRNA stability, splicing, cellular transport, genomic imprinting, X-inactivation, DNA methylation, and translation. Transcriptional interference occurs when RNA polymerase units moving towards one another on opposite strands results in stalled elongation, as is the case with the convergently oriented GAL10 and GAL7 genes in yeast (Prescott and Proudfoot 2002). This can produce the common phenomenon

of inverse expression levels of sense-antisense pairs, or simply block elongation of both transcripts. Transcript pairs can mask cis-elements in each other, causing changes in splicing, stability, polyadenylation, degradation, or any process involving pre-mRNA interacting with proteins or RNAs. The splice variants of a mammalian thyroid hormone receptor are strongly affected by the expression level of an overlapping gene, which is thought to mask the splicing regulatory cis-elements of one form of the thyroid receptor transcripts (Hastings *et al*. 1997). Double-stranded RNA-dependent mechanisms, such as RNA interference (RNAi), are another way for antisense pairs to cause silencing or degradation of specific transcripts. The salt-stress induced expression of an overlapping gene pair in *Arabidopsis* results in the processing of the sense-antisense duplex to produce small interfering RNAs (siRNAs) involved in salt tolerance (Borsani *et al*. 2005). Finally, antisense transcription can play a role in methylation, causing gene silencing, and monoallelic expression. An example of monoallelic expression is X-inactivation in females. The non-coding RNA involved in recruiting the histone-modifying complex is blocked by its antisense transcript, resulting in activation of the X-chromosome expressing the antisense (Ogawa and Lee 2002). In all, antisense transcription appears to play important roles in regulation at nearly all levels of gene expression, and therefore, deserves more attention in transcriptome-wide studies.

In recent years *C. merolae* has gained attention as an emerging model system for studying organelle division, as well as the origin and evolution of photosynthetic eukaryotes. It has simple cellular architecture and its nuclear, plastid, and mitochondrial genomes have been fully sequenced (Matsuzaki *et al*. 2004; Nozaki *et al*. 2007; Ohta, Sato, Kuroiwa 1998; Ohta *et al*. 2003). Also, transformation systems have been developed in *C. merolae*, bringing it into the realm of experimental genetics (Fujiwara *et al*. 2013; Minoda *et al*. 2004; Ohnuma *et al*. 2008).

While some microarray-based analyses of gene expression have been performed on *C. merolae*, they focused on specific processes such as nitrogen-responsive genes or genes associated with organelle division (Fujiwara *et al*. 2009; Imamura *et al*. 2010). Here, I present the first high-throughput sequence analysis of the *C. merolae* transcriptome, focusing on pre-mRNA splicing and antisense transcription. I prepared and sequenced both non-stranded and strand-specific RNAseq libraries for *C. merolae*, from both light and dark phases of its growth cycle. This provided the sequencing depth to accurately measure the levels of pre-mRNA splicing and differential gene expression, identify new introns and alternative splicing events, and assess levels of antisense transcription. Examining these aspects of the transcriptome under light and dark cycles provided insight into the changes in pre-mRNA splicing and gene regulation during the day:night cycle.

4.2 Results and discussion

The *C. merolae* transcriptome was examined by sequencing cDNA libraries made from polyA-selected RNA, extracted half way through the light and dark periods of growth. A total of 818.1 million reads were produced from three lanes of Illumina HiSeq paired-end sequencing (see Table 4.1), of which 578.7 million (70.7%) mapped to the *C. merolae* reference nuclear genome (Matsuzaki *et al*. 2004; Nozaki *et al*. 2007).

4.2.1 Differential expression between light and dark phases

Genome wide differential expression analysis was performed for 4494 annotated *C. merolae* protein-coding genes with above-background levels of expression (see Methods for details). I found a large proportion (94.2%) of *C.* merolae genes expressed during both conditions. This appears to be a common phenomenon in reduced systems. Reduced microsporidian and nucleomorph genomes are observed to maintain high levels of gene

expression for nearly all genes, likely as a mechanism of compensating for inefficient processing or protein function (Grisdale *et al*. 2013; Tanifuji *et al*. 2014). Gene expression was assessed at half way points through the light and dark phases of the 12h:12h light:dark growth cycle, with four biological replicates per condition. We expected a small number of photosynthesis and cell cycle related genes to be prime candidates to undergo changes in gene expression levels between the two conditions examined. Surprisingly, we found that 2485 (55.3%) *C. merolae* genes are differentially expressed (corrected p-value <0.1; see methods for details) between light and dark phases within a single day (Figure 4.1). The average fold change is 1.57 for the 1236 genes with increased expression in the dark relative to the light. The average fold change is 0.57 for the 1249 genes with decreased expression in the dark relative to the light. Therefore, the average change in expression (whether an increase or decrease) between light and dark is approximately 50%.

Of the 2485 genes showing significant changes in expression between day and night, 164 (6.6%) had a log2 fold change of greater than 1.0, meaning they had at least doubled in expression from light to dark. On the other hand, 316 (12.7%) of the 2485 differentially expressed genes had a log2 fold change of less than -1.0, meaning a reduction by at least a half in expression levels between light and dark. The ten genes with the largest fold changes (8.1 - 46.2 fold) between light and dark include: six hypothetical genes, alpha- and beta-tubulin, and two kinases. Although we cannot infer the role of the hypothetical genes with large increases in expression during the dark, the changes in levels of tubulin and kinase transcripts likely reflect their roles in cell cycle progression. In line with these results, a microarray-based study of gene expression during the cell cycle of *C. merolae* found that tubulin genes were induced during the G2-M phase, in the dark (Fujiwara *et al*. 2009). The ten genes with largest negative fold change

(-13.0 – -110.7 fold) between light and dark include: one light harvesting protein, two chlorophyll binding proteins, two kinases whose origin or function is linked with the plastid, one iron-sulfur protein related gene, one aminotransferase, one dual specificity kinase, and one hypothetical gene. As expected, several plastid related genes are among the most strongly down-regulated during the dark, including a light harvesting gene with the single largest change in expression level (-110 fold) between day and night. Therefore, the regulation of photosynthesis or plastid-related genes is detectable at the mRNA level (pre-translation), however, we cannot determine the exact stage of RNA processing at which these transcripts are regulated from expression data alone.

The large number of genes with differing levels of expression between day and night is likely partially the result of cell cycle regulation. In *S. cerevisiae* the expression levels of 400-800 genes have been shown to fluctuate with cell-cycle progression, while approximately 500 genes are cell-cycle regulated in *A. thaliana* (Cho *et al*. 1998; Menges *et al*. 2002; Spellman *et al*. 1998). A previous gene expression analysis in *C. merolae* identified 358 of 4586 genes expressed being cell-cycle regulated (Fujiwara *et al*. 2009). However, even if these same 358 genes were found to be differentially expressed in my analysis, this still leaves over two thousand genes not related to cell-cycle progression that are undergoing changes between day and night. A study of diurnal gene expression changes in *Arabidopsis* showed a high frequency of genes with changes in expression level (Blasing *et al*. 2005). The authors suggest that potentially up to half of the genes expressed in leaves undergo diurnal changes (Blasing *et al*. 2005). Therefore, *C. merolae* gene expression changes are in line with those observed in this well-studied photosynthetic system.

4.2.2 High prevalence of antisense transcription

Strand-specific library preparation allocates reads to the strand from which their associated transcript originated. This allowed me to obtain read counts of sense and antisense mapping reads for all annotated loci. I chose to use the dUTP directional cDNA library preparation method, as it was shown to be the best performing method allowing paired-end sequencing from a thorough analysis of seven directional library preparation methods (Levin *et al*. 2010). Previous strand-specific studies of antisense transcription found that this method can produce anywhere from 1-12% error (Levin *et al*. 2010; Li *et al*. 2013). This error in library preparation likely results from PCR template fragments being sequenced even though they should be much less abundant than amplified fragments, which are the reverse complement of the template. If the library preparation procedure has a high error rate in strand specificity, we would expect at minimum a low level of antisense transcripts mapping to all expressed genes. As a check of strand specificity, I calculated the number of genes with 100 or more sense reads and 0 antisense reads in any single biological replicate, representing genes with high sense coverage but absolutely no antisense expression. A total of 88 genes passed this filter. In addition, I found 479 genes with less than 1% antisense reads out of the total count of sense and antisense reads, suggesting that my sequence libraries have high strand specificity.

A large proportion of genes were found to have detectable levels of antisense transcription. While the majority of genes have less than 10% antisense reads, several hundred have very high counts of antisense reads (see Figure 4.3). In fact, 310 genes have greater than 50% antisense reads in both light and dark, while 510 genes have greater than 50% in one condition but not both. This represents a large number of genes with more antisense than sense expression. If reads are normalized for gene and library size using fragments per kilobase per

million mapped reads (FPKM), I find 1635 genes with greater than 1.0 FPKM: considered a basal level of biologically relevant expression (Mortazavi *et al*. 2008). Given the lack of identified small-RNA processing components in *C. merolae*, it is surprising to find that more than 30% of its genes have antisense transcription with the potential to form cis-NATs. Analyses of cis-NAT levels in several animal and plant species have found between 5-29% of genes with antisense transcription (Chen *et al*. 2004; Jen *et al*. 2005; Katayama *et al*. 2005; Misra *et al*. 2002; Osato *et al*. 2003; Poole *et al*. 2008; Sun *et al*. 2006; Wang, Gaasterland, Chua 2005; Yelin *et al*. 2003). While one study examining antisense transcription in yeast detected antisense transcription in a very large proportion of genes (~90%), the use of different sequencing methods and read count normalization make it difficult to directly compare with the data presented here (Goodman, Daugharthy, Kim 2013). Antisense transcripts forming cis-NATs play major roles in gene regulation through many mechanisms, such as altering transcription, pre-mRNA splicing, mRNA stability, and transport, among others (Borsani *et al*. 2005; Hastings *et al*. 1997; Misra *et al*. 2002; Morrissy, Griffith, Marra 2011; Peters *et al*. 2003; Prescott and Proudfoot 2002). The high variability in estimated levels of antisense transcript levels in diverse species is perhaps not surprising given the recent development of techniques allowing for transcriptome-wide assessment of antisense transcription. Nevertheless, a high frequency of potential cis-NAT forming antisense transcripts in *C. merolae* raises questions regarding the potential function of these transcripts.

To assess potential regulatory roles for antisense transcripts expressed in *C. merolae*, I analyzed their expression levels under light and dark conditions. Genes with fewer than 20 reads of coverage in any replicate were filtered from the dataset, leaving a total of 2070 antisense transcripts to be analyzed. Of these 2070 genes, 1189 (57.4%) were found to have significantly

different expression levels between light and dark conditions. The presence of a large portion of antisense transcripts differentially expressed indicates that antisense transcription in *C. merolae* is likely not just background or messy transcription, and instead suggests that antisense transcripts are playing a role in regulation. Identifying the pathway through which these transcripts function, and which transcripts form cis-NAT pairs, will require analysis of small RNAs specifically. However, preliminary analysis of antisense transcripts by gene ontology enrichment provides evidence that photosynthesis-related genes are down-regulated in the dark by increased levels of antisense transcription (data not shown). Antisense transcripts with increased expression in the dark are enriched for photosynthesis-related genes in the biological process category, while those with decreased expression are not. In addition, functional annotation clustering shows enrichment of photosynthetic genes for the set of antisense transcripts with increased expression in the dark, but not for those with decreased expression.

4.2.3 Intron annotation

C. merolae was originally annotated with a mere 27 introns in 26 genes, suggesting that 99.5% of protein-coding genes lack introns (Matsuzaki *et al*. 2004; Nozaki *et al*. 2007). This contrasts with most other eukaryotes, even those with small genomes. For example, the genomes of *Saccharomyces cerevisiae* (12.5Mb), *Cryptosporidium parvum* (9.1Mb), *Ostreococcus tauri* (12.5Mb), and *Plasmodium falciparum* (22.8Mb) all contain introns in at least 5% of their protein-coding genes (Abrahamsen *et al*. 2004; Derelle *et al*. 2006; Gardner 2002; Goffeau *et al*. 1996; Spingola *et al*. 1999). In order to identify any additional spliceosomal introns in *C. merolae*, the RNAseq data were mapped using Tophat split-read mapping software (see Methods for details; (Kim *et al*. 2013)). All 27 originally annotated introns were found to be spliced during both light and dark conditions. After filtering out false positive new introns with low

depth of coverage or that flank repetitive DNA motifs, I was able to identify an additional 16 introns, raising the total to 43 spliceosomal introns. These newly annotated introns show sequence and read-alignment characteristics that lend confidence to their designation, and strongly suggest that they are not resulting from sequencing artifacts or biological noise. However, not all new introns were found in conserved ORFs. New intron insertion patterns include: seven introns that split a single exon gene into two exons, one intron found in a 3'UTR region, three introns that extend the 5'end of a gene, and five that create new ORFs or non-coding genes. New non-coding genes were annotated based on having a transcriptional unit that is not overlapping with transcribed regions of neighboring annotated genes. However, these non-coding genes typically showed little to no similarity to known genes by BLAST searches and, therefore, were annotated based on transcriptional profile alone.

Two of the newly annotated introns have AT-AC boundaries, suggesting that they require splicing by the U12-dependent spliceosome. The RNAseq data show that these introns in CMI050C and CMJ154C are spliced at extremely low levels (see Figure 4.4). However, these are likely bona fide introns, as CMI050C was covered by over a thousand spliced reads, while CMJ154C was covered by more than fifty spliced reads. To date, no minor spliceosome-specific components have been annotated in *C. merolae*, making the presence of AT-AC introns particularly surprising. The minor spliceosome consists of U11, U12, U4atac, and U6atac snRNAs and a small number of minor spliceosome-specific proteins (Hall and Padgett 1996; Tarn and Steitz 1996a; Tarn and Steitz 1996b). The U5 snRNA, along with a number of proteins assigned to the major spliceosome, are shared with the minor spliceosome (Hall and Padgett 1996; Tarn and Steitz 1996a). Although BLAST searches did not identify any minor spliceosome components in *C. merolae*, it is possible that the level of divergence has rendered it impossible to

find them by simple sequence similarity. On the other hand, it is possible that introns in CMI050C and CMJ154C are spliced by the canonical major spliceosome, as has been shown for a subclass of introns with AT-AC boundaries (Dietrich, Incorvaia, Padgett 1997; Wu and Krainer 1997). The CMI050C and CMJ154C introns do not have the highly conserved canonical motifs that are specific to U12-dependent introns, suggesting an increased likelihood that they are spliced by a similar mechanism as U2-dependent introns (Burge, Padgett, Sharp 1998).

C. merolae introns have several unusual characteristics that may help shed light on the evolution of intron retention in reduced systems. Intron phase refers to the position of introns within (phase-1 and -2) or between (phase-0) codons. Eukaryote-wide, phase-0 introns are predicted to occur most frequently, while phase-2 introns are rarest (Fedorov *et al*. 1992; Long, Rosenberg, Gilbert 1995; Long *et al*. 1998). However, I find more phase-2 (40%) positioned introns than phase-0 (32.5%) introns in *C. merolae* (Table 4.2). While difficult to draw conclusions from the small number of introns in *C. merolae*, it appears that this species does not follow the eukaryotic norm by having an abundance of phase-0 introns. Other interesting characteristics of *C. merolae* introns are that very few are a multiple of three nucleotides (3n introns) in length, and all but six introns contain premature stop codons. Just 11 of 40 introns in protein-coding genes are 3n, allowing read-through if the intron does not contain a stop codon. Interestingly, three of the 3n introns do not contain premature stop codons. These three introns have 5′ splice site motifs that are different from the most common motif (GTAAGT), and they are all shorter than the average intron size in *C. merolae*. Also, two of these 3n introns with no stop codon showed an elevated level of coverage across the intron. This suggests that read-through of these introns happens frequently, possibly resulting in the favoring of shorter intron length and relaxed selection at the 5′ splice site motif. Analysis of introns in *Yarrowia lipolytica*

and *Paramecium tetraurelia* showed an over-representation of introns that cause frame-shifts, and an under-representation of 3n introns (Jaillon *et al*. 2008b; Mekouar *et al*. 2010). Also, the under-represented 3n introns in *P. tetraurelia* are significantly enriched for stop codons, suggesting that introns in *P. tetraurelia* are under selective pressure to introduce premature stop codons into intron-retaining transcripts (Jaillon *et al*. 2008b). A similar pattern seems to be present in *C. merolae*, as very few introns are phase-0 or lack a premature stop codon. This suggests that selective pressures may be favoring introns that result in premature stop codons, potentially as a means of targeting them for degradation (see section 4.2.5 for further discussion).

4.2.4 Pre-mRNA splicing levels

Levels of pre-mRNA splicing were calculated for all 43 junctions using counts of reads split across junctions (spliced) and reads mapping to intronic regions (unspliced). The level of splicing was determined by dividing the number of reads split over a junction by the sum of all reads (both spliced and unspliced) at that junction. In addition, splicing levels were normalized to take into account the higher probability of reads mapping to longer introns (see Methods for details). Initially, these values were calculated from the strand-specific data and non-stranded data individually. Although I found four genes with more than 5% coverage over their intron originating from the antisense strand, statistical analysis provided evidence that, overall, there is not a significant difference between the two datasets. The correlations between strand-specific and non-stranded splicing levels in light and dark conditions were 0.95 and 0.94, respectively. The average splicing levels from the two datasets for the light condition were very similar at 18.24% and 18.27% for stranded and non-stranded data, respectively. The average splicing level from the dark condition showed slightly more variability at 20.25% and 18.09%. Kolmogorov-Smirnov tests showed that the differences in splicing levels between the two data sets are not

significantly different, with p-values of 0.765 and 0.917 for light and dark conditions, respectively. This statistical analysis confirms that the two datasets are not significantly different. Therefore, the datasets were combined, and splicing levels were assessed using this larger dataset.

The levels of splicing were found to be very low in *C. merolae* (Figure 4.4). The average splicing levels of the normalized, combined datasets were 20.8% for the light condition, and 21.2% for the dark condition. Similar levels of splicing were found in the microsporidian parasite *E. cuniculi*, which showed less than 50% splicing efficiency in 30 of its 37 introns (Grisdale *et al*. 2013). These levels of splicing found in *C. merolae* and *E. cuniculi* are unusually low, even when compared with other unicellular eukaryotes. An analysis of splicing in *P. falciparum* found that only a small proportion of intron-containing genes (~5%) have splicing levels below 50% (Sorber, Dimon, DeRisi 2011). In previous work, I found approximately 10-15% of introns analyzed have low splicing levels (<50%) in two fungal species (Appendix A.5) (Grisdale *et al*. 2013). Low splicing efficiencies appear be common to reduced systems, as *C. merolae* and *E. cuniculi* have similarly low levels of splicing, and have independently lost many spliceosomal protein-coding genes (see Table 1.1). Also, both species appear to lack the U1 snRNA. A computational screen for snRNAs did not identify U1 in *C. merolae* or *E. cuniculi* (Davila Lopez, Rosenblad, Samuelsson 2008). In addition, using a biochemical assay to enrich for snRNAs, followed by high-throughput sequencing, we could not identify any U1-like sequences in *C. merolae* (data not shown). We suggest that the loss of spliceosomal machinery likely results in changes in the splicing reaction. As components are lost, fewer interactions will take place. For example, in systems without the U1 snRNA, initial identification of the intron and

binding at the 5′ splice site by the U1 snRNP complex no longer occurs. Changes such as these may result in decreased splicing fidelity that affects splicing efficiency.

4.2.5 Alternative splicing

In addition to the high frequency of intron retention events found for all 43 *C. merolae* splice junctions, more than half of the introns display additional types of alternative splicing. This large proportion of alternatively spliced introns was surprising given that only two SR splicing regulatory protein-coding genes are annotated in *C. merolae*, while at least 10 SR protein-coding genes are found in organisms with high rates of alternative splicing (see Table 1.1). Combining data from all samples and conditions shows that, of the 43 introns, 27 undergo alternative 3′ splice site usage, 6 undergo alternative 5′ splice site usage, 4 use alternative 5′ and 3′ splice sites in the same transcript, and the single multi-intron gene has rare exon skipping events. Fewer events are found in the strand-specific dataset when compared with the non-strand-specific dataset, likely as a result of the large difference in sequencing depth (Table 4.2). The distributions of alternative splicing events are very similar when comparing the data from light and dark conditions (Table 4.2). Therefore, differences in alternative splicing event frequencies do not appear to be playing a role in regulation between day and night. However, in addition to intron retention, other types of alternative splicing events resulting in frame-shifts and premature stop codons are likely having significant effects on the expression of approximately half of the intron-containing genes in *C. merolae* during both light and dark growth.

Although alternative splicing is relatively rare in unicellular eukaryotes, it has been noted in a few well-studied species. The genome of the ciliate *Paramecium tetraurelia* contains thousands of small introns, but alternative splicing is extremely rare, affecting less than 1% of the introns examined (Jaillon *et al*. 2008b). A genome-wide analysis of alternative splicing in the

model green alga *Chlamydomonas reinhardtii* revealed 498 transcripts with 611 alternative splicing events, a relatively small proportion of transcripts in such an intron-rich genome (Labadorf *et al*. 2010). Intron retention events were found to be significantly more frequent than any other type of alternative splicing event in *C. reinhardtii*, a phenomenon that has been noted in land plants as well (Kim, Magen, Ast 2007; Labadorf *et al*. 2010; Wang and Brendel 2006). Similar results were found in two fungi; the intron-rich yeast *Yarrowia lipolytica* has frequent intron retention, while 100% of intron-containing genes in the intron-poor microsporidian parasite *E. cuniculi* display high levels of intron retention (Grisdale *et al*. 2013; Mekouar *et al*. 2010). In contrast, an analysis of a set of intron-containing genes in the malaria parasite *Plasmodium falciparum* shows intron retention events (26.7%) to occur at the same frequency as alternative 3′ splice site usage (26.7%), while alternative 5′ splice site events (46.7%) occur with nearly twice the frequency of the other events (Iriko *et al*. 2009). Although not a universal phenomenon, high frequencies of intron retention events seem to be widespread among unicellular eukaryotes. The numbers of genes affected by alternative splicing in unicellular eukaryotes are typically fewer than in plants and animals. However, it is important to examine this process in a diversity of eukaryotes, as we continue to find biologically relevant examples of alternative splicing in an increasing number of unicellular eukaryotes.

Alternative splicing often results in the introduction of premature termination codons into transcripts, which can mark them for degradation through the nonsense mediated decay pathway (Black 2003; Chang, Imam, Wilkinson 2007; Lareau *et al*. 2007a; Lareau *et al*. 2007b; Maquat 2004; Palusa and Reddy 2010; Zhang *et al*. 1998). It has even been suggested that alternative splicing is linked with mRNA degradation through a mechanism known as RUST, or, regulated unproductive splicing and translation (Lareau *et al*. 2007a; Lareau *et al*. 2007b; Palusa and

Reddy 2010). Studies in *Arabidopsis* show that NMD plays a major role in regulating alternative splice forms, as well as regulating levels of non-coding transcripts (Drechsel *et al*. 2013; Kalyna *et al*. 2011). Knocking down a crucial NMD pathway gene (Upf1) in *P. tetraurelia* results in increased levels of intron-retained transcripts in the mutants relative to wild-type, suggesting that functional NMD is crucial for regulating unspliced transcripts and avoiding the translation of these isoforms (Jaillon *et al*. 2008b). The authors even suggest that the NMD pathway may have co-evolved with introns as a mechanism to regulate splicing patterns (Jaillon *et al*. 2008b). Analysis of alternative splicing in *Y. lipolytica* found nearly all intron-retaining transcripts to contain premature stop codons, suggesting that they are targeted for degradation via the NMD pathway (Mekouar *et al*. 2010). Indeed, inactivation of Upf1 and Upf2 in *Y. lipolytica* results in increased levels of intron-containing isoforms (Mekouar *et al*. 2010). *C. merolae* does encode several NMD-pathway genes, including candidates for Dcp2, Ccr4, Dis3, two ORFs with similarity to Upf1, and a very strong Nmd3 candidate (Matsuzaki *et al*. 2004; Nozaki *et al*. 2007). As *Y. lipolytica* has functional NMD in the absence of Upf3, it is feasible that NMD is active in *C. merolae* without Upf2 and Upf3. Also, the presence of two ORFs annotated as similar to Upf1 in *C. merolae* leaves open the possibility that one or both of these have taken over the roles played by Upf2 and/or Upf3 (Matsuzaki *et al*. 2004; Nozaki *et al*. 2007). Similar to *Y. lipolytica*, a high frequency of *C. merolae* introns (86%) introduce premature stop codons when retained. Therefore, the presence of several NMD-related genes and high levels of stop codon-generating alternative splicing suggests that NMD-targeted alternative splice forms could be an important mechanism of regulation for intron-containing genes in *C. merolae*, and that the RUST mechanism could be active.

4.3 Conclusions

Transcriptome-wide gene expression analysis showed that more than half of *C. merolae* genes have significant changes in expression levels between day and night. I found evidence for several plastid-related genes being strongly down-regulated in the dark. High levels of antisense transcription were found for hundreds of *C. merolae* genes. Surprisingly, antisense transcripts that could form cis-NAT pairs appear to be more prevalent than in several metazoan species examined to date. Also, levels of antisense transcripts of 1189 genes differ significantly between light and dark growth. Antisense transcripts that are up-regulated in the dark are enriched for photosynthesis-related genes, while those up-regulated in the light are not, suggesting a possible role for regulation of photosynthetic genes by their antisense counterparts. Confirming the presence and function of these antisense transcripts in *C. merolae* will require analyses of small RNA populations.

