BUILDING A FRAMEWORK: THE *SARCOCYSTIS NEURONA* GENOME PROJECT

by

JOSHUA SEUNG DEOK BRIDGERS

(Under the Direction of Jessica Kissinger)

ABSTRACT

Sarcocystis neurona is an obligate intracellular parasite and the main agent behind equine protozoan myeloencephalitis a disease of critical importance to the equine industry. To facilitate the *S. neurona* genome project this study developed a bioinformatics framework that included assembly of the apicoplast genome, assembly and testing multiple transcriptome assembly algorithms and evaluating each to identify the superior assembly. Also, a pipeline for annotation of the nuclear genome was developed and its effectiveness was demonstrated via annotation of the *S. neurona* apicoplast genome. Finally, in order to facilitate study of and access to, the *S. neurona* genome, a database, SarcoDB, was created. It is intended that the results of this study will greatly aid in our understanding of the parasite, *Sarcocystis neurona.*

INDEX WORDS: Annotation, Assembly, Apicoplast, DNA, EST, *Eimeria tenella*, Genome, *Sarcocystis neurona*, Transcriptome, *Toxoplasma gondii*

BUILDING A FRAMEWORK: THE *SARCOCYSTIS NEURONA* GENOME PROJECT

by

JOSHUA SEUNG DEOK BRIDGERS

B.S., The University of Arizona, 2002

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTERS OF SCIENCE

ATHENS, GEORGIA

2011

BUILDING A FRAMEWORK: THE *SARCOCYSTIS NEURONA* GENOME PROJECT

by

JOSHUA SEUNG DEOK BRIDGERS

Major Professor: Jessica Kissinger

Committee: Andrew Patterson
Boris Striepen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2011

DEDICATION

To God who has given me the strength to persevere.

ACKNOWLEDGEMENTS

I would like to thank the following people, first, my parents who have supported me in all my endeavors. Next, I would like to thank Dawn Geiser and Joy Winzerling from my first lab. Amanda and Jeff you have always been there for me. Next I would not have been able to write this thesis today without the help and support of so many people at UGA. From the Kissinger laboratory, I would like to thank Dr. Jessica Kissinger, for the opportunity to work in her lab; followed by many of my fellow lab mates, Ranjani, Allyson, Mark, Brian, Betsy, Ganesh, and Haiming with special thanks going to Dr. Jeremy Debarry who has tolerated my incessant questions and brainstorming sessions. I would also like to thank other members of the IOB, Dr. Travis Glenn, Dr. Jeffery Dean, Tim and Anuj. Also thank you to my committee members, Dr. Andrew Patterson and Dr. Boris Striepen. All of this work was made possible by the RCC and the fantastic support offered by their staff, especially Yecheng Huang. Finally, I would like to thank my collaborators at the University of Kentucky, Dr. Dan Howe and Dr. Sriveny Dangoudoubiyam.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND REVIEW

Purpose of the study

The number of genomes being sequenced is increasing rapidly. Despite the leaps in DNA sequencing technology and throughput, the annotation of these genomes is lagging. It is genome annotation, not DNA sequence generation that is the bottleneck in genomics. This study directly confronts this bottleneck. I provide the requisite bioinformatics framework for the *Sarcocystis neurona* genome and transcriptome project. *Sarcocystis neruona* is an intracellular obligate parasite. It belongs to the phylum Apicomplexa which is compose of many parasites, the most notorious of which is *Plasmodium falciparum*. *Sarcocystis neurona* infects a wide range of organisms and is the causative agent of equine protozoan myeloencephalitis, the primary neurological disease in horses. In order to better understand this organism, its nuclear and organellar genomes have been sequenced. However, raw sequence alone is not enough; we need an annotation to provide knowledge of its genes and genomic regions of interest.

Annotation of an entire genome is no trivial task and requires the hard work and dedication of multiple highly skilled researchers. It also requires careful organization, planning, and the right set of tools. This genome project includes the following phases: DNA and RNA processing from appropriate starting material; genome and transcriptome sequencing; genome and transcriptome assembly; transcriptome alignment to the genome; gene prediction; genome

annotation and visualization; and finally database construction. Bioinformatics techniques facilitate nearly every step of this process. In this study, I will discuss my involvement with the *Sarcocystis neurona* genome project. Specifically I have contributed to the construction of a bioinformatics framework to expedite the study of the nuclear genome, creation assembly and annotation pipelines for the transcriptome and annotation of the apicoplast organellar genome.

Life cycle and host range of *Sarcocystis neruona*

A genome project begins with an understanding of the organism and its biology. *Sarcocystis neurona* is an intracellular coccidian parasite that is the primary agent of equine protozoal myeloencephalitis, EPM (Davis, Daft et al. 1991). EPM is a major neurological disease of horses that infects all regions of the central nervous system, CNS (Dubey, Davis et al. 1974). Despite *S. neurona*'s importance to the equine industry little is known about how this organism can infect such a broad range of species.

The lifecycle of *Sarcocystis neurona* is complex and involves a definitive host, intermediate hosts and many different tissue types. The cycle begins with the opossum, the definitive host. The definitive host was proven to be the opossum (*Didelphis viginiana*) for *S. neurona* when sporocysts obtained from opossums were fed to horses and induced EPM (Fenger, Granstrom et al. 1997). Dubey and Lindsay continued this work with interferon-gamma gene knockout mice that were fed sporocysts from opossums. The mice developed neurological symptoms similar to those seen in horses (Dubey, Speer et al. 1998). The opossum becomes infected by ingesting sarcocyst from infected muscle tissue of an intermediate host. Upon digestion of the sarcocysts, bradyzoites, the infectious form of the parasite are released

from the sarcocyst, infiltrating the small intestines. Bradyzoites can then undergo a process called gametogony or reproduction by means of gametes, to form oocysts. These oocysts then undergo sporogony, reproduction by multiple fission, to create sporulated oocysts. The oocysts contain 2 sporocysts which contains four sporozoites for a total of 8 sporozoites per oocyst. The oocysts will often rupture releasing the sporocysts into the intestinal lumen which are then excreted by the opossum into the environment.

Sporocysts are then ingested by an intermediate host in either tainted food or water. The sporocyst then releases the sporozoites in the small intestines. The development of the sporocyst in the lymph node arteries is currently not known. In the arteries, through asexual reproduction, schizonts or meronts are produced. In the schizonts, the nucleus and organelles are repeatedly replicated. Then through cytokinesis the multinucleated schizont divides into numerous daughter cells, all merozoites, in a process called merogony. Merozoites will then emerge from the schizonts and eventually infect the muscle tissue of the intermediate host. Merozoites can then form schizonts and undergo multiple rounds of merogony. Eventually, after a varying number of generations the merozoites transform into a sarcocyst filled with bradyzoites. Upon death of the intermediate host, opossums then consume the sarcocyst tainted muscle tissue thus completing the cycle (Samuel, Kocan et al. 2001). In the case of horses, *S. neurona* does not infect the muscle tissue but rather the central nervous system. Infection occurs from fecal to oral transmission. Horses are thought to be aberrant or dead end hosts as only schizonts and merozoites, not sarcocysts, were found in the infected tissues (Dubey, Lindsay et al. 2001). However, mature sarcocysts were found in the tongue in one case for a young horse (Mullaney, Murphy et al. 2005).

Sarcocystis neurona is known to infect a wide range of intermediate hosts which include, but are not exclusive to, armadillos (Cheadle, Tanhauser et al. 2001), striped skunks (Cheadle, Yowell et al. 2001), raccoons (Dubey, Saville et al. 2001) , sea otters (Lindsay, Thomas et al. 2000), brown-headed cowbirds (Mansfield, Mehler et al. 2008), lemurs (Yabsley, Jordan et al. 2007), rhesus monkeys (Klumpp, Anderson et al. 1994), dogs (Cooley, Barr et al. 2007) and cats (Dubey and Hamir 2000; Dubey, Saville et al. 2000). In order to look at all species which are intermediate hosts to *S. neurona*, one must rise to the level of phylum because *S. neurona* infects species from both *Mammalia* and *Aves* classes. While the reasons for host specificity are largely unknown *S. neurona* may be evolving a generalist infection preference in order to survive as a species.

Equine Protozoan Myeloencephalitis

Sarcocystis neurona is of interest because it is the number one agent for equine protozoan myeloencephalitis (EPM) the most prominent neurological disease in horses yet little is known about it. Therefore, the genome project is of high significance to the equine industry. In the US alone, the equine industry is worth $102 billion according to the American Horse Council's national study (http://www.horsecouncil.org/nationaleconomics.php). The cost treatment for EPM in the US has been estimated to range from $55.3 to $110.8 million per year. This amount does not include more indirect costs such as transport costs, euthanasia, etc. (Dubey, Lindsay et al. 2001). In a 1998 survey on the needs of the US equine industry, EPM was ranked first among infectious diseases (USDA, APHIS report, may 1997). EPM is a debilitating disease that can lead to death. Diagnosis and treatment of this disease remains expensive.

Sarcocystis neurona infects only the brain, and spinal cord in horses. Unlike intermediate hosts of *S. neurona*, it has not been found in muscles. The tissues that can be infected range from the anterior of the cerebrum to the end of the spinal cord. Once a cell is infected, *S. neurona* will undergo cycles of schizogony an undetermined amount of times, but multiple schizonts and hundreds of merozoites have been previously observed in a single neuron (Dubey 1974).

Symptoms of EPM are dependent on the area of the CNS that is infected. If the cerebrum has been infected the horse might show signs of depression, have seizures or undergo behavior changes. Damage to the brainstem and spinal cord can cause the horse to have an unusual gait or be uncoordinated. Other symptoms can include facial nerve paralysis, head tilt, ataxia, tongue paralysis, urinary incontinence, dysphagia and a tendency to lean to one side. Infection of the gray matter can manifest itself as atrophied or weak muscles in limbs (Dubey, Lindsay et al. 2001).

Once infected, EPM is characterized as a debilitating neurological disorder that ends in death. Diagnosis of the disease however still remains difficult. As of this writing, diagnosis of EPM involves testing for the presence of antigen-specific antibodies in cerebrospinal fluid. Current tests have been shown to have a high false positive rate. In tests, 32-89% of animals were found to be seropositive even though EPM affects <1% of horses (Furr, Howe et al. 2011). Once EPM is suspected, treatment should be administered as soon as possible. Treatment involves dihydrofolate reductase inhibitors such as sulfonamides and pyrimethamine (Mayhew 1976; MacKay 1992) over a course of 12 weeks or longer.

Related Species

Often in genome projects, genes and proteins from other closely related species are used to infer genes. Fortunately, *S. neurona* has many related species that are well annotated. *Sarcocystis neurona* belongs to the phylum Apicomplexa. Some prominent members of the Apicomplexa are the parasite species *Plasmodium falciparum*, *Toxoplasma gondii*, *Eimeria tenella*, *Babesia bovis* and *Theileria parvum*. These parasites cause a range of diseases for humans, livestock, wild animals and invertebrates the most notable of which is malaria which is responsible for the death of approximately one million people in 2008 alone (http://www.cdc.gov/MALARIA/). *Sarcocystis neurona* is a member of *Coccidia* (Figure 1.1).

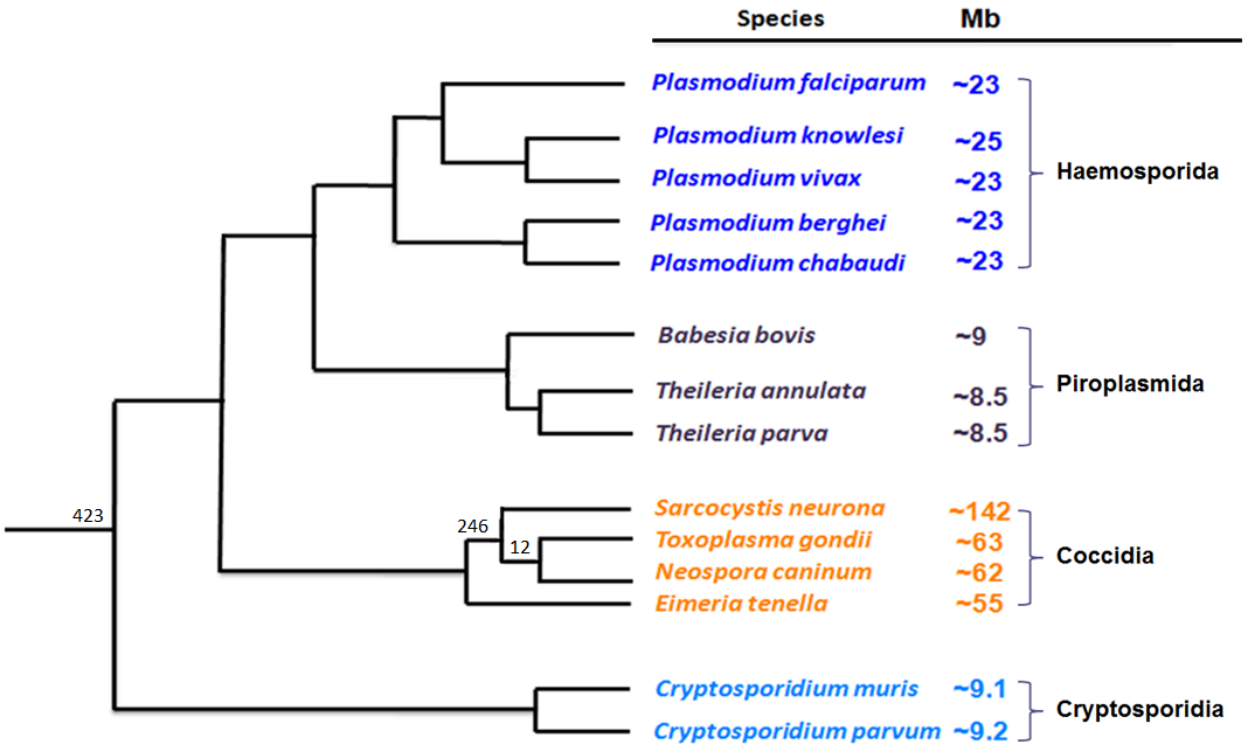

Figure 1.1: Cladogram of apicomplexan relationships
A cladogram of various apicomplexan species with genome sizes grouped by class. Numbers to the left of nodes are estimated divergence times in millions of years (Su, Evans et al. 2003; Okamoto and McFadden 2008). Length of lines is not to scale.

Its sister genera are the generalist parasite, *T. gondii* and a parasite of chickens, *E. tenella*. *Toxoplasma gondii* can infect humans and a wide range of other animals, to cause toxoplasmosis (Weiss and Kim 2007), while *E. tenella* is of major concern to the poultry industry as it causes coccidiosis in young chicks (Shirley 2000). *Sarcocystis neurona* is believed to be approximately equally distant, between *T. gondii* and *E. tenella*. Both *T. gondii* and *E. tenella* along with several other apicomplexans have had their nuclear, mitochondrial and apicoplast genomes sequenced. The *T. gondii* proteome and a set of orthologous proteins shared by related apicomplexan organisms will be utilized to facilitate gene finding. The orthologous proteins were identified (Kuo, Wares et al. 2008) using a custom pipeline involving the program orthoMCL (Li, Stoeckert et al. 2003). The orthologous protein set includes proteins shared by all Apicomplexa except *Cryptosporidium*; this same set including *Cryptosporidium* and a set including all apicomplexan and two ciliate outgroups, *Paramecium tetraurelia* and *Tetrahymena thermophilia.*

Apicomplexan genomes

The nuclear genome

Genomes can be incredibly diverse therefore it is important to understand the characteristics of our genome and apicomplexan genomes in general. Apicomplexan genomes tend to be small (<65Mb) and range from four – fourteen chromosomes. The karyotype of *S. neurona* is not known. The genome project was begun without an exact genome size, karyotype or genetic map. The nuclear genome of *S. neurona* was sequenced and assembled by the University of Kentucky's Advanced Genetic Technologies Center, however for educational

purposes, I also assembled the nuclear genome. The genome is estimated to be between 120-142 Mbp, based on estimations made by the assembly program Newbler and its resulting genome assembly. The May 26th 2011 genome assembly contained 123,739,627 bp of sequence assembled into 172 scaffolds. Contour-clamped homogeneous electric field (CHEF) gel experiments are underway in Dr. Jessica Kissinger's laboratory at the University of Georgia to help determine the karyotype. Previous CHEF gel experiments in Dr. Kissinger's laboratory performed by Dr. Gregorio Cordon and Ranjani Namasivayam suggest that *S. neurona* chromosomes are too large to even enter the gel.

