UNIVERSITY OF CALIFORNIA, SAN DIEGO

Metagenomics of the *Riftia pachyptila* Symbiont

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Marine Biology

by

Julie Christine Robidart

Committee in charge:

Horst Felbeck, chair
Eric Allen
Lihini Aluwihare
Doug Bartlett
Bianca Brahamsha
Victor Nizet

2006

The Dissertation of Julie Christine Robidart is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2006

DEDICATION

To my family, and particularly my mom, for all the support they have provided over the years.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

%GC: percentage of guanine + cytosine nucleotides in the genome

16S: small ribosomal RNA subunit, commonly used to create bacterial phylogenies

AA: amino acid

ACL: ATP citrate lyase

AT: adenine and thymine

BLAST: Basic Local Alignment Search Tool: for finding homologous sequences in the sequence database GenBank

BLASTn: BLAST of a nucleotide sequence against the database of nucleotides

BLASTx: BLAST of a translated query sequence against the amino acid database

tBLASTn: BLAST of possible nucleotide sequences that can comprise an amino acid sequence, against the amino acid sequence database

Bp: basepairs

COG: Clusters of orthologous groups (database within NCBI)

Contig: contiguous stretch of DNA

GC: guanine + cytosine nucleotides

GenBank: an online database of deposited sequences

ITS: internal transcribed spacer region, used in bacterial phylogenetics

Kb: kilobase

LB: Luria-Bertani

Mb: megabase

MLSA: multilocus sequence analysis

Mmo: soluble methane monooxygenases

Nr: non-redundant

Nt: nucleotide

ORF: open reading frame

PCR: polymerase chain reaction

PEP: phosphoenolpyruvate

Pmo: particulate methane monooxygenases

PTS: phosphotransferase system, for responding to a carbon source by changing catabolically

REP-PCR: repetitive extragenic palindromic sequence amplification, used in genome fingerprinting

rTCA: reverse tricarboxylic acid cycle for carbon fixation

RuBisCO: ribulose bisphosphate carboxylase

SNP: single nucleotide polymorphism

Sp. (spp.): species (plural)

Synteny: conserved gene order

TCA: tricarboxylic acid

TRAP: tripartite ATP-independent periplasmic transport system for 4-carbon molecules

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

I am grateful that I have had the chance to work with my advisor, Horst Felbeck, who I consider to be one of the most experienced and knowledgeable chemosymbiosis physiologists in the world. Thank you to Horst for soliciting my help with this metagenome, which has led to many fascinating discussions of metabolism and symbiosis.

Barbara Prezelin peeked my interest in the capabilities of single-celled organisms, while Jim Childress was the first to introduce me to the hydrothermal vent food web. Subsequent conversations with Jim and Pete Girguis solidified my resolve to pursue studies in the field of hydrothermal vent microbiology.

At Scripps, initial collaborations with Cordelia Arndt and Melinda Duplessis introduced me to working with live animals and enzymatic and molecular techniques. Cordelia, Melinda, and Suzanne Dufour were all great labmates and I learned much from all of them.

Jeff Stein, Doug Bartlett, Eric Allen, Kelly Bidle, Bianca Brahamsha, Jay McCarren, Mark Hildebrand and Kim Thamatrakoln have all trained me in molecular biology, and Doug Bartlett has acted as a second advisor to me throughout the years. Without his recommendation, I would not have attended the Marine Biological Laboratory's Microbial Diversity Course, which contributed greatly to my becoming a molecular microbiologist. Eric Allen has only entered into my committee this year, but has been invaluable as a member. I have learned more this last year of graduate school than I have in the last four, and it is due to Eric's patience and willingness to teach elementary techniques (in his mind) to the computer illiterate (me). I do not know where this project would be without him. I would like to thank Bianca Brahamsha and Lihini Aluwihare specifically, for providing a great deal of unsolicited assurances during the most difficult points of committee meetings. Without their confidence in my potential, my endurance may have faltered.

Barbara Campbell and Craig Cary were very involved in my thesis at the beginning. They were incredibly welcoming and allowed me to join them on many cruises and at the

University of Delaware. In Delaware Barb became a good friend, taught me a great deal in the lab and introduced me to metagenomics.

The Azam Lab educates “real microbiologists,” the kinds that use microscopes. The lab as a whole has been essential to my development in this field and on various occasions has recruited me into some incredible philosophical discussions. I am honored to have been “cultured” in this discipline by Farooq, whom I consider to be one its greatest minds.

I am always indebted to my family, for endless love and support, despite their general disinterest in the hard sciences. My father provided me with a scientific mind, and my mother gave me the determination to accomplish whatever goals I set my sights on. I am especially grateful to my mom and sister Emily, for their sound guidance and reassurance over the many years of frantic phone calls.

VITA

1999 B.S., Aquatic Biology with a minor in Spanish; University of California, Santa Barbara

Graduated with Honors and Distinction in the Major. Thesis Project: "The Effects of Ultraviolet Radiation on Carbon Fixation Rates of Natural Assemblages of Marine Phytoplankton in the Santa Barbara Channel."

2006 Ph.D., Marine Biology, specialization microbiology; Scripps Institution of Oceanography

Research Experience:

1997-1999: Dr. Prezelin, University of California, Santa Barbara

Studies of natural phytoplankton populations, computational data analysis

1998-1999: Dr. Jim Childress, University of California, Santa Barbara

HPLC analysis of hemoglobin concentrations in *Riftia pachyptila* tissues

2001: Marine Biological Laboratory, Woods Hole Oceanographic Institution

Microbial Diversity Course: Molecular investigations of sulfate reducing microbes

2002: Dr. Craig Cary and Dr. Barbara Campbell, University of Delaware

Examination of *Alvinella pompejana* metagenome, protein expression

ABSTRACT OF THE DISSERTATION

Metagenomics of the *Riftia pachyptila* Symbiont

by

Julie Christine Robidart

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2006

Professor Horst Felbeck, Chair

Despite the organisms' relative inaccessibility relative to most biological systems, much has been learned regarding the physiology of the *Riftia pachyptila* tubeworm and its chemolithoautotrophic symbiont since its discovery a quarter century ago, but many questions regarding the physiology of this association remain unanswered. Since the symbiont is unculturable and all experiments are performed with dying preparations, one molecular option to approach these queries is metagenomics. Metagenomics is inherently difficult, as sequences are derived from an environmental sample rather than a pure monoclonal culture, and the methods used to create the *Riftia* symbiont metagenome were not optimal for genome

closure. However, the symbionts' metagenome provides a wealth of information regarding its physiology. Investigations in this thesis support the theory that trophosome contains a single species, and genomic heterogeneity is very low within the symbiont populations at 9ºN. Though sequence fractionation is a problem with the symbiont metagenome, much information has been gathered from it regarding the symbionts' metabolic capabilities. It has been discovered that the symbionts can use the reverse TCA cycle for carbon fixation in addition to the Calvin-Benson Cycle, which explains the discrepancy in the hosts' carbon isotope ratios. The symbionts can also function heterotrophically and have a large suite of signal transduction mechanisms to respond to various environments. It appears as though the host can supply both inorganic and organic carbon to the mixotrophic symbionts, which contain various enzymes to break down host cells. These are the most significant of several new insights the symbiont metagenome has provided us, and a large number of new hypotheses have been proposed as a result.

CHAPTER 1

Introduction

The *Riftia pachyptila* symbiosis

INTRODUCTION

The *Riftia* symbiosis

The discovery of the deep sea hydrothermal vents in 1977 was one of the most groundbreaking discoveries in oceanography of the past century (Corliss and Ballard 1977). Geologists were exploring the seafloor for evidence of seafloor spreading to support the theory of plate tectonics when they happened upon a thriving community of unusual organisms on the seafloor. The most conspicuous organism, the giant tubeworm *Riftia pachyptila*, represented most of the biomass at the site. It was soon discovered that *Riftia* has no mouth or digestive tract but obtains nutrition from its chemolithoautotrophic symbionts housed in its trophosome (Felbeck 1981; Cavanaugh *et al.* 1981). The trophosome tissue is in the trunk of the worm, within the body wall. Symbionts reach densities of 10^9 cells/g tissue (Cavanaugh, 1981), 3 orders of magnitude higher than the surrounding sea water. Since the associated symbionts are contained in bacteriocytes and have no connection with the outside environment, they rely on their host to deliver all necessary substrates and to expel waste products before they accumulate to toxic levels. The host, in turn, is completely dependent on the symbiont for organic carbon to fuel its remarkably high growth rate (Lutz *et al.* 1994) and as a result of this dependence, aposymbiotic *Riftia* juveniles do not survive (Bright and Giere 2005).

Riftia pachyptila is in the family Siboglinidae (Rouse 2001) and its symbionts are sulfur-oxidizing gamma proteobacteria. *Riftia* lives in a chitinous tube (Gaill *et al.* 1992) and has a relatively simple body plan [Figure 1.1], consisting of the plume at the anterior end, above the vestimentum (the muscle that holds the worm within the tube). The trunk makes up most of the organism, and is located directly below the vestimentum. The opisthosome is posterior to the trunk (Rouse 2001). The worms settle in areas of diffuse hydrothermal flow, where access to both reduced and oxidized substrates are available over spatiotemporal gradients. Temperature varies from averages of 2 to 25ºC (Chevaldonne *et al.* 1991). Concentrations of oxygen in the diffuse flow regions range from 0 to 100μM, hydrogen sulfide

from 0 to 40μM, and dissolved inorganic carbon from 2 to 11.6mM (Le Bris *et al*. 2006, Shank *et al*. 1998). The worm's plume acts as a gill and permits diffusion of various chemicals into the blood (Anderson *et al.* 2002). Sulfide is normally toxic to metazoans, but *Riftia* has a specialized hemoglobin that binds oxygen and sulfide (and possibly nitrate), to prevent them from poisoning host tissues and from reacting with each other (Arp *et al.* 1987; Hahlbeck *et al.* 2005). The worm also contains carbonic anhydrase, which concentrates carbon dioxide within the blood for transport (Kochevar and Childress 1996; De Cian *et al.* 2003). These substrates are carried *via* the major vascular blood vessel to the symbionts in the trophosome. These redox substrates react with each other and so are rarely encountered concurrently. In the host environment, the symbionts overcome the difficulties consequential to dependence on these reactive substrates. Whereas the sulfide oxidizing bacteria *Beggiatoa* spp., *Thioploca* spp., and others must build sheaths to span gradients within sediments or contain gas vacuoles to keep substrates separated (Larkin and Strohl 1983; McHatton *et al.* 1996), the *Riftia* host provides these functions for the symbionts due to its dual-purpose hemoglobin.

Studies of key enzymes of various pathways indicate that the symbiont uses the Calvin-Benson Cycle for carbon fixation (*via* ribulose bisphosphate carboxylase (or RuBisCO) and phosphoribulokinase), reverse sulfate oxidation for energy generation (APS reductase and ATP sulfurylase), and nitrate as an alternate electron acceptor (*via* nitrate reductase) (Fisher 1990; Hentschel and Felbeck 1993). Energy gained from sulfide oxidation with either oxygen or nitrate is used to fix carbon dioxide *via* the Calvin-Benson Cycle and a form II RuBisCO carboxylase (Robinson *et al.* 2003).

Paradoxically, the carbon isotope ratio of the tubeworm host does not reflect the Calvin-Benson Cycle as the primary source of fixed carbon. Whereas the $\delta^{13}C$ of the adult tubeworm is consistently approximately -11‰, Calvin-Benson Cycle discriminates to a higher degree and generates $\delta^{13}C$ values averaging -25‰ (range -23 to -50‰) in similar chemoautotrophic symbioses. The form II RuBisCO present in the *Riftia* symbiont discriminates to a lesser degree than the typical form I, but the most enriched value assayed

from an organism using form II RuBisCO is -17.8‰ from the alpha proteobacterium *Rhodospirillum rubrum* (Roeske and O'Leary 1985). Various investigators provide insight into the possible reasons for this discrepancy, most of which are based on availability of substrates and their diffusion through the various tissues of the tubeworm host (Scott 2003; Robinson and Cavanaugh 1995; Fisher *et al.* 1990). Tubeworm wet weight correlates well with isotopic signature; smaller worms (with higher surface:volume ratios and therefore increased substrate uptake relative to need) have lighter carbon isotopic values (Fisher *et al.* 1990). The carbon limitation theory suggests that when carbon is low, the enzyme must use what carbon is available, resulting in a decreased ability to discriminate. It has been shown, however, that environmental CO_2 concentrations are sufficient to facilitate uptake by *Riftia* and that symbionts are not carbon-limited (Childress *et al.* 1993). Furthermore, experiments with catheterized plumes that are obstructed from contact with the trophosome show that carboxylation does in fact occur in the blood of the host, and malate and succinate can be formed in this manner (Felbeck *et al.* 2004). If this is the primary step of carbon fixation, the carboxylase required for this reaction may discriminate to a lesser degree, as it does in C4 plants, resulting in heavier isotopic signatures for the association. However, this study and others (Bright *et al.* 2000) indicate that the majority of carbon fixation occurs in the trophosome, so decreased discrimination resulting from carboxylation by the host cannot fully explain the observed carbon isotope ratio.

Attempts to cultivate the symbiont have failed to date. However, experimental manipulations can be performed with isolated bacteriocytes or purified bacteria (De Cian *et al.* 2003; Felbeck and Jarchow 1997). Although the symbionts can be maintained in this manner for hours, it is unknown whether results of these experiments are physiologically relevant. Sophisticated methods have been developed to maintain a pressurized and chemically static environment for the symbiosis while allowing sampling of the surrounding fluids and of the worm itself (Felbeck *et al.* 2004, Girguis and Childress 2006). The intact symbiosis can be maintained under such conditions for several months (Childress, personal communication).

These studies have shown that the host takes up CO_2, not bicarbonate, as its source of inorganic carbon and that it and the symbionts have a large array of H^+-ATPases to maintain an alkaline pH in the worm's coelomic and vascular fluids (Goffredi *et al.* 1997). The association also takes up both HS^- and H_2S (Girguis and Childress 2006), with the less toxic HS^- form being the chemical species that binds to the worm's hemoglobin (Flores *et al.* 2005). *Riftia* can maintain net chemoautotrophy for at least 5.3 h without sulfide in the surrounding sea water (Girguis and Childress 2006). The source of energy used by the symbionts to sustain autotrophy is likely to be hemoglobin-bound sulfide in the coelomic and vascular fluid of the host. Furthermore, the concentration of intracellular sulfur granules decreases when oxygen is depleted. It is hypothesized that these granules are stored when sulfide concentrations are high, and that they can be used as an electron sink when oxygen is absent (Arndt *et al.* 2000). In purified symbiont preparations with phylogenetically similar symbionts, elemental sulfur consumption is concomitant with sulfide production during anoxia (Duplessis *et al.* 2004), adding support to this theory. These data indicate that *Riftia* is a well-buffered system, and metabolic function is well sustained over a range of environmental conditions. This versatility is likely an adaptation to survival in the highly variable diffuse hydrothermal flow region.

Evidence of the direct consumption of symbionts exists through visualization of symbiont degeneration towards the margin of the trophosome lobule (Bosch and Grasse 1984). Symbiont morphology varies from large rod shaped cells at the interior to smaller coccoid cells at the exterior of each lobule (Bright and Sorgo 2003; Bosch and Grasse 1984). Organic carbon is also transferred from symbiont to host *via* translocation, based on studies with purified symbionts and autoradiography (Felbeck and Jarchow 1997; Bright *et al.* 2000). The fixed carbon exuded from symbiont incubations with radiolabelled carbon are succinate, and to a lesser extent glutamate. These are presumed to be the translocation products (Felbeck and Jarchow 1997).

Several lines of evidence indicate that the obligate nature of the symbiosis is only with respect to the host; the symbiont has a free living stage and juvenile tubeworms must therefore be colonized *de novo* every generation (Nussbaumer *et al.* 2006; Cary *et al.* 1993; Feldman *et al.* 1997). In support of this is the incongruency of symbiont and host phylogenies (Feldman *et al.* 1997) suggesting that cospeciation of symbiont and host does not occur. Identical symbionts are found in phylogenetically distinct hosts when they are present in the same hydrothermal region. A second line of evidence is the lack of symbiont detection in host ovaries with a symbiont-specific 16S ribosomal RNA probe (Cary *et al.* 1993), indicating transovarial inheritance is absent in this species. Thirdly, the symbiont does not possess either the A+T (adenosine + thymine) nucleotide enrichment or genome reduction typical of vertically transmitted symbiont genomes (Nelson *et al.* 1984). Vertically transmitted symbionts contain reduced genomes because over evolutionary time unnecessary functionalities become deleted from the genome (Moran 2003). This deletion process results in loss of *mut*Y, which recognizes guanine-adenine mismatches and excises the adenine (Tsai-Lu and Wu 1994). Loss of *mut*Y therefore leads to a more AT-rich genome in obligate symbionts. Additionally, the symbiont 16S sequence has been detected at many sites throughout the hydrothermal region (Cavanaugh *et al.* 2005). Lastly the *Riftia* symbiont genome contains the genetic components for a functional flagellum, although no flagella have been identified in trophosome-associated symbionts (Millikan *et al.* 1999). A functional flagellum may be necessary during a free-living stage.