RNA-seq data allowed for the detection of 16 new spliceosomal introns and confirmed active splicing of the previously annotated 27 introns. Introns in *C.* merolae seem to lack a bias towards phase-0 positioning, and most introns encode stop codons. These introns may be under selective pressure to cause premature stop codons when retained. Pre-mRNA splicing levels are extremely low in *C. merolae*, with high levels of intron retention seen for all junctions. The high levels of intron containing, non-functional transcripts are likely compensated for by NMD or an alternate decay pathway. Also, a large proportion of *C. merolae* introns undergo alternative splicing events, including alternate 5′ and 3′ splice site usage, intron retention, and even exon skipping in the one multi-intron gene. With only two alternative splicing regulating SR proteins annotated in *C. merolae*, discerning the exact mechanism of alternative splicing control will require further experimental work. The low splicing levels found in *C. merolae* are in line with

the transcriptome analysis of the microsporidian parasite *Encephalitozoon cuniculi*, another reduced eukaryotic system. Together, these findings of the lowest levels of splicing efficiency among eukaryotes in two unrelated species with highly reduced spliceosomes, suggests that spliceosomal reduction results in decreased splicing efficiency.

4.4 Materials and methods

4.4.1 Cell culture and RNA preparation

Cultures of *C. merolae* 10D (NIES strain 1332) were obtained from the NIES Microbial Culture Collection (Japan), and grown at 40°C under 90μmol of photons/m^2/s of light on a 12h:12h light:dark cycle. Erlenmeyer flasks (250mL) containing 50mL of modified Allen's media (MA2) media were shaking at 120rpm (Minoda *et al*. 2004). Cells were collected by centrifugation after 3 weeks of growth at the half-way point of both the light and dark cycles. Total RNA was extracted from two biological replicates of *C. merolae* for the light and dark conditions, using the Ambion RNAqueous kit (Ambion, Austin, TX). Extracted RNA was treated with TURBO DNase (Ambion, Austin, TX) to eliminate any contaminating DNA. RNA quality was assessed on an Agilent Bioanalyzer 2100 (Agilent, Santa Clara, CA) and RNA quantity was measured on a Qubit 2.0 fluorometer (Life Technologies Corp., Carlbad, CA).

4.4.2 RNA-seq library preparation

A total of eight Illumina libraries were prepared. Four libraries, two biological replicates of both light and dark condition cultures, were prepared according to the TruSeq library preparation protocol (Illumina, Hayward, CA). A total of 4μg of RNA from each of the four DNase-treated samples was used as starting material. Since the Illumina TruSeq libraries do not retain strand information, four additional libraries were made using the NEXTflex™ directional RNA-seq (dUTP) v2 library preparation kit from Bioo Scientific (Bioo Scientific, Austin, TX).

The directional RNA-seq libraries retain strand information, allowing us to identify whether a given read was transcribed from the plus or minus strand. A total of 2μg of RNA was used as input for the directional RNA-seq libraries. Library quality control and pooling were performed by the Biodiversity Research Centre (BRC) sequencing facility (UBC, Vancouver, BC).

4.4.3 Illumina sequencing and data processing

Paired-end sequencing was performed on an Illumina HiSeq 2000 at the BRC sequencing facility. The four TruSeq libraries were multiplexed and sequenced in two lanes in order to give two technical replicates for each of the biological replicates, and to help avoid bias associated with a particular flow cell or lane (Auer and Doerge 2010). Since standard paired-end RNA-seq does not account for possible antisense and overlapping transcription, I prepared directional RNA-seq libraries for all four samples. The four directional libraries were multiplexed and sequenced on a single lane of the HiSeq 2000.

Raw sequence data was processed and converted to fastq format. Sequence reads were mapped to the *C. merolae* 10D reference genome (Nozaki *et al*. 2007). The short-read aligner Tophat version 2.0.6 (Kim *et al*. 2013) was used for read mapping, allowing only a single alignment for each read. All biological replicates showed high levels of correlation (> 0.98), as has been noted in previous RNA-seq experiments (Bruno *et al*. 2010; Mortazavi *et al*. 2008; Nagalakshmi *et al*. 2008; Wilhelm *et al*. 2008). SAMtools version 0.1.18 (Li *et al*. 2009) was used to process SAM and BAM alignment files. Alignments were visualized using the Integrative Genomics Viewer version 2.0.7 (Robinson *et al*. 2011). Expression levels were measured in the standard fragments per kilobase per million mapped reads (FPKM) format (Mortazavi *et al*. 2008).

4.4.4 Assessing splicing efficiency

Illumina reads were mapped with Tophat, which splits reads over putative splice junctions and outputs alignments in SAM format (Kim *et al*. 2013). A custom Python software package was created specifically to call all possible types of alternative splicing events and provide counts of constitutive and alternative spliced reads. Splicing levels were calculated by dividing the number of reads split across a splice junction by the total number of reads, those either split over the junction or mapping within the intron. This gives a percentage of splicing efficiency, essentially a measure of the proportion of spliced transcripts versus intron-retaining transcripts.

In order to control for the difference in probability that a read will map over a splice junction versus within a large intronic region, splicing levels required normalization. The average insert size of mapped reads was estimated from the data and used as a measure of the frequency of fragmentation during library preparation. For a read to be called as mapping across a splice junction, it must overlap with the single position where the 5′ and 3′ exons joined together. Therefore, the size of fragments sequenced does not affect the mapping of these reads. On the other hand, reads mapping within introns are affected by the frequency at which fragmentation occurred during library preparation, as the higher the frequency of fragmentation (i.e. smaller fragments) the more likely reads from a single transcript will map to the same intron and result in an overestimation of intron retention events. Therefore, intron mapping reads were normalized by multiplying their total by the ratio of insert size over intron size. This normalizes for the number of times (on average) a transcript with an intron would be fragmented within the intron, leading to more than one read arising from the intron of a single transcript. Introns

smaller than the average insert size were not normalized since the probability of reads mapping to these introns is not affected by fragmentation size.

4.4.5 Differential gene expression analysis

All Tophat output alignment files were processed using a custom suite of Python software. The Balmung script was run in order to obtain raw read counts for all *C. merolae* annotated genes, including 80nt of 5′ and 3′ UTR, since *C. merolae* UTRs are not annotated. A value shorter than the length of sequenced reads was chosen as a conservative UTR size, as it will catch nearly all reads that overlap with the annotated gene, while being highly unlikely to overlap with neighboring genes since average intergenic length is approximately 1500bp. The Balmung output file was filtered in order to exclude non protein-coding genes and genes covered by fewer than twenty reads in any biological replicate. This filtered read count file was used as input for the program DESeq2 (Anders and Huber 2010), an R/Bioconductor package (Gentleman *et al*. 2004; R Development 2011). An adjusted p-value cut-off of 0.1, as suggested by the DESeq2 authors, was used to call differential expression between the two conditions tested. The adjusted p-value corrects for multiple comparisons and signifies a 10% false discovery rate. The *C. merolae* gene annotation files were obtained from the EnsemblPlants site (plants.ensembl.org/Cyanidioschyzon_merolae/) (Flicek *et al*. 2014).

4.4.6 Analysis of cis-NATs

BAM alignment files obtained Tophat mapped reads were processed with a custom Python package in order to produce raw read counts for each gene under each condition. Reads mapping in either sense or antisense orientations were filtered into separate files in order to have distinct sense and antisense read counts for each gene. The relative levels of antisense transcription for each gene was calculated by dividing the number of antisense reads by the total

number of mapped reads per gene. Prior to differential expression analysis, genes were filtered based on a minimum coverage of 20 antisense reads. Differential expression analysis was performed with DESeq2, as described above (Anders and Huber 2010). Gene ontology analysis was performed using *Arabidopsis thaliana* gene names obtained by selecting the top hit from BLAST searches of *C. merolae* protein sequences against *A. thaliana* protein sequences.

Table 4.1: Number of reads mapping to *Cyanidioschyzon merolae* reference genome

Library type	Strand-specific	Standard	Total reads mapped
Light	83056253	219371521	302427774
Dark	80652958	195584396	276237354
Total	163709211	414955917	578665128

Table 4.2: Frequency of alternative splicing events in *Cyanidioschyzon merolae* during light and dark growth

Numbers represent how many of the 43 introns display each type of alternative splicing event under light and dark conditions, from two RNAseq library preparation methods.

Library type	**Strand-specific**		**Standard**	
	Light	**Dark**	**Light**	**Dark**
Intron retention	43	43	43	43
Alternative 5′ splice site	1	1	3	5
Alternative 3′ splice site	13	13	23	24
Alternative positions	2	2	3	4
Exon skipping	1	1	1	1

Table 4.3: Characteristics of 40 *Cyanidioschyzon merolae* introns in protein-coding genes

	Phase 0	Phase 1	Phase 2
Total introns	13	11	16
Contain PTC	9	11	14
3n length	3	3	5

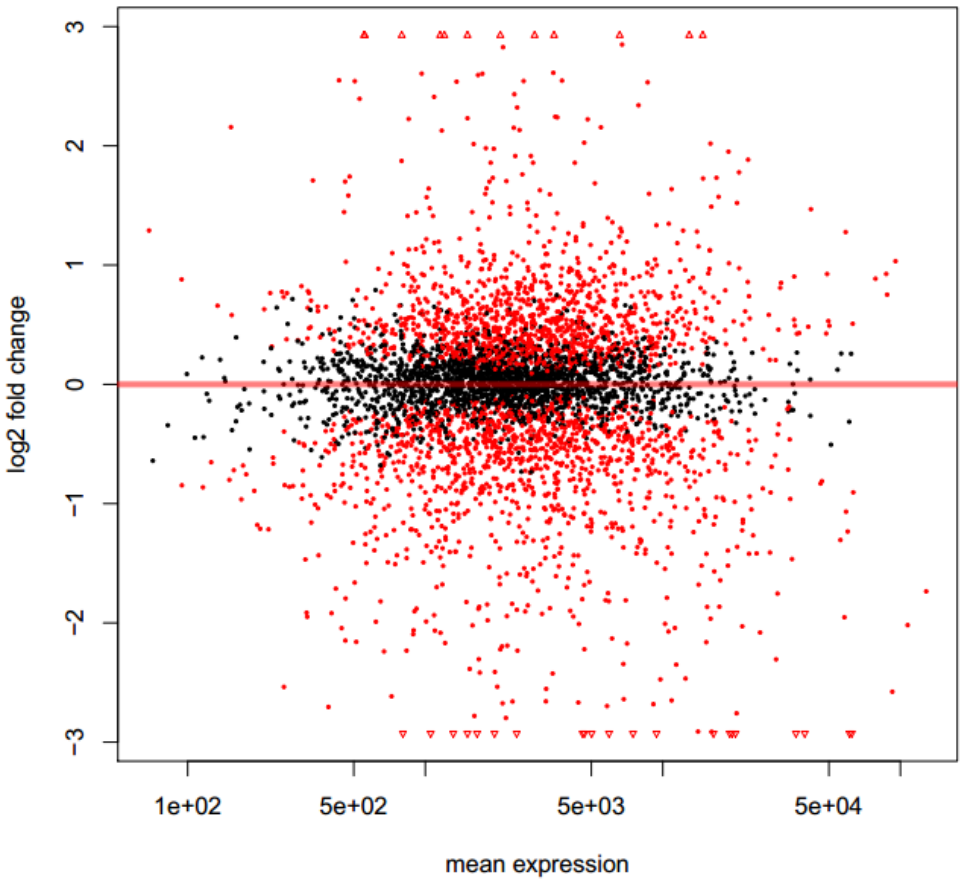


Figure 4.1: Differential expression of 4494 *Cyanidioschyzon merolae* genes during light and dark conditions

Plot of log2 fold change versus mean expression level. Red dots indicate those genes that are differentially expressed and black dots indicate those that are not.

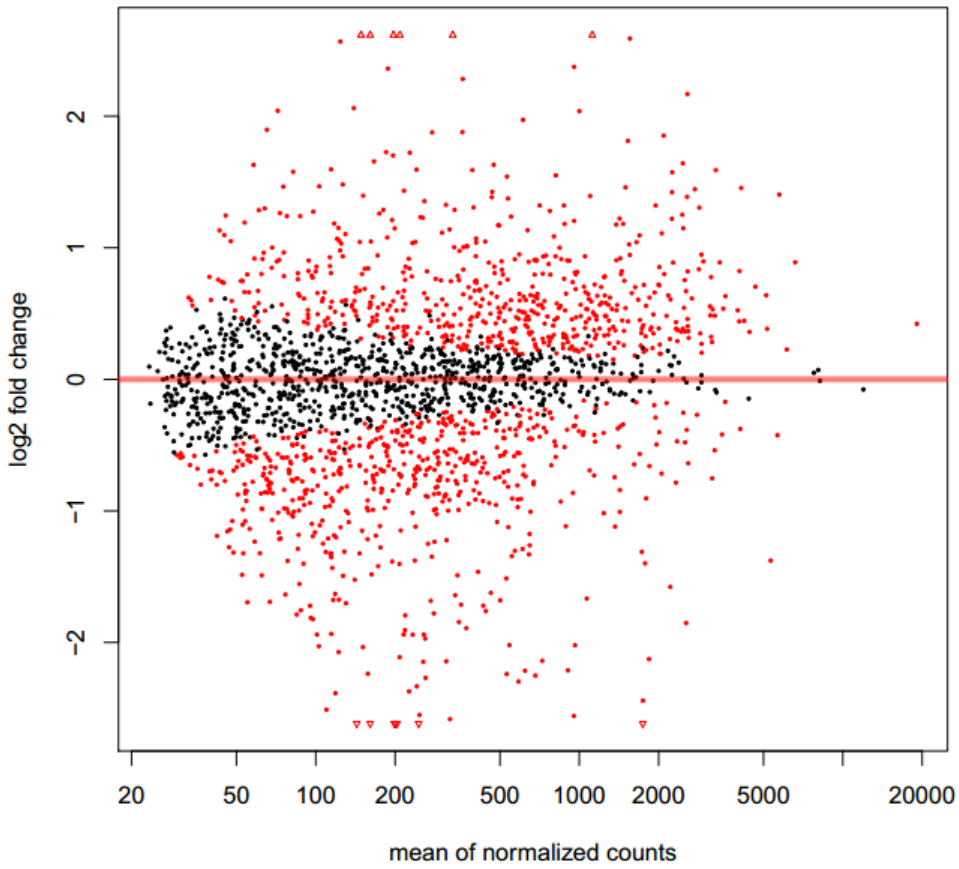


Figure 4.2: Differential expression of *Cyanidioschyzon merolae* antisense transcripts during light and dark conditions

Plot of log2 fold change versus mean expression level. Red dots indicate those genes that are differentially expressed and black dots indicate those that are not.

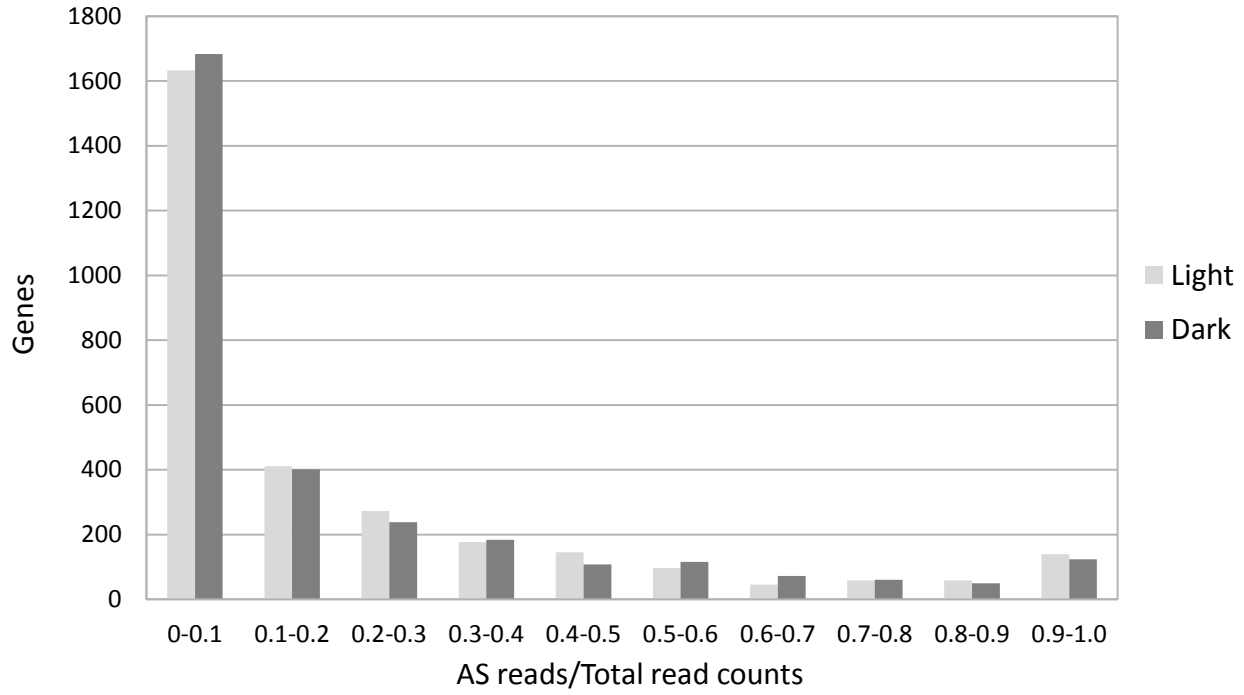


Figure 4.3: Distribution of antisense transcription levels for all *Cyanidioschyzon merolae* genes

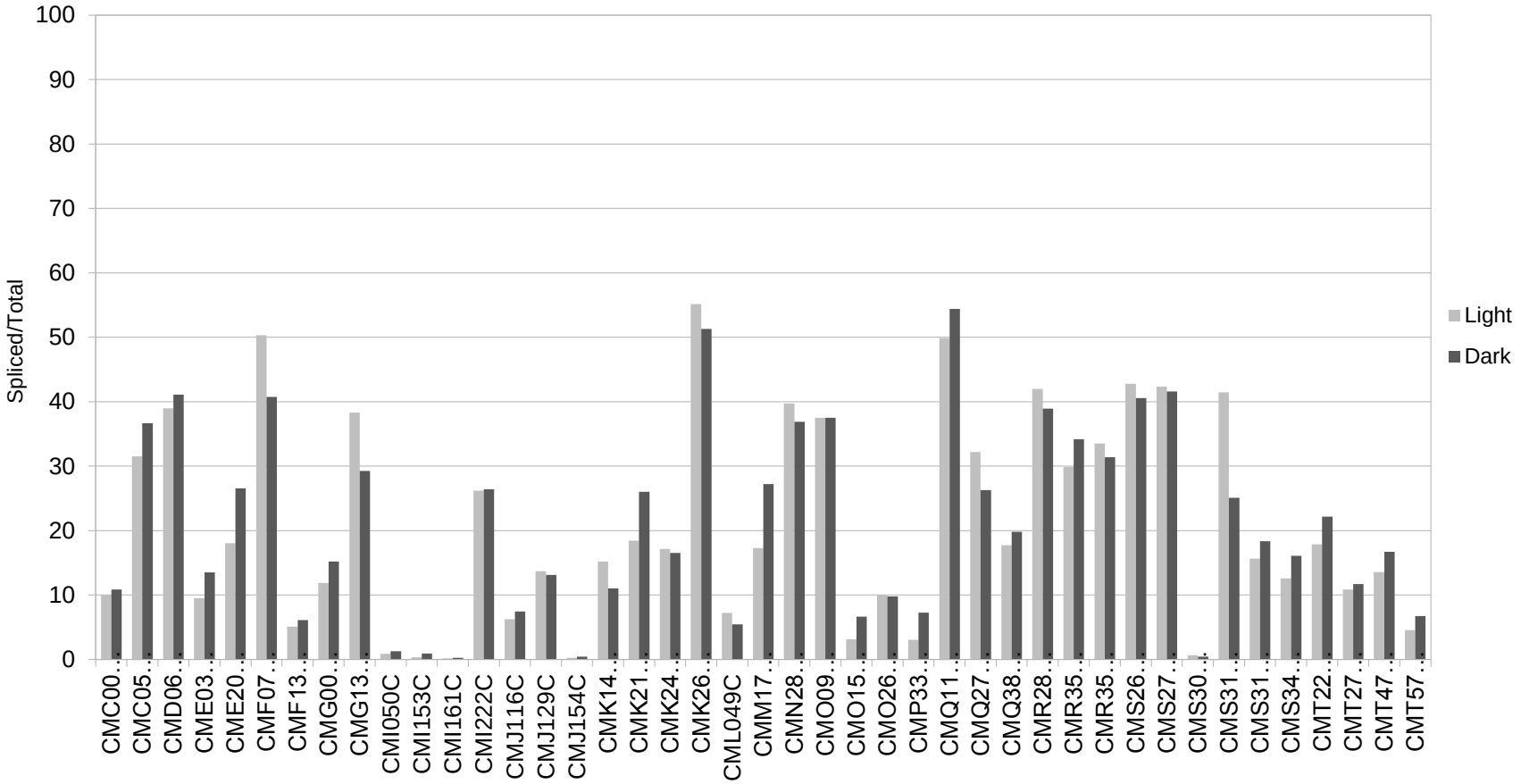


Figure 4.4: Levels of pre-mRNA splicing for all 43 introns in *Cyanidioschyzon merolae* during light and dark

Chapter 5: Conclusion

5.1 Summary

The results reported here shed light on the evolution of transcription and pre-mRNA splicing in eukaryotes. Examining these essential cellular processes in highly reduced systems has provided insight into the limitations of RNA processing machinery and the putative minimal spliceosomal core. The microsporidian *Encephalitozoon cuniculi* has one of the smallest, most gene-poor genomes among free living eukaryotes. Initial work examining transcription and splicing in this microsporidian parasite revealed many unusual features of intracellular and extracellular stage transcripts (Gill *et al*. 2010). We found the first evidence for differences in transcripts between two microsporidian life-stages (Gill *et al*. 2010), which set the ground work for the research described in Chapters 2 and 3. Finding unusually low splicing levels in *E. cuniculi* (Chapter 3) led me to examine the transcriptome of a distantly related eukaryote with a highly reduced spliceosome (Chapter 4). Altogether, these findings have helped shape our understanding of RNA processing following genome reduction.

Investigating the transcriptome of *E. cuniculi* allowed me to identify some of the shortest 5′UTRs discovered in a eukaryote to date (Grisdale and Fast 2011). I found that a large proportion of *E. cuniculi* transcripts have 5′UTRs of less than 20nt in length (see Figure 2.1), and some even lack 5′UTRs altogether (Grisdale and Fast 2011). A few cases of unicellular eukaryotes with short 5′UTRs (<25nt) have been noted, however, the overall trend in eukaryotes is a paucity of 5′UTRs less than 50nt in length (Ghosh *et al*. 1994; Iwabe and Miyata 2001; Liston and Johnson 1999; Lynch, Scofield, Hong 2005; Singh *et al*. 1997). While short 5′UTRs have been shown to be sufficient for translation, those shorter than 30nt in length can negatively

impact translation efficiency (Hughes and Andrews 1997; Lynch, Scofield, Hong 2005; Maicas, Shago, Friesen 1990; van den Heuvel *et al*. 1989). The discovery that the vast majority of *E. cuniculi* 5′UTR lengths fall in the 0-30nt range suggests that short 5′UTRs do not have significantly reduced translation efficiency in this microsporidian; perhaps short 5′UTRs are selected for advantages unrelated to translation. My transcriptome data shows that *E. cuniculi* genes are nearly all expressed at a relatively high rate, during the intracellular stage of its life cycle. The compact nature of *E. cuniculi* UTRs may be prevalent because they promote high rates of expression for most genes, which could be beneficial for rapid growth during the proliferative stage. Constitutive, high expression levels have been noted in other reduced systems, such as those of nucleomorph genomes (Tanifuji *et al*. 2014). My work provides evidence that short 5′UTRs on intracellular stage transcripts of a microsporidian parasite are associated with the high levels of transcription required during proliferative growth, and that long 5′UTRs are reserved for the few developmentally regulated genes.

High-throughput sequencing of the *C. merolae* transcriptome has revealed two surprising results: a high frequency of alternative splicing events among the small number of introns, and high levels of antisense transcription at hundreds of annotated loci. Alternative splicing is a complex process that has mostly been studied in model multicellular organisms, with relatively few unicellular eukaryotes analyzed to date (Curtis *et al*. 2012; Escalante, Moreno, Sastre 2003; Iriko *et al*. 2009; Jaillon *et al*. 2008b; Labadorf *et al*. 2010; Maniatis and Tasic 2002; Mekouar *et al*. 2010). Characterizing alternative splicing events in *C. merolae* provides new data for an emerging model unicellular eukaryote. Although few spliceosomal components have been retained in *C. merolae*, I find several major types of alternative splicing events occurring with widely varying frequencies. Similar to plants and most other unicellular eukaryotes, intron

retention is the most frequently occurring event, and was found at high levels for all 43 introns (see Figure 4.4) (Grisdale *et al*. 2013; Jaillon *et al*. 2008b; Kim, Magen, Ast 2007; Labadorf *et al*. 2010; Mekouar *et al*. 2010; Wang and Brendel 2006). I also find alternative 5′ and 3′ splice site usage in roughly half of the introns, and a single case of exon skipping in the one multi-intron gene. Finding frequent alternative splicing events is particularly surprising given the limited number of alternative splicing regulators encoded in the *C. merolae* genome. As described in organisms from several lineages, the high levels of alternative splicing, notably intron retention, could be playing a regulatory role in gene expression by inducing transcript decay via the NMD pathway (Black 2003; Chang, Imam, Wilkinson 2007; Drechsel *et al*. 2013; Jaillon *et al*. 2008b; Kalyna *et al*. 2011; Lareau *et al*. 2007a; Lareau *et al*. 2007b; Maquat 2004; Mekouar *et al*. 2010; Palusa and Reddy 2010; Zhang *et al*. 1998). Questions still exist as to whether the intron retention events in these reduced systems are regulatory or are simply by-products of low splicing efficiency. Currently, it is not possible to distinguish potential regulatory consequences of intron retention from low splicing efficiency leading to high levels of unspliced transcripts that are tolerated by the cell. Further work deducing the mechanisms of pre-mRNA splicing and alternative splicing in this system with a highly reduced spliceosome should provide clues as to the co-evolution between spliceosomal machinery and introns, as well as the minimal number of spliceosomal components required for regulating alternative splicing.