Sarcocystis.neurona's genome is quite large for an apicomplexan with the next largest sequenced genome of an apicomplexan is *T. gondii* at 62 Mbp. Other species have even smaller genomes such as *Theileria parva* with a genome of only 8.5 Mbp. Genome sizes of major apicomplexan species can be seen in Figure 1.1, Despite *S. neurona's* large genome for an apicomplexan, it is still a small genome when compared with other eukaryotic organisms but genome reduction and gene loss are common within parasite genomes (Kuo and Kissinger 2008; DeBarry and Kissinger 2011). For example, *Cryptosporidium parvum* has lost all pathways for *de novo* synthesis of nucleotides (Striepen, Pruijssers et al. 2004). The *S. neurona* genome is roughly twice the size of its closest neighbor; however there is no evidence, as of yet, for whole genome duplication. Based upon the preliminary results of transcriptome data mapped onto the *S. neurona* genome, then average introns length is longer than those observed in *T. gondii* but not long enough to explain *S. neurona*'s genome size (Table 1.1).

Table 1.1: Apicomplexan Genome Characteristics

Comparison of *S. neurona*, *T. gondii* and *P. falciparum* genomes. * denotes that the value is currently unknown. Number of predicted genes for *S. neurona* is from preliminary work with the gene predictor Augustus (Stanke and Waack 2003). For *T. gondii* and *P. falciparum*, numbers were calculated from ToxoDB (Gajria, Bahl et al. 2008) and PlasmoDB (Aurrecoechea, Brestelli et al. 2009) respectively.

Features	*S. neurona*	*T. gondii*	*P. falciparum*
Size (Mbp)	123.0	63.0	25.0
Number of chromosomes	*	14.0	14.0
Apicoplast genome size (Kbp)	35.0	35.0	34.0
Mitochondrial genome size (Kbp)	*	6.0	6.0
# of predicted genes	9,245.0	7,934.0	6,372.0
Mean protein length	620.0	708.0	752.0
Mean exon size	532.0	730.0	1,506.0
Mean intron size	979.8	600.5	178.7
GC% content	51.4	52.3	19.4

Initial results from the gene prediction programs Augustus (Stanke and Waack 2003) and glimmerHMM (Majoros, Pertea et al. 2004) also hint that there has been no significant expansion of the number of genes in *S. neurona* compared to *T. gondii.* Analysis of the genome's repeat content however provides some clues. When the *de novo* repeat finding program RepeatModeler (Smit and Hubley 2008-2010) was used it found 26.45% of *S. neurona's* genome to contain repetitive elements verses *T. gondii's* 8.59%.

The apicoplast genome

Many eukaryotic organisms contain both a nuclear genome and a mitochondrial genome but within the many apicomplexan species there one additional organellar genomes, known as the apicoplast. It is a relict chloroplast. The apicoplast is of great interest due to the fact that many drugs target the apicoplast. *Plasmodium falciparum's* which is the agent behind malaria was the first to have its apicoplast genome annotated (Wilson, Denny et al. 1996). It

was then followed by two coccidian species, *Toxoplasma gondii* (Kissinger et al., unpublished: database sequence with GenBank accession U87145) and *Eimeria tenella* (Cai, Fuller et al. 2003). Afterwards two piroplasmid species were sequenced and annotated, *Theileria parva* (Gardner, Bishop et al. 2005) and *Babesia bovis* (Brayton, Lau et al. 2007) species that cause the diseases theileriosis and babesiosis in cattle, respectively.

The apicoplast is the result of a secondary endosymbiosis (Roos, Crawford et al. 1999) (Figure 1.2). This refers to an event where a eukaryote engulfed another eukaryote (thought to be a red alga in this case) that had a plastid acquired as a result of a primary endosymbiosis. The result of secondary endosymbiosis is a plastid with three or four membranes.

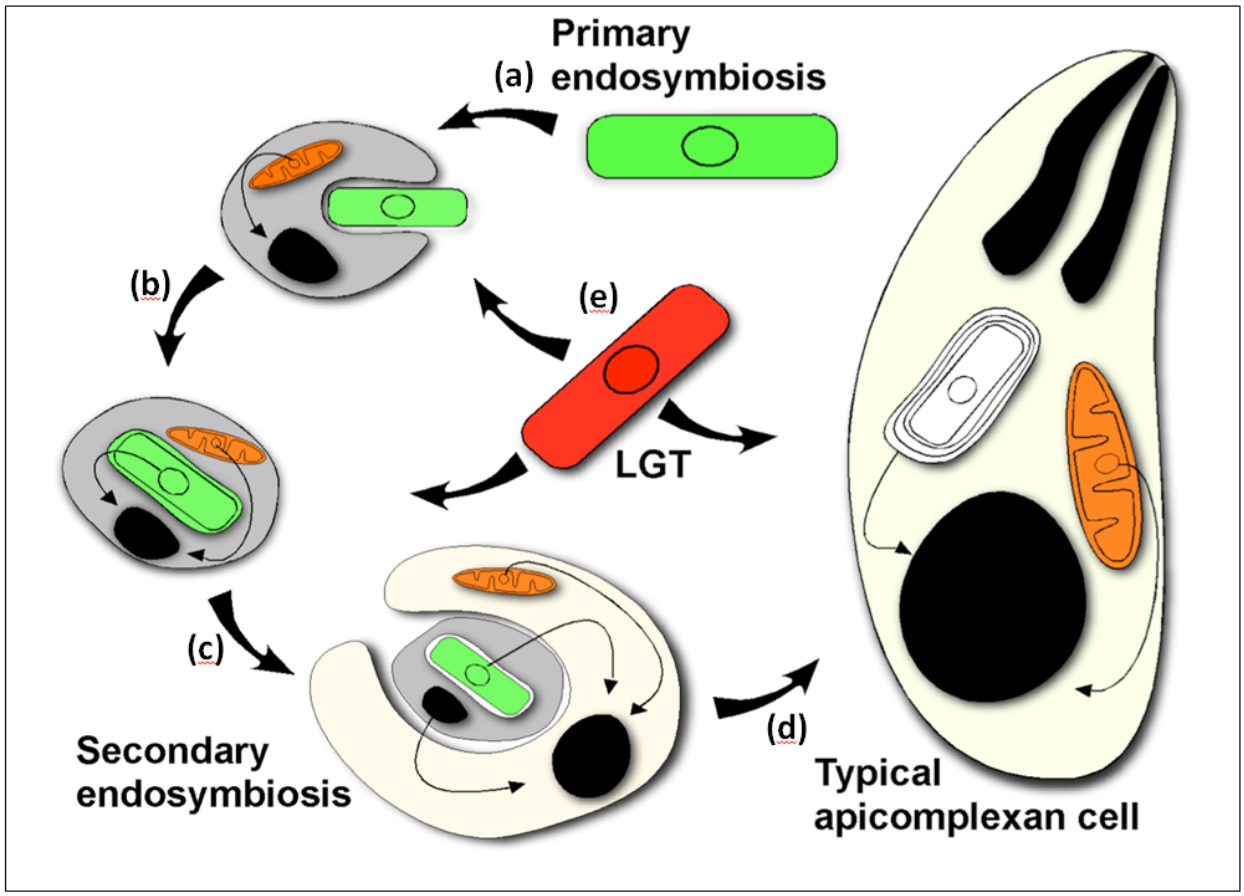

Figure 1.2: Evolution of the apicoplast through secondary endosymbiosis

(a) Primary endosymbiosis of a cyanobacterium (green) by a eukaryotic cell to create a photosynthetic eukaryotic cell; **(b)** Transfer of genes from the chloroplast (green) to the nuclear genome (black) in an algal cell; **(c)** Secondary endosymbiosis of a red alga by another eukaryotic cell. Continued gene transfer from the algal nucleus and chloroplast to the new 'host' nuclear genome; **(d)** Loss of the algal nucleus to give rise to modern apicomplexan cell; **(e)** Lateral gene transfer (LGT) from external sources (red) is occurring throughout evolution of the Apicomplexa. Gene transfer continues from organelles to the nuclear genome. This figure is reproduced in part from (Huang and Kissinger 2006).

With the discovery of a photosynthetic apicomplexan *Chromera* (Moore, Obornik et al. 2008), the apicoplast is thought to be of red algal origin. Furthermore, the structure of the apicoplast is more closely associated with red algae (Blanchard and Hicks 1999; McFadden 2000; Fast, Kissinger et al. 2001; Harper and Keeling 2003). In the case of *S. neurona,* four continuous membranes were found (Tomova, Geerts et al. 2006). The outermost membranes of the apicoplast are thought to be related to the plasma membrane of the eukaryotic endosymbiont. The two innermost membranes are thought to be derived from the primary symbiont.

Despite humanity's best efforts, diseases caused by the Apicomplexa are still prevalent. As previously mentioned the apicoplast is an exciting target for drug development to combat these organisms. The apicoplast is essential for survival in the species that have it (Fichera and Roos 1997; McConkey, Rogers et al. 1997). Apicomplexans treated with antimalarial and anti-coccidian drugs often display a 'delayed death' phenotype. When *Plasmodium falciparum*, for example, was treated with drugs targeting basic housekeeping processes in the apicoplast, the parasite progresses through most of the erythrocyte stage but after the parasite forms multi-nucleated schizonts it was unable to form functional merozoites (Dahl, Shock et al. 2006; Dahl

and Rosenthal 2007; Goodman, Su et al. 2007; Ramya, Mishra et al. 2007; Sidhu, Sun et al. 2007). *Toxoplasma gondii* also has been shown to die when it enters into a host cell if it lacks plastid function (Fichera, Bhopale et al. 1995). Thus, with treatment, the parasite will not instantly die but rather die later in subsequent stages or the next infective cycle, hence the term "delayed death".

Since the apicoplast is unique to the Apicomplexa and has a plant ancestry, this opens the door for many possibilities for medical treatment with low toxicity to the host (McFadden, Reith et al. 1996; McFadden and Roos 1999). Today there are numerous drugs available to use against apicomplexan infections in the field however not all drugs have similar effectiveness for different species and the mechanisms in which they kill parasites are often not fully understood (Dahl and Rosenthal 2008). Furthermore, there have been cases where the parasite has developed drug resistance (Yuthavong, Panijpan et al. 1985).

As often happens after an endosymbiotic event, many genes from the endosymbiont are transferred to the host's nuclear genome (Martin and Herrmann 1998). Although many genes have been transferred to the nuclear genome, the apicoplast is still the target for many proteins encoded by the nuclear genome. These proteins contain a bipartite organellar transit peptide at the N terminus (Waller, Keeling et al. 1998; Waller, Reed et al. 2000) consisting of a signal peptide and a targeting/transit peptide. The signal sequence of *P. falciparum* contains many asparagine and lysine residues which are different than what is found in plants, where they use serine and threonine (Waller, Reed et al. 2000). The transit peptide also usually contains putative Hsp70-binding sites. With this information, the prediction program PlasmoAP (Foth, Ralph et al. 2003) was created. It predicted more than 500 nuclear-encoded apicoplast-

targeted proteins in the *P. falciparum*'s that have an apicoplast targeting sequence at the N terminus. Of the proteins whose function can be inferred from sequence similarity with proteins of known function or structure, the most prominent are enzymes involved with fatty acid, heme, and isoprenoid *de novo* biosynthesis (Ralph, van Dooren et al. 2004). Other proteins are associated with housekeeping functions such as DNA polymerase, DNA gyrase, ribosomal proteins and chaperones. Finally many proteins involved in iron-sulphur cluster synthesis are present as well (Ralph, van Dooren et al. 2004; Sato, Clough et al. 2004).

The apicoplast genome is highly AT-rich. (Gardner, Bishop et al. 2005; Brayton, Lau et al. 2007). It is also small when compared to photosynthetic plastid genomes. Chloroplasts of green algae typically range between 100-200 Kbp, but there are many exceptions (Barbrook, Howe et al. 2010). Overall, the gene content of the apicoplast genome is highly conserved (Sato 2011). The apicoplast genome contains both SSU and LSUrRNAs, RNA polymerase beta chain (*rpoB*), RNA polymerase beta' chain (*rpoC1*) and RNA polymerase beta'' (*rpoC2*). All currently sequenced apicoplast genomes lack the gene for the alpha subunit (*rpoA*) unlike what is seen in other plastid genomes. However it is present in the nuclear genome. The apicoplast genome also contains *EF-Tu*, a *ClpC*-like protein and numerous tRNA species. Another feature in the *P. falciparum* and coccidian apicoplast genomes is the inverted repeat (IR). Each half contains the SSU and LSUrRNAs with 9 tRNAs with the IR is arranged in a head to head manner. In the P*iroplasmida* species, *Babesia bovis* and *Theileria parva* there is no IR (Sato 2011).

Within some genes of coccidian apicoplast genomes there are in-frame UGA and UAA codons. In plastid genomes, there are two proteins that are release factors known as *RF1* and *RF2*, used for translation termination. *RF1* binds to UAA and UAG codons while RF2 binds to

UAA and UGA codons (Scolnick, Tompkins et al. 1968; Nakamura and Ito 1998). Since the apicoplast genome lacks RFs they must be imported and while *Plasmodium* and *Piroplasmida* have RFs none have been found in *T. gondii* or *E. tenella* for *RF2* (Sato 2011). The genes that contain these codons are highly conserved and so UGA codons are thought to encode tryptophan in the apicoplast of species. Other examples of this usage are found in bacteria with extremely A + T rich genomes (Ohama, Inagaki et al. 2008). Therefore, most likely, these genes are not pseudogenes but genes that code for proteins. However there is no direct evidence of what these codons might encode (Sato 2011). *Plasmodium falciparum*'s *rpoC2* gene requires a frame shift in order to produce the correct translation product (Wilson, Denny et al. 1996). *Toxoplasma gondii's rpoC2* gene contains a UAA stop codon in the same region however does not require a frame shift (Kissinger et al., unpublished: database sequence with GenBank accession U87145).

The mitochondrial genome

The third genome within *S. neurona* is the mitochondrial genomes. Apicomplexan mitochondrial genomes are the smallest known containing only 3 protein coding genes and rRNA. *Cryptosporidium parvum* while containing a multi-membraned organelle does not contain a mitochondrial genome (Putignani, Tait et al. 2004). The three protein encoding genes that are found are highly conserved cytochrome genes, *cob coxI* and *coxIII*. As of this writing, the *S. neurona* mitochondrial genome has not been clearly defined. This does not come as a surprise as the mitochondrial genome for *T. gondii* has not been defined for *T. gondii* either. Elucidating the mitochondrial genome is complicated by numerous small mitochondrial sequence insertions into the nuclear genome (NuMts) (Weiss and Kim 2007). These fragments

are found throughout *T. gondii's* nuclear genome. When the *S. neurona* genome was masked by *T. gondii* mitochondrial genes, 6.82% of the genome was masked. Two mitochondrial genes have been found among small scaffolds in the *S. neurona* genome assembly, *coxIII* and *cob*. The gene *coxI* has been found in the transcriptome assembly. Like the apicoplast genome, there appears to be gene loss from the mitochondrial genome to the nuclear genome. Nuclear-encoded proteins are targeted to the mitochondria in *T. gondii.* Also, no tRNAs have been found in the mitochondrial genome but tRNA import into the mitochondrion has been seen for *T. gondii (Esseiva, Naguleswaran et al. 2004).*

Framework for Genome Annotation

Completion of the genome sequence and knowledge of the organism is insufficient information for annotation of the genome. Careful thought and planning are necessary to deal with the large amount of data associated with a genome project. My project was to develop a bioinformatics foundation for the genome annotation of *S. neurona*. The first step was to gather data that can be used to identify genes in the genome. Two important datasets have already been mentioned, the *T. gondii* proteome and the set of conserved orthologous proteins shared among the Apicomplexa and ciliates. The third dataset is the *S. neurona* transcriptome.