Prior to trophosomal development, vestimentiferan larvae contain a digestive tract with ciliary ducts (Jones and Gardiner 1988, 1989). This may be the stage where the symbiont can be acquired from the environment through digestion. Contrary to this theory, thin sections of settled juveniles at various developmental stages show hybridization of symbiont-specific FISH probes to epidermal and mesodermal tissue and infection through the trunk body wall. The dorsal mesentery in proximity to the foregut is proposed as the region that develops into the trophosome, according to this study (Nussbaumer *et al.* 2006). Despite

the host's reliance on the symbiont, it is not transported directly to offspring ("vertically") like many dependent symbionts (Clark *et al*. 2000; Chen *et al*. 1999; Cary and Giovannoni 1993). The horizontal transfer of symbionts is not unique to this symbiosis. Most symbionts are acquired from the environment by their hosts (Nyholm and McFall-Ngai 2004). A variety of associated organisms must therefore overcome the obstacles of symbiont selection from the environment and stabilization of their growth rate, in addition to evasion of host immune reactions that are designed to eliminate foreign organisms. How these feats are overcome with regard to the *Riftia* symbiosis is currently unknown.

A number of vertically transmitted symbiont genomes have been sequenced (Moran 2003). These organisms have undergone genome reduction due to decreased selection in the static host environment and the small size of these genomes requires modest sequencing effort. Such studies have provided insight on the evolution of bacterial symbionts. Since many of these symbionts' closest relatives are pathogens, it is thought that their relationship with their hosts originated as pathogenicity (Douglas and Raven 2002; Moran 2003). This is reflected in the nature of deleted genes: those that confer a deleterious effect on the host are often absent from the reduced genomes of their mutualistic relatives (Ochman and Moran 2001). Due to the free-living life stage of the *Riftia* symbiont and the evidence of infection *via* the epithelium, the genomic component dedicated to pathogen-type interactions will provide insight into the evolution of mutualistic symbioses involving horizontally transferred symbionts.

The symbiont must have extensive regulatory networks to sense its surroundings. Accordingly, Hughes *et al.* sequenced a sensory transduction protein kinase homolog from the symbiont genome (Hughes *et al.* 1997). The trophosome is presumed to be a comfortable environment where nutrients are provided and waste products removed, and the symbiont has simultaneous access to both sulfide and oxygen. Outside of the trophosome, the symbiont must resist poisoning from heavy metals and respond to highly variable physicochemical regimes (Johnson *et al.* 1998a, 1998b). In its free-living form, the symbiont has been detected by PCR in various vent regions, including distant basalt (Cavanaugh *et al.* 2005). Due to its

presence in highly variable conditions, the symbiont's physiology must be extremely adaptable to its environment. We introduce the name "*Endoriftia persephone*" for this microbe, to reflect the dual nature of its life stages: within the nourishing, "motherly" host *Riftia pachyptila*, and outside, in the "hellish" hydrothermal realm. Persephone is the Greek goddess of the seasons. As a young girl she was kidnapped from her mother Demeter, the goddess of fertility (analogous to *Riftia*: one of the most fertile of the known annelids). Hades was the kidnapper and he brought Persephone to hell (analogous to the hydrothermal environment), but she soon became very depressed. Rather than keeping her captive in hell forever, Hades decided he would let Persephone visit her mother once a year. This story explains the seasons in Greek mythology. Demeter's happiness upon her daughter's visit translates into fertile soil and new crops, but when Persephone returns to hell the fertile season ends.

Riftia pachyptila is found along the East Pacific Rise, at sites from 18ºS to 26ºN (Black *et al.* 1994). Hydrothermal vents are ephemeral features and all organisms inhabiting these environments are periodically buried under new basalt when eruptions occur (Shank *et al.* 1988). Accordingly, ecosystems must be re-established post-eruption. Seafloor vent sites are colonized by organisms from neighboring hydrothermal sites and can return to thriving ecosystems within a year (Johnson and Tunnicliffe 1988; Childress 1988; Shank *et al.* 1988). The biogeography of vent-endemic organisms is therefore relatively homogeneous across the East Pacific Rise and *Riftia* larval dispersal rates are sufficient to connect hydrothermal sites genetically (Tyler and Young 2003; Black *et al.* 1994). Perhaps surprisingly, investigations of symbiont biogeography also show little phylogenetic variation (Feldman *et al.* 1997; Di Meo *et al.* 2000) according to signature sequences (rRNA subunits and the ITS region) and genomic fingerprints (REP-PCR patterns). These data indicate that symbiont phylogeny is influenced by host genera, geography, and substrate type (sediment or basalt). DNA-DNA solution hybridizations support the theory of little symbiont variation between sites and that variation is not a direct result of geographic location (Edwards and Nelson 1991). The techniques used in these biogeographic studies however are not sufficiently sensitive to detect a second

symbiont, if less abundant than the primary symbiont. The single study that directly addresses the possibility of a second symbiont concludes that a single species represents at least 90% of the symbiotic component within the trophosome (Distel *et al.* 1988). Morphology of the symbionts within the trophosome varies (Bright and Sorgo 2003) and multiple partner symbioses have been established with sulfide and methane oxidizers in other chemosynthetic systems (Distel *et al.* 1995; Fisher *et al.* 1993). While most experiments indicate that a single symbiont inhabits the trophosome, the existence of a monospecific population in the trophosome is inconclusive to date.

Metagenomics

Environmental genetics is a valuable tool for the *in situ* dissection of microbial ecosystems. Its broad use stems from its accessibility for a variety of projects due to the lack of need for cultured organisms of interest. Though the vast majority of microbes are uncultured, the analysis of environmental DNA has provided important insight into the physiology and functional roles of microbial populations in the context of their environment. This has led to an understanding of microbial evolution and the vast diversity of prokaryotic organisms on Earth, offering a new lens through which to view the tree of life (Hugenholtz *et al.* 1998, Woese 1987).

The field of microbial genomics is growing at an exponential rate due to the ever-decreasing cost of DNA sequencing. Genomics has been instrumental in investigations of microbial physiology when coupled with molecular techniques and has enhanced our understanding of microbial physiology. Just as importantly, genomic analysis creates new hypotheses, which can then be investigated with experimentation on cultured organisms.

Recently, the field of environmental metagenomics has merged environmental genetics with microbial genomics in order to better understand the *in situ* processes that govern ecosystems. Environmental metagenomics (herein referred to as “metagenomics”) concerns the random sequencing of environmental DNA samples with the objective of obtaining the partial genomic components of a biological system. In environments where

microbial complexity is low, full consensus genomes can be assembled with these data and fluxes of nutrients and waste products outlined by examining the key genes for the processes in question. On the other hand, in environments of high complexity, an unassembled metagenome provides a snapshot of the most abundant genomic component of the system but lacks information regarding rare organisms simply because they may be missed in environmental sampling. No matter the environment, undersampling will remain an issue, as it is unfeasible to sample every gene of every cell from the environment; some population structure will certainly be missed in all studies. However, metagenomics inarguably contributes to a deeper understanding of the physiology of uncultured organisms. In this respect, it provides a vast amount of ecological information when considered in the context of the environment sampled.

Another major problem with metagenomics is the sequence fragmentation that results from these studies. Due to undersampling (as compared to genomics with clonal, cultured organisms), the assembly programs have fewer overlapping sequences to create contiguous stretches of DNA. Additionally, members of the population can have insertions or deletions of DNA sequences. In both of these situations, problems are encountered when assembling the raw sequences, resulting in fragmentation. Bioinformatics has not yet fully solved this problem, and much of the curation must be performed manually (Chen and Pachter 2005). Creating scaffolds by linking forward and reverse reads from the same cloned insert is one way of creating a less fragmented genome (Gordon 2004). Comparing the genome to that of a closely related organism with a sequenced genome is another (Eric Allen, personal communication). Genomics as a field has advanced quickly, and closed genomes are now acquired by strategically approaching each genome project. Advancements include the use of partial rather than full genomic DNA digests, and sequencing of larger inserts in conjunction with shotgun sequencing (Fraser and Fleischmann, 1997). These methods have helped decrease the number of unclosed genomes that are published.

Addressing the *Riftia* symbiosis with metagenomics

With respect to the *E. persephone-R. pachyptila* association, all described experimental data involving live organisms were obtained at sea, after collection of the specimens with a submersible. The effort involved in recovery of live worms from the sea floor and their subsequent maintenance aboard ship is arduous. Such cruises are very short (< 2 months) and any subsequent experiments must be performed promptly while at sea. It is therefore difficult to perform long-term experiments and, when attempted, their significance with regard to *in situ* physiology is often called into question.

Our knowledge of *Riftia pachyptila* and *Endoriftia persephone* has increased substantially due to the use of molecular techniques. Many questions regarding the symbiont's physiology and symbiont-host interactions remain unanswered. Much of the experimental evidence that does exist is partial and generates additional inquiries. Molecular approaches are inarguably the preferred route in the difficult field of hydrothermal vent biology, since they allow investigation without the requirement of live organisms. *E. persephone* is a rare case of an uncultured organism that is present in high concentrations *in situ*, representing at least 90% of the microbial component of the trophosome (Distel *et al.* 1988), and high numbers of purified symbionts are easily obtained (Felbeck and Jarchow 1997). The *Riftia pachyptila-Endoriftia persephone* symbiosis is poised for an environmental metagenomics study, to gain insight into these organisms' physiologies and to generate better informed hypotheses for experimental investigations.

Introduction to the thesis

In Chapter 2 of this thesis, I directly address the fragmentation of the *E. persephone* metagenome. Manual curation involved connecting contiguous stretches of DNA (or contigs) and open reading frames (or ORFs) that were broken apart in the first assembly. The first assembly contained 2472 contigs and 4642 ORFs, while the genome size estimation only requires approximately 3000 ORFs (assuming the average of 1075 bp per gene holds true for this genome). This discrepancy necessitated investigation of fragmented ORFs. The first assembly was performed with a low threshold for sequence quality, and therefore many

questionable nucleotide calls and stretches of questionable bases were accepted into the pool of sequences that was assembled automatedly. A second assembly was performed using a higher threshold for sequence quality, and this assembly had fewer partial ORFs. Specifically, many ORFs from the first assembly with similar or identical annotations that were directly next to each other on the same strand of a contig were combined into a single gene in the second assembly, indicating that they were actually partial genes in the first assembly. Comparison of the two assemblies was the main major amendment of the metagenome. Several other methods were used to trim the number of open reading frames down towards the expected outcome if the genome were closed. In sequencing projects, investigators choose to sequence larger inserts in order to gain closure on genomes, but in many cases with environmental genomes closure is not achieved. As stated previously with respect to metagenomes in particular, one that is sequenced from a single species that is purified in abundance is optimal for complete or near-complete closure. Theoretically therefore, closure should be expected after curation of the *E. persephone* metagenome, assuming that the population heterogeneity is not too extensive.

In the next chapter I directly address the population heterogeneity of the symbiont. In order to speculate about metabolic fluxes between symbiont and host, I must be confident that there is only a single species within the trophosome, and that a secondary symbiont has not been overlooked in past studies. The presence of a second symbiont species would interfere with these analyses. As stated above, this has been addressed in the past (Distel *et al.* 1988), but techniques in the field have changed slightly since that study in 1988, where the amplified 16S ribosomal RNA genes from the trophosome were sequenced after a standard PCR amplification with universal bacterial primers. I opted to use a similar approach, but used fewer cycles in the PCR (10 vs. the 25 in the previous study) to ensure that the resultant clone library was not completely dominated by the most abundant organism present. I also used universal archaeal primers in addition to the bacterial primers, and used DNA that was extracted from 4 individual worms' trophosomes without purification of the symbiont fraction.

In a symbiont purification, the sulfur-oxidizing symbionts sink to the lower fraction because they contain intracellular sulfur granules. Any symbiont present that does not contain sulfur would therefore not separate into this fraction and would be detected. In addition to clone library analysis, Chapter 3 includes analysis of the diversity that exists within the sequence database, according to comparison of the sequences that comprise the metagenome. The nature of the trophosome inoculum is unknown. Of those that colonize, is a single organism the source of the entire trophosome population, or do a variety of organisms successfully infect and initiate the population? Analysis of the genotypes within the trophosome brings us closer to this understanding.

Finally, in Chapter 4 I investigate the metabolic component of the metagenome. The *E. persephone* metagenome is used to corroborate past findings with experimental analyses and new possibilities for physiological existence of the symbiont are uncovered. Carbon, sulfur and nitrate response and cycling are analyzed. The functional profile of the genome is compared to that of other pathogens and symbionts in order to gain a better understanding of the metabolic possibilities of this uncultured organism by comparison with well characterized cultured systems. Lastly, the genome is investigated with respect to the metabolic capability of the symbiont when outside of the host. This chapter leads to many new hypotheses and opens up new doors for exploration with regard to culturing attempts.

Many investigators in hydrothermal vent and symbiosis research have been awaiting the publication of the *E. persephone* metagenome due to the advancement of knowledge that the genome provides and the well known effect that genomes have on research within their specific fields. Many molecular techniques depend on genetic knowledge and a full genome provides many additional possibilities for exploration with these techniques. This is especially true with regard to this symbiosis, since molecular analyses are the dominant mode of examination of this uncultured organism. The opportunities that the genome provides are immeasurable, and consequential investigations will likely even further enhance our appreciation of this symbiosis.

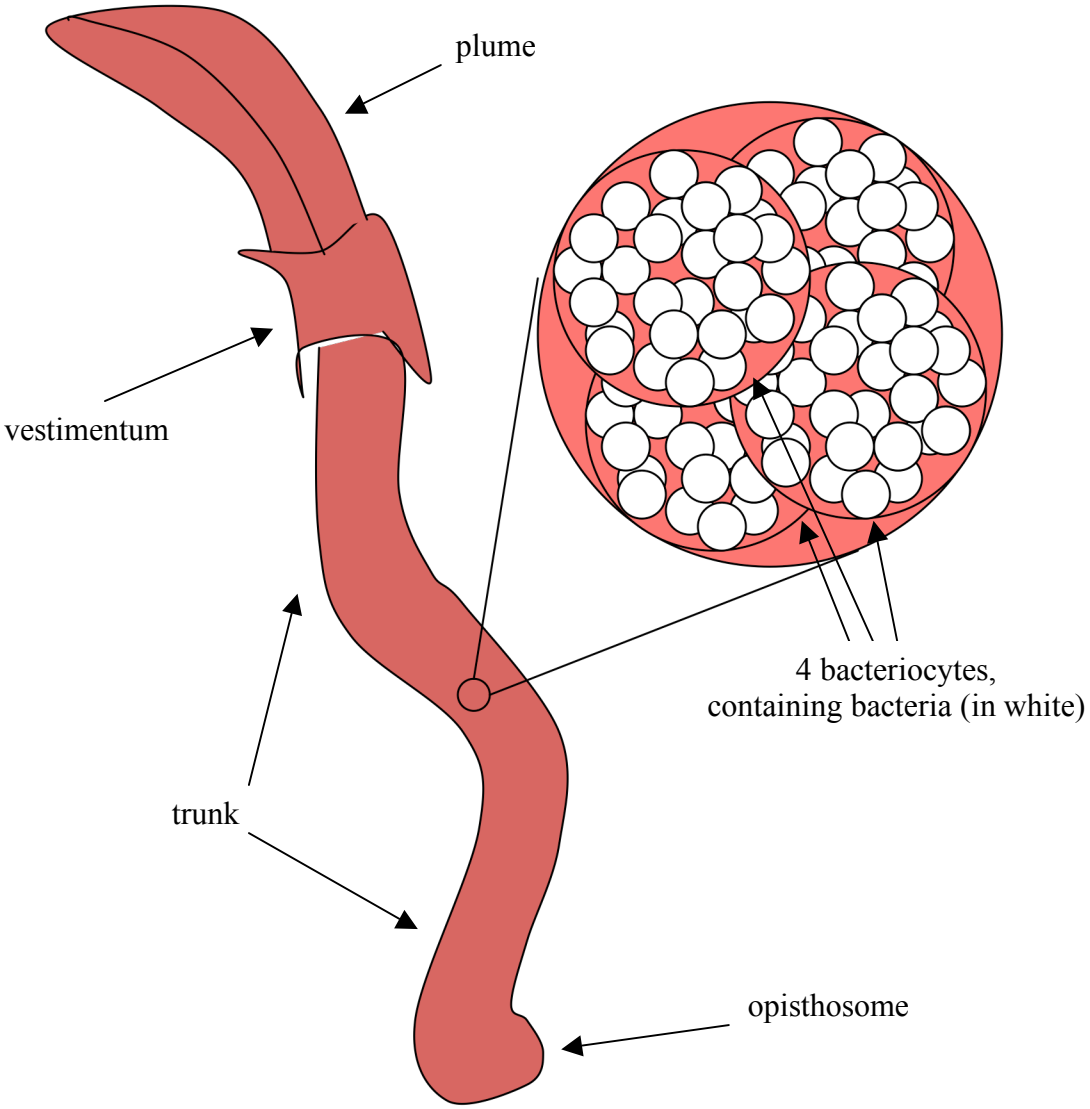


Figure 1.1 Body plan of *Riftia pachyptila*. The trophosome is located within the body wall of the trunk, and is the bacteria-containing tissue. The inset is meant to demonstrate that the bacteriocytes are within the trunk.

REFERENCES

Anderson, A.C., S. Jollivet, S. Claudinot and F.H. Lallier (2002). Biometry of the branchial plume in the hydrothermal vent tubeworm *Riftia pachyptila* (Vestimentifera; Annelida). *Canadian Journal of Zoology* **80**: 320-332.