Transcriptome-wide analyses of *E. cuniculi* and *C. merolae* identified new introns and found low splicing levels at all junctions (see Figures 3.2 and 4.4). These levels of splicing are much lower than in other eukaryotes, including species of the same lineage and similar lifestyle (Grisdale *et al*. 2013). Therefore, I propose that splicing levels in *E. cuniculi* and *C. merolae* are low because the pressures of reductive evolution have altered the canonical splicing pathway in

both species. These two systems have undergone reductive evolution under vastly different constraints. *E. cuniculi* has been under extreme reductive pressures as a result of its intracellular, parasitic lifestyle (Katinka *et al*. 2001; Keeling, P.J.,and Fast,N.M. 2002; Slamovits *et al*. 2004). Although the genome of *C. merolae* is larger than that of *E. cuniculi* by a fair margin, reduction has taken place in this red algal species. While the nature of the reductive pressures is not certain, links between cell size of thermophiles and reduction in genome size have been noted in several lineages (Sabath *et al*. 2013; van Noort *et al*. 2013). That these two species have independently lost many spliceosomal components (see Table 1.1) as a result of different reductive pressures strengthens our conclusion that low splicing efficiency is the result of spliceosomal reduction. Also, it seems likely that the lack of a U1 snRNA in these two species has an effect on the mechanism of splicing, possibly influencing efficiency. Further work examining this loss, and the mechanism of pre-mRNA splicing in *C. merolae*, will help to test this hypothesis (see Future directions).

5.2 Future directions

In collaboration with the Rader lab (UNBC) I am working on characterizing the reduced spliceosome of *C. merolae*. The predicted loss of one of five essential snRNA components of the spliceosome has led us to speculate on the potential mechanism of splicing in the absence of the U1 snRNP molecule. This relates to work in Chapters 3 and 4, as both *E. cuniculi* and *C. merolae* have functional spliceosomes that putatively lack U1 snRNAs: the only characterized examples of splicing in systems missing an snRNA. Using biochemical assays, RNA has been purified for molecules containing the unique 5′ cap structure found on the U1, U2, U4, and U5 snRNAs. I performed modified cDNA library preparations from this purified RNA, and

sequenced these libraries on an Illumina HiSeq machine to obtain high depth. All known snRNAs (U2, U4, U5, U6) were recovered. However, nothing resembling a U1 snRNA was retrieved after searching for transcripts containing U1-specific motifs or conserved secondary structure. This is further evidence that U1 has been lost in *C. merolae*. Upon further inspection of snRNA sequences and intron motifs, we found the potential for novel base-pair interactions between the U5 snRNA and the 5′ splice site of *C. merolae* introns. Therefore, in addition to its canonical pre-mRNA-spliceosome interactions, it appears that the U5 snRNA may be compensating for the lack of U1 snRNA by binding to the 5′ splice site motif and potentially taking over the role of intron recognition. This indicates the potential for plasticity in the spliceosomal machinery and the power of co-evolution between introns and an atypical spliceosome.

The high levels of antisense transcription found in *C. merolae* were unexpected. Frequent antisense transcription covering annotated genes raises the possibility that sense-antisense RNA pairs can form. Typically, sense-antisense transcript pairs are processed into short interfering RNAs (siRNAs). siRNAs are known to play roles in regulating gene expression at many levels, such as: DNA methylation, transcription, pre-mRNA splicing, alternative splicing, mRNA stability, and translation (Borsani *et al*. 2005; Hastings *et al*. 1997; Misra *et al*. 2002; Morrissy, Griffith, Marra 2011; Peters *et al*. 2003; Prescott and Proudfoot 2002). In order to assess whether sense-antisense pairs are being processed into siRNAs in *C. merolae*, a transcriptome-wide small RNA analysis is currently being performed. Typically, formation and function of small RNAs, such as miRNAs and siRNAs, requires proteins such as Dicer and Argonaute (Baulcombe 2005; Vaucheret 2006; Voinnet 2009). While one study has computationally predicted miRNAs in *C. merolae*, to date, no Dicer-like or Argonaute-like genes have been annotated in its genome

(Huang *et al*. 2011; Matsuzaki *et al*. 2004; Nozaki *et al*. 2007). With predicted miRNAs and evidence from my transcriptomic data revealing the presence of sense-antisense pairs that could form siRNAs, it seems plausible that small RNAs are present in *C. merolae*, but are perhaps processed by non-canonical or highly divergent proteins. There are several examples in the literature of small RNAs being processed via non-canonical pathways, including in a Dicer-independent manner (Castellano and Stebbing 2013; Havens *et al*. 2012; Marasovic, Zocco, Halic 2013; Yang, Maurin, Lai 2012). Therefore, it is possible that small RNAs function in *C. merolae*, but are processed via a non-canonical pathway. The small RNA analysis that is underway will answer questions raised in Chapter 4 regarding the presence and potential roles of siRNAs in *C. merolae*. While my research indicates that there are similarities in pre-mRNA splicing in two unrelated eukaryotes in response to reduction, this additional work will contribute to defining the changes in the spliceosome and its interactions with introns as a result of reductive pressures.

Bibliography

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S *et al*. (20 co-authors). 2004. Complete genome sequence of the apicomplexan, Cryptosporidium parvum. Science 304:441-445.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*. (195 co-authors). 2000. The genome sequence of Drosophila melanogaster. Science 287:2185-2195.

Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M. 2010. Comparison and calibration of transcriptome data from RNA-seq and tiling arrays. BMC Genomics 11:383.

Akiyoshi DE, Morrison HG, Lei S, Feng XC, Zhang QS, Corradi N, Mayanja H, Tumwine JK, Keeling PJ, Weiss LM *et al*. (11 co-authors). 2009. Genomic survey of the non-cultivatable opportunistic human pathogen, Enterocytozoon bieneusi. Plos Pathogens 5(1): e1000261.

Anantharaman V, Koonin EV, Aravind L. 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. 30:1427-1464.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106.

Ast G. 2004. How did alternative splicing evolve? Nat. Rev. Genet. 5:773-782.

Auer PL, Doerge RW. 2010. Statistical design and analysis of RNA sequencing data. Genetics 185:405-416.

Baim SB, Sherman F. 1988. mRNA structures influencing translation in the yeast Saccharomyces cerevisiae. Mol. Cell. Biol. 8:1591-1601.

Baulcombe D. 2005. RNA silencing. Trends Biochem. Sci. 30:290-293.

Bentley D. 2002. The mRNA assembly line: Transcription and processing machines in the same factory. Curr. Opin. Cell Biol. 14:336-342.

Bigliardi E, Sacchi L. 2001. Cell biology and invasion of the microsporidia. Microbes Infect. 3:373-379.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. 72:291-336.

Blasing OE, Gibon Y, Gunther M, Hohne M, Morcuende R, Osuna D, Thimm O, Usadel B, Scheible WR, Stitt M. 2005. Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. Plant Cell 17:3257-3281.

Bolte K, Bullmann L, Hempel F, Bozarth A, Zauner S, Maier UG. 2009. Protein targeting into secondary plastids. J. Eukaryot. Microbiol. 56:9-15.

Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK. 2005. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. Cell 123:1279-1291.

Brosson D, Kuhn L, Delbac F, Garin J, Vivares CP, 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.

Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, Sherlock G, Snyder M. 2010. Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. Genome Res. 20:1451-1458.

Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. Mol. Cell 2:773-785.

Cali A, Takvorian PM. 1999. Developmental morphology and life cycles of the microsporidia. In: Wittner M, Weiss LM, editors. The Microsporidia and Microsporidiosis. Washington, DC, USA: ASM Press. p. 85-128.

Castellano L, Stebbing J. 2013. Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues. Nucleic Acids Res. 41:3339-3351.

Cavaliersmith T. 1987. Molecular evolution - eukaryotes with no mitochondria. Nature 326:332-333.

Cavalier-Smith T. 1983. A 6-kingdom classification and a unified phylogeny. In: Schenk HEA, Schwemmler WS, editors. *Endocytobiology. II. Intracellular Space as Oligogenetic*. Berlin: Walter de Gruyter. p. 1027-1034.

Chang YF, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. Annu. Rev. Biochem. 76:51-74.

Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD. 2004. Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res. 32:4812-4820.

Chen Y, Pettis JS, Zhao Y, Liu X, Tallon LJ, Sadzewicz LD, Li R, Zheng H, Huang S, Zhang X *et al*. (15 co-authors). 2013. Genome sequencing and comparative genomics of honey bee

microsporidia, Nosema apis reveal novel insights into host-parasite interactions. BMC Genomics 14:451-2164-14-451.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ *et al*. (11 co-authors). 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2:65-73.

Chung S, Perry R. 1989. Importance of introns for expression of mouse ribosomal protein gene rpL32. Mol. Cell. Biol. 9:2075-2082.

Ciniglia C, Yoon HS, Pollio A, Pinto G, Bhattacharya D. 2004. Hidden biodiversity of the extremophilic Cyanidiales red algae. Mol. Ecol. 13:1827-1838.

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

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

Corradi N, Burri L, Keeling PJ. 2008. mRNA processing in Antonospora locustae spores. Mol. Genet. Genomics 280:565-574.

Corradi N, Pombert JF, Farinelli L, Didier ES, Keeling PJ. 2010a. The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nat Commun 21:77.

Corradi N, Pombert JF, Farinelli L, Didier ES, Keeling PJ. 2010b. The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nat. Commun. 1:77.

Corradi N, Haag KL, Pombert JF, Ebert D, Keeling PJ. 2009. Draft genome sequence of the daphnia pathogen Octosporea bayeri: Insights into the gene content of a large microsporidian genome and a model for host-parasite interactions. Genome Biol. 10(10):R106.

Corradi N, Keeling PJ. 2009. Microsporidia: A journey through radical taxonomical revisions. Fungal Biology Reviews 23:1-8.

Cuomo CA, Desjardins CA, Bakowski MA, Goldberg J, Ma AT, Becnel JJ, Didier ES, Fan L, Heiman DI, Levin JZ *et al*. (13 co-authors). 2012. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. Genome Res. 22(12):2478-2488.

Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y *et al*. (72 co-authors). 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492:59-65.

Dabeva MD, Post-Beittenmiller M, Warner JR. 1986. Autogenous regulation of splicing of the transcript of a yeast ribosomal protein gene. Proc. Natl. Acad. Sci. U. S. A. 83:5854-5857.

Dabeva MD, Warner JR. 1993. Ribosomal protein L32 of Saccharomyces cerevisiae regulates both splicing and translation of its own transcript. J. Biol. Chem. 268:19669-19674.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. USA 103:5320-5325.

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

Davis CA, Grate L, Spingola M, Ares MJ. 2000. Test of intron prediction reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. Nucleic Acids Res. 28:1700-1706.

Day DA, Tuite MF. 1998. Post-transcriptional gene regulatory mechanisms in eukaryotes: An overview. J. Endocrinol. 157:361-371.

De Luca P, Taddei R, Varano L. 1978. Cyanidioschyzon merolae: A new alga of thermal acidic environments. Webbia 33:37-44.

Delbac F, Polonais V. 2008. The microsporidian polar tube and its role in invasion. Subcell. Biochem. 47:208-220.

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R *et al*. (26 co-authors). 2006. Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proc. Natl. Acad. Sci. USA 103:11647-11652.

Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. 27:3219-3228.

Dietrich RC, Incorvaia R, Padgett RA. 1997. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol. Cell 1:151-160.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. Nature 410:1091-1096.

Drechsel G, Kahles A, Kesarwani AK, Stauffer E, Behr J, Drewe P, Ratsch G, Wachter A. 2013. Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. Plant Cell 25:3726-3742.

Dunn, A.M.,and Smith,J.E. 2001. Microsporidian life cycles and diversity: The relationship between virulence and transmission. Microbes Infect. 3:381-388.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. Trends Genet. 19:362-365.

Engebrecht JA, Voelkel-Meiman K, Roeder GS. 1991. Meiosis-specific RNA splicing in yeast. Cell 66:1257-1268.

Escalante R, Moreno N, Sastre L. 2003. Dictyostelium discoideum developmentally regulated genes whose expression is dependent on MADS box transcription factor SrfA. Eukaryot. Cell. 2:1327-1335.

Fedorov A, Suboch G, Bujakov M, Fedorova L. 1992. Analysis of nonuniformity in intron phase distribution. Nucleic Acids Res. 20:2553-2557.

Fewell SW, Woolford JL. 1999. Ribosomal protein S14 of saccharomyces cerevisiae regulates its expression by binding to RPS14B pre-mRNA and to 18S rRNA. Mol. Cell. Biol. 19:826-834.

Fischer J, Tran D, Juneau R, Hale-Donze H. 2008. Kinetics of encephalitozoon spp. infection of human macrophages. J. Parasitol. 94:169-175.

Fitzgerald KD, Semler BL. 2009. Bridging IRES elements in mRNAs to the eukaryotic translation apparatus. Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1789:518-528.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S *et al*. (56 co-authors). 2014. Ensembl 2014. Nucleic Acids Res. 42:D749-55.

Fokin SI, Di Giuseppe G, Erra F, Dini F. 2008. Euplotespora binucleata n. gen., n. sp. (protozoa: Microsporidia), a parasite infecting the hypotrichous ciliate Euplotes woodruffi, with observations on microsporidian infections in ciliophora. J. Eukaryot. Microbiol. 55:214-228.

Franzén O, Jerlström-Hultqvist J, Castro E, Sherwood E, Ankarklev J, Reiner DS, Palm D, Andersson JO, Andersson B, Svärd SG. 2009. Draft genome sequencing of Giardia intestinalis assemblage B isolate GS: Is human giardiasis caused by two different species? PLoS Pathog 5:e1000560.

Franzen C. 2004. Microsporidia: How can they invade other cells? Trends Parasitol. 20:275-279.

Friz CT. 1968. The biochemical composition of the free-living amoebae Chaos chaos, Amoeba dubia and Amoeba proteus. Comp. Biochem. Physiol. 26:81-90.

Fujiwara T, Ohnuma M, Yoshida M, Kuroiwa T, Hirano T. 2013. Gene targeting in the red alga Cyanidioschyzon merolae: Single- and multi-copy insertion using authentic and chimeric selection markers. PLoS One 8:e73608.

Fujiwara T, Misumi O, Tashiro K, Yoshida Y, Nishida K, Yagisawa F, Imamura S, Yoshida M, Mori T, Tanaka K *et al*. (12 co-authors). 2009. Periodic gene expression patterns during the highly synchronized cell nucleus and organelle division cycles in the unicellular red alga Cyanidioschyzon merolae. DNA Res. 16:59-72.

Gardner MJ. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419:498-511.

Gene Ontology Consortium. Creating the gene ontology resource: Design and implementation. Genome Res. 2001 Aug;11(8):1425-33.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*. (25 co-authors). 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 5:R80.

Ghosh S, Jaraczewski JW, Klobutcher LA, Jahn CL. 1994. Characterization of transcription initiation, translation initiation, and poly(A) addition sites in the gene-sized macronuclear DNA molecules of euplotes. Nucleic Acids Res. 22:214-221.

Gill EE, Fast NM. 2007. Stripped-down DNA repair in a highly reduced parasite. BMC Mol. Biol. 20:24.

Gill EE, Fast NM. 2006. Assessing the microsporidia-fungi relationship: Combined phylogenetic analysis of eight genes. Gene 375:103-109.

Gill EE, Lee RC, Corradi N, Grisdale CJ, Limpright VO, Keeling PJ, Fast NM. 2010. Splicing and transcription differ between spore and intracellular life stages in the parasitic microsporidia. Mol. Biol. Evol. 27:1579-1584.

Gilson P, Mcfadden GI. 1995. The chlorarachniophyte - a cell with 2 different nuclei and 2 different telomeres. Chromosoma 103:635-641.

Gilson PR, McFadden GI. 1996. The miniaturized nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. Proc. *Natl*. Acad. Sci. U. S. A. 93:7737-7742.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus. Proc. Natl. Acad. Sci. USA 103:9566-9571.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*. (16 co-authors). 1996. Life with 6000 genes. Science 274:563-567.

Goodman AJ, Daugharthy ER, Kim J. 2013. Pervasive antisense transcription is evolutionarily conserved in budding yeast. Mol. Biol. Evol. 30:409-421.

Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. Annu. Rev. Plant. Biol. 59:491-517.

Graveley BR, Nilsen TW. 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463:457.

Gray MW, Lang BF, Burger G. 2004. Mitochondria of protists. Annu. Rev. Genet. 38:477-524.

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2005. Eukaryotic genome size databases. - Nucleic Acids Res.2007 Jan;35 (Database Issue):D332-8.Epub 2006 Nov 7.

Grisdale CJ, Fast NM. 2011. Patterns of 5' untranslated region length distribution in Encephalitozoon cuniculi: Implications for gene regulation and potential links between transcription and splicing. J. Eukaryot. Microbiol. 58:68-74.

Grisdale C, Bowers L, Didier E, Fast N. 2013. Transcriptome analysis of the parasite Encephalitozoon cuniculi: An in-depth examination of pre-mRNA splicing in a reduced eukaryote. BMC Genomics 14:207.

Gross W, Heilmann I, Lenze D, Schnarrenberger C. 2001. Biogeography of the Cyanidiaceae (rhodophyta) based on 18S ribosomal RNA sequence data. Eur. J. Phycol. 36:275-280.

Hall SL, Padgett RA. 1996. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. Science 271:1716-1718.

Hastings ML, Milcarek C, Martincic K, Peterson ML, Munroe SH. 1997. Expression of the thyroid hormone receptor gene, erbAalpha, in B lymphocytes: Alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels. Nucleic Acids Res. 25:4296-4300.

Havens MA, Reich AA, Duelli DM, Hastings ML. 2012. Biogenesis of mammalian microRNAs by a non-canonical processing pathway. Nucleic Acids Res. 40:4626-4640.

He F, Peltz SW, Donahue JL, Rosbash M, Jacobson A. 1993. Stabilization and ribosome association of unspliced pre-mRNAs in a yeast upf1- mutant. Proc. Natl. Acad. Sci. U. S. A. 90:7034-7038.

Henras A, Henry Y, Bousquet-Antonelli C, Noaillac-Depeyre J, Gelugne JP, Caizergues-Ferrer M. 1998. Nhp2p and Nop10p are essential for the function of H/ACA snoRNPs. EMBO J. 17:7078-7090.

Howe KJ. 2002. RNA polymerase II conducts a symphony of pre-mRNA processing activities. Biochim. Biophys. Acta 1577:308-324.

Huang A, Wu X, Wang G, Jia Z, He L. 2011. Computational prediction of microRNAs and their targets from three unicellular algae species with complete genome sequences. Can. J. Microbiol. 57:1052-1061.

Hughes MJ, Andrews DW. 1997. A single nucleotide is a sufficient 5' untranslated region for translation in an eukaryotic in vitro system. FEBS Lett. 414:19-22.

Hurowitz EH, Brown PO. 2003. Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae. Genome Biol. 5:R2.

Imamura S, Terashita M, Ohnuma M, Maruyama S, Minoda A, Weber AP, Inouye T, Sekine Y, Fujita Y, Omata T *et al*. (11 co-authors). 2010. Nitrate assimilatory genes and their transcriptional regulation in a unicellular red alga Cyanidioschyzon merolae: Genetic evidence for nitrite reduction by a sulfite reductase-like enzyme. Plant Cell Physiol. 51:707-717.

Iriko H, Jin L, Kaneko O, Takeo S, Han ET, Tachibana M, Otsuki H, Torii M, Tsuboi T. 2009. A small-scale systematic analysis of alternative splicing in Plasmodium falciparum. Parasitol. Int. 58:196-199.

Irimia M, Roy SW. 2008. Evolutionary convergence on highly-conserved 3 ' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. Plos Genetics 4.

Irimia M, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. Trends in Genetics 23:321-325.

Iwabe N, Miyata T. 2001. Overlapping genes in parasitic protist giardia lamblia. Gene 280:163-167.

Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Segurens B *et al*. (19 co-authors). 2008b. Translational control of intron splicing in eukaryotes. Nature 451:359-U15.

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J *et al*. (70 co-authors). 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443:818-822.

Jen CH, Michalopoulos I, Westhead DR, Meyer P. 2005. Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation. Genome Biol. 6:R51.

Jorgenson JW, Lukacs KD. 1983. Capillary zone electrophoresis. Science 222:266-272.

Juneau K, Palm C, Miranda M, Davis RW. 2007. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. Proc. Natl. Acad. Sci. USA 104:1522-1527.

Jurica MS, Moore MJ. 2003. Pre-mRNA splicing: Awash in a sea of proteins. Mol. Cell 12:5-14.

Kabran P, Rossignol T, Gaillardin C, Nicaud JM, Neuveglise C. 2012. Alternative splicing regulates targeting of malate dehydrogenase in Yarrowia lipolytica. DNA Res. 19:231-244.

Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, Marshall J, Fuller J, Cardle L, McNicol J *et al*. (13 co-authors). 2011. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. Nucleic Acids Res. .

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J *et al*. (34 co-authors). 2005. Antisense transcription in the mammalian transcriptome. Science 309:1564-1566.

Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P *et al*. (17 co-authors). 2001. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature 414:450-453.

Kaufer NF, Potashkin J. 2000. Analysis of the splicing machinery in fission yeast: A comparison with budding yeast and mammals. Nucleic Acids Res. 28:3003-3010.

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

Keeling PJ. 2003a. Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. Fungal Genetics and Biology 38:298-309.

Keeling PJ. 2003b. Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. Fungal Genet. Biol. 38:298-309.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9:605-618.

Keeling PJ, Slamovits CH. 2004. Simplicity and complexity of microsporidian genomes. Eukaryotic Cell 3:1363-1369.

Keeling PJ, Luker MA, Palmer JD. 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. Mol. Biol. Evol. 17:23-31.

Keeling PJ, Corradi N, Morrison HG, Haag KL, Ebert D, Weiss LM, Akiyoshi DE, Tzipori S. 2010. The reduced genome of the parasitic microsporidian Enterocytozoon bieneusi lacks genes for core carbon metabolism. Genome Biol Evol 12:304-309.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49-63.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.

Kim E, Goren A, Ast G. 2008. Alternative splicing: Current perspectives. Bioessays 30:38-47.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35:125-131.

Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: A compromise solution to the introns-early versus introns-late debate? Biology Direct 1.

Kornblihtt AR, De La Mata M, Fededa JP, Munoz MJ, Nogues G. 2004. Multiple links between transcription and splicing. RNA 10:1489-1498.

Kozak M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283-292.

Kuroiwa T. 1998. The primitive red algae: Cyanidium caldarium and Cyanidioschyzon merolae as model system for investigating the dividing apparatus of mitochondria and plastids. Bioessays 20:344-354.

Labadorf A, Link A, Rogers MF, Thomas J, Reddy AS, Ben-Hur A. 2010. Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii. BMC Genomics 11:114-2164-11-114.

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald JM. 2007. Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. Proc. Natl. Acad. Sci. U. S. A. 104:19908-19913.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

Lapidot M, Pilpel Y. 2006. Genome-wide natural antisense transcription: Coupling its regulation to its different regulatory mechanisms. EMBO Rep. 7:1216-1222.

Lareau LF, Brooks AN, Soergel DA, Meng Q, Brenner SE. 2007a. The coupling of alternative splicing and nonsense-mediated mRNA decay. Adv. Exp. Med. Biol. 623:190-211.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007b. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature 446:926-929.

Larsson JI. 2000. The hyperparasitic microsporidium amphiacantha longa caullery et mesnil, 1914 (microspora: Metchnikovellidae) - description of the cytology, redescription of the species, emended diagnosis of the genus amphiacantha and establishment of the new family amphiacanthidae. Folia. Parasitol. (Praha) 47:241-256.

Lee RC, Gill EE, Roy SW, Fast NM. 2010. Constrained intron structures in a microsporidian. Mol. Biol. Evol. 27:1979-1982.

Lee SC, Corradi N, Byrnes EJ3, Torres-Martinez S, Dietrich FS, Keeling PJ, Heitman J. 2008. Microsporidia evolved from ancestral sexual fungi. Curr. Biol. 18:1675-1679.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods 7:709-715.

Li B, Vilardell J, Warner JR. 1996. An RNA structure involved in feedback regulation of splicing and of translation is critical for biological fitness. Proc. Natl. Acad. Sci. USA 93:1596-1600.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data,Processing Subgroup. 2009. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078-2079.

Li S, Liberman LM, Mukherjee N, Benfey PN, Ohler U. 2013. Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. Genome Res. 23:1730-1739.

Li SW, Feng L, Niu DK. 2007. Selection for the miniaturization of highly expressed genes. Biochem. Biophys. Res. Commun. 360:586-592.

Liston DR, Johnson PJ. 1999. Analysis of a ubiquitous promoter element in a primitive eukaryote: Early evolution of the initiator element. Mol. Cell. Biol. 19:2380-2388.

Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA *et al*. (54 co-authors). 2005. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. Science 307:1321-1324.

Long MY, Rosenberg C, Gilbert W. 1995. Intron phase correlations and the evolution of the intron exon structure of genes. Proc. Natl. Acad. Sci. U. S. A. 92:12495-12499.

Long MY, De Souza SJ, Rosenberg C, Gilbert WE. 1998. Relationship between "proto-splice sites" and intron phases: Evidence from dicodon analysis. Proc. Natl. Acad. Sci. U. S. A. 95:219-223.

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

Luehrsen KR, Walbot V. 1991. Intron enhancement of gene expression and the splicing efficiency of introns in maize cells. Molecular and General Genetics MGG 225:81-93.

Lynch M, Scofield DG, Hong X. 2005. The evolution of transcription-initiation sites. Mol. Biol. Evol. 22:1137-1146.

Maicas E, Shago M, Friesen JD. 1990. Translation of the Saccharomyces cerevisiae tcm1 gene in the absence of a 5'-untranslated leader. Nucleic Acids Res. 18:5823-5828.

Maniatis T, Reed R. 2002. An extensive network of coupling among gene expression machines. Nature 416:499-506.

Maniatis T, Tasic B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418:236-243.