Transcriptome

In order to annotate the *S. neurona* nuclear genome, one of the best datasets one can have is transcriptome data because each sequenced transcript is direct evidence of a gene that can then be mapped back onto the genome. By mapping transcripts to the genome, the location expressed regions, exons, introns, and URTs can be determined, thereby providing

valuable information for annotation of genes. As sequence generation prices continue to fall, transcriptome sequence projects become more cost effective and available to smaller non-model organism species. Despite transcriptome datasets being much smaller than a genome project, transcriptome data still present their own unique challenges including: different assembly programs; alternative splicing; uneven sequencing depth; paralogous transcript;, etc. (Kumar and Blaxter 2010). Chapter 2 focuses on obtaining the 'best' transcriptome assembly from different assembly programs, MIRA (Chevreux, Wetter et al. 1999) and Newbler (Margulies, Egholm et al. 2005). CAP3 (Huang and Madan 1999) was used for an assembly of assemblies.

Computational resources

Annotation requires a bioinformatics infrastructure. In this stage of the project, I developed and configured an analysis pipeline to handle and process the nuclear genome sequence. This pipeline was largely based upon the program and annotation pipeline, MAKER (Cantarel, Korf et al. 2008). Protein and transcriptome data previously collected can be processed and aligned to the genome with this tool. Clustered EST transcripts are aligned via BLASTN (Altschul, Gish et al. 1990) while proteins are aligned by BLASTX. BLAST will only produce high scoring pair (HSP) alignments, so a second program, Exonerate (Slater and Birney 2005), is used to group HSPs corresponding to the same clustered EST or protein and then the ends of the HSP alignments are "polished" or moved to coincide with proper donor and acceptor sites.

As useful as transcriptome and conserved protein data are, there are some limitations. At any given time in the cell, only a small portion of genes are expressed, therefore sequenced

transcriptome data are not guaranteed to cover the entire breadth of genes. Secondly, any proteins specific to *S. neurona* will not be found by searching with orthologous proteins. These limitations necessitate computational gene finding approaches. Current gene finders primarily use complex probabilistic models such as hidden markov models (HMMs) to search for genes. HMMs however, are required to be "trained" by known genes in order to somewhat accurately find genes. In my pipeline, three gene finders are used, two of which are run by MAKER, Augustus and SNAP (Korf 2004). The third, GlimmerHMM, is run manually and is currently being trained.

Annotation

Although gene-finding programs are useful for discovering regions the genome where a gene is encoded, they often struggle with identification of the exact location of exon and intron boundaries. These inaccuracies require manual correction. Furthermore all lines of evidence for a gene need to be doubled-checked by hand in a process called annotation. In order to facilitate efficient annotation, the alignments along with 6-frame start and stop codon locations must be visualized. Once MAKER is finished, it outputs all alignment data and gene predictions in a tabular file format (GFF3) that can be visualized in a separate program called Apollo (Lewis, Searle et al. 2002). Chapter 3 focuses on application of this pipeline to annotate the apicoplast genome.

Database

Given the flood of data being produced through sequence, assembly and analysis projects, it is imperative that the data and analysis results be available to the end user. In order to facilitate dissemination of our data, SarcoDB (http://sarcodb.ctegd.uga.edu/) was built to

provide a user-friendly database for the sequence data generated by the *S. neurona* project. SarcoDB also contains tools to specifically facilitate access to unannotated sequence data. The genomic data can be queried based upon BLAST searches, protein motifs and text searches of key words found in the WUBLAST description headers of hits identified in searches of NCBI's non-redundant database. Additional tools, specifically genome views, are planned for in the future. These tools will not only aid in gene determination and annotation but will be an asset to the *S. neurona* scientific community as a whole. Chapter 4 focuses on the creation of SarcoDB and its tools.

CHAPTER 2

ASSEMBLY OF THE *SARCOCYSTIS NEURONA* TRANSCRIPTOME

Introduction

In this study, it was necessary to deal with the issue of choosing the 'best' transcriptome assembly from different assembly programs and an assembly of assemblies. In the context of this study, the 'best' assembly will be one that maps the best to the genome and represents more of the *Toxoplasma gondii* proteome and conserved orthologs shared by Apicomplexa. By assembling a more complete transcriptome, it is possible to achieve a more accurate annotation for *S. neurona*. Each assembly program works differently and presents particular strengths and weaknesses. In theory, one can assemble not only transcripts, but transcriptome assemblies from different programs to form more complete clustered EST transcripts (Kumar and Blaxter 2010). However, combination of output from multiple assembly programs into a single assembly introduces its own set of problems. Take, for instance, an alignment of reads at some position X. If it is imagined that one-fifth of the reads code for the nucleotide A at that position while four-fifths of the reads codes for the nucleotide T, then the program 1 would assembly those reads into separate clusters. When the assembly from program 1 and the program 2 assembly is assemblied together by program 3, it could align these two clusters. Program 3 looks at the alignment of the two reads, and at position X it is either an A or a T and

essentially with a roll of the dice, picks either A or T in the final assembly when in actuality that position most likely codes for an A. In this study I will compare assemblers to see which one performs better. I will also determine if the assembly of assemblies method can overcome potential drawbacks to produce the 'best' assembly.

Initially, it is necessary to choose from the many available sequence assembly programs. This choice is facilitated by the fact that many of the programs are designed for short-read (~35-100 bp) assembly and not the longer reads produced by 454 pyrosequencing technology (~400 bp) (Miller, Koren et al. 2010). The next limiting factor is the fact that not all assembly programs will accept reads from different technologies, 454 and Sanger in this case. Another discriminator to consider is whether the assembler can natively process the 454 sequencing output files, the Standard Flowgram Format (SFF) file. For each 454-sequencing run, an .SFF file is produced. It contains sequences, quality scores, flowgram information, and information on where to clip off the adapter sequences. A flowgram contains the signal intensity for each nucleotide during pyrosequencing. Of the available non-commercial assemblies that satisfy these criteria, we chose to look at MIRA and Newbler (Margulies, Egholm et al. 2005)(Margulies, Egholm et al. 2005)(Margulies, Egholm et al. 2005). CAP3 (Huang and Madan 1999) was chosen to assembly the assemblies. Key features of each assembler are summarized in Table 2.1 and discussed below.

Table 2.1: Summary of transcript assembly programs used
OLC: Overlap-Layout-Consensus.

Assembler	Type	Splits reads	Source available	Reference
CAP3	OLC	No	No	Huang et al. (1999)
MIRA	OLC	No	Yes	Chevreux et al. (1999)
Newbler 2.5	OLC	Yes	No	Margulies et al. (2005)

Note that all 3 assembly programs in Table 2.1 are of the Overlap-Layout-Consensus (OLC) type and none use the de Bruijn graph approach which is common for short read assemblers (Miller, Koren et al. 2010). The OLC method first pre-computes all K-mers across all reads and then groups reads that share K-mers. This is the overlap stage. In the layout stage, an overlap graph of the reads is made and redundant information is removed. Finally in the consensus stage a multiple sequence alignment is performed on the reads in the overlap graph producing a consensus of aligned reads.

Once the assemblies are made, the question arises as to how it is possible to gauge the quality of an assembly. The best assembler should have the highest percent of assembled sequences that map back to the *S. neurona* nuclear genome sequence with very high coverage and identity. This will show the quality of our assemblyIt should also produce a transcriptome with the highest overall coverage of a proteome from a related species. After the best transcriptome assembly algorithm and parameters are determined, it can be applied to all future transcriptome assemblies from this organism.

Methods

Library preparation

Total RNA for Sanger sequencing was isolated from the SN3 and SN4 strains (Table 2.2) from merozoites grown in monolayers of bovine turbinate (BT) host cells.

Table 2.2: Sequenced strains of *Sarcocystis neurona* in this study

Strain	Host	Location	Year	Reference
SN3	Horse	Panama	1992	(Granstrom, Alvarez et al. 1992)
SN4	Horse	California	1991	(Davis, Speer et al. 1991)

For SN4 merozoites, cDNA was synthesized using the Creator SMART cDNA library construction kit and cloned into pDNR-LIB vector. For SN3 merozoites, cDNA was synthesized by oligo d(T) priming and cloned into pBlueScript phagemid within the Uni-ZAP XR lambda vector. Sanger sequencing was performed at the Washington University Genome Sequencing Center. Library construction was performed in Dr. Dan Howe's laboratory at the University of Kentucky. All total RNA for 454 sequencing was isolated from SN3 strain merozoites grown in monolayers of bovine turbinate (BT) host cells. Total RNA was isolated with Trizol and enriched for Poly(A) using the NucleoTrap kit for all samples. For schizont samples, *S. neurona* merozoites were allowed to infect BT host cells. After 2, 8, 24, or 64 hours from the introduction of merozoites to BT host cells, total rRNA was extracted. All samples were prepared for sequencing by the GS FLX Titanium Sequencing Kit XLR70 and sequenced by the GS FLX System. Sequencing was performed by the University of Kentucky's Advanced Genetic Technologies Center on a Roche GS FLX genome sequencer.

Read preparation

All sequenced Sanger reads were edited to remove vector sequences using VectorNTI (http://www.invitrogen.com/site/us/en/home/LINNEA-Online-Guides/LINNEA-Communities/Vector-NTI-Community/vector-nti-software/what-is-vector-NTI.html). Previous test assemblies of 454 reads were screened for bovine contamination. Reads in clusters identified as bovine were removed from the dataset. Screening of Sanger reads was performed by Dr. Dan Howe's laboratory at the University of Kentucky. Screening of 454 reads was performed by the University of Kentucky's Advanced Genetic Technologies Center. A summary of the initially screened libraries can be seen in Table 2.3

Table 2.3: Statistics of *Sarcocystis neurona* transcriptome datasets

S. neurona lifestage	Sequence technology	# raw reads	# bases	# trimmed reads	# trimmed bases	Avg. len. of trimmed reads
SN3 merozoites	Sanger	8,056	3,903,242	8,003	3,881,942	485.06
SN3 schizonts	Sanger	1,601	897,848	1,532	877,614	572.86
SN4 merozoites	Sanger	6,413	2,620,359	6,332	2,572,532	406.27
SN3 merozoites	454 FLX Titanium	654,246	159,813,625	643,836	158,065,205	245.51
SN3 merozoites	454 FLX Titanium	105,636	43,147,599	104,915	42,934,666	409.23
SN3 schizonts 2hr	454 FLX Titanium	120,199	48,202,085	24,430	9,959,296	407.67
SN3 schizonts 8hr	454 FLX Titanium	90,838	35,810,134	13,714	5,361,866	390.98
SN3 schizonts 24hr	454 FLX Titanium	105,742	40,358,089	33,050	12,943,630	391.64
SN3 schizonts 64hr	454 FLX Titanium	120,343	46,871,470	90,075	35,398,947	392.99
Total		1,213,074	381,624,451	925,887	271,995,698	293.77

Preparation of the 454 sequence reads was more complicated. Like the Sanger cDNA dataset, the 454 cDNA dataset contained bovine contamination. But unlike the Sanger reads, the 454 reads contained adaptor sequences that had to be removed. Both MIRA and Newbler have options to remove vector sequences. However, each program searches for vector sequences differently. To solve this issue, an assembly performed by Newbler using the *Bos taurus* unigene file, build #98 (Nov. 10, 2010) was used as the vector file. The 454ReadStatus.txt output file can be used to determine which reads were used in the assembly and which were not (potential vector contamination sequences). Based on this information, only reads used by the Newbler assembly were input into the MIRA program. This was accomplished by changing the output of the 454 sequencing, .SFF files, to only contain reads used by the Newbler assembly. Because MIRA does not natively accept .SFF files they need to be converted into fasta and qual files using the program sff_extract (http://bioinf.comav.upv.es/sff_extract/) written by the author of MIRA. During extraction of the fasta files, adapter sequences are automatically clipped from the reads using information in the .SFF files. Newbler automatically removes adapters used in GS FLX Titanium Sequencing Kits. Poly-A tails are automatically clipped by Newbler and sff_extract. Via this processing, identical set of vector screened reads with removed adapter sequences and poly-A tails were input into MIRA and Newbler.

Transcriptome assemblies

The data described above were input to MIRA 3.0 with the following parameters: denovo, est, 454, sanger and accurate. This tells MIRA to run a *de novo* transcriptome assembly using 454 and Sanger reads in accurate (most stringent) mode. MIRA allows custom

parameters. Because quality scores were not available for the Sanger reads, the wants_quality_file, quality_clip, and enforce_presence_of_qualities parameters were set to "no". All other parameters were set to default. For 454 reads, default settings were used. Since poly-A tails were removed in pre-processing, the clip_polyat parameter was set to "no". Finally, the number_of_threads parameter was set to "3" to allow for parallel processing of the program. Newbler assembly was performed using Newbler 2.5 in cDNA mode. The *Bos taurus* cDNA unigene file, build #98 (Nov. 10, 2010) was used as the vector file. Due to concerns about low coverage of the transcriptome the –urt option was turned on. Normally, Newbler constructs clusters from a consensus sequence of 2 or more reads, ignoring any single read ends. However with the –urt option active, clusters can be extended by the ends of single reads. All other options were set to default. CAP3 was used for the assembly of assemblies. Clusters from the MIRA and Newbler assemblies were compiled and input into the program CAP3. Note that the compiled file includes MIRA singletons. However, these are not singletons in the traditional sense (reads that were not clustered), but are single reads that did not overlap and were heavily filtered (reads that are too short, thought to be chimeric, or low quality values) by the MIRA program. The CAP3 assembly was performed using default parameters. MIRA and Newbler assemblies were run on a 96GB RAM quad-core Xeon. CAP3 was run on a 16GB RAM quad-core Xeon. The MIRA and Newbler assemblies would not successfully complete on the 16GB RAM quad-core Xeon machine.

Transcriptome analysis

The various assemblies were aligned to the *S. neurona* nuclear genome using BLAT version 0.34 (Kent 2002) with default parameters. The assemblies were aligned to protein

sequences using NCBI BLAST version 2.2.20. TBLASTN was performed with the various transcriptome assemblies as the database. All BLAST jobs had an expect value cutoff of 1e-5. The *T. gondii* proteome was downloaded from ToxoDB release 6.3 for strain ME49. The set of apicomplexan orthologous genes used was generated by the Kissinger laboratory (Kuo, Wares et al. 2008). For each orthologous gene cluster that had genes from all apicomplexan species, the *T. gondii* sequence was extracted from the cluster and used in subsequent analyses.

Results

The metrics of an assembly are often used to evaluate its quality. However, metrics alone may not accurately predict the best assembly (Bradnam 2011). Based solely on statistics, MIRA appears to be the superior assembler with a higher N50 score and longer cluster length (Table 2.4). N50 represents the length of the shortest cluster where the cumulative length of clusters greater in size is at least 50% of the total length of all clusters. This metric is loosely related to the completeness of a genome assembly. With a more complete assembly, a large majority of sequence are in larger scaffolds giving a higher N50 count but if the assembly is fragmented then a large proportion of the genome is in small scaffolds giving a lower N50 count. However, the picture becomes murkier as other metrics are considered. Ideally, an assembler would use as many reads as possible for the assembly and in this case Newbler performs best. Another means of gauging the assembly is assessment of the cumulative length of the clusters (Figure 2.1). This metric is useful to visualize the distribution of cluster lengths in an assembly.

