2018

A GENOMIC APPROACH TO THE COMPLEX RELATIONSHIP BETWEEN AN APICOMPLEXAN ENDOSYMBIONT AND ITS HOST

Christopher Paight
University of Rhode Island, cpaight12@gmail.com

A GENOMIC APPROACH TO THE COMPLEX RELATIONSHIP BETWEEN AN APICOMPLEXAN ENDOSYMBIONT AND ITS HOST

BY

CHRISTOPHER PAIGHT

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

BIOLOGICAL AND ENVIRONMENTAL SCIENCES

UNIVERSITY OF RHODE ISLAND

2018

DOCTOR OF PHILOSOPHY DISSERTATION

OF

CHRISTOPHER PAIGHT

APPROVED:

Dissertation Committee:

Major Professor Christopher Lane

Jason Kolbe

Ying Zhang

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2018

ABSTRACT

Nephromyces, a genus in the phylum Apicomplexa, has recently been described as having a mutualistic relationship with its host: tunicates in the Molgulidae family (Saffo et al. 2010). If true, *Nephromyces* would be the only known example of a mutualistic apicomplexan genus. In addition to the possible switch to mutualism, *Nephromyces* is one of a few apicomplexan groups containing bacterial endosymbionts. To test the hypothesis that endosymbiotic bacteria facilitated the transition of *Nephromyces* from parasitism, the metabolic capabilities of *Nephromyces* and its bacterial endosymbionts need to be determined. The transition from obligate parasite to endosymbiont is predicted to involve different selective pressures leading to wide spread genomic changes. Identifying these changes will lead to a better understanding of the dynamics between the different biological players in this system.

Using data from Illumina HiSeq, we have assembled and annotated the transcriptomes of *Nephromyces* and *Cardiosporidium cionae.* Using data from a combination of platforms; Illumina MiSeq, HiSeq, and Pacific Biosciences, we have partially assembled a pan-genome for *Nephromyces* and have assembled the genomes of its bacterial endosymbionts. Using amplicon sequencing, we have estimated the genetic diversity and prevalence of multispecies infections of *Nephromyces* and its bacterial endosymbionts in its host *Molgula manhattensis.* In addition to the implementation of next-generation sequencing technologies, this work is also based on laboratory cultures and species isolation experiments.

With the aforementioned data we are able to describe the transcriptome

of *Nephromyces* and *Cardiosporidium* as well as the genomes of all three bacterial endosymbionts, providing a basic overview of the metabolism of this system. *Nephromyces* and *Cardiosporidium* both encode a complete purine degradation pathway, which enables them to break uric acid into pyruvate and glycine, additionally *Nephromyces* is also able to create malate from uric acid. This could represent the primary route of carbon, nitrogen and energy acquisition in *Nephromyces*. The genomes of the bacterial endosymbionts are severely reduced, but relatively enriched for vitamin and amino acid biosynthesis (at least in the Betaproteobacteria and Bacteroidetes symbionts). It is likely that the bacterial endosymbionts are supplementing vitamins and amino acids to the limited diet of uric acid found in *Nephromyces*. Our amplicon data reveals that nearly all *M. manhattensis* are infected with multiple species of *Nephromyces*. The community of *Nephromyces* forms a tightly integrated system of metabolic inter-dependencies based of the different bacterial endosymbionts.

ACKNOWLEDGMENTS

Without the support, guidance, and mentorship of my advisor Dr. Christopher Lane this work would not have been possible. The growth I have achieved both as a person and as a scientist is in large part due to his excellent mentorship. The input I received from my committee members, Dr. Jason Kolbe, Dr. Ying Zhang, and Dr. Marta Gomez-Chiarri was invaluable in the development of how I thought of this system.

I would also like to thank my colleagues in the lab for our scientific discussions and sharing new methods and strategies for dealing with problems. In particular I would like to thank Liz Hunter, Katie Neville, and Brandon Seah for their direct contributions to this work, and Kristina Terpis for making the lab run smoothly.

I would never have gotten to this point without the dedication of Jennifer Paight in teaching me to read, the sense of scientific curiosity instilled by Daniel Paight, and the patience and encouragement from Rachelle Amundson. Their continued support has pushed me to reach my highest potential. I would also like to thank Dr. Therese Markow for giving me my start in science, and encouraging me to attend graduate school. Finally I would like to thank Robert August for getting me through the rough spots and bad days.

PREFACE

The data chapters 2-4 have been prepared for submission as manuscripts and the manuscript format is in use. Chapter 2 has been submitted to the journal Genome Biology and Evolution (GBE) as a research article it includes the following sections in order; abstract, introduction, material and methods, results, discussion, acknowledgments, and references. Chapter 3 will be submitted to the journal Proceeding of the National Academy of Sciences (PNAS) as a research article it includes the following sections in order; abstract, introduction, results, discussion, material and methods, and references. Chapter 4 will be submitted to the Journal of Eukaryotic Microbiology (J.Euk.Microbiol.) as a research article it includes the following sections in order; abstract, introduction, material and methods, results, discussion, acknowledgments, and references.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Apicomplexa is a large, diverse phyla thought to be composed entirely of parasites of metazoans (Morrison 2009). It is believed that every species of metazoan is host to at least one species of apicomplexan parasite. Some apicomplexans are of particular significance to humans including *Plasmodium,* the etiological agent of malaria. *Plasmodium* is estimated to have driven the evolution of 4.3 percent of the human genome (Whitfield 2002; McManus et al. 2017). Another apicomplexan of note is *Toxoplasma,* which is able to infect all warm-blooded animals and is estimated to infect a third of the global human population (Wilking et al. 2016). *Cryptosporidium* is a waterborne apicomplexan responsible for 90% of occurrences of severe diarrhea in children under five, and a continuing challenge for municipal water treatment (Sow et al. 2016). In addition to the medical significance of apicomplexans, there has been research into apicomplexans on the effects of parasitism on an organisms genome, due to their long 700-900 million year history as a lineage of obligate parasites (Kuo et al. 2008).

Obligate parasites face different challenges than free-living organisms, resulting in different evolutionary pressures and unusual life histories as well as dramatic genomic changes in parasitic lineages (Janouskovec & Keeling 2016). One of the problems faced by parasites is the host's immune system. The need to evade the host's immune system results in a complex evolutionary arms race

between host and parasite, and is a core component of the red queen hypothesis (van Valen 1973). Apicomplexans have developed numerous and diverse strategies for evading their host's immune system, including the *var* family of genes in *Plasmodium.* Var proteins are cell surface antigens which are capable of reorganizing into a wide number of protein conformations to keep the host from recognizing the infection (Kyes et al. 2007). Another strategy, used by *Toxoplasma* is to suppress immune response by silencing signal pathways and by forming the latent bradyzoite cyst stage, which causes chronic infection (Blader & Saeij 2014).

In addition to evasion of the host's immune system, intracellular parasites have ready access to an abundance of pre-formed metabolites. Access to these pre-formed metabolites leads to one of the most common and pronounced consequences of parasite genome evolution, the loss of many basic biosynthesis pathways, which are critical in free-living organisms (Janouskovec & Keeling 2016). This loss of biosynthesis capabilities is particularly pronounced in Apicomplexa. Amino acid biosynthesis, vitamin and cofactor biosynthesis, purine synthesis, purine degradation, and fatty acid biosynthesis have all been lost in Apicomplexa (Woo et al. 2015). Because these losses are observed throughout the phylum, it was believed that these losses occurred early in apicomplexan evolution. However, there has been a recent proposal that these losses are a continuous gradual process (Zarowiecki & Berriman 2015). In either case, loss of biosynthetic pathways creates a dependence on the host for not only primary

carbon and nitrogen, but also a dependence on salvaging the hosts premade metabolites. This has led to a number of intricate and elegant strategies for host manipulation.

As a consequence of the extraction of nutrients and metabolites, a parasite's growth and reproduction imposes a cost to their host. The total impact on the host by the parasite is known as virulence (Read 1994). If a parasite's virulence is too great the host will die, either directly due to the parasite, or as a consequence of being weakened; *i.e.,* the host is too weak to find food, the host becomes easy prey for predators, or the weakened immune system makes the host vulnerable to other pathogens. If the host dies before the parasite can complete its life cycle, or before it can infect a new host, then the parasites fitness falls to zero. This leads to a complex balance between parasite growth and transmission to a new host. Factors involved in transmission and virulence include host genotype, parasite genotype, host health, parasite load, as well as external factors such as other parasites infecting the same host simultaneously (Frank 1996a). The interplay of these factors creates a dynamic relationship between host and parasite and has led to a wide variety of strategies (Alizon et al. 2009). One common strategy, adopted by many parasite species, is to lower their virulence to the host (Cressler et al. 2016). Parasites often achieve this by lowering their reproduction levels. Lower parasite density means lower virulence to the host, and as long as there is still good transmission to other hosts, this low-density strategy is often successful. A different strategy is

exemplified by *Plasmodium. Plasmodium* merozoites proliferate within the host's liver cells through schizogony by simultaneously inhibiting cell death (thereby avoiding immunity) until parasite levels are high enough to cause cell death, releasing release sporozoites into the bloodstream. The high density of sporozoites overwhelms the immune system and creates a high likelihood that a mosquito feeding on the host will ingest blood with sporozoites, maximizing transmission success. After simultaneous release sporozoites re-infect the liver to repeat the cycle again. This cycle causes the episodic fevers seen in malaria patients. The episodic overwhelming of the host immune system causes the high virulence found with *Plasmodium* infections. *Plasmodium falciparum* is particularly virulent, even among *Plasmodium* species. Some of the virulence of *P. falciparum* has been attributed to the large number of other human parasites found in the same locations. With the presence of other parasites the likelihood of multiple parasitic infections increases, which then leads to the virulence of all pathogens present being cumulative. Rather than a long sustained infection with low probability of transmission over a longer time, with the possibility of host death from other pathogens, *P. falciparum,* has adopted a high density, high virulence infection combined with periods hidden from the immune system, which maximizes transmission over a short period of time.

Although transmission strategies are diverse there are predictable patterns of a disease's epidemiology and the type of strategy a parasite is using. Parasites with high virulence are characterized by low prevalence in a population

or by low prevalence with sporadic outbreaks (Frank 1996b). Parasites using this strategy may reach high cellular densities in an effort to maximize transmission before the host dies. Alternatively, high-sustained prevalence in a host population indicates low virulence. Parasites using this strategy often maintain low cell densities with lower transmission rates over a longer period of time. Parasite virulence and transmission strategies are not dichotomous, but rather a continuum. Parasites with the highest, sustained prevalence being the least virulent and parasites with the lowest prevalence being the most virulent. Of course such predictions about virulence only apply to parasites that have co-evolved with their host, and does not apply to parasites infecting a new or incidental host. In these instances, virulence is often extremely high and often results in the host death before transmission, leading to a self-limiting infection pattern, *i.e*. Ebola in humans.

The relationship between prevalence and virulence is important for this work because *Nephromyces* mutualistic relationship with their hosts, *Molgula* tunicates was solely based on the nearly 100% year round infection prevalence (Saffo et al. 2010). As infection prevalence only predicts virulence, and not host/endosymbiont relationship, this characterization may have been unfounded and premature. Characterization of mutualism based solely on infection prevalence for *Nephromyces* is particularly surprising, because it becomes the only mutualistic genus in a phylum composed entirely of parasites. However, there are other indicators, besides prevalence, suggesting that the relationship

between *Nephromyces* and its host is unusual. Unlike other pathogens with low virulence, *Nephromyces* reaches extremely high cell densities inside its host. This high cell density becomes apparent when you compare *Nephromyces* to its closest known relative *Cardiosporidium cionae,* a blood parasite that infects a number of solitary tunicates outside of the Molgulidae family (Ciancio et al. 2008; Kumagai et al. 2010). As with *Nephromyces*, *C. ciona* has a high sustained infection prevalence reaching ~95% in *Ciona intestinalis*. Such high, sustained prevalence indicates that, like *Nephromyces, C. ciona* is largely avirulent. However, *Nephromyces* reaches over an order of magnitude higher cell densities than *Ciona.* In two such closely related organisms with closely related hosts with similar epidemiology in other respects, the difference in relative cell densities is striking. Typically, the higher the parasite load the greater the virulence, but paradoxically *Nephromyces* can remain avirulent and reach extremely high cell densities. Rather than focus on the proposed mutualistic relationship between *Nephromyces* and its host, which remains unclear, this work will focus on the apparent paradox in *Nephromyces'* epidemiology.

In order to consider the unusual epidemiology of *Nephromyces*, it is necessary to examine its other life history traits. The phylum Apicomplexa has a tremendous amount of variation in hosts, cell types infected, transmission methods, host manipulation strategies, life cycles, reproduction, and morphology (Roos 2005). Even with so much diversity, *Nephromyces* stands out as unusual for an apicomplexan. One of the most unusual aspects of *Nephromyces'* biology is

where it lives. *Nephromyces* is only found and completes its entire life cycle inside the renal sac (Saffo 1982). Specifically, *Nephromyces* is found in the lumen of the renal sac and, unlike other apicomplexans, is extra-cellular, with no part of its life cycle inside or joined to its host's cells. The renal sac is a large, ductless, structure present only in tunicates in the Molgulidae family (Goodbody 1965). The function of the renal sac has not been determined, and despite its name, the renal sac does not function as a typical kidney (Saffo 1978). The renal sac was named for the large deposits of uric acid and calcium oxalate, nitrogenous compounds that are the major constituents of kidney stones (Saffo & Lowenstam 1978). Localized deposits of uric acid are not exclusive to *Molgula* tunicates and many ascidians have crystallized uric acid deposits located in various tissues, but the deposits in *Molgula* are by far the largest (Lambert et al. 1998).

Another unusual aspect about *Nephromyces* is the presence of bacterial endosymbionts. Even though it is not unusual for Eukaryotes to have bacterial endosymbionts, it is unusual in the phylum Apicomplexa. The only other apicomplexan known to harbor a bacterial endosymbiont is *C. ciona.* Bacterial endosymbionts are a common way for an organism to add novel functionality to its metabolism, and the acquisition and maintenance of bacterial endosymbionts is a major driver of eukaryotic evolution Prominent examples include the alphaproteobacterium that became the mitochondria and the cyanobacterium that gave rise to the chloroplast (John & Whatley 1975; Mereschkowsky 1905). More recent bacterial endosymbionts provide their hosts with a wide variety of

metabolic capabilities including vitamin and co-factor biosynthesis, amino acid biosynthesis, methanogenesis, photosynthesis, and protection from parasitoids. (Moran et al. 2005; Gijzen et al. 1991; Marin et al. 2005; Hansen et al. 2012). These functions allow their hosts to colonize new habitats and take advantage of novel food sources.

One of the consequences of a parasitic lifestyle is the loss of biosynthetic capabilities, and bacterial endosymbionts can supplement a host's metabolism. It was hypothesized that *Nephromyces* bacterial endosymbionts were an important factor in *Nephromyces*' colonization of the renal sac and its paradoxical epidemiology (Saffo et al. 2010). Therefore, it was necessary to determine how *Nephromyces* bacterial endosymbionts were contributing to the host's metabolome. One hypothesis suggested that bacterial endosymbionts of *Nephromyces* are capable of degrading the abundant amounts of uric acid in the renal sac(Saffo et al. 2010). The degradation of uric acid was also proposed as the host benefit that made *Nephromyces* mutualistic instead of parasitic.

A previous study had found three different bacterial endosymbionts in *Nephromyces:* an alphaproteobacteria, a betaproteobacteria, and a bacteroidetes (Seah et al. 2011). This study also detailed how these different bacterial endosymbionts were never found together in the same *Nephromyces* cell. No explanation of how a species of a single-celled organism could maintain three different endosymbionts without the endosymbionts ever being together was given. What this study failed to recognize is there were multiple species of

Nephromyces inside the same renal sac, and that different *Nephromyces* species contained a single type of bacterial endosymbiont (Chapter 4).

Organisms harboring multiple endosymbionts are not uncommon (Bennett & Moran 2013; Moran et al. 2008; Gruwell et al. 2010). Many organisms that are dependent on bacterial endosymbionts contain two or three different endosymbionts. Multiple endosymbionts are often required due to the evolutionary consequences of a free-living bacteria becoming an endosymbiont (Wernegreen 2017, 2015; Mccutcheon & Moran 2011; Moran 1996). One driver of bacterial endosymbionts' evolution is a tiny population size relative to free-living bacterial. Another is when only a few, or just a single bacterium, is vertically transmitted to subsequent host generations. Small population size, coupled with extreme bottlenecks repeated every host generation, produces profound effects from genetic drift and results in an accelerated Muller's ratchet (Moran 1996). One of the consequences of the accelerated Muller's ratchet on bacterial endosymbionts is a severe reduction of all non-essential genes. Some of the genes commonly lost are DNA repair genes (Kuwahara et al. 2007). The loss of DNA repair genes combined with the effects of genetic drift leads to high mutation rates, a low ratio of synonymous/non-synonymous mutations, and an AT bias. Over time this results in endosymbiont genomes, which are small, gene poor, and AT rich (Moran 2002).

The genomic instabilities of bacterial endosymbionts can quickly make them a burden and a liability to the host. As bacterial endosymbionts decrease in

function, the host must support their symbionts to a greater and greater degree. If a host is dependent on their symbionts, this can result in reduced fitness if the endosymbiont requirements outpace the host's ability to meet them. A common solution to the problem of symbiont degradation is acquiring additional bacterial endosymbionts. In such cases either the original bacterial endosymbiont is replaced in favor of the new endosymbiont, or both bacterial endosymbionts can be maintained together (McCutcheon & Moran 2007). When both bacterial endosymbionts are retained, the metabolisms of the bacterial endosymbionts can become tightly intertwined. One such example from McCutcheon and Moran (2007) involves the endosymbionts of the glassy winged sharpshooter *Homalodisca vitripennis*, which maintains two bacterial endosymbionts *Sulcia muelleria* and *Baumannia cicadellinicola.* These two endosymbionts show tight metabolic integration, with *S. muelleria* capable of biosynthesizing 8 of 10 essential amino acids, menaquinone, and fabF for fatty acid biosynthesis. *Baumannia cicadellinicola* is capable of biosynthesizing the remaining two essential amino acids, a number of vitamins and cofactors, but not menaquinone, and has the rest of the genes needed for fatty acid biosynthesis except fabF (McCutcheon & Moran 2007).

The genus *Nephromyces* contains three different bacterial endosymbionts, and although all three types of bacterial endosymbiont are regularly found in the same renal sac, no species of *Nephromyces* is known to contain more than one type of bacterial endosymbiont. It is currently unknown if the different

endosymbionts are providing the same metabolic functions or not. If the bacterial endosymbionts are providing different functions, it is unclear if there is any interaction between the endosymbionts, as is seen in *H. vitripennis.* A significant difference between the two systems is that *H. vitripennis* contains multiple symbionts, and *Nephromyces* only contain one endosymbiont per organism. Given the very different habitats of *Cardiosporidium* and *Nephromyces* and the rarity of endosymbionts in apicomplexans, it seemed likely that the bacterial endosymbionts might have contributed to *Nephromyces* ability to colonize the renal sac and even to the unusual epidemiology of *Nephromyces*.

In order to explore; one, how *Nephromyces* is able to remain avirulent and reach such high cell densities, two, how *Nephromyces* was able to thrive in the unusual renal sac environment, three, what the relationship between *Nephromyces* and its host is, four, determine what effects this possibly mutualistic relationship had on *Nephromyces* genome, five, the role of the bacterial endosymbionts, and six, the differences between *Nephromyces* and other apicomplexans, we used a combination of next generation sequencing, culturing, and amplicon sequencing. Using Illumina HiSeq we sequenced and assembled the transcriptomes of *Nephromyces*, *Cardiosporidium,* their respective bacterial endosymbionts, *Molgula manhattensis,* and *Ciona intestinalis* (Chapter 2). Using a combination of Illumina MiSeq, HiSeq, and Pacific Biosciences we sequenced and partially assembled *Nephromyces* genome, and assembled two of *Nephromyces* bacterial endosymbionts genomes (alphaproteobacteria and bacteroidetes)

(Chapter 3). Using amplicon sequencing targeting 18s rRNA, 16s rRNA, and COI, we identified the diversity of *Nephromyces* and its bacterial endosymbionts (Chapter 4). With these data, in combination with culturing and isolation experiments, we were able to make substantial progress on characterizing the biology of both *Nephromyces* and *Cardiosporidium*, as well as their bacterial endosymbionts. This includes the unusual epidemiology, how *Nephromyces* survives in the renal sac, the role of the bacterial endosymbionts, and how *Nephromyces* compares to *Cardiosporidium* and other apicomplexans. In addition, we have uncovered some unexpected results, including a highly unusual basis for an apicomplexan metabolism and genus specific co-dependent species complex.

Chapter 2

Nephromyces encodes a urate metabolism pathway and peroxisomes, demonstrating these are not ancient losses of apicomplexans

by

Christopher Paight[1], Claudio H. Slamovits[2], Mary Beth Saffo[3] & Christopher E Lane[1]*

is submitted to the journal Genome Biology and Evolution

[1] Department of Biological Sciences, University of Rhode Island, Kingston RI, 02881, USA.

[2] Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada

[3] Smithsonian National Museum of Natural History, Washington, DC 20560, USA.

CHAPTER 2

Abstract

The Phylum Apicomplexa is a quintessentially parasitic lineage, whose members infect a broad range of animals. One exception to this may be the apicomplexan genus *Nephromyces*, which has been described as having a mutualistic relationship with its host. Here we analyze transcriptome data from *Nephromyces* and its parasitic sister taxon, *Cardiosporidium*, revealing an ancestral purine degradation pathway thought to have been lost early in apicomplexan evolution. The predicted localization of many of the purine degradation enzymes to peroxisomes, and the *in silico* identification of a full set of peroxisome proteins, indicates that loss of both features in other apicomplexans occurred multiple times. The degradation of purines is thought to play a key role in the unusual relationship between *Nephromyces* and its host. Transcriptome data confirm previous biochemical results of a functional pathway for the utilization of uric acid as a primary nitrogen source for this unusual apicomplexan.

Key words: Apicomplexan, tunicates, Peroxisomes, Purine degradation, *Nephromyces*, *Cardiosporidium*

Introduction

Apicomplexans are most well known for being parasites of humans and livestock. Species in the genus *Plasmodium*, for instance, are the etiological agents of malaria. Apicomplexan species show tremendous variation in transmission

methods, life cycles, host range, host manipulation strategies, cell-types infected, metabolic capabilities, immune evasion strategies, and virulence (Roos 2005; Reid et al. 2012; Kemp et al. 2013; Cardoso et al. 2016). Because of this variability, there are few apicomplexan characteristics shared throughout the phylum. Among the few universal apicomplexan features are a parasitic life-history and an inability to degrade purines (Janouskovec & Keeling 2016). *Nephromyces*, a derived apicomplexan genus of uncertain phylogenetic placement, appears to be an exception to both of these traits.

Nephromyces was misclassified as a fungus for more than a 100 years, based on long hyphal-like cell structures, flagellated spores interpreted by some as chytrid zoospores and cell walls made of a chitin (Giard 1888). It was not until the application of molecular methods that *Nephromyces* was confirmed as a member of the derived apicomplexans (Saffo et al. 2010). Although some analyses have tentatively placed it sister to adeleids, coccidia, or piroplasmida, the precise phylogenetic position of *Nephromyces* remains unresolved (Saffo et al. 2010; Janouškovec et al. 2015). *Nephromyces* species are monoxenous (infecting a single host) and are found exclusively in the Molgulidae family of tunicates (Saffo & Davis 1982). In a phylum composed of obligate parasites, the feature that distinguishes *Nephromyces* is its apparent mutualistic relationship with its tunicate hosts. The mutualistic relationship has been inferred based primarily on the nearly 100% infection rate and lack of clearance from the host (Saffo 1978, 1988, 1990, Saffo et al. 2010). We use this label with caution, given how complex

host-symbiont dynamics can be, how the costs and benefits of both "harmful" and "beneficial" symbioses can be difficult to determine, and how they can vary with genomic changes in hosts and symbionts (Leung & Poulin 2008; Saffo 2104; Mushegian & Ebert 2016).

A shift in lifestyle from obligate parasite to mutualistic symbiont is quite rare, and completely unknown from deep within a eukaryotic lineage with such a long evolutionary history of parasitism. One common consequence of a parasitic lifestyle is a loss of genes essential to free living organisms (Greganova et al. 2013; Janouškovec et al. 2015; Zarowiecki & Berriman 2015; Petersen et al. 2015). In an intracellular environment, if precursor molecules can be scavenged, there is less selective pressure to maintain biosynthesis pathways, and many are consequently lost (Keeling 2004; Sakharkar et al. 2004; Morrison et al. 2007). In phyla such as Apicomplexa, these losses can be extreme and over half of the genes found in their photosynthetic sister group, chromerids, have been lost in apicomplexans (Woo et al. 2015).

With so many basic metabolic functions lost, and with such dependence on the host, it is difficult to see how the relationship between host and parasite could change to a mutualistic interaction. However, one way for an organism to rapidly change its metabolic capabilities is to take on a bacterial symbiont. *Nephromyces* has done just that, leading to the hypothesis that bacterial endosymbionts inside *Nephromyces* perform some of the metabolic functions lost in Apicomplexa, and potentially contribute something beneficial to the tunicate

host (Saffo 1990; Saffo et al. 2010). Bacterial endosymbionts are common across the tree of life (although rare in apicomplexans) and perform a wide variety of functions for their hosts (Nowack & Melkonian 2010). These include amino acid metabolism and vitamin metabolism (Moran et al. 2005), nitrogen metabolism (Lopez-Sanchez et al. 2009), defense (De Souza et al. 2009), chemotrophic energy production (Urakawa et al. 2005), and photosynthesis (Marin et al. 2005), to name a few.

A tempting hypothesis for the functional role of *Nephromyces* bacterial endosymbionts is the break down of purines to urea in the purine degradation pathway (Saffo 1990). In support of this hypothesis *Nephromyces* infected tunicates have quite high levels of the enzyme urate oxidase, which catalyzes conversion of uric acid to 5-hydroxyisourate, but the enzyme is undetectable in uninfected tunicates (Mahler et al. 1955; Saffo 1988). Coupled with the fact that all known apicomplexans and tunicates have lost the purine degradation pathway, these data were suggestive of a bacterial contribution to purine degradation.

In a yet unexplained quirk of tunicate biology, many tunicate species have localized deposits of uric acid (Lambert et al. 1998; Saffo & Lowenstam 1978; Goodbody 1965). Storage as a form of excretion, nitrogen storage for future release, and structural support, are among the proposed functions of tunicate urate deposits (Goodbody 1965; Saffo 1988; Lambert et al. 1998). Tunicates in the Molgulidae family have the largest uric acid deposits, which are localized to a

specialized, ductless structure, called a renal sac (Saffo & Lowenstam 1978). These uric acid deposits occur regardless of infection status, indicating a tunicate origin of these purine deposits. Despite the name, the renal sac has many features (most notably, the absence of any ducts or macroscopic openings) atypical for an excretory organ, and its biological function has yet to be determined. *Nephromyces* infects feeding molgulid tunicates after the post-metamorphic onset of host feeding and completes its entire lifecycle within the renal sac. Four factors led to the conclusion that the bacterial endosymbionts within *Nephromyces* are the source of urate oxidase activity in this system: 1) the colonization of *Nephromyces* within a structure with high concentrations of urate, 2) the absence of urate oxidase activity in the molgulid hosts (Saffo, 1988, 1991), 3) the high urate oxidase activity found in *Nephromyces* (including its bacterial symbionts: Saffo, 1988, 1991), coupled with 4) the lack of obvious ultrastructural evidence of peroxisomes in *Nephromyces* (Saffo, 1990).

It is logical to think that the addition of bacterial endosymbionts to *Nephromyces* might have been key to colonizing this novel purine–rich niche, and is how *Nephromyces* escaped the "evolutionary dead end" of a parasitic lifestyle. In order to test this directly, and examine the metabolic relationships between the tunicate host, *Nephromyces*, and its bacterial endosymbionts, we sequenced the community transcriptome. To identify possible evolutionary or physiological changes involved in coevolution of *Nephromyces* with its molgulid hosts, we also sequenced the transcriptome of a sister taxon of *Nephromyces*, *Cardiosporidium*

cionae (Ciancio et al., 2008; Saffo et al., 2010), an apicomplexan parasite found in the blood in a broad range of non-molgulid ascidian hosts, including *Ciona intestinalis, Styela clava, Halocynthia roretzi,* and *Ascidiella aspersa* (Ciancio et al. 2008; Dong et al. 2006). Interestingly, *Cardiosporidium cionae* also harbors bacterial endosymbionts, which allows for a more direct comparison between *Nephromyces* and *Cardiosporidium*.

Here we confirm the exceptionally high levels of urate oxidase activity in tunicates with *Nephromyces,* and extend this result to include high expression levels of all the genes in the purine degradation pathway (xanthine dehydrogenase, urate oxidase, 5-hydroxyisourate hydrolase, 2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase, and allantoinase). The breakdown of purines starts by conversion to xanthine. Xanthine then enters the ureide pathway and the enzyme xanthine dehydrogenase catalyzes the reaction of xanthine to urate (Xi et al. 2000; Nishino et al. 2008). Urate oxidase catalyzes the oxidation of uric acid to 5-hydroxyisourate. Following conversion the enzyme 5-hydroxyisourate hydrolase catalyzes 5-hydroxyisourate to 5-hydroxy-2-oxo-4-ureido-2,5-dihydro-1H-imidazole-5-carboxylate (Kahn & Tipton 1998). This is further processed into (s)-allantoin by the enzyme 2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase (Jung et al. 2006). Allantoinase catalyzes (s)-allantoin into allantoate. From this point there are a few different pathways with different endpoints that organisms are able to shuttle allantoate to (Cusa et al. 1999). A common end point is to process allantoate into urea and

ureidoglycolate, to be further converted into carbon dioxide and ammonia. Alternatively, ureidoglycolate can be converted to glyoxylate, or the urea may be excreted as waste (Schultz et al. 2001; Werner et al. 2009).

We confirm that all the genes necessary for purine degradation are encoded by *Nephromyces* and *Cardiosporidium,* and not their endosymbiotic bacteria. Whereas the expression of urate oxidase by *Nephromyces* and *Cardiosporidium* is unexpected, a parallel issue is where the urate oxidase is physically located in the cell, given that apicomplexans reportedly lack peroxisomes (Schlüter et al. 2006). Urate oxidase activity is restricted to peroxisomes in eukaryotes, due to the numerous toxic byproducts that are produced in the break down of uric acid. Research into peroxisomes in Apicomplexa has a complex and contradictory history, with studies reporting both the presence (Kaasch & Joiner 2000; Gabaldon et al. 2016) and absence (Ding et al. 2000; Schlüter et al. 2006; Gabaldon 2010) of peroxisomes in Apicomplexa. Recent work by Moog *et. al* (2017) and Ludewig-Klingner et al. (2018) demonstrates compelling support for peroxisomes in coccidians. Both studies present comprehensive bioinformatic (and also proteomic, in part) evidence for the presence of peroxisomal biogenesis factors (peroxins) and typical peroxisomal metabolic enzymes (including predicted relevant targeting signals) in coccidians (Moog et al. 2017; Ludewig-Klingner et al. 2018). However, neither paper provides explicit experimental evidence (for example, microscopic) for the formation of peroxisomes in these organisms. Although direct evidence is

still absent, both studies point to (Lige et al. 2009)and their identification of peroxisome-like vesicles in *T. gondii*, for possible microscopic support.

Our data demonstrate that *Nephromyces* encodes a complete purine degradation pathway and a number of proteins predicted to be targeted to, or involved in, peroxisome biogenesis, maintenance and protein import, providing novel support of peroxisomes in Apicomplexa. Additionally, we propose the functional significance of purine degradation in *Nephromyces*, and reject the hypothesis that bacterial endosymbionts facilitated an escape from parasitism by providing genes in the purine degradation pathway.

Methods

Molgula manhattensis collection and laboratory culture

Molgula manhattensis tunicates were collected from a dock in Greenwich Bay, Rhode Island (41°39'22.7"N 71°26'53.9"W) on July 2014. For transcriptomic analysis, a single renal sac was separated from one tunicate, and all extraneous tissue removed. The intact renal sac was placed in liquid nitrogen for 5 min and then stored at -80°C for later RNA extraction. Gonads were dissected from five, sexually mature, *M. manhattensis*, collected from the same population in Greenwich Bay, Rhode Island August 2014. Eggs and sperm were mixed with sterile seawater and divided evenly between two petri dishes. Plates were incubated at room temperature for two days with daily 100% water changes. Tunicate larvae attached to the bottom and sides of the petri dishes by day three.

By day four, larvae had metamorphosed into adults and were actively feeding. Plates were moved to an incubator at 18° C with a 24 hr. dark cycle to limit growth of contaminants. Tunicates were fed by 100% water exchange with cultures of *Isochrysis galbana* and *Chaetoceros gracilis* three days a week. After several weeks tunicates were moved to aerated beakers to meet their increased nutrient and gas exchange requirements. Feeding regimen remained the same except that food volume was increased with tunicate growth. Tunicates were grown for six months until they were ~10mm across. Each renal sac was placed into a 1.5ml Eppendorf tube and flash frozen in liquid nitrogen. PCR screens confirmed *Nephromyces* was absent from lab-raised individuals. Lab grown tunicates were split into two groups. Renal sacs were harvested from three tunicates to use as transcriptome controls. A second group was infected with *Nephromyces* oocysts. Oocysts were collected from a wild *M. manhattenensis* and serially diluted by 50x to limit co-infections from multiple species, and raised for genomic analysis.