Maquat LE. 2004. Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. Nat. Rev. Mol. Cell Biol. 5:89-99.

Marasovic M, Zocco M, Halic M. 2013. Argonaute and triman generate dicer-independent priRNAs and mature siRNAs to initiate heterochromatin formation. Mol. Cell 52:173-183.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18:1509-1517.

Marquez Y, Brown JWS, Simpson C, Barta A, *Kalyna* M. 2012. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res. 22:1184-1195.

Marshall AN, Montealegre MC, Jimenez-Lopez C, Lorenz MC, van Hoof A. 2013. Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. PLoS Genet. 9:e1003376.

Martin W, Herrmann RG. 1998. Gene transfer from organelles to the nucleus: How much, what happens, and why? Plant Physiol. 118:9-17.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. U. S. A. 99:12246-12251.

Matsuzaki M, Misumi O, Shin-i T, Maruyama S, Takahara M, Miyagishima S, Mori T, Nishida K, Yagisawa F, Yoshida Y *et al*. (42 co-authors). 2004. Genome sequence of the ultra-small unicellular red alga Cyanidioschyzon merolae 10D. Nature 428:653-657.

McCracken S, Fong N, Yankulov K, Ballantyne S, Pan G, Greenblatt J, Patterson SD, Wickens M, Bentley DL. 1997. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature 385:357-361.

McCutcheon JP, Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13-26.

McFadden GI. 1999. Plastids and protein targeting. J. Eukaryot. Microbiol. 46:339-346.

McGuire, AM, Pearson, MD, Neafsey, DE, Galagan, JE. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol. 9:R50.

Mekouar M, Blanc-Lenfle I, Ozanne C, Da Silva C, Cruaud C, Wincker P, Gaillardin C, Neuveglise C. 2010. Detection and analysis of alternative splicing in Yarrowia lipolytica reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. Genome Biol. 11:R65-2010-11-6-r65. Epub 2010 Jun 23.

Menges M, Hennig L, Gruissem W, Murray JA. 2002. Cell cycle-regulated gene expression in Arabidopsis. J. Biol. Chem. 277:41987-42002.

Minoda A, Sakagami R, Yagisawa F, Kuroiwa T, Tanaka K. 2004. Improvement of culture conditions and evidence for nuclear transformation by homologous recombination in a red alga, Cyanidioschyzon merolae 10D. Plant Cell Physiol. 45:667-671.

Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE *et al*. (30 co-authors). 2002. Annotation of the Drosophila melanogaster euchromatic genome: A systematic review. Genome Biol. 3:RESEARCH0083.

Misumi O, Matsuzaki M, Nozaki H, Miyagishima S, Mori T, Nishida K, Yagisawa F, Yoshida Y, Kuroiwa H, Kuroiwa T. 2005. Cyanidioschyzon merolae genome. A tool for facilitating comparable studies on organelle biogenesis in photosynthetic eukaryotes. Plant Physiol. 137:567-585.

Mitrovich QM, Tuch BB, Guthrie C, Johnson AD. 2007. Computational and experimental approaches double the number of known intron in the pathogenic yeast Candida albicans. Genome Res. 17:492-502.

Moore CE, Archibald JM. 2009. Nucleomorph genomes. Annu. Rev. Genet. 43:251-264.

Moore MJ, Query CC, Sharp PA. 1993. Splicing of precursors to mRNA by the spliceosome. In: Gesteland RF, Atkins JF, editors. RNA World. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. p. 303-357.

Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science 323:379-382.

Morrissy AS, Griffith M, Marra MA. 2011. Extensive relationship between antisense transcription and alternative splicing in the human genome. Genome Res. 21:1203-1212.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods 5:621-628.

Muhia DK, Swales CA, Eckstein-Ludwig U, Saran S, Polley SD, Kelly JM, Schaap P, Krishna S, Baker DA. 2003. Multiple splice variants encode a novel adenylyl cyclase of possible plastid origin expressed in the sexual stage of the malaria parasite Plasmodium falciparum. J. Biol. Chem. 278:22014-22022.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344-1349.

Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, Fluhr R. 2004. Intron retention is a major phenomenon in alternative splicing in Arabidopsis. Plant J. 39:877-885.

Neugebauer KM. 2002. On the importance of being co-transcriptional. J. Cell. Sci. 115:3865-3871.

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

Nissim-Rafinia M, Kerem B. 2005. The splicing machinery is a genetic modifier of disease severity. Trends in Genetics 21:480-483.

Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, Maruyama S, Nishida K, Yagisawa F, Yoshida Y, Fujiwara T *et al*. (18 co-authors). 2007. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae. BMC Biology 5:28.

Ogawa, Y, Lee, JT. 2002. Antisense regulation in X inactivation and autosomal imprinting. Cytogenet. Genome Res. 99:59-65.

Ohnuma M, Yokoyama T, Inouye T, Sekine Y, Tanaka K. 2008. Polyethylene glycol (PEG)-mediated transient gene expression in a red alga, Cyanidioschyzon merolae 10D. Plant Cell Physiol. 49:117-120.

Ohta N, Sato N, Kuroiwa T. 1998. Structure and organization of the mitochondrial genome of the unicellular red alga Cyanidioschyzon merolae deduced from the complete nucleotide sequence. Nucleic Acids Res. 26:5190-5298.

Ohta N, Matsuzaki M, Misumi O, Miyagishima S, Nozaki H, Tanaka K, Shin-i T, Kohara Y, Kuroiwa T. 2003. Complete sequence and analysis of the plastid genome of the unicellular red alga Cyanidioschyzon merolae. DNA Res. 10:67-77.

Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohtomo Y, Murakami K *et al*. (13 co-authors). 2003. Antisense transcripts with rice full-length cDNAs. Genome Biol. 5:R5.

Palusa SG, Reddy AS. 2010. Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine (SR) genes with nonsense-mediated decay. New Phytol. 185:83-89.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genet. 40:1413-1415.

Patron NJ, Waller RF, Keeling PJ. 2006. A tertiary plastid uses genes from two endosymbionts. J. Mol. Biol. 357:1373-1382.

Pesole G, Liuni S, Grillo G, Saccone C. 1997. Structural and compositional features of untranslated regions of eukaryotic mRNAs. Gene 205:95-102.

Peters NT, Rohrbach JA, Zalewski BA, Byrkett CM, Vaughn JC. 2003. RNA editing and regulation of drosophila 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts. RNA 9:698-710.

Peyretaillade E, Goncalves O, Terrat S, Dugat-Bony E, Wincker P, Cornman RS, Evans JD, Delbac F, Peyret P. 2009. Identification of transcriptional signals in Encephalitozoon cuniculi widespread among microsporidia phylum: Support for accurate structural genome annotation. BMC Genomics 10:607-2164-10-607.

Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007. Rapid, transcript-specific changes in splicing in response to environmental stress. Mol. Cell 27:928-937.

Pombert JF, Xu J, Smith DR, Heiman D, Young S, Cuomo CA, Weiss LM, Keeling PJ. 2013. Complete genome sequences from three genetically distinct strains reveal a high intra-species genetic diversity in the microsporidian Encephalitozoon cuniculi. Eukaryotic Cell 12(4): 503-511.

Pombert JF, Selman M, Burki F, Bardell FT, Farinelli L, Solter LF, Whitman DW, Weiss LM, Corradi N, Keeling PJ. 2012. Gain and loss of multiple functionally related, horizontally transferred genes in the reduced genomes of two microsporidian parasites. PNAS 109:12638-12643.

Poole RL, Barker GL, Werner K, Biggi GF, Coghill J, Gibbings JG, Berry S, Dunwell JM, Edwards KJ. 2008. Analysis of wheat SAGE tags reveals evidence for widespread antisense transcription. BMC Genomics 9:475-2164-9-475.

Prescott EM, Proudfoot NJ. 2002. Transcriptional collision between convergent genes in budding yeast. Proc. Natl. Acad. Sci. USA 99:8796-8801.

Proudfoot NJ. 2003. Dawdling polymerases allow introns time to splice. Nat. Struct. Biol. 10:876-878.

Proudfoot NJ, Furger A, Dye MJ. 2002. Integrating mRNA processing with transcription. Cell 108:501-512.

van Rossum G, Drake FL (eds). 2001. Python Reference Manual, PythonLabs, Virginia, USA. Available at www.python.org.

R Development CT. 2011. R: A Language and Environment for Statistical Computing, Reference Index Version 2. 14. 1.

Reyes-Prieto A, Weber AP, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. Annu. Rev. Genet. 41:147-168.

Ringner M, Krogh M. 2005. Folding free energies of 5'-UTRs impact post-transcriptional regulation on a genomic scale in yeast. PLoS Comput Biol 1:e72.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrated genomics viewer. Nat. Biotechnol. 29:24-26.

Russell CB, Fraga D, Hinrichsen RD. 1994. Extremely short 20-33 nucleotide introns are the standard length in Paramecium tetraurelia. Nucleic Acids Res. 22:1221-1225.

Russo A, Siciliano G, Catillo M, Giangrande C, Amoresano A, Pucci P, Pietropaolo C, Russo G. 2010. hnRNP H1 and intronic G runs in the splicing control of the human rpL3 gene. Biochim. Biophys. Acta 1799:419-428.

Sabath N, Ferrada E, Barve A, Wagner A. 2013. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome Biol. Evol. 5:966-977.

Sasaki Y, Ishikawa J, Yamashita A, Oshima K, Kenri T, Furuya K, Yoshino C, Horino A, Shiba T, Sasaki T *et al*. (11 co-authors). 2002. The complete genomic sequence of Mycoplasma penetrans, an intracellular bacterial pathogen in humans. Nucleic Acids Res. 30:5293-5300.

Scheid P. 2007. Mechanism of intrusion of a microspordian-like organism into the nucleus of host amoebae (vannella sp.) isolated from a keratitis patient. Parasitol. Res. 101:1097-1102.

Schellenberg MJ, Ritchie DB, MacMillan AM. 2008. Pre-mRNA splicing: A complex picture in higher definition. Trends Biochem. Sci. 33:243-246.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. 2000. Drosophila dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101:671-684.

Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 18:88-103.

Sebé-Pedrós A, Irimia M, del Campo J, Parra-Acero H, Russ C, Nusbaum C, Blencowe BJ, Ruiz-Trillo I. 2013. Regulated aggregative multicellularity in a close unicellular relative of metazoa. eLIFE 2: e01287.

Siegel TN, Hon CC, Zhang Q, Lopez-Rubio JJ, Scheidig-Benatar C, Martins RM, Sismeiro O, Coppee JY, Scherf A. 2014. Strand-specific RNA-seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. BMC Genomics 15:150-2164-15-150.

Singh U, Rogers JB, Mann BJ, Petri WA,Jr. 1997. Transcription initiation is controlled by three core promoter elements in the hgl5 gene of the protozoan parasite Entamoeba histolytica. Proc. Natl. Acad. Sci. U. S. A. 94:8812-8817.

Skelly DA, Ronald J, Connelly CF, Akey JM. 2009. Population genomics of intron splicing in 38 Saccharomyces cerevisiae genome sequences. Genome Biol Evol 1:466-478.

Slamovits CH, Williams BAP, Keeling PJ. 2004. Transfer of Nosema locustae to Antonospora locustae n comb based on molecular and ultrastructural data. J. Eukaryot. Microbiol. 51:307-213.

Slamovits CH, Fast NM, Law JS, Keeling PJ. 2004. Genome compaction and stability in microsporidian intracellular parasites. Curr. Biol. 14:891-896.

Sorber K, Dimon MT, DeRisi JL. 2011. RNA-seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. Nucleic Acids Res. 39:3820-3825.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 9:3273-3297.

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

Staley JP, Guthrie C. 1998. Mechanical devices of the spliceosome: Motors, clocks, springs, and things. Cell 92:315-326.

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. Gene 344:1-20.

Sun M, Hurst LD, Carmichael GG, Chen J. 2006. Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcription and organismic complexity. Genome Res. 16:922-933.

Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29:288-299.

Tanifuji G, Onodera NT, Moore CE, Archibald JM. 2014. Reduced nuclear genomes maintain high gene transcription levels. Mol. Biol. Evol. 31:625-635.

Tarn WY, Steitz JA. 1996a. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell 84:801-811.

Tarn WY, Steitz JA. 1996b. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science 273:1824-1832.

Taupin V, Metenier G, Delbac F, Vivares CP, Prensier G. 2006. Expression of two cell wall proteins during the intracellular development of Encephalitozoon cuniculi: An immunocytochemical and in situ hybridization study with ultrathin frozen sections. Parasitology 132:815-825.

Terui S, Suzuki K, Takahiashi H, Itoh R, Kuroiwa T. 1995. High synchronization of chloroplast division in the ultramicro-alga Cyanidioschyzon merolae by treatment with both light and aphidicolin. J. Phycol. 31:958-961.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5:123-135.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-seq. Bioinformatics 25:1105-1111.

Tsaousis AD, Kunji ER, Goldberg AV, Lucocq JM, Hirt RP, Embley TM. 2008. A novel route for ATP acquisition by the remnant mitochondria of Encephalitozoon cuniculi. Nature 453:553-556.

Tuller T, Ruppin E, Kupiec M. 2009. Properties of untranslated regions of the S. cerevisiae genome. BMC Genomics 10:391.

van den Heuvel JJ, Bergkamp RJ, Planta RJ, Raue HA. 1989. Effect of deletions in the 5'-noncoding region on the translational efficiency of phosphoglycerate kinase mRNA in yeast. Gene 79:83-95.

Van Eden ME, Byrd MP, Sherrill KW, Lloyd RE. 2004. Demonstrating internal ribosome entry sites in eukaryotic mRNAs using stringent RNA test procedures. RNA 10:720-730.

van Noort V, Bradatsch B, Arumugam M, Amlacher S, Bange G, Creevey C, Falk S, Mende DR, Sinning I, Hurt E *et al*. (11 co-authors). 2013. Consistent mutational paths predict eukaryotic thermostability. BMC Evol. Biol. 13:7-2148-13-7.

Vaucheret H. 2006. Post-transcriptional small RNA pathways in plants: Mechanisms and regulations. Genes Dev. 20:759-771.

Vavra, J. & Larsson, J.I.R. 1999. Structure of the microsporidia. In: Wittner, M. (ed.), the microsporidia and microsporidiosis. The Microsporidia and Microsporidiosis .

Venkatesh B, Gilligan P, Brenner S. 2000. Fugu: A compact vertebrate reference genome. FEBS Lett. 476:3-7.

Vilardell J, Warner JR. 1997. Ribosomal protein L32 of Saccharomyces cerevisiae influences both the splicing of its own transcript and the processing of rRNA. Mol. Cell. Biol. 17:1959-1965.

Vinogradov AE. 1999. Intron-genome size relationship on a large evolutionary scale. J. Mol. Evol. 49:376-384.

Voinnet O. 2009. Origin, biogenesis, and activity of plant microRNAs. Cell 136:669-687.

Wahl MC, Will CL, Luhrmann R. 2009. The spliceosome; design principles of a dynamic RNA machine. Cell 136:701-718.

Wang BB, Brendel V. 2004. The ASRG database: Identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. Genome Biol. 5:R102.

Wang B, Brendel V. 2006. Genomewide comparative analysis of alternative splicing in plants. Proceedings of the National Academy of Sciences 103:7175-7180.

Wang ET. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456:470-476.

Wang G, Cooper TA. 2007. Splicing in disease: Disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. 8:749-761.

Wang XJ, Gaasterland T, Chua NH. 2005. Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. Genome Biol. 6:R30.

Wang Z, Gerstein M, Snyder M. 2009. RNA-seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet. 10:57-63.

Warner JR, Mitra G, Schwindinger WF, Studeny M, Fried HM. 1985. Saccharomyces-cerevisiae coordinates accumulation of yeast ribosomal-proteins by modulating messenger-rna splicing, translational initiation, and protein-turnover. Mol. Cell. Biol. 5:1512-1521.

Wilhelm BT, Marguerat S, Goodhead I, Bahler J. 2010. Defining transcribed regions using RNA-seq. Nat Protoc 5:255-266.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453:1239-1243.

Williams BA, Slamovits CH, Patron NJ, Fast NM, Keeling PJ. 2005. A high frequency of overlapping gene expression in compacted eukaryotic genomes. Proc. Natl. Acad. Sci. USA 102:10936-10941.

Wittner M, Weiss LM. 1999. The microsporidia and microsporidosis. Washington, D. C.: ASM Press.

Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. Nature 337:283-285.

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S *et al*. (134 co-authors). 2002. The genome sequence of Schizosaccharomyces pombe. Nature 415:871-880.

Wu Q, Krainer AR. 1997. Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. RNA 3:586-601.

Xia X, Holcik M. 2009. Strong eukaryotic IRESs have weak secondary structure. PLoS ONE 4:e4136.

Yang JS, Maurin T, Lai EC. 2012. Functional parameters of dicer-independent microRNA biogenesis. RNA 18:945-957.

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R *et al*. (16 co-authors). 2003. Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. 21:379-386.

Yoon HS, Hackett JD, Pinto G, Bhattacharya D. 2002. The single, ancient origin of chromist plastids. Proc. Natl. Acad. Sci. U. S. A. 99:15507-15512.

Zauner S. 2000. Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph. Proc. Natl. Acad. Sci. USA 97:200-205.

Zhang J, Sun X, Qian Y, Maquat LE. 1998. Intron function in the nonsense-mediated decay of beta-globin mRNA: Indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. RNA 4:801-815.

Zhang Z, Dietrich FS. 2005. Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. Nucleic Acids Res. 33:2838-2851.

Appendices

Appendix A Supplementary tables and figures of chapter 2

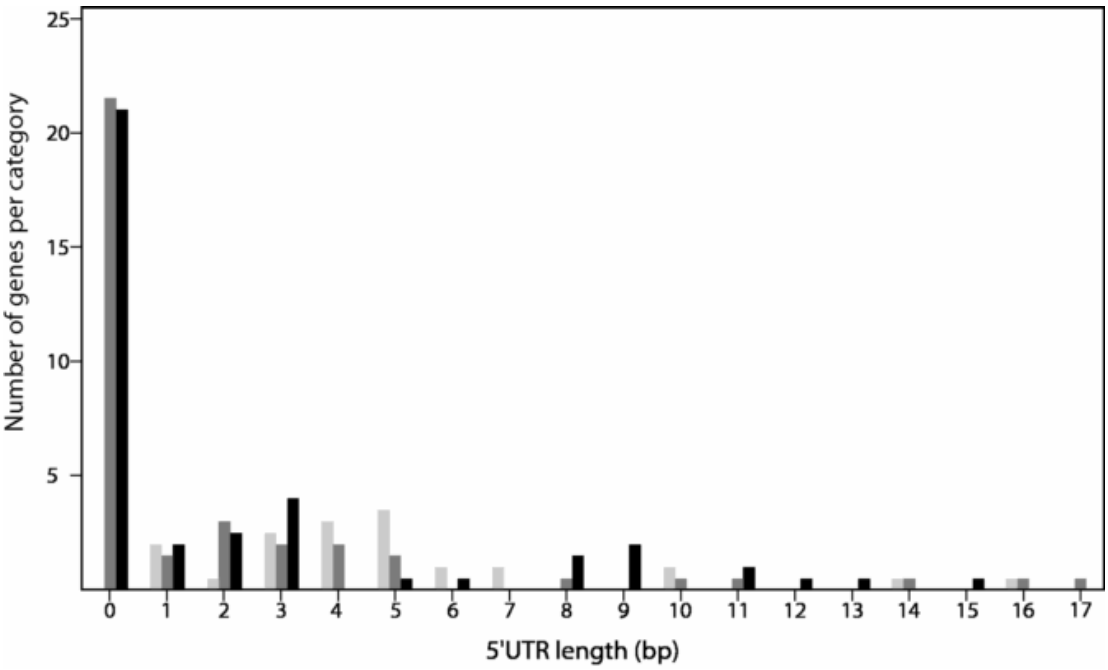


Figure A.1: The number of *Encephalitozoon cuniculi* genes examined at each 5' untranslated region (UTR) length between 0 and 20 bp

Light gray bars are intron-containing genes, dark gray bars are intron-lacking ribosomal protein genes, and black bars are intron-lacking non-ribosomal protein genes.

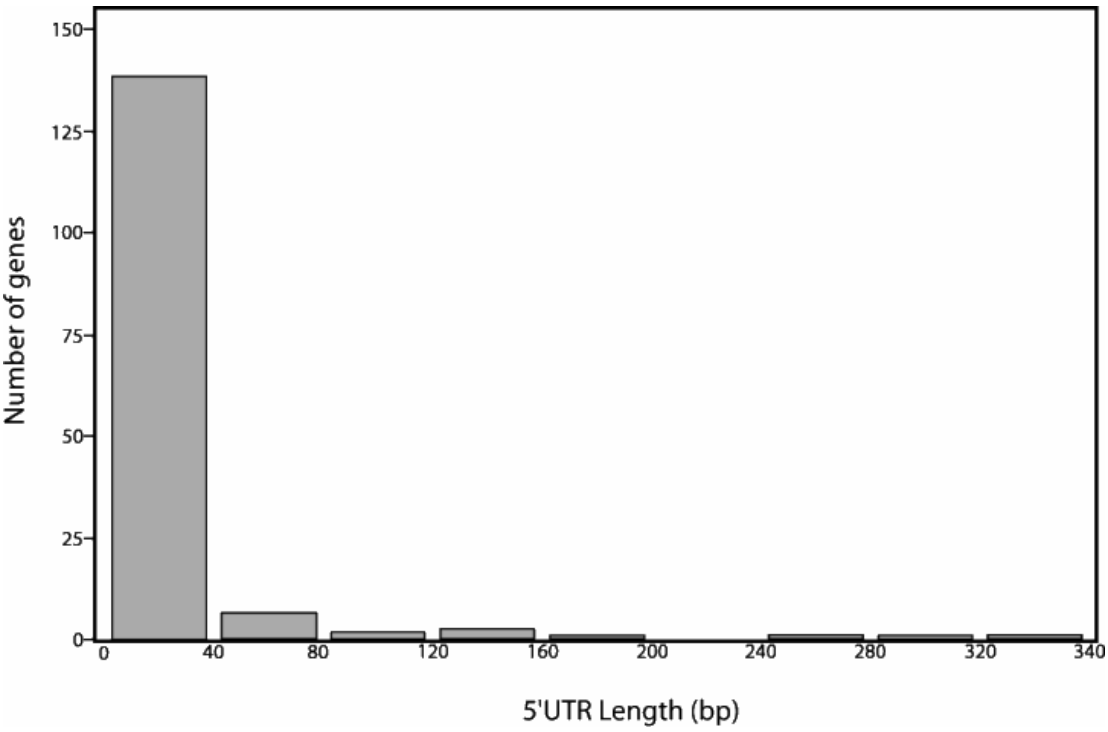


Figure A.2: A histogram of lengths of 5' untranslated regions (UTRs) for 155 genes of *Encephalitozoon cuniculi*

Appendix B Supplementary tables and figures of chapter 3

Table B.1: Gene expression levels in FPKM are shown for all 1985 *Encephalitozoon cuniculi* genes at three post-infection time-points

Gene	T1	T2	T3
ECU01_0220	27.71	24.71	29.15
ECU01_0230	502.24	651.38	692.9
ECU01_0240	109.01	196.57	231.19
ECU01_0250	623.29	429.48	362.07
ECU01_0280	151.33	138.27	149.6
ECU01_0290	81.79	80.36	84.56
ECU01_0310	1975.88	1093.74	1032.7
ECU01_0320	59.23	45.98	45.2
ECU01_0330	300.3	258.53	269.87
ECU01_0340	143.41	141.53	140.28
ECU01_0350	121.07	102.43	95.24
ECU01_0360	409.15	294.9	298.27
ECU01_0370	327.82	315.5	302.31
ECU01_0380	107.22	113.36	115.86
ECU01_0390	37.85	55.03	71.79
ECU01_0400	265.76	539.75	545.63
ECU01_0405	2	4.77	11.96
ECU01_0410	106.4	96.2	104.86
ECU01_0420	675.94	946.02	925.52
ECU01_0430	76.57	60.96	66.72
ECU01_0440	2274.85	1642.97	1569.77
ECU01_0450	42.24	44.06	41.14
ECU01_0460	6264.84	4569.79	4478.74
ECU01_0470	237.68	604.09	652.88
ECU01_0480	116.95	111.27	108.66
ECU01_0490	252.84	372.93	394.99
ECU01_0500	123.58	112.12	99.21
ECU01_0510	154.65	219.83	223.51
ECU01_0520	30.95	80.29	99.52
ECU01_0525	221.29	207.29	214.15
ECU01_0530	62.26	76.36	83.16
ECU01_0540	310.03	232.45	228.01
ECU01_0545	59.01	63.4	62.1
ECU01_0550	228.95	189.1	198.61
ECU01_0555	4298.59	6167.52	6037.58
ECU01_0560	149.02	146.75	137.12
ECU01_0570	67.7	64.86	65.89
ECU01_0580	176.43	148.55	152.42
ECU01_0590	625.53	732.42	662.26
ECU01_0600	204.77	162.89	161.29
ECU01_0610	56.29	43.47	39.27