Arndt, C., F. Gaill and H. Felbeck (2001). Anaerobic sulfur metabolism in thiotrophic symbioses. *Journal of Experimental Biology* **204**: 741-750.

Arp, A.J., J.J. Childress and R.D. Vetter (1987). The sulfide-binding protein in the blood of the vestimentiferan tubeworm, *Riftia pachyptila*, is the extracellular hemoglobin. *Journal of Experimental Biology* **128**: 139-158.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389-3402.

Bosch, C. and P.P. Grasse (1984). Cycle partiel des bacteries chimiautotrophes symbiotiques et leaurs rapports avec les bacteriocytes chez *Riftia pachyptila* Jones (Pogonophore Vestimentifere). II. L'evolution des bacteries symbiotiques at des bacteriocytes. *CR Hebd Seanc Acad Sci, Paris Ser III* **299**: 413-419.

Bright, M., H. Keckelts and C. R. Fisher (2000). An autoradiographic examination of carbon fixation, transfer and utilization in the *Riftia pachyptila* symbiosis. *Marine Biology* **136**: 621-632.

Bright, M. and A. Sorgo (2003). Ultrastructural reinvestigation of the trophosome in adults of *Riftia pachyptila* (Annelida, Siboglinidae). *Invertebrate Biology* **122**: 345-366.

Bright, M. and O. Giere (2005). Microbial Symbiosis in Annelida. *Symbiosis* **38**: 1-45.

Cary, S. C., W. Warren, E. Anderson and S. J. Giovannoni (1993). Identification and localization of bacterial endosymbionts in hydrothermal vent taxa with symbiont-specific polymerase chain reaction amplification and *in situ* hybridization techniques. *Molecular Marine Biology Biotechnology* **2**(1): 51-62.

Cary, S.C. and S.J. Giovannoni (1993). Transovarial inheritance of endosymbiotic bacteria in clams inhabiting deep-sea hydrothermal vents and cold seeps. *Proceedings of the National Academy of Sciences* **90**: 5695-5699.

Cavanaugh, C.M., S.L. Gardiner, M.L. Jones, H.W. Jannasch, J.B. Waterbury (1981). Prokaryotic cells in the hydrothermal vent tube worm *Riftia-pachyptila* Jones: Possible chemoautotrophic symbionts. *Science* **213**: 340-342.

Cavanaugh, C.M., T.L. Harmer, A.D. Nussbaumer, and M. Bright (2005). Environmental transmission in the hydrothermal vent Vestimentiferan – chemoautotroph symbiosis: Stalking the wild symbiont. 3rd International Symposium on Hydrothermal Vent and Seep Biology, La Jolla, USA.

Chen, K. and L. Pachter (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLOS Computational Biology* **1**(2):0106-0112.

Chen, X.A., S. Li and S. Aksoy (1999). Concordant evolution of a symbiont with its host insect species: molecular phylogeny of genus *Glossina* and its bacteriome-associated endosymbiont, *Wigglesworthia glossinidia. Journal of Molecular Evolution* **48**: 49-58.

Chevaldonne, P., D. Desbruyeres and M. Le aitre (1991). Time series of temperature from three deep-sea hydrothermal vent sites. *Deep Sea Research* A **38**:1417-1430.

Childress, J.J. (1988). Biology and chemistry of a deep-sea hydrothermal vent on the Galapagos Rift: the Rose Garden in 1985, and introduction. *Deep Sea Research* **35**: 1677-1680.

Childress, J.J., R.W. Lee, N.K. Sanders, H. Felbeck, D.R. Oros, A. Toulmond, D. Desbruyeres, M.C. Kennicutt and J.M. Brooks (1993). Inorganic carbon uptake in hydrothermal vent tubeworms facilitated by high environmental pCO_2. *Nature* **362**: 147-149.

Clark, M.A., N.A. Moran, O. Baumann and J.J. Wernegreen (2000). Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. *Evolution* **54**: 517-525.

Corliss, J.B. and R.D. Ballard (1977). Oases of life in the cold abyss. *National Geographic* **152**: 441-454.

De Cian, M.C., A.C. Anderson, X. Bailly and F.H. Lallier (2003). Expression and localization of carbonic anhydrase and ATPases in the symbiotic tubeworm *Riftia pachyptila. Journal of Experimental Biology* **206**: 399-409.

De Cian, M.C., A.C. Andersen, J.Y. Toullec, I. Biegala, J.C. Caprais, B. Shillito and F.H. Lallier (2003). Isolated bacteriocyte cell suspensions from the hydrothermal vent tubeworm *Riftia pachyptila*, a potent tool for cellular physiology in a chemoautotrophic symbiosis. *Marine Biology* **142**: 141-151.

Di Meo, C.A., Wilbur, A.E., W.E. Holben, R.A. Feldman, R.C. Vrijenhoek and S.C. Cary (2000). Genetic variation among endosymbionts of widely distributed vestimentiferan tubeworms. *Applied and Environmental Microbiology* **66**: 651-658.

Distel, D.L., D.J. Lane, G.J. Olsen, S.J. Giovanonni, B. Pace, N.R. Pace, D.A. Stahl and H. Felbeck (1988). Sulfur-oxidizing bacterial endosymbionts: analysis of phylogeny and specificity by 16S rRNA sequences. *Journal of Bacteriology* **170**: 2506-2510.

Distel, D.L. and H. Felbeck (1988). Pathways of inorganic carbon fixation in the endosymbiont-bearing lucinid clam *Lucinoma aequizonata*. Part 1. Purification and characterization of the endosymbiotic bacteria. *Journal of Experimental Zoology* **247**: 1-10.

Distel, D.L., H.K. Lee and C.M. Cavanaugh (1995). Intracellular coexistence of methano-and thiotrophic bacteria in a hydrothermal vent mussel. *Proceedings of the National Academy of Sciences* **92**: 9598-9602.

Douglas, A.E. and J.A. Raven (2002). Genomes at the interface between bacteria and organelles. *Philosophical Transactions of the Royal Society of London* **358**: 5-18.

Duplessis, M., W. Ziebis, O. Gros, A. Caro, J.C. Robidart and H. Felbeck (2004). Respiration strategies utilized by the gill endosymbiont from the host lucinid *Codakia orbicularis* (Bivalvia: Lucinidae). *Applied and Environmental Microbiology* **70**: 4144-4150.

Edwards, D.B. and D.C. Nelson (1991). DNA-DNA solution hybridization studies of the bacterial symbionts of hydrothermal vent tube worms (*Riftia pachyptila* and *Tevnia jerichonana*). *Applied and Environmental Microbiology* **57**: 1082-1088.

Ewing B. and P. Green (1998). Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**:186-194.

Ewing B., L. Hillier, M. Wendl, P. Green (1998). Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**:175-185.

Felbeck H (1981). The chemoautotrophic potential of the hydrothermal vent tube worm, *Riftia pachyptila* Jones (Vestimentifera). *Science* **213**: 336-338.

Felbeck, H. and J. Jarchow (1997). Carbon release from purified chemoautotrophic bacterial symbionts of the hydrothermal vent tubeworm *Riftia pachyptila. Physiological Zoology* **71**: 294-302.

Felbeck, H., C. Arndt, U. Hentschel and J.J. Childress (2004). Experimental application of vascular and coelomic catheterization to identify vascular transport mechanisms for inorganic carbon in the vent tubeworm, *Riftia pachyptila*. *Deep-Sea Research I* **51**: 401-411.

Feldman, R.A., M.B. Black, C.S. Cary, R.A. Lutz and R.C. Vrijenhoek (1997). Molecular phylogenetics of bacterial endosymbionts and their vestimentiferan hosts. *Molecular Marine Biology Biotechnology* **6**: 268-277.

Fisher, C.R., M.C. Kennicutt and J.M. Brooks (1990). Stable carbon isotopic evidence for carbon limitation in hydrothermal vent vestimentiferans. *Science* **247**: 1094-1096.

Fisher, C.R. (1990). Chemoautotrophic and methanotrophic symbioses in marine invertebrates. *Reviews in Aquatic Sciences* **2**: 399-426.

Fisher, C.R., J.M. Brooks, J. Vodenichar, J. Zande, J.J. Childress and R.A. Burke (1993). The co-occurrence of methanotrophic and chemoautotrophic sulfur-oxidizing symbionts in a deep sea mussel. *Marine Ecology* **14**: 277-289.

Fraser, C. M. and R. D. Fleischmann (1997). Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18**:1207-1216.

Gaill, F., J. Persson, J. Sugiyama, R. Vuong and H. Chanzy (1992). The chitin system in the tubes of deep sea hydrothermal vent worms. *Journal of Structural Biology* **109**: 116-128.

Girguis, P.R. and J.J. Childress (2006). Metabolite uptake, stoichiometry and chemoautotrophic function of the hydrothermal vent tubeworm *Riftia pachyptila*: responses to environmental variations in substrate concentrations and temperature. *Journal of Experimental Biology* **209**: 3516-3528.

Goffredi, S.K., J.J. Childress, N.T. Desaulniers, R.W. Lee, F.H. Lallier and D. Hammond (1997). Inorganic carbon acquisition by the hydrothermal vent tubeworm *Riftia pachyptila* depends upon high external pCO_2 and upon proton-equivalent ion transport by the worm. *Journal of Experimental Biology* **200**: 883-896.

Gordon, D. (2004). Viewing and editing assembled sequences using Consed. In *Current Protocols in Bioinformatics*, A. D. Baxevanis and D. B. Davison, eds. New York: John Wiley & Co., 11.2.1-11.2.43.

Hahlbeck, E., M.A. Propesel, F. Zal, J.J. Childress, H. Felbeck (2005). Proposed nitrate binding by hemoglobin in *Riftia pachyptila* blood. *Deep Sea Research I* **52**: 1885-1895.

Hentschel, U., and H. Felbeck (1993). Nitrate respiration in the hydrothermal vent tubeworm *Riftia pachyptila. Nature* **366**: 338-340.

Hugenholtz, P., B.M. Goebel and N.R. Pace (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology* **180**: 4765-4774.

Hughes, D.S., H. Felbeck and J.S. Stein (1997). A histidine protein kinase homolog from the endosymbiont of the hydrothermal vent tubeworm Riftia pachyptila. *Applied and Environmental Microbiology* **63**: 3494-3498.

Johnson, K.S., J.J. Childress, R.R. Hessler, A.C.M. Sakamoto and C.L. Beehler (1988a). Chemical and biological interactions in the Rose Garden hydrothermal vent field. *Deep Sea Research Part A: Oceanography Research Papers* **35**: 1723-1744.

Johnson, K.S., J.J. Childress and C.L. Beehler (1988b). Short-term temperature variability in the Rose Garden Eastern Pacific Ocean hydrothermal vent field: An unstable deep-sea environment. *Deep Sea Research Part A: Oceanography Research Papers* **35**: 1711-1722.

Johnson, H.P. and V. Tunnicliffe (1988). Time-lapse photography of a hydrothermal system: A successful one-year deployment. *Eos* **69**: 1024-1026.

Kochevar, R.E. and J.J. Childress (1996). Carbonic anhydrase in deep-sea chemoautotrophic symbioses. *Marine Biology* **125**: 375-383.

Larkin, J.M. and W.R. Strohl (1983). Beggiatoa, Thiothrix and Thioploca. *Annual Review of Microbiology* **37**: 341-367.

Le Bris, N., B. Govenar, C. Le Gall and C.R. Fisher (2006). Variability of physico-chemical conditions in 9º50'N EPR diffuse flow vent habitats. *Marine Chemistry* **98**: 167-182.

Lutz, R.A., T.M. Shank, D.J. Fornari, R.M. Haymon, M.D. Lilley, K.L. Von Damm, D. Desbruyeres (1994). Rapid growth at deep-sea vents. *Nature* **371**: 663-664.

Markert, S. C. Arndt, H. Felbeck, D. Becher, S.M. Sievert, M. Hugler, D. Albrecht, J. Robidart, S. Bench, R.A. Feldman, M. Hecker and T. Schweder (*in press*). Approaching the uncultured endosymbiont of *Riftia pachyptila* by physiological proteomics. *Science.*

McHatton, S.C., J.P. Barry, H.W. Jannasch and D.C. Nelson (1996). High nitrate concentrations in vacuolate, autotrophic marine *Beggiatoa* spp. *Applied and Environmental Microbiology* **62**: 954-958.

Millikan, D.S. H. Felbeck and J.S. Stein (1999). Identification and characterization of a flagellin gene from the endosymbiont of the hydrothermal vent tubeworm *Riftia pachyptila*. *Applied and Environmental Microbiology* **65**: 3129-3133.

Moran, N.A. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Current Opinion in Microbiology* **6**: 512-518.

Nelson D.C., J.B. Waterbury, and H.W. Jannasch (1984). DNA base composition and genome size of the prokaryotic symbiont of *Riftia pachyptila* (Pogonophora). *FEMS Microbiology Letters* **24**: 267-271.

Nussbaumer, A.D., C.R. Fisher and M. Bright (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* **441**: 345-348.

Nyholm, S.V. and M.J. McFall-Ngai (2004). The winnowing: Establishing the squid-vibrio symbiosis. *Nature Reviews Microbiology* **2**: 632-642.

Ochman, H. and N.A. Moran (2001). Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science* **292**: 1096-1099.

Robinson, J.J. and C.M. Cavanaugh (1995). Expression of Form I and Form II RuBisCO in chemoautotrophic symbioses: Implications for the interpretation of stable carbon isotope values. *Limnology and Oceanography* **40**: 1496-1502.

Robinson, J.J., K.M. Scott, S.T. Swanson, M.H. O'Leary, K. Horken, F.H. Tabita and C.M. Cavanaugh (2003). Kinetic isotope effect and characterization of form II RuBisCO from the chemoautotrophic endosymbionts of the hydrothermal vent tubeworm *Riftia pachyptila. Limnology and Oceanography* **48**: 48-54.

Roeske, C.A. and M.H. O'Leary (1985). Carbon isotope effects on the enzyme-catalyzed carboxylation of ribulose bisphosphate carboxylase from *Rhodospirillum rubrum. Biochemistry* **24**: 6275-6284.

Rouse, G.W. (2001). A cladistical analysis of *Siboglinidae caullery* 1914 (Polychaeta, Annelida): formerly the phyla Pogonophora and Vestimentifera. *Zool. J. Linn. Soc.* **132**: 55-80.

Scott, K.M. (2003). A δC-13-based carbon flux model for the hydrothermal vent chemoautotrophic symbiosis *Riftia pachyptila* predicts sizeable CO_2 gradients at the host-symbiont interface. *Environmental Microbiology* **5**: 424-432.

Shank, T.M., D.J. Fornari, K.L. Von Damm, M.D. Lilley, R.M. Haymon, R.A. Lutz (1988). Temporal and spatial patterns of biological community development at nascent deep-sea hydrothermal vents (9º 50'N, East Pacific Rise). *Deep Sea Research II* **45**: 465-515.

Tsai-Lu, J. and A. Lu (1994). *Escherichia coli mut*Y-dependent mismatch repair involves DNA polymerase I and a short repair tract. *Molecular and General Genetics* **244**: 444-450.

Tyler, P.A. and C.M. Young (2003). Dispersal at hydrothermal vents: a summary of recent progress. *Hydrobiologica* **503**: 9-19.

CHAPTER 2

Manual Curation of the *Riftia pachyptila* Symbiont Metagenome

ABSTRACT

The E. persephone metagenome is that of an environmental sample rather than a clonal cultured organism. Problems with the methods used in the sequencing project and in the heterogeneous nature of the population resulted in difficulties with assembly and consequential fragmentation of the metagenome. Many partial open reading frames with the same annotation existed consecutively on a contiguous stretch of DNA (or contigs), indicating that it was necessary to curate the metagenome in order to link some of these artificially truncated open reading frames (or ORFs). Several methods were used in the curation process. The most successful method included the creation of a new assembly with a higher base calling threshold and the use of a different ORF-calling programs. The first assembly of the metagenome included 2472 contiguous stretches of DNA, containing 4642 open reading frames. The final version post-curation contained 1971 contigs and 4004 open reading frames, better reflecting the estimated 3.2 megabase genome size of the symbiont.

INTRODUCTION

The relationship between the hydrothermal tubeworm *Riftia pachyptila* and its gamma proteobacterial symbiotic partner *Endoriftia persephone* has stimulated interest from various fields of science since its discovery just twenty-five years ago (Felbeck 1981; Cavanaugh *et al.*1981). Numerous studies have yielded a wealth of information about the physiological coexistence of the two organisms (Felbeck 1985; Childress *et al.* 1993; Girguis *et al.* 2002; Felbeck *et al.* 2004). However collection and maintenance of the symbiosis is expensive and difficult. Experiments are therefore rare and results may not be environmentally relevant, as organisms contained in an artificial environment inevitably perish. Though culturing efforts are unsuccessful to date, purified symbionts act as source material for experimentation and can be maintained for hours (Felbeck and Jarchow, 1998). This is also not sufficient for most investigators' needs, and the question of ecological significance remains.

One route to provide physiological insight into *Endoriftia persephone* is through genomic analysis. Cultivation independent genomic analysis of a single species from the environment is hampered in complex systems due to undersampling of all but the most abundant organisms (Allen and Banfield, 2005). In the case of *Endoriftia persephone*, however, the tubeworm host acts as a culture vessel and bacterial concentrations within the trophosome reach approximately 10^9 cells per gram wet weight (Cavanaugh *et al.* 1981). Studies indicate that this is a highly specific association and the entire microbial assemblage consists of the symbiont alone (Cary *et al.* 1993; Distel *et al*. 1988). Since the symbiont is enclosed within the body wall of the host and the trophosome fills the volume of the trunk, there is very little possibility of contamination.