Cardiosporidium cionae collection, isolation and concentration

Ciona intestinalis were collected from docks in Snug Harbor RI (41.3890° N, 71.5201° W), in August 2017. Tunics were removed and the body wall was opened to allow access to the heart. A sterile syringe was used to remove cardiac blood as cleanly as possible. Blood was kept at 4° C until *Cardiosporidium* infection was verified using Giemsa stain to visualize *Cardiosporidium.* Heavily

infected samples were pooled together and centrifuged at 500 g for 5 minutes. The resulting supernatant was removed and the samples were frozen in liquid nitrogen and stored at -80° C. Samples with low rates of infection were enriched for *Cardiosporidium* using sucrose gradients (Ogedengbe et al. 2015; Arrowood and Sterling 2016). Gradients of 20, 25, 30, 35, 40% sucrose solutions in phosphate buffer were layered together. Approximately 5 ml of tunicate blood was added to the column and centrifuged at 500 g for 30 mins at 4° C. The 25% and 30% layers were collected (based on visual screens showing high *Cardiosporidium* cell density and low tunicate cell density), washed in PBS, pelleted and then frozen in liquid nitrogen and stored at -80° C.

RNA Extraction

RNA extraction buffer (Zymo Research LLC. Irvine, CA) was added to samples and ground with a pestle. Following grinding, the Zymo Quick-RNA kit (Zymo Research LLC. Irvine, CA) was used and the manufacturer's protocol was followed. RNA was converted to cDNA and sequenced at the School of Medicine Genome Resource Center, University of Maryland. Five separate paired-end RNA libraries (two from infected renal sac, and three uninfected renal sac) were multiplexed on one lane of the Illumina HiSeq platform, resulting in 326,299,923; 327,957,761 and 316,754,780 reads for the three renal sacs without *Nephromyces*, and 40,606,230 from the wild *M. manhattensis* renal sac. For *Cardiosporidium,* three samples of *C. intestinalis* blood were used: one with

unseparated blood, one enriched with cells collected at the 25% sucrose gradient, and one enriched with cells from the 30% sucrose gradient were multiplexed on one lane of the Illumina HiSeq platform, resulting in 92,250,706; 109,023,104 and 110,243,954 respectively. Transcriptome data was assembled and proteins were predicted with the Trinity/Trinotate pipeline version 2.4.0 run on the server at Brown University Center for Computation and Visualization (Haas et al. 2014). Reads assembled into 115,457; 388,535 109,446 contigs from infected, uninfected samples, and *C. intestinalis* respectively. Protein sequences were predicted using Transdecoder (Haas et al. 2014). Transcriptome completeness was assessed with Busco v3 against the Eukaryotic reference data sets (Simão et al. 2015).

Genomic DNA Extraction

The renal sacs from 8 lab grown *M. manhattensis* individuals were dissected and their renal fluid was pooled in a 1.5ml Eppendorf tube. Contents were centrifuged at 8000 g for 5 min to pellet *Nephromyces* cells, and following centrifugation the renal fluid was discarded. Five hundred microliters of CTAB buffer with 5ul of proteinase K and ceramic beads were added to the pelleted *Nephromyces* cells. The sample was placed in a bead beater for 3 min. and then on a rotator for 1.5hrs at room temp. Five hundred microliters of chloroform was added, mixed gently and centrifuged for 5 min. The top layer was removed and 2x the sample volume of ice cold 100% EtOH and 10% sample volume of 3M sodium acetate were added to the sample and incubated a -20^{o}C overnight. The sample

was centrifuged at 16000Xg for 30min. and the liquid was removed. Ice cold 70% EtOH was added and centrifuged at 16000xg for 15min. Liquid was removed and sample air dried for 2 min. DNA was re-eluted in 50ul of deionized water.

Illumina Sequencing

A nanodrop (2000c, ThermoScientific) was used to assess DNA purity and DNA concentration, and an agarose gel was run to assess genomic DNA fragmentation. Following quality control, an Illumina library was constructed. Library prep and sequencing were done at the URI Genomics and Sequencing Center (URIGSC). The completed library was sequenced on the Illumina MiSeq platform at the URIGSC and the HiSeq platform at the University of Baltimore sequencing center on three lanes.

Pacific Biosciences Sequencing

Using the contents of 150 (done in batches of 10 then pooled) *M. manhattensis* renal sacs, the same DNA extraction protocol was performed as for Illumina sequencing. DNA was sequenced using three SMRT cells on the Pacific Biosciences platform at the University of Baltimore sequencing center.

Illumina sequence data assembly

One MiSeq lane and three lanes of HiSeq, all from the same library, were trimmed using Trimmomatic (Bolger et al. 2014) and then assembled using Spades assembler (Bankevich et al. 2012) on the URI server BlueWaves.

Pacific Biosciences sequence data assembly

Pacific Biosciences reads were error corrected using pbsuite/15.8.24

(English et al. 2012) on the Brown University server, Oscar. Reads were then assembled using Canu (Koren et al. 2014). Contigs generated by Canu were combined with Illumina MiSeq/HiSeq short reads with Abyss v2.02 (Jackman et al. 2017).

Sequence annotation

Genes in the urate pathway were identified initially using KEGG GhostKOALA and KASS and subsequently by BLASTP searches against NCBI's nr protein database (Kanehisa et al. 2016). All candidate genes were screened using InterProScan to predict function (Finn et al. 2017). A curated database of phylogenetically representative species with good quality annotations for the three purine degradation genes and malate synthase were downloaded from NCBI. These genes were then used to construct gene trees.

Sequences were aligned with MAFFT (Katoh & Standley 2013) using FFT-NS-i. Maximum likelihood phylogenetic trees were constructed performed with RAxML (Stamatakis 2014) using the GAMMA model with 1000 seed trees and 1000 bootstrap replicates. Trees were viewed and modified using Figtree (v1.4.0, http://tree.bio.ed.ac.uk/software/figtree/).

Protein sequences were used to search against PeroxisomeDB (Schlüter et al. 2009) and BLAST hits lower than e^{-20} were retained and used in a BLASTP query against NCBI's Refseq protein database (Schlüter et al. 2009). Additional peroxisomal genes were identified with KAAS (Moriya et al. 2007). As many of

these peroxisome genes are encoded by *M. manhattensis*, all copies that had a closest hit to opisthokonta or bacteria were removed. Transcripts from uninfected *M. manhattensis* were used to screen additional tunicate genes using cd-hit at a 90% identity level (Li & Godzik 2006). Remaining genes were tested for signal motifs and subcellular location predictions with Wolf PSORTII, Ppero, TargetP, topcons, and Predotar (Supplementary Table 1)(Nakaia & Horton 1999; Wang et al. 2017; Emanuelsson et al. 2007; Bernsel et al. 2009; Small et al. 2004).

Nephromyces specific RNAseq reads were mapped to our genomic assembly using bowtie2 (Langmead & Salzberg 2012) with the –very sensitive flag set. Following mapping, Bedtools (Quinlan & Hall 2010) was used to quantify coverage across contigs, which were separated based on coverage levels. Contigs identified as *Nephromyces* were annotated using Maker2 with *ab initio* gene predictions from Augustus (Holt & Yandell 2011; Stanke et al. 2004).

Results

The contents of a single renal sac from an individual *Molgula manhattensis* resulted in 195,694 transcripts from *M. manhattensis*, *Nephromyces*, and the bacterial endosymbionts. After binning by species, 60,223 transcripts were attributed to *Nephromyces*. The cardiac fluid from 40 infected *Ciona intestinalis* individuals resulted in 109,446 transcripts, including 15,541 *Cardiosporidium* transcripts. The BUSCO algorithm was used to assess the completeness of the transcriptomes and reported 81.8% complete transcripts and 6.3% partial for the *Nephromyces* data and 69.7% complete and 11.9% partial for *Cardiosporidium*.

The *Nephromyces* genome assembly consisted of 1176 contigs greater than 5kb with a maximum length of 287,191 bp and an average length of 36 kb (Paight *et al*, in prep). This dataset was used to search for purine degradation genes to determine their genomic context. All of the purine degradation genes, as well as malate synthase, were predicted and annotated in the genome by Maker2. All genes but URAD contained introns, and neighboring genes on the identified contigs had top BLAST hits to apicomplexans in all cases (Table 1), indicating that they are encoded in the *Nephromyces* genome, not the endosymbiotic bacteria or host *Molgula manhattensis*. Phylogenetic trees for xanthine dehydrogenase, uric oxidase, malate synthase and allantoicase consistently resolved the monophyly of *Nephromyces, Cardiosporidium,* and Chromerids (Figure 1). Chromerids are the photosynthetic and the closest free-living relatives of Apicomplexa (Moore et al. 2008), indicating a vertical inheritance of this pathway from the common ancestor of apicomplexans.

The presence of urate oxidase also provides further support for peroxisomes in some lineages of Apicomplexa (Moog et al. 2017; Ludewig-Klingner et al. 2018), because urate oxidase activity is confined to peroxisomes in eukaryotes (Usuda et al. 1994). In addition to urate oxidase, *Nephromyces* and *Cardiosporidium* encode more peroxisome-associated proteins than *Plasmodium,* and nearly the same complement of genes encoded by *Toxoplasma* (Table 2). There are a few notable differences between *Toxoplasma* and *Nephromyces/Cardiosporidium*, including the absence of PEX3, PEX16, VLACS,

Table 1. Genomic context of the annotated purine degradation genes and malate synthase, in the *Nephromyces* genomic assembly. The phylogenetic affiliation of neighboring genes on each contig was identified by top hit against the NCBI nr database using BLASTp. Every contig encoding a target gene included other apicomplexan genes, and genes that did not hit apicomplexans had no strong affinity for other organisms.

Gene	Introns in gene	Contig	Contig length (kb)	Predicted Genes on Contig	Genes with top apicomplexan BLAST hits
XDH	4	Neph_3686418	24.5	4	2
UO	7	Neph_3687015	82.5	12	7
uraH	2	Neph_3685393	94.3	6	4
URAD	0	Neph_3687674	30.9	6	5
ALLC	10	Neph_3687655	116.3	16	11
MS	6	Neph_3671841	7	2	1

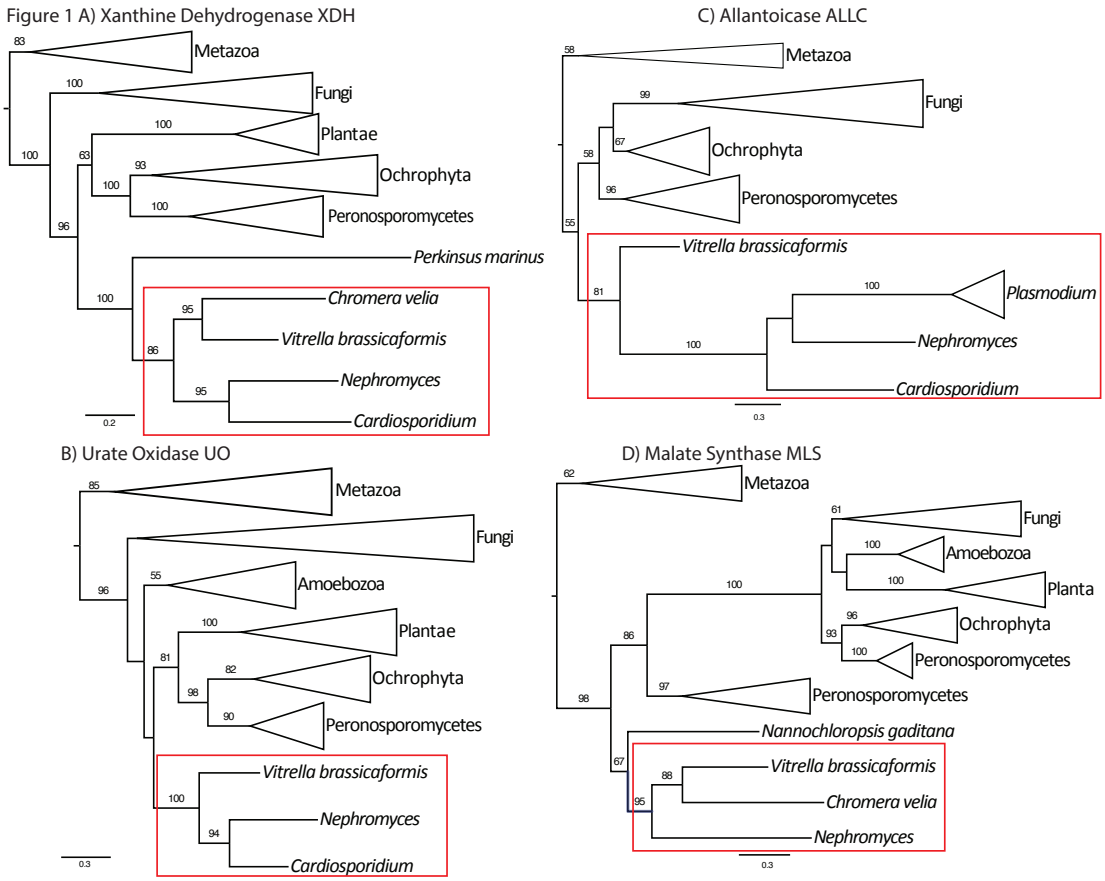


Figure 1. Maximum likelihood protein trees of A) Xanthine Dehydrogenase, B) Urate oxidase, C) Allantoicase D) Malate synthase. Genes A-C are involved in purine degradation and their position supports an ancestral apicomplexan purine degradation pathway in *Nephromyces/Cardiosporidium*. Malate synthase, D), acts on glyoxylate and acetyl-CoA to produce malate to complete the pathway. Stramenopiles are paraphyletic in the malate synthase phylogeny, possibly indicating a deep gene duplication. Whereas the support for deeper nodes is variable among all four genes, there is consistent support for a monophyletic origin of *Nephromyces/Cardiosporidium* genes with Chromerids (red box). Major lineages have been collapsed for presentation. Support values are percentage bootstrap support above 50%.

Table 2. Peroxiomal genes identified in *Nephromyces* and *Cardiosporidium* and their functional catagory. (X) denotes presence of gene and (-) absence. *Vitrella brassicaformis* (Vb) *Chromera velia* (Cv) *Cardiosporidium (C) Nephromyces (N) Toxoplasma gondii* (Tg) *Plasmodium falciparum (Pf) Cryptosporidium parvum (Cp).* Table modified based on Ludwig-Klinger et al. (2017).

	Protein	Abbr.	Vb	Cv	C	N	Tg	Pf	Cp
Glyoxylate pathway									
	Isocitrate lyase	ICL	x	x	-	-	-	-	-
	Malate synthase	MLS	x	x		x	-	-	-
	Citrate synthase	CS	x	x	x	x	x	x	-
	Aconitase	ACO	x	x	x	x	x	x	-
	Malate dehydrogenase	MDH	x	x	x	x	x	x	x
Peroxisome									
	Biogenesis factor 1	Pex1	x	x	-	x	x	-	-
	Biogenesis factor 2	Pex2	x	x	x	x	x	-	-
	Biogenesis factor 3	Pex3	x	x	-	-	x	-	-
	Ubiquitin carrier protein	Pex4	x	x	x	x	x	x	x
	Biogenesis protein 5	Pex5	x	x	x	x	x	-	-
	Biogenesis protein 6	Pex6	x	x	x	x	x	-	-
	Biogenesis protein 7	Pex7	x	x	x	x	x	-	-
	Biogenesis protein 10	Pex10	x	x	x	x	x	-	-
	Biogenesis factor 11	Pex11	x	x	x	x	x	-	-
	Biogenesis protein 12	Pex12	x	x	-	x	x	-	-
	Biogenesis factor 13	Pex13	-	-	-	-	-	-	-
	Membrane protein 14	Pex14	x	x	x	x	x	-	-
	Membrane protein 15	Pex15	-	-	-	-	-	-	-
	Biogenesis factor 16	Pex16	x	x	-	-	x	-	-
	Membrane protein receptor	Pex19	x	-	-	-	-	-	-
	Biogenesis protein 22	Pex22	x	x	x	x	x	x	x
	Biogenesis factor 26	Pex26	-	-	-	-	-	-	-
	Membrane channel	PMP22	x	x	x	x	-	-	-
	Membrane protein 4	PMP27	-	-	-	-	-	-	-
	ATP/ADP-transporter	PMP34	x	x	x	x	x	-	-
	Fatty acid ABC-transporter	PMP70	x	x	x	x	x	-	-
	ROS metabolism	MPV17	x	x	x	x	x	x	-
	Protein	Abbr.	Vb	Cv	C	N	Tg	Pf	Cp
Fatty acid oxidation									
a-oxidation	2-Hydroxyacyl-CoA lyase	HPCL2	-	x	-	-	-	x	-
	Phytanoyl-CoA hydrolase	PHYH	x	x	-	-	-	-	-
B-oxidation	a-Methylacyl-CoA-racemase	AMACR	-	-	-	-	-	-	-
	Acyl-CoA-oxidase	ACOX	x	x	x	x	x	-	-
	Multifunctional protein	DBP	x	x	x	x	x	-	-
	Sterole carrier protein 2	SCPX	-	x	-	-	x	-	-

	Multifunctional protein	PBE	-	-	-	-	-	-	-
	Acetyl-CoA acyltransferase 1	ACAA1	x	x	x	x	x	-	-
	2, 4-dienoyl-CoA reductase	PDCR	x	x	x	x	x	-	-
	d(3, 5)-d(2, 4)-dienoyl-CoA isomerase	ECH	x	x	x	x	x	-	-
	ATP-binding cassette, subfamily D	ABCD	x	x	x	x	x	-	-
	Long-chain acyl-CoA synthetase	ACSL	x	x	x	x	x	x	-
	Solute carrier family 27, member 2	VLACS	x	x	-	-	x	-	-
Other oxidation	Acyl-CoA thioesterase 8	PTE	-	-	-	-	-	-	-
	Nucleoside disphosphate-linked m.	NUDT19	-	x	-	-	-	-	-
Amino acid metabolism							-	-	
	Multifunctional protein	AGT	x	x	x	x	x	-	-
	D-Amino-acid oxidase	DAO	-	-	-	-	-	-	-
	Isocitrate dehydrogenase	IDH	x	x	x	x	x	x	-
	N1-acetylpolyamine oxidase	PAOX	x	x	-	-	-	-	-
	L-Pipecolate oxidase	PIPOX	x	x	-	-	-	-	-
	hydroxymethylglutaryl-CoA lyase	HMGCL	x	x	x	x	x	-	-
	(S)-2-hydroxy-acid oxidase	HAO	x	x	x	x	-	-	-
Antioxidant system							-	-	
Hydrogen peroxide metabolism	Catalase	CAT	x	-	x	x	x	-	-
	Superoxide dismutase	SOD	-	-	x	x	x	x	x
	Nitric-oxide synthase, inducible	INOS	-	-	-	-	-	-	-
	Peroxiredoxin 1	PRDX1	x	x	-	x	x	x	x
	Peroxiredoxin 5	PRDX5	-	-	-	-	-	-	-
Glutathione metabolism									
	Glutathione S-transferase kappa 1	GSTK1	x	x	x	x	-	-	-
	Protein	Abbr.	Vb	Cv	C	N	Tg	Pf	Cp
Etherphospholipid biosynthesis									
	Dihydroxyacetone phosphate acyltr.	DHAPAT	x	x	x	x	x	-	-
	Alkyldihydroxyacetonephosphate syn	AGPS	x	x	x	-	-	-	-
	Fatty acyl-CoA reductase	FAR	x	x	-	x	x	-	-
Purine metabolism									
	Xanthine dehydrogenase	XDH	x	x	x	x	-	-	-
	Urate oxidase	UO	x	x	x	x	-	-	-
Retinol metabolism									
	Dehydrogenase/reductase SDR family	DHRS4	x	x	x	x	-	-	-
Sterol precursor biosynthesis									
	Mevalonate kinase	MVK	-	-	-	-	-	-	-
	Phosphomevalonate kinase	PMVK	-	-	-	-	-	-	-

and SCPX in *Nephromyces/Cardiosporidium* and the absence of PMP22, GSTK1, DHRS4, XDH, and UO in *Toxoplasma.* Additionally, *Nephromyces* encodes a copy of Malate synthase (MLS) absent in both *Cardiosporidium* and *Toxoplasma*. Malate synthase is a key gene in the glyoxylate cycle, a pathway maintained in the photosynthetic *Chromera velia* and *Vitrella brassicaformis*, but lost in all other apicomplexans (Ludewig-Klingner et al. 2018). *Nephromyces/Cardiosporidium* also encode the enzyme serine-pyruvate transaminase (AGXT), which also uses glyoxylate as a substrate. AGXT converts glyoxylate into glycine and pyruvate and is often localized to peroxisomes, however the localization of AGXT in *Nephromyces/Cardiosporidium* is unclear (Supplementary Table 1).

Discussion

The recent scrutiny by Moog et al (2017), and Ludewig-Klingner et al. (2018) has built a case for the presence of peroxisomes in some apicomplexan lineages. While some apicomplexans may have lost peroxisomes, it seems likely that this loss is not a universally shared trait in the phylum. Despite the extensive search for peroxisome-associated functions in apicomplexans, no genes involved in purine degradation were found in other sequenced apicomplexan genomes, with the lone exception of allantoicase in *Plasmodium* (Gardner et al. 2002). Our *in silico* predictions indicate a complete purine degradation pathway in *Nephromyces* and *Cardiosporidium.* In addition to highly expressed transcripts for

the genes involved, all of the identified purine degradation genes and MLS have been located on genomic contigs from *Nephromyces*. Based on neighboring genes and the presence of introns in the *Nephromyces* genes matching the expressed transcripts, these contigs almost certainly originate from the *Nephromyces* genome (Table 1). Additionally, none of the purine degradation transcripts attributed to *Nephromyces* were detected in uninfected tunicates (Table 3). Phylogenetic trees of purine degradation genes are poorly supported at an inter-phylum level, indicating a rapid evolutionary rate. Whereas most genes are phylogenetically uninformative across the spectrum of eukaryotes, these gene trees have strong support for monophyly of purine degradation genes from *Nephromyces* and *Cardiosporidium* with Chromerids (Figure 1). The combination of gene trees, expression only when *Nephromyces* is present, and preliminary genomic assemblies strongly suggest that these genes were present since the divergence of Apicomplexa and Chromerida and have been vertically transmitted. Thus, these genes have been subsequently lost across apicomplexans, possibly multiple times. Although the exact placement of *Nephromyces* and *Cardiosporidium* is not certain (Saffo et al. 2010), multi-gene phylogenies place them in the subclass Hematozoa (Muñoz et al. in prep), suggesting that purine degradation was independently lost multiple times in Apicomplexa as well as maintained long after apicomplexans had become obligate parasites.

The presence of predicted purine degradation genes in *Nephromyces* and *Cardiosporidium,* adds a function not previously demonstrated in apicomplexan

peroxisomes (Table 2; Moog et al 2017; Ludewig-Klingner et al. 2018). While *Toxoplasma* and *Cardiosporidium/Nephromyces* share many of the same peroxisomal marker genes, no copy of PEX3 has been found in *Cardiosporidium/Nephromyces.* PEX3 (along with PEX10, PEX12, and PEX19) is one of the 4 genes reportedly required for peroxisome function (Schluter et al. 2006). However, the fundamentals of peroxisome biology have been described from a limited set of eukaryotes, and organisms such as ciliates have peroxisomes but also lack PEX3 (Ludewig-Klingner et al. 2017). Therefore, PEX3 may not be critical to peroxisome function for alveolates, and possibly other under-studied eukaryotic lineages. Extreme sequence conservation of PEX3 and PEX19 is only found in opisthokonta and sequence divergence in other lineages may indicate alternative functions (Hua et al. 2015).

Two other genes (Sterol carrier protein 2, SCPX and Solute carrier family 27, member 2, VLACS) missing from *Cardiosporidium/Nephromyces,* but found in *Toxoplasma*, are involved in β-fatty acid oxidation. Both *Cardiosporidium/Nephromyces* encode the seven other β-fatty acid oxidation genes encoded in *Toxoplasma*, suggesting β-fatty acid oxidation forms part of the functional capabilities of the *Cardiosporidium/Nephromyces* peroxisome*.* Fatty acid oxidation is often a central component of peroxisome function and has been hypothesized to be the impetus for the evolution of peroxisomes (Speijer 2011).

Based on transcript abundance, purine degradation in *Nephromyces* peroxisomes appears to be heavily utilized. Only 0.13% of genes had a higher

transcription rate than urate oxidase in our data from wild collected *Nephromyces,* and the other genes in the purine degradation pathway are among the most highly expressed transcripts in both wild and lab grown *Nephromyces* samples (Table 3). This result aligns with the previously reported high levels of urate oxidase protein in the renal sac of infected *Molgula* (Saffo 1988), indicating that the expression levels reported here do translate to protein. Much of this pathway is expressed over the 99th percentile of all transcripts in *Nephromyces,* which corresponds to the top 100 genes. Expression of purine degradation genes in *Cardiosporidium* is far lower, and in the 70-90 percentile range (Table 3). Such high expression in *Nephromyces* represents an enormous metabolic investment and it is unlikely that these transcripts go largely untranslated.

Both *Nephromyces* and *Molgula manhattensis* encode xanthine dehydrogenase, and are able to convert xanthine into uric acid. Since we have identified the tunicate host as the source of purines, this raises the question of why *Nephromyces* is expressing xanthine dehydrogenase in the 97.87th percentile, compared with similarly high tunicate expression (93.64th percentile). Although the percentile ranking between these two organisms cannot be directly compared, such high xanthine dehydrogenase expression in *Nephromyces* is surprising. It seems unlikely that so much xanthine dehydrogenase production is needed to convert only endogenous purines of *Nephromyces*. However, xanthine is only detected in the renal sac in small quantities, not nearly as abundant as uric acid, and xanthine dehydrogenase activity is restricted to the renal wall, not the

Gene	Wild Neph	Lab grown Neph 1	Lab grown Neph 2	Cardio Fraction 1	Cardio Fraction 2	Cardio Fraction 3	Mm	Uninfected Mm 1	Uninfected Mm 2	Uninfected Mm 3
xanthine dehydrogenase	97.87	93.17	94.83	none	76.88	69.5	93.64	N/A	N/A	N/A
urate oxidase	99.87	99.44	99.54	86.75	87.24	70.98	-	-	-	-
5-hydroxyisourate hydrolase	99.16	91.31	88.41	87.67	83.27	79.1	-	-	-	-
OHCU decarboxylase	93.38	-	-	-	-	-	-	-	-	-
allantoinase	99.09	98.38	98.23	73.61	90.32	71.89	-	-	-	-
amindohydrolase	99.75	79.25	89.18	87.43	92.27	92.08	-	-	-	-
malate synthase	59.17	93.81	93.11	-	-	-	-	-	-	-
serine-pyruvate transaminase	99.85	99.57	99.79	84.64	80.81	77.79	85.65	91.17	71.85	75.05

Table 3. Expression percentile ranking of purine degradation genes, from total expressed transcripts in *Nephromyces (*Neph*), Cardiosporidium (*Cardio*)* and *Molgula (*Mm*)*. The wild *Nephromyces* and *Molgula manhattensis* data originate from the same RNA extraction and were bioinformatically separated. Data was also generated from laboratory grown tunicates, artificially infected with *Nephromyces* (Lab grown Neph 1 & 2). *Cardiosporidium* fractions represent 1) unfiltered pericardial fluid, 2) the 25% and 3) 30% fractions extracted from a sucrose gradient, and may contain different proportions of *Cardiosporidium* life stages. The three uninfected *Molgula manhattensis* were raised from gametes in the lab and never exposed to *Nephromyces* infection. The (-) denotes the transcript was not recovered in that dataset whereas (N/A) indicates the transcript was assembled, but the transcripts per million (TPM) was <1.

renal lumen (Nolfi 1970). One possible explanation is that *Nephromyces* exports its xanthine dehydrogenase into the renal wall in order to drive the production of xanthine from hypoxanthine before the purine salvage enzymes adenine phosphoribosyltransferase and hypoxanthine-guanine phosphoribosyltransferase can salvage hypoxanthine into adenine and guanine.

High expression of purine degradation genes in *Nephromyces* is clear, but the purpose is uncertain. It does indicate purine degradation is an important pathway for *Nephromyces,* however, the functional significance is not immediately obvious. Pathway analysis predicts that *Nephromyces* is able to convert xanthine into urea and ureidoglycolate, however neither compound is biologically useful without further conversion. We propose that the products of purine degradation in *Nephromyces* are converted to glyoxylate.

One possible route is the conversion of ureidoglycolate into glyoxylate. There are two known enzymes able to catalyze this conversion: ureidoglycolate lyase, found in fungi and bacteria, which catalyzes (s)-ureidoglycolate to glyoxylate and urea, and ureidoglycolate amidohydrolase, found in plants and bacteria, which catalyzes (s)-ureidoglycolate to glyoxylate, carbon dioxide, and ammonia (Muñoz et al. 2006; Percudani et al. 2013; Wells & Lees 1991; Werner et al. 2010; Shin et al. 2012; Serventi et al. 2010). Both ureidoglycolate lyase and ureidoglycolate amidohydrolase are amidohydrolases - hydrolases that use amide bonds as substrates. No orthologs to either ureidoglycolate lyase and

ureidoglycolate amidohydrolase have been found in the *Nephromyces* transcriptome. However, an amidohydrolase is present, which is predicted to be structurally similar to the ureidoglycolate amidohydrolase found in *Arabidopsis*, including similar location and number of zinc binding domains. This amidohydrolase also has a similarly high expression level as the other purine degradation enzymes (Table 3). In order to determine if the amidohydrolase found in *Nephromyces* is capable of catalyzing (s)-ureidoglycolate, functional assays will need to be performed.

While the functionality of this particular amidohydrolase has yet to be determined, its ability to act on an (s)-ureidoglycolate is an attractive hypothesis for a few reasons. One, there are two known enzymes capable of breaking the amide bond in (s)-ureidoglycolate that have independently evolved: ureidoglycolate lyase and ureidoglycolate amidohydrolase. This pathway has not been widely explored across eukaryotes and the modification to a class of molecules able to break amide bonds to accommodate the structure of (s)-ureidoglycolate may not be a complex evolutionary step. Two, (s)-ureidoglycolate is unstable and will spontaneously convert to glyoxylate, albeit without the stereospecific conversion present when catalyzed by ureidoglycolate amidohydrolase (Werner et al. 2010). Spontaneous conversion of glyoxylate results in a 50% loss of efficiency *versus* enzymatic conversion, presumably creating strong evolutionary pressure to enzymatically degrade (s)-ureidoglycolate to maintain stereochemistry.

Glyoxylate is a common substrate for a number of enzymes including glyoxylate oxidase, which catalyzes glyoxylate with water and oxygen to form oxalate and hydrogen peroxide (Kasai et al. 1963). Notably, no copy of glyoxylate oxidase has been identified in *Nephromyces*, which is surprising given that another common component of the renal sac is calcium oxalate (Saffo & Lowenstam 1978). We have not identified any genes suggesting that *Nephromyces* or its bacterial endosymbionts can produce or process oxalate. Calcium oxalate is also found in uninfected hosts indicating that the tunicate is the source. Another enzyme that uses glyoxylate as a substrate, which is present in *Nephromyces/Cardiosporidium*, is serine-pyruvate transaminase (AGXT), which can be localized to peroxisomes or mitochondria, and catalyzes glyoxylate to glycine and pyruvate (Takada & Noguchi 1985). An alternative enzyme for processing glyoxylate is malate synthase (MLS), which is also targeted to the peroxisome and missing from apicomplexans, including *Cardiosporidium*, but is found in *Nephromyces* (Figure 1).

Malate synthase is one of two genes integral to the glyoxylate cycle, an alternative pathway for part of the citrate cycle. In the glyoxylate cycle, isocitrate is converted into glyoxylate and succinate by isocitrate lyase (McFadden & Howes 1965). Glyoxylate is combined with acetyl-CoA to create malate (Molina et al. 1994). This cycle allows for the creation of glucose from fatty acids directly (Kornberg & Krebs 1957). The presence of malate synthase indicates at least a piece of the glyoxylate cycle is present in *Nephromyces*. No copy of isocitrate lyase

is predicted from the *Nephromyces* transcriptome, and only a small fragment of a possible isocitrate synthase has been identified in *Cardiosporidium*. However, under the model proposed here, the generation of glyoxylate is from uric acid, and isocitrate synthase would not be required.

Both AGXT and MLS (in *Nephromyces*) show similarly high expression as the purine degradation genes (Table 3), which is consistent with our proposed uric acid to glyoxylate pathway. In particular, AGXT is among the most highly expressed *Nephromyces* transcripts, with consistently higher expression than MLS, possibly indicating it is the primary route of glyoxylate conversion. The products of AGXT, glycine and pyruvate, are versatile substrates and used by a number of pathways. Glycine is the simplest amino acid and an essential component of many important biological compounds, as a nitrogen source in a readily useable form. Pyruvate is extremely versatile and involved in several critical biological pathways. A non-inclusive list includes amino acid biosynthesis, acetyl-CoA biosynthesis, fatty acid biosynthesis, and the citric acid cycle. These pathways represent both carbon and energy acquisition (Figure 2). Additionally, *Nephromyces* has the ability to use MLS to convert glyoxylate and acetyl-CoA into malate, a compound central to the citric acid cycle, allowing for another mechanism of carbon and energy acquisition (Figure 2).