Gene	T1	T2	T3
ECU01_0620	148.7	179.13	174.06
ECU01_0630	33.57	34	38.25
ECU01_0640	73.27	114.75	112.53
ECU01_0650	168.1	313.37	323.72
ECU01_0670	91.97	91.35	82.95
ECU01_0680	487.87	535.47	518.29
ECU01_0690	125.18	256.88	261.88
ECU01_0700	161.6	163.03	165.52
ECU01_0710	145.63	107.89	88.53
ECU01_0720	280.14	212.13	191.03
ECU01_0730	160.7	170.21	184
ECU01_0750	103.59	100.43	104.8
ECU01_0760	191.91	169.5	176.72
ECU01_0770	44.37	43.11	49.14
ECU01_0775	6.85	6.14	2.56
ECU01_0780	34.67	36.01	35.22
ECU01_0790	645.3	557.13	504.47
ECU01_0800	85.26	273.76	340.6
ECU01_0810	371.83	231.53	263.22
ECU01_0820	9119.49	24264.8	23713.85
ECU01_0830	203.39	451.8	489.38
ECU01_0840	187.52	178.4	194.35
ECU01_0850	109.25	105.18	110
ECU01_0860	36.07	40.42	40.35
ECU01_0870	146.4	265.24	296.84
ECU01_0880	106.88	71.11	69.54
ECU01_0890	219.5	150.51	139.3
ECU01_0900	41.78	85.62	82.04
ECU01_0910	142.75	138.05	160.64
ECU01_0920	2018.08	1033.54	926.33
ECU01_0940	65.82	48.22	58.1
ECU01_0950	247.1	195.55	184.54
ECU01_0955	1041.26	1013.37	921.65
ECU01_0960	315.13	232.16	222.43
ECU01_0970	699.96	910.43	918.39
ECU01_0975	59.53	57.53	52.29
ECU01_0980	144.63	131.94	128.96
ECU01_0990	60.09	58.57	50.81
ECU01_1000	151.24	184.29	172.18
ECU01_1010	749.18	711.9	714.41
ECU01_1020	82.94	98.97	99.41
ECU01_1030	129	143.35	146.7
ECU01_1040	65.6	63.54	57.16
ECU01_1050	60.74	69.35	67.91
ECU01_1060	15.72	16.29	18.03
ECU01_1070	77.96	111.2	132.99

Gene	T1	T2	T3
ECU01_1080	178.94	172.75	159.01
ECU01_1090	73.88	77.88	77.58
ECU01_1095	96.73	124.44	120.03
ECU01_1100	336.5	333.33	329.47
ECU01_1110	228.92	189.44	188.13
ECU01_1115	193.17	138.51	115.24
ECU01_1120	42.21	40.63	39.12
ECU01_1130	151.81	289.17	289.69
ECU01_1140	89.89	79.54	80.38
ECU01_1150	64.5	67.52	65.01
ECU01_1160	156.75	126.14	112.44
ECU01_1170	56.26	64.72	64.15
ECU01_1180	24.61	19.99	20.38
ECU01_1190	55.31	46.22	45.06
ECU01_1200	125.1	125.17	119.2
ECU01_1210	72.46	64.82	60.21
ECU01_1220	39.94	24.78	23.48
ECU01_1230	539.79	726.64	742.65
ECU01_1240	80.25	87.76	80.33
ECU01_1250	124.28	183.03	182.99
ECU01_1260	110.86	231.2	261.86
ECU01_1270	5147.88	14157.15	14514.58
ECU01_1280	833.07	650.19	648.81
ECU01_1290	282.01	252.21	264.68
ECU01_1300	47.68	45.81	44.93
ECU01_1310	124.58	104.8	94.11
ECU01_1320	143.99	182.31	170.89
ECU01_1330	226.5	186	178.95
ECU01_1340	15.05	23.28	26.45
ECU01_1350	122.26	121.37	131.55
ECU01_1360	331.91	261.84	260.08
ECU01_1370	277.68	590.54	639.82
ECU01_1375	57.14	89.91	91.95
ECU01_1380	46.7	72.01	74.11
ECU01_1390	51.03	86.89	104.52
ECU01_1400	135.54	81.17	78.37
ECU01_1410	123.76	100.5	96.08
ECU01_1420	403.75	271.64	278.67
ECU02_0090	19.27	16.96	15.66
ECU02_0100	1876.46	1432.78	1611.45
ECU02_0110	53.25	64.83	75.36
ECU02_0120	134.42	98.08	107.18
ECU02_0130	32.09	62.39	68.64
ECU02_0140	786.18	505.31	508.39
ECU02_0150	240.76	756.65	833.03
ECU02_0160	38.9	65.48	79.94

Gene	T1	T2	T3
ECU02_0170	129.76	573.37	538.41
ECU02_0180	162.33	480.84	511.11
ECU02_0190	374.15	1560.69	1532.04
ECU02_0200	87.99	76.17	76.69
ECU02_0210	133.65	102.5	97.79
ECU02_0215	206.51	204.06	221.22
ECU02_0220	167.75	120.72	130.21
ECU02_0230	40.47	47.83	46.97
ECU02_0240	251.38	207.55	200.62
ECU02_0250	110.23	242.09	257.36
ECU02_0260	253.57	598.01	661.94
ECU02_0270	74.54	178.44	239.92
ECU02_0280	115.77	430.25	485.09
ECU02_0290	132.57	139.86	140.06
ECU02_0300	127.23	94.35	99.99
ECU02_0310	126.11	85.67	85.12
ECU02_0320	64.87	39.27	38.28
ECU02_0330	255.72	240.21	226.6
ECU02_0340	378.35	328.56	327.75
ECU02_0350	54.5	43.18	40.84
ECU02_0355	4288.63	2324.98	2315.38
ECU02_0360	230.74	247.45	219.11
ECU02_0370	47.91	37.64	38.42
ECU02_0380	139.35	139.71	133.9
ECU02_0390	224.33	160.29	173.84
ECU02_0400	59.29	52.07	46.73
ECU02_0410	450.61	281.27	253.62
ECU02_0420	84.93	77.81	77.14
ECU02_0430	82.46	90.42	94.17
ECU02_0440	417.24	353.09	319.62
ECU02_0450	272.93	246.6	260.7
ECU02_0460	225.13	171.32	163.25
ECU02_0470	131.2	123.97	111.93
ECU02_0480	258.63	202.21	222.89
ECU02_0490	144.78	218.36	192.18
ECU02_0500	43.01	94.2	89.64
ECU02_0510	63.02	54.37	50.39
ECU02_0520	552.78	391.76	386.18
ECU02_0525	5.27	4.89	3.94
ECU02_0530	19.65	27.83	31.31
ECU02_0535	97.31	236.44	191.1
ECU02_0540	91.76	240.9	237.31
ECU02_0550	31.64	35.57	31.03
ECU02_0560	88.28	92.97	86.71
ECU02_0565	68.02	73.25	72.77
ECU02_0570	63.88	69.42	71.33

Gene	T1	T2	T3
ECU02_0580	690.69	1046.1	1080.99
ECU02_0590	43.05	36.41	34.51
ECU02_0600	99	115.09	106.92
ECU02_0610	1945.85	1460.98	1289.3
ECU02_0615	105.01	80.68	66.29
ECU02_0630	173.76	144.57	147.47
ECU02_0640	120.01	102.59	98.41
ECU02_0650	222.75	246	251.85
ECU02_0660	102.26	116.77	122.35
ECU02_0670	204.52	227.13	238.94
ECU02_0680	151.62	95.42	81.14
ECU02_0690	69.13	65.08	56.91
ECU02_0700	147.89	92.43	92.7
ECU02_0705	146.47	177.72	184.96
ECU02_0710	98.59	100.44	101.34
ECU02_0720	3797.29	3012.71	3165.36
ECU02_0730	100.46	77.22	84.84
ECU02_0760	556.24	661.48	556.15
ECU02_0770	2609.62	1766.46	1469.58
ECU02_0780	183.63	162.46	150.79
ECU02_0785	236.54	139.9	121.87
ECU02_0800	2985.95	1849.68	1698.34
ECU02_0810	393.05	261.61	274.73
ECU02_0820	49.16	42.73	42.65
ECU02_0830	80.31	72.9	64.3
ECU02_0840	404.06	318.83	308.32
ECU02_0850	206.21	622.93	754.57
ECU02_0860	84.22	220.14	214.18
ECU02_0865	75.31	34.81	32.71
ECU02_0870	78.79	70.32	65.63
ECU02_0875	0.48	1.25	0.91
ECU02_0890	77.67	93.22	86.85
ECU02_0900	1536.97	892.23	849.49
ECU02_0910	99.44	74.65	68.53
ECU02_0920	230.46	173.24	179.47
ECU02_0930	324.72	210.62	220.16
ECU02_0940	268.98	271.93	241.53
ECU02_0950	58.63	84.78	90.24
ECU02_0960	634.93	768.09	824.67
ECU02_0970	55.64	56.37	63.14
ECU02_0980	73.24	74.24	64.42
ECU02_0990	116.69	447.03	492.45
ECU02_1000	73.34	86.74	87.13
ECU02_1010	260.51	288.98	304.47
ECU02_1020	412.02	361.57	310.44
ECU02_1030	105.79	119.94	125.35

Gene	T1	T2	T3
ECU02_1040	19.8	11.44	7.93
ECU02_1045	2.03	4.84	4.55
ECU02_1050	356.95	234.73	235.36
ECU02_1060	33.99	30.47	28.02
ECU02_1070	99.23	70.98	73.25
ECU02_1080	1553.63	1054.49	1060.4
ECU02_1090	63.56	44.3	43.03
ECU02_1100	2487.89	1231.82	1401.58
ECU02_1110	106.43	68.01	71.49
ECU02_1120	38.25	35.07	39.58
ECU02_1130	48.64	49.71	44.6
ECU02_1140	670.69	614.74	580.04
ECU02_1150	142.17	110.03	106.62
ECU02_1160	125.12	80.7	74.6
ECU02_1170	334.92	217.7	223.65
ECU02_1180	96.08	70.66	75.56
ECU02_1190	263.07	288.83	298.85
ECU02_1200	190.26	145.38	162.42
ECU02_1210	215.6	161.99	146.53
ECU02_1220	56.6	48.07	41.49
ECU02_1230	147.3	125.79	132.75
ECU02_1240	96.37	86.57	82.29
ECU02_1250	61.02	44.73	42.4
ECU02_1260	34.87	27.06	28.58
ECU02_1270	49.68	39.69	48.11
ECU02_1280	73.19	64.95	64.08
ECU02_1290	78.39	69.64	71.06
ECU02_1300	82.37	131.83	155.61
ECU02_1310	169.89	167.89	165.09
ECU02_1320	25.42	39.06	43.8
ECU02_1330	258.22	193.14	183.87
ECU02_1340	152.8	107.79	104.34
ECU02_1350	69.53	54.19	52.66
ECU02_1360	246.34	226.89	225.16
ECU02_1370	320.1	1042.51	986.81
ECU02_1380	319.56	420.8	384.44
ECU02_1390	72.21	64.22	67.6
ECU02_1400	132.91	117.57	110.23
ECU02_1410	63.38	101.93	116.61
ECU02_1420	80.68	73.58	92.19
ECU02_1430	215.1	211.91	208.21
ECU02_1440	120.69	96.08	106.89
ECU02_1450	60.09	46.04	43.33
ECU02_1460	190.77	175.3	196.11
ECU02_1470	242.03	428.06	464.57
ECU02_1480	96.61	86.77	87.2

Gene	T1	T2	T3
ECU02_1485	6.14	9.08	15.76
ECU02_1490	226.24	167.97	167.56
ECU02_1495	341.31	241.23	234.89
ECU02_1500	227.66	167.2	154.84
ECU02_1510	21.69	18	20.2
ECU02_1520	363.72	323.36	305.69
ECU02_1530	81.47	55.56	61.79
ECU02_1540	7.16	6.65	7.4
ECU03_0090	11.8	15.89	16.13
ECU03_0100	23.35	10.21	8.53
ECU03_0110	1.16	2.64	4.42
ECU03_0120	117.66	89.82	121.09
ECU03_0130	31.45	43.63	39.93
ECU03_0150	30.05	46.41	43.98
ECU03_0160	143.18	123	126.01
ECU03_0170	24.55	46.03	55.38
ECU03_0180	72.88	181.37	220.62
ECU03_0190	57.67	44.2	51.69
ECU03_0200	32.13	23.42	27.64
ECU03_0210	144.81	130.41	149
ECU03_0220	659.63	408.57	405.85
ECU03_0230	2407.02	1476.15	1329.04
ECU03_0240	127.11	122.28	118.28
ECU03_0250	77.28	105.99	109.01
ECU03_0255	515.97	400.05	452.77
ECU03_0260	66.64	83.04	84.55
ECU03_0270	202.91	223.04	233.34
ECU03_0280	44.62	118.42	123.41
ECU03_0290	87	76.95	76.79
ECU03_0300	48.6	45.44	43.64
ECU03_0305	40.23	47.94	49.73
ECU03_0310	1776.16	1350.31	1289.97
ECU03_0320	1446.64	779.05	757.93
ECU03_0325	6.94	7.59	11.68
ECU03_0330	37.88	31.64	27.68
ECU03_0340	111.75	105.23	100.35
ECU03_0350	19.49	15.73	16.93
ECU03_0355	4.93	7.06	5.9
ECU03_0360	195.68	95.06	98.41
ECU03_0370	40.72	35.34	32.32
ECU03_0380	542.2	304.48	306.32
ECU03_0390	181.32	117.69	129.43
ECU03_0400	51.11	32.11	33.92
ECU03_0410	73.12	55.12	55.8
ECU03_0420	160.6	140.69	151.14
ECU03_0430	68.93	55.02	55.43

Gene	T1	T2	T3
ECU03_0440	174.45	113.02	113.56
ECU03_0450	227.42	163.64	157.65
ECU03_0460	55.46	42.24	46.39
ECU03_0470	37.86	34.42	37.76
ECU03_0480	65.65	49.7	54.23
ECU03_0490	143.7	101.31	89.22
ECU03_0500	94.87	133.9	131.81
ECU03_0510	134.16	215.33	200.01
ECU03_0520	3020.2	1618.09	1991.12
ECU03_0530	86.47	104.12	95.92
ECU03_0540	76.9	51.9	52.11
ECU03_0550	71.31	92.47	85.63
ECU03_0560	14.81	23.23	21.94
ECU03_0570	12.17	22.61	17.68
ECU03_0580	175.7	148.24	143.92
ECU03_0590	236.63	179.65	187.23
ECU03_0600	67.59	61.54	54.41
ECU03_0610	58.86	61.18	53.51
ECU03_0620	46.47	48.27	50.77
ECU03_0630	205.65	197.91	218.21
ECU03_0640	54.77	37.11	39.62
ECU03_0650	1013.22	481.83	447.21
ECU03_0660	166.35	173.2	167.51
ECU03_0670	55.57	42.73	34.85
ECU03_0680	51.71	78.83	71.91
ECU03_0690	177.44	163.2	168.2
ECU03_0700	92.8	69.79	63.42
ECU03_0710	2119.87	1391.9	1303.5
ECU03_0715	95.55	69.86	67.49
ECU03_0730	1450.3	871.38	880.53
ECU03_0750	277.74	210.52	217.65
ECU03_0760	219.39	166.8	185.38
ECU03_0770	94.8	142.4	179.47
ECU03_0780	97.39	94.96	86.64
ECU03_0790	135.89	107.69	105.62
ECU03_0800	180.53	127.78	128.03
ECU03_0810	76.84	61.92	78.24
ECU03_0830	108.01	251.47	270.15
ECU03_0840	367.19	271.06	228.01
ECU03_0850	295.44	197.33	192.15
ECU03_0860	60.68	33.06	26.54
ECU03_0870	77.47	84.82	75.67
ECU03_0880	92.22	111.19	116.06
ECU03_0890	33.26	49.96	42.55
ECU03_0900	179.69	187.32	186.69
ECU03_0910	71.02	72.95	68.73

Gene	T1	T2	T3
ECU03_0920	25.53	28.21	27.66
ECU03_0930	315.46	398.11	401.42
ECU03_0940	254.28	252.97	250.71
ECU03_0950	1903.61	1102.1	1047.81
ECU03_0960	180.12	137	114.31
ECU03_0970	73.21	65.87	64.01
ECU03_0980	41.17	45.24	48.81
ECU03_0990	84.38	165.68	164.86
ECU03_1000	39.86	56.02	58.25
ECU03_1010	567.95	390.36	388.6
ECU03_1020	131.9	100.78	87.35
ECU03_1030	97.47	74.69	74.47
ECU03_1040	480.07	423.35	378.11
ECU03_1050	245.21	201.85	207.41
ECU03_1060	842.66	500.23	472.92
ECU03_1070	217.72	156.05	147.21
ECU03_1080	243.63	205.6	190.47
ECU03_1090	100.26	94.01	94.03
ECU03_1095	3	2.39	4.49
ECU03_1100	56.8	90.42	93.21
ECU03_1110	31.58	24.25	23.89
ECU03_1120	160.43	207.03	184.27
ECU03_1130	74.68	108.38	106.97
ECU03_1140	131.74	191.79	167.91
ECU03_1150	1465.95	1026.42	969.97
ECU03_1160	274.74	288.49	278.51
ECU03_1170	249.15	265.11	297.7
ECU03_1175	37.88	96.49	119.02
ECU03_1180	36.39	56.3	52.37
ECU03_1190	1082.46	684.3	678.96
ECU03_1200	69.22	51.02	59.74
ECU03_1220	947.74	578.35	547.38
ECU03_1230	90.88	68.49	58.67
ECU03_1240	36.51	31.29	28.78
ECU03_1250	164.52	147.57	139.63
ECU03_1260	958.9	955.72	895.6
ECU03_1270	180.73	178.27	164.85
ECU03_1280	75.67	68.56	70.44
ECU03_1290	141.15	341.79	358.66
ECU03_1300	89.89	73.05	78.49
ECU03_1305	5.92	4.71	2.95
ECU03_1310	32.27	27.95	23.33
ECU03_1320	59.45	45.4	51.61
ECU03_1330	350.78	381.45	397.51
ECU03_1340	36.39	31.93	31.97
ECU03_1350	177.56	196.23	196.41

Gene	T1	T2	T3
ECU03_1360	81.22	67.34	77.3
ECU03_1370	215.61	203.43	204.26
ECU03_1380	50.67	42.75	47.14
ECU03_1390	29.02	33.98	28.9
ECU03_1395	264.37	369.09	381.28
ECU03_1400	58.26	43.27	41.52
ECU03_1410	143.86	175.51	189.3
ECU03_1420	58.42	93.2	94.65
ECU03_1430	470.04	441.33	421.49
ECU03_1440	107	101.65	103.59
ECU03_1450	92.39	75.13	72.14
ECU03_1460	3364.22	4001.57	3647.08
ECU03_1470	42.5	30.53	34.12
ECU03_1480	80.99	52.21	59.6
ECU03_1490	2838.08	1714.71	1600.23
ECU03_1500	61.37	83.39	87.64
ECU03_1505	209.41	716.81	869.28
ECU03_1510	54.12	162.73	191.04
ECU03_1520	37.8	64.47	69.36
ECU03_1530	74.16	71.77	73.96
ECU03_1535	259.18	500.07	508.37
ECU03_1540	204.48	172.74	184.88
ECU03_1550	307.43	226.27	225.29
ECU03_1560	224.3	215.59	233.91
ECU03_1570	997.78	845.43	778.68
ECU03_1580	285.61	235.54	226.99
ECU03_1590	48.99	51.96	53.09
ECU03_1600	103.85	107.42	88.55
ECU03_1610	184.63	324.53	362.65
ECU03_1620	10.44	10.93	11.37
ECU04_0120	186.4	133.92	137.16
ECU04_0130	299.75	446.87	497.64
ECU04_0140	1536.8	977.8	1067.19
ECU04_0150	71.19	61.91	56.16
ECU04_0155	64.23	62.63	59.44
ECU04_0160	132.25	116.79	116.14
ECU04_0165	47.65	68.91	62.57
ECU04_0170	829.08	1434.67	1421.01
ECU04_0180	298.37	223.18	253.07
ECU04_0190	48.1	40.07	42.4
ECU04_0195	2.92	3.1	2.91
ECU04_0200	10.69	14.18	14.05
ECU04_0210	87.64	77.39	73.85
ECU04_0230	17.11	9.47	10.72
ECU04_0240	28.81	23.4	22.68
ECU04_0250	71.22	62.03	65.01

Gene	T1	T2	T3
ECU04_0260	76.48	109.07	105.82
ECU04_0270	396.18	354.76	337.73
ECU04_0280	39.47	38.31	35.68
ECU04_0290	15.8	16.51	12.88
ECU04_0300	203.42	212.71	205.15
ECU04_0310	279.25	238.84	245.22
ECU04_0320	84.74	93.92	81.47
ECU04_0330	1708.68	1059.85	952.18
ECU04_0340	115.95	106.54	99.59
ECU04_0350	160.25	146.84	139.88
ECU04_0360	123.99	98.38	82.01
ECU04_0370	50.36	37.86	35.18
ECU04_0380	15.99	27.07	26.85
ECU04_0400	113.58	126.75	133.61
ECU04_0410	46.04	47.01	45.12
ECU04_0415	159.62	145.19	136.21
ECU04_0420	340.75	392.43	375.99
ECU04_0430	144.41	118.43	122.63
ECU04_0440	35.26	30.91	31.24
ECU04_0450	1300.8	827.02	809.96
ECU04_0455	23.15	17.44	21.54
ECU04_0460	90.35	87.93	92.19
ECU04_0470	149.57	144.16	132.19
ECU04_0480	178.58	146.08	154.94
ECU04_0485	9.93	12.68	13.62
ECU04_0490	71.13	79.31	79.72
ECU04_0500	97.16	87.79	101.48
ECU04_0510	433.49	381.28	354.9
ECU04_0520	153.27	128.28	136.59
ECU04_0530	187.38	174.64	161.45
ECU04_0540	72.61	72.87	75.42
ECU04_0550	50.56	37.62	40.62
ECU04_0560	113.51	70.57	72.77
ECU04_0570	130.7	104.05	100.59
ECU04_0575	231.8	153.89	130.64
ECU04_0580	865.57	747.39	770.42
ECU04_0590	206.63	167.54	166.62
ECU04_0600	315.94	286.56	249.18
ECU04_0610	53.71	67.81	86.9
ECU04_0620	82.07	80.41	89.59
ECU04_0630	197.46	271.67	308.94
ECU04_0640	1619.47	813.05	857.11
ECU04_0650	149.02	145.01	129.89
ECU04_0660	37.12	51.87	54.57
ECU04_0670	38.14	39.02	34.7
ECU04_0680	60.52	155.58	154.69

Gene	T1	T2	T3
ECU04_0690	34.65	56.76	56.35
ECU04_0700	115.73	94.86	97.82
ECU04_0710	151.07	236.54	249.34
ECU04_0715	121.19	377.58	432.22
ECU04_0720	316.62	578.2	591.1
ECU04_0725	159.88	149.12	175.6
ECU04_0730	139.75	130.75	120.94
ECU04_0740	1287.96	709.32	658.16
ECU04_0750	368.98	285.67	289.82
ECU04_0760	65.88	65.96	58.4
ECU04_0770	54.92	53.52	54.41
ECU04_0780	427.14	350.15	378.67
ECU04_0790	233.74	237.79	249.07
ECU04_0800	962.2	868.14	816.12
ECU04_0810	90.38	54.4	48.89
ECU04_0820	251.01	172.66	159.7
ECU04_0830	46.01	33.28	31.79
ECU04_0840	19.29	17.93	18.29
ECU04_0850	135.98	104.68	96.7
ECU04_0860	254.96	195.73	181.61
ECU04_0870	35.21	31.07	31.39
ECU04_0880	28.89	29.2	28.56
ECU04_0890	128.47	117.62	108.2
ECU04_0900	189.63	160.35	154.38
ECU04_0910	68.22	73.88	65.83
ECU04_0915	1.76	6.29	7.88
ECU04_0920	527.44	438.49	421.13
ECU04_0925	36.75	22.78	21.08
ECU04_0930	113.43	82.98	75.06
ECU04_0940	374.82	423.9	394.19
ECU04_0950	283.48	256.9	207.43
ECU04_0960	135.38	122.91	109.81
ECU04_0970	24.97	79.05	73.55
ECU04_0980	298.91	343.55	340.49
ECU04_0990	160.35	136.95	145.47
ECU04_1000	353.99	394.37	355.86
ECU04_1010	237.8	155.69	142.8
ECU04_1015	804.18	415.93	401.51
ECU04_1020	857.03	584.05	587.01
ECU04_1030	245.62	171.45	167.42
ECU04_1040	151.55	132.52	132.42
ECU04_1050	268.18	262.49	245.99
ECU04_1060	115.06	149.81	164.09
ECU04_1070	51.76	38.8	36.29
ECU04_1080	152.85	321.96	321.46
ECU04_1090	342.72	315.95	287.26

Gene	T1	T2	T3
ECU04_1100	7208.68	4412.06	4410.56
ECU04_1110	23.83	27.26	24.09
ECU04_1120	45.98	30.12	38.03
ECU04_1130	157.75	184.47	186.03
ECU04_1140	258.26	194.98	201.66
ECU04_1150	621.06	328.45	333.47
ECU04_1160	33.02	66.65	77.42
ECU04_1170	219.3	125.41	145.67
ECU04_1180	133.37	82.33	90.39
ECU04_1190	913.41	655.04	704.38
ECU04_1200	169.78	125.69	125.78
ECU04_1210	77.27	105.55	112.41
ECU04_1220	88.21	55.8	66.55
ECU04_1230	81.97	66.87	56.71
ECU04_1240	347.04	311.64	302.58
ECU04_1250	79.7	69.05	69.09
ECU04_1260	87.07	76.49	78.59
ECU04_1270	15.11	12.03	13.6
ECU04_1280	191.66	145.74	145.49
ECU04_1290	90.5	94.83	94.11
ECU04_1300	198.03	150.88	146.67
ECU04_1310	1822.29	862.81	836.84
ECU04_1320	272.51	141.01	146.56
ECU04_1330	81.29	60.46	59.73
ECU04_1340	123.61	100.19	96.57
ECU04_1350	67.49	47.78	41.76
ECU04_1355	1287.23	744.02	741.93
ECU04_1357	117.8	88.65	92.15
ECU04_1360	69.05	50.65	51.81
ECU04_1370	75.78	51.23	56.48
ECU04_1380	2050.38	1365.34	1329.34
ECU04_1390	43.49	43.19	35.89
ECU04_1400	77.28	60.22	56.03
ECU04_1405	1.97	0.78	2.95
ECU04_1410	54.1	124.78	140.42
ECU04_1415	8	10.34	8.97
ECU04_1420	203.42	178.17	163.6
ECU04_1422	197.1	301.43	320.83
ECU04_1425	110.09	113.02	120.47
ECU04_1430	60.86	69.93	72.11
ECU04_1435	68.64	86.38	85.04
ECU04_1440	125.25	72.84	76.84
ECU04_1450	107.95	72.05	73.42
ECU04_1460	30.18	21.63	23.6
ECU04_1470	109.94	100.64	102.15
ECU04_1480	227.84	534.83	536.24