Although the high concentration of symbionts within the trophosome is a great advantage for DNA purification, the sample is an environmental one, and is a mixed population of heterogeneous organism types. Heterogeneity exists within the sample as a result of its direct collection from the environment. While such heterogeneity provides the opportunity for population genomic analysis, it also complicates genomic assembly and

subsequent sequence based analyses. Genomic assembly algorithms were initially designed with sequences from cultured organisms, which are clonal and have very little, if any, sequence variation (Chen and Pachter 2005). Only recently have bioinformaticians attempted to tackle the difficulties of environmental metagenomics, and there are still many barriers to overcome before algorithms are perfected for processing environmental whole genome shotgun sequences. Fortunately, the *Endoriftia persephone* metagenome is derived from a population of individuals within a single species, rather than a complex community made up of populations of several species (Robidart *et al. in prep*.). In this respect, the trophosome represents an ideal environment to examine using metagenomic techniques. Its simplicity relative to most microbial systems makes it a very approachable project, and it offers a great initiation for bioinformaticians interested in metagenomics in natural ecosystems. Whereas genome closure for most organisms from complex microbial communities is impossible with current methodologies, environmental simplicity makes closure with the *E. persephone* metagenome an obtainable objective.

MATERIALS AND METHODS

Preparation of symbiont DNA.

Worms were collected from 2600m depth at 9º North, East Pacific Rise aboard the R/V Nadir by the DSRV Nautile in March 1996. Symbionts were isolated with a Percoll density gradient centifugation as described previously (Felbeck and Jarchow, 1998) and DNA was purified in agarose plugs according to Stein *et al.* 1996.

Library construction.

High molecular weight DNA was purified from plugs with QbioGene's Geneclean Kit, according to the manufacturer's instructions. Purified DNA was digested to completion (with Bcl I, Bam HI, Xho I or Nde I), end-repaired, dephosphorylated, and TA cloned into the PCR4-TOPO cloning vector (Invitrogen, Carlsbad, CA) according to manufacturer's instructions. The ligations were transformed into chemically competent TOP10 One Shot *E. coli* cells

(Invitrogen, Carlsbad, CA). Transformations were plated on LB/agar plates containing 50µg/ml Kanamycin and grown for 14-20 hours at 37°C. Colonies were picked into 96-well plates containing 150µl LB + Kanamycin (50µg/ml final concentration) followed by 22-26 hours of growth at 37°C. After growth, 65ml of 50% glycerol (30% final glycerol concentration) was added to each well prior to storage at -80°C. This work was performed at SIO, Amersham and SymBio Corp.

Sequencing and assembly.

Subclone plasmid DNA was amplified by two different methods during the project. The first method was an alkaline lysis purification of plasmid DNA. Subclone cultures (2-10mll) were inoculated into 2ml of LB + Kanamycin (50µg/ml) in deep 96-well plates. Following 24 hours of growth in a shaker at 37°C, cultures were lysed with NaOH/SDS and plasmid DNA was purified *via* ethanol precipitation or *via* a bind/elute protocol using MultiScreen-FB 96-well filter plates (Millipore, Billerica, MA). The second method amplified plasmid DNA directly from the frozen cultures using the TempliPhi Amplification kit (Amersham Biosciences, Piscataway, NJ) generally according to manufacturer's instructions, but using longer extension (30°C) incubations of 14-18 hours.

Amplified plasmid DNA (300-400ng of purified DNA or 3µl of TempliPhi reaction directly) was used in 30-cycle, 20µl sequencing reactions with DynamicET Dye Terminator Cycle Sequencing Kit reagents (Amersham Biosciences, Piscataway, NJ) and modified M13 forward (GTTTTCCCAGTGACGACGTTGTA) or M13 reverse (TGAGCGGATAACAATTTCACAGGA) primers. Sequencing reactions were purified and dried *via* ethanol precipitation in 96-well plates and resuspended in 10µl of water or loading solution (DYEnamic ET sequencing reagents). Resuspended sequencing products were run on MegaBACE 1000 or MegaBACE 4000 (Amersham Biosciences, Piscataway, NJ) capillary electrophoresis instruments.

In collaboration with Amersham Pharmaceuticals and SymBio Corporation, a total of 720 96-well plates (69,120 sub-clones) were sequenced bi-directionally, yielding over 130

thousand raw sequences. Raw (.rsd) sequence files were processed using the Sequence Analyzer program (Amersham Biosciences, Piscataway, NJ) to generate analyzed (.esd) files, which were used in subsequent analyses with the Phred/Phrap/Consed software package (Ewing and Green, 1998; Ewing *et al.* 1998; and Gordon *et al.* 1998, 2001; Gordon 2004). Briefly, vector sequence and bases below a phred 10 quality score (a probability of error for an observed base if those bases were called randomly) were removed from all sequences and only the resulting 97,345 sequences longer than 30 bases were used in the assembly. Those assembled sequences comprised 44,897,623 bases, or approximately 14X coverage of the estimated 3.2 Mb genome (Nelson *et al*. 1984). The final assembly resulted in 2472 contiguous stretches of DNA ("contigs"), the longest of which was 10.5 kb. Contig consensus sequences were used to identify 9258 open reading frames (or ORFs) using the MAGPIE genome analysis software package in collaboration with the Scripps Genome Center (see below).

Gene Identification and Function Assignment.

Analysis of the post-assembly *Endoriftia persephone* genome sequence data was performed *via* the MAGPIE genome sequence analysis software, which uses a series of computational sequence analysis tools to execute a semi-automated annotation of assembled, partially assembled, or raw genome sequence data (Gopal *et al.* 2003; Gaasterland and Sensen, 1996). MAGPIE accepts sequence data in fasta format as input. It then identifies, extracts, and assesses all possible open reading frames (ORFs) for protein coding potential, using a combination of sequence alignment searches, protein functional pattern searches, and statistical analysis of sequence composition properties. MAGPIE assigns putative functions to predicted proteins by integrating significant annotation keywords extracted from matching database protein sequences and patterns to select statistically likely functional descriptions. These descriptions can later be manually edited by human curators, if desired. Sequence analysis results are presented to users visually through graphics embedded in HTML pages. A hierarchical organization of input data, extracted data, and computed analysis allows users to

browse multiple levels of detail, including overall genome features and individual gene properties, as well as fine-grained alignments and pattern matching comparisons at the nucleotide or amino acid level. In addition to web-based interaction, output from a MAGPIE genome analysis can be exported to tab-separated tables for follow-up analysis *via* spreadsheets, external databases or stand-alone computational tools.

To execute MAGPIE analysis of the *Endoriftia persephone* genome sequence data, all 2472 DNA sequence contigs were loaded into MAGPIE. The program was specifically configured for marine microbial genomes, integrating BLASTx (Altschul *et al.* 1997) sequence alignment output from both traditional and metagenomic sequence databases with PFAM (Bateman *et al.* 2004) and TIGRFAM (Haft *et al.* 2003) pattern matching. These tools were run against all possible open reading frames, established using an embedded algorithm called spliceORF (Gaasterland and Sensen, 1996). In brief, this algorithm parses genomic DNA into a series of non-stop codons preceded by a start codon (ATG, GTG or TTG) and terminated by a stop codon (TAA, TGA, or TAG). The minimum length of such an open reading frame, which is user-configurable, was set to 70 amino acids. Potential coding sequences were also examined for prokaryotic signal peptide leader sequences using SignalP 3.0 (Bendtsen *et al.* 2004) and transmembrane domains with TMHMM (Sonnhammer *et al.* 1998). Transfer RNA sequences were identified using tRNAscan-SE (Lowe and Eddy, 1997). The annotation process was carried out on a networked system of three 8-CPU Sun V880s each with 32 GB RAM, six 4-CPU Sun V440s each with 16 GB RAM, and a cluster of 5 dual-CPU Apple G5s each with 4 GB RAM, with output stored on two 5.2 terabyte Apple RAID arrays.

Manual Curation.

SpliceORF identified 9258 open reading frames in the metagenome. The first stage of filtering involved excluding overlapping ORFs. Then a BLAST (Basic Local Alignment Search Tool: Atschul *et al*. 1997) was run between the remaining ORFs and the GenBank non-redundant (nr) database (Benson *et al.* 2000). The output was sorted by e value and ORF were returned to the dataset if they did not overlap more than 200bp with a gene with a better

BLAST hit. The remaining overlaps were then added or discarded based on their length, orientation (inner vs. outer ORF), and the frequency of rare codons in the sequence. The initial 9258 ORFs were decreased to 4342 in this process.

The original assembly was verified with a second "cleaner" assembly using a phred value of >20 and a trim length of 50bp. ORFs were identified with GLIMMER (Salzberg *et al.* 1998) and CRITICA (Badger and Olsen 1999). BLAST analysis between ORFs from the two assemblies confirmed a first group of contigs and allowed linkages of partial ORFs and of some separated contigs from the first assembly [Figure 2.1]. The smaller of these ORFs were renumbered as the original annotation number of the ORF belonging to the longest of the linked contigs. The remaining genes, belonging only to the original assembly, were left in the metagenome file. These were either confirmed or rejected as belonging to the genome dataset by performing BLASTp of the same amino acid file, against itself. The output was run through MSPcrunch (Sonnhammer and Durbin 1994), and the resultant list of "second-best" hits was sorted by percent identification. ORFs that hit at >90% identification against another ORF within the genome were investigated and were either maintained or discarded.

Redundant ORFs resulting from frame shifts in one of the assemblies were filtered by performing a reciprocal tBLASTn of the metagenome against the contigs from a second, more stringent assembly. Overlapping ORFs were investigated and retained or discarded based on the following criteria: Those present in both assemblies by BLASTp and tBLASTn were retained, those absent from either the first or the second assembly and those with low tBLASTn percentage of identical nucleotide to amino acid hits ("% ID's") to the second assembly were BLASTed against GenBank nr. The subset of these ORFs with spurious BLAST hits were discarded.

Many more partial ORFs were linked by performing BLAST of the original dataset against GenBank nr. The output was sorted by accession number and genes with the same accession number that did not overlap with respect to the comparison gene were either merged (if alignment to the best hit showed that there are no missing amino acids between the

two) or assigned as “a,” “b,” etc. of the same original contig number as stated above. Any additional ORFs from the smaller contigs were given new numbers corresponding to the newly linked contig. An Excel file was created to serve as the key to the new numbering to clear up discrepancies with the first version of the genome.

MAGPIE annotations of eukaryotic hits were analyzed further by classifying up- and downstream genes and sorting by percent guanosine + cytosine (“%GC”) and codon usage. Those determined to be of eukaryotic origin were filtered prior to finalizing the list of ORFs. All ORFs with top BLAST hits annotated as eukaryotic were also screened. Those that did not appear in the second assembly were discarded.

RESULTS

General metagenome characteristics.

The *Riftia* symbiont genome size is estimated to be 3.2 Mb (Nelson *et al.* 1984) and sequencing was performed to 14X coverage. The average insert size of sequenced clones is 1900bp, though the restriction enzyme used for most digestions is an AT-rich six-mer (Bcl I: site TGATCA), which is estimated to cut approximately every 5000bp in the 60% GC symbiont metagenome (Nelson *et al.* 1984). The first assembly of the sequence fragments resulted in 2472 contigs total. Unfortunately, most digests were not visualized on agarose gels. We therefore infer that there was likely biased digestion and cloning, resulting in a more fragmented assembly. Several screens were used with the goal of condensing the number of contigs by correctly connecting those that were fragmented, and obtaining full-length ORFs for better annotations and quantification. The screening process is outlined below.

Comparison with the new assembly.

The higher base calling requirements in the second assembly were much more conservative and resulted in a much “cleaner” metagenome. The average open reading frame (“ORF”) size in the first assembly is 474 basepairs, and 610 basepairs in the second. The average microbial gene size is 1075 basepairs (calculated by averaging the gene sizes of all

currently sequenced prokaryotes from http://www.jgi.doe.gov). It appears that neither assembly approached the expected average gene size [Figure 2.2]. The average contig size in the first assembly is 1555 basepairs and 1077 basepairs in the second.

The more stringent base call threshold in the second assembly resulted in many more discarded base calls from questionable chromatogram peaks. This had two effects on the assembly: a) Due to the lower number of incorrect base calls, numerous sequences were better aligned. b) On the other hand, sequences that were called correctly but the corresponding chromatogram was questionable were discarded, occasionally resulting in shorter sequences than in the first assembly. In many cases this was reflected in the size of the corresponding contigs. The overall number of contigs for this second assembly is much the same as the first, but the confidence in each individual base is much higher and the number of frame shifts and artificial stops in the second was lower. The second assembly was used as a tool to modify the first (as described below), since both have their individual advantages and ORFs/contigs that weren't assembled in one may have been assembled in the other version [Figure 2.1].

The first version of the genome (from the initial assembly, after the first screening) contains 4342 ORFs and 2472 contigs. After comparison with the second assembly, "Version2" had an increase in the number of ORFs to 4492 total, and 2119 contigs. 1427 ORFs were combined, resulting in 558 total ORFs, because they hit to the same gene in the second assembly, likely due to differences in ORF prediction. 354 of these ORFs were relabeled with a new contig number because in the second assembly a partial ORF on that contig was identified as the same gene as a partial ORF on another contig from the original assembly. 995 ORFs were recruited from the new assembly and incorporated into the new metagenome file. 65 of these were added because they were on the same contig as an ORF that replaced partial ORFs in the original assembly, but they didn't hit to an ORF from the first. 930 were added because they were present only in the second assembly. In 896 cases a gene was retained from the first assembly though it was not homologous to any ORFs in the

second assembly. The first- or second-version-specific ORFs were compared *via* tBLASTn to the more rigorous assembly and discarded if they overlapped with any other better-annotated ORFs and their BLAST hits were spurious.

Comparison with the GenBank non-redundant database.

Evaluation with respect to GenBank's non-redundant database (GenBank nr) yielded 450 additional partial ORFs that did not overlap but hit to the same accession number. 199 of these had been identified in the previous screen using the new assembly. 37 of them were fused because one ended within a few amino acids of where the next began, with respect to their subject hit, when aligned. The remainder of these ORFs were labeled as "a" or "b" of the lower original ORF number. There were 198 total ORFs that overlapped with each other, according to their positions with respect to their subject hit. 67 of these had been changed in the first screen, 30 were changed based on amino acid alignments (100% amino acid identity across the overlapping region) with the subject gene, and 101 were left as is, since their sequences did not align properly with >60% nucleotide identity. Once all the GenBank nr same accession number hits were identified, the number of ORFs decreased to 4277 and the number of contigs to 1975. This edition of the genome is "Version3/4."

Using BLASTp to compare the resultant fasta file to itself, and MSPcrunch to discard redundant hits screened out an additional 74 ORFs and 4 contigs. This resulted in 4203 ORFs and 1971 contigs.

260 possible eukaryotic hits were identified in the original assembly by GenBank annotation. 58 of these had already been modified in previous curation steps and no longer had eukaryotic homologues as best hits in GenBank. 199 ORFs were discarded: 177 of them did not appear in the cleaner assembly and so were likely the result of poor quality sequence (misidentification of an ORF or simply bad sequence of low confidence). The other 22 were small genes that had no orthologs in GenBank upon further investigation. Since they had appeared eukaryotic in the original assembly they were discarded so that they did not skew the analyses. This final version has 4004 ORFs and 1971 contigs. Post-curation, the

metagenome has an average gene size of 625 basepairs, and an ORF frequency distribution that better reflects the second assembly though it is a modification of the first [Figure 2.3].

DISCUSSION

A critical improvement in the field of whole genome shotgun (WGS) sequencing is the transition to randomization in sequencing (Fraser and Fleischmann, 1997). One such change is from the use of DNA restriction enzymes to shearing and other random fractionation techniques. Restriction digests have the limitation of sequence specificity; if the restriction site does not occur within a long stretch of DNA, end sequencing of the fragment will leave the interior portion of the DNA unsequenced due to limitations in the size of average sequence reads. These noncontiguous output sequences will therefore yield an unassembled genome and the complete genetic repertoire of the organism will remain unknown. Additionally, there is a cloning bias with respect to the size of inserts. Smaller inserts are preferentially cloned because they are less likely to contain a full gene, whose product may have a deleterious effect on the recombinant host organism. Various groups have invested a considerable amount of time and money on closing genomes that result from non-random cloning methods (Tettelin *et al.* 1999). In most cases this is no longer necessary with the use of random shearing to small insert sizes. Randomization has the added benefit of read depth correlations. The chance of a certain genomic region being sampled is virtually identical to the chance of any other region being sampled (Fraser and Fleischmann, 1997), and so if one finds four times more rRNA genes, for example, one can say with some confidence that there are four copies of that gene in the genome. This read depth analysis is useful in metagenomic projects because it aids in assignment of certain gene sequences (of key enzymes, for example) to specific 16S ribosomal RNA signatures, when the genomes are not closed. It is a method used to find out “who” is doing “what” in the environment. Due to the enzyme-based size selection and cloning bias (which results in incomplete genomic representation in sequence libraries), this is not possible in non-random sequencing projects.