The hypothesized conversion of uric acid to glyoxylate in *Nephromyces* creates several possibilities. First, it allows for the metabolic waste product, uric acid, to be converted into glycine, pyruvate, and malate (Figure 2). Second, it

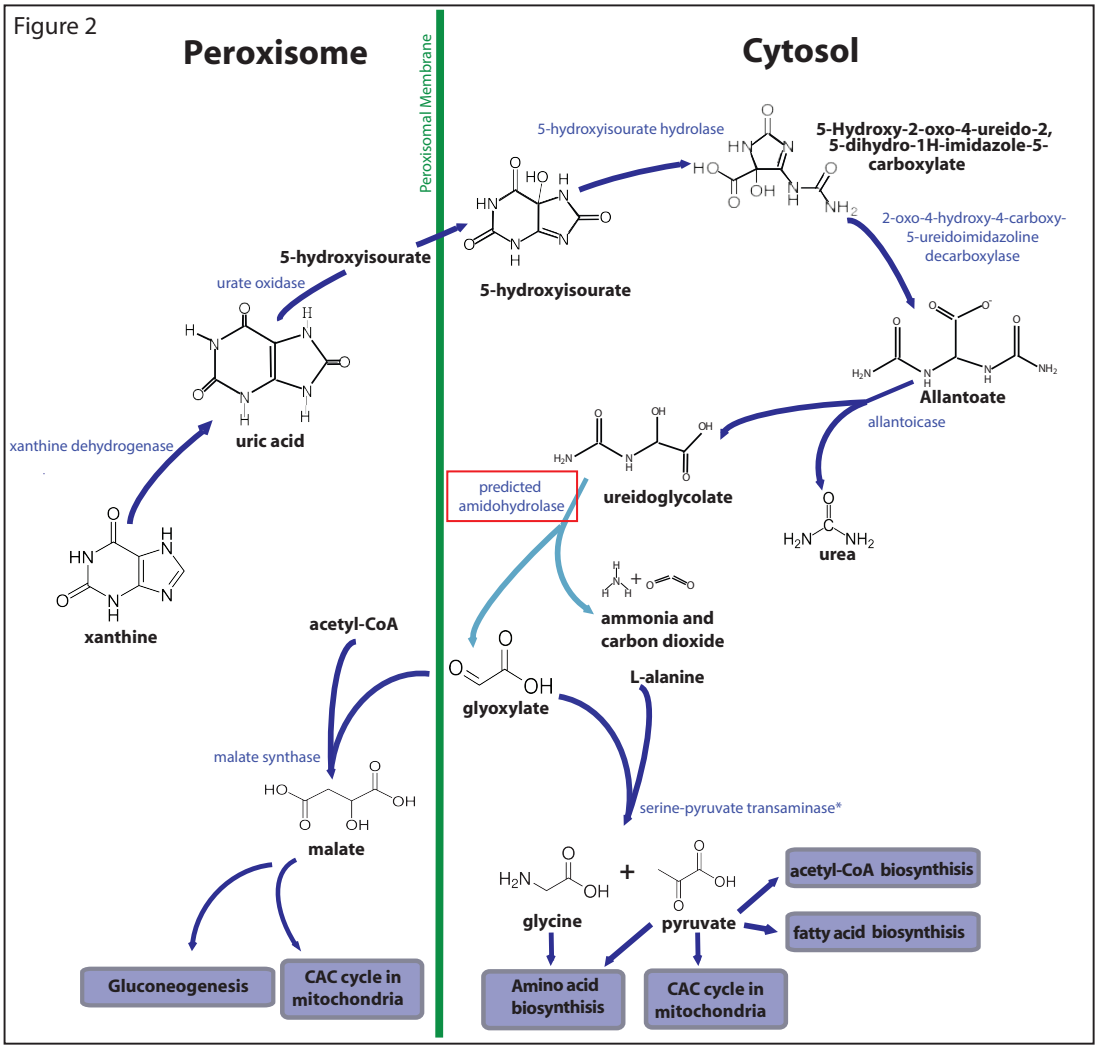


Figure 2. Predicted purine degradation pathway in *Nephromyces*, within the peroxisome and cytosol. Dark blue arrows represent enzymes identified in the *Nephromyces* transcriptome. The light blue arrow represents the highly expressed amidohydrolase (red box) predicted to convert ureidoglycolate into glyoxylate. Enzymes on the left side are localized to peroxisomes, the right side to the cytosol, with the green vertical line representing the peroxisomal membrane. The predicted pathway is able to convert uric acid into glyoxylate, and subsequent conversion by serine-pyruvate transaminase (AGXT) or malate synthase, creates glycine and pyruvate or malate respectively. The * by AGXT indicates ambiguous predicted localization, to either peroxisomes or mitochondria

provides an explanation for the exceptionally high expression of the purine degradation pathway. Third, it gives *Nephromyces* access to a primary carbon, nitrogen, and an energy source at no cost to its host. And finally, this change in primary carbon, nitrogen, and energy could conceivably reduce the impact of *Nephromyces* on its host, allowing *Nephromyces* densities to increase while decreasing virulence. Reduction in virulence would have been a necessary first step toward mutualism.

Uric acid as a primary carbon and energy source is not completely unknown. Bacterial species have been found in chicken hutches that were able to grow solely on uric acid (Rouf & Lomprey 1968; Thong-On et al. 2012), and some species of fungi are able to grow on media solely containing uric acid (Middelhoven et al. 1989). However, this is a novel substrate for an apicomplexan to grow on, and while it is unlikely that *Nephromyces* could survive on uric acid alone, it is a promising base for both carbon and nitrogen acquisition. It is possible that the *Nephromyces* bacterial endosymbionts (Sabree et al. 2009; Potrikus & Breznak 1980) are contributing to the proposed purine to glucose pathway, but that is not currently supported by our data.

As the adaptive significance of uric acid deposits in tunicates, and particularly in *Molgula*, are unknown, it is difficult to speculate on the effects of *Nephromyces* uric acid degradation to the host. If these renal sac deposits are a form of excretion by storage, as has been hypothesized (Goodbody 1965), then

having a symbiont that is capable of digesting uric acid may be beneficial simply by digesting an indigestible metabolite and converting uric acid into urea. Alternatively, once the uric acid has been broken down, the tunicate may benefit from metabolites derived from uric acid previously unavailable to the tunicate. If *Nephromyces* is overexpressing xanthine dehydrogenase in order to outcompete adenine phosphoribosyltransferase and hypoxanthine-guanine phosphoribosyltransferase, diverting hypoxanthine from purine salvage to purine degradation, there could be a potential cost to the host under purine-limited conditions.

Our data demonstrate that both the proposed mutualistic *Nephromyces* and parasitic *Cardiosporidium* encode the genes for purine degradation, which have been lost in other apicomplexans sequenced to date. Additionally, these genes share a common ancestry with chromerid genes, indicating they are not the product of a recent horizontal gene transfer from bacteria. These data also add support to the growing body of evidence that indicate the presence of peroxisomes in apicomplexans. *Nephromyces* and *Cardiosporidium* are predicted to have peroxisomes and, unlike any other apicomplexan, are capable of preforming both purine degradation and part of the glyoxylate cycle. The presence of purine degradation, AGXT, and MLS allow for the intriguing possibility of conversion of uric acid into a primary nitrogen, carbon and energy source. This predicted metabolic activity would be a completely novel substrate for an apicomplexan and may have been an important factor in the reduction of

virulence in *Nephromyces*.

Acknowledgements

Support for this project was provided by National Institute of Health, grant AI124092, as well as National Science Foundation award 1541510. CHS is supported by NSERC (Discovery grant RGPIN/05754-2015).

Anderson TJC et al. 2000. Microsatellite Markers Reveal a Spectrum of Population Structures in the Malaria Parasite *Plasmodium falciparum*. Mol Biol Evol. 17:1467–1482. doi: 10.1093/oxfordjournals.molbev.a026247.

Arnott A, Barry AE, Reeder JC. 2012. Understanding the population genetics of *Plasmodium vivax* is essential for malaria control and elimination. Malar. J. 11:14. doi: 10.1186/1475-2875-11-14.

Arrowood MJ, Sterling CR. 2016. Isolation of *Cryptosporidium* Oocysts and Sporozoites Using Discontinuous Sucrose and Isopycnic Percoll Gradients Author (s): Michael J . Arrowood and Charles R . Sterling Published by : Allen Press on behalf of The American Society of Parasitologists St. J. Parasitol. 73:314–319.

Bankevich A et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477. doi: 10.1089/cmb.2012.0021.

Bennett GM, Moran NA. 2013. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome Biol. Evol. 5:1675–1688. doi: 10.1093/gbe/evt118.

Bernsel A, Viklund H, Hennerdal A, Elofsson A. 2009. TOPCONS: Consensus prediction of membrane protein topology. Nucleic Acids Res. 37:465–468. doi: 10.1093/nar/gkp363.

Bolger A, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina

sequence data. Bioinformatics. 30:2114.

Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. 2018. Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. bioRxiv. 390013. doi: 10.1101/390013.

Brayton KA et al. 2007. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. 3:1401–1413. doi: 10.1371/journal.ppat.0030148.

Brown AM V. et al. 2018. Comparative Genomics of Wolbachia–Cardinium Dual Endosymbiosis in a Plant-Parasitic Nematode. Front. Microbiol. 9:1–21. doi: 10.3389/fmicb.2018.02482.

Brunner JL, Collins JP. 2009. Testing assumptions of the trade-off theory of the evolution of parasite virulence. Evol. Ecol. Res. 11:1169–1188.

Bushnell B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner.

Callahan BJ et al. 2016. DADA2 : High-resolution sample inference from Illumina amplicon data. Nat. Methods. 1–7. doi: 10.1038/nmeth.3869.

Cardoso R, Soares H, Hemphill A, Leitão A. 2016. Apicomplexans pulling the strings: manipulation of the host cell cytoskeleton dynamics. Parasitology. 1–14. doi: 10.1017/S0031182016000524.

Ciancio A et al. 2008. Redescription of *Cardiosporidium cionae* (Van Gaver and Stephan, 1907) (Apicomplexa: Piroplasmida), a plasmodial parasite of ascidian haemocytes. Eur. J. Protistol. 44:181–196. doi: 10.1016/j.ejop.2007.11.005.

Cusa E et al. 1999. Genetic Analysis of a Chromosomal Region Containing Genes Required for Assimilation of Allantoin Nitrogen and Linked Glyoxylate Metabolism in *Escherichia coli* Genetic Analysis of a Chromosomal Region Containing Genes Required for Assimilation of Allantoin. 181:7479–7484.

Ding M, Clayton C, Soldati D. 2000. *Toxoplasma gondii* catalase: are there peroxisomes in toxoplasma? J. Cell Sci. 113 (Pt 1:2409–2419.

Dong LC et al. 2006. First report on histology and ultrastructure of an intrahemocytic paramyxean parasite (IPP) from tunicate *Halocynthia roretzi* in Korea. Dis. Aquat. Organ. 72:65–69. doi: 10.3354/dao072065.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2:953–971. doi: 10.1038/nprot.2007.131.

English AC et al. 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS One. 7:1–12. doi: 10.1371/journal.pone.0047768.

Escalante a a, Ayala FJ. 1995. Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes. Proc. Natl. Acad. Sci. U. S. A. 92:5793–7. doi: 10.1073/pnas.92.13.5793.

Finn RD et al. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 45:D190–D199. doi: 10.1093/nar/gkw1107.

Frank S a. 1996. Host-symbiont conflict over the mixing of symbiotic lineages.

Proc. Biol. Sci. 263:339–344. doi: 10.1098/rspb.1996.0052.

Frank S a. 1996. Models of parasite virulence. Q. Rev. Biol. 71:37–78. doi: 10.1086/419267.

Frank SA. 1992. A kin selection model for the evolution of virulence. Proc R Soc L. B. 250:195–197.

Gabaldon T. 2010. Peroxisome diversity and evolution. Philos. Trans. R. Soc. B Biol. Sci. 365:765–773. doi: 10.1098/rstb.2009.0240.

Gabaldon T, Ginger ML, Michels PAM. 2016. Peroxisomes in parasitic protists. Mol. Biochem. Parasitol. 209:33–45. doi: 10.1016/j.molbiopara.2016.02.005.

Gardner MJ et al. 2002. Europe PMC Funders Group Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 419:3–9. doi: 10.1038/nature01097.Genome.

Giard MA. 1888. On *Nephromyces*, a new genus of Fungi parasitic in the kidney of the Molgulidæ. Ann. Mag. Nat. Hist. Ser. 6. 1:386–388. doi: 10.1080/00222938809460753.

Goodbody BYI. 1965. Nitrogen Excretion in Ascidiacea. Enzyme. 34:299–305.

Greganova E, Steinmann M, Mäser P, Fankhauser N. 2013. In silico ionomics segregates parasitic from free-living eukaryotes. Genome Biol. Evol. 5:1902–1909. doi: 10.1093/gbe/evt134.

Gruwell ME, Hardy NB, Gullan PJ, Dittmar K. 2010. Evolutionary relationships

among primary endosymbionts of the mealybug subfamily phenacoccinae (Hemiptera: Coccoidea: Pseudococcidae). Appl. Environ. Microbiol. 76:7521–7525. doi: 10.1128/AEM.01354-10.

Guillou L et al. 2013. The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41:597–604. doi: 10.1093/nar/gks1160.

Haas BJ et al. 2014. *reference generation and analysis with Trinity*. doi: 10.1038/nprot.2013.084.De.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491. doi: 10.1186/1471-2105-12-491.

Hua R, Gidda SK, Aranovich A, Mullen RT, Kim PK. 2015. Multiple Domains in PEX16 Mediate Its Trafficking and Recruitment of Peroxisomal Proteins to the ER. Traffic. 16:832–852. doi: 10.1111/tra.12292.

Hyde JE. 2007. Targeting purine and pyrimidine metabolism in human apicomplexan parasites. Curr. Drug Targets. 8:31–47. doi: 10.2174/138945007779315524.

Jackman SD et al. 2017. ABySS 2 . 0 : resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 27:768–777. doi: 10.1101/gr.214346.116.Freely.

Jackson AP. 2015. The evolution of parasite genomes and the origins of

parasitism. Parasitology. 142:S1–S5. doi: 10.1017/S0031182014001516.

Janouškovec J et al. 2015. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. Proc. Natl. Acad. Sci. U. S. A. 112:10200–7. doi: 10.1073/pnas.1423790112.

Janouskovec J, Keeling PJ. 2016. Evolution: Causality and the origin of parasitism. Curr. Biol. 26:R174–R177. doi: 10.1016/j.cub.2015.12.057.

Jung D-K et al. 2006. Structural and functional analysis of PucM, a hydrolase in the ureide pathway and a member of the transthyretin-related protein family. Proc. Natl. Acad. Sci. 103:9790–9795. doi: 10.1073/pnas.0600523103.

Kaasch AJ, Joiner KA. 2000. Targeting and Subcellular Localization of. J. Biol. Chem. 275:1112–1118.

Kada S, Lion S. 2015. Superinfection and the coevolution of parasite virulence and host recovery. J. Evol. Biol. 28:2285–2299. doi: 10.1111/jeb.12753.

Kahn K, Tipton PA. 1998. Spectroscopic characterization of intermediates in the urate oxidase reaction. Biochemistry. 37:11651–11659. doi: 10.1021/bi980446g.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44:D457–D462. doi: 10.1093/nar/gkv1070.

Kappmeyer LS et al. 2012. Comparative genomic analysis and phylogenetic position of *Theileria equi*. BMC Genomics. 13:603. doi: 10.1186/1471-2164-13-603.

Kasai T, Suzuki I, Asai T. 1963. Glyoxylate Oxidation in Acetobacter with Refenence to the formation of Oxalic Acid. Applyed Microbiol. 9:49–58. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1152716&tool=pmcentrez&rendertype=abstract.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780. doi: 10.1093/molbev/mst010.

Keeling PJ. 2004. Reduction and compaction in the genome of the apicomplexan parasite *Cryptosporidium parvum*. Dev. Cell. 6:614–616. doi: 10.1016/S1534-5807(04)00135-2.

Kemp LE, Yamamoto M, Soldati-Favre D. 2013. Subversion of host cellular functions by the apicomplexan parasites. FEMS Microbiol. Rev. 37:607–631. doi: 10.1111/1574-6976.12013.

Kleiner M et al. 2018. Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. Proc. Natl. Acad. Sci. 115:E5576–E5584. doi: 10.1073/pnas.1722325115.

Klindworth A et al. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41:1–11. doi: 10.1093/nar/gks808.

Koren S et al. 2014. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 1–11. doi:

10.1101/gr.215087.116.Freely.

Kornberg HL, Krebs HA. 1957. Synthesis of cell constituents from C2-units by a modified tricarboxylic acid cycle. Nature. 179:988–991. doi: 10.1038/179988a0.

Kumagai A et al. 2011. Soft tunic syndrome in the edible ascidian *Halocynthia roretzi* is caused by a kinetoplastid protist. Dis. Aquat. Organ. 95:153–161. doi: 10.3354/dao02372.

Kuwahara H et al. 2007. Reduced Genome of the Thioautotrophic Intracellular Symbiont in a Deep-Sea Clam, *Calyptogena okutanii*. Curr. Biol. 17:881–886. doi: 10.1016/j.cub.2007.04.039.

Laczny CC et al. 2015. VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 3:1–7. doi: 10.1186/s40168-014-0066-1.

Lalremruata A et al. 2017. Species and genotype diversity of *Plasmodium* in malaria patients from Gabon analysed by next generation sequencing. Malar. J. 16:1–11. doi: 10.1186/s12936-017-2044-0.

Lambert CC, Lambert G, Crundwell G, Kantardjieff K. 1998. Uric acid accumulation in the solitary ascidian *Corella inflata*. J. Exp. Zool. 282:323–331. doi: 10.1002/(SICI)1097-010X(19981015)282:3<323::AID-JEZ5>3.0.CO;2-O.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357. http://dx.doi.org/10.1038/nmeth.1923.

Lee KS et al. 2011. *Plasmodium knowlesi*: Reservoir hosts and tracking the

emergence in humans and macaques. PLoS Pathog. 7. doi: 10.1371/journal.ppat.1002015.

Leung TLF, Poulin R. 2008. Parasitism, commensalism, and mutualism: Exploring the many shades of symbioses. Vie Milieu. 58:107–115.

Li J et al. 1997. Regulation and trafficking of three distinct 18 S ribosomal RNAs during development of the malaria parasite. J. Mol. Biol. 269:203–213. doi: 10.1006/jmbi.1997.1038.

Li W, Godzik A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22:1658–1659. doi: 10.1093/bioinformatics/btl158.

Lige B, Jayabalasingham B, Zhang H, Pypaert M, Coppens I. 2009. 'Role of an Ancestral D-Bifunctional Protein Containing' 'Two Sterol-Carrier Protein-2 Domains in Lipid Uptake and Trafficking in *Toxoplasma*'. Mol. Biol. Cell. 20:658–672.

Lopez-Sanchez MJ et al. 2009. Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. PLoS Genet. 5. doi: 10.1371/journal.pgen.1000721.

Ludewig-Klingner A-K, Michael V, Jarek M, Brinkmann H, Petersen J. 2018. Distribution and Evolution of Peroxisomes in Alveolates (Apicomplexa, Dinoflagellates, Ciliates). Genome Biol. Evol. 10:1–13. doi: 10.1093/gbe/evx250.

Mahler BYHR, The W, Assistamze T, Germille OF. 1955. Studies On Uricase. J. Biol. Chem.

Marin B, Nowack ECM, Melkonian M. 2005. A plastid in the making: Evidence for a second primary endosymbiosis. Protist. 156:425–432. doi: 10.1016/j.protis.2005.09.001.

Mccutcheon JP, Moran NA. 2011. in symbiotic bacteria. Nat. Publ. Gr. 10:13–26. doi: 10.1038/nrmicro2670.

McCutcheon JP, Moran NA. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc. Natl. Acad. Sci. U. S. A. 104:19392–19397. doi: 10.1073/pnas.0708855104.

McFadden B, Howes W. 1965. Crystallization and Some Properties of Isocitrate Lyase from *Pseudomonas indigofera**. J. Biochem. 58:116–122.

Middelhoven WJ, De Hoog GS, Notermans S. 1989. Carbon assimilation and extracellular antigens of some yeast-like fungi. Antonie Van Leeuwenhoek. 55:165–175. doi: 10.1007/BF00404756.

Molina I, Pellicer M, Badia J, Aguilar J, Baldoma L. 1994. Molecular Characterization of *Escherichia coli* Malate Synthase G: Differentiation with the Malate Synthase A Isoenzyme. Eur. J. Biochem. 224:541–548. doi: 10.1111/j.1432-1033.1994.00541.x.

Moog D, Przyborski JM, Maier UG. 2017. Genomic and proteomic evidence for the presence of a peroxisome in the apicomplexan parasite *Toxoplasma gondii* and

other Coccidia. Genome Biol. Evol. 1–39. doi: 10.1093/gbe/evx231/4596564.

Moore RB et al. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. Nature. 452:900–900. doi: 10.1038/nature06871.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. 93:2873–2878. doi: 10.1073/pnas.93.7.2873.

Moran NA, Dunbar HE, Wilcox JL. 2005. Regulation of transcription in a reduced bacterial genome: Nutrient-provisioning genes of the obligate symbiont *Buchnera aphidicola*. J. Bacteriol. 187:4229–4237. doi: 10.1128/JB.187.12.4229-4237.2005.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and Evolution of Heritable Bacterial Symbionts. Annu. Rev. Genet. 42:165–190. doi: 10.1146/annurev.genet.41.110306.130119.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:182–185. doi: 10.1093/nar/gkm321.

Morrison HG et al. 2007. Genomic Minimalism in the Early Diverging Intestinal Parasite *Giardia lamblia*. Science (80-.). 317:1921–1926. doi: 10.1126/science.1143837.

Muñoz A, Raso MJ, Pineda M, Piedras P. 2006. Degradation of ureidoglycolate in French bean (*Phaseolus vulgaris*) is catalysed by a ubiquitous ureidoglycolate urea-lyase. Planta. 224:175–184. doi: 10.1007/s00425-005-0186-8.

Mushegian AA, Ebert D. 2016. Rethinking 'mutualism' in diverse host-symbiont

communities. BioEssays. 38:100–108. doi: 10.1002/bies.201500074.

Nakaia K, Horton P. 1999. PSORT : a program for detecting sorting signals in proteins and predicting their subcellular localization. Comput. CORNER. 0004:1336–1337.

Nishino T, Okamoto K, Eger BT, Pai EF, Nishino T. 2008. Mammalian xanthine oxidoreductase - Mechanism of transition from xanthine dehydrogenase to xanthine oxidase. FEBS J. 275:3278–3289. doi: 10.1111/j.1742-4658.2008.06489.x.

Nolfi JR. 1970. Biosynthesis of uric acid in the tunicate, *Molgula manhattensis*, with a general scheme for the function of stored purines in animals. Comp. Biochem. Physiol. 35:827–842. doi: 10.1016/0010-406X(70)90078-2.

Nowack ECM, Melkonian M. 2010. Endosymbiotic associations within protists. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365:699–712. doi: 10.1098/rstb.2009.0188.

Ogedengbe ME, Qvarnstrom Y, da Silva AJ, Arrowood MJ, Barta JR. 2015. A linear mitochondrial genome of *Cyclospora cayetanensis* (Eimeriidae, Eucoccidiorida, Coccidiasina, Apicomplexa) suggests the ancestral start position within mitochondrial genomes of eimeriid coccidia. Int. J. Parasitol. 45:361–365. doi: 10.1016/j.ijpara.2015.02.006.

Percudani R, Carnevali D, Puggioni V. 2013. Ureidoglycolate hydrolase, amidohydrolase, lyase: How errors in biological databases are incorporated in

scientific papers and vice versa. Database. 2013:1–9. doi: 10.1093/database/bat071.

Petersen G, Cuenca A, Møller IM, Seberg O. 2015. Massive gene loss in mistletoe (*Viscum viscaceae*) mitochondria. Sci. Rep. 5:1–7. doi: 10.1038/srep17588.

Philip LA, Tullman-Ercek, George G. 2006. NIH Public Access. Annu. Rev. Microbiol. 373–395. doi: 10.1146/annurev.micro.60.080805.142212.The.

Potrikus CJ, Breznak JA. 1980. Uric acid-degrading bacteria in guts of termites (*Reticulitermes flavipes* (Kollar)). Appl. Environ. Microbiol. 40:117–124. doi: 16345587.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842. doi: 10.1093/bioinformatics/btq033.

Rao Q et al. 2015. Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly *Bemisia tabaci*. BMC Genomics. 16:1–13. doi: 10.1186/s12864-015-1379-6.

Reid AJ et al. 2012. Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and neospora caninum: Coccidia differing in host range and transmission strategy. PLoS Pathog. 8. doi: 10.1371/journal.ppat.1002567.

Reid AJ. 2014. Large, rapidly evolving gene families are at the forefront of host-parasite interactions in Apicomplexa. Parasitology. 1–14. doi: 10.1017/S0031182014001528.

Roos DS. 2005. Themes and Variations in Apicomplexan Parasite Biology. Science (80-.). 309:72–73. doi: 10.1126/science.1115252.

Rouf MA, Lomprey RF. 1968. Degradation of uric acid by certain aerobic bacteria. J. Bacteriol. 96:617–622.

Sabree ZL, Kambhampati S, Moran N a. 2009. Nitrogen recycling and nutritional provisioning by *Blattabacterium*, the cockroach endosymbiont. Proc. Natl. Acad. Sci. U. S. A. 106:19521–19526. doi: 10.1073/pnas.0907504106.

Saffo M. 1990. Symbiosis within a symbiosis: Intracellular bacteria within the endosymbiotic protist *Nephromyces*. Mar. Biol. 107:291–296. doi: 10.1007/BF01319828.

Saffo M, McCoy A, Rieken C, Slamovits C. 2010. *Nephromyces*, a beneficial apicomplexan symbiont in marine animals. Proc. Natl. Acad. Sci. U. S. A. 107:16190–16195. doi: 10.1073/pnas.1002335107.

Saffo MB. 1982. DISTRIBUTION OF THE ENDOSYMBIONT *NEPHROMYCES* GIARD WITHIN THE ASCIDIAN FAMILY MOLGULIDAE. Biol Bull. 162:95–104.

Saffo MB. 1988. Nitrogen Waste or Nitrogen Source? Urate Degradation in the Renal Sac of Molgulid Tunicates. Biol. Bull. 175:403. doi: 10.2307/1541732.

Saffo MB, Davis WL. 1982. Modes of Infection of the Ascidian *Molgula manhattensis* by Its Endosymbiont *Nephromyces* Giard. Biol. Bull. 162:105. doi: 10.2307/1540974.

Saffo MB, Lowenstam H a. 1978. Calcareous deposits in the renal sac of a molgulid

tunicate. Science. 200:1166–1168. doi: 10.1126/science.200.4346.1166.

Sakharkar KR, Kumar Dhar P, Chow VVTK. 2004. Genome reduction in prokaryotic obligatory intracellular parasites of humans: A comparative analysis. Int. J. Syst. Evol. Microbiol. 54:1937–1941. doi: 10.1099/ijs.0.63090-0.

Schluter A et al. 2006. The evolutionary origin of peroxisomes: An ER-peroxisome connection. Mol. Biol. Evol. 23:838–845. doi: 10.1093/molbev/msj103.

Schlüter A et al. 2006. The evolutionary origin of peroxisomes: An ER-peroxisome connection. Mol. Biol. Evol. 23:838–845. doi: 10.1093/molbev/msj103.

Schlüter A, Real-Chicharro A, Gabaldón T, Sánchez-Jiménez F, Pujol A. 2009. PeroxisomeDB 2.0: An integrative view of the global peroxisomal metabolome. Nucleic Acids Res. 38:800–805. doi: 10.1093/nar/gkp935.

Schultz AC, Nygaard P, Saxild HH. 2001. Functional analysis of 14 genes that constitute the purine catabolic pathway in *Bacillus subtilis* and evidence for a novel regulon controlled by the PucR transcription activator. J. Bacteriol. 183:3293–3302. doi: 10.1128/JB.183.11.3293-3302.2001.

Seah B, Saffo MB, Cavanaugh CM. 2011. A Tripartite Animal-Protist-Bacteria Symbiosis: Culture-Independent and Phylogenetic Characterization. Harvard.

Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 30:2068–2069. doi: 10.1093/bioinformatics/btu153.

Serventi F et al. 2010. Chemical basis of nitrogen recovery through the ureide pathway: Formation and hydrolysis of S-ureidoglycine in plants and bacteria. ACS

Chem. Biol. 5:203–214. doi: 10.1021/cb900248n.

Shin I, Percudani R, Rhee S. 2012. Structural and functional insights into (S)-ureidoglycine aminohydrolase, keyenzyme of purine catabolism in *Arabidopsis thaliana*. J. Biol. Chem. 287:18796–18805. doi: 10.1074/jbc.M111.331819.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212. doi: 10.1093/bioinformatics/btv351.

Small I, Peeters N, Legeai F, Lurin C. 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics. 4:1581–1590. doi: 10.1002/pmic.200300776.

De Souza DJ, Bézier A, Depoix D, Drezen JM, Lenoir A. 2009. *Blochmannia* endosymbionts improve colony growth and immune defence in the ant *Camponotus fellah*. BMC Microbiol. 9:1–8. doi: 10.1186/1471-2180-9-29.

Speijer D. 2011. Oxygen radicals shaping evolution: Why fatty acid catabolism leads to peroxisomes while neurons do without it: FADH2/NADH flux ratios determining mitochondrial radical formation were crucial for the eukaryotic invention of peroxisomes and catabolic tissue. BioEssays. 33:88–94. doi: 10.1002/bies.201000097.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313. doi:

10.1093/bioinformatics/btu033.

Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic Acids Res. 32:309–312. doi: 10.1093/nar/gkh379.

Takada Y, Noguchi T. 1985. Characteristics of alanine: glyoxylate aminotransferase from *Saccharomyces cerevisiae*, a regulatory enzyme in the glyoxylate pathway of glycine and serine biosynthesis from tricarboxylic acid-cycle intermediates. Biochem. J. 231:157–63. doi: 10.1042/bj2310157.

Thong-On A et al. 2012. Isolation and Characterization of Anaerobic Bacteria for Symbiotic Recycling of Uric Acid Nitrogen in the Gut of Various Termites. Microbes Environ. 27:186–192. doi: 10.1264/jsme2.ME11325.

Urakawa H et al. 2005. Hydrothermal vent gastropods from the same family (Provannidae) harbour E- and y -proteobacterial endosymbionts. Environ. Microbiol. 7:750–754. doi: 10.1111/j.1462-2920.2005.00753.x.

Usuda N et al. 1994. Uric acid degrading enzymes, urate oxidase and allantoinase, are associated with different subcellular organelles in frog liver and kidney. J. Cell Sci. 107 (Pt 4:1073–81. http://www.ncbi.nlm.nih.gov/pubmed/8056832.

Wang J, Wang Y, Gao C, Jiang L, Guo D. 2017. PPero, a computational model for plant PTS1 type peroxisomal protein prediction. PLoS One. 12:1–12. doi: 10.1371/journal.pone.0168912.

Wells XE, Lees EM. 1991. Ureidoglycolate amidohydrolase from developing

french bean fruits (Phaseolus vulgaris [L.].). Arch. Biochem. Biophys. 287:151–159. doi: 10.1016/0003-9861(91)90400-D.

Wernegreen JJ. 2015. Endosymbiont evolution: Predictions from theory and surprises from genomes. Ann. N. Y. Acad. Sci. 1360:16–35. doi: 10.1111/nyas.12740.

Wernegreen JJ. 2017. In it for the long haul: evolutionary consequences of persistent endosymbiosis. Curr. Opin. Genet. Dev. 47:83–90. doi: 10.1016/j.gde.2017.08.006.

Werner AK, Romeis T, Witte CP. 2010. Ureide catabolism in *Arabidopsis thaliana* and Escherichia coli. Nat. Chem. Biol. 6:19–21. doi: 10.1038/nchembio.265.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York http://ggplot2.org.

Woo YH et al. 2015. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. Elife. 4:1–41. doi: 10.7554/eLife.06974.

Xi H, Schneider BL, Reitzer L. 2000. Purine catabolism in *Escherichia coli* and function of xanthine dehydrogenase in purine salvage. J. Bacteriol. 182:5332–5341. doi: 10.1128/JB.182.19.5332-5341.2000.

Zarowiecki M, Berriman M. 2015. What helminth genomes have taught us about parasite evolution. Parasitology. 142 Suppl:S85-97. doi: 10.1017/S0031182014001449.

Chapter 3

Comparative transcriptomics of two apicomplexans (*Nephromyces* and *Cardiosporidium cionae)* and the genomes of *Nephromyces* bacterial endosymbionts

by

Christopher Paight[1], Elizabeth S. Hunter[1], Claudio H. Slamovits[2], Mary Beth Saffo[1,3] & Christopher E Lane[1]*

will be submitted to the journal Proceedings of the National Academy of Sciences

[1] Department of Biological Sciences, University of Rhode Island, Kingston RI, 02881, USA.

[2] Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada

[3] Smithsonian National Museum of Natural History, Washington, DC 20560, USA.

CHAPTER 3

Abstract

Parasitism has been referred to as an "evolutionary dead end", because the transition to parasitism is unidirectional. Once an organism becomes an obligate parasite, it is likely to remain an obligate parasite due to the loss of metabolic pathways. *Nephromyces* is a genus in the parasitic phylum Apicomplexa, but has an apparent mutualistic relationship with *Molgula* tunicates. Support for a mutualistic relationship is based largely on a nearly 100% host infection rate with no known clearance of *Nephromyces*. Because transition away from obligate parasite is so rare, little is known about the evolutionary steps involved in such a transition - particularly in lineages with such a long history of obligate parasitism as Apicomplexa. In order to examine this unusual evolutionary transition, we sequenced transcriptomes from *Nephromyces* and its parasitic sister taxon, *Cardiosporidium cionae*, which is an excellent model for what *Nephromyces* might have looked like as a parasite. Both *C. ciona* and *Nephromyces* have tunicate hosts and bacterial endosymbionts, but each maintains a different lifestyle. A comparison of *Nephromyces, C. ciona,* and their endosymbionts will be presented with a focus on system dynamics, relationships, and clues to how this transition occurred.