Table 2.4: Clustered EST assembly metrics

	CAP3	MIRA	Newbler 2.5
Number of clusters	24,962.00	40,564.00	30,288.00
Total Bases	19,150,573.00	31,727,233.00	17,791,495.00
Number of large clusters (>=500 bp)	12,162.00	8,918.00	11,472.00
Max cluster length	18,050.00	11,628.00	18,044.00
N50	907.00	1,328.00	745.00
Reads used	72,849.00	582,717.00	875,254.00
Average cluster length	767.19	782.15	587.41

The cumulative cluster length is calculated by adding the lengths of the largest clusters first in descending order and plotting them against the number of clusters (Figure 2.1). As you can see both the Newbler and MIRA assemblies follow similar distributions of cluster lengths through the beginning and middle section of the graph. Towards the end of the graph, MIRA's assembly begins to have a larger cumulative cluster length indicating that the MIRA assembly contains many more small clusters. The CAP3 assembly consistently maintains a higher cumulative cluster length throughout the distribution of clusters. This suggests that there have been extensions to the transcripts by performing the assembly of assemblies while the number of small clusters was consistent with the other assemblies.

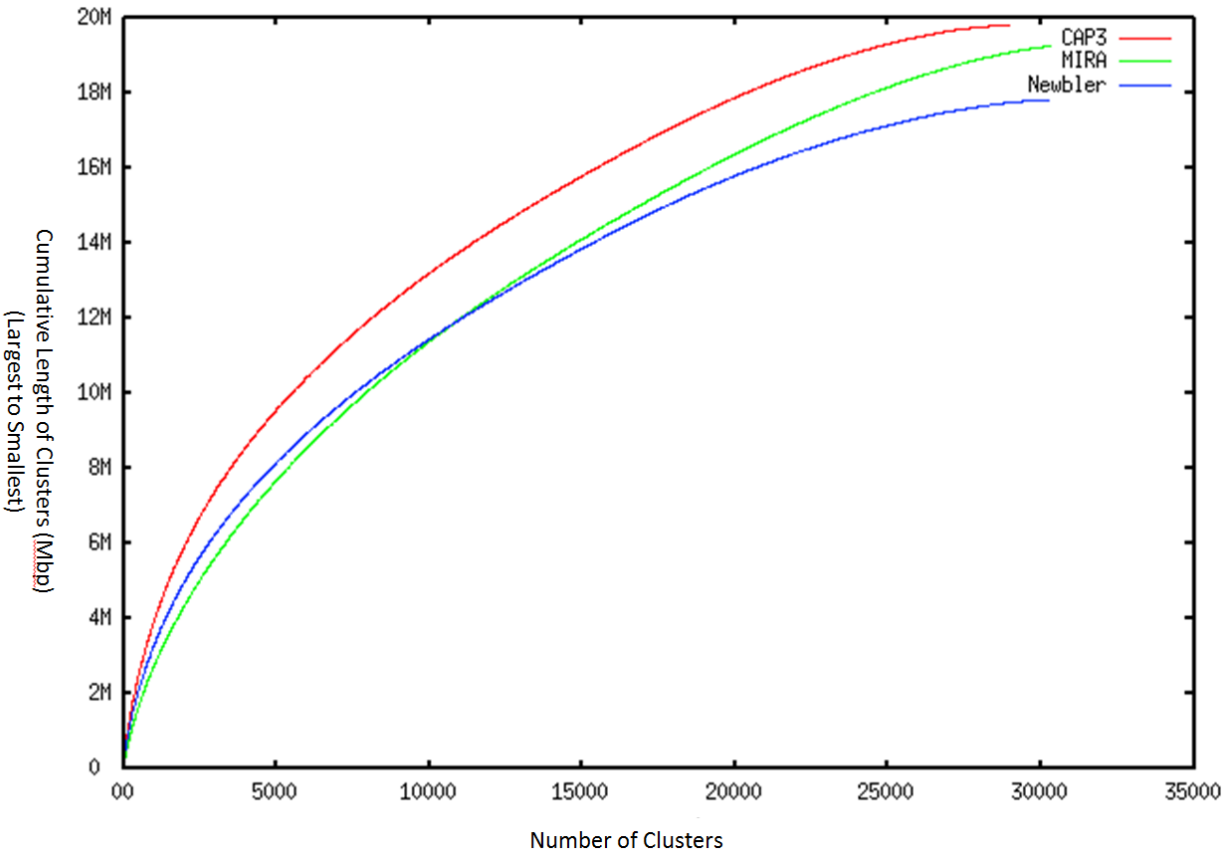

Figure 2.1: Cumulative cluster lengths of different assemblies

Clusters were ordered by length and the cumulative length of all clusters was plotted. Initial slope of the curve shows proportion of longer cluster while the slope at the end shows the proportion of shorter clusters. The Y-axis represents the cumulative size of all clusters from largest to smallest. The X-axis represents the number of clusters.

Table 2.5 shows the results of aligning the EST clusters to the *S. neurona* genome with the program BLAT (Kent 2002). The Newbler assembly clearly outperformed MIRA's assembly but did not perform as well as the CAP3 assembly for percent of clusters that hit to the genome. Lastly, Newbler shows a higher percent of each cluster's bases mapping to the genome. The next analysis was to align the clustered ESTs to *T. gondii*'s proteome (Table 2.6). Again Newbler outperformed MIRA, and in this CAP3 as well in all metrics.

Table 2.5: EST BLAT hits to *Sarcocystis neurona* genome

	CAP3	MIRA	Newbler
% of clusters hit	98.02	94.78	95.73
% of clusters' bases covered	97.75	96.27	98.89
% of clusters hit with ≥ 80% coverage and 92% identity	93.58	86.48	92.54
% of clusters' bases covered	98.89	98.59	99.47

Table 2.6: TBLASTN hits of EST clusters to 7,993 *Toxoplasma gondii* proteins

	CAP3	MIRA	Newbler
No. of proteins hit	3,570.00	3,512.00	3,604.00
No. of HSP hits with ≥80% coverage	401.00	314.00	411.00
% of bases covered	94.29	93.67	94.45

HSP: High Scoring Pairs. E-value cutoff of 1e-5.

Table 2.7: TBLASTN hits of EST clusters to 1,088 *Toxoplasma gondii* conserved orthologs

	CAP3	MIRA	Newbler
No. of orthologs hit	926.00	909.00	925.00
No. of HSP hits with ≥80% coverage	158.00	117.00	166.00
% of bases covered	94.21	93.23	94.06

HSP: High Scoring Pairs. E-value cutoff of 1e-5.

Finally, the different assemblies were aligned to the set of orthologous proteins (Table 2.7). Similar to the alignment of the assemblies with the *T. gondii* proteome, Newbler outperformed MIRA. Newbler and CAP3 produce similar numbers of total orthologs that were hit by their respective assemblies but the Newbler assembly had more HSPs with ≥80% coverage.

Discussion

Three different *de novo* transcriptome assembly algorithms were tested. Each assembler has its own strengths and weaknesses. Newbler is very straightforward to use, natively supports .SFF files, and takes advantage of flowgram information for more accurate basecalling. Newbler can screen for vector sequences very quickly compared to BLAST-based programs such as SeqClean, but proved less than 100% accurate as some individual bovine clustered ESTs were found. MIRA, while more complicated to use, has many more options and parameters for an advanced user to customize. In addition, it allows for specific parameters to be individually applied for each type of read technology. It is also feasible for some algorithms to work better with certain types of data, therefore users should cross compare assemblers for new datasets.

Every sequencing technology contains limitations and biases. For 454 one major limitation is homopolymer runs. Newbler addresses this source of error by analyzing the flowgrams for each read, but during the conversion of .SFF files to fasta and qual files, flowgram information is lost. There were 17 instances where the Newbler assembly had clusters with > 80% coverage to the orthologous protein dataset and the CAP3 assembly did not. There were 9 instances of the inverse. When I investigated those 9 instances with a multiple sequence alignment of each cluster from each assembly, the CAP3 assembly had clusters that were longer, however when this procedure was repeated for the group of 17 instances, there were basecalling error within homopolymer runs. MIRA, using only qual scores, is at a disadvantage when assembling. This led to misassemblies that manifested themselves in EST clusters with

frameshifts. Without flowgram information to make more accurate base calls, the MIRA and CAP3 assemblies are somewhat hobbled.

Interestingly, only 926 of 1,088 orthologs from *T. gondii* were hit by the CAP3 *S. neurona* EST clusters. The assembly of assemblies had little added benefit in creating additional clustered ESTs that mapped to the conserved orthologs dataset. Furthermore, only ~half of *T. gondii's* predicted proteins were hit by the *S. neurona* transcriptome. This suggests that *S. neurona* has evolved significantly from *T. gondii* and lost these 162 conserved orthologous proteins or, that the transcriptome is incomplete. It is likely that the transcriptome is incomplete as only merozoite and schizont cDNA was produced leaving other S*. neurona* lifecycles un-sampled.

CHAPTER 3

THE APICOPLAST GENOME OF *SARCOCYSTIS NEURONA*

Introduction

Sarcocystis neurona belongs to the parasitic protistan phylum Apicomplexa. *Sarcocystis* and other members of Apicomplexa have medical, veterinary and economic importance as they cause a wide array of disease that affects humans, livestock, wild animals and invertebrates. Some prominent members of Apicomplexa are *Plasmodium falciparum, Toxoplasma gondii, Eimeria tenella, Babesia bovis* and *Theileria parva*. Most apicomplexans have an organelle which is non-photosynthetic but derived from chloroplasts, named the apicoplast. Notable exceptions are the genus C*ryptosporidium* (Zhu, Marchewka et al. 2000), *Gregarina* (Toso and Omoto 2007), and C*olpodella* (Singh, Alam et al. 2010). This chapter will focus on the assembly and annotation of *S. neurona*'s apicoplast genome. Because the apicoplast is a potent drug target, by sequencing and annotation the apicoplast genome we acquired a greater understanding of the apicoplast and with the fully sequenced apicoplast genome of *S. neurona* we have a greater understanding of not only *S. neurona* but of other coccidans.

Methods

Cloning, sequencing, and assembly

Bovine turbinate cell monolayers were infected with *S. neurona* (SN3 strain) merozoites in the laboratory of Dr. Dan Howe at the University of Kentucky. Egressed merozoites were then harvested. DNA libraries were then prepared for sequencing using GS FLX Titanium sequencing kit XLR70. Paired-end libraries of length 3 Kbp and 8 Kbp were generated using the GS FLX titanium rapid library preparation kit, and GS FLX titanium library paired end adaptor kit. Sequencing was performed by the University of Kentucky's Advance Genetic Technologies Center on a Roche GS FLX genome sequencer.

Sequence assembly was performed using Newbler 2.5.3 with the –large parameter turned off. *S. neurona* is cultured in bovine turbinate cell therefore, a vector screening file consisting of the UMD 3.1 assembly of the *Bos taurus* reference genome was used to screen out any contaminates. All other parameters were set to default. The assembly produced the whole genome for *S. neurona.* In order to find the apicoplast genome, candidates were identified by screening for scaffolds that were approximately 25-35 Kbp in length and contained over 20 tRNAs as identified by tRNAscan-SE (Lowe and Eddy 1997). We screened for a range of sizes due to the effect of the IR during assembly as seen with the *T. gondii* apicoplast genome assembly where the IR had collapsed in half on top of itself. The results of this screening produced one scaffold of size 34488 that would be identified as the apicoplast genome. The scaffold was made of two contigs one of length 5,780 bp and the other 24,002 bp joined by a paired-end read creating a gap of 4,706 bp.

In order to complete the assembly we had to identify the contigs. By comparing the two contigs against *Toxoplasma gondii's* apicoplast genome with BLAST it was evident that the smaller contig was part of the IR but at roughly half the size. As hypothesized the IR had collapsed during assembly. The remaining contig was identified as the other section of the apicoplast genome. The average coverage of the smaller contig (502.8X) was more than twice the coverage of the larger contig (192.5X) indicating the IR had indeed collapsed during assembly. Even though the IR had collapsed into a separate contig from the rest of the apicoplast genome, small pieces on the ends of the larger contig could be aligned to the IR. The 5' end of the larger contig contained the end of LSUrRNA, while the 3' end of the larger contig contained a portion of the beginning of the LSUrRNA. Using this information we were able to fill in the gaps of the apicoplast genome. However, due to the sequence similarity of the IR we were not able to split it apart in the assembly. To complete the assembly, the reverse complement of the IR was inserted.

Annotation

Sequence analysis and annotation was performed using programs Maker (Cantarel, Korf et al. 2008) and Apollo (Lewis, Searle et al. 2002), tRNAscan-SE (Lowe and Eddy 1997) and ARAGORN (Laslett and Canback 2004). MAKER was used to scan the apicoplast genome against all apicoplast-encoded protein sequences from *P. falciparum, E. tenella* and *T. gondii*. Proteins were extracted for *P. falciparum* and *T. gondii* from PlasmoDB and ToxoDB respectively (Gajria, Bahl et al. 2008; Aurrecoechea, Brestelli et al. 2009). *Eimeria tenella* proteins were extracted from the GenBank record NC_004823.1. Output from MAKER was loaded into Apollo where coding regions were visualized and aligned to UAA or UAG stop codons and ATG start codons.

Apollo however did not allow for alternative codon usage and all UGA codons were considered as stop codons. Hypothetical protein sequences had to be manually corrected for this error after the sequences were output by Apollo. SSUrRNA and LSUrRNA were found by comparisons to *T. gondii* via BLASTN. The tRNAs were found by tRNAscan-SE server using default parameters for a Mito/Chloroplast source (http://lowelab.ucsc.edu/tRNAscan-SE/). The tRNA-Leu in the region after *rps4 was* found by tRNAscan-SE with search mode: Cove only. tRNAs were further confirmed by the program ARAGORN.