The quality of sequences is also of importance, as is the knowledgeable use of base calling and assembly programs such as Phred and Phrap (Ewing and Green 1998; Ewing *et al*. 1998; Gordon *et al*. 2001). These are just some of the factors that must be considered when determining a sequencing strategy. At the inception of the *Riftia* symbiont genome project in 1999, investigators had just begun to master the art of sequencing (Fraser and Fleischmann 1997), and only a few environmental metagenomic projects were underway (DeLong 2005). Though the estimated coverage at the termination of the sequencing effort is 14X the estimated genome size, the metagenome that resulted from these efforts was fragmented into 2472 contiguous stretches of DNA (or "contigs"), with several partial open reading frames, due to the misguided use of complete restriction digests for fractionation and the poor quality of the reads that went into the assembly. The ORF-calling program correctly identified putative ORFs from the consensus sequences, but the quality of the sequences entered into this program contained many incorrect bases and therefore resulted in artificial frame shifts and stop codons. Likewise, many contigs and ORFs were disconnected in this original assembly due to the presence of a restriction site within the ORF [Figure 2.4, 2.5]. The strain variation within the population may add to these difficulties with the assembly.

In order to perform comparative genomic analyses gene quantification is necessary (Tringe et al. 2005). Enumeration is dubious with the original assembly, as larger ORFs would likely be fragmented more often and therefore estimates on their numbers would be falsely inflated. This may be avoided by quantifying ORFs within functional categories (e.g. amino acid metabolism), and calculating the percentage of the total number of ORFs that fall into that category as compared to all other categories from the rest of the genome. Unfortunately this is unreliable because it assumes that genes that are a part of amino acid metabolism, for instance, are on average the same size as genes for nucleotide metabolism. Since there are no estimates of gene size within functional categories, this is not a fair supposition. Manual curation of the genome was therefore necessary before performing any quantitative analysis.

The curated version gives the investigators increased confidence in quantification, and serves as more dependable data for annotation purposes.

The most effective curation step involved comparisons with the more rigorous assembly. The version of the original assembly that we used as starting material had been annotated previously and some of the ORF numbers had been used in publication (Markert *et al. in press*), so modification of this assembly was preferred over use of the second assembly. The decreased number of "true" ORFs (excluding the nonsense ORFs) in the second assembly primarily reflects more reliable base calling. Artificial frame shifts and stop codons were less likely to be encountered. In the vast majority of cases the newly connected ORFs were adjacent to each other in the original assembly and annotations were similar if not identical. In other cases, replacement of the original ORFs with the correct ORFs from the new assembly often times resulted in connection or scaffolding of multiple contigs from the original assembly. This is likely the result of discarding low quality base calls also, which are more frequent at the end of the sequences. If a single low quality sequence from one partial ORF and a single high quality sequence from the respective partial ORF span the two partial genes they will not be assembled because they may not align properly. Once the low quality sequence is discarded, the high quality sequence will align with the other partial gene sequences.

This comparison increased the total number of ORFs because some that were present in the first assembly were not present in the second, and vice-versa, largely due to frame shifts in the same coding sequence or difference in ORF calling between the two assemblies. In some cases the ORFs from the first assembly may have been low quality bases, and therefore were not present in the second. These genes were included in the metagenome. It was therefore necessary to perform a tBLASTn of the ORFs against the stringent assembly. This localized each onto a contig. All were analyzed with respect to other ORFs on the same contig and compared with GenBank nr to determine its legitimacy. If an

ORF that was only present in one assembly overlapped with a longer ORF with a stronger annotation, it was discarded.

Identification of partial genes whose BLAST annotation top hits were to the same accession number in GenBank nr is essential for quantification purposes, as these are likely partial ORFs that make up one gene. Specifically, the number of partial ORFs in this category that did not overlap with respect to their accession number subject hit was significant. The fact that 44% of these ORFs were previously united in the screen against the new assembly gives us confidence in this method. In many cases these partial ORFs abutted each other with respect to the subject hit due to the presence of an intervening restriction site, and these were fused. Unfortunately, since end sequence of the insert tends to be the poorer quality sequence, it is likely that the partial ORFs that did not connect in these cases were just up- or downstream from a restriction site. Since the sequence surrounding those sites is likely low quality, the bases may have been discarded in both assemblies (i.e. phred quality <10). Such contigs were relabeled as “a,” “b,” etc. of the same ORF number to indicate that the intervening sequence is missing.

The sequences identified in the GenBank nr comparison that overlapped with respect to their subject hit required more extensive evaluation. The number of these overlaps that were identified as a single ORF in the first screen was high (34%). We are therefore confident in this screening step. In all of the remaining cases the “cleaner” version of the ORF from the second assembly was aligned with the “cleaner” version of the overlapping ORF and the subject hit. In 34 cases these were found to be the same gene and were fused. In 101 cases, however, the genes were different and did not align with each other at >60% nucleotide identity over 20 base pairs as the threshold. Only a portion of these genes aligned with the subject hit, but the genes as a whole did not align. The portions that did align were not the same in the metagenome ORFs, so they may be conserved domains of genes or the result of misassemblies.

Many ORFs from one assembly did not correspond to an ORF in the second assembly. These were retained when encountered. It is likely that due to a frame shift or truncation in one, the two ORFs were not recognized as belonging to a single gene. The retention of such ORFs resulted in a lot of redundancy in the dataset [Figure 2.6]. The next step of curation involved using BLAST and MSPcrunch (Sonnhammer and Durbin 1994) to compare the fasta file to itself, and list the second-best (or non-self) hits. Sorting through this list aided with disposal of much of the repeated sequence in the genome dataset.

Deciding what genes may be eukaryotic is difficult. The host has a lower GC content than the symbiont (Jennings and Halanych, 2005; Nelson *et al.* 1984), and when sorted by %GC most ORFs fall within a single peak at 60%, which is the expected GC content of the symbiont [Figure 2.7]. Unfortunately, because of the small size of the average contig in the metagenome and the presence of some AT-rich DNA in the symbiont (see Chapter 2), this is not the most reliable strategy for identification of host "contamination." Due to GenBank's bias towards prokaryotes, using a phylogenetic analysis with top BLAST hits (such as the DarkHorse algorithm (Podell and Gaasterland, submitted)) does not suffice either, since this requires that there are multiple eukaryotic top hits for a gene to be categorized as eukaryotic. TETRA is a program that analyses the tetranucleotide frequency of contigs, but this analysis is unreliable for contigs <50kb (Teeling *et al.* 2004) and no significant findings were observed with respect to the <10kb contigs in the metagenome. Although GenBank contains relatively few eukaryotic sequences, top BLAST hit annotation is likely the most dependable method because it would identify the most possible eukaryotic genes. Therefore this was used to identify host DNA. Very little actual eukaryotic sequence was identified, however. 256 ORFs had eukaryote sequences as the best hits, but 68% of the genes identified initially as of possible eukaryotic origin were not present in the cleaner assembly, but were in the first. These (usually small) ORFs identified in the first assembly were miscalls owing to the simple requirements for ORF identification in SpliceORF (Gopal and Gaasterland, 2001). They were not identified as ORFs using the HMM-dependent GLIMMER and CRITICA algorithms. The

remaining possible eukaryotic ORFs in this screen were either discarded because no significant homologs were found in GenBank, or they had been altered in previous screens and were newly annotated as prokaryotic genes. The discarded hits with no homologs in GenBank only made up 7% of the eukaryotic hits (21 total) and were mostly present as the sole partial genes on small contigs. They were initially identified as eukaryotic sequence based on small fragments of the ORF, but when the entire ORF was taken into account these homologies were no longer apparent. These were discarded because they may actually be eukaryotic sequence. Since they did not have significant homologs in GenBank they were not going to affect the state of the genome if they were present. In general, this method of identifying eukaryotic hits aided in disposing of bad sequences that had been retained from the first assembly, but was not very successful in identifying eukaryotic DNA, likely because there is little to no eukaryotic DNA in the metagenome due to the specificity of the symbiont purification process (Distel and Felbeck, 1988).

The final version of the metagenome contains 4004 potential ORFs and 1971 contigs. This is a 6.3% and a 20.3% decrease in the number of ORFs and the number of contigs, respectively, since the initial ORF selection. The genome size is estimated to be 3.2 Mb, and with the assumption of 1075bp per gene, the number of ORFs is not much closer to what we would expect when we take the genome size into account. If genes that begin with start codons are exclusively counted, however, the total number of ORFs is reduced to 3347. If the 1075bp average is the same for the genome, 3347 ORFs is calculated to be an approximately 3.6 Mb genome. Moreover, the revised metagenome has an average gene size of 625 basepairs, vs. 474 basepairs in the first. This longer ORF size results in less false duplications of ORF calls, and aids in numeration. Though the average gene size now better reflects the average prokaryotic gene size, the small size indicates that many remain partial [Figure 2.3]. With respect to quantification, the extent of this problem is minimized when one considers the size of the remaining parts of those ORFs (on average 1075bp – 625bp = 550bp). In order for the remaining partial ORF to be called by ORF-calling algorithms, it must

have a start codon in the same frame as the stop codon, and it must contain an HMM signature. Furthermore, for the single gene to be quantified as two, the second fragment must have an acceptable annotation. The likelihood of these three circumstances occurring is reduced due to the longer average ORF length in the final version of the metagenome.

Due to the current need for bioinformatic methodologies that are specialized for metagenomes by managing population-level heterogeneity, the number of open reading frames resulting from their automated assembly is often overestimated (Chen and Pachter 2005). Artificial frame shifts are a common problem, especially in low read depth regions. Though the individual reads may contain expected ORF sizes, when these are combined into a consensus sequence, frame shifts and stop codons appear as a result of the artificial nature of the composite genome of the organisms sampled (Chen and Pachter 2005). This problem is plainly apparent in the *E. persephone* metagenome, when assessing the open reading frame composition of individual contigs. Many sequential ORFs in the original assembly have similar, if not identical annotations. The result is an artificial overrepresentation of ORF numbers, which makes any quantification of genes uncertain. The effect of the manual curation of the genome, and especially of the inclusion of information from the second assembly, was a decrease in the number of ORFs and more confidence in quantification for comparative analyses.

Manual curation also aided in connection of contigs that were separate in the first assembly. This is helpful in analyses of certain pathways, as genomic location and orientation can provide insight into function; genes that cluster together on a genome often times have complementary roles in metabolism. Additionally, the connection of contigs gives the investigators further confidence that the genome has been sequenced to near-totality and missed genes are rare. Finally, the reduced number of contigs will prove valuable once fosmids are sequenced in an effort to close the composite genome.

Misassembly also occurs in population metagenomics due to heterogeneity in gene content. If one organism has an extra gene that intervenes in an area of synteny within the

population, that gene is going to interrupt the assembly of the gene cluster. No such insertions were identified. It is likely that this contributed to fractionation of the metagenome, and such genes would be encountered if pursued.

In order to verify the representation of most of the symbiont genome in this dataset, gene sets were interrogated and specific pathways were analyzed to ensure that all components required to make that pathway fully functional are present. The full length 5S-16S-23S sequences are present. The full complement of genes necessary to biosynthesize a functional flagellum and ATPase were present, as were all enzymes of functional reductive and oxidative TCA cycles, glycolysis and the Calvin-Benson Cycle. A full complement of tRNA synthetases is present, excluding the cysteine tRNA synthetase gene (*cys*S), but the genomic absence of an annotated *cys*S is not unprecedented (Lipman *et al.* 2000). The number of identified tRNA synthetases in the metagenome (55) is comparable to the closest sequenced relatives *Thiomicrospira crunogena* and *Methylococcus capsulatus* (54 and 55, respectively). Every gene that is investigated and should be present is accounted for. These lines of evidence give the investigators confidence that coverage of the *E. persephone* genome is essentially complete.

There is little doubt that the current version of the metagenome, in its fragmented form, is not ideal for analysis. However, this form suffices for investigation of the metabolic potential of the symbiont. We also would argue that this version is satisfactory for quantitative analysis providing absolute numbers are not overstated. Unfortunately, it is possible that the metabolic potential of the symbiont is underestimated, due to the possible absence of key enzymes that are present in the genome, but are absent in the dataset due to fractionation and undersampling. The occurrence of all genes necessary for a variety of pathways, however, increases our confidence in the genome coverage. Provided that analysis is performed with respect to the genes that are present, and assumptions are not made due to the absence of genes, analysis is dependable.

A full composite genome has many advantages. For instance, promoter mapping can identify potentially co-occurring metabolic pathways when associated and when free-living. Better orientation within the genome and synteny with respect to related organisms can provide insight into potentially horizontally transferred material by identifying surrounding recombination sites. Insertions and deletions of genes, as well as the presence of any extrachromosomal elements could be identified. Sequences of full open reading frames would allow better phylogenetic analysis with respect to related symbionts, and analysis of genome reduction in the forthcoming genome of the related obligate symbiont *Calyptogena magnifica*. Genome closure is the next goal with respect to the symbiont metagenome.

The unclosed state of the *E. persephone* composite genome may arise from a variety of factors. Theoretically, assembly would be expected due to the low heterogeneity in the population (Chapter 3). The algorithms created for assembly of clonal genomes can manage single nucleotide polymorphisms when they are rare, as is the case with the symbiont metagenome. Once a more rigorous assembly was attempted by increasing the requirements for the base-calling algorithm and ORF-calling programs, the quality of the metagenome increased substantially. However, assembly remained unobtainable. It is likely that this is due to a combination of low sequence quality [Figure 2.8], non-randomness of genomic sampling as a result of the use of complete digests and insert size, and to a lesser extent, the genomic heterogeneity of the population. Certainly, with the estimated 14X coverage of the symbiont genome, the techniques used could not achieve genome closure, but several promising possibilities for genome closure exist. Many of the contigs are separated due to restriction sites at the ends of the contigs. Sequencing fosmids would undoubtedly connect many, if not all of these contigs and is recommended as the next step in this metagenome project. Another possibility is end-sequencing of some fosmids in order create better scaffolds for analysis in Consed (Gordon 2004). Since many of the contigs end in restriction sites, it might be possible to connect abutting partial sequences this way if the size of the fosmid is taken into account during analysis. A third method would be to scaffold with respect to a

phylogenetically similar organism with a full, closed genome. Few environmental gamma proteobacteria have sequenced genomes (~77% are human pathogens), but the database is getting larger daily. Newly-entered related organisms include *Nitrosococcus* and the γ3 symbiont from the *Olavius algarvensis* symbiosis. There are likely syntenic regions of each of these genomes that can be used as scaffolds to aid in connection of contigs. Each of these possibilities should be investigated to aid further in completion of the *Riftia* symbiont composite metagenome.

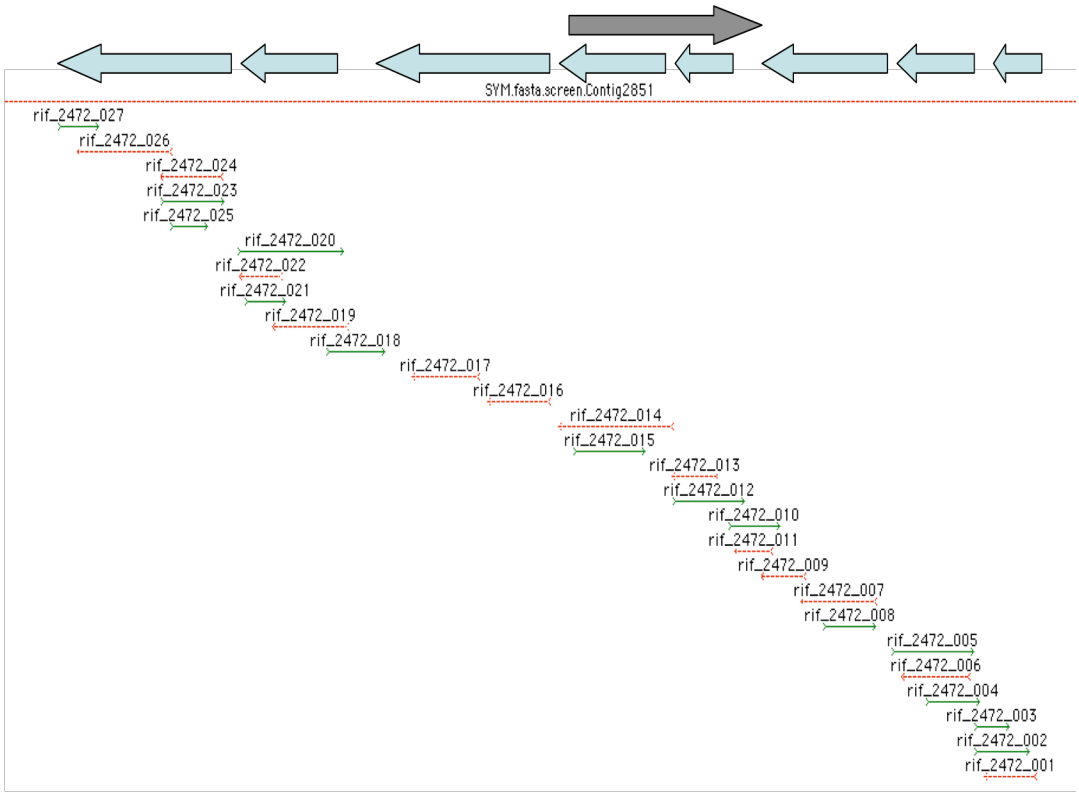


Figure 2.1 Alignment of partial sequences from the first assembly with the corresponding full-length contig from the second. The thick arrows at top correspond with the ORFs called in the cleaner version. The DNA sequences (represented by small arrows) below are the ORFs called in the first assembly. This is just one example of a contig whose component sequences were truncated more often in the first assembly than the second. Nine total ORFs were called in the second version, while 27 were called in the first. The first assembly was revised using data from the second. The entire length of "SYM.fasta.screen.Contig2851" (the cleaner assembly's version of contig rif 2472 in the metagenome) is 7289 basepairs compared to the old assembly's 9865 basepair equivalent contig.