Introduction

In 2010, Saffo et al. characterized the apicomplexan *Nephromyces* as

having a mutualistic relationship with its host *Molgula* tunicates. The significance of a mutualistic genus within a group of >6000 obligate parasites, some of which cause massive human mortality, prompted us to initiate a genomic investigation of *Nephromyces*. The unique evolutionary pressures caused by an obligate parasitic lifestyle, namely host immune system evasion and the ready abundance of pre-formed host metabolites, lead to predictable patterns of gene family losses and expansions. Typically this consists of the expansion of gene families related to host immune system evasion and other parasitism related functions, and the subsequent loss of many of the core biosynthetic genes due to their presence in the intracellular environment [1–5]. Due to these pressures, obligate parasitism is often an evolutionary dead end.

Besides *Nephromyces*, the phylum Apicomplexa is composed entirely of obligate metazoan parasites. As a result of an estimated 800 million years of evolution as obligate parasites, [6], many of the genomic patterns associated with parasitism have been described from the apicomplexan lineages. Gene expansions can be seen in the plasmodium var protein family, which are involved in host manipulation, evasion and in the expansion of rhoptry, microneme, and dense granule proteins. The list of core biosynthetic pathways lost in apicomplexans includes purine biosynthesis, purine degradation, biosynthesis of many amino acids, and vitamin biosynthesis. These losses make the parasite dependent on the host, not only for primary carbon and nitrogen, but also for any metabolites it can no longer generate by either *de novo* synthesis or by

conversion. High demand on the host for these metabolites to fuel parasite growth increases the cost of infection, thereby increasing virulence. Parasites must maintain a delicate balance between transmission, virulence, and host immune system evasion.

The trade offs in this balance have been described in detail [7–11], but one common solution many parasites adopt is maintaining low relative abundance inside the host. Higher parasite abundance will increase the cost to that host, and increase virulence. If the parasites kill the host before completing their lifecycle or before transmission to a new host, their fitness falls to zero. Similarly, if parasites have a high prevalence in a population and high lethality, they risk decimating their host population. High-sustained infection prevalence is a good indicator of low virulence, and low virulence is often achieved by self-limited reproduction by the parasites. In this way, *Nephromyces* stood out as a very atypical parasite. *Nephromyces* has a nearly 100% infection rate, sustained almost year-round. Unexpectedly, based on typical host / parasite dynamics, *Nephromyces* also reaches very high cell densities. These atypical epidemiological factors were the basis for the Saffo et al. 2010 conclusion that *Nephromyces* must be mutualistic. In order to reach high cell densities while maintaining low virulence, *Nephromyces* was predicted to produce something of high value to the host, to offset the cost associated with maintaining such high densities of an obligate parasite.

The unusual epidemiology of *Nephromyces* becomes more apparent when

contrasted with its parasitic sister taxon, *Cardiosporidium cionae.*

Cardiosporidium, first described in 1907 by Van Gaver and Stephan, and later described by Ciancio et al 2008 is a blood parasite found in solitary non-molgulidae ascidian hosts, including *Ciona intestinalis*. *Cardiosporidium* quickly reaches and maintains ~95% infection prevalence by late July. In contrast to *Nephromyces, Cardiosporidium* cell densities remain low, with orders of magnitude difference in cell densities (based on DNA extraction quantities as a proxy for cell density). Virulence in both of these apicomplexans is thought to be low based on histological work by Ciancio et al. 2008 and Saffo and Nelson 1982. Low virulence is also predicted from the high-sustained infection prevalence. The contrast in cell densities between *Nephromyces* and *Cardiosporidium,* along with the lack of apparent virulence, indicates an unusual relationship between *Nephromyces* and its host.

In addition to being sister taxa, *Nephromyces* and *Cardiosporidium* share a number of other traits. Both organisms are monoxenous, with ascidians as the only host, both have infective stages that are transmitted through seawater, localize within the pericardium of the host and each harbor a monophyletic bacterial endosymbiont species that has been maintained since *Nephromyces* and *Cardiosporidium* diverged (Figure 3). These similarities make *Cardiosporidium* an ideal organism to compare with *Nephromyces* in order to resolve the genomic changes taking place behind the transition from obligate parasitism to a mutualistic host-symbiont relationship.

However, there are some key differences, besides the epidemiological factors, between *Nephromyces and Cardiosporidium*, including host species. *Cardiosporidium* infects several genera of tunicates including *Ciona, Halocynthia, Styela, Ascidiella* and possibly others [12, 13], while *Nephromyces* is restricted to the Molgulidae family of tunicates [14]. Interestingly, *Cardiosporidium* has not been found in any Molgulidae tunicates. Another key difference is that *Cardiosporidium* is an intracellular blood parasite, while *Nephromyces* is extracellular (another unusual trait for an apicomplexan). Additionally, *Nephromyces* is exclusively found in a Molgulidae specific structure called the renal sac, a large ductless structure of unknown function [15, 16].

Despite its name, the renal sac does not seem to function as a typical renal organ, but was named for the large deposits of crystallized uric acid and calcium oxalate within it. Many ascidians have localized deposits of uric acid, but tunicates in the Molgulidae family have the largest [17, 18]. While the function of these deposits in the tunicate remain unclear, previous work demonstrated that *Nephromyces* is able to degrade uric acid because it retains the ancestral purine degradation genes lost in all other apicomplexans (Chapter 2). Transcriptome data and pathway analysis suggest that uric acid may be the primary source of carbon and nitrogen for *Nephromyces* (Chapter 2). Uric acid is an atypical source of carbon and nitrogen, but it is not unheard of. There are several species of bacteria and fungi which can be cultured on media containing only uric acid [19–21]. Due to their unusual environment inside the renal sac and because

Nephromyces is extracellular, this organism may not have access to all the required pre-formed metabolites. However, the genus *Nephromyces* is reported to maintain three different bacterial endosymbionts [22].

In addition to the monophyletic alphaproteobacteria, *Nephromyces* also harbors two other bacterial endosymbionts: a Betaproteobacteria and a Bacteroidetes. Acquisition and maintenance of bacterial endosymbionts is a common way for eukaryotes to gain new metabolic pathways and capabilities. The functional capabilities of bacterial endosymbionts exploited by eukaryotic hosts, for example include amino acid metabolism and vitamin metabolism [23], nitrogen metabolism [24], defense [25], chemotrophic energy production [26], and photosynthesis [27], to name a few. While bacterial endosymbionts are common in many protist lineages, they are rare in the phylum Apicomplexa. The only known apicomplexans to contain bacterial endosymbionts are *Cardiosporidium* and *Nephromyces*. This limited distribution to apicomplexans with ascidian hosts may be due to an unknown aspect of ascidian biology. Previous speculation that *Nephromyces'* bacterial endosymbionts are responsible for the observed high levels of purine degradation have recently been rejected (Chapter 2). However, the bacterial endosymbionts are likely instrumental to *Nephromyces'* ability to colonize the renal sac.

In order to examine the claim of mutualism, characterize the relationships involved in this tripartite endosymbiosis, and determine how *Nephromyces* achieves low virulence with high cell density, we sequenced the transcriptomes

of *Molgula manhattensis, Ciona intestinalis, Nephromyces, Cardiosporidium,* and their bacterial endosymbionts. Additionally, to better understand the dynamics of the renal sac and the interplay between *Nephromyces* and its bacterial endosymbionts, we sequenced and partially assembled the *Nephromyces* genome and the genomes of all three types of the bacterial endosymbionts.

Results

RNA Sequencing

The contents of a single renal sac from an individual *Molgula manhattensis* yielded 32 ng/µl RNA by Qubit and resulted in 195,694 transcripts from *M. manhattensis*, *Nephromyces*, and the bacterial endosymbionts. After binning by species, 60,223 transcripts were attributed to *Nephromyces*. The cardiac fluid from 40 infected *Ciona intestinalis* individuals resulted in 109,446 transcripts, including 15,541 *Cardiosporidium* transcripts.

The large number of transcripts identified from *Nephromyces* was due to multiple species infection of a single host. Clustering sequences together resulted in 26938 transcripts at 90%, 23850 at 80%, 21762 at 70%, 19540 at 60%, 16668 at 50%. Due to the multi species nature of *Nephromyces* infections, the transcriptome is a pan-genome assembly rather than a precise uni-species dataset, but we estimate that there are between 8000 and 12000 unique transcripts in *Nephromyces*. Kyoto Encyclopedia of Genes and Genomes (KEGG) functionally predicts 13336 transcripts in the full dataset. The tool BUSCO was

used to assess the completeness of *Nephromyces* transcriptome resulting in 81.8% complete transcripts and 6.3% partial.

RNA extractions for *Cardiosporidium* samples yielded 164 ng/µl for non-sucrose gradient separated blood, 48 ng/µl from cells taken from the 25% layer, and 24 ng/µl from cells taken from the 30% layer. These resulted in 97,417,356 reads from the non-sucrose separated sample, 115,085,369 reads from cells at the 25% layer, and 116,393,114 reads from the 30% layer. Separated by species, 3877 transcripts were from *C. intestinalis*, 16,663 transcripts were from *Cardiosporidium*, 1,689 transcripts were from the bacterial endosymbiont of *Cardiosporidium*. KEGG functionally predicts 9,775 total transcripts for *Cardiosporidium*, and BUSCO analysis for reports 69.7% complete and 11.9% partial coverage. BUSCO analysis for the bacterial endosymbiont transcriptome resulted in 14.8% complete and 19.6% partial against bacterial_od9.

Nephromyces genome

The *Nephromyces* genome assembly remains highly fragmented and consists of 1176 contigs greater than 5kb with a maximum length of 287,191 bp and an average length of 36 kb.

α-proteobacteria genome (Nαe)

Two different alphaproteobacteria endosymbionts were recovered from our genomic data and assembled into a draft genome. The presence of two closely related alpha proteobacteria genomes both with high AT bias (25% GC content)

and regions of low complexity have limited our ability to assemble these genomes completely (Table 4). The two genomes assemble into 11 contigs ranging in size from 13 kb to 312 kb for a combined length of 995,540 and an average of 90,503. Based on transcriptome sequencing we estimate that the genome is largely complete. The draft genome contains 844 predicted coding sequences, 35 tRNAs matching all codons, and 4 rRNAs. Of the 546 predicted genes have KASS annotations (119 Genetic information processing, 32 Carbohydrate metabolism, 30 Energy metabolism, 29 Cellular processes, 26 Nucleotide metabolism, 26 Metabolism of cofactors and vitamins, 25 environmental information processing, 15 lipid metabolism, 12 amino acid metabolism, 15 unclassified).

Bacteroidetes Genome (Nbe)

Nephromyces bacteroidetes genome is circular, 494,352 nucleotides long, and extremely AT rich (22% GC content) (Table 4). The genome contains 503 predicted genes, 31 tRNAs predicted to recognize all codons, and 4 rRNAs. 391 of the predicted genes have KASS annotations (110 Genetic information processing, 40 Carbohydrate metabolism, 38 Energy metabolism, 31 amino acid metabolism, 21 Metabolism of cofactors and vitamins, 17 Nucleotide metabolism, 11 unclassified, 10 Lipid metabolism, 9 Cellular processes).

Betaproteobacteria Genome (Nβe)

The Betaproteobacteria genome is circular and 866,396 bp long with 30% GC content (Table 4). It contains 880 predicted genes, 40 tRNAs, and 4 rRNAs

	Nephromyces's Alphaproteobacterial genome (Nαe)	*Nephromyces's* Bacteroidetes Genome (NBe)	*Nephromyces's* Betaproteobacterial Genome (Nβe)
contigs	11	1	1
size in BP	995,540	494,352	866,396
CDS	844	503	880
tRNA	35	31	40
rRNA	4	4	4
GC content	25%	22%	30%

Table 4) Genomic assembly statistics for *Nephromyces* three bacterial endosymbionts. Nαe and Nbe were assembled from a hybrid assembly of Illumina HiSeq and Pacific Biosciences data, while Nβe was assembled with Pacific Biosciences. Sequence data contained multiple closely related strains and these assemblies represent pan-genomes of the various symbionts.

(two identical 16s copies and 2 identical 23s copies). 753 of the 880 predicted genes have KAAS annotations (156 Genetic information processing, 61 Carbohydrate metabolism, 47 Energy metabolism, 11 Cellular processes, 45 Nucleotide metabolism, 68 Metabolism of cofactors and vitamins, 39 environmental information processing, 18 lipid metabolism, 62 amino acid metabolism, 14 unclassified).

Bacterial Genome BUSCO Analysis

BUSCO analysis against bacteria_odb9 for the alphaproteobacteria endosymbionts resulted in 99 complete BUSCO copies, 97 single copy and 2 duplicate, and 5 fragmented copies, and is predicted to be 66.9% complete and 3.4% incomplete. For the bacteroidetes endosymbiont BUSCO returned 88 complete single copy, no duplicates, and 10 fragmented for a 59.5% complete and 6.8% incomplete. The Betaproteobacteria endosymbiont had 116 complete BUSCO copies, all single copy (78.4%), 15 incomplete (10.1%), and 17 missing (11.5%).

Nephromyces and *Cardiosporidium* metabolic pathway characterization

Endocytosis

Both *Nephromyces/Cardiosporidium* are predicted to encode genes for clathrin-dependant endocytosis. An additional 25 genes related to endocytosis

are predicted in *Nephromyces/Cardiosporidium* than in *P. falciparum.* Many of the additional genes found in *Nephromyces/Cardiosporidium* are in the VPS and CHMP protein families and form part of the ESCRT machinery important in the biogenesis of multivesicular bodies. Multivesicular bodies transport ubiquitinated proteins to lysosomes for degradation. Other proteins not found in *P. falciparum* include a number of genes in the AP2 complex, which are accessory proteins in clathrin-mediated endocytosis. The AP2 complex plays an important role in the regulation of the assembly of clathrin-coated vesicles.

Carbohydrate Metabolism

Basic carbon metabolism in *Nephromyces/Cardiosporidium* is similar to other apicomplexans and encodes the complete pathways for the citrate acid cycle, glycolysis, gluconeogenesis, and the pentose phosphate pathway. Interestingly, *Nephromyces/Cardiosporidium* are predicted to encode far more genes involved with inositol phosphate metabolism than either *P. falciparum* or *T. gondii.* Despite the absence of genes involved in the synthesis of myo-inositol in *P. falciparum* or *T. gondii* there is support that myo-inositol is used in intracellular calcium signaling in these two organisms. How *P. falciparum* or *T. gondii* are able to use myo-inositol without being able to synthesize it is unclear, but it is presumed that they have divergent and unrecognizable myo-inositol biosynthesis genes. However, both *Nephromyces/Cardiosporidium* are predicted to be able to synthesize myo-inositol and these genes are readily identifiable as orthologous to myo-inositol biosynthesis genes in other organisms. Both

Nephromyces/Cardiosporidium have a copy of serine-pyruvate aminotransferase (AGXT), which catalyzes glyoxylate to glycine and pyruvate. Additionally, *Nephromyces* encodes malate synthase, which, in conjunction with acetyl-CoA, forms malate from glyoxylate.

Fatty Acid Metabolism

Nephromyces/Cardiosporidium are predicted to be able to perform fatty acid initiation and elongation in both the mitochondrial and cytoplasmic pathways, as well as elongation in the endoplasmic reticulum. Additionally, *Nephromyces/Cardiosporidium* are predicted to encode D-glycerate 3-kinase (GLYK), aldehyde dehydrogenase (ALDH), alcohol dehydrogenase (ADH), glycerol kinase (glpK), glycerol-3-phosphate O-acyltransferase (GPAT1), and 1-acyl-sn-glycerol-3-phosphate acyltransferase (plsC), and thus are able create triglycerides from glucose.

Nucleic acid metabolism

Unique among apicomplexans, *Nephromyces/Cardiosporidium* contain a complete pathway for the biosynthesis of inosine monophosphate (IMP). IMP is a purine and the starting molecule for the biogenesis of guanine and adenine. *De novo* biosynthesis of purines has been lost in all sequenced apicomplexans. Other apicomplexans are capable of scavenging precursor molecules to IMP and converting them into IMP, but *Nephromyces/Cardiosporidium* encode the entire IMP biosynthesis pathway from 5-Phosphoribosyl diphosphate (PPRP). The genes involved in this pathway include amidophosphoribosyltransferase (purF),

phosphoribosylamine---glycine ligase (purD), phosphoribosylformylglycinamidine synthase (purL), phosphoribosylformylglycinamidine cyclo-ligase (purM), phosphoribosylaminoimidazole carboxylase (PAICS), phosphoribosylaminoimidazole-succinocarboxamide synthase (purC), adenylosuccinate lyase (purB), IMP cyclohydrolase (purH). From IMP both *Nephromyces/Cardiosporidium* have the metabolic machinery for the biosynthesis of adenine and guanine. In addition to the biosynthesis of purines, *Nephromyces/Cardiosporidium* are also the only known apicomplexans capable of purine degradation (Chapter 2).

Biosynthesis of Amino Acids

Nephromyces is predicted to be able to synthesis 10 amino acids (alanine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, methionine, serine, and threonine). *Cardiosporidium* is predicted to be able to synthesis the same amino acids with the exception of cysteine. In addition to the complete pathways *Nephromyces/Cardiosporidium* have partial pathways for synthesis of phenylalanine from phenylpyruvate and can convert tyrosine from phenylalanine. Both *Nephromyces/Cardiosporidium* encode branched-chain amino acid aminotransferase, which adds the final amine group to valine, leucine, and isoleucine.

Vitamin and cofactor synthesis

Nephromyces has the predicted biosynthetic capabilities to produce

riboflavin, acetyl CoA, nicotinate, folate, retinol, vitamin E, heme, and ubiquinone. *Cardiosporidium* has similar predicted biosynthetic capabilities, but is only capable of synthesizing 6-Geranylgeranyl-2,3-dimethylbenzene-1,4-diol and lacks the final two enzymes in the production of vitamin E. Both *Nephromyces/Cardiosporidium* encode a copy of lipoyl synthase (lipA), but lack lipoyl(octanoyl) transferase (lipB) in the lipoic acid synthesis pathway. Lipoic acid is an essential cofactor involved in the citric acid cycle, pyruvate dehydrogenase complex, 2-oxoglutarate dehydrogenase complex, branched-chain oxoacid dehydrogenase, and acetoin dehydrogenase.

Endosymbiont pathways (Figure 3)

Carbohydrate metabolism

The α-proteobacteria endosymbiont of *Nephromyces* (Nαe) has an extremely reduced carbohydrate metabolism, including all of the genes involved with gluconeogenesis and glycolysis. The only carbohydrate metabolism genes present are a complete citrate acid cycle and pyruvate dehydrogenase E1 component (aceE), dihydrolipoamide dehydrogenase (pdhD), and pyruvate dehydrogenase E2 component (aceF), which converts pyruvate to acetyl-CoA. With such severe reduction in carbohydrate metabolism, pyruvate appears to be the only carbon source the alphaproteobacteria is capable of processing.

The bacteroidetes endosymbiont (Nbe) has a similarly reduced carbohydrate metabolism as in the α-proteobacteria, however the reduction is

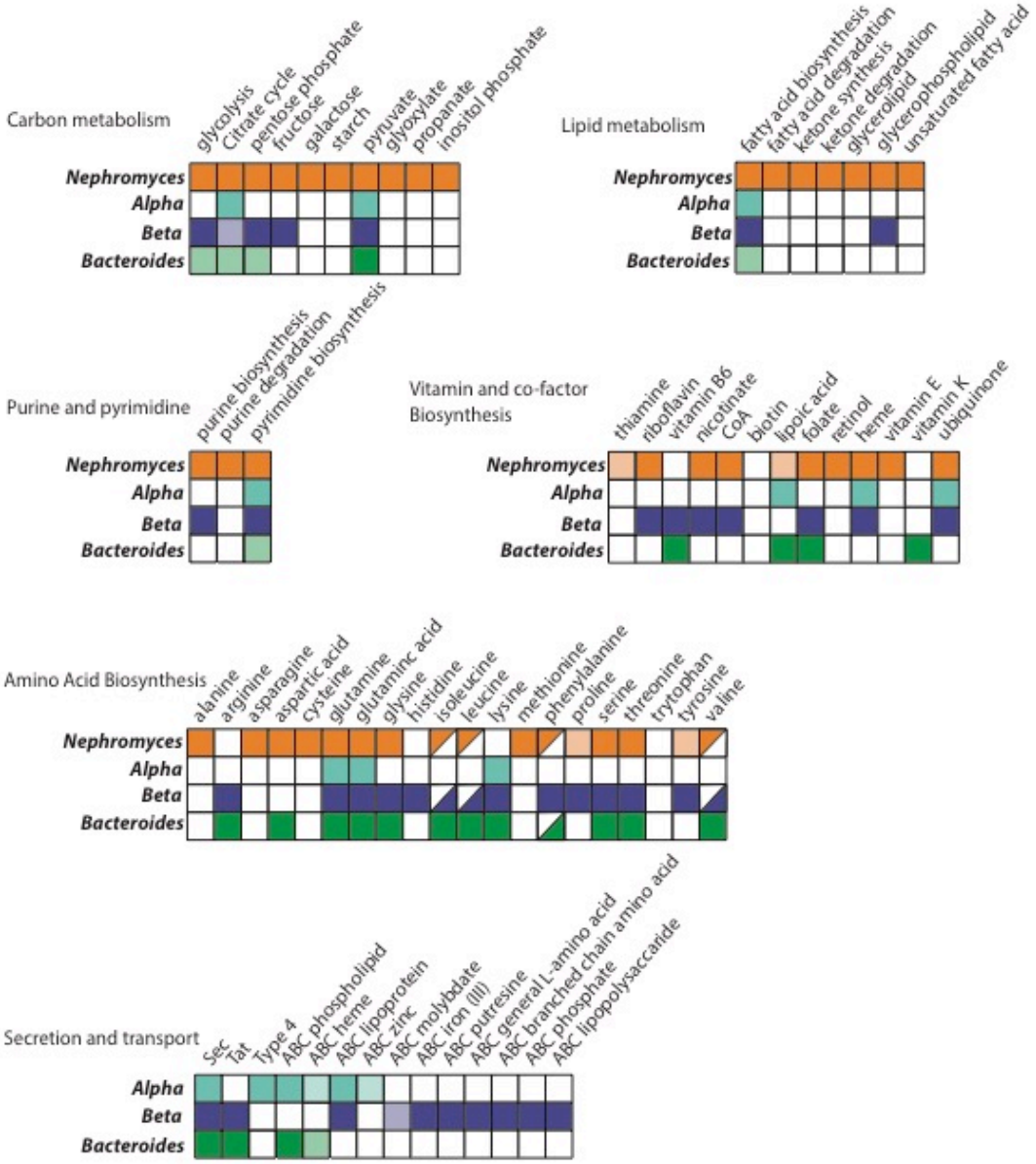


Figure 3) Metabolic pathway capabilities of *Nephromyces* (orange) and its bacterial endosymbionts (alpha=teal, beta=purple, bacteroidetes=green) solid colored boxes indicated a complete pathway, light shaded boxes indicated a partial pathway, and white boxes indicate the pathway is not present.

not as extreme. Having lost gluconeogenesis the bacteroidetes endosymbiont can process fructose into phosphoenolpyruvate and contains pyruvate dehydrogenase E1 component (aceE), dihydrolipoamide dehydrogenase (pdhD), and pyruvate dehydrogenase E2 component (aceF) to convert pyruvate into acetyl-CoA. While the full citrate cycle is incomplete the partial cycle from 2-oxoglutarate to oxaloacetate is complete, as well as the reductive pentose phosphate pathway from glyceraldehyde-3P to ribulose-5P.

Nephromyces β-proteobacteria endosymbiont (Nβe) has the most complete carbohydrate metabolism encoding the complete non-oxidative pentose phosphate pathway, the citrate cycle from 2-oxoglutarate to oxaloacetate, and the core glycolysis module involving three carbon compounds.

Fatty Acid metabolism

Paradoxically, while the enzymes involved with fatty acid biosynthesis (from malonyl-CoA in Nαe and Nβe and from acetyl-CoA in Nbe) are present in all three types of bacterial endosymbiont, all the genes involved in fatty acid degradation have been lost in every symbiont. Nαe and Nβe have a reduced glycerophospholipid metabolism and must convert phosphatidate to synthesis phosphatidylethanolamine, phosphatidylglycerol, and phosphatidylserine, while Nβe is able to *de novo* synthesize glycerophospholipids from glycerone.

Nucleic acid metabolism

Purine metabolism is similarly reduced in both Nαe and Nbe. Both endosymbionts lack all genes in the IMP biosynthesis pathway, as well as the ability to convert IMP to guanine or adenine. With so few purine biosynthesis capabilities, all of the guanine and adenine for DNA replication must be obtained as preformed nucleobases. Nβe purine biosynthesis is complete encoding adenylosuccinate lyase (purB), and IMP cyclohydrolase (purH) and is able to synthesize IMP from 5-Phosphoribosyl diphosphate (PRPP) (with the histidine synthesis pathway) as well as the genes required to convert IMP to both guanine and adenine. However, none of the three endosymbionts encode any genes involved in purine degradation.

Both Nαe and Nβe encode the necessary genes for pyrimidine biosynthesis. However, because Nbe is missing several genes involved in pyrimidine biosynthesis it seems likely that Nbe is dependent on *Nephromyces* for both purines and pyrimidines

Biosynthesis of Amino Acids

Nαe is only capable of synthesizing the amino acids glutamine, glutamic acid, and lysine. Of these three amino acids, lysine is the only amino acid that *Nephromyces* is unable to synthesize.

Nbe is capable of synthesizing 11 amino acids: arginine, aspartic acid, glutamine, glutamic acid, glycine, isoleucine, leucine, lysine, serine, threonine, and

valine. Bacteroidetes is also able to synthesis phenylpyruvate, but lacks the ability to the ability to synthesis phenylalanine. Biosynthesis of arginine, isoleucine, leucine, lysine, and valine are not present in the *Nephromyces* transcriptome and may represent the bacteroidetes contribution. Additionally, Bacteroidetes synthesis of phenylpyruvate, but inability to synthesis phenylalanine compliments *Nephromyces* synthesis of phenylalanine from phenylpyruvate, but inability to synthesis phenylpyruvate.

Nβe encodes the genes for 11 amino acids: arginine, glutamine, glutamic acid, glycine, histidine, lysine, phenylalanine, proline, serine, threonine, and tyrosine. Nβe also encodes all the genes for synthesis of isoleucine, leucine, and valine, but lacks the last gene in the pathway branched-chain amino acid aminotransferase. However, *Nephromyces* encodes branched-chain amino acid aminotransferase and may be able to complete isoleucine, leucine, and valine by adding the final amine group.

Vitamin and cofactor synthesis

Nαe only encodes genes for the biosynthesis three vitamins and co-factors; heme, ubiquinone, and lipoic acid. Lipoic acid has been experimentally shown to be exclusively synthesized in the apicoplast in apicomplexans; additionally all lipoic acid used in the mitochondria needs to be scavenged from the host. Lipoic acid may be an important product produced by the alphaproteobacteria. In contrast Nbe is capable of biosynthesis of vitamin B6,

lipoic acid, folate and vitamin K. Nβe is able to synthesize most cofactors including riboflavin, vitamin B6, nicotinate, coenzyme-A, folate, heme, and ubiquinone.

Secretion and Transporters

All three symbionts encode at least a partial bacterial sec secretion system. Nβe is the most complete missing only secM. Nαe lacks secE, secG, and secM. Nbe is the most incomplete missing secB, secD, secE, secM with so many genes missing it is not clear if the sec secretion system is functional in Nbe. Both Nβe and Nbe have the twin-arginine translocation pathway. Nαe has TatC, but lacks TatA and is therefore incomplete [28]. While Nαe is missing the Tat transport system, Nαe does encode the type 4 bacterial secretion system.

In addition to the more general secretion systems there are also more specific ABC transporters. Nbe has the fewest ABC transporters only phospholipid and possibly a heme transporter. Nαe also encodes a phospholipid and heme transporter in addition to a lipoprotein transporter and possibly a zinc transporter. Nβe's genome contain the most ABC transporters, including Iron(III), putracine, General L-amino acid, branched-chain amino acid, phosphate, lipoprotein, lipopolysaccharide, and possibly a molybdate transporter.

Bacterial phylogeny

Nephromyces and *Cardiosporidium* alphaproteobacteria endosymbionts

are monophyletic (98 bootstrap support) and are in the genus *Rickettsia* (94 bootstrap support). *Nephromyces* bacteroidetes endosymbiont is sister to the genus *Pedobacter* (87 bootstrap support) in the family Sphingobacteriaceae. The Betaproteobacteria fall within the genus *Bordetella* (98 bootstrap support) (Figure 4).

Orthology

Nephromyces had 21,762 genes when clustered at 70% of these 20881 were assigned to orthogroups. 39.7% of orthogroups made contained *Nephromyces* and there were eight genus specific orthogroups containing 21 genes (Figure 5). 3,455 orthogroups were shared by *Nephromyces*, Apicomplexa, and Chromerids. 421 orthogroups were shared between *Nephromyces* and Apicomplexa (Figure 6). 218 orthogroups were shared between *Nephromyces* and Chromerids that were not found in Apicomplexa. *Cardiosporidium* had 7,395 genes, of which 6,977 were assigned to orthogroups. 31.5% of orthogroups contained *Cardiosporidium* and there was one species-specific orthogroup containing five genes. 2,778 orthogroups were shared between *Cardiosporidium*, Apicomplexa, and Chromerids. 236 orthogroups were shared between *Cardiosporidium* and Apicomplexa and not found in Chromerid. 219 orthogroups were shared between *Cardiosporidium* and Chromerids that were not found in Apicomplexa. *Nephromyces* and *Cardiosporidium* had 3, 106 shared orthogroups. *Nephromyces* had 1,178 orthogroups not found in *Cardiosporidium.* 289

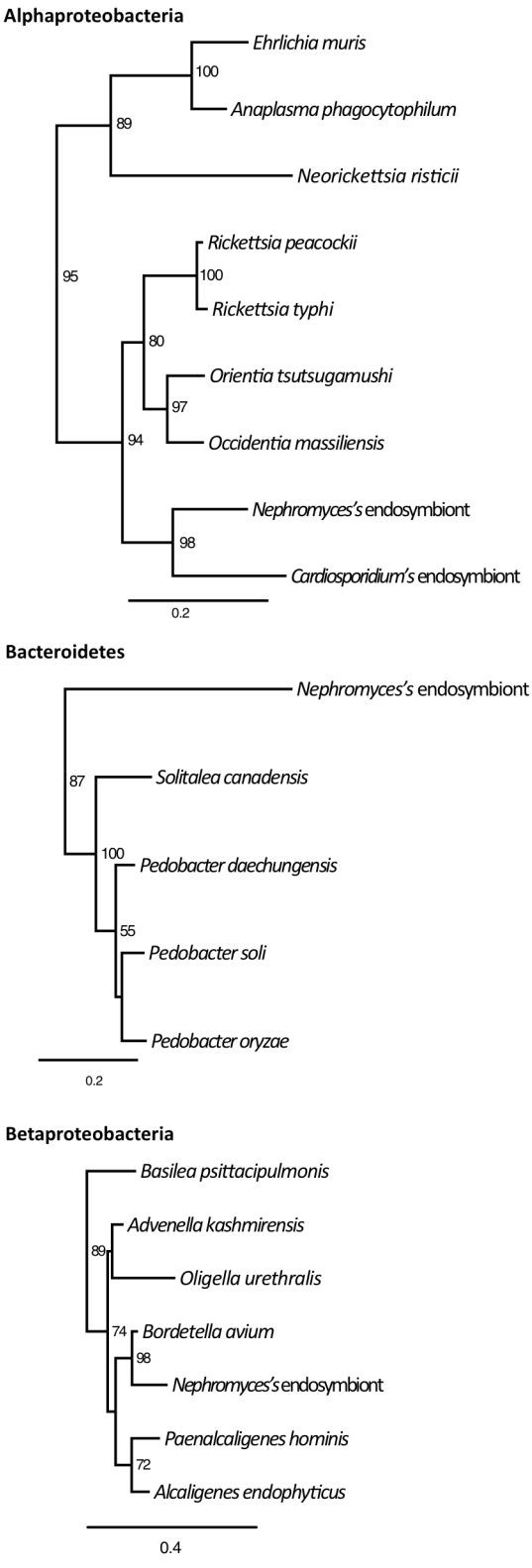


Figure 4) Subsets from larger maximum liklihood16s rRNA trees composed of Alphaproteobacteria, Bacteroidetes, and Betaproteobacteria sequences. Nodes are labeled with percentage bootstrap support values. *Nephromyces* and *Cardiosporidium* alphaproteobacteria endosymbionts are monophyletic and sister to the family rickettsiaceae. *Nephromyces* bacteroidetes endosymbiont is in the family sphingobacteriaceae. *Nephromyces* betaproteobacteria endosymbiont is sister to the genus *Bordetella*.