Confirmation of *rps4* fragment insert

Due to the possibility that the *rps4* insert could be an error in the assembly, the region of the insert was amplified and sequenced. Confirmation of the *rps4* fragment insert was performed by Dan Howe's Lab at the University of Kentucky. Two primer pairs were designed to amplify the 238 bp *rps4* fragment located between the ORF-G and LSUrRNA genes of the *S. neurona* apicoplast genome. The two forward primers, F1 (5'-TGCTGGAATTGTATTCCT-3') and F2 (5'-TCCTCCTACTAATCTAATAG-3') were located on ORF-G. One of the reverse primers, R1 (5'-ACGGTATTTTAATCTTAC-3') was located on the *rps4* fragment, while the other reverse primer R3 (5'-GAGCCACTGATTTGTAATCAG-3') was located on the LSUrRNA gene. A summary of the primer sequences is located in Table 3.1. The amplification schematic provided in Figure 3.1

Table 3.1: Primers used in PCR amplification of *rps4* fragment insert

Primer name	Sequence	Location	Primer combination	Expected product size
Forward 1	5′-TGCTGGAATTGTATTCCT-3′	29,198-29,215	F1-R1	510 bp
Forward 2	5′-TCCTCCTACTAATCTAATAG-3′	29,431-29,450	F2-R1	277 bp
Reverse 1	5′-ACGGTATTTTAATCTTAC-3′	29,691-29,708	F1-R3	635 bp
Reverse 3	5′-GAGCCACTGATTTGTAATCAG-3′	29,813-29,833	F2-R3	402 bp

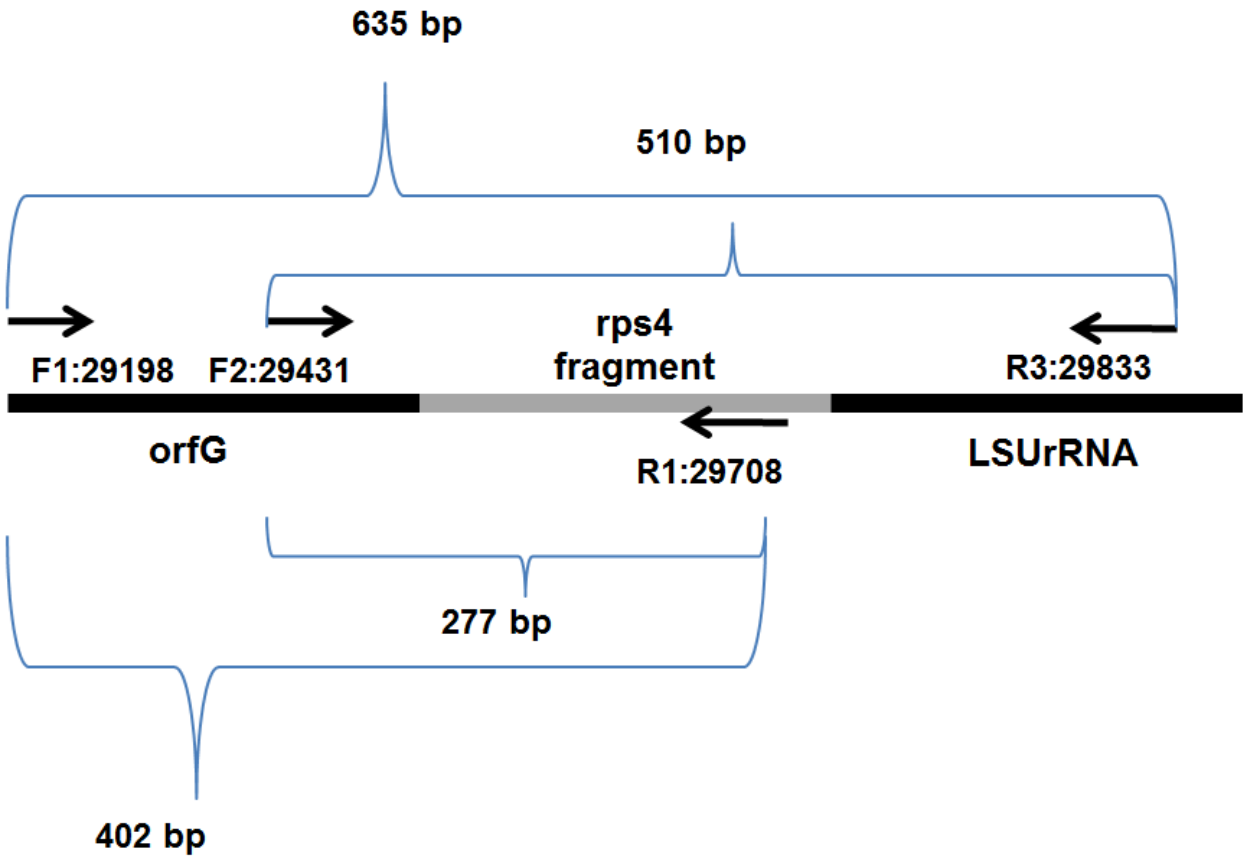

Figure 3.1: Amplification scheme to verify the *rps4* insert fragment

Polymerase chain reaction (PCR) was performed using Verbatim high-fidelity DNA polymerase (Thermo Fisher Scientific Inc., Pittsburgh, PA). Three 25 μl PCR reactions (*S. neurona* genomic DNA; *S. neurona* apicoplast DNA; or no DNA) were set up for each of the primer pair combinations. The cycling conditions included an initial denaturation at 95^{o}C for 3 min, 35 cycles of denaturation at 95^{o}C for 30 sec, annealing as shown in the table and extension at 68 ^{o}C for 1 min, followed by a final extension at 68 ^{o}C for 2 min. The amplified PCR products were analyzed on 1.5% agarose gels. The 635 bp product amplified using primer pair F1-R3 was purified using a PCR purification kit (Qiagen, Valencia, CA) and was sequenced in both directions at Advanced Genetic Technologies Center, University of Kentucky.

Results

The apicoplast sequence of 35,004 Kbp was determined along with gene content (Figure 3.2).

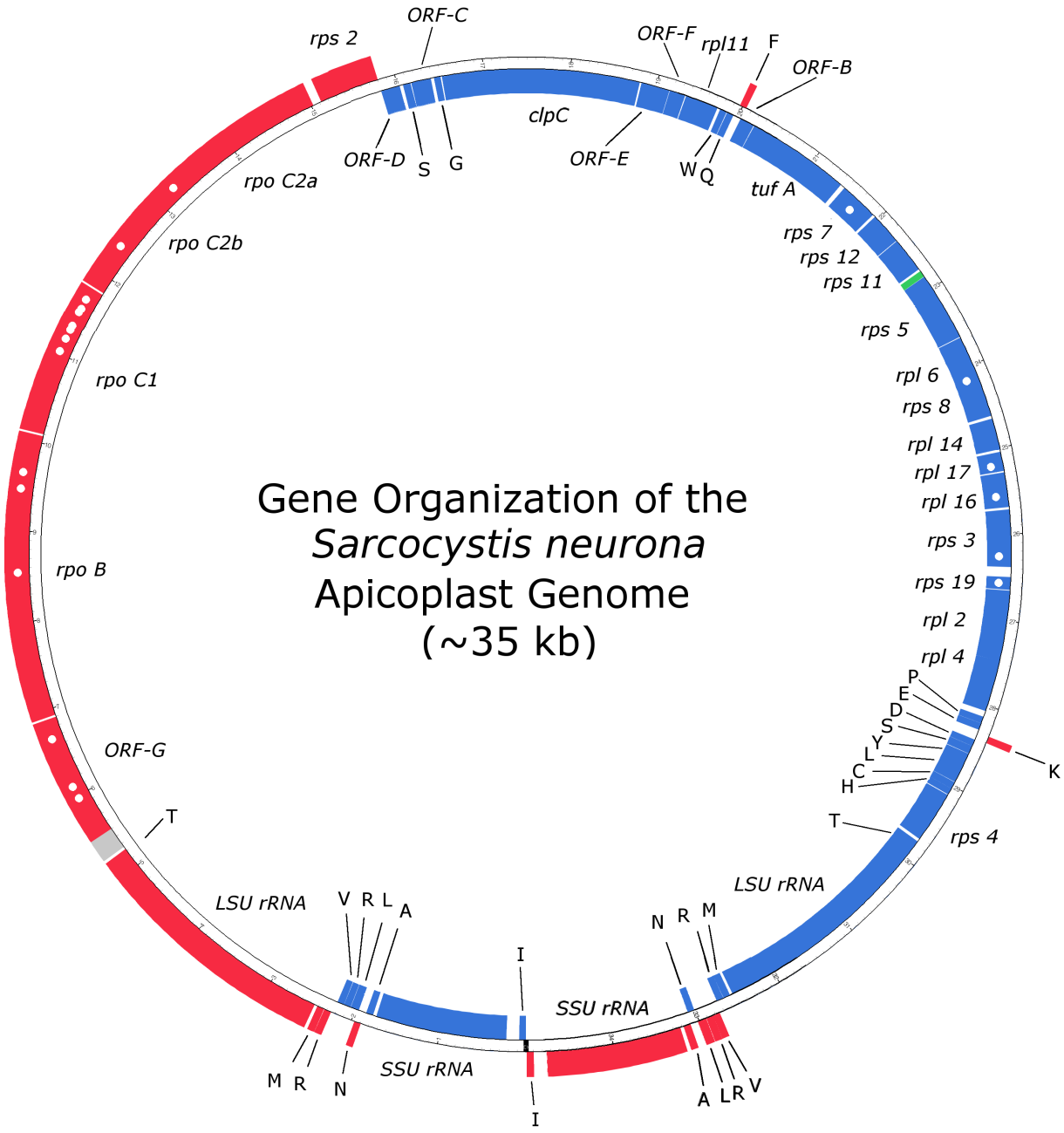

Figure 3.2: The annotated apicoplast genome of *Sarcocystis neurona*

The center of the inverted repeat is located at the bottom of the circle denoted with a thick black line. Red and blue indicate differing gene orientations. Green signifies the location of the missing *rpl36* gene in other species. Gray depicts the location of the *rps4* fragment insert. In-frame UGA codons are shown as white circles.

The apicoplast genomes size is similar to other species, *T. gondii* (34,996 bp) and *E. tenella* (34,750 bp) (Cai, Fuller et al. 2003). The apicoplast genome is highly AT-rich (78%) as in other apicomplexan species. Gene content and organization is very similar to *T. gondii* and *E. tenella* (Table 3.2).

Table 3.2: Comparison of apicoplast genome gene content

Chart of the gene content in the apicoplast genomes of *S. neurona*, *T. gondii* and *P. falciparum.* "o" denotes the presence of the gene in the genome while "x" means it is not found. * Present in the genome twice, once complete and once as a 239 bp insert. ** Hypothetically expressed as two proteins

Gene	*S. neurona*	*T. gondii*	*P. falciparum*
SSUrRNA	o	o	o
LSUrRNA	o	o	o
rps4	o*	o	o
rpl4	o	o	o
rpl23	**x**	**x**	o
rpl2	o	o	o
rps19	o	o	o
rps3	o	o	o
rpl16	o	o	o
rps17	o	o	o
rpl14	o	o	o
rps8	o	o	o
rpl6	o	o	o
rps5	o	o	o
ORFA	**x**	**x**	o
rpl36	**x**	o	o
rps11	o	o	o
rps12	o	o	o
rps7	o	o	o
tufA	o	o	o
ORFB	o	o	o
rpl11	o	o	o
ORFF	o	o	o

ORFE	o	o	o
clpC	o	o	o
ORFC	o	o	o
ORFD	o	o	o
rps2	o	o	o
rpoC2	o**	o	o
rpoC1	o	o	o
rpoB	o	o	o
ORFG	o	o	o

Since the apicoplast genome is so AT-rich, there is a clear bias in the third position of codons for A or T (Figure 3.3). The first and third row of every second nucleotide block corresponds to A or T at the third nucleotide position. Looking at the first and third row you can see biases as high as 95% for some codons. However when looking at the relationship between apicoplast encoded tRNAs and the most frequently utilized codons many of the most frequently utilized codons are not found. The codon utilization for *T. gondii* is identical. This suggests that the apicoplast must import these other nuclear-encoded tRNAs.

			-|	.t.		|	.c.		|	.a.		|	.g.		|
t ALL:	9025	41%	t.. |	Phe	613	91% |	Ser	200	43% |	Tyr	473	95% |	Cys	69	93% | ..t
t 1st:	2501	34%	t.. |	Phe	58	9% |	Ser	8	2% |	Tyr	27	5% |	Cys	5	7% | ..c
t 2nd:	2848	38%	t.. |	Leu	809	82% |	Ser	131	28% |	OCH	24	ALL |	Trp	36	71% | ..a
t 3rd:	3676	50%	t.. |	Leu	28	3% |	Ser	3	1% |	AMB	2	ALL |	Trp	15	29% | ..g
			-|			|			|			|			|
c ALL:	1817	8%	c.. |	Leu	89	9% |	Pro	110	70% |	His	73	96% |	Arg	23	13% | ..t
c 1st:	629	8%	c.. |	Leu	0	0% |	Pro	2	1% |	His	3	4% |	Arg	5	3% | ..c
c 2nd:	983	13%	c.. |	Leu	63	6% |	Pro	45	29% |	Gln	188	93% |	Arg	10	5% | ..a
c 3rd:	205	3%	c.. |	Leu	0	0% |	Pro	0	0% |	Gln	15	7% |	Arg	3	2% | ..g
			-|			|			|			|			|
a ALL:	9471	43%	a.. |	Ile	620	70% |	Thr	167	51% |	Asn	669	94% |	Ser	116	25% | ..t
a 1st:	3277	44%	a.. |	Ile	30	3% |	Thr	2	1% |	Asn	41	6% |	Ser	7	2% | ..c
a 2nd:	2886	39%	a.. |	Ile	232	26% |	Thr	151	46% |	Lys	983	97% |	Arg	141	77% | ..a
a 3rd:	3308	45%	a.. |	Met	78	ALL |	Thr	5	2% |	Lys	34	3% |	Arg	1	1% | ..g
			-|			|			|			|			|
g ALL:	1962	9%	g.. |	Val	114	50% |	Ala	108	68% |	Asp	142	96% |	Gly	90	32% | ..t
g 1st:	1018	14%	g.. |	Val	3	1% |	Ala	3	2% |	Asp	6	4% |	Gly	5	2% | ..c
g 2nd:	708	10%	g.. |	Val	103	45% |	Ala	47	30% |	Glu	190	92% |	Gly	155	56% | ..a
g 3rd:	236	3%	g.. |	Val	8	4% |	Ala	1	1% |	Glu	16	8% |	Gly	27	10% | ..g
			-|			|			|			|			|

Figure 3.3: Codon usage of *Sarcocystis neurona* apicoplast genes
Codons highlighted in yellow represent codons encoded by tRNA genes found in the apicoplast genome.

As found in other *Coccidia* and *Plasmodia*, there is an inverted repeat composed of rRNA and tRNA genes (Sato 2011). Within the coding sequence of several genes, there are 1 or more UGA codons (Figure 3.2, open circle) which are predicted to be translated into tryptophan (Wilson 2002). This pattern of codon reassignment for the UGA codon was initially found in the AT-rich genome of *Mycoplasma capricolum* (Oba, Andachi et al. 1991).

Despite these similarities, there are some differences present. In between the genes *ORFG* and LSUrRNA, there is an insert corresponding to the first 240 bp of the *rps4* protein coding sequence. The full *rps4* gene of *S. neurona* is 600 bp in length. Note the whole *rps4 gene* flanks the other LSUrRNA on the other side of the IR. In order to rule out possible misassembly issues, individual reads for this region were manually inspected. They had an average coverage of 116X and verified the observation. As final proof, the region was amplified by PCR and sequenced to confirm the presence of the insert by Dr. Dan Howe's Lab at the University of Kentucky (Figure **3.4**).

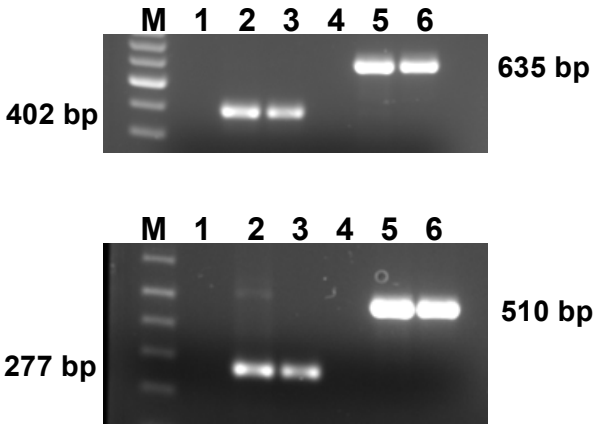

Figure 3.4: Electrophoretic analysis of *rps4* fragment insert
Lane M: GeneRuler 100 bp ladder. Lane 1 and 4: Negative control. Lane 2 and 5: *S. neurona* apicoplast DNA as template. Lane 3 and 6. *S. neurona* total genomic DNA prep.