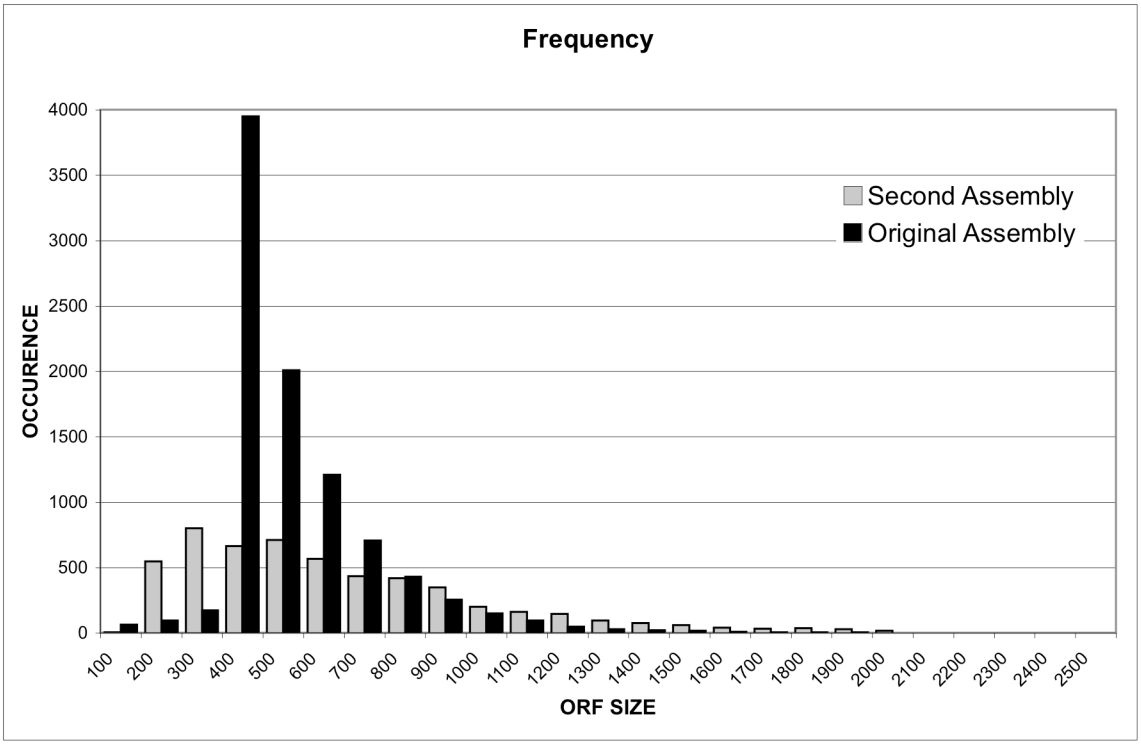


Figure 2.2 ORF size frequency distributions for the two metagenome assemblies. Sizes are in nucleotides. The sharp peak at 400-500 basepairs in the first assembly is not present in the second because there are fewer partial ORFs in the second. The second, "cleaner" assembly was used to modify the first.

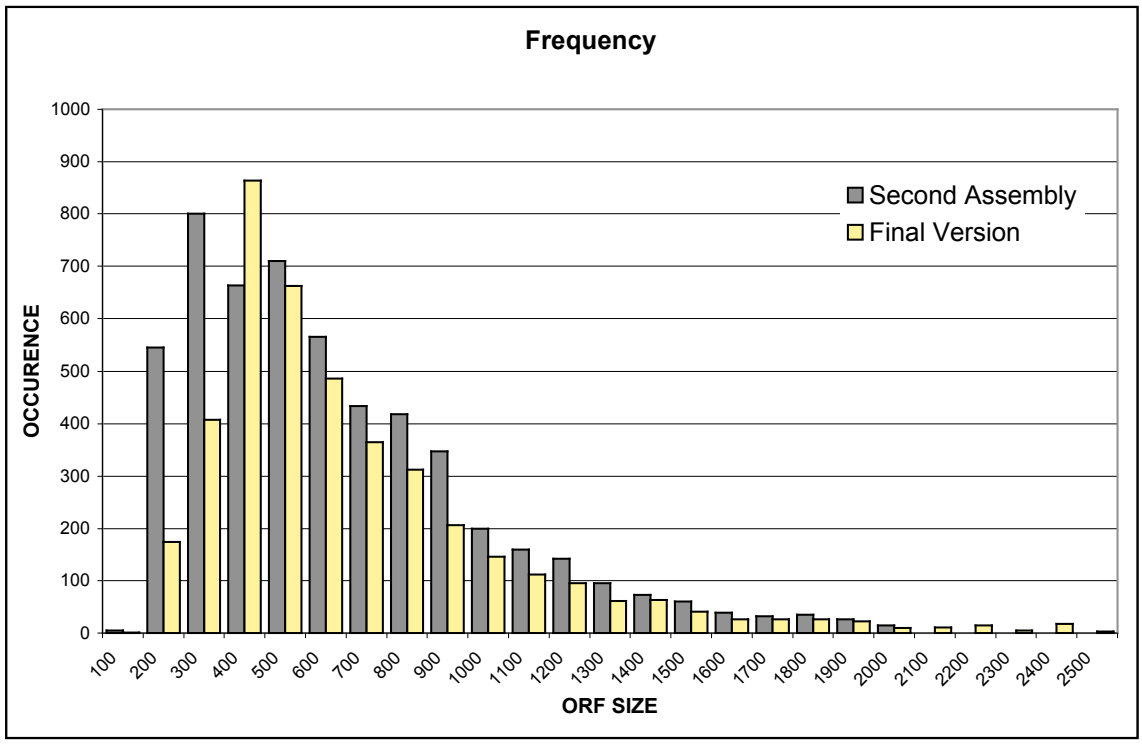


Figure 2.3 Comparison of ORF size profiles of the second with the final version of the metagenome after manual curation. The final version of the metagenome resembles the second assembly more than the first.

Figure 2.4 A hypothetical scenario that results in genome fragmentation. This is the case for fragmentation of the ATP citrate lyase ORFs, among many others. The top and bottom sequences correspond to separate contigs. A partial ORF from each contig hit to the same accession number in GenBank nr. When these partial ORFs are aligned with the sequence of that hit (boxed), the reason for the fragmentation becomes clear. At each end of the contigs is the sequence of the restriction site for Bcl I (TGATCA), the enzyme used in genome digestion in creation of the metagenome. The partial ORFs that hit to this sequence were located at the ends of the contigs and the 6 overlapping nucleotides were not enough for the two to co-assemble. These two contigs can now be connected.

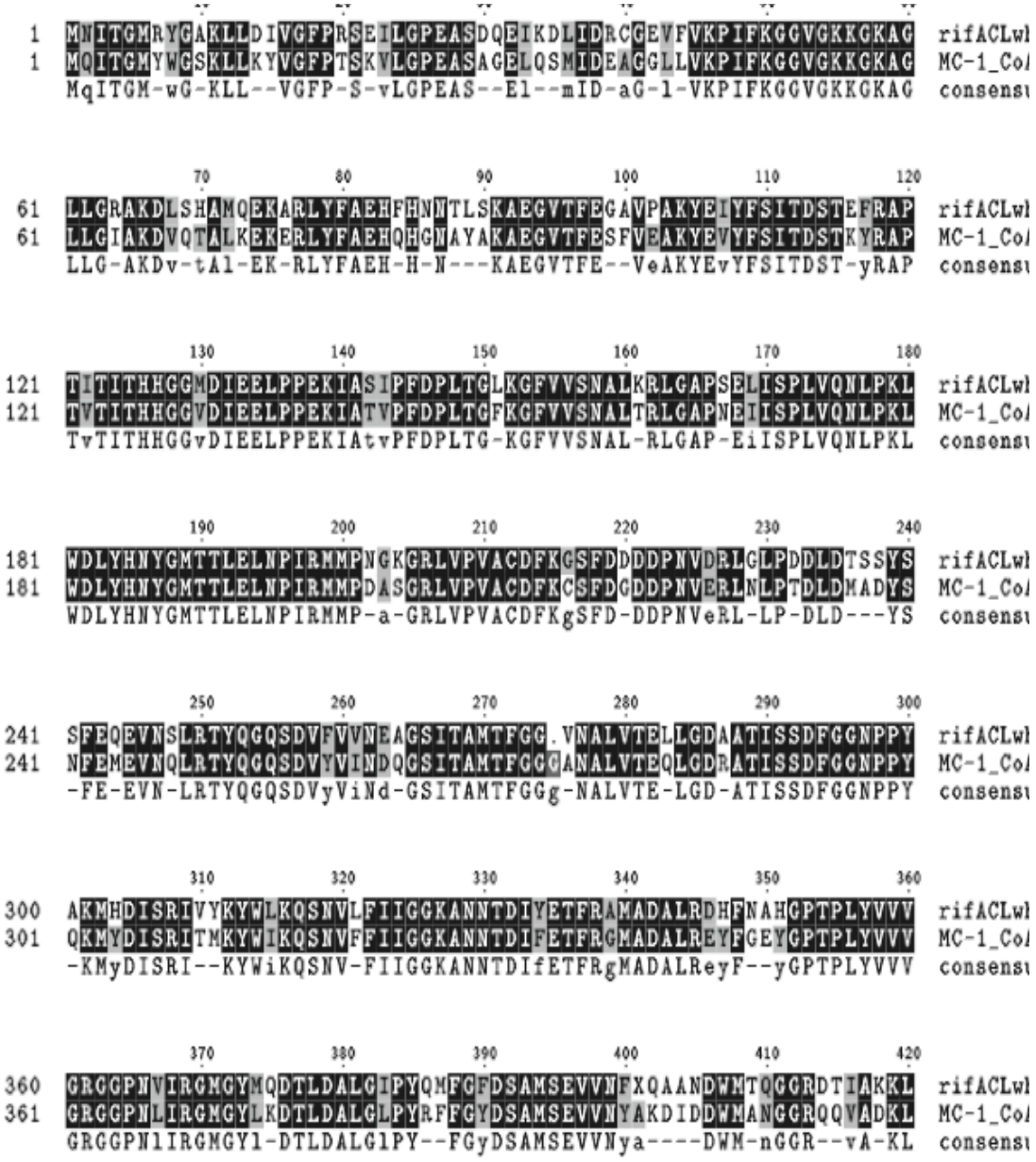


Figure 2.5 Manually curated amino acid alignment of *E. persephone* putative ATP citrate lyase with the *Magnetococcus* MC-1 homologue. "rifACLw" is the *E. persephone* sequence, and "MC-1_Co" is the *Magnetococcus* MC-1 sequence. All sequence is present in the metagenome but does not correlate well until it is aligned manually.

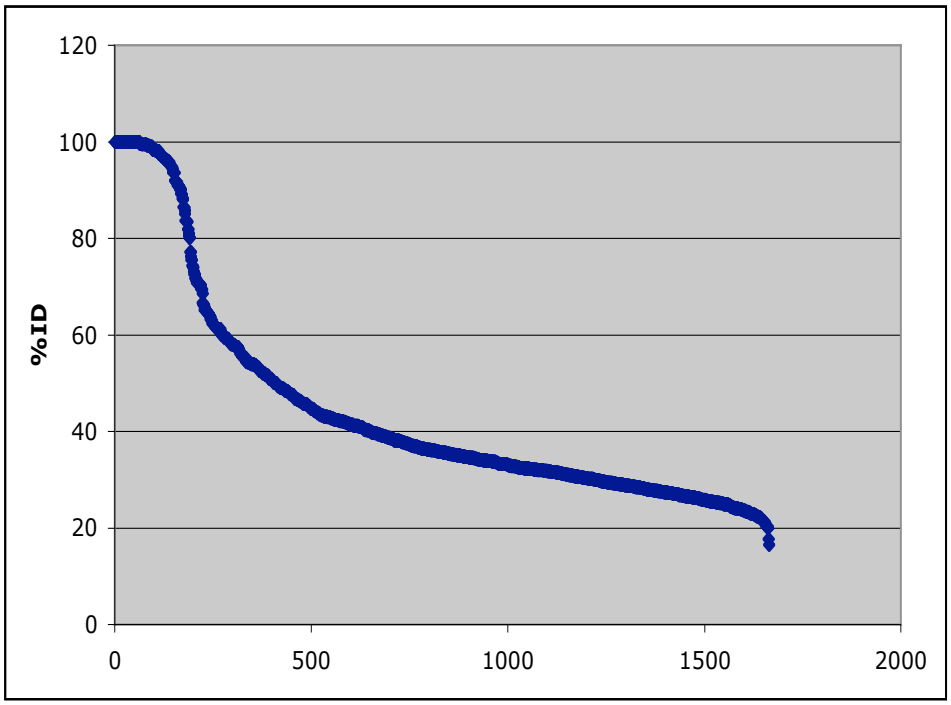


Figure 2.6 Scatterplot of non-self hits to ORFs within the "Version4" fasta file, vs. %ID to that ORF. All ORFs >90% ID were evaluated, most were duplications

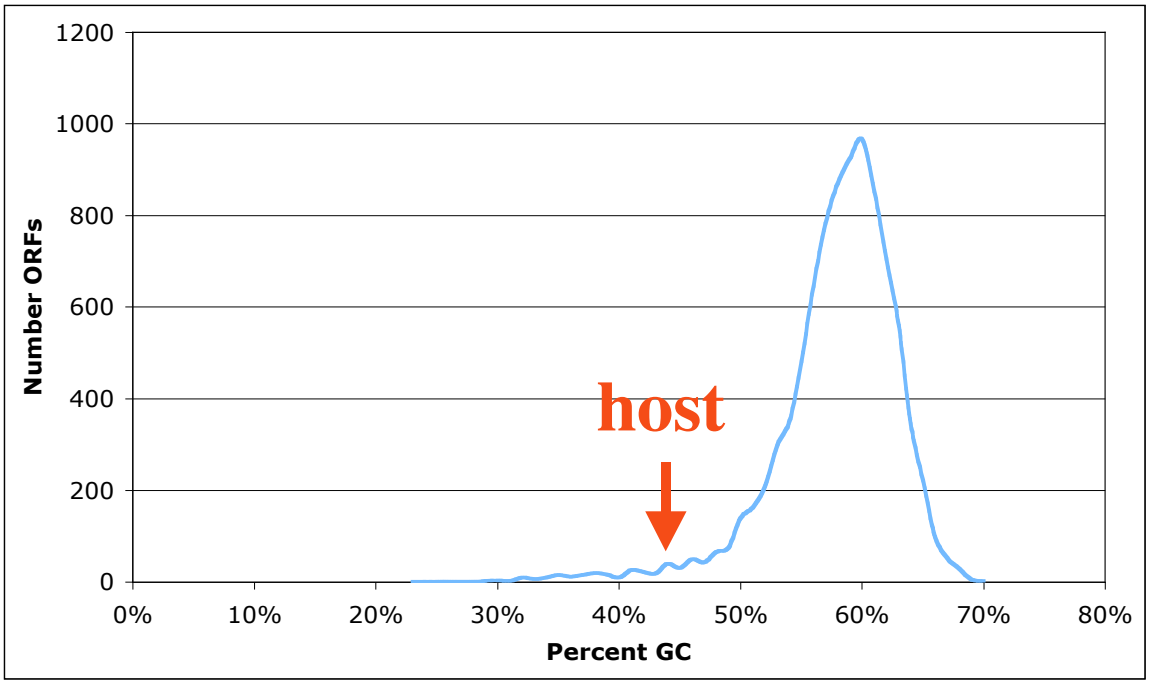


Figure 2.7 Percent GC vs. number of ORFs. The vast majority of ORFs from the metagenome have ~60% GC, represented by the sharp peak. This is the only major peak on the graph, supporting the theory of a single symbiont. The noise at 40-50% GC is likely due to horizontally transferred material in the symbiont genome, the majority of which contains on average 45% GC. Host DNA is also approximately 45% GC and this graph therefore verifies that there is low host contamination in the metagenome.