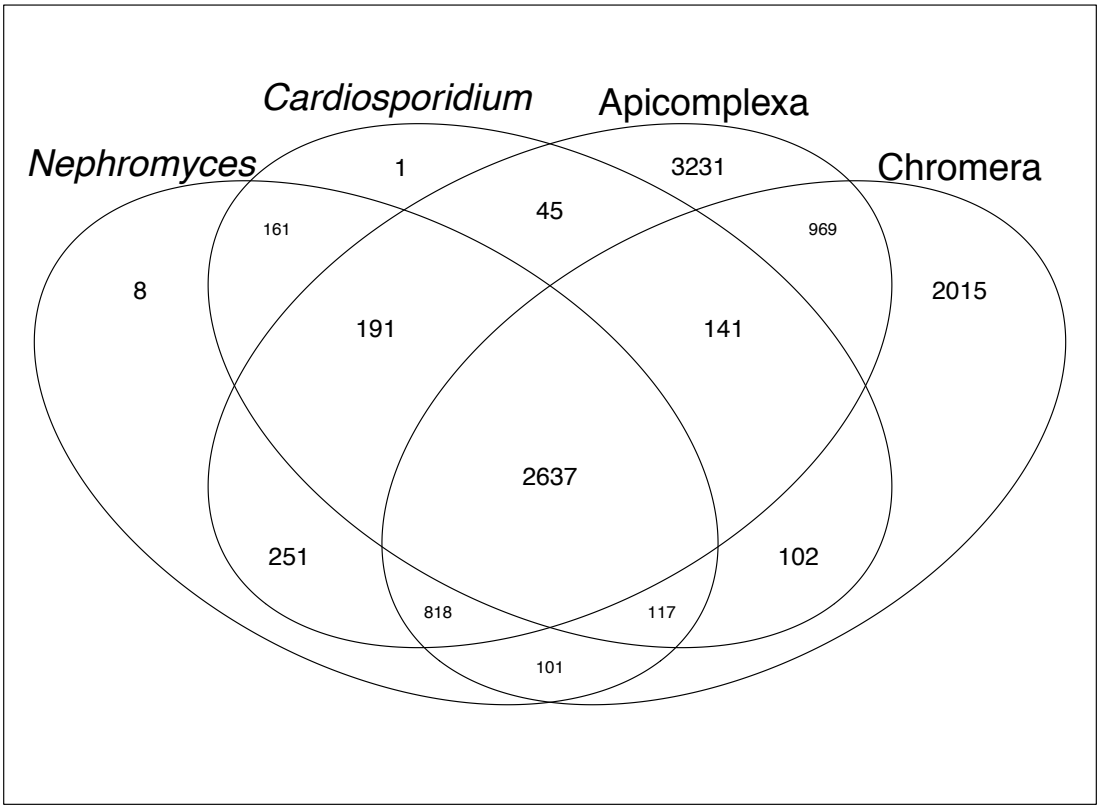


Figure 5) Collapsed Venn Diagram of orthologous gene groups from *Nephromyces* (70% identity level), *Cardiosporidium,* Chromera (*V. brassicaformis, C. velia),* and Apicomplexa (*C. parvum*, *G. niphandrodes*, *B. bovis*, *T. parva*, *P. falciparum*, *C. cayetanensis*, *E. brunetti*, *E. falciformis*, *E. tenella*, *H. hammondi*, *N. caninum*, *S. neurona*, *T. gondii).* Orthology was predicted with OrthoFinder.

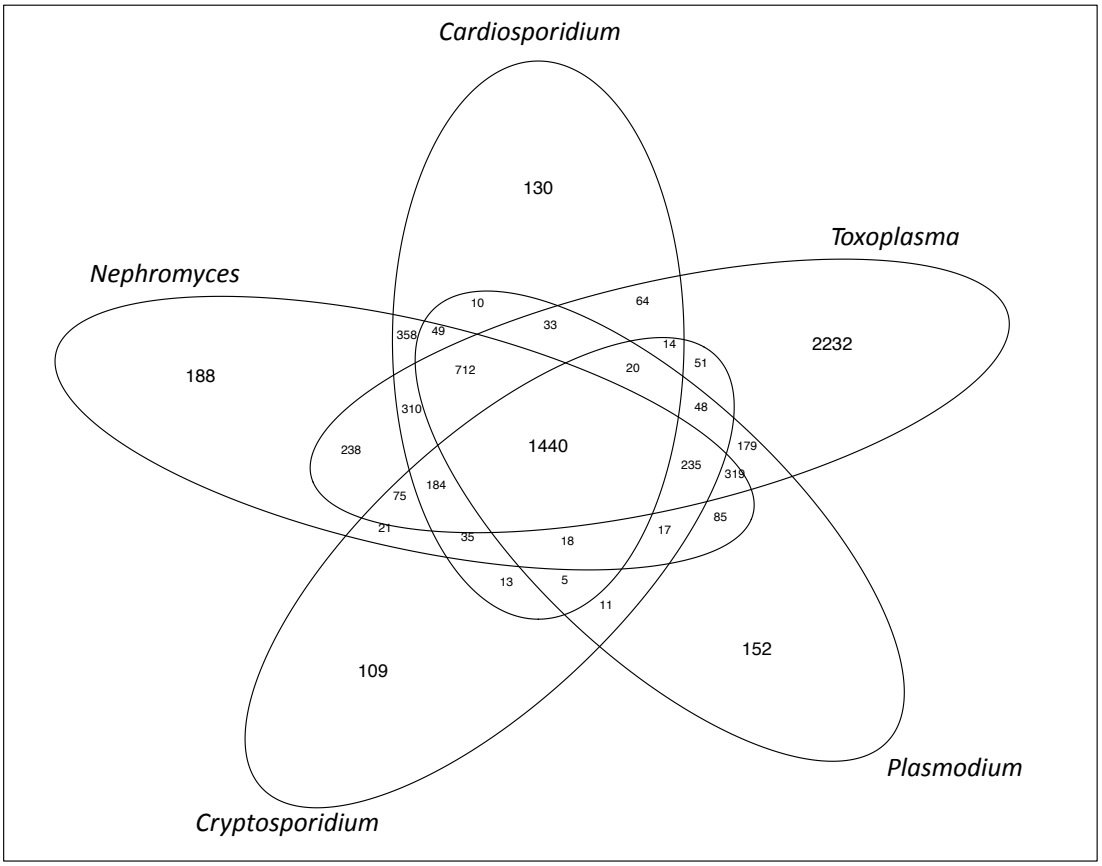


Figure 6) Venn Diagram of orthologous gene groups from *Nephromyces* (70% identity level), *Cardiosporidium*, *Toxoplasma*, *Cryptosporidium* and *Plasmodium*. Orthology was predicted with OrthoFinder.

orthogroups were found in *Cardiosporidium* and not in *Nephromyces*.

Discussion

Our initial aim in sequencing and comparing the transcriptomes of *Nephromyces/Cardiosporidium* was to better characterize mutualism in *Nephromyces*. A mutualistic relationship was described based on the unusual epidemiology *Nephromyces*, but high infection prevalence is not proof of mutual benefit. While our current efforts do not conclusively characterize *Nephromyces'* relation to its host, the data do provide insight into *Nephromyces'* atypical lifestyle. By comparing *Nephromyces* (a proposed mutualist) to *Cardiosporidium* (a blood parasite) we are able to extricate the evolutionary effects of a changing relationship from the evolutionary effects of an ascidian host.

Since all of the sequencing on *Nephromyces* has been on samples containing multiple *Nephromyces* species, our *Nephromyces* data is therefore a pan-transcriptome/genome. This approach limits our results and conclusions. This is particularly evident in our genomic assemblies, which are highly fragmented and almost certainly poly-species chimeric. In addition to problems with assembling the genomes of closely related organisms, we are also unable to estimate gene family expansions or reductions in *Nephromyces*. While these limitations are significant, the pan-transcriptome/genome does reflect the natural biology of *Nephromyces*. All of the renal sacs sampled from the host *M. manhattensis* to date have contained multiple *Nephromyces* infections. Efforts to

culture single species isolates in the lab have met with limited success. This indicates that sustained *Nephromyces* infection is dependent on contributions from the community of species and endosymbionts.

Notably, by sequencing multiple *Nephromyces* species we were able to recover the transcriptomes/genomes of all three of *Nephromyces* endosymbionts. Recovering multiple bacterial endosymbiont types provides key insights into how this system works, and outweighs the disadvantages of a pan-transcriptome/genome approach.

The transcriptomes for *Nephromyces/Cardiosporidium* are largely complete (estimated by BUSCO), but *Cardiosporidium* is estimated to be about 10% less complete than *Nephromyces*. We have taken this difference into account when comparing these two data sets. Both *Nephromyces/Cardiosporidium* encode an estimated eight thousand genes, which is a large number of genes for an apicomplexan. This estimate is similar to the number of genes in the most gene rich apicomplexan, *Toxoplasma,* which also encodes eight thousand genes. This is interesting given the phylogenetic placement of *Nephromyces/Cardiosporidium* in the hematozoa. Hematozoa, which contains plasmodiidae and piroplasmida lineages, have some of the smallest genomes with the least number of genes of any sequenced apicomplexans. This high gene number in both *Nephromyces/Cardiosporidium* may indicate that greater biosynthesis and metabolic capabilities are necessary for living in an ascidian host.

The two most striking observations made over the course of this

exploration involve purine metabolism. The first is purine degradation; both *Nephromyces/Cardiosporidium* have the metabolic capabilities to convert xanthine into glyoxylate. Glyoxylate can then be converted, with serine-pyruvate aminotransferase (AGXT), into glycine and pyruvate; *Nephromyces* additionally encodes malate synthase (MLS), which combines Glyoxylate and acetyl-CoA, into malate. This is the proposed primary route of carbon, nitrogen, and energy acquisition for *Nephromyces* (Chapter 2). This pathway is absent in all other sequenced apicomplexans, but it appears the enzymes in this pathway were retained from the last common ancestor of Apicomplexa, and not a more recent horizontal gene transfer.

Nephromyces/Cardiosporidium also purine metabolism is *de novo* purine biosynthesis. While some apicomplexan lineages have one or two genes to synthesize inosine monophosphate (IMP) from immediate precursors, *Nephromyces/Cardiosporidium* are predicted to encode the entire *de novo* purine synthesis from 5-Phosphoribosyl diphosphate (PRPP). As the inability to synthesize purines has been widely targeted for drug development against other apicomplexan species, its presence in *Nephromyces/Cardiosporidium* is surprising [29].

The presence of both purine degradation and purine synthesis could be critical to *Nephromyces'* unusual epidemiology. By obtaining the bulk of the required carbon, nitrogen and energy from a metabolic waste product (i.e. tunicates lack the enzymatic ability to degrade purines past uric acid),

Nephromyces is able to limit its impact on the host while still reaching high cellular densities. *De novo* synthesis of purines indicates that neither *Nephromyces/Cardiosporidium* is dependent of the host for IMP. In fact, these purine degradation and biosynthesis pathways may have been the integral factors that allowed *Nephromyces* to leave the intracellular environment and colonize the renal sac.

Another critical factor in *Nephromyces'* ability to survive in the renal sac is likely its bacterial endosymbionts. The α-proteobacteria endosymbionts found in *Nephromyces* and *Cardiosporidium* are monophyletic, which indicates they have been maintained and vertically transmitted since the divergence of *Nephromyces* and *Cardiosporidium.* In addition to the α-proteobacteria, *Nephromyces* has acquired a β-proteobacteria and a Bacteroidetes endosymbiont. The α-proteobacteria and Bacteroidetes symbionts show a marked reduction in carbon metabolism with the Nαe, only encoding genes for the citric acid cycle. Bacteroidetes is only capable of processing three carbon compounds and encodes a partial citric acid cycle. Such pronounced reduction suggests that *Nephromyces* provides its symbionts a limited 'diet'. In both symbionts, carbon metabolism may be dependent on pyruvate, which is one of the products of AGXT. Related to this limited carbon metabolism, all three of the bacterial endosymbionts paradoxically encode complete fatty acid biosynthesis, but lack fatty acid degradation. Presumably, the fatty acid biosynthesis is for the construction of membranes, but without fatty acid degradation these symbionts are incapable of

processing fatty acids as a carbon source. Both Nαe and Nβe lack complete pathways for the creation of glycerophospholipids, and both contain phospholipid ABC transporters. This could indicate a dependence on *Nephromyces* for phospholipids.

None of the three endosymbionts of *Nephromyces* contain any genes involved in purine degradation. The absence of this entire pathway in the genomes of the endosymbionts is further support that the high levels of uric oxidase detected are from *Nephromyces* and not from any of its bacterial endosymbionts. Similarly, Nβe and Nαe do not encode any genes involved in *de novo* purine biosynthesis, including genes for the conversion from IMP to adenine and guanine. Nβe can likely synthesize purines from PPRP through the histidine biosynthesis pathway and contains the genes to synthesize adenine and guanine from IMP. The total lack of purine biosynthesis genes in both Nβe and Nαe makes these symbionts dependent on *Nephromyces* for both adenine and guanine. If *Nephromyces* were incapable of *de novo* purine biosynthesis then the entire renal sac community would be dependent on either the tunicate host for all purines, or on Nβe. This would be a significant burden on the host and would markedly increase the cost of infection, which does not align with *Nephromyces'* strategy of low virulence and high-density infection of its host. This argument adds support to the prediction that *Nephromyces/Cardiosporidium* are able to synthesize purines.

Given the reduced genomes and correspondingly reduced metabolic

capabilities, Nβe and Nbe encode a large proportionally high number of genes for synthesizing amino acids, vitamins, and co-factors. Together Nβe and Nbe could provide *Nephromyces* with all but one essential amino acid (tryptophan). Nβe is predicted to synthesize leucine, isoleucine, and valine up to the last step, where the final amine group is added. Conversely, *Nephromyces* only encodes the last step in the conversion of these three amino acids. Similarly, Nbe encodes a partial biosynthetic pathway for phenylalanine, which seems to be complemented by *Nephromyces.* Nαe is capable of synthesizing three amino acids and only one, which is an essential amino acid, is not encoded by *Nephromyces* (lysine). Vitamin and cofactor biosynthesis in Nαe is also limited, synthesizing heme, ubiquinone, and lipoic acid, with lipoic acid being the only product *Nephromyces* may be incapable of synthesizing itself.

With such limited vitamin and amino acid metabolism encoded in the Nαe genome it is unlikely that *Nephromyces* is maintaining Nαe just for lysine biosynthesis, but from the data we are unable to propose what particular function Nαe serves *Nephromyces.* In addition, all of the limited species infections of *Nephromyces* we have been able to culture so far have an Nαe type of symbiont. While neither the frequency of Nαe or the limited species cultures are strong support for the Nαe symbiont being essential, it does suggest that Nαe may have an important role in *Nephromyces* metabolism.

As more apicomplexan genomes are sequenced it is becoming apparent that while they do share a large core subset of proteins, the differential losses and

expansions are very lineage specific. This is likely due to adaptations required for specific host biology. Each lineage displays a characteristic patchwork of different gene losses and expansions. Many of these lineages contain orthologs with the Chromerids that are not found in other apicomplexan lineages. *Nephromyces* and *Cardiosporidium* have retained both purine biosynthesis and degradation, which has been lost in all other apicomplexan lineages. There may be something particular to ascidian biology that necessitates retaining and expressing these purine metabolism pathways. *Nephromyces* and *Cardiosporidium* share the vast majority of their genes with each other, as well as encoding the majority of the commonly shared apicomplexan genes. With so many metabolic similarities between *Cardiosporidium* and *Nephromyces*, we were unable to detect any clear differences related to *Nephromyces'* proposed mutualistic relationship.

While there are no obvious differences between *Nephromyces* and *Cardiosporidium*, we are severely limited by a lack of lineage specific proteomic work. In other apicomplexans the gene families that modulate host immunity are highly lineage specific. In *Nephromyces/Cardiosporidium,* the mechanisms of host manipulation are entirely unknown. Without a greater understanding of how both *Cardiosporidium* and *Nephromyces* interact with their host at a proteomic level, we are unable to conclusively say that there is a difference between these two organisms.

Methods

Molgula manhattensis collection

Molgula manhattensis tunicates were collected from a dock in Greenwich Bay, Rhode Island (41°39'22.7"N 71°26'53.9"W) in July 2014. A single renal sac was separated from one tunicate, and all extraneous tissue removed. The intact renal sac was placed in liquid nitrogen for 5 min and then stored at -80°C.

Cardiosporidium cionae collection, isolation, and concentration

Ciona intestinalis were collected from Matunuck Marina, RI (41.3890° N, 71.5201° W), in August 2017. Tunics were removed and the body wall was opened to allow access to the heart. A sterile syringe was used to remove cardiac blood as cleanly as possible. Blood was kept at 4° C until *Cardiosporidium* infection was verified using Giemsa stain to visualize *Cardiosporidium.* Heavily infected samples were pooled together and centrifuged at 500g for 5 minutes. The resulting supernatant was removed and the samples were frozen in liquid nitrogen and stored at -80° C. Samples with high rates of infection were enriched for *Cardiosporidium* using sucrose gradients [30, 31]. Gradients of 20, 25, 30, 35, 40% sucrose solutions in phosphate buffer were layered together. Approximately 5 ml of tunicate blood was added to the column and centrifuged at 1750g for 30 mins at 4° C. The 25% and 30% layers were collected (based on visual screens showing high *Cardiosporidium* cell density and low tunicate cell density), washed with PBS twice, pelleted and then frozen in liquid nitrogen and stored at -80° C.

RNA Extraction

RNA extraction buffer (Zymo Research LLC. Irvine, CA) was added to samples and ground with a pestle. Following grinding, the Zymo Quick-RNA kit (Zymo Research LLC. Irvine, CA) was used and the manufacturer's protocol was followed. RNA was converted to cDNA and sequenced at the School of Medicine Genome Resource Center, University of Maryland. One paired-end RNA library was run on one lane of the Illumina HiSeq platform. Resulting in 40,606,230 from the *M. manhattensis* renal sac. For *Cardiosporidium,* three samples of *C. intestinalis* blood were used: one with unseparated blood, one enriched with cells collected at the 25% sucrose gradient, and one enriched with cells from the 30% sucrose gradient were multiplexed on one lane of the Illumina HiSeq platform, resulting in 92,250,706, 109,023,104, and 110,243,954 reads respectively. Transcriptome data was assembled and proteins were predicted with Trinity/Trinotate pipeline version 2.4.0 run on the server at Brown University Center for Computation and Visualization [32]. Reads assembled into 145674 and 109,446 contigs from *M. manhattensis* and *C. intestinalis* respectively. Protein sequences were predicted using Transdecoder [32]. Blastp was used to identify bacterial sequences from assembled transcripts against NCBI's refseq and binned. Remaining Eukaryotic sequences were separated with blastp against a custom database of alveolate and ascidian transcriptomes. Trimmed reads were mapped back to each of the six bins (*Nephromyces*, *M. manhattensis*, *Nephromyces*

bacteria, *Cardiosporidium*, *C. intestinalis*, and *Cardiosporidium* bacteria) and then reassembled independently in Trinity. The *Nephromyces* transcriptome was composed of multiple *Nephromyces* species, and CD-hit was used to cluster transcripts based on 50 percent identity. Transcriptome completeness was assessed with Busco v3 against the Eukaryotic and bacterial reference data sets [33]. Transcripts were annotated using Interproscan [34].

DNA Extraction

The renal sacs from 8 lab grown *M. manhattensis* individuals were dissected and their renal fluid was pooled in a 1.5ml Eppendorf tube. Contents were centrifuged at 8000g for 5 min. to pellet *Nephromyces* cells, and following centrifugation the renal fluid was discarded. 500µl of CTAB buffer with 5ul of proteinase K and ceramic beads were added to the pelleted *Nephromyces* cells. The sample was placed in a bead beater for 3 min. and then on a rotator for 1.5hrs at room temp. 500µl of chloroform was added, mixed gently and centrifuged for 5 min. The top layer was removed and twice the sample volume of ice cold 100% EtOH and 10% sample volume of 3M sodium acetate were added to the sample and incubated a -20C overnight. The sample was centrifuged at 16000g for 30 min. and the liquid was removed. Ice cold 70% EtOH was added and centrifuged at 16000g for 15 min. Liquid was removed and sample air dried for 2 min. DNA was re-eluted in 50ul of deionized water.

Illumina Sequencing

A nanodrop (2000c, Thermo Scientific) was used to assess DNA purity and

DNA concentration, and a genomic gel was run to assess DNA fragmentation. Following quality control, an Illumina library was constructed. Library prep and sequencing were done at the URI Genomics and Sequencing Center (URI GSC). The completed library was sequenced on the Illumina MiSeq platform at the URI GSC and the HiSeq platform at the University of Baltimore sequencing center on three lanes.

Pacific Biosciences Sequencing

Using the contents of 150 *M. manhattensis* renal sacs (done in batches of 10 then pooled), the same DNA extraction protocol was performed as for Illumina sequencing. DNA was sequenced using three SMRT cells on the Pacific Biosciences platform at the University of Baltimore sequencing center.

Illumina assembly

One MiSeq lane and three lanes of HiSeq, all from the same library, were trimmed using Trimmomatic [35] and then assembled using Spades [36] assembler on the URI server BlueWaves.

Pacific Biosciences assembly

Pacific Biosciences reads were error corrected using pbsuite/15.8.24 [37] on the Brown University server, Oscar. Reads were then assembled using Canu [38]. Contigs generated by Canu were combined with Illumina MiSeq/HiSeq short reads with Abyss v2.02 [39]. *Nephromyces* contigs were identified by mapping *Nephromyces* transcriptome reads using Bowtie2. Contigs with greater than 90x coverage as assessed with bedtools [40] were binned as *Nephromyces*.

Bacterial endosymbiont genome assembly

Using the contigs from the Abyss assembly bacterial contigs were initially identified by hexemers using VizBin [41]. Transcriptomic reads that were identified as bacterial were mapped using Bowtie2 [42]. Bacterial contigs were separated based on a 90x coverage threshold with bbmap. Binned bacterial contigs were preliminarily annotated with Prokka [43]. Resulting annotations were run through KEGG GhostKoala to assign and separate by taxonomy. Taxon separated contig bins were merged and scaffolded using PBJelly from the PBsuite of tools [37]. Trimmed Illumina MiSeq and HiSeq reads were remapped to resulting contigs to insure accurate assembly using Bowtie2. Final assembled bacterial genomes were re-annotated with Prokka with a genus specific database.

Bacterial phylogeny

16s rRNA sequences from *Nephromyces* bacterial endosymbiont genomes, predicted by rRNAammer, and 16S rRNA sequences from *Cardiosporidium* transcriptome were used in the phylogenetic analysis. All 16s rRNA rickettsiales, sphingobacteriaceae, and alcaligenaceae sequences with a minimum length of 1300bp available on NCBI's refseq were downloaded separately. Sequences were aligned with MAFFT [44] with G-INS-I and trimmed to length in Geneious 6. Maximum likelihood trees of the alignments were generated with RAxML v 8.2.0 using the GTRCAT model run for 10000 generations with 100 generation burn in [45].

Orthology

The following apicomplexan and Chromerid transcriptomes were downloaded from EuPathDB: *C. parvum* Iowa, *G. niphandrodes*, *B. bovis* T2Bo, *T. parva* Muguga, *P. falciparum* 3D7, *C. cayetanensis*, *E. brunetti* Houghton, *E. falciformis* Bayer Haberkorn, *E. tenella* Houghton, *H. hammondi* HH34, *N. caninum* LIV, *S. neurona* SN3, *T. gondii* ME49, *C. velia* CCMP2878, *V. brassicaformis* CCMP3155. These transcriptomes were combined with *Cardiosporidium* and *Nephromyces* (clustered at the 70% identity) Orthofinder v. 2.2.6 was used to assign transcripts to orthologous groups.

1. Reid AJ (2014) Large, rapidly evolving gene families are at the forefront of host-parasite interactions in Apicomplexa. Parasitology 1–14. https://doi.org/10.1017/S0031182014001528
2. Janouskovec J, Keeling PJ (2016) Evolution: Causality and the origin of parasitism. Curr Biol 26:R174–R177. https://doi.org/10.1016/j.cub.2015.12.057
3. Woo YH, Ansari H, Otto TD, et al (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. Elife 4:1–41. https://doi.org/10.7554/eLife.06974
4. Zarowiecki M, Berriman M (2015) What helminth genomes have taught us about parasite evolution. Parasitology 142 Suppl:S85-97. https://doi.org/10.1017/S0031182014001449
5. Jackson AP (2015) The evolution of parasite genomes and the origins of parasitism. Parasitology 142:S1–S5. https://doi.org/10.1017/S0031182014001516
6. Escalante a a, Ayala FJ (1995) Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes. Proc Natl Acad Sci U S A 92:5793–7. https://doi.org/10.1073/pnas.92.13.5793
7. Frank SA (1992) A kin selection model for the evolution of virulence. Proc R Soc L B 250:195–197
8. Frank S a (1996) Host-symbiont conflict over the mixing of symbiotic lineages. Proc Biol Sci 263:339–344. https://doi.org/10.1098/rspb.1996.0052
9. Frank S a (1996) Models of parasite virulence. Q Rev Biol 71:37–78. https://doi.org/10.1086/419267
10. Brunner JL, Collins JP (2009) Testing assumptions of the trade-off theory of the evolution of parasite virulence. Evol Ecol Res 11:1169–1188
11. Kada S, Lion S (2015) Superinfection and the coevolution of parasite virulence and host recovery. J Evol Biol 28:2285–2299. https://doi.org/10.1111/jeb.12753
12. Ciancio A, Scippa S, Finetti-Sialer M, et al (2008) Redescription of *Cardiosporidium cionae* (Van Gaver and Stephan, 1907) (Apicomplexa: Piroplasmida), a plasmodial parasite of ascidian haemocytes. Eur J Protistol 44:181–196. https://doi.org/10.1016/j.ejop.2007.11.005
13. Kumagai A, Suto A, Ito H, et al (2011) Soft tunic syndrome in the edible ascidian *Halocynthia roretzi* is caused by a kinetoplastid protist. Dis Aquat Organ 95:153–161. https://doi.org/10.3354/dao02372
14. Saffo MB (1982) DISTRIBUTION OF THE ENDOSYMBIONT *NEPHROMYCES GIARD* WITHIN THE ASCIDIAN FAMILY MOLGULIDAE. Biol Bull 162:95–104
15. Saffo MB, Lowenstam H a (1978) Calcareous deposits in the renal sac of a molgulid tunicate. Science 200:1166–1168. https://doi.org/10.1126/science.200.4346.1166
16. Goodbody BYI (1965) Nitrogen Excretion in Ascidiacea. Enzyme 34:299–305
17. Saffo MB (1988) Nitrogen Waste or Nitrogen Source? Urate Degradation in the Renal Sac of Molgulid Tunicates. Biol Bull 175:403. https://doi.org/10.2307/1541732
18. Lambert CC, Lambert G, Crundwell G, Kantardjieff K (1998) Uric acid accumulation in the solitary ascidian *Corella inflata*. J Exp Zool 282:323–331. https://doi.org/10.1002/(SICI)1097-010X(19981015)282:3<323::AID-JEZ5>3.0.CO;2-O
19. Rouf MA, Lomprey RF (1968) Degradation of uric acid by certain aerobic bacteria. J Bacteriol 96:617–622
20. Thong-On A, Suzuki K, Noda S, et al (2012) Isolation and Characterization of Anaerobic Bacteria for Symbiotic Recycling of Uric Acid Nitrogen in the Gut of Various Termites. Microbes Environ 27:186–192. https://doi.org/10.1264/jsme2.ME11325
21. Middelhoven WJ, De Hoog GS, Notermans S (1989) Carbon assimilation and extracellular antigens of some yeast-like fungi. Antonie Van Leeuwenhoek 55:165–175. https://doi.org/10.1007/BF00404756
22. Seah B, Saffo MB, Cavanaugh CM (2011) A Tripartite Animal-Protist-Bacteria Symbiosis: Culture-Independent and Phylogenetic Characterization. Harvard
23. Moran NA, Dunbar HE, Wilcox JL (2005) Regulation of transcription in a reduced bacterial genome: Nutrient-provisioning genes of the obligate symbiont *Buchnera aphidicola*. J Bacteriol

187:4229–4237. https://doi.org/10.1128/JB.187.12.4229-4237.2005
24. Lopez-Sanchez MJ, Neef A, Peret J, et al (2009) Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. PLoS Genet 5:. https://doi.org/10.1371/journal.pgen.1000721
25. De Souza DJ, Bézier A, Depoix D, et al (2009) *Blochmannia* endosymbionts improve colony growth and immune defence in the ant *Camponotus fellah*. BMC Microbiol 9:1–8. https://doi.org/10.1186/1471-2180-9-29
26. Urakawa H, Dubilier N, Fujiwara Y, et al (2005) Hydrothermal vent gastropods from the same family (Provannidae) harbour E- and y -proteobacterial endosymbionts. Environ Microbiol 7:750–754. https://doi.org/10.1111/j.1462-2920.2005.00753.x
27. Marin B, Nowack ECM, Melkonian M (2005) A plastid in the making: Evidence for a second primary endosymbiosis. Protist 156:425–432. https://doi.org/10.1016/j.protis.2005.09.001
28. Philip LA, Tullman-Ercek, George G (2006) NIH Public Access. Annu Rev Microbiol 373–395. https://doi.org/10.1146/annurev.micro.60.080805.142212.The
29. Hyde JE (2007) Targeting purine and pyrimidine metabolism in human apicomplexan parasites. Curr Drug Targets 8:31–47. https://doi.org/10.2174/138945007779315524
30. Ogedengbe ME, Qvarnstrom Y, da Silva AJ, et al (2015) A linear mitochondrial genome of *Cyclospora cayetanensis* (Eimeriidae, Eucoccidiorida, Coccidiasina, Apicomplexa) suggests the ancestral start position within mitochondrial genomes of eimeriid coccidia. Int J Parasitol 45:361–365. https://doi.org/10.1016/j.ijpara.2015.02.006
31. Arrowood MJ, Sterling CR (2016) Isolation of *Cryptosporidium* Oocysts and Sporozoites Using Discontinuous Sucrose and Isopycnic Percoll Gradients Author (s): Michael J . Arrowood and Charles R . Sterling Published by : Allen Press on behalf of The American Society of Parasitologists St. J Parasitol 73:314–319
32. Haas BJ, Papanicolaou A, Yassour M, et al (2014) reference generation and analysis with Trinity
33. Simão FA, Waterhouse RM, Ioannidis P, et al (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351
34. Finn RD, Attwood TK, Babbitt PC, et al (2017) InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45:D190–D199. https://doi.org/10.1093/nar/gkw1107
35. Bolger A, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114
36. Bankevich A, Nurk S, Antipov D, et al (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021
37. English AC, Richards S, Han Y, et al (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS One 7:1–12. https://doi.org/10.1371/journal.pone.0047768
38. Koren S, Walenz B, Berlin K, et al (2014) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 1–11. https://doi.org/10.1101/gr.215087.116.Freely
39. Jackman SD, Vandervalk BP, Mohamadi H, et al (2017) ABySS 2 . 0 : resource-efficient assembly of large genomes using a Bloom filter. Genome Res 27:768–777. https://doi.org/10.1101/gr.214346.116.Freely
40. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033
41. Laczny CC, Sternal T, Plugaru V, et al (2015) VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 3:1–7. https://doi.org/10.1186/s40168-014-0066-1
42. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357
43. Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153

44. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
45. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Chapter 4

Amplicon Sequencing reveals hyper diversity and universal multi-species infection prevalence in the genus *Nephromyces*

by

Christopher Paight[1], & Christopher E Lane[1]*

will be submitted to the Journal of Eukaryotic Microbiology

[1] Department of Biological Sciences, University of Rhode Island, Kingston RI, 02881, USA.

CHAPTER 4

Abstract

Monogenus multi-species infections are common in many Apicomplexans. However, these multi-species infections are often overlooked unless specifically targeted. Using amplicon primers designed to target 18s rRNA, COI, and 16s rRNA we attempted to quantify the species diversity and incidence of multi-species infections of *Nephromyces* in the tunicate host *Molgula manhattensis* collected in a limited geographic area from the waters of Rhode Island. Our data indicate that *Nephromyces* is hyper diverse and multispecies infections are nearly universal.

Introduction

Nephromyces is a genus of Apicomplexa with a symbiotic relationship with their hosts, tunicates in the family molgulidae. First described in 1874 by de Lacaze-Duthiers, *Nephromyces* was given several "identities" until it was finally placed in Apicomplexa using molecular phylogenetics (Saffo et al. 2010). Part of the confusion over its taxonomic affinity was because *Nephromyces* inhabits the renal sac, a structure unique to the molgulid tunicates. While the function of the renal sac in not understood, it contains high levels of uric acid and calcium oxalate (Saffo and Lowenstam 1978). Based on the metabolic capacity of *Nephromyces,* it appears to use uric acid for the purpose of primary carbon and nitrogen acquisition (Chapter 2). In order to supplement a diet of uric acid,

Nephromyces relies on bacterial endosymbionts for the biosynthesis of metabolites from pathways missing from its genome (Chapter 3). Three different types of bacterial endosymbionts have been found in the genus *Nephromyces,* an alphaproteobacteria in *Rickettsia,* a betaproteobacteria in *Bordetella,* and a bacteroidetes in the family sphingobacteriaceae (Chapter 3). Despite genomic data, which indicates that these different types of bacteria are not functionally equivalent, no species of *Nephromyces* has been shown to have more than one type of bacterial endosymbiont (Seah et al. 2011).

Based on preliminary genomic and transcriptomic sequencing of *Nephromyces* it became apparent that there was a surprising amount of genetic diversity in the genus. In addition to the high levels of genetic diversity, *Nephromyces* also had high incidences of multi-species infections within individual renal sacs. Attempts to culture single species infections and limited species (3-5 isolates) infections in the lab were met with mixed success, but even limited species populations did poorly compared to the cultures that contained species numbers that better approximated wild samples.