PCR product sequence confirms the presence the *rps4* fragment insert. Multiple sequence alignment of the PCR products, the insert and the original gene show no mutations of the *rps4* gene fragment insert from the original gene suggesting it is a relatively recent event (Figure 3.5).

```
PCR_rps4_insert1-LSUR3      -------GGG---------------GAGGGTCAT-----------CCTTT 17
PCR_rps4_insert2-ORFG       TTTTTGAGAGCCCATTGTTTAATCAGAAGGTTATGGGTTCAAATCCCTTT 50
Apicoplast_rps4_insert      --------------------------------------------------
Apicoplast_rps4             --------------------------------------------------


PCR_rps4_insert1-LSUR3      AT-AGTTGATTAA-TTTTATCAAATAGATCTAATTTTTACTAATAATGGG 65
PCR_rps4_insert2-ORFG       ATCAGTTGATTAAATTTTATCAAATAGATCTAATTTTTACTAATAATGGG 100
Apicoplast_rps4_insert      ----------------------------------------------ATGGG 5
Apicoplast_rps4             ----------------------------------------------ATGGG 5
                                                                           *****

PCR_rps4_insert1-LSUR3      AAAATATTTAGGTGCTAAACTTAAAAAATTACGGTATTTTAATCTTACTT 115
PCR_rps4_insert2-ORFG       AAAATATTTAGGTGCTAAACTTAAAAAATTACGGTATTTTAATCTTACTT 150
Apicoplast_rps4_insert      AAAATATTTAGGTGCTAAACTTAAAAAATTACGGTATTTTAATCTTACTT 55
Apicoplast_rps4             AAAATATTTAGGTGCTAAACTTAAAAAATTACGGTATTTTAATCTTACTT 55
                            **************************************************

PCR_rps4_insert1-LSUR3      TTTTATCTGGATTTTCTACTAAATTATTAAAAAAAGAGGCTTGTTTGATA 165
PCR_rps4_insert2-ORFG       TTTTATCTGGATTTTCTACTAAATTATTAAAAAAAGAGGCTTGTTTGATA 200
Apicoplast_rps4_insert      TTTTATCTGGATTTTCTACTAAATTATTAAAAAAAGAGGCTTGTTTGATA 105
Apicoplast_rps4             TTTTATCTGGATTTTCTACTAAATTATTAAAAAAAGAGGCTTGTTTGATA 105
                            **************************************************

PCR_rps4_insert1-LSUR3      AAAAAAAAAAGTAAAATTTCTGTCTTTCTATCTAAATTATTAGAAAAACA 215
PCR_rps4_insert2-ORFG       AAAAAAAAAAGTAAAATTTCTGTCTTTCTATCTAAATTATTAGAAAAACA 250
```

```
Apicoplast_rps4_insert      AAAAAAAAAAGTAAAATTTCTGTCTTTCTATCTAAATTATTAGAAAAACA 155
Apicoplast_rps4             AAAAAAAAAAGTAAAATTTCTGTCTTTCTATCTAAATTATTAGAAAAACA 155
                            **************************************************

PCR_rps4_insert1-LSUR3      AAAATTAAAATATAATTATGGACTTAAAGAAAATCAAATTAAAAAATATT 265
PCR_rps4_insert2-ORFG       AAAATTAAAATATAATTATGGACTTAAAGAAAATCAAATTAAAAAATATT 300
Apicoplast_rps4_insert      AAAATTAAAATATAATTATGGACTTAAAGAAAATCAAATTAAAAAATATT 205
Apicoplast_rps4             AAAATTAAAATATAATTATGGACTTAAAGAAAATCAAATTAAAAAATATT 205
                            **************************************************

PCR_rps4_insert1-LSUR3      TAAAATATATTAAAATTTTAAAAATATTTAATTTAGAATGAAACTATATA 315
PCR_rps4_insert2-ORFG       TAAAATATATTAAAATTTTAAAAATATTTAATTTAGAATGAAACTATATA 350
Apicoplast_rps4_insert      TAAAATATATTAAAATTTTAAAAATATTTAATTTA--------------- 240
Apicoplast_rps4             TAAAATATATTAAAATTTTAAAAATATTTAATTTAGT-TCAAATTATTGA 254
                            **********************************

PCR_rps4_insert1-LSUR3      AATA--TTTATATTC---TTTAC--GAAATA-----------ATCAATTA 347
PCR_rps4_insert2-ORFG       AATA--TTTATATTC---TTTAC--GAAATA-----------ATCAATTA 382
Apicoplast_rps4_insert      ---------------------------------------------------
Apicoplast_rps4             ATTACGTTTAGATGCAAGTTTATTTAGAATAGGATTTAGTAGATCTATTA 304

PCR_rps4_insert1-LSUR3      ATTAATACTATTAGATTA------GTAGGAGGATTT------AATTTAAA 385
PCR_rps4_insert2-ORFG       ATTAATACTATTAGATTA------GTAGGAGGATTT------AATTTAAA 420
Apicoplast_rps4_insert      ---------------------------------------------------
Apicoplast_rps4             ATCAAGCTAAACAATTTATCAACCATGGACATATTTTTGTTAATTTTAAA 354

PCR_rps4_insert1-LSUR3      CACAATAAA----TAAAT-TAATTTTTAAACAAGATAATTTTATTT-TCT 429
PCR_rps4_insert2-ORFG       CACAATAAA----TAAAT-TAATTTTTAAACAAGATAATTTTATTT-TCT 464
Apicoplast_rps4_insert      ---------------------------------------------------
Apicoplast_rps4             CTAGTTAAAAATCCAAATATGCTTATTACAGAAAAAGATTTAATTTATAT 404

PCR_rps4_insert1-LSUR3      TATATATTTATAGATTAAATGCTTTATCTTATTTAAAAAAATTTAAACAA 479
PCR_rps4_insert2-ORFG       TATATATTTATAGATTAAATGCTTTATCTTATTTAAAAAAATTTAAACAA 514
Apicoplast_rps4_insert      ---------------------------------------------------
Apicoplast_rps4             TA-ATCCACAAAAAGTAAGTA--TTATTTTAATTTGTAGAATT-AATTTA 450

PCR_rps4_insert1-LSUR3      CCTGATTGGTGTTTT------TTTGAATTATCTGAATTTGCTTTTGATG- 522
PCR_rps4_insert2-ORFG       CCTGATTGGTGTTTT------TTTGAATTATCTGAATTTGCTTTTGATG- 557
Apicoplast_rps4_insert      ---------------------------------------------------
Apicoplast_rps4             TTTTATAGATATTATAAGAAATTTAATTTTTCTTTATATAATATTTGTGC 500

PCR_rps4_insert1-LSUR3      ---ATATTTCATATT---ATTCAA----------TACCTATAAATCT-TT 555
PCR_rps4_insert2-ORFG       ---ATATTTCATATT---ATTCAA----------TACCTATAAATCT-TC 590
Apicoplast_rps4_insert      ---------------------------------------------------
Apicoplast_rps4             AGAATTTTTAAAATTTAAATTAAAAAAAGAATTTTATTTTAAAATTTATT 550

PCR_rps4_insert1-LSUR3      CTTTTACTAAATATAAAAAATTTTTATTTAATTTAGGAAAAATTTTTCAG 605
PCR_rps4_insert2-ORFG       --TTTACTAAATAAAGAAA----------------GG------------- 609
Apicoplast_rps4_insert      ---------------------------------------------------
Apicoplast_rps4             TTTTTAATTCATTTAAAGAATCTT----------ATATTAAATTATTTTA 590

PCR_rps4_insert1-LSUR3      CCAAA 610
PCR_rps4_insert2-ORFG       -----
Apicoplast_rps4_insert      -----
Apicoplast_rps4             T---- 591
```

Figure 3.5: Alignment of the *rps4* gene and the *rps4* fragment insert

The multiple sequence alignment is composed of 4 sequences: *rps4*, *rps4* fragment insert, and the 635 bp product of PCR amplification sequence both ways. Highlighted in yellow is the alignment of the *rps4* insert. Highlighted in blue is ambiguity in the length of the homopolymer run in the *rps4* insert

The region from 5,950 bp to 6,845 bp contains several tightly packed tRNAs. The order and identity of these tRNAs are identical in *T. gondii* and *E. tenella*. *Plasmodium falciparum* has one exception, a strand flip of tRNA-K. *Sarcocystis neurona* has an identical gene order for this region with *T. gondii* except for the omission of the tRNA-Met. However, the missing tRNA does not mean the apicoplast genome is without a tRNA-Met. Within the IR there are two additional tRNA-MET genes. Therefore, the loss of one tRNA-Met should not be detrimental to the fitness of *S. neurona* as the apicoplast genome still contains a full array of all 20 tRNAs. In this same region, like *P. falciparum* and *T. gondii* but not *E. tenella* one of the tRNA-Leu (UAA) genes contains a group-I type self-splicing intron. This type has been observed in cyanobacteria and plastids (Kuhsel, Strickland et al. 1990). The *S. neurona* apicoplast genome is also missing the *rpl36* gene, however a gap nearly the same size as *rpl36* exists where *rpl36* is present in other apicoplast genomes. This gap does not contain an appropriate ORF and is filled with UAA stop codons on both strands. A translated BLASTX search of the gap against nr with the e-value cutoff raised to 10,000 does not return any significant hits. The *rpl36* is not found in *S. neurona* nuclear genome.

Another gene that differs slightly is the *rpoC2* gene. In *T. gondii,* the *rpoC2* gene is composed of an ORF followed by a one in-frame UAA codon and UAG codon followed by an ORF of the same frame without a start methionine. In *P. falciparum* a frame shift must occur to produce the correct translation. The relative location of the in-frame stop codons and frame

shift in *P. falciparum* are all in the same region of poorly conserved sequence (Sato 2011). The protein coding sequence of the ORFs is similar to other *rpoC2* genes. In *S. neurona* however, the second ORF contains a start methionine. The gene *rpoC2* is split into two parts in *Theliera parva* and *Babesia bovis (Gardner, Bishop et al. 2005; Brayton, Lau et al. 2007)*. A summary of the *rpoC2* gene can be seen in Figure 3.6

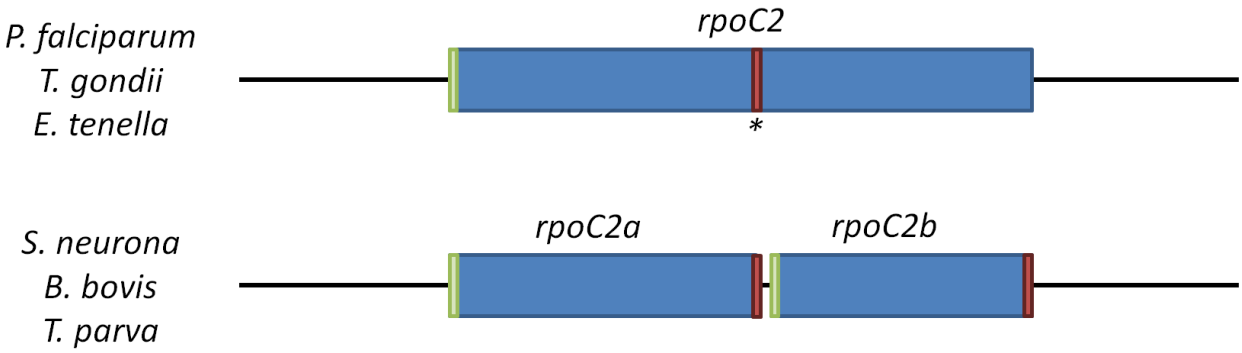

Figure 3.6: Diagram of *rpoC2* gene across apicomplexan species
Organisms are group depending on if *rpoC2* is split. The green box represents the start methionine. The red box signifies a non-UGA stop codon. The asterisk shows the location where a frame slip must occur to produce the correct translation product.

Discussion

We have determined the sequence and annotation of the 35 Kbp apicoplast genome for *Sarcocystis neurona*. Overall, gene content and organization is very similar to other *Coccidia* with a few differences. Between the gene ORFG and the LSUrRNA there is an insertion of a fragment of the *rps4* gene. One feature common to plasmodium and coccidian apicoplast genomes is the inverted repeat. The IR sequence is so similar that they collapse on themselves during assembly of shotgun reads. Due to this similarity it is possible that *rps4* which flanks the 3' end of IR had a recombination event which allowed part of the *rps4* gene to be inserted on the 5' end of the IR. The *rps4* insert aligns nearly perfectly with the *rps4* gene.

There is some ambiguity of the sequence of the *rps4* insert. Within *rps4* there is a homopolymer run of 11 A's however in the *rps4* insert there is a run of 10 or 11 A's. The 454 assembly program which uses flowgram information to predict the sequence determined there were 10 A's but sequencing of PCR product of that region determined there were 11 A's. If there are only 10 A's the *rps4* insert contains a stop UAA codon at position 144 of the insert, however if there are 11 A's there is no stop codon and translation could run into *ORFG* out of frame producing a dead mRNA. Due to the limitations of 454 sequencing of homopolymer runs it is difficult to definitively determine the sequence without expression data.

The *rpoC2* gene differs somewhat with *T. gondii*. While both contain a UAA codon in the center of the coding sequence, *S. neurona* contains an AUG codon in a different frame after the UAA codon as seen in Figure 3.7.

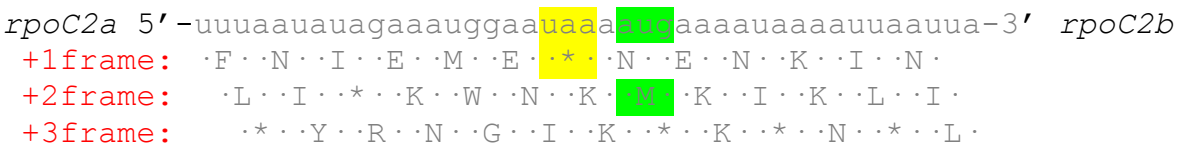

Figure 3.7: Three-frame translation of *rpoC2a* and *rpoC2b* gap
Highlighted in yellow is the stop codon for *rpoC2a.* Highlighted in green is the hypothetical start codon for *rpoC2b.*

There is no experimental evidence to support what the exact protein product is for *S. neurona.* One option is that *S. neurona* contains two separate rpoC2 subunits. In *T. parva, rpoC2* is annotated as two separate proteins *rpoC2a* and *rpoC2b* (Gardner, Bishop et al. 2005). Another option is that around the region of the UAA codon a frame slip occurs as the RNA is being translated. This would produce a single *rpoC2* protein. Finally it is possible that like *T. gondii*,

the UAA and UAG codon could be translated into an as of yet, unknown, amino acid. Despite not knowing the exact products of this gene, it is clear that this region of the *rpoC2* is undergoing evolution within the Apicomplexa.

It is curious that *rpl36* gene is not found in either in the apicoplast genome or the nuclear genome. This protein is highly conserved in bacteria but is not present in *Archaea* or *Animalia*. It is found in other *Coccidia*, *Plasmodia* and in the *Piroplasmida* apicoplast genomes. Also, there were no sequences were found to be similar to *T. gondii*'s *rpl36* protein coding sequence in the *S. neurona* apicoplast genome. However, the sequence location where the gene *rpl36* is conserved in other *Coccidia*, *Plasmodia* and *Piroplasmida* remains. It is possible that *rpl36* is not required for *S. neurona*. This gene encodes a ribosomal protein that is part of the 60S subunit. It has been shown in *E. coli* that *rpl36* can be knocked out and yet still survive, albeit with 40-50% slower cell growth (Maeder and Draper 2005). Maeder and Draper theorize due to the highly basic nature of the protein (~30% arginine or lysine), it acts as "mortar" filling in the gaps of the ribosome. It is feasible for *S. neurona* to not require the *rpl36* gene for proper ribosome function. Another possibility is that the *S. neurona* nuclear genome is incomplete and the region that contains the *rpl36* gene in the nuclear genome was not assembled into the genome scaffold sequence. If there is an *rpl36* gene encoded by the nuclear genome the gene product could be imported into the apicoplast if it also contained a transit peptide. Whether *S. neurona* has evolved a new protein to replace the function of rpl36, or it has not been sequenced yet, or the ribosome functions without *rpl36*, remains to be seen.