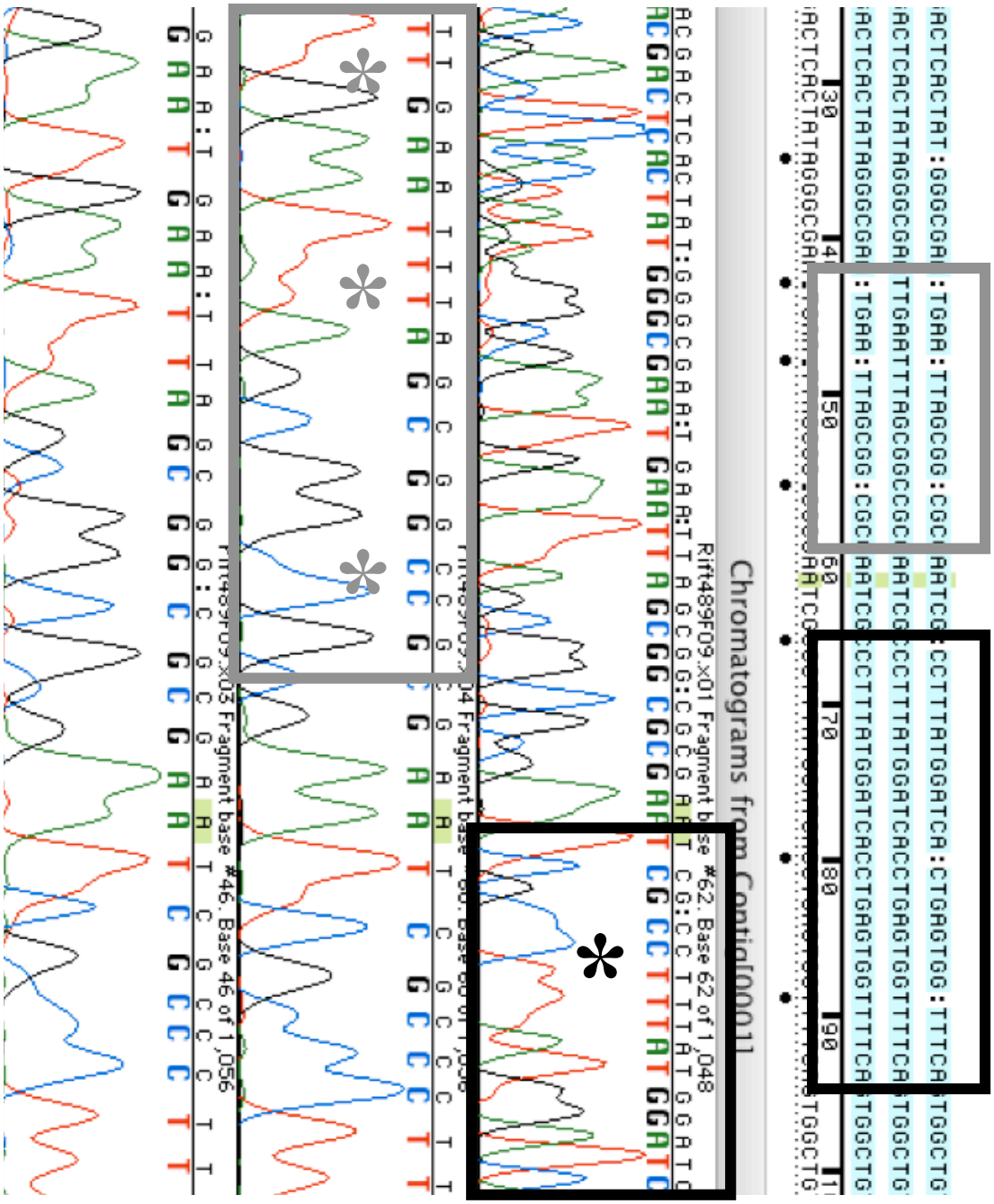


Figure 2.8 An example of sequencing errors from three forward sequencing reactions from the same clone. All sequences were used to generate the consensus in the first assembly. These sequences were trimmed in the second assembly. The chromatograms below correspond to the first three sequences above. ":" indicates a space in the sequence. The fourth sequence is the consensus. Asterisks mark nucleotide miscalls in the corresponding chromatograms. In the grey boxes, the middle sequence is different than the other two. Just downstream (black boxes), the upper sequence is different.

ACKNOWLEDGMENTS

Many thanks to Eric Allen for his knowledgeable advice in the manual curation process. Thank you to Craig Young for the generous donation of time at sea (and Horst Felbeck for the invitation), and the captain and crew of the R/V Atlantis and DSR/V Alvin for collections and help during cruises. SymBio Corp. and Amersham Biosciences performed the sequencing and assembly and The Scripps Genome Center provided analyses. Shellie Bench and Sheila Podell provided valuable discussions and analysis.

REFERENCES

Allen, E. A. and J. F. Banfield (2005). Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology* **3**: 489-498.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389-3402.

Badger, J.H. and G.J. Olsen (1999). CRITICA: coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution* **16**: 512-524.

Bateman**,** A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall and E.L.L. Sonnhammer (2004). The Pfam protein families database. *Nucleic Acids Research*, **32**(Database issue): D138-41.

Bendtsen, J.D., H Nielsen, G. von Heijne, S. Brunak (2004). Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, **340**(4): 783-795.

Benson, D.A., D.J. Lipman, I. Karsch-Mizrachi, J. Ostell, B.A. Rapp and D.L. Wheeler (2000). GenBank. *Nucleic Acids Research* **28**: 15-18.

Cary, S. C., W. Warren, E. Anderson, and S. J. Giovannoni (1993). Identification and localization of bacterial endosymbionts in hydrothermal vent taxa with symbiont-specific polymerase chain reaction amplification and *in situ* hybridization techniques. *Molecular Marine Biology Biotechnology* **2**(1): 51-62.

Cavanaugh, C. M., S. L. Gardiner, M.L. Jones, H.W. Jannasch, and J.B. Waterbury (1981). Procaryotic cells in the hydrothermal vent tube worm. *Science* **213**: 340-342.

Chen, K. and L. Pachter (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLOS Computational Biology* **1**(2):0106-0112.

Childress, J. J., R. W. Lee, *et al.* (1993). Inorganic carbon uptake in hydrothermal vent tubeworms facilitated by high environmental pCO_2. *Nature* **362**:147-149.

DeLong, E. F. (2005). Microbial community genomics in the ocean. *Nature Reviews Microbiology* **3**:459-469.

Distel, D.L., D.J. Lane, G.J. Olsen, S.J. Giovanonni, B. Pace, N.R. Pace, D.A. Stahl and H. Felbeck (1988). Sulfur-oxidizing bacterial endosymbionts: analysis of phylogeny and specificity by 16S rRNA sequences. *Journal of Bacteriology* **170**: 2506-2510.

Distel, D.L. and H. Felbeck (1988). Pathways of inorganic carbon fixation in the endosymbiont-bearing lucinid clam *Lucinoma aequizonata*. Part 1. Purification and characterization of the endosymbiotic bacteria. *Journal of Experimental Zoology* **247**: 1-10.

Ewing B. and P. Green (1998). Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**:186-194.

Ewing B., L. Hillier, M. Wendl, P. Green (1998). Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**:175-185.

Felbeck, H. (1981). Chemoautotrophic potential of the hydrothermal vent tube worm, *Riftia pachyptila* Jones (Vestimentifera). *Science* **213**: 336-338.

Felbeck, H. (1985). Carbon dioxide fixation in the hydrothermal vent tube worm *Riftia pachyptila* (Jones). *Physiological Zoology* **58**: 272-281.

Felbeck, H. and J. Jarchow (1998). Carbon release from purified chemoautotrophic bacterial symbionts of the hydrothermal vent tubeworm *Riftia pachyptila. Physiological Zoology* **71**: 294-302.

Felbeck, H., C. Arndt, *et al.* (2004). Experimental application of vascular and coelomic catheterization to identify vascular transport mechanisms for inorganic carbon in the vent tubeworm, *Riftia pachyptila. Deep-Sea Research Part I-Oceanographic Research Papers* **51**(3): 401-411.

Fraser, C. M. and R. D. Fleischmann (1997). Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18**: 1207-1216.

Gordon D., C. Abajian and P. Green (1998). Consed: a graphical tool for sequence finishing. *Genome Research* **8**: 195-202.

Gordon D., C. Desmarais and P. Green (2001). Automated finishing with Autofinish. *Genome Research* **11**: 614-625.

Gordon, D. (2004). Viewing and Editing Assembled Sequences Using Consed. In *Current Protocols in Bioinformatics*, A. D. Baxevanis and D. B. Davison, eds. New York: John Wiley & Co., 11.2.1-11.2.43.

Gaasterland, T. and C.W. Sensen (1996). Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie*, **78**: 302-310.

Girguis, P. R., J. J. Childress, J. K. Freytag and H. R. Stuber (2002). Effects of metabolite uptake on proton-equivalent elimination by two species of deep-sea vestimentiferan tubeworm, *Riftia pachyptila* and *Lamellibrachia* cf *luymesi*: proton elimination is a necessary adaptation to sulfide-oxidizing chemoautotrophic symbionts. *Journal of Experimental Biology* **205**: 3055-3066.

Gopal, S. and T. Gaasterland (2001). Automated genome annotation and comparative genomics. In *Biotechnology vol. 5b: Genomics and Bioinformatics*, C. Sensen ed. Wiley-VCH: Weinheim, Germany.

Gopal, S., G.A. Cross, and T. Gaasterland (2003). An organism-specific method to rank predicted coding regions in *Trypanosoma brucei*. *Nucleic Acids Research*, **31**:5877-5885.

Haft, D.H., J.D. Selengut, and O. White (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research*, 2003. **31**: 371-373.

Jennings R.M. and K.M. Halanych (2005). Mitochondrial genomes of *Clymenella torquata* (Malanidae) and *Riftia pachyptila* (Siboglinidae): Evidence for conserved gene order in Annelida. *Molecular Biology and Evolution* **22**: 210-222.

Lipman, R.S.A., K.R. Sowers and Y. Hou (2000). Synthesis of cysteinyl-tRNA^{Cys} by a genome that lacks the normal cysteine-tRNA synthetase. *Biochemistry* **39**: 7792-7798.

Lowe, T.M. and S.R. Eddy (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**(5): 955-64.

Nelson, D., J.B. Waterbury and H.W. Jannasch (1984). DNA base composition and genome size of the prokaryotic symbiont of *Riftia pachyptila* (Pogonophora). *FEMS Microbiology Letters* **24**: 267-271.

Podell, S. and T. Gaasterland, submitted. DarkHorse: A method for genome-wide prediction of horizontal gene transfer.

Robidart, J.C., S. Bench, R. Feldman, E.A. Allen, A. Novoradovsky, S. Podell, T. Gaasterland, and H. Felbeck, *in prep*. The duality of a facultative symbiont: population genomic analysis of the *Riftia pachyptila* endosymbiont.

Salzberg, S. L., A. L. Delcher, S. Kasif and O. White (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acid Research* **26**(2):544-548.

Sonnhammer, E.L. and R.M. Durbin (1994). A workbench for large scale sequence homology analysis. *Computer Applications in the Biosciences* **10**: 301-307.

Sonnhammer, E.L., G. von Heijne and A. Krogh (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol.* **6**:175-82.

Stein J.L., T.L. Marsh, K.Y. Wu and E.F. Delong (1996). Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology* **178**: 591-599.

Teeling, H., J. Waldmann, T. Lombardot, M. Bauer and F. Glockner (2004). TETRA: A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**:163-170.

Tettelin, H., D. Radune, S. Kasif, H. Khouri, and S. Salzberg (1999). Optimized multiplex PCR: Efficiently closing a whole-genome shotgun sequening project. *Genomics* **62**:500-507.

Tringe S.G., C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, E.M. Rubin (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554-557.

CHAPTER 3

On the Specificity of the *Riftia pachyptila-Endoriftia persephone* Association

ABSTRACT

Reexamination of the symbiotic component of the *Riftia pachyptila* trophosome is necessary due to the fragmented nature of the *Endoriftia persephone* metagenome and the advancements of molecular microbial diversity analyses since the last publication that dealt with the symbiotic composition of the trophosome. A low cycle number PCR was used to amplify 16S ribosomal RNA from DNA extracted from trophosomes of 4 worms from 9ºN, East Pacific Rise, and a clone library was generated from the products. No 16S rRNA sequences aside from that of *E. persephone* were recovered in this process, corroborating the results of past studies. Methanotroph- and archaea-specific 16S rRNA PCR primers, in addition to methane monooxygenase-specific primers supported this finding. Percent guanine plus cytosine content and the presence of distinct, solitary versions of known single-copy genes showed that the metagenome contains DNA from a single species. Multilocus sequence analysis using metagenomic data reveals that this species is virtually monoclonal, with a 0.29% single nucleotide polymorphic component, on average.

INTRODUCTION

The *Endoriftia persephone* metagenome is a composite genome from an environmental sample. The metagenome, by definition, belongs to the symbiont population as a whole, but does not belong to a single organism. Due to this and the non-random library preparation, problems with the assembly have resulted in a fragmented genome reconstruction. The fragmented nature of the *Endoriftia persephone* metagenome leaves uncertainty as to whether all recovered sequences are from one specific symbiont species. 16S rRNA phylogenies are used to elucidate prokaryotic phylogenies and distinguish between species (Hugenholtz *et al.* 1998, Woese 1987). If one 16S ribosomal RNA subunit is present in the metagenome, the possibility remains that a second 16S rRNA would have been sequenced had the investigators performed random sampling of the DNA, or additional sequencing.

In the purification process the bacteriocytes are lysed and symbionts are purified from the resultant material *via* Percoll density gradient centrifugation. In this procedure, only cells containing sulfur granules are concentrated at the bottom of the tube due to their high density. If another symbiont that does not contain sulfur granules is present in the trophosome very little, if any, of that symbiont's DNA would be collected with the main symbiont's DNA. These factors make further investigation of the metagenome and of the starting material (native trophosome tissue) necessary.

Genomic heterogeneity within the symbiont community is also of interest. Are the symbionts monoclonal as a result of one organism's infection and success in the trophosome? Alternatively, does the trophosome contain a population of individuals within a species? Results from previous work (Feldman and Black 1997; Di Meo *et al.* 2001) suggest that there is little variation in the symbiont species, within and between worms, in spite of differences in hydrothermal collection sites. Furthermore, the work of Di Meo *et al.* shows that symbionts from different sites along the East Pacific Rise show very little variation at the strain level.

Since the focus of these studies was symbiont biogeography, the sensitivities of the techniques used were not sufficient to detect different strains or a second species, if present in lower abundance in a single trophosome. The results of these two studies, therefore, only apply to the dominant symbionts within the associations sampled. Morphological differences are indeed observed in symbionts from a single trophosome (Bright and Sorgo 2003), but this may be an indication of cell cycle rather than phylogenetic diversity. Furthermore, fluorescent *in situ* hybridization (FISH) experiments with symbiont specific probes on dissected trophosome tissue have been published (Cary *et al.* 1993; Nussbaumer *et al.* 2006), and though unhybridized cells have been visualized, this does not necessarily signify the presence of another species or strain (Amann and Ludwig 2000).

At first glance the metagenome appears to contain a variety of symbionts. When catalogued by species, of the open reading frames identified, the majority (13.7%) have top BLAST (Basic Local Alignment Search Tool: Atschul *et al.* 1997) hits closest to related genes in the *Methylococcus capsulatus* genome. The next most common hit is to *Thiomicrospira crunogena* (10.6%), followed by *Thiobacillus denitrificans* (7.8%) and *Magnetococcus* sp. MC-1 (5.5%). This is somewhat surprising, as the latter two organisms are an epsilon and an alpha proteobacterium, respectively. The genome as a whole does not resemble any specific, related group of organisms within the division Bacteria [Figure 3.1]. One might expect more hits to other gamma proteobacteria, considering the relatively high number of gamma proteobacteria in GenBank (Benson *et al.* 2000). It is possible, however, that this finding is not unusual and may be the case for most environmental gamma proteobacterial genomes. It has long been known that phylogeny does not directly indicate metabolic capability; organisms that group together phylogenetically can have widely different metabolic capabilities (Feil 2004). It therefore follows that the genes that code for these variant pathways will look distinct from the organism's classified phylogenetic group. This degree of genomic diversity therefore requires further investigation. Additionally, the relatively high similarity to a methanotroph (*Methylococcus capsulatus*) provokes interest, though no methanotroph-like genes could be

recovered from the metagenome. Methanotrophs exist as symbionts in similar symbioses (Fisher *et al.* 1986), and Fisher and Childress (1984) demonstrated that trophosome tissue is capable of methane uptake. Moreover, bacteriocytes contain several enclosed bodies that have been recognized as mitochondria, but have the same ultrastructural characteristics as methanotrophic bacteria.

Given this information, the assumption cannot be made that the DNA sequenced is from a single organism. In order to reconstruct metabolic interactions between symbiont and host it is imperative to establish or exclude the presence of a possible secondary symbiont. Further investigation into the presence of additional symbiont(s) and into the genomic heterogeneity of the symbionts is necessary prior to metagenomic analysis.

MATERIALS AND METHODS

The *Riftia* symbiont metagenome.

Methods used in sequencing, assembly and annotation of the metagenome have been outlined elsewhere (Chapter 2).

Sample collection and DNA extraction.

Specimens were collected from 9ºN with the DSRV Alvin on the May 2000 cruise aboard the R/V Atlantis. Bulk trophosome was dissected aseptically and stored at –80ºC in a Ziploc bag until DNA extraction. Trophosome tissue was thawed and homogenized in phosphate buffered saline (137mM NaCl, 2.7mM KCl, 10mM Na_2HPO_4, 2mM KH_2PO_4), then treated with proteinase K (500 mg/ml) for 30 min at 65ºC. Samples were then extracted repeatedly with equal volumes phenol-chloroform (80:1) until proteins were visually undetectable, then extracted once with chloroform:isoamyl alcohol (24:1). Extracted DNA was precipitated with 100% 0.8 vol isopropanol at room temperature for 30 min. After centrifugation at 12,000 X g the samples were washed with 70% ethanol, recentrifuged, and ultimately eluted in 10mM Tris, pH 8. The same process was used to extract DNA from

vestimentum tissue of two worms. DNA of purified symbionts was prepared from samples from 9ºN, 1996 as outlined elsewhere (Robidart 2006).

Amplification of the low GC contig.