Gene	T1	T2	T3
ECU04_1490	127.21	131.22	142.17
ECU04_1500	85.83	64.67	70.19
ECU04_1510	237.62	178.62	176.21
ECU04_1520	50.2	26.03	31.7
ECU04_1530	97.7	79.57	70.92
ECU04_1540	209.15	231.96	211.06
ECU04_1550	125.89	81.81	85.21
ECU04_1560	1297.51	1007.21	933.91
ECU04_1570	408	245.41	258.04
ECU04_1580	77.33	50.82	45.38
ECU04_1590	101.25	144.85	163.82
ECU04_1600	57.78	62.4	61.93
ECU04_1605	337.07	214.97	195.12
ECU04_1610	214.37	184.01	187.34
ECU04_1620	202.69	316.66	311.99
ECU04_1625	304.84	252.97	268.93
ECU04_1630	80.46	51.33	68.05
ECU05_0060	55.8	48.62	45.97
ECU05_0070	27.88	23.04	22.19
ECU05_0080	51.53	66.6	63.85
ECU05_0085	57.23	157.28	167.42
ECU05_0087	50.31	176.92	153.55
ECU05_0090	517.92	568.8	599.69
ECU05_0100	55.77	45.88	48.48
ECU05_0110	128.69	262.61	325.77
ECU05_0120	66.99	51.72	52.94
ECU05_0140	964.6	1066.5	1041.26
ECU05_0150	4356.38	2857.27	2935.79
ECU05_0160	62.18	59.5	65
ECU05_0180	105.17	77.21	91.19
ECU05_0185	259.94	224.79	207.13
ECU05_0190	70.15	54.46	57.64
ECU05_0200	155.63	222.81	276.36
ECU05_0210	206.98	761.83	905.24
ECU05_0220	116.17	82.47	84.46
ECU05_0230	12.6	13.11	7.3
ECU05_0240	60.79	46.27	44.18
ECU05_0250	2784.1	1678.61	1463.88
ECU05_0260	209.83	366.28	425.42
ECU05_0270	235.31	240.93	243.06
ECU05_0275	131.81	112.42	108.79
ECU05_0280	285.72	222.34	241.92
ECU05_0290	360.78	379.6	362.74
ECU05_0300	23.65	23.03	24.18
ECU05_0310	260.28	204.34	188.37
ECU05_0320	49.83	91.81	92.55

Gene	T1	T2	T3
ECU05_0330	30.8	30.52	31.75
ECU05_0340	141.65	253.62	269.39
ECU05_0350	145.01	124.78	127.73
ECU05_0360	188.44	115.18	111.76
ECU05_0370	427.64	293.24	277.46
ECU05_0380	72.89	56.98	51.08
ECU05_0390	87.03	75.94	68.14
ECU05_0400	76.88	73.02	66.62
ECU05_0405	288.25	321.13	285.44
ECU05_0420	27.01	18.58	21.45
ECU05_0430	131.69	82.84	83.01
ECU05_0435	203.82	165.51	143.82
ECU05_0440	166.77	134.68	128.22
ECU05_0450	72.84	146.44	170.33
ECU05_0460	72.33	126.4	151.36
ECU05_0465	45.01	67.59	64.93
ECU05_0470	529.92	437.22	426.53
ECU05_0480	112.71	94.22	98.95
ECU05_0490	17.19	24.67	29.61
ECU05_0495	476.65	492.18	478.33
ECU05_0500	285.08	230.86	238.84
ECU05_0510	280.8	236.28	218.32
ECU05_0520	110.69	111.58	99.79
ECU05_0530	95	113.06	106.53
ECU05_0540	46.6	48.13	43.86
ECU05_0550	75.73	63.44	57.06
ECU05_0560	326.11	268.32	253.65
ECU05_0570	84.34	77.43	84.62
ECU05_0580	188.93	148.23	132.98
ECU05_0590	117.1	494.36	560.24
ECU05_0595	10.4	13.81	15.57
ECU05_0600	1524.38	1197.95	1155.14
ECU05_0610	70.78	88.11	84.77
ECU05_0620	180.99	162.55	146
ECU05_0630	231.43	187.35	190.83
ECU05_0640	241.81	187.67	163.2
ECU05_0650	420.56	807.14	884
ECU05_0660	75.9	62.95	59.3
ECU05_0670	1637.95	1109.6	994.66
ECU05_0680	81.29	66.21	62
ECU05_0690	120.37	108.16	104.26
ECU05_0700	73.52	60.82	64.49
ECU05_0710	180.08	298.93	322.55
ECU05_0720	65.78	75.23	74.81
ECU05_0730	35.23	61.99	67.72
ECU05_0740	130.1	156.25	167.47

Gene	T1	T2	T3
ECU05_0750	96.16	59.36	53.54
ECU05_0760	54.24	50.9	48.64
ECU05_0770	178.32	171.67	151.55
ECU05_0780	124.72	103.55	102.61
ECU05_0785	18.36	15.8	19.05
ECU05_0790	40.42	38.42	37.57
ECU05_0800	237.12	211.34	197.65
ECU05_0810	118.71	127.21	117.6
ECU05_0820	104.62	114.27	115.72
ECU05_0830	93.26	105.29	114.78
ECU05_0840	118.79	198.01	204.17
ECU05_0850	64.56	61.33	67.09
ECU05_0860	237.93	244.39	255.54
ECU05_0870	148.52	98.88	92.53
ECU05_0880	177.67	188.44	186.06
ECU05_0883	9.86	16.48	17.7
ECU05_0885	413.07	302.42	277.81
ECU05_0890	118.92	109.22	104.26
ECU05_0900	1728.68	1229.02	1078.83
ECU05_0920	1491.71	883.86	805.97
ECU05_0930	10.47	19.63	17.23
ECU05_0940	130.67	85.61	84.72
ECU05_0950	113.08	110.66	99.58
ECU05_0960	154.78	120.65	107.4
ECU05_0970	36.52	30.62	31.68
ECU05_0980	43.95	41.68	44.28
ECU05_0990	24.9	24.61	23.14
ECU05_1000	48.55	48.16	42.94
ECU05_1010	279.44	216.16	211.66
ECU05_1020	145.47	103.92	104.57
ECU05_1030	418.73	293.66	284.88
ECU05_1040	1099.16	887.11	881.02
ECU05_1050	58.99	68.13	65.2
ECU05_1060	122.41	83.35	75.81
ECU05_1070	130.33	119.66	110
ECU05_1075	96.23	79.65	73.01
ECU05_1080	544.57	582.33	559.23
ECU05_1090	104.95	91.28	90.38
ECU05_1100	103.42	106.49	104.61
ECU05_1110	47.34	55.19	49.64
ECU05_1130	115.71	103.51	101.92
ECU05_1140	57.06	43.58	40.21
ECU05_1150	53.32	40.97	39.22
ECU05_1160	140.55	212.03	199.26
ECU05_1165	461	563.38	504.08
ECU05_1170	467.3	335.18	337.18

Gene	T1	T2	T3
ECU05_1180	1332.97	1539.61	1403.53
ECU05_1190	131.44	131.38	138.86
ECU05_1200	179.5	273.62	267.66
ECU05_1210	61.65	85.65	101.85
ECU05_1220	39.06	84.55	99.4
ECU05_1230	212.84	166.73	161.65
ECU05_1240	80.78	125.11	130.51
ECU05_1250	95.65	165.76	176.46
ECU05_1260	446.23	433.75	417.16
ECU05_1270	76.57	83.58	86.69
ECU05_1280	148.96	124.51	112.93
ECU05_1290	40.22	51	43.85
ECU05_1300	45.46	49.93	45.93
ECU05_1310	558.96	453.9	442.23
ECU05_1320	157.05	131.79	113.83
ECU05_1330	368.68	244.96	266.3
ECU05_1340	466.29	405.72	412.24
ECU05_1350	70.26	40.46	37.38
ECU05_1355	6.1	5.34	3.65
ECU05_1360	107.85	116.84	121.45
ECU05_1370	39.43	39.79	43.92
ECU05_1380	185.09	157.24	155.24
ECU05_1390	37.32	53.66	60.18
ECU05_1400	431.59	444.74	410.55
ECU05_1410	80.25	76.51	84.87
ECU05_1420	37.03	39.39	44.31
ECU05_1430	171.15	130.42	143.95
ECU05_1440	1266.64	879.63	884.75
ECU05_1450	139.47	100.9	106.41
ECU05_1460	50.29	43.95	48.16
ECU05_1470	76.37	51.46	50.47
ECU05_1480	132.08	103.76	95.21
ECU05_1490	1141.12	524.05	463.66
ECU05_1495	60.6	27.64	38.26
ECU05_1500	186.3	127.68	130.65
ECU05_1510	126.47	98.17	89.96
ECU05_1520	536.33	355.7	350.18
ECU05_1530	100.89	88.71	74.5
ECU05_1540	395.59	490.21	478.52
ECU05_1550	2374.63	1815.72	1733.09
ECU06_0080	54.63	58.32	54.72
ECU06_0090	14.62	16.11	16.85
ECU06_0100	22.43	16.25	14.25
ECU06_0110	40.04	29.18	34.63
ECU06_0120	217.88	199.14	194.14
ECU06_0130	63.26	46.99	39.83

Gene	T1	T2	T3
ECU06_0140	84.08	63.36	65.13
ECU06_0150	50.27	43.79	45.08
ECU06_0155	256.29	202.34	193.56
ECU06_0157	270.78	205.16	206.12
ECU06_0160	173.17	122.5	126.16
ECU06_0170	73.07	56.15	62.49
ECU06_0175	186.15	148.2	177.08
ECU06_0180	177.72	153.83	139.39
ECU06_0185	73.04	83.52	70.72
ECU06_0190	121.32	139.83	136.77
ECU06_0200	61.99	55.29	58.12
ECU06_0210	589.72	447.86	395.01
ECU06_0220	84.65	80.42	81.52
ECU06_0240	606.54	2671.76	2938.79
ECU06_0250	659.76	2521.93	2683.03
ECU06_0260	200.58	197.34	201.71
ECU06_0270	138.32	141.38	130.55
ECU06_0280	306.23	225.44	224.14
ECU06_0290	138.77	109.43	118.28
ECU06_0300	76.59	171.99	181.71
ECU06_0310	105.55	101.07	94.15
ECU06_0320	211.96	116.94	96.37
ECU06_0330	248.69	137.67	119.04
ECU06_0340	167.27	122.97	120.97
ECU06_0350	105.27	86.77	86.87
ECU06_0360	415.42	322.64	283.36
ECU06_0380	21.82	42.57	44.58
ECU06_0390	71.2	75.26	76.71
ECU06_0400	158.59	155.87	146.3
ECU06_0405	6.08	8.88	13.65
ECU06_0410	87.52	78.33	66.46
ECU06_0415	80.86	48.07	39.52
ECU06_0420	278.8	216.21	208.49
ECU06_0430	164.94	104.44	114.12
ECU06_0435	87.03	116.32	111.24
ECU06_0440	44.29	40.11	39.98
ECU06_0450	141.17	126.62	125.52
ECU06_0460	22.56	21.03	25.64
ECU06_0470	199.49	160.44	153.84
ECU06_0480	86.02	60.11	54.16
ECU06_0490	42.85	27.58	31.05
ECU06_0500	41.54	65.78	72.12
ECU06_0510	531.51	351.91	322.77
ECU06_0520	109.77	84.37	80.57
ECU06_0530	85.85	126.83	138.45
ECU06_0540	11.47	12.76	13.47

Gene	T1	T2	T3
ECU06_0550	249.42	208.12	207.31
ECU06_0560	141.5	144.12	147.79
ECU06_0570	101.17	175.96	186.73
ECU06_0575	655.41	2178.73	2276.28
ECU06_0580	203.7	890.78	1173.63
ECU06_0590	117.93	100.4	95.68
ECU06_0600	316.93	306.51	281.05
ECU06_0605	157.37	138.46	122.76
ECU06_0610	101.77	84.03	79.4
ECU06_0620	274.17	243.76	243.85
ECU06_0630	54.03	85.26	86.96
ECU06_0640	277.48	878.8	912.61
ECU06_0650	699	402.17	362.08
ECU06_0660	85.08	109	105.98
ECU06_0670	230.14	867.69	1106.42
ECU06_0680	259.97	233.17	234.56
ECU06_0700	159.94	138.17	145.78
ECU06_0710	36.13	50.88	69.33
ECU06_0720	70.52	94.56	99
ECU06_0730	1406.88	1113.29	1241.43
ECU06_0735	54.99	54.05	59.82
ECU06_0750	907.74	580.93	546.5
ECU06_0760	224.8	165.83	148.21
ECU06_0770	62.45	178.99	187.95
ECU06_0780	125.37	98.68	96.53
ECU06_0790	306.34	223.61	207.87
ECU06_0800	214.06	223.39	205.02
ECU06_0810	151.93	99.53	88.71
ECU06_0820	66.76	42.31	46.85
ECU06_0830	215.4	208.29	207.84
ECU06_0840	119.19	105.19	96.22
ECU06_0850	72.78	62.75	66.9
ECU06_0860	362.98	321.78	298.44
ECU06_0870	316.48	303.36	320.85
ECU06_0880	85.83	98.51	89.26
ECU06_0890	125.54	84.7	76.07
ECU06_0905	29.99	50.93	73.28
ECU06_0910	44.12	55.69	53.77
ECU06_0920	19.32	18.59	18.49
ECU06_0930	136.4	147.04	153.9
ECU06_0935	824.27	685.64	646.68
ECU06_0940	206.41	166.91	164.45
ECU06_0950	64.24	107.73	119.79
ECU06_0960	201.94	194.83	212.28
ECU06_0970	60.49	36.66	33.73
ECU06_0980	76.35	50.62	54.84

Gene	T1	T2	T3
ECU06_0990	952.26	567.13	523.12
ECU06_1000	37.53	44.61	41.79
ECU06_1010	101.24	95.37	90.07
ECU06_1020	143.19	135.01	127.84
ECU06_1030	69.93	70.58	69.27
ECU06_1040	21.9	23.57	21.73
ECU06_1050	127.18	102.24	99.95
ECU06_1060	137.35	69.65	85.51
ECU06_1070	32.24	21.01	20.59
ECU06_1090	159.09	329.11	367.31
ECU06_1110	2042.47	991.72	925.79
ECU06_1120	1591.59	909.05	805.92
ECU06_1130	160.43	114.56	106.19
ECU06_1135	2079.24	1332.97	1225.73
ECU06_1140	201.91	116.97	107.68
ECU06_1150	172.69	114.98	125.38
ECU06_1160	46.49	48.62	57.13
ECU06_1170	114.42	84.03	83.02
ECU06_1180	571.17	544.43	528.35
ECU06_1190	118.97	79.44	86.73
ECU06_1200	53.29	93.09	117.09
ECU06_1210	29.13	72.84	66.66
ECU06_1215	114.67	90.71	87.22
ECU06_1220	178.53	166.46	161.31
ECU06_1230	101.02	69.89	64.07
ECU06_1240	81.69	52.62	55.95
ECU06_1250	155.98	125.69	126.48
ECU06_1255	64.08	43.17	51.63
ECU06_1260	134.09	87.26	92.77
ECU06_1270	1242.01	871.78	809.59
ECU06_1280	256.74	256.09	252.23
ECU06_1290	122.83	132.2	147.66
ECU06_1295	8.24	4.28	8.57
ECU06_1300	110.33	66.73	76.63
ECU06_1310	57.82	64.67	77.79
ECU06_1320	13.05	8.39	7.3
ECU06_1330	93.41	71.56	64.16
ECU06_1340	172.75	143.39	148.99
ECU06_1350	497.61	315.76	277.94
ECU06_1360	41.53	35.97	36.54
ECU06_1370	145.97	109	110.16
ECU06_1380	72.64	207.49	214.15
ECU06_1390	139.5	109.67	107.52
ECU06_1400	180.04	119.94	109.58
ECU06_1405	378.59	224.52	218.59
ECU06_1410	28.49	21.49	25.29

Gene	T1	T2	T3
ECU06_1430	58.27	52.35	55.16
ECU06_1440	171.08	119.48	112.07
ECU06_1445	1888.51	1056.16	1037.96
ECU06_1450	201.56	202.96	210.07
ECU06_1460	63.07	116.65	128.74
ECU06_1470	69.93	110.8	119.83
ECU06_1480	52.47	52.56	56.61
ECU06_1490	51.92	47.61	42.17
ECU06_1500	196.07	286.38	296.66
ECU06_1510	109.21	92.81	90.05
ECU06_1520	162.55	184.44	189.15
ECU06_1530	731.66	378.21	375.92
ECU06_1540	44.09	92.17	102.23
ECU06_1550	295.69	192.35	192.3
ECU06_1560	57.19	52.95	50.25
ECU06_1570	271.5	203.92	217.02
ECU06_1575	15.99	42.97	31.41
ECU06_1580	74.15	253.07	266.33
ECU06_1590	386.37	522.94	535.64
ECU06_1600	16.19	18.23	23.85
ECU06_1610	40.32	32.07	30.23
ECU06_1620	14.24	10.18	11.04
ECU07_0070	5.37	4.11	4.7
ECU07_0080	203.12	157.21	160.76
ECU07_0090	60.18	46.52	46.09
ECU07_0100	94.67	71.6	60.41
ECU07_0110	2678.9	1474.57	1360.18
ECU07_0120	47.43	32.55	40.63
ECU07_0130	702.96	498.82	455.6
ECU07_0140	37.52	31.32	31.26
ECU07_0150	92.54	51.92	44.41
ECU07_0160	152.82	116.71	105.95
ECU07_0170	43.09	74.91	96.88
ECU07_0180	56.28	56.03	62.84
ECU07_0190	245.59	283.83	298.72
ECU07_0200	58.45	49.94	55.47
ECU07_0210	382.15	290.18	278.43
ECU07_0220	65.39	45	43.64
ECU07_0230	28.36	25.16	22.27
ECU07_0235	14.99	7.96	5.98
ECU07_0240	266.56	133.99	145.01
ECU07_0250	181.91	269.44	298.07
ECU07_0260	51.9	64.42	65.72
ECU07_0270	27.64	38.39	38.93
ECU07_0280	86.74	75.37	82.65
ECU07_0290	1061.91	1298.44	1373.07

Gene	T1	T2	T3
ECU07_0300	340.6	298.06	315.92
ECU07_0310	110.55	108.36	108.12
ECU07_0320	61.52	59.33	59.4
ECU07_0330	273.28	230.86	265.92
ECU07_0340	35.05	29.79	30.52
ECU07_0350	44.12	36.89	37.6
ECU07_0360	147.1	149.42	145.94
ECU07_0370	38.08	45.35	52.39
ECU07_0380	60.59	53.18	52.54
ECU07_0390	55.19	53.46	53.86
ECU07_0400	733.78	2727.32	3013.21
ECU07_0410	310.4	353.05	347.91
ECU07_0420	118.19	186.85	220.36
ECU07_0430	98.46	127.06	112.27
ECU07_0440	42.6	51.11	52.05
ECU07_0450	67.51	67.58	65.24
ECU07_0455	190.2	166.81	145.77
ECU07_0460	294.01	201.87	224.26
ECU07_0470	87.4	60.11	79.58
ECU07_0480	120.38	126.73	146.64
ECU07_0490	42.36	31.22	28.96
ECU07_0500	43.54	165.51	181.94
ECU07_0510	53.76	66.94	73.81
ECU07_0520	144.38	91.29	106.13
ECU07_0530	2305.24	1848.39	1865.55
ECU07_0540	71.87	47.33	47.53
ECU07_0550	130.86	101.44	93.28
ECU07_0560	70.33	68.36	62.4
ECU07_0570	66.26	69.02	69.83
ECU07_0580	62.4	75.34	78.82
ECU07_0590	205.75	263.74	286.14
ECU07_0600	44.44	34.05	40.07
ECU07_0620	347.34	279	261.28
ECU07_0630	83.83	72.28	70.45
ECU07_0640	87.29	60.98	62.9
ECU07_0650	372.25	346.3	350.91
ECU07_0660	394.98	333.36	341.88
ECU07_0670	147.8	123.89	133.08
ECU07_0680	227.48	192.67	173.5
ECU07_0690	65.61	50.06	49.09
ECU07_0700	240.97	332.44	335.31
ECU07_0710	133.49	115.6	125
ECU07_0720	201.12	165.11	158.18
ECU07_0730	87.4	74.54	66.37
ECU07_0740	467.83	1339.12	1052.65
ECU07_0750	375.1	337.51	300.3

Gene	T1	T2	T3
ECU07_0760	555.59	325.81	420.92
ECU07_0770	76.61	69.41	80.09
ECU07_0780	36.16	24.95	23.45
ECU07_0790	143.44	122.18	121.5
ECU07_0800	200.76	178.92	202.12
ECU07_0810	100.49	96.55	112.24
ECU07_0820	1649.57	1021.19	1033.67
ECU07_0830	70.11	54.84	51.28
ECU07_0840	158.8	96.75	93.15
ECU07_0850	40.78	29.86	31.98
ECU07_0860	143.67	230.9	223.67
ECU07_0865	59.68	47.74	42.9
ECU07_0870	227.53	175	162.81
ECU07_0880	24.44	19.9	20.46
ECU07_0885	5.33	8.49	3.99
ECU07_0890	53.1	52.61	51.58
ECU07_0900	117.5	156.42	170.33
ECU07_0910	88.04	71.49	75.17
ECU07_0920	221.51	662.61	762.17
ECU07_0930	1813.17	902.7	902.68
ECU07_0940	101.26	88.69	100.78
ECU07_0950	312.36	222.54	238.03
ECU07_0960	196.87	189.66	184.95
ECU07_0965	235.09	221.06	207.29
ECU07_0970	449.63	370.86	384.65
ECU07_0980	36.81	32.6	36.72
ECU07_0985	48.5	33.01	36.28
ECU07_0990	97.7	151.9	179.64
ECU07_1000	84.02	225.6	259.95
ECU07_1005	2173.38	1620.97	1618.82
ECU07_1010	272.77	204.92	174.41
ECU07_1020	53.89	60.1	54.56
ECU07_1030	28.68	86.01	97.54
ECU07_1040	645.56	555.62	548.78
ECU07_1050	132.64	117.29	110.67
ECU07_1060	47.11	62.38	56.38
ECU07_1070	47.43	65.04	70.56
ECU07_1080	228.27	398.63	363.64
ECU07_1090	256.34	501.71	538.67
ECU07_1100	125.66	109.23	118.39
ECU07_1110	125.84	140.89	130.83
ECU07_1120	36.01	30.57	30.69
ECU07_1130	173.18	156.82	136.98
ECU07_1140	35.56	29.39	27.48
ECU07_1150	63.41	50.98	52.44
ECU07_1160	126.56	79.68	104.38

Gene	T1	T2	T3
ECU07_1170	88.9	98.11	98.76
ECU07_1180	429.91	368.59	325.1
ECU07_1190	1382.61	1046.62	1170.78
ECU07_1200	152.18	104.81	93.61
ECU07_1210	421.99	391.92	382.35
ECU07_1220	101.2	94.25	97.3
ECU07_1230	331.97	296.63	281.07
ECU07_1240	30.37	19.82	32.26
ECU07_1250	432.24	493.64	496.4
ECU07_1260	47.09	149.74	156.35
ECU07_1270	36.62	41.97	42.2
ECU07_1280	719.01	912.35	973.96
ECU07_1290	87.75	89.97	85.67
ECU07_1300	486.75	371.22	376.61
ECU07_1310	116.24	91.7	91.32
ECU07_1320	347.8	241.77	214.08
ECU07_1330	246.61	191.79	177.94
ECU07_1340	279.07	211.29	201.55
ECU07_1350	579.91	569.61	548.6
ECU07_1360	31.01	64.4	67.85
ECU07_1370	185.02	228.8	221.3
ECU07_1380	1662.68	1133.57	1029.24
ECU07_1390	106.65	100.6	104.71
ECU07_1410	2498.74	1424.03	1310.72
ECU07_1420	721.15	618.41	625
ECU07_1430	317.16	224.13	204.07
ECU07_1440	77.63	95.44	85.13
ECU07_1450	204.79	150.84	130.28
ECU07_1460	2564.95	1651.64	1523.72
ECU07_1470	88.36	145.7	140.65
ECU07_1475	8.11	5.65	15.17
ECU07_1480	132.48	142.73	148.26
ECU07_1485	0.99	1.57	2.95
ECU07_1490	68.7	87.7	71.95
ECU07_1495	2.44	1.46	2.74
ECU07_1500	42.05	40.44	36.99
ECU07_1523	11.82	12.61	13.26
ECU07_1525	57.97	47.74	44.87
ECU07_1530	94.55	172.1	183.21
ECU07_1550	162.04	558.09	606.77
ECU07_1560	38.46	69.06	77.6
ECU07_1570	291.27	230.99	225.5
ECU07_1580	29.44	23.25	26.59
ECU07_1590	56.34	70.71	72.66
ECU07_1600	238.66	691.66	887.37
ECU07_1610	42.72	33.28	33.59