To quantify the biological diversity and the incidence of multispecies *Nephromyces* infections found in molgulid tunicates, we used an amplicon sequencing approach. Because polymorphic 18S rDNA sequences have been reported in *Plasmodium* (Li et al. 1997), we targeted the Cytochrome Oxidase I (CO1) mitochondrial gene, as well as the 18S. In order to account for the endosymbiotic diversity within the *Nephromyces* population, we also targeted the

bacterial 16S rRNA. Genomic data indicate that members of *Rickettsia*, *Bordetella*, and sphingobacteriaceae are endosymbionts of *Nephromyces* isolates, but their diversity is unknown.

Methods

Fifty *Molgula manhattensis* tunicates were collected from a single floating dock located in Greenwich Bay, RI (41° 39' 11.009" N 71° 27' 5.843" W), over a period of 4 weeks in the summer of 2016. Renal sacs were dissected out of the animals and contents were collected by a micropipette and placed in 1.5 ml Eppendorf tubes. Dissecting tools were sterilized in a 10% bleach solution for 15 min and then rinsed between tunicates. Sample tubes were immediately frozen in liquid nitrogen for five minutes and subsequently stored at -80° C.

DNA was extracted using the method described in (Chapter 2). Extracted DNA was stored at -20° C. The 18S rRNA primers and CO1 primers were designed to target *Nephromyces* based on available genomic data (Chapter 3). The universal 16S rRNA primers from (Klindworth et al. 2013) were used to amplify the bacterial endosymbionts from *Nephromyces*. The Illumina adaptor sequence was added to the start of each primer resulting in the following sequences 18Sf (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGGTAATTCCAGCTCC), 18Sr (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGCTTTCGCAGTAGTYYGTCTTT), CO1f (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGYGGWGTAGGWSCWGGWTGGA),

CO1r
(GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTTCWGGATGWCCAAARAA)
16Sf (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG),
16Sr
(GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC).
For each sample, PCR was performed with all three primer sets with the following cycle 94° C 2 min (94° C 30 sec, 55° C 30 sec, 72° C 45 sec) x 35, 72° C 5 min. Resulting PCR products were visually inspected on an agarose gel and quantity estimated with a nanodrop. Twenty microliters of PCR product from each of the three primer sets was pooled into a single tube corresponding to an individual tunicate. The pooled sample was cleaned using the ampure bead purification with a 0.7% solution of Agencourt AMPure beads (Beckman Coulter). The addition of well specific adaptors, library preparation, and sequencing was done at the URI genomic sequencing center on the Illumina MiSeq platform.

Sequence data were de-multiplexed prior to analysis. Bduck from the bbmap suit of tools was used to bin reads based on CO1 primers (Bushnell 2014). The universal 18S rRNA and 16S rRNA primers were too conserved for reliable binning based on primers, so reads were screened against the PR2 database using the NCBI's magicblast (Boratyn et al. 2018). Sequences with an 85% ID and 35% coverage were classified as 18S sequences and binned into a new file composed of 18S reads. Adaptors and primer sequences were remove from the forward and reverse reads from each of the three read sets using bduck.

Cleaned and binned read sets were individually processed in R using dada2 with the pool="pseudo" setting (Callahan et al. 2016). Assembled 18S and 16S were assigned taxonomies with the PR2 database (Guillou et al. 2013). Cytochrome oxidase 1 (CO1) sequences were assigned taxonomy using BLASTx against NCBI's refseq_protein database. All 18S and CO1 sequences that were not apicomplexan were removed from the count table, taxonomy table, and sequence files. Remaining sequences were aligned with MAFFT (Katoh and Standley 2013) to 16S rRNA sequences from the three known bacterial endosymbionts found in *Nephromyces.* Reference sequences were trimmed to the amplicon sequence length and CD-hit was used to cluster sequences with 85% sequence identity. All bacterial sequences, which did not cluster were deemed contamination and removed from count table, taxonomy table, and sequences file.

Remaining 18S and CO1 sequences were aligned with MAFFT and trimmed to the same length. Sequences were clustered at 100%, 99%, 98%, 97%, 96%, 95%, 94% sequence identity levels using CD-hit (Li and Godzik 2006). Sequences from 18S, CO1 clusters and 16S bins corresponding to endosymbiont type were processed individually in R. Figures were made in R using ggplot (Wickham 2016).

Results

The amplicon sequencing run resulted in 25,895,690 reads with average reads per sample of 137,743. After binning there were 4,930,010 CO1 reads,

6,491,918 18S reads, and 14,468,228 16S reads. Following assembly in dada2 and decontamination there were 1,876,107 sequences corresponding to 329 amplicon sequence variants (ASVs) for 18S, 1,522,378 sequences corresponding to 188 ASVs for CO1, and 62,905 sequences with 152 ASVs for 16S.

Of the 329 18S ASVs there is an average of 79.26 ASVs per tunicate individual with a max of 145 ASVs and a min of 34 ASVs. When clustered at the more taxonomically relevant 98% identity there are 23 clusters with an average of 5.1 per tunicate individual with a max of 11 and a min of 3 (Figure 7). The most common ASVs were seen in 59.57% of samples the least common in 2.12%, when clustered at 98% these numbers rise to 86.17% and 2.12%.

There is a total of 188 CO1 ASVs with an average of 52.8 ASVs per tunicate a max of 101 and a min of 16. When clustered to the 98% identity level there are a total of 26 clusters an average of 6.24 and 10/2 max/min (Figure 8). The most common AVSs were found in 53% of tunicates sampled and the rarest ASVs were in 2%. After clustering at 98% the most common clusters were in 75.5% of tunicates the rarest in 2%.

From a total of 152 16s ASVs classified as *Nephromyces* bacterial endosymbionts 49 were recovered from *Bordetella*, 89 from *Rickettsia*, and 14 from sphingobacteriaceae. The average number of 16S ASVs per tunicate is 13.94 with a max of 49 and a min of 0 (Figure 9). Forty percent of samples contained all three types of bacterial endosymbiont and 90% contained at least two of the bacterial endosymbiont types. There were two samples, which did not contain

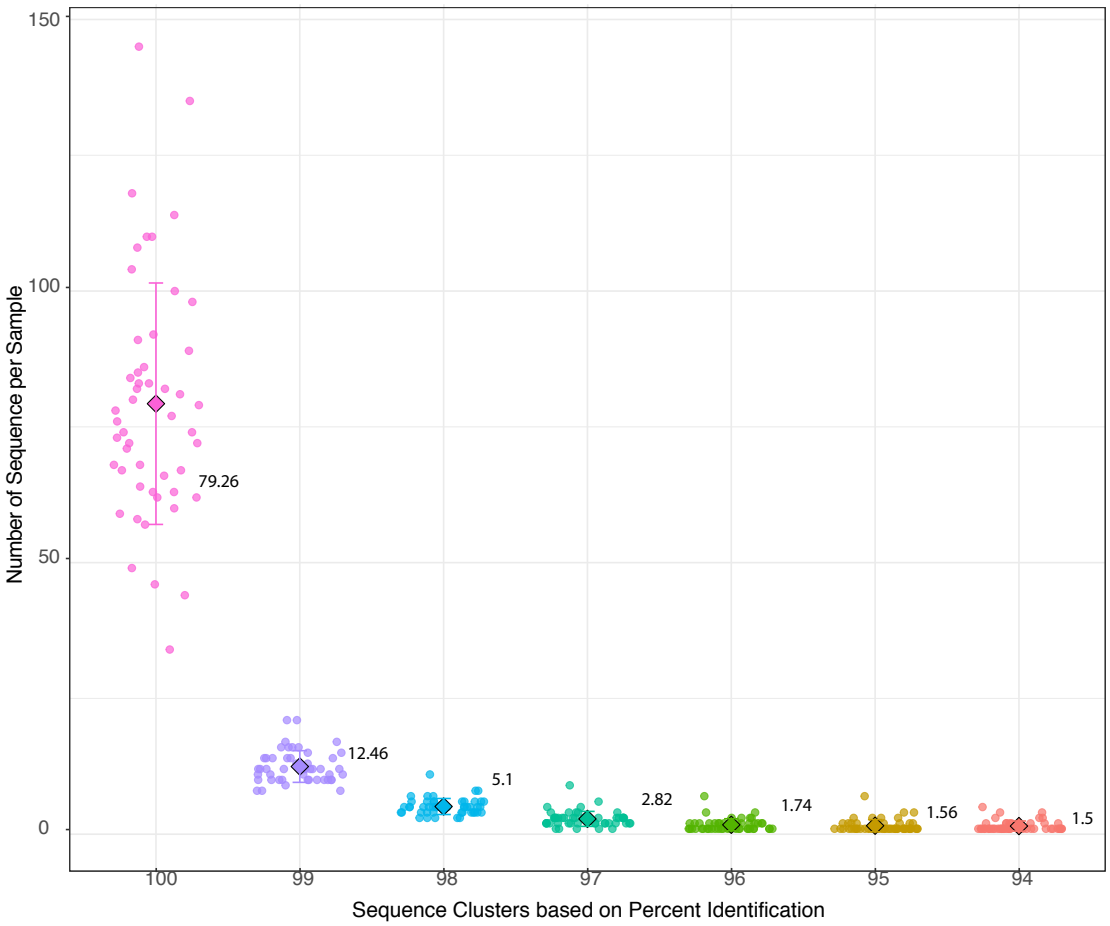


Figure 7) Grouped scatterplot of *Nephromyces* 18s rRNA ASV's clustered at different percent identity levels. Each dot represents the number of ASV's from an individual tunicate sample, squares indicate mean per sample (labeled to the side), and error bars indicate one standard deviation. Without clustering the average number of ASV's is 79.26 per sample and max of 145 and a min of 34. When clustered at 98% identity the average number of sequences per sample falls to 5.1 with a max of 11 and a min of 3.

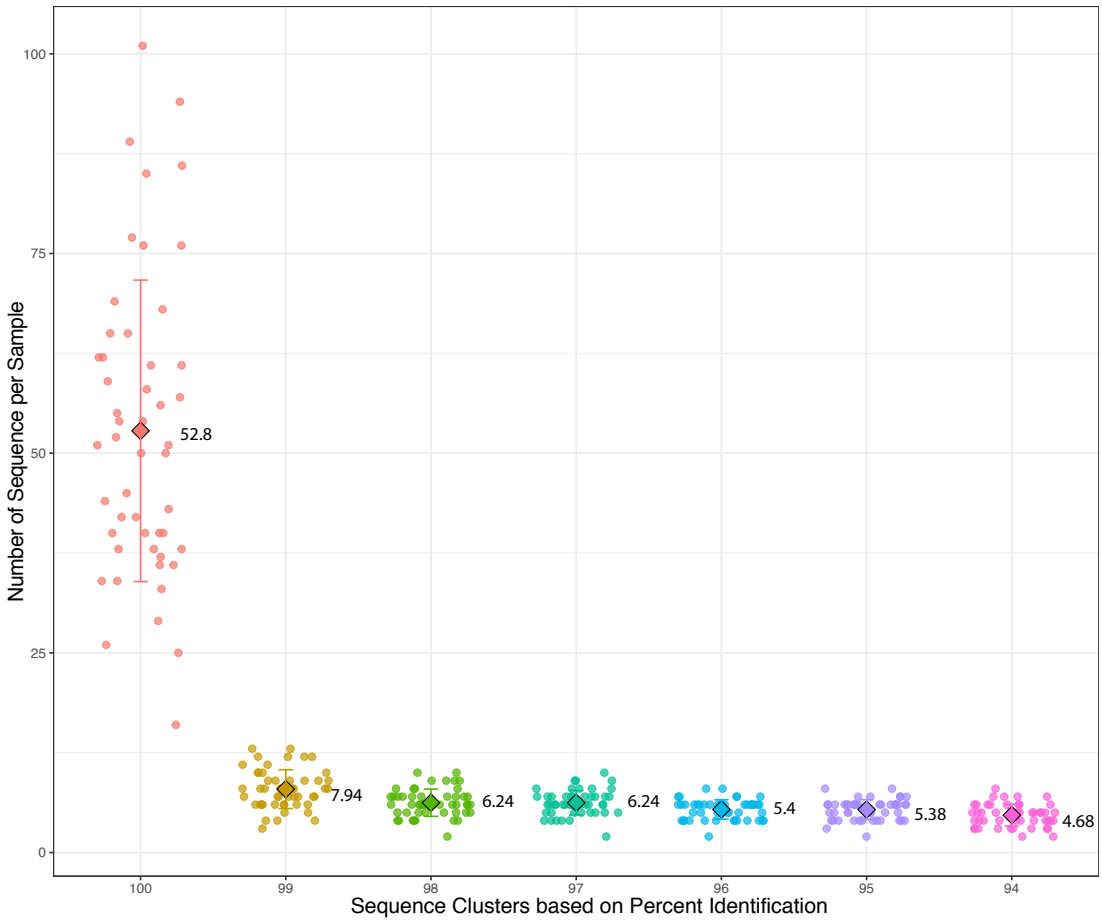


Figure 8) Grouped scatterplot of *Nephromyces* COI ASV's clustered at different percent identity levels. Each dot represents the number of ASV's from an individual tunicate sample, squares indicate mean per sample (labeled to the side), and error bars indicate one standard deviation. Without clustering the average number of ASV's is 52.8 per sample and max of 101 and a min of 16. When clustered at 98% identity the average number of sequences per sample falls to 6.24 with a max of 10 and a min of 2.

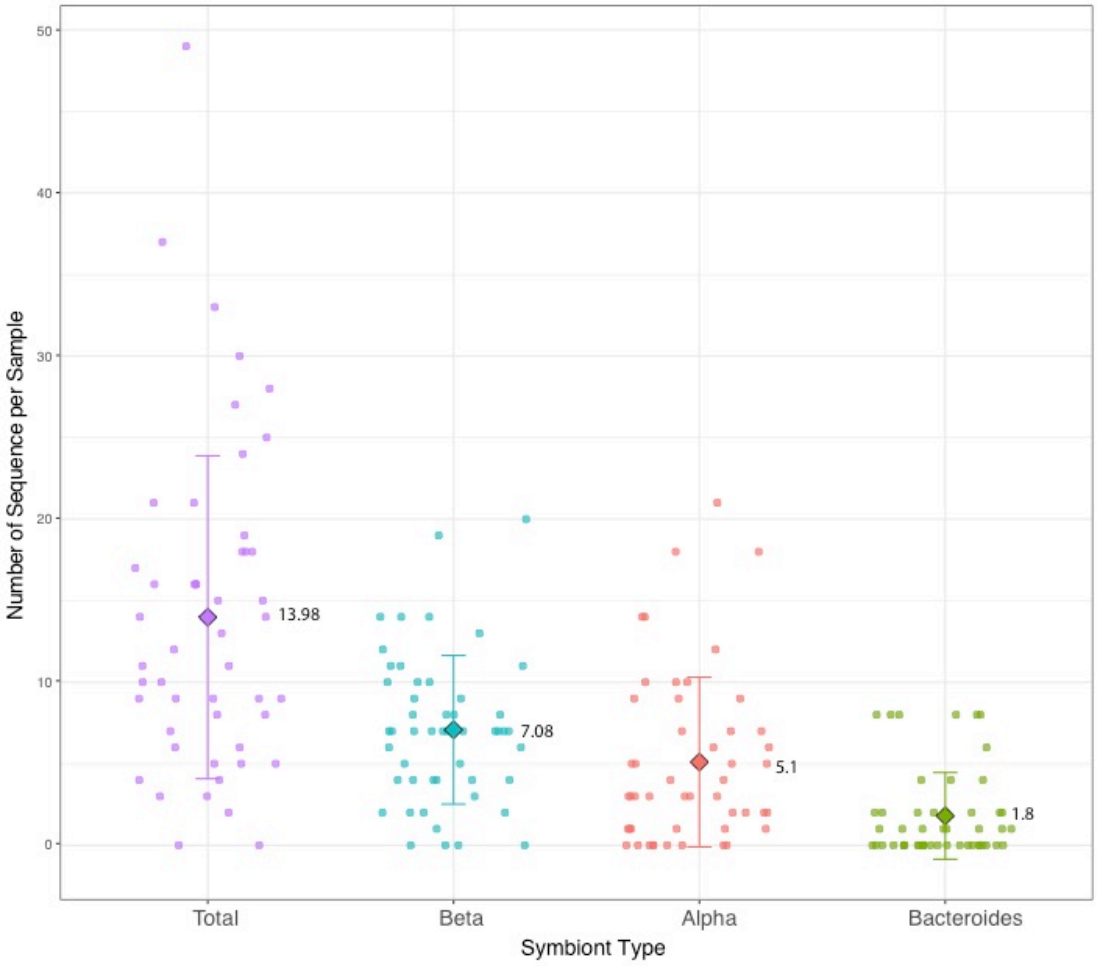


Figure 9) Grouped scatterplot of 16s rRNA ASV's from *Nephromyces* bacterial endosymbionts. Purple is all bacterial combined, blue is the *Bordetella* only, red is *Rickettsia*, and green in sphingobacteriaceae. Each dot represents the number of ASV's from an individual tunicate sample, squares indicate mean per sample (labeled to the side), and error bars indicate one standard deviation. An average of 13.98 bacterial ASV's per sample with 7.08 *Bordetella*, 5.1 *Rickettsia*, and 1.8 sphingobacteriaceae.

any ASVs assigned as *Nephromyces* bacterial endosymbionts.

Of the 25 samples that contained two bacterial endosymbionts the *Bordetella* type was in 23 of the samples, *Rickettsia* was in 21 of the samples, and the sphingobacteriaceae type was present in 6 of the samples. *Bordetella* and *Rickettsia* were found together in 76% (19) of renal sacs with two detected bacterial endosymbiont types*, Bordetella* and sphingobacteriaceae in 16% (4), and *Rickettsia* and sphingobacteriaceae in 8% (2).

Discussion

The high numbers of ASVs obtained in this study reveal that *Nephromyces* is extremely diverse, and in all instances, *Nephromyces* infections are multi-species infections. There was one renal sac that only contained a single 18S rRNA sequence, but that renal sac had multiple CO1 sequences and is likely the result of poor 18S PCR amplification of this sample. The diversity observed from amplicon sequencing is extreme, but it is supported by genomic and transcriptomic sequencing on *Nephromyces* (Chapter 2), as well as by cloning of full-length 18S sequences. Despite the substantial diversity, the results are consistent across multiple datasets. The different number of 18S and CO1 ASVs indicate that each *Nephromyces* species encodes multiple copies of 18S with different sequences, as previously described from *Plasmodium* (Li et al. 1997). While CO1 is also likely found in multiple copies in a cell, based on genomic data we do not have any

reason to suspect that CO1 copies differ in sequence. A similar level of diversity is observed among the 16S ASVs, which is consistent with a vertically transmitted bacterial endosymbiont.

When ASVs for 18S and CO1 are clustered at a 98% sequence identity to give an approximate species number, we estimate that there are an average of ~5 different *Nephromyces* species per *M. manhattensis* host, with some *M. manhattensis* containing as many as 10-11 *Nephromyces* species. Whereas multispecies apicomplexan infections are relatively common (Anderson et al. 2000; Lee et al. 2011; Arnott et al. 2012; Lalremruata et al. 2017), the diversity found among *Nephromyces* infecting a single host species is striking. The precise reasons for such high diversity are not known. However, we hypothesize that high levels of diversity may be due to the dependence of *Nephromyces* species on essential amino acids, co-factors, and vitamins produced by their bacterial endosymbionts.

Based on the different metabolic capabilities of the *Nephromyces* bacterial endosymbionts, we postulated that *Nephromyces* might be dependent on metabolites produced by bacterial endosymbionts in conspecifics (Chapter 2). The high proportion of tunicates containing at least two of the bacterial endosymbiont types, 90%, supports this hypothesis. Only two samples did not contain any bacterial endosymbionts and three contained only one type of bacterial endosymbiont. This may be due to sampling error and the endosymbionts were either not amplified in the initial PCR, or one endosymbiont

was preferentially amplified.

In samples where only two types of bacterial endosymbionts were detected the *Bordetella/Rickettsia* pairing were the most common (76% of renal sacs with two bacterial types). Based on the genomes of the endosymbionts *Bordetella* encodes the most complete vitamin and amino acid biosynthesis capabilities, and is predicted to be providing essential amino acids and vitamins to *Nephromyces*. Despite a largely complete genome assembly of the *Rickettsia* symbiont, its functional role is not clear. It encodes the least complete amino acid and vitamin synthesis of any of the three endosymbiont types found in *Nephromyces*, and all the amino acids and vitamin biosynthesis capabilities are also encoded by both the *Bordetella* and sphingobacteriaceae. However, given the high prevalence of *Rickettsia* in our samples and its presence in 82% of renal sacs with only two types of symbionts, it does appear that the *Rickettsia* symbionts are providing an essential function not encoded by either the *Bordetella* or the sphingobacteriaceae.

Based on previous (Seah et al. 2011) and our own microscopy results (Figure 10), no *Nephromyces* individual has been observed containing multiple types of bacterial endosymbiont. If our hypothesis is correct and *Nephromyces* needs metabolites from multiple symbiont types, *Nephromyces* species with dual endosymbionts could survive in hosts without conspecifics, as has been repeatedly observed in insects (McCutcheon and Moran 2007; Bennett and Moran 2013; Rao et al. 2015; Brown et al. 2018). One possibility is that since the

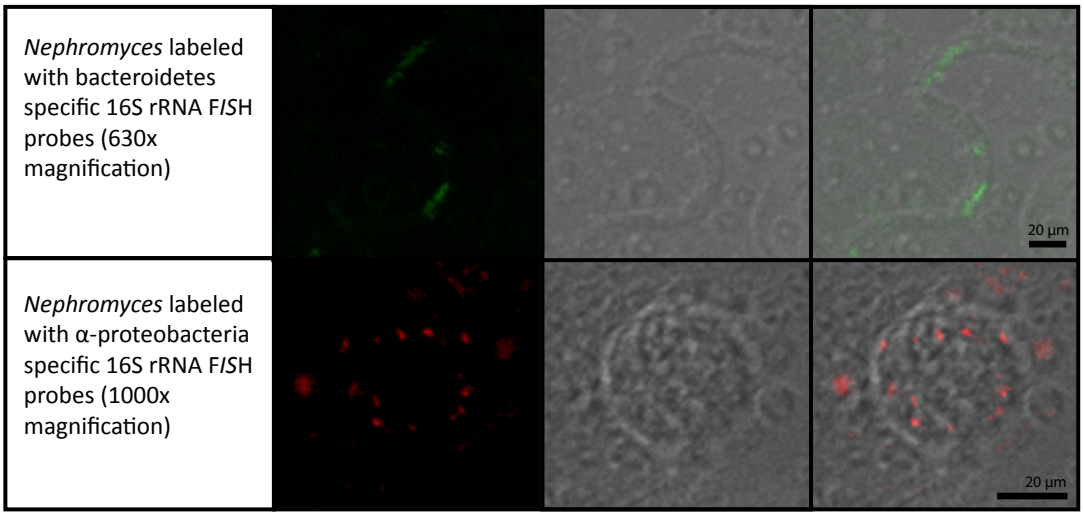


Figure 10) Photos of *Nephromyces* taken on a Zeiss confocal microscope labeled with two 16S rRNA F*IS*H probes, one targeting sphingobacteriaceae (green, top, 630x) and *Rickettsia* (red, bottom, 1000x). Images from left to right were captured under fluorescence, TPMT, and both overlaid. The top pictures are of the tachyzoite life stage and the bottom show the oocyst or merozoite stage. Bacterial endosymbionts have been observed in all of *Nephromyces* life stages, but no *Nephromyces* has been shown to contain multiple types of bacterial endosymbiont. (Courtesy of Liz Hunter)

bacterial endosymbionts of *Nephromyces* are vertically transferred through the oocyst stage, there simply may not be enough space for multiple endosymbiont cells. In insects with dual endosymbionts the bacteria are typically vertically transmitted through a much larger egg cell than the oocyst stage of *Nephromyces*. Alternatively, it may be disadvantageous to carry multiple endosymbionts when multispecies infections are universal in molgulid renal sacs. Given the extreme Muller's ratchet known to occur in bacterial endosymbionts (Moran 1996), it may be evolutionarily cheaper to maintain one bacterial endosymbiont and rely on conspecifics with other types of bacterial endosymbiont. Such a system could not evolve unless there was a high probability of a multispecies infection of any given host.

In a system with so much co-dependence there is the potential for "cheaters" to develop and indeed we have found *Nephromyces* species, that do not seem to contain any bacterial endosymbiont and are presumed to parasitize the system. The absence of bacterial endosymbionts is based on single species isolates and fluorescent *in situ* hybridization (FISH) microscopy (Figure 6), but as the lack of signal in FISH microscopy is not definitive of absence, this has not yet been confirmed.

Given the high sequence diversity among our *Nephromyces* datasets, universal prevalence of multispecies infections within individual renal sacs, and that 90% of renal sac samples contain at least two types of bacterial endosymbionts, *Nephromyces* may not be capable of establishing single species

infections in the wild. Combined with the metabolic capabilities of the different bacterial endosymbiont types (Chapter 2), *Nephromyces* likely exists as a co-dependent species complex or species swarm. This raises interesting questions about the evolution of co-dependence and co-evolution of *Nephromyces* species that require further exploration.

Anderson T. J. C., Haubold B., Williams J. T., Estrada-Franco J. G., Richardson L., Mollinedo R., Bockarie M., Mokili J., Mharakurwa S., French N., Whitworth J., Velez I. D., Brockman A. H., Nosten F., Ferreira M. U. & Day K. P. 2000. Microsatellite Markers Reveal a Spectrum of Population Structures in the Malaria Parasite *Plasmodium falciparum*. *Mol Biol Evol* [Internet], **17**:1467–1482. Available from: http://mbe.oxfordjournals.org/cgi/content/abstract/17/10/1467

Arnott A., Barry A. E. & Reeder J. C. 2012. Understanding the population genetics of *Plasmodium vivax* is essential for malaria control and elimination. *Malar. J.* [Internet], **11**:14. Available from: http://www.malariajournal.com/content/11/1/14

Bennett G. M. & Moran N. A. 2013. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biol. Evol.*, **5**:1675–1688.

Boratyn G. M., Thierry-Mieg J., Thierry-Mieg D., Busby B. & Madden T. L. 2018. Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. *bioRxiv* [Internet], :390013. Available from: https://www.biorxiv.org/content/early/2018/08/14/390013

Brown A. M. V., Wasala S. K., Howe D. K., Peetz A. B., Zasada I. A. & Denver D. R. 2018. Comparative Genomics of *Wolbachia–Cardinium* Dual Endosymbiosis in a Plant-Parasitic Nematode. *Front. Microbiol.* [Internet], **9**:1–21. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2018.02482/full

Bushnell B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner.

Callahan B. J., Mcmurdie P. J., Rosen M. J., Han A. W., Johnson A. J. A. & Holmes S. P. 2016. DADA2 : High-resolution sample inference from Illumina amplicon data. *Nat. Methods* [Internet], :1–7. Available from: http://dx.doi.org/10.1038/nmeth.3869

Guillou L., Bachar D., Audic S., Bass D., Berney C., Bittner L., Boutte C., Burgaud G., De Vargas C., Decelle J., Del Campo J., Dolan J. R., Dunthorn M., Edvardsen B., Holzmann M., Kooistra W. H. C. F., Lara E., Le Bescot N., Logares R., Mahé F., Massana R., Montresor M., Morard R., Not F., Pawlowski J., Probert I., Sauvadet A. L., Siano R., Stoeck T., Vaulot D., Zimmermann P. & Christen R. 2013. The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.*, **41**:597–604.

Katoh K. & Standley D. M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**:772–780.

Klindworth A., Pruesse E., Schweer T., Peplies J., Quast C., Horn M. & Glöckner F. O. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.*, **41**:1–11.

Lalremruata A., Jeyaraj S., Engleitner T., Joanny F., Lang A., Bélard S., Mombo-Ngoma G., Ramharter M., Kremsner P. G., Mordmüller B. & Held J. 2017. Species and genotype diversity of *Plasmodium* in malaria patients from Gabon analysed

by next generation sequencing. *Malar. J.*, **16**:1–11.
Lee K. S., Divis P. C. S., Zakaria S. K., Matusop A., Julin R. A., Conway D. J., Cox-Singh J. & Singh B. 2011. *Plasmodium knowlesi*: Reservoir hosts and tracking the emergence in humans and macaques. *PLoS Pathog.*, **7**.
Li J., Gutell R. R., Damberger S. H., Wirtz R. A., Kissinger J. C., Rogers M. J., Sattabongkot J. & McCutchan T. F. 1997. Regulation and trafficking of three distinct 18 S ribosomal RNAs during development of the malaria parasite. *J. Mol. Biol.*, **269**:203–213.
Li W. & Godzik A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**:1658–1659.
McCutcheon J. P. & Moran N. A. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc. Natl. Acad. Sci. U. S. A.* [Internet], **104**:19392–19397. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2148300&tool=pmcentrez&rendertype=abstract
Moran N. A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci.* [Internet], **93**:2873–2878. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.93.7.2873
Rao Q., Rollat-Farnier P. A., Zhu D. T., Santos-Garcia D., Silva F. J., Moya A., Latorre A., Klein C. C., Vavre F., Sagot M. F., Liu S. S., Mouton L. & Wang X. W. 2015. Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly *Bemisia tabaci*. *BMC Genomics* [Internet], **16**:1–13. Available from:
Saffo M. B. & Lowenstam H. a. 1978. Calcareous deposits in the renal sac of a molgulid tunicate. *Science*, **200**:1166–1168.
Saffo M., McCoy A., Rieken C. & Slamovits C. 2010. *Nephromyces*, a beneficial apicomplexan symbiont in marine animals. *Proc. Natl. Acad. Sci. U. S. A.*, **107**:16190–16195.
Seah B., Saffo M. B. & Cavanaugh C. M. 2011. A Tripartite Animal-Protist-Bacteria Symbiosis: Culture-Independent and Phylogenetic Characterization. Harvard.
Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Available from: http://ggplot2.org

CHAPTER 5

CONCLUSION

The impetus for this research was the characterization of *Nephromyces* symbiotic relationship with its host (Saffo *et al* 2010). In a phylum composed of parasites *Nephromyces* alone provided greater benefits to its host than costs. This characterization was premature and no work was done to demonstrate the details of the relationship between *Nephromyces* and *Molgula* tunicates. As is the case with most symbiosis, the relationship between organisms is complex and can be dependent on external factors. Despite the extensive research into *Nephromyces* there is not enough support to either confirm or refute the classification of *Nephromyces*/*Molgula* relationship as mutualistic. However, the claim of mutualism was entirely based on the unusual epidemiology of *Nephromyces* and our work does provide insight into how *Nephromyces* is able to maintain nearly universal prevalence, reach extremely high cell densities, and remain avirulent.

Central to *Nephromyces* epidemiology is *Nephromyces* ability to degrade purines into glyoxylate and subsequently convert glyoxylate into malate, pyruvate, and glycine. This pathway provides a primary nitrogen, carbon and energy source from uric acid. Malate is part of the citrate cycle critical in the generation of ATP and NADH or it can be converted into oxaloacetate for gluconeogenesis. Pyruvate is extremely versatile and is easily converted into fatty acids, acetyl-CoA, or gluconeogenesis. Glycine, the most basic amino acid, can be

processed into other amino acids and represents nitrogen capture. Tunicates have lost the ability to degrade purines past uric acid essentially making uric acid nitrogenous waste. By using a metabolic waste product, for the host, as a primary substrate, *Nephromyces* can presumably utilize uric acid at little to no cost to the host. This novel substrate decouples the relationship between high parasite growth and virulence, which *Nephromyces* to reach high cell densities and remain avirulent.

With the exception of malate synthase *Cardiosporidium* encodes the same purine degradation capabilities as *Nephromyces* and is capable of converting uric acid into pyruvate and glycine. However, due to the different hosts, *Cardiosporidium* has access to far smaller quantities of uric acid than *Nephromyces* and therefore must rely on additional sources of nutrition from the host. *Molgula* storage and concentration of uric acid to the renal sac has enabled *Nephromyces* to develop its unusual uric acid based metabolism.

It has been proposed that the mutualistic benefit to its host is the processing of indigestible uric acid. This may be the case, but it is likely an oversimplification. First, it is unclear if *Molgula* ever recovers anything back from the uric acid imported into the renal sac. It is possible that valuable metabolites like amino acids or vitamins are exported out of the renal sac, but this has not been demonstrated. Second, the purpose of the renal sac has not been established. Sequestration of uric acid to the renal sac may have developed over time as a way of ridding *Molgula* of an apicomplexan blood parasite; by providing

a parasite with a metabolic waste product as an alternative to infecting blood cells. In this case, *Molgula* benefits from losing a parasite, but this relationship is hardly mutualistic, it is more of a clever host defense mechanism. Third, Nephromyces has been shown to express xanthine dehydrogenase at high levels (99 percentile of all gene expression). Hypoxanthine is the interchange between purine recycling and purine degradation. Xanthine dehydrogenase converts hypoxanthine to xanthine and xanthine to uric acid; this represents the beginning of purine degradation. The competing enzyme, hypoxanthine-guanine phosphoribosyltransferase (HGPRT), converts hypoxanthine to inosine monophosphate (IMP) and from IMP to adenine or guanine. Since the source of uric acid within the renal sac has been shown to be from the tunicate it is curious that *Nephromyces* would have such high expression of xanthine dehydrogenase (Saffo 1988). A possible explanation of *Nephromyces* high expression of xanthine dehydrogenase is as a form of host manipulation. By outcompeting host production of HGART, *Nephromyces* forces greater production of uric acid than may be ideal for the host. This is potentially a cost to the host, particularly in times when purines are scarce

Almost as surprising as an apicomplexan with a uric acid based metabolism was where the genes in the purine degradation pathway come from. All other sequenced apicomplexans have lost the purine degradation pathway. It was thought that the high levels of uric oxidase measured inside the renal sac originated from the *Nephromyces* bacterial endosymbionts. Our data conclusively

show that the genes involved in purine degradation are encoded in *Nephromyces* and *Cardiosporidium* genome. Additionally, these genes are not the result of a gene transfer event, but are the genes that were present when Apicomplexa split with the Chromerids. This pathway had previously been attributed to the bacterial endosymbionts, but is in fact encoded by *Nephromyces/Cardiosporidium.* If the bacterial endosymbionts are not being utilized for purine degradation then they must be contributing in another way.