Chapter 4

SARCODB: DATABASE RESOURCE FOR THE EUKARYOTIC PROTIST PATHOGEN *SARCOCYSTIS NEURONA*

Introduction

Sarcocystis neurona is an intracellular coccidian parasite and is the primary agent for equine protozoal myeloencephalitis (EPM) (Davis, Daft et al. 1991). The genome and transcriptome for *S. neurona* have been sequenced but in this current generation of DNA sequencing scientists are being overwhelmed with data. Data needs to be organized and associated with tools that can aid in gene annotation and identifying genomic regions of interest for *Sarcocystis neurona*. SarcoDB (http://sarcodb.ctegd.uga.edu/) provides a user-friendly database that houses genomic data for *S. neurona* and closely related species. It also provides tools to analyze and explore the data. Data can be queried based upon BLAST searches, protein motifs and text searches of WUBLAST queries against NCBI's non-redundant database. These tools will not only aid in gene determination and annotation, but are also an asset to the *S. neurona* scientific community as a whole.

Data Content

Version 1.0 of SarcoDB contains the finalized assembly of the nuclear genome for the SN3 strain of *S. neurona.* It also includes the apicoplast genome sequence. As of this writing,

the mitochondrial genome has not been found, however parts of mitochondrial genes that are conserved among the Apicomplexa (cytochrome genes) are found among small, assembled scaffolds of the assembly. The mitochondrial genome of the closest sequenced relative of *S. neurona*, *Toxoplasma gondii* has also eluded full annotation. In addition to the genome sequence, SarcoDB contains clustered ESTs. There are various sets of clustered ESTs for the SN3 and SN4 (Davis, Speer et al. 1991) strain of *S. neurona,* each covering a different transcriptome assembly for the different life cycle stages of *S. neurona*. Table 4.1 contains a summary of all genomic and EST data available in SarcoDB.

Table 4.1: Summary of data located in SarcoDB

Text in parenthesis represents the sequencing technology used. Clustered ESTs are dataset created by Newbler 2.5 assembling various raw EST datasets. Early-mid schizont references clustered ESTs of the 2hr., 8hr. and 24hr. dataset. Late schizont represents clustered ESTs from the 64hr. schizont and SN3 schizont dataset. SN3 merozoites are clustered ESTs of all SN3 merozoites datasets. All SN3 is clustered ESTs of all SN3 datasets. SN4 merozoite is clustered ESTs SN4 dataset. *Sarcocystis falcatula* is clustered ESTs of all *Sarcocystis falcatula* ESTs from NCBI.

Type	Strain	Source	Number of sequences	Number of basepairs
Genome	SN3	Merozoite (454)	172	123,793,627
ORFs > 50			1,480,269	153,009,373
ORFs > 100			473,435	82,573,781
Clustered ESTs	SN3	Early-mid schizont	11,383	5,491,685
ORFs > 50			38,255	2,955,385
ORFs > 100			5,759	758,539
Clustered ESTs	SN3	Late schizont	7,179	3,781,693
ORFs > 50			31,174	2,509,923
ORFs > 100			5,726	764,903
Clustered ESTs	SN3	SN3 merozoites	23,876	16,748,383
ORFs > 50			156,304	13,617,675
ORFs > 100			34,565	5,240,677
Clustered ESTs	SN3	All SN3	37,507	25,088,694
ORFs > 50			223,423	19,269,525

ORFs > 100			48,304	7,236,972
Clustered ESTs	SN4	All SN4	439	256,718
ORFs > 50			2,141	169,845
ORFs > 100			368	48,330
Clustered ESTs	SF1	*Sarcocystis falcatula*	520	335,147
ORFs > 50			2,748	216,839
ORFs > 100			459	62,211
Raw ESTs	SN3	Merozoite (Sanger)	8,056	3,903,242
Raw ESTs	SN3	Schizont (Sanger)	1,601	897,848
Raw ESTs	SN4	Merozoite (Sanger)	6,413	2,620,359
Raw ESTs	SN3	Merozoite (454)	643,836	158,065,205
Raw ESTs	SN3	Merozoite (454)	104,915	42,934,666
Raw ESTs	SN3	2 hr. schizont (454)	24,430	9,959,296
Raw ESTs	SN3	8 hr. schizont (454)	13,714	5,361,866
Raw ESTs	SN3	24hr. schizont (454)	33,050	12,943,630
Raw ESTs	SN3	64hr. schizont (454)	90,075	35,398,947

In addition, results from pre-computed analyses have been added to the database. The database contains 6-frame open reading frames (ORFs) of greater than 50aa and 100aa for each clustered EST consensus sequence and genome data set respectively. ORFs (often called “the poor man’s gene finder”) can aid in initial the search for genes. Open reading frames are continuously being destroyed due to random DNA mutation, therefore any ORFs that remain constant within the genome must be under selective pressures that maintain them. By searching for ORFs, one is essentially searching through all exons greater than 150 bp. This method is limited because many putative ORFs are random statistical noise and not associated with real exons. Also, by using this method, potential exons shorter than 150 bp are ignored. SarcoDB also acts as a repository for all available sequence data. Users have the option to download bulk sequences, single FASTA sequences or subsets of FASTA sequences by specifying genomic locations, clustered ESTs and ORF datasets. Raw ESTs are also available for download.

Data-Mining Tools

At present, SarcoDB has multiple tools for analyzing and retrieving data, with a focus on tools that find genes in unannotated but assembled genome sequence. One method for finding genes is based on protein similarity. Towards that end, SarcoDB contains WUBLAST databases. Each sequence dataset has been formatted into a BLAST-searchable database. This allows the user to find DNA or protein sequences with similarity to the nuclear and apicoplast genomes, or clustered ESTs and ORF datasets. Users can adjust the following parameters: BLAST program, expect value, cutoff score, strand, number of descriptions and alignments, graphical output and width of HSP alignments.

Another useful data mining tool is the text search of BLAST results. All genome sequence data has been compared to the NCBI non-redundant protein sequence database. Each WUBLAST hit was then input into the program MSPcrunch to determine the best hits for each location as opposed to the best hits overall. Each NCBI text description of each BLAST hit was parsed in a searchable index in the database. This index allows the user to easily search for a protein description and see if there was a WUBLAST hit among the datasets. For example, the user could enter the word “actin” and any genome sequence regions that had a hit in the NCBI protein database that contained the word “actin” in their descriptor would be displayed. This tool effectively permits users to quickly find regions of an unannotated genome with high similarity to a known protein in the GenBank. Last, is the user-defined motif search tool. While many related protein sequences diverge, conversed amino acids in certain positions are seen. The user can specify their own amino acid patterns using PERL regular expressions to search

against protein and ORF datasets. For example: "CC6+RK" searches for two consecutive cysteine residues followed by a stretch of hydrophobic amino acids followed by the amino acid arginine and lysine. For more information please visit the SarcoDB website. By using regular expressions, the user can indicate very specific amino acid patterns that are only limited by the user's imagination. The homepage and layout of each tool can be seen in Figure 4.1.

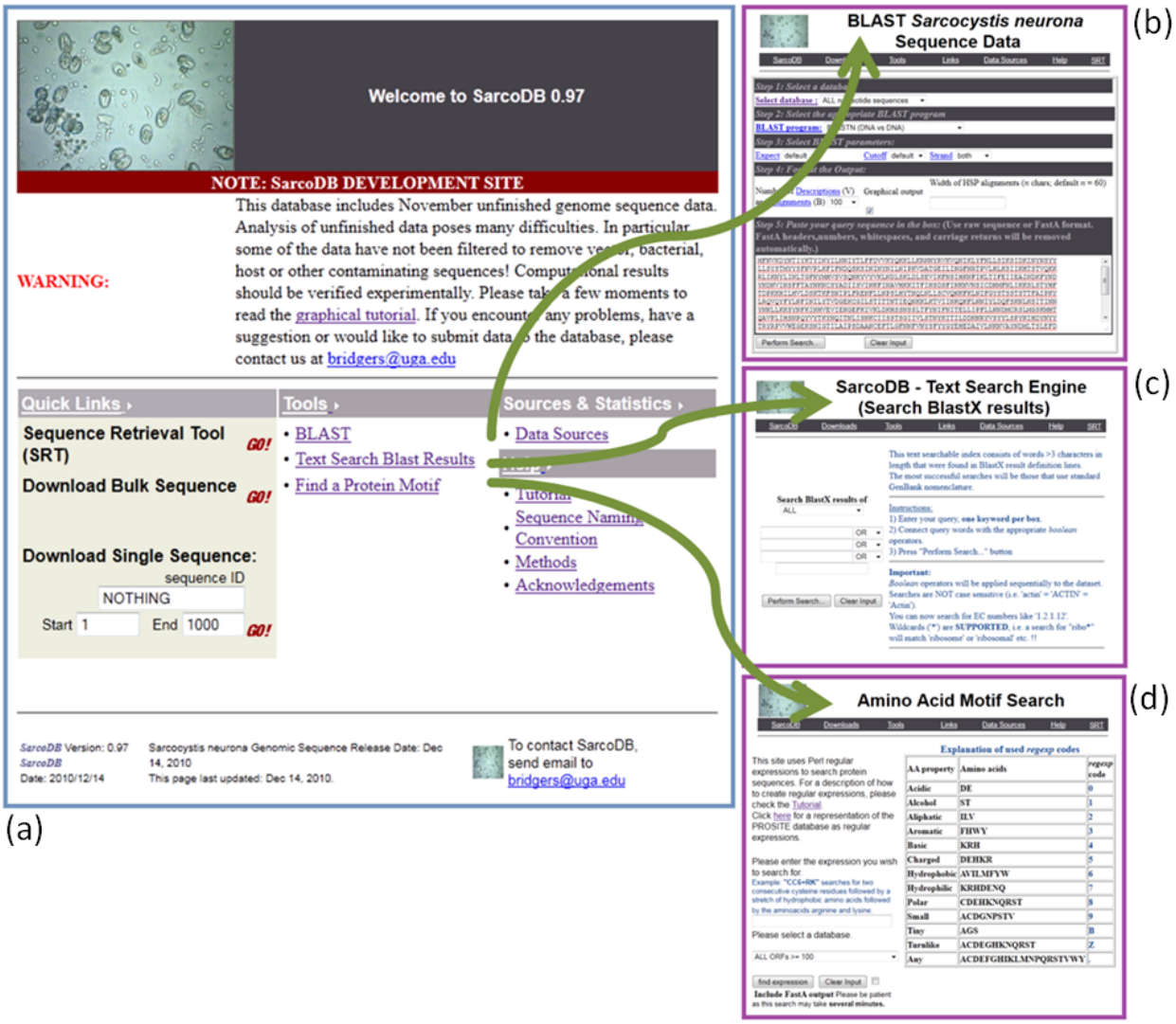

Figure 4.1: Screenshots highlighting the homepage and various tools in SarcoDB

(a) SarcoDB homepage. (b) BLAST tool. User can input protein or nucleotide sequence that can be BLASTed against any sequence data loaded into SarcoDB (c) Text Search of BLASTX results tool. The tool perform a keyword search of all BLASTX results of any sequence

loaded in SarcoDB verses the NCBI nr database. (d) Tool that will search all protein and ORF sequences in SarcoDB for protein motifs. Queries can be formatted as a regular expression.

Methods

SarcoDB creation

The SarcoDB database and web interface were created using the YourDB application (Kissinger, unpublished software) on a server running Linux and Apache HTTP Server. WUBLAST databases were created using the program, xdformat. ORFs greater than 50aa and 100aa were created with custom YourDB PERL scripts. WUBLAST BLASTX version 2.2.6 (Gish 1996-2004) was used to generate results used for the text search engine, against the June 15th 2011 version of nr from NCBI. The following parameters were used for the BLASTX search: matrix=BLOSUM62 V=100 B=1000000 -hspmax=1000000 W=4 T=18 -span1 -gi E=1e-3 -rdmask=seg. The WU-BLAST BLASTX results were then parsed by MSPcrunch with default parameters (Sonnhammer and Durbin 1994). MSPcrunch output was loaded into SarcoDB using the YourDB application.

Transcriptome assembly datasets

Based on results from Chapter 2, ESTs were clustered using Newbler 2.5 in cDNA mode of assembly with default parameters. Please refer to Chapter 2 for information on the sequencing of the transcriptome. Early-mid schizont transcriptome is composed of clustered ESTs of the 2hr., 8hr. and 24hr. dataset. Late schizont transcriptome is the clustered ESTs of 64hr. schizont and SN3 schizont dataset. SN3 merozoites is clustered ESTs of all SN3 merozoites datasets. All SN3 is clustered ESTs of all SN3 datasets. SN4 merozoite is clustered ESTs of the SN4 dataset. *Sarcocystis falcatula* is clustered ESTs of all *Sarcocystis falcatula* ESTs available

from NCBI on July 7th 2010. All transcriptomes were then loaded onto SarcoDB via YourDB custom scripts.

Future Directions

Future plans include adding new datasets to facilitate annotation, specifically the additional clustered ESTs produced by methods described in Chapter 2. Eventually, an annotation will be generated and SarcoDB will have outlived its usefulness. At that stage it is possible the data will be integrated into EuPathDB (http://eupathdb.org/).

CHAPTER 5

CONCLUDING REMARKS

This work focuses on the *Sarcocystis neurona* genome project. I have provided an assembly and annotation of the nuclear transcriptome and the organellar apicoplast genome and created a database for community-based investigation of available data. In this thesis, I have detailed my contributions to this project. I have constructed a foundation and framework that facilitates the overall genome project, and provides necessary and critical data and resources for both the genome project and the research community. As an academic exercise, I was able to assemble the *S. neurona* nuclear genome. After the nuclear genome was sequenced and assembled by the University of Kentucky, the project moved to the next phase of data acquisition. In Chapter 2, I show that, after comparison of assemblers, Newbler 2.5's transcriptome assembly consistently outperformed MIRA's. It was also the most accurate as measured by more accurate base calling, and a higher coverage of the proteome of *T. gondii.* I show that 454 and Sanger reads could be clustered by Newbler 2.5, which can then be used for annotation.

Once the transcriptome and necessary datasets were prepared, they were input into a custom pipeline that I developed specifically for *Sarcocystis neurona*. In Chapter 3, the foundation for annotation established in Chapter 2 was put to the test. Results of my pipeline produced annotations and alignments that are ready for manual editing. Using this pipeline, I

was able to produce annotation data for the organellar apicoplast genome, a potential drug target in the fight against apicomplexan parasites. During the course of this work, numerous differences in the in the *S. neurona* apicoplast genome were determined when compared to other coccidian apicoplast genomes. There are examples of gene loss and gain. Data can now begin to paint a picture of the evolution of the apicoplast genome, which is converging with other apicomplexan species. Chapter 3 shows how the pipeline aids in annotation. Now that the apicoplast genome is annotated and the pipeline tested, we can focus on the nuclear genome. In addition to the steps taken for the apicoplast annotation, further attention on nuclear genome tools is needed.

With the large amount of data being produced, it is necessary to organize the data in a database for access and investigation. Toward this end, SarcoDB is presented in Chapter 4. However SarcoDB is not just a repository for sequences. It also contains a set of tools directly integrated with the data that can aid the *S. neurona* scientific community. In chapter 3, analysis of the apicoplast was facilitated by SarcoDB via the search for 'missing' apicoplast genes that may have been transferred to the nuclear genome. SarcoDB is expected to grow, accumulating sequencing data. We hope this database will become an important tool to aid researchers in the analysis of the *S. neurona* genomes and with this understanding it will aid us in our research of all of Apicomplexa.

In conclusion, in this study I have been involved with all aspects of the *Sarcocystis neurona* genome project. I was able to assemble the nuclear and apicoplast genomes. I assembled multiple transcriptome data sets and proved which assembly algorithm performs the best. I have developed a pipeline for genome annotation and it effectiveness was shown

with the annotation of *S. neurona*'s apicoplast genome. Finally, in order facilitate *S. neurona* research, SarcoDB was created and maintained with the aim of greatly aiding our understanding and research of the obligate intracellular parasite, *Sarcocystis neurona.*

REFERENCES

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Aurrecoechea, C., J. Brestelli, et al. (2009). "PlasmoDB: a functional genomic database for malaria parasites." Nucleic Acids Res **37**(Database issue): D539-543.