Gene	T1	T2	T3
ECU07_1620	278.52	237.01	241.87
ECU07_1630	59.92	55.96	58.2
ECU07_1640	274.71	273.34	282
ECU07_1650	105.77	98.47	90.67
ECU07_1660	147.11	103.51	105.12
ECU07_1670	57.08	43.66	39.66
ECU07_1680	194.02	142.39	141.6
ECU07_1690	101.4	91.46	88.75
ECU07_1700	1795.11	1119.69	1024.38
ECU07_1710	48.63	63.24	63.55
ECU07_1720	71.03	60.27	57.57
ECU07_1730	65.48	58.23	61.26
ECU07_1740	237.62	217.91	207.58
ECU07_1750	93.03	64.64	65.81
ECU07_1760	205.32	137.52	122.81
ECU07_1770	117.51	72.65	91.3
ECU07_1780	31.8	20.79	15.29
ECU07_1790	26.3	19.71	17.6
ECU07_1800	13.45	11.91	9.88
ECU07_1805	86.05	57.71	59.69
ECU07_1810	114.48	81.42	79.96
ECU07_1820	1640.01	1200.52	1093.16
ECU07_1830	466.94	353.98	341.75
ECU07_1840	442.46	388.2	361.82
ECU07_1850	99.26	103.36	96.25
ECU07_1860	239.4	188.78	168.07
ECU07_1870	54.78	68.63	67.94
ECU07_1880	50.75	43.43	44.52
ECU07_1890	50	65.45	62.11
ECU08_0060	57.24	68.5	59.95
ECU08_0070	1459.99	1371.57	1320.07
ECU08_0080	63.81	58.98	56.41
ECU08_0090	235.58	160.68	157.08
ECU08_0100	61.13	50.7	54.66
ECU08_0110	48.75	53.21	46.1
ECU08_0120	383.49	375.1	392.16
ECU08_0130	29.06	34.71	31.93
ECU08_0135	303.54	290.54	252.69
ECU08_0140	172.49	230.86	228.86
ECU08_0145	260.5	244.99	274.25
ECU08_0150	90.87	218.56	218.26
ECU08_0160	103.12	64.04	68.04
ECU08_0170	145.99	137.49	145.19
ECU08_0180	195.63	172.16	164.05
ECU08_0190	67.84	58.28	64.2
ECU08_0200	97.97	75.04	71.79

Gene	T1	T2	T3
ECU08_0210	194.55	243.75	266.76
ECU08_0215	2	0.8	2.99
ECU08_0230	41.48	71.65	80.42
ECU08_0240	188.18	206.54	202.95
ECU08_0250	83.77	67.4	73.12
ECU08_0260	75.55	48.6	53.54
ECU08_0265	130.39	109.04	104.79
ECU08_0270	155.59	122.12	125.58
ECU08_0280	604.69	575.83	520.28
ECU08_0290	52.23	37.36	38.97
ECU08_0300	126.18	89.6	90.92
ECU08_0310	106.95	64.77	56.9
ECU08_0320	117.89	71.82	70.87
ECU08_0325	382.39	265.09	308.58
ECU08_0330	271.78	182.67	196.87
ECU08_0340	182.74	190.39	196.1
ECU08_0360	114.98	79.15	80.1
ECU08_0370	2533.98	1494.93	1477.89
ECU08_0380	178.86	141.41	161.44
ECU08_0390	509.03	309.61	310.27
ECU08_0400	54.49	45.41	51.75
ECU08_0410	3867.87	3666.04	3592.97
ECU08_0420	298.53	293.03	286.48
ECU08_0430	2404.72	1759.05	1751.22
ECU08_0440	172.22	165.96	161.91
ECU08_0450	108.17	92.48	87.41
ECU08_0460	91.45	77.13	73.97
ECU08_0470	2481.94	1382.32	1293.43
ECU08_0480	173.49	151.33	135.92
ECU08_0490	362.28	269.12	260.66
ECU08_0495	248.65	159.76	150.48
ECU08_0500	68.14	51.53	43.17
ECU08_0510	28.86	19.25	22.15
ECU08_0520	158.13	123.44	122.33
ECU08_0530	257.22	243.58	230.27
ECU08_0540	3414.49	2538.41	2590.11
ECU08_0545	102.85	99.55	105.44
ECU08_0550	406.45	383.58	383.25
ECU08_0555	13.05	12.52	17.53
ECU08_0560	39.87	109.3	129.75
ECU08_0570	22.66	22.68	22.18
ECU08_0575	1	2.39	7.48
ECU08_0580	337.58	259.3	245.87
ECU08_0590	109.67	81.02	80.85
ECU08_0610	115.56	95.94	89.76
ECU08_0620	73.29	69.3	63.44

Gene	T1	T2	T3
ECU08_0630	19.23	10.84	13.87
ECU08_0640	35.11	33.33	29.71
ECU08_0650	64.38	65.06	59.62
ECU08_0655	40.98	48.94	42.62
ECU08_0660	39.46	40.57	37.01
ECU08_0665	32.53	29.82	19.18
ECU08_0670	169.69	139.66	126.99
ECU08_0680	397.98	448.68	414.08
ECU08_0690	44.65	43.96	40.13
ECU08_0700	264.88	280.29	284.79
ECU08_0710	111.36	88.7	91.2
ECU08_0720	135.94	279.24	330.35
ECU08_0730	269.93	822.17	924.73
ECU08_0740	195.51	253.68	263.37
ECU08_0750	73.14	71.73	63.56
ECU08_0760	32.13	28.39	28.07
ECU08_0765	298.17	224.75	175.61
ECU08_0770	32.07	31.24	30.71
ECU08_0780	139.08	128.3	134.75
ECU08_0790	180.95	221.27	238.99
ECU08_0800	391.82	457.79	382.87
ECU08_0810	160.31	221.01	213.71
ECU08_0830	2518.61	1535.89	1431.49
ECU08_0840	52.95	51.22	51.81
ECU08_0850	150.96	152.27	144.67
ECU08_0860	76.99	229.62	263.36
ECU08_0870	1937.12	1109.66	1040.66
ECU08_0880	59.22	41.31	34.69
ECU08_0890	56.18	54.84	55.08
ECU08_0900	215.62	139.28	137.8
ECU08_0910	223.79	157.49	159.31
ECU08_0920	229.54	230.74	209.96
ECU08_0930	169.76	146.99	140.59
ECU08_0935	4.93	14.13	16.23
ECU08_0937	2.57	10.23	11.54
ECU08_0940	164.76	112.07	106.57
ECU08_0950	81.73	188.84	150.59
ECU08_0960	10.41	17.96	15.57
ECU08_0970	88.59	119.15	120.67
ECU08_0980	289.86	198.78	183.21
ECU08_0985	93.06	74.08	78.1
ECU08_0990	251.17	230.1	213.52
ECU08_1000	33.83	39.15	36.21
ECU08_1010	205.25	189.49	189.92
ECU08_1020	361.33	327.39	318.88
ECU08_1030	38.18	43.87	37.07

Gene	T1	T2	T3
ECU08_1040	839.83	497.76	446.72
ECU08_1050	414.32	287.37	304.79
ECU08_1060	2018.6	992.57	938.3
ECU08_1070	871.88	507.48	506.43
ECU08_1080	630.11	411.7	442.04
ECU08_1090	85.39	67.13	72.62
ECU08_1100	272.61	293.07	290.13
ECU08_1110	1863.39	1233.49	1140.04
ECU08_1120	40.12	43.22	40
ECU08_1130	71.7	106.33	101.23
ECU08_1140	47.06	57.72	64.95
ECU08_1150	120.59	110.92	109.25
ECU08_1165	99.69	94.25	105.35
ECU08_1170	104.14	87.76	98
ECU08_1180	117.44	92.19	91.57
ECU08_1190	229.65	217.59	200.05
ECU08_1200	28.48	47.44	50.65
ECU08_1210	184	562.36	626.3
ECU08_1220	968.81	732.52	761.24
ECU08_1230	87.44	121.63	139.33
ECU08_1240	30.94	24.31	24.17
ECU08_1250	230.68	236.39	225.61
ECU08_1260	32.14	29.04	32.55
ECU08_1280	29.89	29.67	19.79
ECU08_1290	136.11	130.27	120.51
ECU08_1300	182.74	491.22	532.9
ECU08_1320	216.15	239.22	211.59
ECU08_1330	101.23	112.87	106.96
ECU08_1340	64.01	77.18	72.14
ECU08_1350	87.13	87.53	98.21
ECU08_1370	89	99.54	110.76
ECU08_1380	205.38	494.78	513.56
ECU08_1390	123.62	504.21	550.46
ECU08_1400	110.47	86.39	91.04
ECU08_1410	38.69	53.75	69.52
ECU08_1420	86.87	81.35	77.88
ECU08_1425	73.77	45.18	41.68
ECU08_1430	108.76	87.23	90.27
ECU08_1440	160.9	157.79	149.09
ECU08_1445	23.2	14.4	19.42
ECU08_1450	82.88	72.2	77.67
ECU08_1470	81.04	150.59	149.88
ECU08_1480	24.81	61.05	62.97
ECU08_1490	103.52	278.68	277.3
ECU08_1500	35.54	56.5	68.63
ECU08_1510	436.69	266.36	268.97

Gene	T1	T2	T3
ECU08_1520	60.99	57.87	69.5
ECU08_1530	197.54	162.36	162.97
ECU08_1540	307.08	210.17	190.72
ECU08_1550	2714.73	12059.29	11562.46
ECU08_1555	124.92	45.4	56.89
ECU08_1560	43.51	37.73	37.48
ECU08_1570	2259.2	1364.79	1282.56
ECU08_1580	385.36	362.69	394.07
ECU08_1590	34.48	33.65	30.66
ECU08_1600	106.4	96.72	90.08
ECU08_1610	48.51	36.66	40.86
ECU08_1620	49.16	41.55	44.77
ECU08_1630	180.92	148.04	162.8
ECU08_1640	429.21	349.58	357.01
ECU08_1650	71.14	57.05	53.5
ECU08_1660	168.74	92.55	80.69
ECU08_1670	659.42	396.97	383.87
ECU08_1680	89.27	107.56	122.15
ECU08_1690	119.75	98.28	118.46
ECU08_1700	92.97	156.98	176.79
ECU08_1710	79.69	54.81	54.32
ECU08_1720	76.17	211.75	243.82
ECU08_1730	160.33	373.78	453.93
ECU08_1740	145.07	90.02	98.61
ECU08_1750	20.5	14.35	15.66
ECU08_1760	55.49	59.12	54.29
ECU08_1770	122.28	96.73	92.54
ECU08_1780	1995.53	1099.79	996.69
ECU08_1790	79.55	81.13	82.67
ECU08_1810	1117.3	5445.29	5658.12
ECU08_1820	106.58	87.47	87.54
ECU08_1830	42.04	27.59	37.16
ECU08_1840	42.17	35.19	37.34
ECU08_1850	136.54	138.86	150.18
ECU08_1860	19.43	17	18.49
ECU08_1865	9.86	9.42	7.38
ECU08_1870	415.63	512.07	441.34
ECU08_1880	35.92	35.58	34.36
ECU08_1885	168.49	1017.8	765.6
ECU08_1890	40.86	68.04	72.91
ECU08_1900	56.25	46.82	41.06
ECU08_1910	693.86	432.94	417.86
ECU08_1915	24.65	58.08	76.71
ECU08_1920	132.71	131.18	133.91
ECU08_1930	917.92	1306.73	1379.33
ECU08_1950	160.9	260.31	283.46

Gene	T1	T2	T3
ECU08_1960	175.28	192.55	215.16
ECU08_1970	431.24	497.92	512.05
ECU08_1980	51778.04	42881.99	37961.71
ECU08_1990	652.51	641.26	643.3
ECU08_2000	154.71	99.89	99.41
ECU08_2010	1542.32	902.4	898.21
ECU08_2020	173.96	467.75	564.85
ECU08_2030	20.94	50	58.74
ECU08_2040	92.33	114.95	110.39
ECU08_2050	108.44	94.34	97.77
ECU08_2060	13.9	9.69	8.16
ECU08_2070	18.14	12.95	10.43
ECU09_0020	50.98	27.87	30.62
ECU09_0030	122.12	91.28	94.12
ECU09_0040	747.35	446.23	430.41
ECU09_0050	603.41	370.13	371.09
ECU09_0060	103.91	97.48	94.84
ECU09_0070	148.55	137.08	130.97
ECU09_0080	8.97	18.69	13.43
ECU09_0090	66.19	56.12	56.96
ECU09_0100	177.24	139.18	147.25
ECU09_0110	47.84	51.49	54.37
ECU09_0120	27.45	33.92	33.22
ECU09_0130	511.74	498.11	507.73
ECU09_0140	135.7	111.71	109.95
ECU09_0150	58.37	50.36	51.35
ECU09_0160	191.17	178.82	177.03
ECU09_0170	203.36	229.56	219.33
ECU09_0180	107.02	119.93	117.86
ECU09_0190	69.4	81.43	97.28
ECU09_0200	56.85	110.19	135.14
ECU09_0210	55.74	39.87	42
ECU09_0220	11.68	6.29	7.8
ECU09_0230	74.23	63.45	69.19
ECU09_0240	91.53	75.58	71.15
ECU09_0250	35.34	25.59	24.92
ECU09_0260	106.48	87.68	83.07
ECU09_0270	164.16	113.65	110.15
ECU09_0275	267.87	205.7	201.9
ECU09_0280	83.59	87.05	88.46
ECU09_0290	571.58	496	523.32
ECU09_0300	119.46	140.71	150.13
ECU09_0310	39.13	34.05	31.76
ECU09_0320	97.29	82.06	78.53
ECU09_0330	416.89	489.31	481.39
ECU09_0340	78.77	84.29	88.98

Gene	T1	T2	T3
ECU09_0350	67.4	52.72	56.21
ECU09_0360	31.71	53.57	56.58
ECU09_0370	97.44	80.36	70.68
ECU09_0380	70.82	40.92	42.73
ECU09_0390	195.96	157.25	164.79
ECU09_0395	1137.43	645.72	667.09
ECU09_0400	99.9	74.14	70.51
ECU09_0405	0.6	0.48	0.9
ECU09_0410	72.18	87.27	84.57
ECU09_0420	70.1	88.88	81.29
ECU09_0425	63.24	53.82	46.77
ECU09_0430	117.73	120.93	117.66
ECU09_0440	3296.38	2692.41	2926.06
ECU09_0450	2155.25	1535.68	1672.51
ECU09_0460	181.18	149.9	145.54
ECU09_0470	434.44	278.76	274.64
ECU09_0480	685.64	508.59	470.29
ECU09_0490	122.09	78.38	83.02
ECU09_0500	380.93	281.29	254.72
ECU09_0510	70.92	56.8	55.25
ECU09_0520	185.37	169.09	160.46
ECU09_0530	15.52	15.17	15.84
ECU09_0550	556.21	461.13	440.29
ECU09_0560	78.59	71.73	69.29
ECU09_0570	268.98	189.82	192.16
ECU09_0590	42.1	44.6	39.83
ECU09_0600	105.18	95.12	84.67
ECU09_0610	42.34	35.84	32.98
ECU09_0620	240.07	168.49	167.21
ECU09_0625	409.1	428.78	404.87
ECU09_0630	176.64	191.94	184.91
ECU09_0640	87.96	100.83	97.96
ECU09_0650	64.28	119.22	113.26
ECU09_0660	34.87	43.49	38.76
ECU09_0670	78.36	81.76	88.64
ECU09_0680	601.9	435.58	391.69
ECU09_0690	246.4	216.78	218.45
ECU09_0700	85.99	79.93	80.81
ECU09_0710	176.76	165.94	148.56
ECU09_0720	395.39	378.27	407.43
ECU09_0730	28.73	19.31	18.41
ECU09_0740	278.44	227.85	239.41
ECU09_0750	46.93	45.23	44.71
ECU09_0760	169.64	134.3	139.42
ECU09_0770	70.16	41.84	41.97
ECU09_0780	102.62	73.54	79.02

Gene	T1	T2	T3
ECU09_0790	152.34	117.03	104.82
ECU09_0800	64.77	47.44	49.87
ECU09_0810	52.14	42.01	36.88
ECU09_0820	227.8	450.58	494.88
ECU09_0830	65.46	52.93	57.26
ECU09_0840	38.61	36.06	38.57
ECU09_0850	127.35	108.46	113.89
ECU09_0860	137.07	164.55	152.16
ECU09_0870	93.6	126.17	120.37
ECU09_0880	54.41	81.49	73.1
ECU09_0890	400.8	323.16	316.8
ECU09_0900	36.7	29.12	27.2
ECU09_0910	231.72	156.77	158.68
ECU09_0920	39.81	27.36	28.65
ECU09_0925	294.16	201.24	214.47
ECU09_0930	95.25	121.2	116.26
ECU09_0940	1006.41	1057.37	1096.65
ECU09_0945	588	440.55	414.01
ECU09_0950	89.03	66.41	70.87
ECU09_0960	399.13	305.44	258.11
ECU09_0970	20.92	24.41	21.86
ECU09_0980	35.15	43.3	39.21
ECU09_0990	73.53	93.51	86.78
ECU09_1000	983.6	589.94	559.82
ECU09_1005	210.54	161.72	145.49
ECU09_1010	467.54	350.84	338.21
ECU09_1020	182.61	124.36	105.73
ECU09_1030	82	57.02	61.61
ECU09_1040	55.31	80.27	91.68
ECU09_1050	66.29	62.84	66.66
ECU09_1060	71.86	80.72	73.95
ECU09_1070	124.58	121.83	111.15
ECU09_1080	100.59	92.66	89.49
ECU09_1100	157.5	152.89	133.32
ECU09_1110	112.1	120.33	96.66
ECU09_1120	80.63	87.28	82.29
ECU09_1160	26.74	25.24	22.14
ECU09_1170	55.13	78.96	89.15
ECU09_1180	228.82	397.22	386.4
ECU09_1190	40.58	69.32	76.4
ECU09_1195	362.58	298.99	303.63
ECU09_1200	995.37	732.17	700.27
ECU09_1210	33.84	60.38	56.34
ECU09_1220	662.84	480.69	469.42
ECU09_1230	26.92	21.98	25.05
ECU09_1240	76.51	64.83	60.45

Gene	T1	T2	T3
ECU09_1250	735.47	421.91	414.68
ECU09_1255	135.21	93.32	97.08
ECU09_1260	133.66	115.05	105.38
ECU09_1270	33.72	29.81	29.77
ECU09_1275	1204.89	640.06	627.06
ECU09_1280	188.98	193.42	186.77
ECU09_1290	339.52	307.32	273.06
ECU09_1300	92.8	115.56	125.5
ECU09_1310	88.07	91.59	84.33
ECU09_1320	67.4	140.76	149.99
ECU09_1330	105.94	97.15	106.64
ECU09_1340	216.88	165.64	165.31
ECU09_1350	2053.74	1056.3	1006.98
ECU09_1360	163.5	113.35	107.73
ECU09_1370	1483.9	1302.77	1246.32
ECU09_1375	806.17	1688.32	1542.72
ECU09_1380	131.06	108.07	106.67
ECU09_1390	242.75	161.32	167.53
ECU09_1395	4.93	7.06	13.28
ECU09_1400	49.45	135.01	166.72
ECU09_1410	160.71	128.07	104.87
ECU09_1420	61.52	73.77	72.49
ECU09_1430	576.54	210.72	203.47
ECU09_1440	50.15	80.73	86.47
ECU09_1450	132.17	166.54	169.62
ECU09_1460	114.78	74.14	69.54
ECU09_1470	7938.22	4293.6	3894.51
ECU09_1480	50.71	35.14	36.28
ECU09_1490	97.73	100.47	100.82
ECU09_1500	25.92	21.99	19.9
ECU09_1510	87.5	90.24	81.12
ECU09_1520	21.19	11.67	16.35
ECU09_1525	56.23	45.68	35.03
ECU09_1530	2.25	0.6	0
ECU09_1550	38.37	50.47	48.43
ECU09_1560	729.51	1041.24	984.42
ECU09_1570	39.84	30.24	31.19
ECU09_1580	235.4	193.3	186.18
ECU09_1590	132.81	161.82	155.37
ECU09_1600	217.02	155.92	150.64
ECU09_1610	117.46	77.92	62.23
ECU09_1615	25.08	21.7	25.02
ECU09_1620	104.22	97.76	101.94
ECU09_1630	28.77	22.07	24.71
ECU09_1640	144.47	186.56	192.05
ECU09_1650	247.86	289.39	269.47

Gene	T1	T2	T3
ECU09_1660	86.13	96.8	112.84
ECU09_1670	72.99	76.37	76.91
ECU09_1680	594.68	576.69	568.67
ECU09_1690	199	171.75	164.62
ECU09_1695	725.59	484.41	450.58
ECU09_1700	130.84	137.69	142.89
ECU09_1710	84.46	72.01	81.04
ECU09_1720	297.31	298.66	275.16
ECU09_1730	182.39	159.73	152.15
ECU09_1740	128.36	153.78	160.75
ECU09_1750	30.92	28.14	28.36
ECU09_1760	127.94	120.95	107.95
ECU09_1770	175.05	123.31	122.82
ECU09_1780	61.75	41.64	40.69
ECU09_1785	4	0	8.97
ECU09_1790	170.35	156.95	167.34
ECU09_1800	61.65	51.94	52.63
ECU09_1805	387.86	297.32	276.57
ECU09_1810	63.7	51.51	54.52
ECU09_1820	293.73	263.43	268.36
ECU09_1830	107.57	117.87	133.51
ECU09_1840	373.2	445.71	411.82
ECU09_1845	61.83	88.76	75.83
ECU09_1850	62.99	71.43	60.79
ECU09_1860	43.54	30.1	28.35
ECU09_1870	184.3	128.61	126.62
ECU09_1880	47.4	129.13	158.45
ECU09_1890	75.82	85.96	84.7
ECU09_1900	25.72	25.99	22.23
ECU09_1910	112.57	69.82	70.38
ECU09_1920	136.41	79.76	76.36
ECU09_1930	37.91	24.81	28.49
ECU09_1940	44.42	81.78	90.68
ECU09_1950	1132.54	1453.66	1354.2
ECU09_1960	55.13	58.56	60.85
ECU09_1970	54.36	51.35	54.98
ECU09_1980	55.73	127.8	141.58
ECU09_1990	95.04	78.21	94.58
ECU09_2000	92.23	244.01	288.63
ECU09_2010	63.58	39.83	45.09
ECU10_0140	228.85	294.9	295.3
ECU10_0150	107.61	80.66	107.31
ECU10_0155	1.97	3.14	0
ECU10_0160	921.76	564.1	537.06
ECU10_0170	96.35	111.45	119.48
ECU10_0180	16.39	13.43	15.64

Gene	T1	T2	T3
ECU10_0190	539.85	523.4	539.58
ECU10_0195	3.94	3.92	5.9
ECU10_0200	72.79	71.71	73.16
ECU10_0210	69.23	42.87	47.1
ECU10_0220	246.5	161.29	181.54
ECU10_0230	109.71	101.95	89.74
ECU10_0240	460.03	295.24	308.76
ECU10_0250	149.86	123.93	115.35
ECU10_0260	315.08	522.7	527.85
ECU10_0270	161.07	120.45	106.92
ECU10_0280	115.27	81.71	81.48
ECU10_0290	27.5	38.51	52.82
ECU10_0295	12.29	17.47	18.38
ECU10_0300	37	42.19	43.93
ECU10_0310	82.63	83.53	97.04
ECU10_0320	95.86	106.39	112.02
ECU10_0330	35.98	54.84	51.54
ECU10_0340	736.06	1013	1022.59
ECU10_0345	134.71	82.69	92.69
ECU10_0350	81.14	121.33	133.37
ECU10_0355	102.37	101.71	121.04
ECU10_0360	334.61	233.58	236.3
ECU10_0370	142.21	200.87	177.92
ECU10_0380	284.32	365.75	334.8
ECU10_0390	167.22	169.61	166.74
ECU10_0400	2049.59	1194.76	1099.99
ECU10_0410	256.19	271.85	258.47
ECU10_0420	106.45	90.94	93.58
ECU10_0430	142.7	135.51	119.56
ECU10_0440	140.7	101.49	94.52
ECU10_0445	8.38	7.36	8.22
ECU10_0450	36.63	65.39	56.52
ECU10_0460	31.4	37.76	38.21
ECU10_0470	40.46	45.29	42.32
ECU10_0475	16.22	22.59	21.23
ECU10_0480	232.2	168.7	171.19
ECU10_0490	317.76	275.3	254.43
ECU10_0500	77.96	59.65	58.28
ECU10_0510	152.95	138.67	131.49
ECU10_0520	162.67	127.93	126.42
ECU10_0530	153.94	114.59	106.13
ECU10_0540	1292.23	806.52	832.37
ECU10_0550	424.55	460.87	438.37
ECU10_0560	101.02	84.05	85.72
ECU10_0570	46.34	37.12	40.79
ECU10_0580	35.91	50.89	46.43

Gene	T1	T2	T3
ECU10_0590	16.96	23.63	24.33
ECU10_0600	376.08	241.84	238.42
ECU10_0610	68.33	61	55.04
ECU10_0620	384.38	288.88	289.6
ECU10_0630	710.86	426.79	390.35
ECU10_0635	85.74	39.26	29.91
ECU10_0640	56.27	76.65	69.16
ECU10_0650	114.33	167.8	163.53
ECU10_0660	47.72	47.09	45.2
ECU10_0670	254.12	136.81	141.14
ECU10_0680	88.92	64.59	61.98
ECU10_0690	53.37	35.68	37.52
ECU10_0700	100.76	114.79	126.34
ECU10_0710	105.18	81.07	81.2
ECU10_0720	52.46	44.47	42
ECU10_0730	34.32	30.11	33.88
ECU10_0740	65.48	54.89	59.82
ECU10_0750	132.57	104.32	97.28
ECU10_0760	143.72	127.38	118.2
ECU10_0770	267.3	261.91	230.75
ECU10_0790	41.91	36.78	36.31
ECU10_0800	272.97	186.9	187.3
ECU10_0810	205.76	180.45	162.05
ECU10_0820	415.12	292.03	313.21
ECU10_0830	288.38	227.38	189.29
ECU10_0840	83.57	70.8	62.89
ECU10_0850	70.63	58.36	57.22
ECU10_0860	94.53	53.73	48.97
ECU10_0870	48.25	74.22	72.26
ECU10_0880	331.55	387.24	412.84
ECU10_0890	603.05	340.84	326.54
ECU10_0900	91.71	67.61	66.38
ECU10_0910	638.36	338.75	325.37
ECU10_0920	250.21	161.7	179.66
ECU10_0930	52.52	36.98	39.42
ECU10_0940	463.33	338.83	322.14
ECU10_0950	75.77	49.15	47.47
ECU10_0960	133.41	91.69	100.55
ECU10_0970	250.14	206.85	190.3
ECU10_0980	62.44	128.78	116.78
ECU10_0990	1615.03	959.66	908.02
ECU10_1000	47.98	75.6	76.26
ECU10_1010	135.15	180.35	205
ECU10_1020	107.84	94.45	87.94
ECU10_1030	286.72	189.76	181.65
ECU10_1040	288.72	248.11	255.99