Using 16s rRNA we have determined that the α-proteobacteria in *Nephromyces* and *Cardiosporidium* are monophyletic and therefore present when *Nephromyces*/*Cardiosporidium* lineages split. This indicates that there may be some aspect of ascidian biology that makes maintaining a bacterial endosymbiont worthwhile. Particularly as bacterial endosymbionts are not found in other apicomplexan lineages. We have yet to determine exactly what the critical function of the α-proteobacteria is, however because it is maintained in *Cardiosporidium,* which is intracellular, and in *Nephromyces,* which is extracellular in the renal sac, the function seems to be not exclusively connected to renal sac biology.

We do not have any genomic data on the *Cardiosporidium* α-proteobacteria endosymbiont at this time, making any comparisons between *Nephromyces* and *Cardiosporidium* α-proteobacteria endosymbiont is preliminary. The genome from the α-proteobacteria in *Cardiosporidium* will need to be sequenced and assembled for more robust analysis. Preliminarily based on

RNAseq data, *Cardiosporidium* α-proteobacteria appears to have more biosynthetic capabilities than *Nephromyces*'. This includes several essential amino acids and vitamins not present in the α-proteobacteria of *Nephromyces*.

In addition to α-proteobacteria in the genus *Rickettsia*, the genus *Nephromyces* also maintains a β-proteobacteria in the genus *Bordetella* and a Bacteroidetes bacterial endosymbiont in the family Sphingobacteriaceae. We have assembled the complete genome for *Nephromyces* Sphingobacteriaceae endosymbiont and *Nephromyces Bordetella*. NBe and Nβe have the biosynthesis capabilities for a number of amino acids and vitamins, which are not encoded in the *Nephromyces* genome. Providing *Nephromyces* with amino acids and vitamins eliminates the need for *Nephromyces* to scavenge those metabolites from the host. Presumably, this reduces *Nephromyces* dependence on the host and also provides a reliable source of these metabolites, which may not be available in the renal sac.

Despite the three types of bacterial endosymbionts (Nαe, NBe, Nβe) being inside different *Nephromyces* species, we do see some similar patterns to the dual endosymbiont example in glassy winged sharpshooters. While we do not see the single pathway integration where one symbiont produces fabF and the other produces the remainder of the pathway. The lack of overlapping functions seems to indicate that despite being in different *Nephromyces* species, that the close proximity in the renal sac is sufficient to allow for metabolite exchange between bacterial endosymbionts in conspecific *Nephromyces* species. The result is

completely unexpected and represents an unusual evolutionary quirk for the community inside the renal sac. If the renal sac community is in close enough to allow for the development of non-overlapping functions in bacterial endosymbionts in different species, it must be concluded that conspecific *Nephromyces* species are frequently exchanging metabolites. Based on our isolation and culturing experiments, we hypothesize that *Nephromyces* may be incapable of existing in isolation without conspecific *Nephromyces* species which contain a different type of bacterial endosymbiont than their own. We have not found a *Nephromyces* species containing two different types of bacterial endosymbionts, however we can't conclusively say that *Nephromyces* species with dual endosymbionts don't exist.

It remains unclear why a system dependent on conspecifics would develop when maintaining multiple endosymbionts would eliminate the need for competing *Nephromyces* species and guarantee that whatever host *Nephromyces* infected would be able to be colonized independent of conspecifics. Perhaps the cost of maintaining multiple endosymbionts is greater than the cost of sharing. Indeed we have found that some *Nephromyces* species do not maintain any endosymbiont and presumably parasitize the community, i.e. relying on the products of other *Nephromyces'* bacterial endosymbionts. In order for such a system to develop any given *Nephromyces* species must colonize a renal sac where there will be complimentary *Nephromyces* species. The rate at which this happens needs to be greater than the cost of maintaining two endosymbionts

otherwise we would presumably see dual endosymbiont *Nephromyces* species.

A number of factors seem to contribute to the development of this co-dependent species complex. First, the infection prevalence of *Nephromyces* is nearly 100% nearly year round. Second, *Nephromyces* seems capable of infecting at any point in the tunicates adult life stage. Third, we have uncovered a staggering amount of *Nephromyces* diversity. Some individual *M. manhattensis* contain as many as 11 distinct *Nephromyces* species. Based on amplicon sequencing data there is an estimated 60 species of *Nephromyces* in Greenwich Bay, RI alone. The extreme amount of both species and sequence diversity in *Nephromyces* is not observed in *Cardiosporidium* and is likely connected to the unusual renal sac community dynamics. With so many species, each with a lineage of vertically inherited bacterial endosymbionts, we predict that even bacterial endosymbionts of the same type may differ widely in their metabolic capabilities. Presumably bacterial endosymbionts, even within the same taxa, could contribute different metabolites to the renal sac community. Given the tremendous amount of diversity, our sequencing of just a few of the different bacterial endosymbionts is insufficient to develop a complete picture of the intricacies of this system. It is likely that different bacterial endosymbionts within the same type may differ in metabolic capabilities.

Adding to our uncertainties we do not currently know if there are any genetic barriers preventing reproduction between different *Nephromyces* species. Sexual reproduction occurs inside the renal sac in the presence of multiple other

Nephromyces species; interbreeding between species seems likely, unless there are strong genetic barriers between species. Indeed, the proximity and the interdependence of the system in general seems to indicate a great deal of interspecific breeding.

As our genomic assemblies of *Nephromyces* are incomplete, due in large part to the difficulties in assembling a metagenome of closely related species, we are not able to say how *Nephromyces* unusual epidemiology, environment, and community composition has affected its genome. Given the difficulties with *Nephromyces* sequencing and assembling, the genome of *Cardiosporidium* is a more attractive target. There are plans to sequence the genome, but currently we do not have any genomic data for *Cardiosporidium*. We do have good transcriptomic data for both *Nephromyces* and *Cardiosporidium,* meaning we can compare the protein coding genes of *Nephromyces* and *Cardiosporidium* to each other, as well as to other apicomplexans.

Nephromyces and *Cardiosporidium* have very similar metabolic capabilities based on KEGG pathway analysis. This includes purine degradation to glyoxylate and then to glycine and pyruvate, which shows that this pathway was being utilized before *Nephromyces* colonized the renal sac. Based on expression data, this pathway is more important for *Nephromyces* than *Cardiosporidium.* Expression of these genes in *Cardiosporidium* is still high (80-90th percentile, the highest expression of all *Cardiosporidium* genes). Such high expression suggests that *Cardiosporidium* has access to high concentrations of uric acid inside

tunicate blood cells, which may still represent an important source of carbon and nitrogen.

Despite the differences in epidemiology, habitat, and life histories between *Nephromyces* and *Cardiosporidium,* their metabolomes are quite similar, including the complete apicomplexan infection machinery. While *Nephromyces* may have a mutualistic relationship with its host, it is entering the renal sac like an apicomplexan parasite. Related to this infection machinery, both *Cardiosporidium* and *Nephromyces* have the same known dense granule, microneme, and rhoptry proteins. Proteins from these organelles have been shown in other apicomplexans to be important for invasion, immune evasion, and host manipulation. Based on Orthofinder analysis we find surprisingly few lineage-specific genes that might be involved in dealing with an ascidian immune system. We also find few genes without orthologous in either *Nephromyces* or *Cardiosporidium* compared to other apicomplexans. It is possible that these genes without orthologous are involved in the specific challenges imposed by intracellular ascidian life cycle in *Cardiosporidium* and the renal sac for *Nephromyces*. However, because these genes do not have known orthologs studied in other species, we are unable to determine function bioinformatically.

Nephromyces and *Cardiosporidium* have a large number of protein coding genes and metabolic capabilities. With respect to gene number and function *Nephromyces*/*Cardiosporidium* are similar to *Toxoplasma.* This is surprising because phylogenetically *Nephromyces*/*Cardiosporidium* belong to

hematozoa, the lineage that contains *Plasmodium*, *Babesia*, and *Theileria*. The other members of hematozoa have lost peroxisomes, which we have demonstrated in *Nephromyces*/*Cardiosporidium.* Additionally, hematozoa show more metabolic reduction. *Babesia*, and *Theileria* in particular have the smallest genomes of any apicomplexan and the fewest number of protein coding genes (Kappmeyer et al. 2012; Brayton et al. 2007). The placement of *Nephromyces*/*Cardiosporidium* in hematozoa adds to the growing body of support that, these gene losses in Apicomplexa were gradual and lineage specific, with losses occurring later than predicted. These gene losses often occur in parallel in different lineages confounding assumptions based on the most parsimonious solutions. Indeed we see in nearly every apicomplexan lineage, lineage specific orthologs present only in chromerids and lost in all other apicomplexans. In the case of *Nephromyces*/*Cardiosporidium* lineage specific orthologs include purine degradation and *de novo* purine biosynthesis. These retained pathways appear to have particular significance to *Nephromyces*/*Cardiosporidium,* but lineage specific gene retention is consistent with the general patterns observed across the phyla Apicomplexa.

This work represents a step toward fully understanding the complexities of this unusual system, but leaves many questions unresolved. First, sequencing the genome of both *Cardiosporidium* and *Cardiosporidiums* bacterial endosymbiont would allow for more robust comparison to *Nephromyces Rickettsia* endosymbiont. This would provide a better understanding of the

evolutionary history of both *Nephromyces* and its endosymbionts. Secondly, the biochemical pathways were based on bioinformatics with minimal confirmation at the protein level, the presented pathways need to be confirmed. Another step would be to show that uric acid is central to the metabolism of *Nephromyces.* A potential method to demonstrate this pathway is by injecting isotope labeled uric acid into the renal sac, and then using a new method for identifying the proteins from a specific organism in a metaproteomic sample (Kleiner et al. 2018). If this could be adapted to this system we could potentially confirm uric acid as the primary carbon and nitrogen source for *Nephromyces*, and determine the metabolites exchanged with the bacterial endosymbiont. This could also show if any of the carbon or nitrogen from uric acid makes its way across the renal wall back to the tunicate. If useful metabolites are exported or leaked out of the renal sac this would be the best support yet that the relationship between *Nephromyces* and *Molgula* is in fact mutualistic.

APPENDICES

Appendix Table 1) Nephromyces peroxisomal-related genes Identified PeroxDB and KAAS. Complete transcripts were run through Wolf PSORTII, Ppero, TargetP, Topcons and Predotar to identify possible signal motifs inclding PTS1

Gene	PSORTII	Ppero pts1	TargetP	topcons	predotar
MLS	cyto: 8, extr: 5, chlo: 1	No	_	Yes	possibly plastid
MLS	cyto: 11, nucl: 2, vacu: 1	No	_	Yes	none
MLS	chlo: 5.5, chlo_mito: 5.3	No	M	No	possibly mitochondrial
MLS	chlo: 9, mito: 4, plas: 1	No	M	No	possibly mitochondrial
MLS	cyto: 8, nucl: 3, vacu: 1, g	No	_	No	none
MLS	cyto: 7, nucl: 3, chlo: 2, g	No	_	No	none
MLS	cyto: 8, extr: 5, chlo: 1	No	_	No	possibly plastid
MLS	chlo: 6, cyto: 3, mito: 3, r	No	M	No	possibly mitochondrial
MLS	cyto: 6, extr: 3, chlo: 1, n	No	_	No	possibly plastid
MLS	chlo: 5.5, chlo_mito: 5.3	No	M	No	possibly mitochondrial
MDH	extr: 9, chlo: 2, nucl: 1, c	No	S	No	none
PEX1	chlo: 5, nucl: 2, cyto: 2, p	No	S	No	none
PEX2	nucl: 13.5, cyto_nucl: 7.5	No	_	No	none
PEX2	nucl: 6, cyto: 5, mito: 1, e	No	_	No	none
PEX4	cyto: 5, chlo: 3, nucl: 2, p	No	_	No	none
PEX4	cyto: 5, chlo: 3, nucl: 2, p	No	_	No	none
PEX5	mito: 7, chlo: 4, nucl: 2, c	No	M	No	possibly mitochondrial
PEX5	plas: 5, nucl_plas: 4.5, cy	No	_	No	none
PEX7	nucl: 7, cyto: 7	No	_	No	none
PEX7	nucl: 14	No	_	No	none
PEX7	nucl: 14	No	_	No	none
PEX7	extr: 7, nucl: 4, chlo: 2, c	No	_	No	none
PEX7	extr: 7, nucl: 4, chlo: 2, c	No	_	No	none
PEX10	nucl: 12, extr: 2	No	S	No	possibly ER
PEX12	plas: 7, E.R.: 4, cyto: 1, m	No	_	No	none
PEX12	nucl: 3, extr: 3, vacu: 2, E	No	_	No	none
PEX12	nucl: 3, extr: 3, vacu: 2, E	No	_	No	none
PEX14	nucl: 10, cyto: 3, plas: 1	No	_	No	none
PEX14	pero: 11, cyto_nucl: 2, nu	No	_	No	none
MPV17	E.R.: 5, plas: 3, chlo: 2, va	Perhaps yes	_	No	possibly plastid
MPV17	plas: 8, vacu: 4, E.R.: 2	No	S	No	none
MPV17	chlo: 4, nucl: 3, vacu: 2, E	No	S	No	ER
ABCD	plas: 9, vacu: 2, nucl: 1, c	No	_	No	none
ABCD	plas: 5, vacu: 4, E.R.: 2, n	No	_	No	none
ABCD	plas: 9, mito: 2, chlo: 1, E	No	_	No	none
ABCD	chlo: 4, vacu: 4, plas: 2, r	No	_	No	none
ABCD	chlo: 4, vacu: 4, plas: 2, r	No	_	No	none
ABCD3	chlo: 7, cyto: 3, mito: 3, p	No	_	No	none

ABCD3	plas: 8, vacu: 3, E.R.: 2, r	No	_	No	possibly ER
ACAA1	cyto: 8, chlo: 2, nucl: 1, n	No	_	No	none
ACAA1	chlo: 4, cyto: 3.5, golg: 3,	No	_	No	none
ACAA1	chlo: 4, cyto: 3.5, golg: 3,	No	_	No	none
ACAA1	cyto: 8, chlo: 2, nucl: 1, n	No	_	No	none
ACAA1	chlo: 11.5, chlo_mito: 7,	No	M	No	mitochondrial
ACAA1	chlo: 9, mito: 3, nucl: 2	No	_	No	none
ACAA1	chlo: 11.5, chlo_mito: 7,	No	M	No	mitochondrial
ACAA1	chlo: 5, nucl: 4, cyto: 2, n	No	_	No	none
ACAA1	cyto: 6, chlo: 4, nucl: 2, p	No	_	No	none
ACAA1	cyto: 6, chlo: 4, nucl: 2, p	No	_	No	none
ACAA1	chlo: 5, nucl: 4, cyto: 2, n	No	_	Yes	none
ACAA1	chlo: 5, nucl: 4, cyto: 2, n	No	_	No	none
ACAA1	nucl: 7, cyto: 4, mito: 2, c	No	_	No	none
ACAA1	nucl: 7, cyto: 4, mito: 2, c	No	_	No	none
ACAA1	nucl: 7, cyto: 4, mito: 2, c	No	_	No	none
ACAA1	nucl: 7, cyto: 4, mito: 2, c	No	_	No	none
ACAA1	nucl: 7, cyto: 4, mito: 2, c	No	_	No	none
ACAA1	nucl: 7, cyto: 4, mito: 2, c	No	_	No	none
ACAA1	nucl: 7, cyto: 4, mito: 2, c	No	_	No	none
ACAA1	nucl: 6, mito: 4, cyto: 2, c	No	_	No	none
ACAA1	nucl: 6, mito: 4, cyto: 2, c	No	_	No	none
ACOX	cyto: 9, nucl: 2, mito: 2, p	No	_	No	none
ACOX	cyto: 7.5, cyto_nucl: 5, E.	No	_	No	none
ACOX	cyto: 7.5, cyto_nucl: 5, E.	No	_	No	none
ACOX	cyto: 9, nucl: 2, mito: 2, p	No	_	No	none
ACOX	pero: 11, cyto: 2, golg: 1	No	_	No	none
ACOX	pero: 11, cyto: 2, golg: 1	Yes	_	No	none
ACOX	pero: 11, cyto: 2, golg: 1	Yes	_	No	none
ACOX	pero: 11, cyto: 2, golg: 1	Yes	_	No	none
ACOX	pero: 11, cyto: 2, golg: 1	Yes	_	No	none
ACOX	pero: 11, cyto: 2, golg: 1	Yes	_	No	none
ACOX	pero: 11, cyto: 2, golg: 1	Yes	_	No	none
AGXT	chlo: 11, mito: 3	No	M	No	possibly mitochondrial
AGXT	cyto: 7, mito: 4, chlo: 1, e	No	_	No	none
AGXT	cyto: 8, mito: 4, chlo: 1, e	No	_	No	none
AGXT	chlo: 9, mito: 2, pero: 2,	No	M	No	possibly mitochondrial
AGXT	chlo: 9, mito: 2, pero: 2,	No	M	No	possibly mitochondrial
AK	nucl: 8, cyto: 2.5, cyto_E.	No	_	No	none
AK	nucl: 8, cyto: 2.5, cyto_E.	No	_	No	none
AK	cyto: 7.5, cyto_E.R.: 5, ch	No	_	No	none
AK	cyto: 5, nucl: 4, chlo: 2, n	No	_	No	none
ANT	nucl: 5, chlo: 4, cyto: 3, e	No	_	No	none

ANT	nucl: 6, cyto: 4, chlo: 2, e	No	_	No	none
ANT	nucl: 5, cyto: 3.5, cyto_E.	No	_	No	none
ANT	nucl: 6, cyto: 4, chlo: 2, e	No	_	No	none
ANT	nucl: 7, cyto: 3, chlo: 1, n	No	_	No	none
CAT	pero: 12, cyto: 1, golg: 1	Yes	_	No	none
CAT	pero: 12, cyto: 1, golg: 1	Yes	_	No	none
CAT	pero: 12, cyto: 1, golg: 1	Yes	_	No	none
CAT	nucl: 7, pero: 2, chlo: 1, r	No	_	No	none
CAT	nucl: 7, cyto: 2, pero: 2, c	No	_	No	none
CAT	nucl: 7, cyto: 2, pero: 2, c	No	_	No	none
CAT	nucl: 7, cyto: 2, pero: 2, c	No	_	No	none
CHY	chlo: 8, nucl: 2, extr: 2, p	No	S	No	possibly ER
CHY	chlo: 8, nucl: 2, extr: 2, p	No	S	No	possibly ER
CHY	chlo: 10, mito: 3, nucl: 1	No	M	No	none
CHY	chlo: 10, mito: 3, nucl: 1	No	M	No	none
CHY	nucl: 6, cyto: 3, chlo: 1, n	No	_	No	none
CPK	cyto: 12, nucl: 1, pero: 1	No	_	No	none
DECR2	cyto: 5, chlo: 4, extr: 2, n	No	_	No	none
DECR2	chlo: 6.5, chlo_mito: 4.5,	Yes	S	No	ER
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 10, cyto: 3, golg: 1	Yes	_	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, chlo: 3, cyto: 3, g	Yes	_	No	none
DECR2	pero: 7, chlo: 3, cyto: 3, g	Yes	_	No	none
DECR2	pero: 7, chlo: 3, cyto: 3, g	Yes	_	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, chlo: 3, cyto: 3, g	Yes	_	No	none
DECR2	cyto: 5, chlo: 4, extr: 2, n	No	_	No	none
DECR2	chlo: 13, plas: 1	Perhaps yes	S	No	possibly plastid
DECR2	chlo: 11, extr: 2, cyto: 1	Perhaps yes	S	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	Yes	none
DECR2	pero: 10, cyto: 3, golg: 1	Yes	_	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	pero: 7, cyto: 3.5, cyto_n	Yes	M	No	none
DECR2	chlo: 13, plas: 1	Perhaps yes	S	No	possibly plastid
DECR2	chlo: 11, extr: 2, cyto: 1	Perhaps yes	S	No	none
DECR2	pero: 11, chlo: 1, nucl: 1,	Perhaps yes	_	No	none

DECR2	pero: 11, chlo: 1, nucl: 1,	Perhaps yes	_	No	none
DHRS4	pero: 9, cyto: 4, nucl: 1	Yes	_	No	none
DHRS4	pero: 9, cyto: 4, nucl: 1	Yes	_	No	none
DHRS4	chlo: 7, E.R.: 2.5, extr: 2,	No	S	No	ER
DHRS4	cyto: 12, mito: 1, cysk_nu	No	_	No	none
DHRS4	cyto: 10, chlo: 2, nucl: 1.5	No	_	No	none
DHRS4	extr: 5, chlo: 3, vacu: 3, n	No	S	No	ER
DHRS4	extr: 5, chlo: 3, vacu: 3, n	No	S	No	ER
DHRS4	chlo: 6, E.R.: 2.5, extr: 2,	No	S	No	ER
DHRS4	chlo: 7, E.R.: 2.5, extr: 2,	No	S	No	ER
DHRS4	chlo: 8, pero: 3, cyto: 2.5	No	_	No	none
DHRS4	pero: 6, cyto: 3.5, cyto_n	Yes	_	No	none
DHRS4	pero: 11, cyto: 3	Yes	_	No	none
DHRS4	chlo: 12, extr: 2	No	_	No	none
DHRS4	chlo: 8, cyto: 2.5, cyto_nu	Yes	_	No	none
DJP1	cyto: 8, nucl: 4, mito: 1, p	No	_	No	none
DJP1	cyto: 8, nucl: 4, mito: 1, p	No	_	No	none
DJP1	nucl: 10, plas: 2, chlo: 1,	No	_	Yes	none
DJP1	nucl: 10, cyto: 4	No	_	No	none
DJP1	nucl: 6, cyto: 6, plas: 1, p	No	_	No	none
DJP1	nucl: 12, cyto: 1, plas: 1	No	_	No	none
E1.3.3.6	chlo: 6, nucl: 4, cyto: 2, n	No	M	No	none
E1.3.3.6	extr: 7, chlo: 4, vacu: 1, g	No	_	No	possibly mitochondrial
E1.3.3.6	mito: 7, chlo: 5, nucl: 1, c	No	_	No	none
E1.3.3.6	chlo: 7, mito: 4, nucl: 2, p	No	_	No	none
ECH	mito: 8, nucl: 5, chlo: 1	No	_	No	none
ECH	mito: 8, nucl: 5, chlo: 1	No	_	No	none
ECH	pero: 4, E.R.: 3, plas: 2, cy	Yes	_	No	none
ECH	pero: 5, E.R.: 3, plas: 2, n	Yes	_	No	none
ECH	pero: 5, E.R.: 3, plas: 2, n	Yes	_	No	none
ECH	cyto: 6, E.R.: 3, nucl: 2, cl	No	_	No	none
ECH	pero: 5, E.R.: 3, plas: 2, cy	Yes	_	No	none
ECH	chlo: 5, cyto: 3.5, cyto_nu	No	_	No	none
ECH	E.R.: 3, pero: 3, nucl: 2.5,	Yes	_	No	none
ECH	chlo: 4, cyto: 4, extr: 3, p	No	_	No	none
ECH	E.R.: 3, pero: 3, nucl: 2.5,	No	_	No	none
ECH	chlo: 4, cyto: 4, plas: 2, e	Yes	_	No	none
ECH	mito: 8, nucl: 5, chlo: 1	No	_	No	none
ECH1	cyto: 5, extr: 4, E.R.: 2.5,	No	_	No	none
ECH1	pero: 4, chlo: 3, E.R.: 3, p	Yes	_	No	none
ECH1	cyto: 4, extr: 3, E.R._plas	No	_	No	none
ECH1	chlo: 6, cyto: 3.5, cyto_nu	No	_	No	none
ECH1	pero: 4, chlo: 3, E.R.: 3, p	Yes	_	No	none

ECH2	pero: 12, cyto: 2	Yes	_	No	none
ECH2	pero: 12, cyto: 2	Yes	_	No	none
fabG	pero: 6, cyto: 3.5, cyto_n	Yes	M	No	none
fabG	pero: 6, cyto: 3.5, cyto_n	Yes	M	No	none
FACL	nucl: 4, mito: 4, chlo: 3, c	No	_	No	none
FACL	nucl: 4, mito: 4, chlo: 3, c	No	_	No	none
FACL	nucl: 4, mito: 4, chlo: 3, c	No	_	No	none
FACL	nucl: 4, mito: 4, chlo: 3, c	No	_	No	none
FACL	chlo: 6, extr: 4, vacu: 1.5,	No	_	No	none
FACL	chlo: 6, extr: 4, vacu: 1.5,	No	_	No	none
FACL	nucl: 4, mito: 4, chlo: 3, c	Yes	_	No	none
FACL	nucl: 4, mito: 4, chlo: 3, c	No	_	No	none
FAR	chlo: 11, mito: 2, nucl: 1	No	S	No	none
FBPA	cyto: 7, pero: 3, chlo: 2, r	No	_	No	possibly mitochondrial
FBPA	cyto: 7, pero: 3, chlo: 2, r	No	_	No	possibly mitochondrial
FBPA	cyto: 7, pero: 3, chlo: 2, r	No	_	No	possibly mitochondrial
FBPA	cyto: 7, pero: 3, chlo: 2, r	No	_	No	possibly mitochondrial
FBPA	cyto: 7, pero: 3, chlo: 2, r	No	_	No	possibly mitochondrial
FBPA	plas: 4.5, nucl_plas: 4, ch	No	_	No	none
FR	cyto: 6, cysk: 6, nucl: 2	Perhaps yes	_	No	none
FR	cyto: 6, cysk: 6, nucl: 2	No	_	No	none
FR	chlo: 8, vacu: 3, nucl: 2, r	No	S	No	none
FR	chlo: 8, vacu: 3, nucl: 2, r	No	S	No	none
FR	chlo: 4, nucl: 3.5, cyto_nu	No	_	No	possibly mitochondrial
FR	chlo: 3, extr: 3, nucl: 2, cy	No	S	No	none
FR	chlo: 10, cyto: 2, nucl: 1,	No	S	No	ER
FR	chlo: 10, cyto: 2, nucl: 1,	No	S	No	ER
FR	chlo: 3, extr: 3, nucl: 2, cy	No	S	No	none
FR	chlo: 7, cyto: 2, plas: 2, n	No	S	No	possibly ER
FR	chlo: 7, cyto: 2, plas: 2, n	No	S	No	possibly ER
G6PI	plas: 8, cyto: 2, E.R.: 2, ch	No	_	No	none
G6PI	plas: 8, cyto: 2, E.R.: 2, ch	No	_	No	none
GAPDH	cyto: 14	No	M	No	possibly mitochondrial
GAPDH	pero: 6, cyto: 3.5, cyto_n	Perhaps yes	M	No	none
GAPDH	cyto: 5.5, cyto_nucl: 3.5,	No	M	No	none
GAPDH	pero: 6, cyto: 3.5, cyto_n	Perhaps yes	M	No	none
GAPDH	cyto: 14	No	M	No	none
GAPDH	cyto: 13, mito: 1	No	M	No	none
GAPDH	cyto: 13, mito: 1	No	M	Yes	none
GAPDH	cyto: 14	No	M	No	possibly mitochondrial
GAPDH	cyto: 14	No	M	No	none
GAPDH	cyto: 5.5, cyto_nucl: 3.5,	No	M	No	none
GCDH	mito: 4, chlo: 3, nucl: 2, c	No	_	No	none

GK	cysk: 10, cyto: 3, nucl: 1	No	_	No	none
GNPAT	pero: 4, E.R.: 3, plas: 2, c	Perhaps yes	S	No	ER
GNPAT	E.R.: 3, chlo: 2, nucl: 2, cy	No	S	No	ER
GNPAT	pero: 11, cyto: 2, nucl: 1	Yes	_	No	none
GNPAT	cyto: 4, E.R.: 3, nucl: 2, m	No	S	Yes	ER
GNPAT	cyto: 4, E.R.: 3, nucl: 2, m	No	S	No	ER
GNPAT	pero: 4, E.R.: 3, nucl: 2, p	No	_	No	none
GNPAT	chlo: 6, E.R.: 3, plas: 2, pe	Perhaps yes	S	No	none
GOT	cyto: 10, nucl: 1, mito: 1,	No	_	Yes	none
GOT	cyto: 11, mito: 1, cysk: 1,	No	_	No	none
GOT	cysk: 8, cyto: 4, nucl: 2	No	_	No	none
GOT	cyto: 8, cysk: 3.5, cysk_p	No	_	No	none
GOT	cyto: 10, cysk: 3, mito: 1	No	_	No	none
GOT	chlo_mito: 7.33333, mitc	No	M	No	possibly mitochondrial
GPD	pero: 8, cyto: 3.5, cyto_n	Yes	_	No	none
GPD	pero: 8, cyto: 3.5, cyto_n	Yes	_	No	none
GPD	cyto: 4.5, nucl: 3, cyto_E.	No	_	No	none
GPD	nucl: 5, chlo: 3, cyto: 3, p	No	_	No	possibly plastid
GPD	cyto: 6, chlo: 4, nucl: 2, g	No	_	No	possibly plastid
GPD	cyto: 7, chlo: 5, vacu: 1, c	No	_	No	none
GSTK1	chlo: 6, cyto: 3.5, cyto_n	No	_	Yes	none
GSTK1	pero: 4, cyto: 3.5, cyto_n	No	_	No	none
GSTK1	pero: 10, nucl: 2, chlo: 1,	Yes	_	No	none
GSTK1	chlo: 4, pero: 4, cyto_nuc	Perhaps yes	_	No	none
GSTK1	pero: 4, cyto: 3.5, cyto_n	No	_	No	none
GSTK1	chlo: 6, cyto: 3.5, cyto_n	No	_	No	none
GSTK1	pero: 10, nucl: 2, chlo: 1,	No	_	No	none
GTO	nucl: 4, cyto: 4, mito: 4, c	No	_	No	possibly mitochondrial
GTO	cyto: 13, nucl: 1	No	M	No	possibly plastid
GTO	mito: 6, chlo: 3, cyto: 3, r	No	M	No	plastid
HAO	cysk: 12, chlo: 1, cyto: 1	No	_	No	none
HAO	cysk: 12, chlo: 1, cyto: 1	No	_	No	none
HAO	cysk: 12, chlo: 1, cyto: 1	No	_	No	none
HAO	cyto: 6, cysk: 3, chlo: 2, n	No	_	No	none
HAO	cysk: 9, nucl: 4, chlo: 1	No	_	No	none
HAO	cyto: 8, cysk: 2.5, cysk_p	No	_	No	none
HAO	cysk: 13, cyto: 1	No	_	Yes	none
HAO	cyto: 8, mito: 3, chlo: 2, p	No	_	Yes	none
HAO	cyto: 8, mito: 3, chlo: 2, p	No	_	No	none
HAO	cyto: 12, nucl: 2	No	_	No	none
HEX1	cyto: 9, E.R.: 3, golg: 2	No	_	No	none
HEX1	cyto: 9, nucl: 2, cysk: 2, g	No	_	No	none
HEX1	cyto: 6, cysk: 4, extr: 2, c	No	_	No	none

HEX1	cyto: 6, cysk: 4, extr: 2, c	No	_	No	none
HEX1	cyto: 10, extr: 2, cysk: 1,	No	_	No	none
HEX1	nucl: 4, cyto: 4, extr: 2, c	No	_	No	none
HEX1	nucl: 4, cyto: 4, extr: 2, c	No	_	No	none
HEX1	cyto: 8, E.R.: 3, golg: 2, m	No	_	No	none
HEX1	cyto: 8, golg: 3, E.R.: 2, cl	No	_	No	none
HEX1	cyto: 8, E.R.: 3, golg: 2, m	No	_	No	none
HEX1	cyto: 8, golg: 3, E.R.: 2, cl	Yes	_	No	none
HEX1	cyto: 9, E.R.: 3, golg: 2	No	_	No	none
HK	cyto: 7, chlo: 3, nucl: 3, g	Perhaps yes	_	No	none
HK	cyto: 4, chlo: 3, nucl: 2, e	No	_	No	none
HK	cyto: 8, nucl: 2, extr: 2, c	No	_	No	none
HK	cyto: 8, chlo: 2, cysk_pla	No	_	No	none
HK	chlo: 5.5, nucl: 5, chlo_m	Perhaps yes	_	No	none
HK	cyto: 8, nucl: 2, extr: 2, c	Perhaps yes	_	No	none
HK	cyto: 7, chlo: 3, nucl: 3, g	Perhaps yes	_	No	none
HMGCL	cyto: 9, cysk: 3, nucl: 1, g	No	_	No	none
HMGCL	cyto: 9, cysk: 3, nucl: 1, g	No	_	No	none
HMGCL	cyto: 5, extr: 5, chlo: 2, n	No	_	No	none
HMGCL	mito: 7, chlo: 6, nucl: 1	No	M	No	possibly mitochondrial
HMGCL	cyto: 9, cysk: 3, nucl: 1, g	No	_	No	none
HMGCL	cyto: 10, nucl: 2, vacu: 1,	No	_	No	none
HMGCL	cyto: 10, nucl: 2, vacu: 1,	No	_	No	none
HMGCL	cyto: 10, nucl: 2, vacu: 1,	No	_	No	none
HMGCL	cyto: 10, nucl: 2, vacu: 1,	No	_	No	none
HPR	cyto: 10, nucl: 1, extr: 1,	No	_	No	none
HSD17B4	extr: 8, mito: 2, E.R.: 1.5,	No	S	No	ER
HSD17B4	chlo: 7, extr: 3, cyto: 2, n	No	_	No	possibly plastid
HSD17B4	pero: 10, chlo: 2, nucl: 1.	Yes	_	No	possibly plastid
HSD17B4	pero: 10, chlo: 2, nucl: 1.	Perhaps yes	_	No	possibly plastid
HSD17B4	extr: 5, chlo: 4, cyto: 4, n	No	_	No	possibly plastid
HSD17B4	pero: 9, nucl: 2.5, cyto_n	Yes	_	No	mitochondrial
HSD17B4	pero: 8, nucl: 3.5, cyto_n	No	_	No	none
HSD17B4	pero: 8, nucl: 3.5, cyto_n	No	_	No	none
HSD17B4	pero: 10, cyto: 2, nucl: 1,	Perhaps yes	_	No	none
HSD17B4	pero: 12, cyto: 2	Perhaps yes	_	No	none
HSD17B4	pero: 12, cyto: 2	Perhaps yes	_	No	none
HSD17B4	pero: 12, cyto: 2	No	_	No	none
HSD17B4	pero: 12, cyto: 2	Yes	_	No	none
HSD17B4	pero: 13, cyto: 1	Yes	_	No	none
HSD17B4	pero: 13, cyto: 1	Yes	_	No	none
HSD17B4	pero: 10, cyto: 3, nucl: 1	No	_	No	none
HSD17B4	pero: 9, cyto: 3, nucl: 2	Yes	_	No	none