Barbrook, A. C., C. J. Howe, et al. (2010). "Organization and expression of organellar genomes." Philos Trans R Soc Lond B Biol Sci **365**(1541): 785-797.

Blanchard, J. L. and J. S. Hicks (1999). "The non-photosynthetic plastid in malarial parasites and other apicomplexans is derived from outside the green plastid lineage." J Eukaryot Microbiol **46**(4): 367-375.

Bradnam, K. (2011). The Assemblathon: A genome assembly challenge. CSHL Biology of Genomes. Cold Spring Harbor.

Brayton, K. A., A. O. Lau, et al. (2007). "Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa." PLoS Pathog **3**(10): 1401-1413.

Cai, X., A. L. Fuller, et al. (2003). "Apicoplast genome of the coccidian *Eimeria tenella*." Gene **321**: 39-46.

Cantarel, B. L., I. Korf, et al. (2008). "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes." Genome Res **18**(1): 188-196.

Cheadle, M. A., S. M. Tanhauser, et al. (2001). "The nine-banded armadillo (*Dasypus novemcinctus*) is an intermediate host for *Sarcocystis neurona*." Int J Parasitol **31**(4): 330-335.

Cheadle, M. A., C. A. Yowell, et al. (2001). "The striped skunk (*Mephitis mephitis*) is an intermediate host for *Sarcocystis neurona*." Int J Parasitol **31**(8): 843-849.

Chevreux, B., T. Wetter, et al. (1999). "Genome Sequence Assembly Using Trace Signals and Additional Sequence Information." Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)(99): 45-46.

Cooley, A. J., B. Barr, et al. (2007). "*Sarcocystis neurona* encephalitis in a dog." Vet Pathol **44**(6): 956-961.

Dahl, E. L. and P. J. Rosenthal (2007). "Multiple antibiotics exert delayed effects against the *Plasmodium falciparum* apicoplast." Antimicrob Agents Chemother **51**(10): 3485-3490.

Dahl, E. L. and P. J. Rosenthal (2008). "Apicoplast translation, transcription and genome replication: targets for antimalarial antibiotics." Trends Parasitol **24**(6): 279-284.

Dahl, E. L., J. L. Shock, et al. (2006). "Tetracyclines specifically target the apicoplast of the malaria parasite *Plasmodium falciparum*." Antimicrob Agents Chemother **50**(9): 3124-3131.

Davis, S. W., B. N. Daft, et al. (1991). "*Sarcocystis neurona* cultured in vitro from a horse with equine protozoal myelitis." Equine Vet J **23**(4): 315-317.

Davis, S. W., C. A. Speer, et al. (1991). "In vitro cultivation of *Sarcocystis neurona* from the spinal cord of a horse with equine protozoal myelitis." J Parasitol **77**(5): 789-792.

DeBarry, J. and J. Kissinger (2011). "Jumbled Genomes: Missing Apicomplexan Synteny." Molecular Biology and Evoluiton **28**.

Dubey, J. P. (1974). "Letter: Toxoplasmosis in horses." J Am Vet Med Assoc **165**(8): 668.

Dubey, J. P., G. W. Davis, et al. (1974). "Equine encephalomyelitis due to a protozoan parasite resembling *Toxoplasma gondii*." J Am Vet Med Assoc **165**(3): 249-255.
Dubey, J. P. and A. N. Hamir (2000). "Immunohistochemical confirmation of *Sarcocystis neurona* infections in raccoons, mink, cat, skunk, and pony." J Parasitol **86**(5): 1150-1152.
Dubey, J. P., D. S. Lindsay, et al. (2001). "A review of *Sarcocystis neurona* and equine protozoal myeloencephalitis (EPM)." Vet Parasitol **95**(2-4): 89-131.
Dubey, J. P., W. J. Saville, et al. (2000). "Completion of the life cycle of *Sarcocystis neurona*." J Parasitol **86**(6): 1276-1280.
Dubey, J. P., W. J. Saville, et al. (2001). "*Sarcocystis neurona* infections in raccoons (*Procyon lotor*): evidence for natural infection with sarcocysts, transmission of infection to opossums (Didelphis virginiana), and experimental induction of neurologic disease in raccoons." Vet Parasitol **100**(3-4): 117-129.
Dubey, J. P., C. A. Speer, et al. (1998). "Isolation of a third species of *Sarcocystis* in immunodeficient mice fed feces from opossums (*Didelphis virginiana*) and its differentiation from *Sarcocystis falcatula* and *Sarcocystis neurona*." J Parasitol **84**(6): 1158-1164.
Esseiva, A. C., A. Naguleswaran, et al. (2004). "Mitochondrial tRNA import in *Toxoplasma gondii*." J Biol Chem **279**(41): 42363-42368.
Fast, N. M., J. C. Kissinger, et al. (2001). "Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids." Mol Biol Evol **18**(3): 418-426.
Fenger, C. K., D. E. Granstrom, et al. (1997). "Experimental induction of equine protozoal myeloencephalitis in horses using *Sarcocystis* sp. sporocysts from the opossum (Didelphis virginiana)." Vet Parasitol **68**(3): 199-213.
Fichera, M. E., M. K. Bhopale, et al. (1995). "In vitro assays elucidate peculiar kinetics of clindamycin action against *Toxoplasma gondii*." Antimicrob Agents Chemother **39**(7): 1530-1537.
Fichera, M. E. and D. S. Roos (1997). "A plastid organelle as a drug target in apicomplexan parasites." Nature **390**(6658): 407-409.
Foth, B. J., S. A. Ralph, et al. (2003). "Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*." Science **299**(5607): 705-708.
Furr, M., D. Howe, et al. (2011). "Antibody coefficients for the diagnosis of equine protozoal myeloencephalitis." J Vet Intern Med **25**(1): 138-142.
Gajria, B., A. Bahl, et al. (2008). "ToxoDB: an integrated *Toxoplasma gondii* database resource." Nucleic Acids Res **36**(Database issue): D553-556.
Gardner, M. J., R. Bishop, et al. (2005). "Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes." Science **309**(5731): 134-137.
Gish, W. (1996-2004). from http://blast.wustl.edu.
Goodman, C. D., V. Su, et al. (2007). "The effects of anti-bacterials on the malaria parasite *Plasmodium falciparum*." Mol Biochem Parasitol **152**(2): 181-191.
Granstrom, D. E., O. Alvarez, Jr., et al. (1992). "Equine protozoal myelitis in Panamanian horses and isolation of *Sarcocystis neurona*." J Parasitol **78**(5): 909-912.

Harper, J. T. and P. J. Keeling (2003). "Nucleus-Encoded, Plastid-Targeted Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) Indicates a Single Origin for Chromalveolate Plastids." Mol Biol Evol.

Huang, J. and J. C. Kissinger (2006). Lateral and intracellular gene transfer in the Apicomplexa: the scope and functional consequences. Genome Evolution in Eukaryotic Microbes. L. A. Katz and D. Bhattacharya, Oxford University Press.

Huang, X. and A. Madan (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-877.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-664.

Klumpp, S. A., D. C. Anderson, et al. (1994). "Encephalomyelitis due to a *Sarcocystis neurona*-like protozoan in a rhesus monkey (*Macaca mulatta*) infected with simian immunodeficiency virus." Am J Trop Med Hyg **51**(3): 332-338.

Korf, I. (2004). "Gene finding in novel genomes." BMC Bioinformatics **5**: 59.

Kuhsel, M. G., R. Strickland, et al. (1990). "An ancient group I intron shared by eubacteria and chloroplasts." Science **250**(4987): 1570-1573.

Kumar, S. and M. L. Blaxter (2010). "Comparing de novo assemblers for 454 transcriptome data." BMC Genomics **11**: 571.

Kuo, C. H. and J. C. Kissinger (2008). "Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*." BMC Evol Biol **8**: 108.

Kuo, C. H., J. P. Wares, et al. (2008). "The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees." Mol Biol Evol **25**(12): 2689-2698.

Laslett, D. and B. Canback (2004). "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences." Nucleic Acids Res **32**(1): 11-16.

Lewis, S. E., S. M. Searle, et al. (2002). "Apollo: a sequence annotation editor." Genome Biol **3**(12): RESEARCH0082.

Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-2189.

Lindsay, D. S., N. J. Thomas, et al. (2000). "Biological characterisation of *Sarcocystis neurona* isolated from a Southern sea otter (*Enhydra lutris nereis)*." Int J Parasitol **30**(5): 617-624.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**: 955-964.

MacKay, R. (1992). "Equine protozoal myeloencephalitis." Compend Contin Educ Vet **14**(1359-1367).

Maeder, C. and D. E. Draper (2005). "A small protein unique to bacteria organizes rRNA tertiary structure over an extensive region of the 50 S ribosomal subunit." J Mol Biol **354**(2): 436-446.

Majoros, W. H., M. Pertea, et al. (2004). "TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders." Bioinformatics **20**(16): 2878-2879.

Mansfield, L. S., S. Mehler, et al. (2008). "Brown-headed cowbirds (*Molothrus ater*) harbor *Sarcocystis neurona* and act as intermediate hosts." Vet Parasitol **153**(1-2): 24-43.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Martin, W. and R. G. Herrmann (1998). "Gene transfer from organelles to the nucleus: how much, what happens, and Why?" Plant Physiol **118**(1): 9-17.

Mayhew, I. G. (1976). Equine protozoal myeloencephalitis. 22nd Annual Convention of the American Association of Equine Practitioners. Dallas, TX**:** 107-114.
McConkey, G. A., M. J. Rogers, et al. (1997). "Inhibition of *Plasmodium falciparum* protein synthesis. Targeting the plastid-like organelle with thiostrepton." J Biol Chem **272**(4): 2046-2049.
McFadden, G. I. (2000). "Mergers and acquisitions: malaria and the great chloroplast heist." Genome Biol **1**(4): REVIEWS1026.
McFadden, G. I., M. E. Reith, et al. (1996). "Plastid in human parasites." Nature **381**(6582): 482.
McFadden, G. I. and D. S. Roos (1999). "Apicomplexan plastids as drug targets." Trends Microbiol **7**(8): 328-333.
Miller, J. R., S. Koren, et al. (2010). "Assembly algorithms for next-generation sequencing data." Genomics **95**(6): 315-327.
Moore, R. B., M. Obornik, et al. (2008). "A photosynthetic alveolate closely related to apicomplexan parasites-original article." Nature **451**(7181): 959-963.
Mullaney, T., A. J. Murphy, et al. (2005). "Evidence to support horses as natural intermediate hosts for *Sarcocystis neurona*." Vet Parasitol **133**(1): 27-36.
Nakamura, Y. and K. Ito (1998). "How protein reads the stop codon and terminates translation." Genes to Cells **3**(5): 265-278.
Oba, T., Y. Andachi, et al. (1991). "Translation in vitro of codon UGA as tryptophan in *Mycoplasma capricolum*." Biochimie **73**(7-8): 1109-1112.
Ohama, T., Y. Inagaki, et al. (2008). "Evolving genetic code." Proc Jpn Acad Ser B Phys Biol Sci **84**(2): 58-74.
Okamoto, N. and G. I. McFadden (2008). "The mother of all parasites." Future Microbiol **3**: 391-395.
Putignani, L., A. Tait, et al. (2004). "Characterization of a mitochondrion-like organelle in *Cryptosporidium parvum*." Parasitology **129**(Pt 1): 1-18.
Ralph, S. A., G. G. van Dooren, et al. (2004). "Tropical infectious diseases: metabolic maps and functions of the *Plasmodium falciparum* apicoplast." Nat Rev Microbiol **2**(3): 203-216.
Ramya, T. N., S. Mishra, et al. (2007). "Inhibitors of nonhousekeeping functions of the apicoplast defy delayed death in *Plasmodium falciparum*." Antimicrob Agents Chemother **51**(1): 307-316.
Roos, D. S., M. J. Crawford, et al. (1999). "Origin, targeting, and function of the apicomplexan plastid." Curr Opin Microbiol **2**(4): 426-432.
Samuel, W. M., A. A. Kocan, et al. (2001). Parasitic diseases of wild mammals. Ames, Iowa State University Press.
Sato, S. (2011). "The apicomplexan plastid and its evolution." Cell Mol Life Sci **68**(8): 1285-1296.
Sato, S., B. Clough, et al. (2004). "Enzymes for heme biosynthesis are found in both the mitochondrion and plastid of the malaria parasite *Plasmodium falciparum*." Protist **155**(1): 117-125.
Scolnick, E., R. Tompkins, et al. (1968). "Release factors differing in specificity for terminator codons." Proc Natl Acad Sci U S A **61**(2): 768-774.
Shirley, M. W. (2000). "The genome of *Eimeria* spp., with special reference to *Eimeria tenella*--a coccidium from the chicken." Int J Parasitol **30**(4): 485-493.

Sidhu, A. B., Q. Sun, et al. (2007). "In vitro efficacy, resistance selection, and structural modeling studies implicate the malarial parasite apicoplast as the target of azithromycin." J Biol Chem **282**(4): 2494-2504.
Singh, S., M. M. Alam, et al. (2010). "Distinct external signals trigger sequential release of apical organelles during erythrocyte invasion by malaria parasites." PLoS pathogens **6**(2): e1000746.
Slater, G. S. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison." BMC Bioinformatics **6**(1): 31.
Smit, A. and R. Hubley. (2008-2010). "RepeatModeler Open-1.0." from http://www.repeatmasker.org.
Sonnhammer, E. L. and R. Durbin (1994). "A workbench for large-scale sequence homology analysis." Comput Appl Biosci **10**(3): 301-307.
Stanke, M. and S. Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel." Bioinformatics **19 Suppl 2**: ii215-225.
Striepen, B., A. J. P. Pruijssers, et al. (2004). "Gene transfer in the evolution of parasite nucleotide biosynthesis." Proceedings of the National Academy of Sciences, USA **101**(9): 3154-3159.
Su, C., D. Evans, et al. (2003). "Recent expansion of *Toxoplasma* through enhanced oral transmission." Science **299**(5605): 414-416.
Tomova, C., W. J. Geerts, et al. (2006). "New comprehension of the apicoplast of *Sarcocystis* by transmission electron tomography." Biol Cell **98**(9): 535-545.
Toso, M. A. and C. K. Omoto (2007). "*Gregarina* niphandrodes may lack both a plastid genome and organelle." J Eukaryot Microbiol **54**(1): 66-72.
Waller, R. F., P. J. Keeling, et al. (1998). "Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*." Proc Natl Acad Sci U S A **95**(21): 12352-12357.
Waller, R. F., M. B. Reed, et al. (2000). "Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway." Embo J **19**(8): 1794-1802.
Weiss, L. M. and K. Kim (2007). *Toxoplasma gondii* : the model apicomplexan : perspectives and methods. London ; Burlington, Mass., Elsevier Academic Press.
Weiss, L. M. and K. Kim (2007). *Toxoplasma gondii* : the model apicomplexan : perspectives and methods. Amsterdam ; Boston, Elsevier/Academic Press.
Wilson, R. J. (2002). "Progress with parasite plastids." J Mol Biol **319**(2): 257-274.
Wilson, R. J., P. W. Denny, et al. (1996). "Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*." J Mol Biol **261**(2): 155-172.
Yabsley, M. J., C. N. Jordan, et al. (2007). "Seroprevalence of *Toxoplasma gondii*, *Sarcocystis neurona*, and *Encephalitozoon cuniculi* in three species of lemurs from St. Catherines Island, GA, USA." Vet Parasitol **144**(1-2): 28-32.
Yuthavong, Y., B. Panijpan, et al. (1985). "Biochemical aspects of drug action and resistance in malaria parasites." Southeast Asian J Trop Med Public Health **16**(3): 459-472.
Zhu, G., M. J. Marchewka, et al. (2000). "*Cryptosporidium parvum* appears to lack a plastid genome." Microbiol Mol Biol Rev **146**: 315-321.