Gene	T1	T2	T3
ECU10_1045	478.85	390.77	405.88
ECU10_1050	394.91	245.54	242.21
ECU10_1060	71.92	108.18	105.2
ECU10_1070	6612.44	3472.61	3069.65
ECU10_1080	86.17	88.33	85.58
ECU10_1090	178.09	138.35	126.6
ECU10_1100	97.19	86.7	94.37
ECU10_1110	706.36	401.24	394.26
ECU10_1115	8.11	17.75	6.07
ECU10_1120	96.76	75.05	70.08
ECU10_1130	55.01	40.06	37.02
ECU10_1140	108.78	123.49	112.81
ECU10_1150	95.14	117.33	108.45
ECU10_1155	83.42	156.25	160.71
ECU10_1160	91.09	82.32	80.99
ECU10_1170	17.96	15.3	21.24
ECU10_1180	178.97	194.83	177.58
ECU10_1190	53.38	87.96	103.22
ECU10_1210	46.62	36.77	32.94
ECU10_1220	144.28	111.98	106.91
ECU10_1230	77.87	46.56	57.19
ECU10_1240	144.97	114.66	111.88
ECU10_1250	21.3	39.22	40.53
ECU10_1260	126.38	140.7	146.21
ECU10_1270	28	26.92	28.46
ECU10_1280	73.24	60.08	57.05
ECU10_1290	130.89	112.08	117.33
ECU10_1300	2175.21	1503.81	1310.99
ECU10_1320	136.71	139.73	133.78
ECU10_1330	38.54	32.2	36.09
ECU10_1340	126.26	114.5	107.68
ECU10_1350	50.63	39.77	42.18
ECU10_1360	145.99	256.48	278.39
ECU10_1370	29.56	33.23	34.45
ECU10_1380	81.52	96.43	86.79
ECU10_1390	101.77	122.5	117.47
ECU10_1400	108.72	138.95	150.73
ECU10_1410	58.55	108.62	122.81
ECU10_1420	324.68	367.18	377.07
ECU10_1430	98.43	131.9	145.84
ECU10_1440	67.13	79.75	90.12
ECU10_1450	417.38	435.2	417.29
ECU10_1460	206.86	180.22	185.13
ECU10_1465	41.44	33.86	38.07
ECU10_1470	32.04	48	57.15
ECU10_1480	37.5	63.22	65.74

Gene	T1	T2	T3
ECU10_1490	96.33	127.02	157.14
ECU10_1500	92.77	234.45	268.11
ECU10_1505	66.32	50.07	55.2
ECU10_1510	28.68	21.15	20.74
ECU10_1520	87.72	117.79	124.55
ECU10_1530	67.34	55.02	58.62
ECU10_1540	156.72	121.5	106.74
ECU10_1550	39.19	27.98	27.01
ECU10_1560	37.29	27.61	28.48
ECU10_1575	1219.83	692.29	669.42
ECU10_1580	77.61	61.37	64.4
ECU10_1590	296.21	329.11	330.64
ECU10_1600	83.49	63	69.32
ECU10_1610	86.53	55.86	56.08
ECU10_1620	144.96	124.48	134.99
ECU10_1630	88.02	74.82	78.05
ECU10_1640	104.79	72.29	68.74
ECU10_1650	114.07	134.31	143.69
ECU10_1660	8504.6	22258.57	22702.63
ECU10_1680	323.8	427.53	507.78
ECU10_1690	48.91	66.77	69.53
ECU10_1695	34.87	19.29	31.82
ECU10_1710	133.57	99.11	95.04
ECU10_1720	61.44	52.55	54.34
ECU10_1730	196	297.61	320.58
ECU10_1740	515.95	498.6	499.55
ECU10_1750	66.85	66.27	76.83
ECU10_1760	107.17	85.23	84.11
ECU10_1780	132.39	142.29	145.44
ECU10_1790	198.86	142.51	150.94
ECU10_1800	93.18	68	64.33
ECU10_1810	30.32	36.8	36.01
ECU10_1820	19.56	16.69	16.03
ECU10_1830	24.69	23.14	23.32
ECU11_0060	63.12	57.14	39.13
ECU11_0070	63.67	58.49	53.99
ECU11_0080	29.08	28.9	29.54
ECU11_0090	9.45	11.71	8.27
ECU11_0100	115.01	85.94	83.05
ECU11_0110	18.43	8.09	10.34
ECU11_0120	66.94	70.17	66.81
ECU11_0130	1202.15	1076.8	1025.07
ECU11_0140	206.83	128.66	133.79
ECU11_0150	937.25	673.49	670.77
ECU11_0170	109.46	98.32	105.69
ECU11_0180	51.76	41.25	37.75

Gene	T1	T2	T3
ECU11_0190	172.9	176.52	174.03
ECU11_0200	63.71	42.76	44.35
ECU11_0210	87.68	69.24	69.27
ECU11_0220	81.54	63.91	65.96
ECU11_0225	881.17	519.14	508.08
ECU11_0230	104.05	106.73	107.53
ECU11_0240	179.6	145.52	153.56
ECU11_0250	60.45	46.53	46.99
ECU11_0260	62.74	69.36	63.29
ECU11_0270	81.97	103.01	107.08
ECU11_0280	119.68	85.62	80.22
ECU11_0290	103.07	73.02	79.53
ECU11_0300	82.3	44.8	43.91
ECU11_0310	303.63	253.51	240.13
ECU11_0320	58.81	98.04	109.9
ECU11_0330	171.47	181.36	175.26
ECU11_0340	22.15	40.51	52.25
ECU11_0350	92.66	69.47	62.72
ECU11_0360	88.34	98.43	66.96
ECU11_0370	42.39	40.55	38.19
ECU11_0373	92.57	210.56	211.68
ECU11_0375	6.22	6.37	18.61
ECU11_0380	2.98	6.41	6.69
ECU11_0390	73.04	75.15	76.77
ECU11_0400	254.01	291.27	302.74
ECU11_0410	413.51	434.59	454.06
ECU11_0415	123.23	160.89	191.76
ECU11_0420	187.09	226	214.58
ECU11_0430	30.3	47.83	53.32
ECU11_0440	25.87	26.19	26.41
ECU11_0450	60.9	80.11	71.41
ECU11_0460	136.93	134.02	135.16
ECU11_0470	95.17	61.95	58.96
ECU11_0480	244.27	213.71	191.59
ECU11_0490	77.67	84.77	80.46
ECU11_0500	108.24	86.45	75.99
ECU11_0505	322.96	232.99	204.79
ECU11_0510	960.49	2159.24	2382.24
ECU11_0520	146.55	119.45	112.14
ECU11_0530	1083.78	716.31	713.79
ECU11_0540	67.07	50.95	49.82
ECU11_0550	290.73	364.84	370.81
ECU11_0560	147.25	132.8	110.17
ECU11_0570	496.61	563.22	561.16
ECU11_0580	121.76	169.95	158.05
ECU11_0585	73.35	124.46	129.55

Gene	T1	T2	T3
ECU11_0590	102.12	161.15	146.65
ECU11_0600	340.55	192.39	178.43
ECU11_0610	87.52	80.47	74.33
ECU11_0620	115.67	98.99	109.14
ECU11_0630	109.95	102.25	93.47
ECU11_0635	380.75	313.89	284.39
ECU11_0640	132.17	149.25	140.15
ECU11_0650	34.84	41.82	40.77
ECU11_0660	182.48	132.6	133.64
ECU11_0670	157.57	107.82	100.21
ECU11_0680	286.93	239.58	230.62
ECU11_0690	133.74	336.92	338.71
ECU11_0700	237.67	164.66	152.93
ECU11_0710	166.61	131.7	126.5
ECU11_0720	1345.95	994.5	949.26
ECU11_0730	284.25	236.76	189.06
ECU11_0740	30.47	28.65	25.95
ECU11_0750	329.75	267.95	240.69
ECU11_0760	61.77	43.4	41.57
ECU11_0770	124.41	106.31	103.18
ECU11_0780	1422.77	831.59	795.13
ECU11_0790	330.86	383.3	390.98
ECU11_0800	44.11	43.12	44.58
ECU11_0810	311.85	256.09	255.28
ECU11_0820	220.48	172.62	173.58
ECU11_0830	237.25	125.8	147.86
ECU11_0840	150.76	191.36	188.58
ECU11_0850	163.46	162.31	138.55
ECU11_0860	154.03	111.05	110.97
ECU11_0870	669.38	807.01	795.78
ECU11_0880	307.92	216.27	207.4
ECU11_0890	401.54	405.77	396.53
ECU11_0900	65.07	63.06	63.93
ECU11_0910	348.17	229.35	222.83
ECU11_0920	122.06	83.96	74.27
ECU11_0935	36.67	19.92	27.69
ECU11_0940	187.11	120.66	98.87
ECU11_0950	224.85	184.67	186.7
ECU11_0960	114.42	122.8	121.39
ECU11_0970	68.29	48.05	48.92
ECU11_0980	435.76	470.15	461.04
ECU11_0990	37.14	35.78	35.89
ECU11_1000	31.13	33.12	37.8
ECU11_1010	65.83	73.81	63.96
ECU11_1020	170.16	113.13	116.33
ECU11_1030	49.22	52.63	50.13

Gene	T1	T2	T3
ECU11_1040	50.41	36.81	37.45
ECU11_1050	57.06	138.4	154.54
ECU11_1060	141.61	127.56	129.38
ECU11_1065	110.29	107.64	110.01
ECU11_1070	32.6	59.45	67.62
ECU11_1080	377.48	306.66	310.24
ECU11_1090	54.83	46.54	44.94
ECU11_1100	363.88	297.23	275.16
ECU11_1110	60.96	60.02	51.45
ECU11_1120	103.52	83.56	80.09
ECU11_1130	86.91	72.13	63.38
ECU11_1140	91.69	97.66	92.93
ECU11_1150	181.21	270.21	273.81
ECU11_1160	70	65.56	63.34
ECU11_1170	157.3	135.14	129.93
ECU11_1180	23.99	24.36	21.45
ECU11_1190	93.09	78.41	74.12
ECU11_1200	90.32	157.96	173.58
ECU11_1205	107.71	67.03	78.83
ECU11_1210	319.8	744.36	801.55
ECU11_1220	47.88	41.79	42.21
ECU11_1230	109.47	108.84	105.25
ECU11_1240	157.46	152.92	142.43
ECU11_1250	328.91	314.06	307.28
ECU11_1255	7	4.77	2.99
ECU11_1260	83.41	59.44	60.7
ECU11_1270	282.77	191.11	177.16
ECU11_1290	54.16	57.17	43.12
ECU11_1300	347.84	1358.24	1570.45
ECU11_1310	45.48	53.21	58.72
ECU11_1320	79.15	88.86	80.82
ECU11_1330	220.66	412.87	450.28
ECU11_1340	431.04	334.06	353.99
ECU11_1350	307.95	291.51	287.96
ECU11_1360	133.49	103.13	79.39
ECU11_1370	96.63	97.16	93.91
ECU11_1390	157.85	129.89	112.44
ECU11_1400	171.88	140.34	137.98
ECU11_1410	23.9	23.38	23.63
ECU11_1420	121.75	92.6	126.13
ECU11_1425	174.28	155.54	123.25
ECU11_1430	88.05	57.75	57.12
ECU11_1440	200.68	519.15	574.17
ECU11_1450	482.24	417.28	424.02
ECU11_1460	1378.66	846.36	848.47
ECU11_1470	240.14	252.65	247.15

Gene	T1	T2	T3
ECU11_1480	113.09	93.08	96.22
ECU11_1490	59.82	66.84	57.85
ECU11_1500	52.37	54.78	44.33
ECU11_1510	34.59	79.93	96.97
ECU11_1520	181.59	202.48	221.4
ECU11_1530	73.81	66.26	65.02
ECU11_1540	183.72	174.85	186.77
ECU11_1560	45.54	55.83	77.27
ECU11_1570	84.65	69.42	68.9
ECU11_1580	76.84	53.11	55.19
ECU11_1590	262.07	262.18	251.39
ECU11_1600	75.78	59.32	63.19
ECU11_1610	120.09	105.54	100.93
ECU11_1630	182.25	131.52	125.87
ECU11_1640	971.28	826.03	877.59
ECU11_1650	672.46	448.95	455.33
ECU11_1660	118.86	111.61	106.78
ECU11_1670	561.5	483.4	435.7
ECU11_1680	445.53	474.24	483.34
ECU11_1690	904.69	1187.08	1213.36
ECU11_1700	68.16	100.07	117.52
ECU11_1710	93.48	183.2	201.86
ECU11_1720	1172.24	941.93	958.44
ECU11_1723	143.09	185.95	205.92
ECU11_1725	502.54	490.71	417.95
ECU11_1730	176.01	199.11	209.55
ECU11_1740	74.08	69.98	64.4
ECU11_1750	127.7	95.63	95.76
ECU11_1755	92.41	68.57	64.8
ECU11_1760	211.31	176.88	187.61
ECU11_1770	138.89	97.39	104.03
ECU11_1780	85.6	303.46	302.86
ECU11_1790	157.82	166.96	165.32
ECU11_1800	66.57	97.07	98.13
ECU11_1810	62.71	75.28	83.18
ECU11_1820	308.1	245.36	279.74
ECU11_1830	623.78	337.41	401.59
ECU11_1840	77.56	230.71	259.17
ECU11_1850	234.1	886.62	903.95
ECU11_1860	85.41	91	89.52
ECU11_1870	702.29	962.48	1035.71
ECU11_1880	178.44	164.94	174.08
ECU11_1890	178.91	144.22	137.92
ECU11_1900	110.22	125.6	125.72
ECU11_1910	117.72	89.55	95.05
ECU11_1920	273.18	310.91	307.56

Gene	T1	T2	T3
ECU11_1930	101.45	80.83	71.91
ECU11_1935	70.38	48.9	38.66
ECU11_1940	79.18	51.87	51.78
ECU11_1950	125.19	342.05	397.65
ECU11_1960	465.75	491.45	526.99
ECU11_1970	47.21	39.62	36.28
ECU11_1980	32.41	41.73	41.75
ECU11_1990	274.78	362.31	384.03
ECU11_2000	93.5	82.81	81.51
ECU11_2010	140.84	136.72	131.34
ECU11_2020	69.52	83.17	82.84
ECU11_2030	112.05	271.21	322.78
ECU11_2033	124.67	94.72	100.61
ECU11_2037	5.56	5.53	5.1
ECU11_2040	107.79	84.05	91.24

Table B.2: RNA decay genes in *Encephalitozoon cuniculi*

Table of six key RNA decay pathway genes found in *E. cuniculi*. Gene names in yeast and are shown, as well as the protein BLAST e-values.

Gene	*E. cuniculi* name	Similarity to yeast gene (e-value)
Upf1/Nam7	ECU10_1640	9.00E-154
Dcp2	ECU07_1630	2.00E-25
Dis3	ECU03_0700	6.00E-127
Dhh1	ECU09_1640	2.00E-117
Ccr4	ECU11_0770	1.00E-78
Nmd5	ECU10_0620	3.00E-31

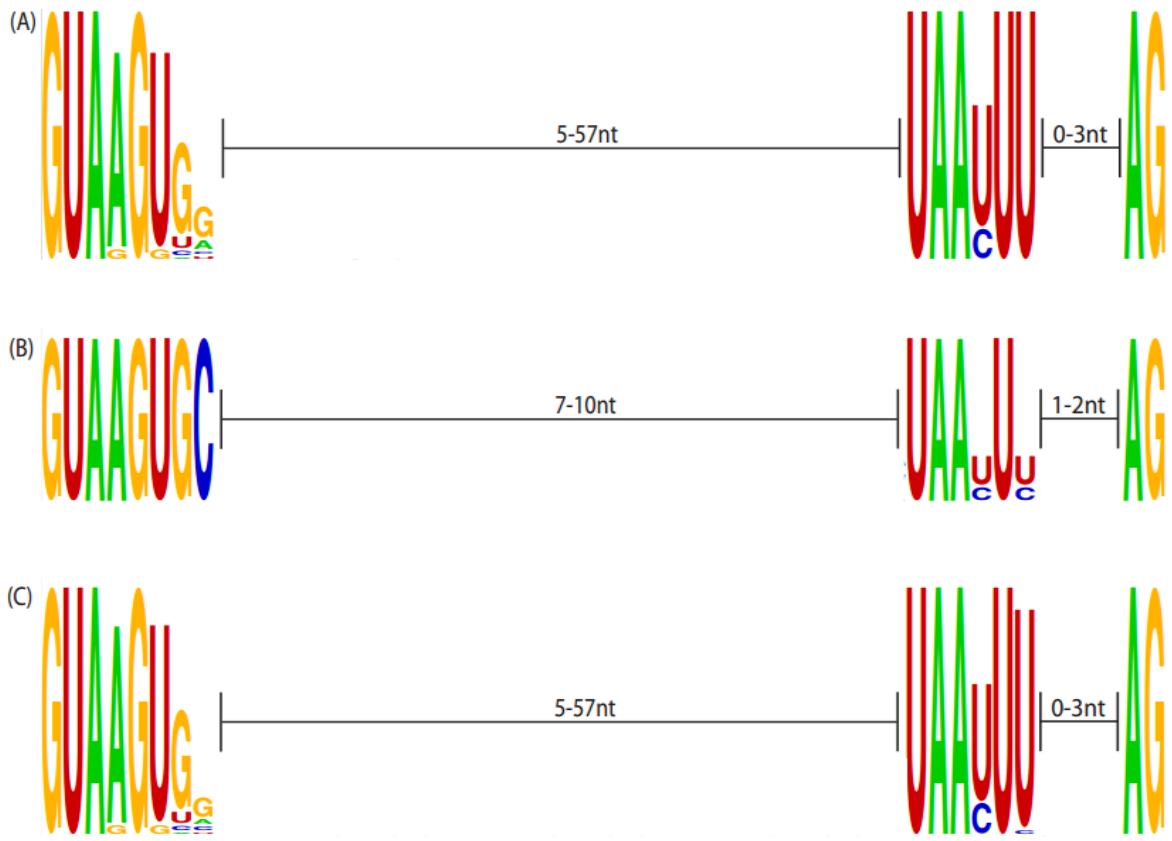
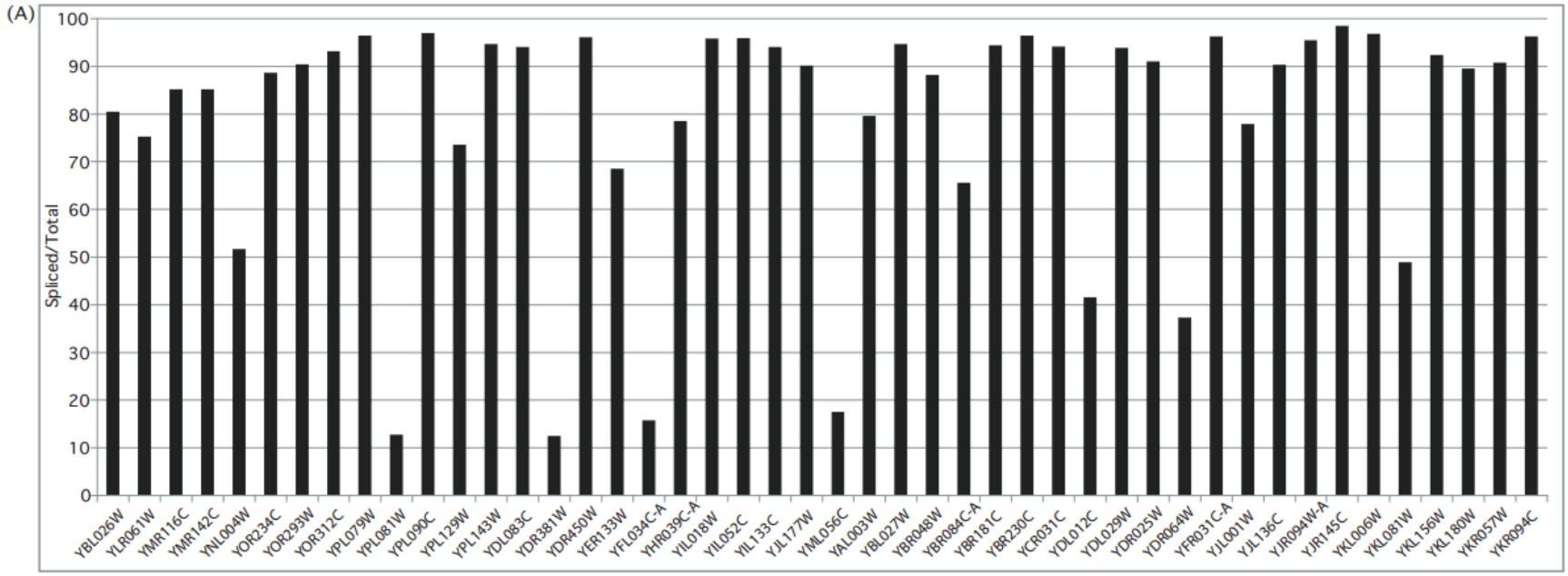


Figure B.1: Intron motifs of *Encephalitozoon cuniculi introns*

(A) Weblogo of 34 *E. cuniculi* intron motifs, showing strict 5' splice site, branch point, and 3' AG. (B) Weblogo of three recently discovered introns, with intron motifs that are consistent with currently annotated introns. (C) Combined old and new data for a total of 37 introns, showing very little change from (A).


(A)
Spliced/Total
100
90
80
70
60
50
40
30
20
10
0
YBL026W
YLR061W
YMR116C
YMR142C
YNL004W
YOR234C
YOR293W
YOR312C
YPL079W
YPL081W
YPL090C
YPL129W
YPL143W
YDL083C
YDR381W
YDR450W
YER133W
YFL034C-A
YHR039C-A
YIL018W
YIL052C
YIL133C
YJL177W
YML056C
YAL003W
YBL027W
YBR048W
YBR084C-A
YBR181C
YBR230C
YCR031C
YDL012C
YDL029W
YDR025W
YDR064W
YFR031C-A
YJL001W
YJL136C
YJR094W-A
YJR145C
YKL006W
YKL081W
YKL156W
YKL180W
YKR057W
YKR094C

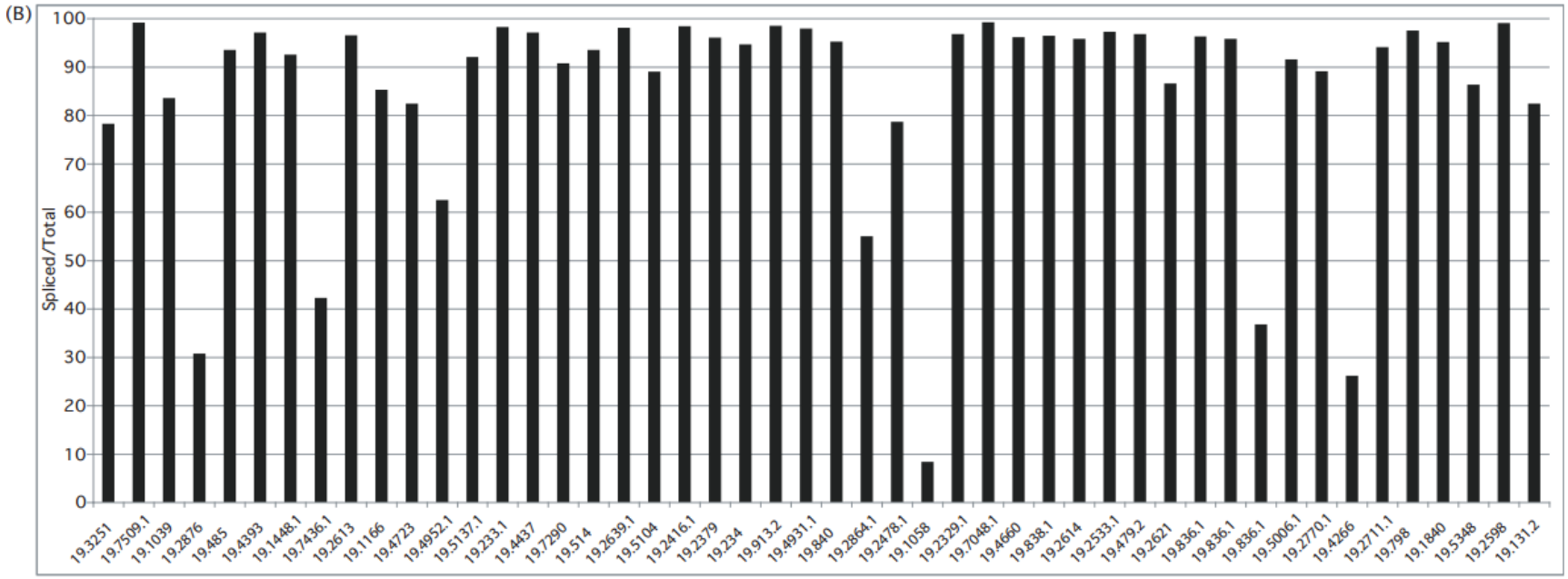


Figure B.2: Splicing levels in two fungal species

Levels of splicing found for 46 *Saccharomyces cerevisiae* introns (A) and 48 *Candida albicans* introns (B). Splicing level was measured by counting the number of spliced and unspliced transcripts and then dividing spliced by total transcripts to give a percentage of splicing