IDI	cyto: 7, mito: 2, chlo: 1, r	No	_	No	none
katE	nucl: 5, mito: 5, cyto: 1, p	No	_	No	none
katE	nucl: 5, mito: 5, cyto: 1, p	No	_	No	none
katE	pero: 12, cyto: 2	Yes	_	No	none
katE	nucl: 9, mito: 4, extr: 1	No	_	No	none
katE	nucl: 5, mito: 5, cyto: 1, p	No	_	No	none
katE	nucl: 11, mito: 2, plas: 1	No	_	No	none
katE	pero: 12, cyto: 2	Yes	_	No	none
katE	nucl: 11, mito: 2, plas: 1	No	_	No	none
katE	nucl: 11, mito: 3	No	_	No	none
katE	nucl: 5, mito: 5, cyto: 1, p	No	_	No	none
katE	nucl: 5, mito: 5, cyto: 1, p	No	_	No	none
katE	nucl: 5, mito: 5, cyto: 1, p	No	_	No	none
katE	nucl: 5, cyto: 2, plas: 2, p	No	_	No	none
katE	cyto: 7, nucl: 2, pero: 2, r	No	_	No	none
LYS4	cyto: 7, chlo: 3, nucl: 2, v	No	S	No	none
LYS4	nucl: 8, cyto: 5, extr: 1	No	_	No	none
LYS4	cyto: 8, chlo: 3, nucl: 2, c	No	_	No	none
LYS4	cyto: 11, nucl: 2, cysk: 1	No	_	No	none
LYS4	nucl: 7, cyto: 4, chlo: 2, n	No	_	No	mitochondrial
LYS4	nucl: 5, chlo: 4, cyto: 4, v	No	_	No	none
LYS4	cyto: 7, chlo: 3, nucl: 2, v	No	S	No	none
LYS4	chlo: 8, cyto: 5, cysk: 1	No	_	No	possibly plastid
LYS4	mito: 7.5, chlo: 6, cyto_n	No	M	No	mitochondrial
LYS4	cyto: 5, mito: 4, chlo: 2, r	No	_	No	mitochondrial
LYS4	cyto: 5, mito: 4, chlo: 2, r	No	_	No	mitochondrial
LYS4	cyto: 8, chlo: 5, mito: 1	No	_	No	possibly mitochondrial
LYS4	nucl: 4, cyto: 4, chlo: 3, n	No	_	No	mitochondrial
LYS4	chlo: 10.5, chlo_mito: 7.5	No	M	No	mitochondrial
MDAR	mito: 8, chlo: 6	Yes	M	No	mitochondrial
MDAR	cysk: 8, nucl: 4, chlo: 1, c	No	_	No	none
MDAR	mito: 8, chlo: 6	No	M	No	mitochondrial
MDAR	cysk: 8, nucl: 4, chlo: 1, c	No	_	No	none
MYA2	cyto: 11, nucl: 2, vacu: 1	No	_	No	none
MYA2	cyto: 11, nucl: 2, vacu: 1	No	_	No	none
NUDT12	chlo: 7, extr: 3, cyto: 2, n	No	_	No	none
OPDC-PRT	extr: 9, cyto: 2, nucl: 1, m	No	_	No	none
OPDC-PRT	extr: 9, cyto: 2, nucl: 1, m	No	_	No	none
PAOX	cyto: 8, nucl: 3, chlo: 1, p	No	M	No	possibly mitochondrial
PAOX	nucl: 5, chlo: 4, cyto: 3, n	Yes	_	No	none
PAOX	nucl: 9, chlo: 2, cysk: 2, c	No	_	No	none
PDCR	pero: 10, chlo: 2, cyto: 1.	Perhaps yes	_	No	none
PDCR	pero: 11, cyto: 2, nucl: 1	Yes	_	No	none

HSD17B4	pero: 6, E.R.: 3, plas: 2, c	Yes	_	No	none
HSD17B4	pero: 12, nucl: 1, cyto: 1	No	_	No	none
HSD17B4	pero: 8, nucl: 3.5, cyto_n	No	_	No	none
HSD17B4	pero: 12, nucl: 1, cyto: 1	No	_	No	none
HSD17B4	pero: 12, cyto: 2	Yes	_	No	none
HSD17B4	pero: 12, cyto: 2	Perhaps yes	_	No	none
HXGPRT	cyto: 7, nucl: 3, cysk: 2, c	No	_	No	none
HXGPRT	cyto: 7, nucl: 3, cysk: 2, c	No	_	No	none
HXGPRT	cyto: 7, nucl: 3, cysk: 2, c	No	_	No	none
HXGPRT	cyto: 7, nucl: 3, cysk: 2, c	No	_	No	none
HXGPRT	cyto: 7, nucl: 3, cysk: 2, c	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	Perhaps yes	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	Yes	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	Perhaps yes	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
HXGPRT	cyto: 11, nucl: 1, extr: 1,	No	_	No	none
I5PDH	cysk: 11, cyto: 3	No	_	No	none
I5PDH	cysk: 11, cyto: 3	No	_	No	none
IDH	cysk: 11, cyto: 3	No	_	No	none
IDH	chlo_mito: 6, chlo: 5.5, n	No	_	No	none
IDH	chlo_mito: 6, chlo: 5.5, n	No	_	Yes	none
IDH	cysk: 14	No	_	No	none
IDH	cysk: 14	No	_	No	none
IDH	pero: 13, cyto: 1	Yes	_	No	none
IDH	pero: 13, cyto: 1	Yes	_	No	none
IDH	cysk: 11, cyto: 3	No	_	No	none
IDH1	nucl: 6, cyto: 3, chlo: 2, c	No	_	No	none
IDH1	extr: 11, cysk_plas: 1.333	No	S	No	possibly ER
IDH1	extr: 10, cysk_plas: 1.333	No	S	No	possibly ER
IDH1	pero: 12, cyto: 2	Yes	_	No	none
IDH1	cysk: 9, chlo: 2, cyto: 2, n	No	_	No	none
IDH1	cyto: 7.5, cyto_E.R.: 4.5,	No	_	No	none
IDH1	cyto: 5.5, chlo: 5, cyto_E.	No	_	No	none
IDI	chlo: 5, cyto: 3, mito: 3, r	No	_	No	possibly mitochondrial
IDI	chlo: 5, mito: 5, cyto: 2, r	No	_	No	mitochondrial

PDCR	pero: 11, cyto: 2, nucl: 1	Yes	_	No	none
PECR	extr: 4, mito: 3, cyto: 2.5	No	S	No	ER
PECR	extr: 4, mito: 3, cyto: 2.5	No	S	No	ER
PEPCK	chlo: 3, mito: 2, plas: 2, E	No	_	No	none
PEPCK	chlo: 3, mito: 2, plas: 2, E	No	_	No	none
PEPCK	chlo: 6, extr: 3, vacu: 3, r	No	_	No	none
PEPCK	chlo: 3, extr: 3, cyto: 2, E	No	_	No	none
PEPCK	chlo: 6, extr: 3, vacu: 3, r	No	_	No	none
PEPCK	chlo: 3, extr: 3, cyto: 2, E	No	_	No	none
PNC1	chlo: 10, cyto: 2, nucl: 1,	No	_	No	none
PNC1	chlo: 6, cyto: 3, extr: 3, n	No	_	No	none
PNC1	chlo: 6, cyto: 3, extr: 3, n	No	_	No	none
PNC1	cyto: 10, cysk: 3, golg: 1	No	_	No	none
PRDX1	chlo: 4, nucl: 3, extr: 3, cy	No	_	No	none
PRDX1	cyto: 9, cysk: 2, chlo: 1, n	No	_	No	none
PRDX1	chlo: 4, nucl: 3, extr: 3, cy	No	_	No	none
PRDX1	nucl: 7, cyto: 6, plas: 1	No	_	No	none
PRDX1	cyto: 6, nucl: 3, extr: 3, cl	No	_	No	none
PRDX1	cyto: 6, nucl: 3, extr: 3, cl	No	_	No	none
PXMP2	plas: 6, chlo: 5, mito: 2, p	No	M	No	possibly mitochondrial
SHBP	cyto: 7, E.R.: 3, extr: 2, ch	No	_	No	none
SHBP	cyto: 7, E.R.: 3, extr: 2, ch	No	_	No	none
SLC25A17	plas: 9, golg: 2, cyto: 1, v	No	_	No	none
SLC25A17	plas: 9, golg: 2, cyto: 1, v	No	_	No	none
SOD2	mito: 5.5, nucl: 4.5, cyto_	No	M	No	mitochondrial
SOD2	mito: 6, nucl: 5.5, cyto_n	No	M	No	mitochondrial
SOD2	cyto: 7, chlo: 2, nucl: 2, n	No	S	No	possibly mitochondrial
SOD2	cyto: 7, chlo: 2, nucl: 2, n	No	S	Yes	possibly mitochondrial
SOD2	chlo: 4, vacu: 3, mito: 2,	No	M	No	possibly mitochondrial
SOD2	mito: 12, chlo: 2	No	M	No	mitochondrial
SOD2	chlo: 4, vacu: 3, mito: 2,	No	M	No	possibly mitochondrial
SOD2	chlo: 4, vacu: 3, mito: 2,	No	M	No	possibly mitochondrial
SOD2	chlo: 4, vacu: 3, mito: 2,	No	M	No	possibly mitochondrial
SOD2	mito: 6, cyto: 4, pero: 2,	No	_	No	none
TPI	chlo: 9, nucl: 2, extr: 2, cy	No	_	No	none
TPI	chlo: 6, cyto: 4, nucl: 2, e	No	S	Yes	none
TPI	cyto: 4, nucl: 3.5, mito: 3	No	_	No	none
TPI	cyto: 10, chlo: 2, mito: 2	No	_	No	none
TPI	cyto: 10, chlo: 2, mito: 2	No	_	No	none
TPI	cyto: 10, chlo: 2, mito: 2	No	_	No	none
TPI	cyto: 8, nucl: 4, chlo: 1, e	No	_	No	none
TPN	nucl: 6.5, cyto_nucl: 5.5,	No	_	No	none
TPN	chlo: 10, mito: 2, nucl: 1,	No	S	No	ER

TPN	chlo: 10, mito: 2, nucl: 1,	No	S	No	ER
TPN	chlo: 12, mito: 1, E.R._va	No	_	No	possibly plastid
TPN	E.R.: 14	No	_	Yes	possibly plastid
TPN	nucl: 8, mito: 3, chlo: 1, c	No	M	No	possibly mitochondrial
TPN	nucl: 8, mito: 3, chlo: 1, c	No	M	No	possibly mitochondrial
TR	chlo: 8, mito: 6	No	M	No	possibly mitochondrial
TR	cyto: 6.5, cyto_nucl: 4, cl	Yes	_	No	none
TR	cyto: 6.5, cyto_nucl: 4, cl	Yes	_	No	none
TR	cyto: 6.5, cyto_nucl: 4, cl	Yes	_	No	none
TR	cyto: 6.5, cyto_nucl: 4, cl	Yes	_	No	none
TR	cyto: 6.5, cyto_nucl: 4, cl	Yes	_	No	none
TR	chlo: 4, cyto: 3, extr: 3, n	No	_	No	none
TR	chlo: 6, extr: 3, cyto: 2, n	No	_	No	none
TR	chlo: 6, extr: 3, cyto: 2, n	No	_	No	none
TR	E.R.: 5, cyto: 3, mito: 2, c	Yes	_	No	none
TR	chlo: 12, mito: 2	No	M	No	possibly mitochondrial
TR	chlo: 8, mito: 6	No	M	No	possibly mitochondrial
uaZ	cyto: 11, cysk: 2, chlo: 1	No	_	No	none
UGP	cyto: 7, nucl: 5, vacu: 1, g	No	_	No	none
UGP	cyto: 7, nucl: 5, vacu: 1, g	No	_	No	none
UGP	cyto: 10, nucl: 2, chlo: 1,	No	_	No	none
UGP	cyto: 6, nucl: 4, chlo: 1, v	No	_	No	none
UGP	chlo: 7, cyto: 3, nucl: 2, n	No	_	No	none
UGP	chlo: 5, nucl: 4, cyto: 3, n	No	_	No	none
UGP	chlo: 7, cyto: 3, nucl: 2, n	No	_	No	none
UGP	chlo: 5, nucl: 4, cyto: 3, n	No	_	No	none
UOX	cyto: 10, nucl: 3, pero: 1	No	_	No	none
UOX	cyto: 10, nucl: 3, pero: 1	No	_	No	none
VPS1	cyto: 9, nucl: 4, chlo: 1	No	_	No	none
XDH	nucl: 9, cyto: 2, chlo: 1, p	No	_	No	none
XDH	nucl: 9, cyto: 2, chlo: 1, p	No	_	No	none
ZADH2	chlo: 4, nucl: 4, cyto: 4, e	No	_	No	none
ZADH2	cyto: 5, nucl: 4, cysk: 4, g	No	_	No	none
ZADH2	cyto: 7, chlo: 2, nucl: 2, n	No	_	No	none
ZADH2	cyto: 9, nucl: 2, mito: 2, E	No	_	No	none
ZADH2	extr: 10, chlo: 3, mito: 1	No	_	No	none
ZADH2	extr: 9, chlo: 4, mito: 1	No	_	No	none
ZADH2	mito: 4.5, chlo_mito: 4.5	No	_	No	none
ZADH2	mito: 4.5, chlo_mito: 4.5	No	_	No	none
ZADH2	chlo: 6, extr: 6, cyto: 1, v	Yes	_	No	none
ZADH2	chlo: 10, mito: 2, extr: 2	No	M	No	none

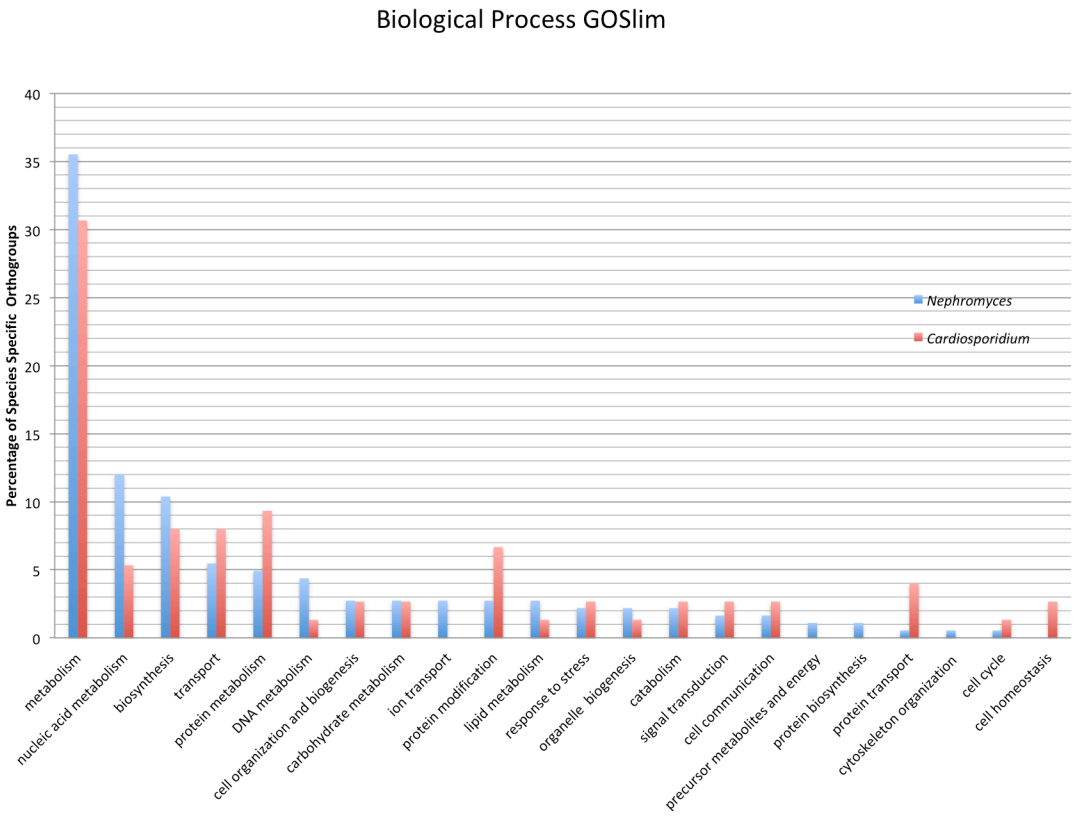
Appendix Table 2) *Cardiosporidium* peroxisomal-related genes Identified PeroxDB and KAAS. Complete transcripts were run through Wolf PSORTII, Ppero, TargetP, Topcons and Predotar to identify possible signal motifs inclding PTS1

Gene	PSORTII	Ppero pts1	TargetP	topcons	predotar
PEX4	chlo: 6, nucl: 5, cyto: 2,	No	_	No	none
PEX5	mito: 5, chlo: 4, nucl: 2	No	_	No	none
PEX5	cyto: 7, nucl: 4, plas: 3	No	_	No	none
PEX5	mito: 5, chlo: 3, cyto: 3	No	_	No	none
PEX6	cyto: 7, nucl: 3, chlo: 2,	Perhaps yes	_	No	none
PEX7	chlo: 7, nucl: 3, mito: 3	No	_	No	ER
PEX7	chlo: 7, nucl: 3, mito: 3	No	_	No	none
PEX7	chlo: 7, nucl: 3, mito: 3	No	_	No	none
PEX7	chlo: 7, nucl: 3, mito: 3	No	_	No	none
PEX7	chlo: 7, nucl: 3, mito: 3	No	_	No	none
PEX7	chlo: 7, nucl: 3, mito: 3	No	_	No	none
PEX7	chlo: 7, nucl: 3, mito: 3	No	_	Yes	none
PEX7	chlo: 7, nucl: 3, mito: 3	No	_	No	none
PEX10	cyto: 6, nucl: 4, chlo: 1,	No	_	No	none
PEX10	cyto: 6, nucl: 4, chlo: 1,	No	_	No	none
PEX10	cyto: 6, nucl: 4, chlo: 1,	No	_	Yes	none
PEX10	mito: 7, chlo: 5, nucl: 2	No	_	Yes	none
PEX14	cyto: 6, chlo: 3, nucl: 2,	No	_	No	none
PEX14	chlo: 7, cyto: 2, nucl: 1,	No	_	No	none
MPV17	mito: 4, E.R.: 4, chlo: 3,	No	M	No	possibly mitochondrial
ABCD	vacu: 10, golg: 2, plas:	No	_	No	possibly plastid
ACAA1	nucl: 3, mito: 3, E.R.: 2.	No	_	No	none
ACAA1	chlo: 13.5, chlo_mito: 7	No	M	No	none
ACAA1	cyto: 8, chlo: 3, nucl: 1,	No	_	Yes	none
ACAA1	nucl: 4.5, cyto_nucl: 3.5	No	C	No	none
ACAA1	chlo: 12, nucl: 1, plas: 1	No	_	No	none
ACAA1	chlo: 12, nucl: 1, plas: 1	No	_	No	none
ACAA1	nucl: 4.5, cyto_nucl: 4,	No	C	No	none
ACOX	chlo: 9, cyto: 2, plas: 2,	No	_	Yes	none
ACOX	pero: 4, E.R.: 3, plas: 2,	Perhaps yes	_	No	none
AGPS	cyto: 7, cysk: 3, mito: 2	No	_	No	none
AGPS	cysk: 8, nucl: 3, cyto: 3	No	_	No	none
AGPS	cyto: 7, cysk: 4, nucl: 3	No	_	No	none
AGPS	cyto: 7, cysk: 3, mito: 2	No	_	No	none
AGXT	cyto: 6, chlo: 4, mito: 3	No	_	No	none
ANT	chlo: 3, nucl: 3, mito: 3	No	_	No	none
ANT	chlo: 3, nucl: 3, mito: 3	No	_	No	none
CAT	pero: 13, cyto: 1	Yes	_	No	none
CPK	cyto: 5, nucl: 4, plas: 2,	No	S	No	none

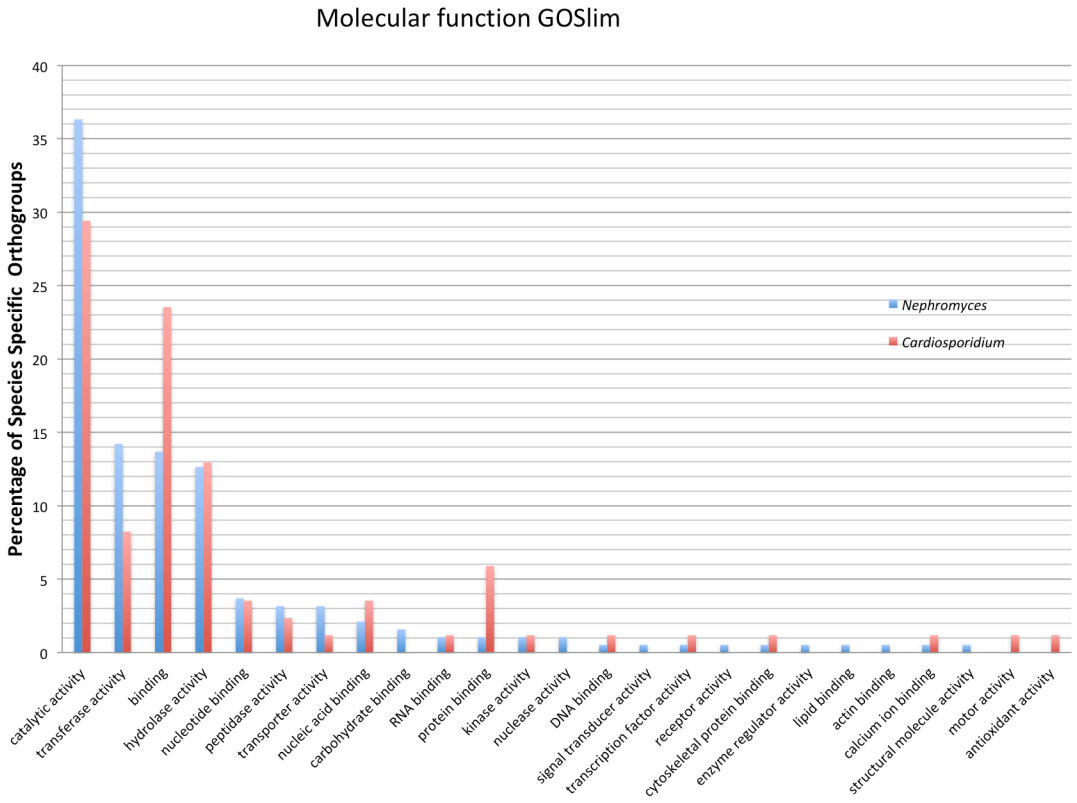
CPK	cyto: 5, nucl: 4, plas: 2,	No	S	No	none
CPK	cyto: 5, nucl: 4, plas: 2,	No	S	No	none
CPK	cyto: 5, nucl: 4, plas: 2,	No	S	No	none
CPK	chlo: 4, nucl: 3, cyto: 2,	No	_	No	ER
DJP1	nucl: 11, chlo: 1, plas: 1	No	_	No	none
DJP1	nucl: 6, cyto: 3, mito: 3	No	_	No	none
DJP1	mito: 5, nucl: 3.5, cyto:	No	_	No	none
DJP1	cyto: 6, nucl: 3, mito: 2	No	_	No	none
DJP1	cyto: 5.5, golg: 4, cyto_	No	_	No	none
E1.3.3.6	chlo: 8, mito: 3, cyto: 2	No	M	No	none
ECH	mito: 5, chlo: 2, nucl: 2	No	_	No	none
ECH	cyto: 7, chlo: 2, nucl: 2,	Yes	_	No	none
ECH	nucl: 6, mito: 5, chlo: 2	No	_	No	none
ECH	nucl: 6, mito: 5, chlo: 2	No	_	No	none
EHHADH	chlo: 12, cyto: 1, mito:	No	M	No	none
FBPA	chlo: 5, cyto: 4, extr: 3,	No	_	Yes	none
FBPA	cyto: 10, chlo: 2, nucl: 1	No	_	No	none
FR	cyto: 8, nucl: 3, extr: 1,	No	_	No	none
FR	cyto: 8, nucl: 3, extr: 1,	No	_	No	none
FR	chlo: 4, nucl: 4, mito: 4	No	_	No	none
FR	chlo: 4, nucl: 4, mito: 4	No	_	No	none
FR	chlo: 4, nucl: 4, mito: 4	No	_	No	none
FR	chlo: 4, nucl: 4, mito: 4	No	_	No	none
FR	chlo: 13, cyto: 1	No	S	No	none
G6PI	cyto: 5, chlo: 4, nucl: 3,	No	M	No	mitochondrial
G6PI	cyto: 5, chlo: 4, nucl: 4,	No	M	No	none
GAPDH	cyto: 13, mito: 1	No	M	No	none
GAPDH	cysk: 8, cyto: 5, chlo: 1	No	_	No	none
GK	cysk: 11, chlo: 2, cyto: 1	Yes	_	No	none
GK	cysk: 11, chlo: 2, cyto: 1	Perhaps yes	_	No	none
GOT	chlo: 10, mito: 4	No	M	No	none
GOT	pero: 11, cyto: 2, nucl:	Yes	_	No	none
GOT	pero: 11, cyto: 2, nucl:	Perhaps yes	_	No	none
GPD	plas: 8, golg: 3, E.R.: 2,	No	S	No	none
GPD	plas: 8, golg: 3, E.R.: 2,	No	S	No	ER
GSTK1	pero: 4, cyto_nucl: 4, n	Yes	_	No	ER
GSTK1	pero: 5, cyto: 4.5, cyto_	Yes	_	No	none
HAO	pero: 12, cyto: 2	Yes	_	No	none
HEX1	cyto: 11, extr: 1, cysk: 1	No	_	No	none
HEX1	cyto: 11, extr: 1, cysk: 1	No	_	No	none
HK	chlo: 4, E.R.: 3, cyto: 2.5	No	_	No	none
HK	cyto: 5, chlo: 3, nucl: 2,	No	_	No	none
HK	cyto: 5.5, E.R.: 4, cyto_	No	_	No	none

HK	cyto: 5, golg: 3, mito: 2	No	_	No	none
HK	nucl: 12, cyto: 1, mito:	No	_	Yes	none
HK	cyto: 5, golg: 3, mito: 2	No	_	No	none
HMGCL	nucl: 6, cyto: 5, chlo: 2,	No	_	No	none
HMGCL	nucl: 6, cyto: 5, chlo: 2,	No	_	No	none
HPR	cyto: 13, chlo: 1	No	_	No	mitochondrial
HPR	cyto: 13, nucl: 1	No	_	No	none
HSD17B4	chlo: 8, cyto: 4, nucl: 1,	No	M	No	none
HSD17B4	cyto: 8, chlo: 3, pero: 2	No	M	No	possibly mitochondrial
HSD17B4	pero: 10, cyto: 2.5, cyt	Yes	_	No	possibly mitochondrial
HSD17B4	pero: 6, cyto: 4.5, cyto_	Yes	M	No	none
HSD17B4	pero: 9, cyto: 2.5, cyto_	Yes	_	No	possibly mitochondrial
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	none
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	mitochondrial
HSD17B4	chlo: 8.5, chlo_mito: 7,	No	M	No	mitochondrial
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	mitochondrial
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	mitochondrial
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	mitochondrial
HSD17B4	pero: 11, cyto: 2, nucl:	Yes	_	No	mitochondrial
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	none
HSD17B4	chlo: 8.5, chlo_mito: 7,	No	M	No	mitochondrial
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	mitochondrial
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	mitochondrial
HSD17B4	chlo: 8, nucl: 2, pero: 2	Yes	M	No	mitochondrial
HSD17B4	chlo: 14	No	S	Yes	none
I5PDH	nucl: 5, cyto: 4, chlo: 2,	No	_	No	none
IDH	pero: 12, cyto: 2	Perhaps yes	_	No	none
IDH1	mito: 7, chlo: 6, cyto_n	No	M	No	possibly mitochondrial
IDH1	chlo: 6, cyto_nucl: 3, nu	No	_	No	none
IDI	chlo: 9, mito: 2, nucl: 1	No	M	No	none
katE	pero: 13, cyto: 1	Yes	_	No	none
katE	pero: 9, nucl: 2, chlo: 1	No	_	No	none
katE	pero: 8, cyto: 3, nucl: 1	No	_	Yes	none
LYS4	cyto: 3, nucl_plas: 3, nu	No	_	No	none
LYS4	chlo: 10, cyto: 4	No	_	No	none
LYS4	cyto: 6, cysk: 4, chlo: 3,	No	_	No	none
LYS4	mito: 7, chlo: 4, nucl: 2	No	M	No	none
LYS4	mito: 7.5, cyto_mito: 4	No	M	No	none
MDAR	cyto: 4, chlo: 3, mito: 3	No	S	Yes	ER
MYA2	nucl: 8, cyto: 3, chlo: 1,	No	_	No	none
MYA2	cyto: 7, nucl: 5, chlo: 1,	No	_	No	none
MYA2	cyto: 6, nucl: 5, chlo: 1,	No	_	No	none
MYA2	cyto: 6, nucl: 5, chlo: 1,	No	_	No	none

OPDC-PRT	chlo: 8, nucl: 3, cyto: 2,	No	_	No	ER
PDCR	pero: 7, cyto: 3.5, cyto_	Yes	_	No	possibly plastid
PDCR	pero: 7, cyto: 3.5, cyto_	Yes	_	No	possibly plastid
PDCR	pero: 7, cyto: 3.5, cyto_	Yes	_	No	possibly plastid
PDCR	pero: 7, cyto: 3.5, cyto_	Yes	_	No	possibly plastid
PEPCK	cyto: 3, E.R.: 3, nucl: 2,	No	_	Yes	none
PNC1	cyto: 10, nucl: 1, extr: 1	No	_	No	none
PRDX1	chlo: 9, cyto: 2, nucl: 1,	No	_	No	none
PXMP2	mito: 7, chlo: 5, nucl: 1	No	M	Yes	ER
SHBP	cyto: 10, plas: 1, extr: 1	No	_	No	none
SOD2	chlo: 9, extr: 3, nucl: 1,	No	S	No	ER
SOD2	cyto: 9, chlo: 2, nucl: 2,	No	S	No	ER
SOD2	chlo: 5, mito: 4, E.R.: 2,	No	M	No	possibly mitochondrial
SOD2	cyto: 8, nucl: 2, plas: 2,	No	_	No	none
TPI	chlo: 4, extr: 4, cyto: 2,	No	S	No	none
TPI	chlo: 4, extr: 4, cyto: 2,	No	S	No	ER
TPN	cyto: 5, nucl: 4, chlo: 2,	No	_	No	none
TPN	cyto: 7, chlo: 2, nucl: 2,	No	_	No	none
TR	vacu: 4, chlo: 3, mito: 2	No	S	No	none
TR	chlo: 5, mito: 3, vacu: 3	No	S	No	ER
TR	cyto: 6, E.R.: 5, mito: 2,	No	_	No	ER
TR	cyto: 6, E.R.: 5, mito: 2,	No	_	No	none
uaZ	pero: 9, cyto: 2.5, cyto_	Yes	_	No	none
uaZ	pero: 11, cyto: 2, golg:	Yes	_	No	none
uaZ	pero: 9, cyto: 2.5, cyto_	Yes	_	No	none
ZADH2	chlo: 6, nucl: 2.5, cyto:	No	_	Yes	none
ZADH2	chlo: 6, nucl: 2.5, cyto:	No	_	Yes	none
ZADH2	chlo: 9, extr: 2, nucl: 1,	No	M	Yes	possibly plastid
ZADH2	chlo: 6, cyto: 6, nucl: 1,	No	_	Yes	none



Appendix Figures 1) Comparison of biological process GOSlim terms from lineage specific orthologous genes from both *Nephromyces* (blue) and *Cardiosporidium* (red), respectively. GOSlim terms are grouped by their major function, showing no clear pattern of gene losses by functional category between *Cardiosporidium* and *Nephromyces*.



Appendix Figure 2) Comparison of molecular function GOSlim terms from lineage specific orthologous genes from both *Nephromyces* (blue) and *Cardiosporidium* (red), respectively. GOSlim terms are grouped by their major molecular function, showing no clear pattern of gene losses by functional category between *Cardiosporidium* and *Nephromyces.*