Primers were designed based on the low GC contig (LowGC_F: 5'-GTT CGW AAA CGT CCT GG-3' and LowGC_R: 5'-GGT TCS ACT TYA TCW CCC AT-3') and the *Riftia* pachyptila host 18S rRNA gene (*Riftia*18S_F: 5'-AAA CGG CTA CCA CAT CCA AG-3' and *Riftia*18S_R: 5'-GAC CCC AGA TCC AAC TAC GA-3'). 50ng DNA was amplified from 2 vestimentum samples and two whole trophosome samples (without symbiont density separation). Amplifications were performed in 50 ml reactions with 2U ExTaq DNA polymerase (TakaraMirusBio) in a reaction that included 1 mM final concentration primers, 1X PCR buffer and 1.6mM dNTP mix (supplied by TakaraMirusBio). The cycling reaction started with a 2 min initial denaturation at 94ºC followed by 30 cycles of 1 min at 94ºC, 1 min at 56ºC and 1.5 min at 72ºC. The final extension was at 72ºC for 10 minutes. Products were visualized by agarose gel electrophoresis and sequenced at the UCSD Rebecca and John Moores Cancer Center.

Creation of the bacterial 16S rRNA library.

50ng DNA from 4 individual worm trophosomes and a single vestimentum sample were used for separate PCR amplifications. Amplifications were performed as stated above but with the bacterial 16S primers 27F and 1541R. Cycling conditions included an initial 1 min denaturation at 94ºC, then 10 cycles of 1 min denaturation 94ºC, 1 min annealing at 50ºC, then 2 min extension at 72ºC. The final extension was for 10 min at 72ºC. Amplified products were ligated and cloned with the TA cloning kit and TOP10 chemically competent cells according to the manufacturer's instructions (Invitrogen Corp.). After overnight growth, plates were kept at 4ºC for 3 hours to encourage blue coloration of β-glucosidase positive colonies. White recombinant colonies were regrown overnight on GeneScreen Plus positively charged nylon filters (PerkinElmer). Filters were washed and DNA was fixed onto the membranes as outlined in Sambrook and Russel 2001. Hybridization with a biotinylated *E. persephone*

specific probe (sequence 5'-GGC CAT GTC TCC CGA CGT TTG-3' published by Cary *et al*. 1993) were performed overnight at 56ºC. An Anti-DIG-AP conjugate and CDP-Star (Roche) were used for detection using Roche protocols, and drained blots were imaged with X-ray film after 10 min exposures. The clones that did not hybridize to the symbiont-specific probe were grown overnight again and boil-lysed in 50 ml water for 10 min at 100ºC. 10 ml of the lysis was included in a 50 ml PCR reaction using M13 forward and reverse primers (Invitrogen Corp.) and the same conditions as the PCR above, excluding the cycling conditions. The cycling reaction started with a 2 min initial denaturation at 94ºC followed by 30 cycles of 1 min at 94ºC, 1 min at 55ºC and 2 min at 72ºC. The final extension was at 72ºC for 10 minutes. Amplified products were visualized after agarose gel electrophoresis and ethidium bromide staining. One product less than, and all products larger than 200bp were extracted from the gel using Qiagen's Gel Extraction Kit and were end-sequenced on an ABI sequencer using M13 primers at the Moores Cancer Center.

Amplification of methanotrophic DNA.

Primers were designed by aligning the 16S rRNA genes of the *Bathymodiolus azoricus* and the *Bathymodiolus puteosurpentis* symbionts to *E. persephone* and the *Riftia pachyptila* 18S rRNA and mitochondrial 16S rRNA sequence. The 16S genes of all methanotrophic *Bathymodiolus* symbionts are nearly identical. The objective was to amplify methanotroph-specific 16S rRNA sequence, and not *E. persephone* or host sequence. 16S primer sequences are MethylF: 5'-AAG ATT AGC TTG TTG GTG AGG TAA AAG CTC-3' and MethylR: 5'-TTA TGA GAT TAG CTT GCT CTC GCG AGG TTG-3'. Additionally, primers were designed to amplify the soluble methane monooxygenase gene by aligning the sequences of 2 *Methylomonas*, 3 *Methylosinus,* 2 *Methylocella* and 2 *Methylocystis* species. Primer sequences are *mmo*X_388F: 5'-GCS GAR CAG AAG AAC GGC TAT CT-3', *mmo*X_564R: 5'-ACG CGC TTC ATG CCC TTC CA-3', *mmo*X_541F:5'-TGG AAG GGC ATG AAG CGC GT-3' and *mmo*X_911R: 5'-GTC TTW ACC CAW GGC TCG ACC TTG AA-3'. The "universal methanotroph *pmo*A" published probe (Auman *et al*. 2000) and a modified version

with higher T_m were used as the forward primers: *pmo*A167Fa: 5'-AAC TTC TGG GGH TGG AC-3' and *pmo*A167Fb (modified): 5'-AAC TCC TGG GGH TGG ACN TAY TTC CC-3'. Reverse primers were designed based on 9 methanotrophic species' *pmo*A genes and include *pmo*A342R: 5'GCR ATV RYY GGC CAG TT-3', *pmo*A486Ra: 5'-CMA CGT CBT TRC CGA ATG T-3', *pmo*A486Rb: 5'-CAA CGT CBT TRC CGA ATG T-3', *pmo*A486Rc: 5'-CMA CGT CBT TAC CGA ATG T-3' and *pmo*A486Rd: 5'-CMA CGT CBT TRC CGA A-3'. Samples used were whole trophosome DNA extractions. PCR conditions were the same as insert amplification from the cloned organisms described above, except extension times were decreased to 2 minutes and annealing temperature used was 48ºC in the first reaction, 44ºC in the second.

Amplification of Archaeal 16S rRNA.

The same DNA extractions were used for amplification of Archaeal 16S rRNA genes using the same conditions as the cloned insert amplification, but with 2 differences. The annealing temperature used was 52ºC and the primers were archaea-specific A571F and UA1204R (Baker et al. 2003). As a positive control, a small piece of a frozen glycerol stock of *Halobacterium* sp. was sampled with a sterile toothpick and boil lysed in 50 ml water. 10 ml of this sample was used in PCR reactions. Bacterial 16S primers mentioned above were used to determine whether PCR inhibition existed from the trophosome extracted DNA.

Multilocus Sequence Analysis (MLSA).

A list of non-self BLAST hits to other ORFs within the genome was generated as discussed elsewhere (Chapter 2) and sorted by % identity (%ID). Genes included in MLSA were not present on a list of all of these hits with identity >15%. Many traces that compose the selected ORFs were aligned in Sequencher Version 4.5 (GeneCodes Corp.) and their chromatograms were used to investigate any questionable base calls. The assembled consensus of all genes analyzed was used as a query against the "cleaned" traces (those that had been quality checked and trimmed). The output alignments were used to quantify the

heterogeneity of each gene, by dividing the number of single nucleotide polymorphisms by the total number of nucleotides in that gene.

RESULTS

Assignment of the low GC contig.

18% of the DNA that makes up the largest 25 contigs is the metagenome has < 50% GC, while the average content is approximately 60% [Figure 2.7]. Amplification of the largest (10 kb) low GC contig from trophosome tissue, but not vestimentum (bacteriocyte-free) tissue was successful. The failed amplification from vestimentum DNA was not due to PCR inhibition, because 18S rRNA was amplified from the same samples [Figure 3.3]. The low GC material, therefore, is likely horizontally transferred DNA in the symbiont, the source of which is unknown.

Bacterial 16S identification.

From the bacterial 16S rRNA library, many clones were obtained from all samples, while those obtained from amplified vestimentum were negative. Two hundred and eighty-five total recombinant colonies with inserts were obtained and these were screened with an *E. persephone* specific 16S probe. Forty-five spots from the filter did not hybridize. Amplification of inserts from the corresponding clones showed that all either contained small inserts or were the expected size of a 16S rRNA gene. These were end-sequenced and determined to be *E. persephone* 16S rRNA sequences. One of the small inserts was sequenced but did not have any homologs in the GenBank non-redundant database.

Methanotroph identification.

All samples amplified with methanotroph 16S-specific primers. Six representative gel-purified products were sequenced and all sequences recovered were the *E. persephone* 16S rRNA sequence. Only non-specific amplification was observed with any combination of methane monooxygenase (*mmo* nor *pmo* forward and reverse) primers. None were of the

expected size (they were at least 200bp greater than expected). Sequence results of two of the smallest of such products confirmed that these were not methane monooxygenases.

Archaeal 16S identification.

Amplification from DNA extracted from whole (symbiont + host) trophosome tissue with universal archaeal primers was not successful, though amplification of bacterial 16S genes was positive, indicating that inhibition of PCR was not a factor.

Multilocus sequence analysis.

Interrogation of hundreds of single nucleotide polymorphisms (SNPs) from several chromatograms provides confidence that only “supported SNPs” (SNPs that occur in more than one trace sequence) from high read depth genes (>10 traces) warrant investigation. In these cases, 100% of the unsupported SNPs analyzed had questionable base calls at that position so were likely miscalled nucleotides. Chromatograms from genes with lower read depth were analyzed before accepting or rejecting a SNP. The average nucleotide heterogeneity, calculated as the percentage of single nucleotide polymorphisms (SNPs), from the selected genes is 0.29%, and ranges from 0% to 1.49%.

DISCUSSION

Prokaryotic small subunit ribosomal RNA (16S rRNA) sequences are used to construct microbial phylogenies and to identify prokaryotic species from environmental samples (Woese 1987). The 16S rRNA complement of the metagenome was analyzed post-assembly, and only a single sequence is present. However, in construction of the metagenome, the threshold used for assembly of the traces was intended to increase assembly; variations within the traces were allowed into the assembly of the representative consensus sequences. This is a problem when managing non-coding genes, as small sequence variations are much more significant than mutations in coding genes, which are often synonymous. This especially applies with respect to 16S ribosomal RNA phylogenies, where minor differences can signify the presence of a second species. In the metagenome

assembly, any 16S rRNA sequences that are present in low abundance would assemble with the main symbiont's 16S sequence, but their sequences would be "lost" in the creation of the consensus sequence. In other words, if four sequences are from organism A and one sequence is from organism B, the consensus sequence in the metagenome would be organism A's, since the traces that are compiled to create the consensus are dominated by that organism's sequences. This consideration necessitates analysis of each of the traces that compose the 16S contig. Forty-four total traces compose this contig, and they show 100% identity to the 16S rRNA sequences of the *Riftia* symbiont in GenBank. Moreover, this contig contains the partial 23S rRNA and therefore the entire internal spacer (ITS) region for the symbiont, and the traces assembled to form these regions also show 100% identity. These results indicate very low or no strain-level variation in the symbiont community used to create the metagenome. However, the possibility of the presence of an unsequenced phylogenetic signature remains.

One genomic signature of an organism is its GC (guanine+cytosine) content (Abe *et al.* 2003). Microbial genomes as a whole have an essentially stable %GC (percentage of guanines + cytosines relative to the total nucleotide composition), and therefore the presence of a second symbiont with a different GC content within the metagenome can be easily identified. Upon interrogation, all open reading frames from the metagenome contained relatively constrained GC contents and fell within a sharp peak on a graph of %GC vs. number of ORFs [Figure 2.7]. The peak at 60% is attributed to symbiont DNA and has the expected GC content, as determined previously (Nelson *et al.* 1986). However, this graph does show some noise at ~45% GC. Surprisingly, 18% of the DNA contained within the twenty-five largest contigs has a relatively low GC content. Amplifications with primers designed from this low GC sequence were successful using DNA from trophosome tissue (which contains symbiont and host DNA), but not from vestimentum (bacteriocyte-free host tissue). To demonstrate that this is not due to PCR inhibition, 18S ribosomal RNA was successfully amplified with specific primers from the vestimentum DNA [Figure 3.3]. Sequencing of the

amplification products confirmed the primer specificity and the presence of the low GC material in the host's trophosome, but not the vestimentum tissue. The low GC material is discussed further in Chapter 4 of this thesis. The noise in Figure 3.3 can also be attributed to host contamination in the metagenome, since the host GC content is approximately 45% (Jennings and Halanych 2005). This graph therefore also serves as verification of the relative purity of the metagenome DNA, since the peak at 60% GC represents most of the material sequenced. This peak supports the presence of a single symbiont but does not rule out the possibility of another with a similar GC content, nor does it demonstrate the absence of a second symbiont in the trophosome tissue prior to symbiont preparation by Percoll density gradient centrifugation.

Small subunit ribosomal RNA phylogenetics is aided with the use of low-cycle number PCR amplifications, which are commonly used to access phylogenies of bacterial symbionts in multiple partner associations (Dubilier *et al.* 1999). The increased fidelity of the low cycle number PCR is due to the exponential amplification of genetic material that occurs with these reactions. The product from usual PCR amplifications is dominated by the most frequently occurring 16S sequence(s) of that community and is biased in that manner. Less abundant members of the community can be missed in analyses. Low cycle number PCR amplifications can circumvent this problem; with less cycles, amplification still occurs exponentially but the increase in the most abundant sequences does not reach the point of absolute numerical dominance early in the cycling reaction (Acinas *et al.* 2005; von Wintzingerode *et al.* 1997; Suzuki and Giovannoni *1*996). This method is much more sensitive because minor members of the community are obtainable. Though very rare members remain inaccessible with this technique, it is adequate for recovery of most organisms of interest. This is the method selected to determine whether a single bacterial species resides within the trophosome of *Riftia pachyptila*. The DNA used in this amplification was purified from homogenized trophosome tissue without the density gradient separation in order to sample a possible second symbiont without sulfur inclusions. Nylon filters containing DNA from all two hundred

and eighty-five colonies obtained from cloning of these PCR products were screened with an *E. persephone* specific 16S probe. All unhybridized clones were amplified, all products >200 bp were sequenced and none contained new (non-*E. persephone*) 16S rRNA genes.

Primers designed to amplify 16S rRNA sequences from known methanotrophic symbionts, to the exclusion of the *E. persephone* 16S rRNA, only amplified *E. persephone* 16S sequences from the whole trophosome sample. *E. persephone* 16S rRNA amplified likely as a result of unspecific primer annealing to the *E. persephone* gene due to the absence of a better match for primer hybridization. "Universal" methane monooxygenase (particulate and soluble) primers failed to amplify from trophosome DNA, indicating the absence of methanotrophic symbionts.

Archaeal 16S rRNA primers were unsuccessful in amplifications from trophosome DNA, though they successfully amplified an archaeal positive control [Figure 3.5]. The primers used were designed to specifically amplify all published archaea in GenBank (including Korarchaeotes and Nanoarchaeotes) but no published bacteria or eukaryotes (Baker et al. 2003). These results, when taken with the bacterial 16S library results, support the hypothesis of a specific single-species association in the *Riftia pachyptila* symbiosis.

Multilocus sequence analysis (MLSA) is commonly used to elucidate population structure within microbial communities and between strains of a particular species (Feil 2004; Gevers *et al.* 2005). Genes chosen for MLSA should have the following characteristics: 1) The genes should be ubiquitous in the organisms sampled. 2) The genes should be present in single copy. If a gene is present in multiple copies as a result of gene duplication, one copy may be evolving to perform a different function, in which case it would be evolving much faster and would artificially skew heterogeneity analyses. When investigating metagenomes, the use of single copy genes is also necessary because without a closed genome, it is difficult to discern between paralogs (the result of duplications) and orthologs (homologous genes from other organisms) from the population. 3) The genes must not have a function where

recombination confers a selective advantage, such as in antigens of pathogenic microbes. The increased genetic variation of such loci will also artificially increase estimates.

Thirty ORFs with these characteristics were chosen for MLSA with the *E. persephone* population [Table 3.1]. Difficulties are presented when determining whether a gene within a metagenome is single copy. Since the genome is not closed, it is impossible to determine for certain whether variations are due to population structure or to heterogeneity from duplicated and diverged genes from a single genome. To avoid this problem, genes known to be single-copy from sequenced genomes were chosen for analysis. All single-copy genes investigated (*mreB*, *recA*, *dnaA, dnaK, rpoB*) had a single representative assembled sequence in the metagenome, but in the case of *dnaK* the ORF was artificially fragmented into 2, one representing the 5' end and the other the 3' end of the gene, and they are sequential ORFs on the same strand of a single contig. The *rpoB* gene was also artificially fragmented. Of the remaining genes chosen, none contained homologs in the metagenome that showed greater than 15% identity to the consensus sequence, and all are the only representative sequences for their specific accession number hits from a BLAST against GenBank's non-redundant database. Some genes are "housekeeping" type genes which are less susceptible to viable duplication (Yanai *et al.* 2000), and some provide metabolic functions so are presumably more variable as a result of their required functionality under a dynamic range of substrate conditions. A diversity of genes, rather than a subset of genes (e.g. housekeeping genes only), was selected in order to calculate the heterogeneity representative of the entire genome. The consensus ORFs were aligned with the raw sequences that comprise that ORF, and the number of supported nucleotide polymorphisms was counted.

The average read depth of the genes chosen was 18.6 and the average nucleotide heterogeneity is 0.29%. Surprisingly, 32% of the genes analyzed have no polymorphic nucleotides. More than 95% of mutations are synonymous. These results were unexpected because this low degree of heterogeneity should not preclude genome assembly. The lack of

random sequencing and the low quality of sequences recovered are therefore the likely causes of fragmentation of the metagenome.

When evaluated with respect to the only other metagenomic population structure available, that of the acid mine drainage (AMD) biofilm (Tyson *et al.* 2003), the diversity observed with *E. persephone* is comparable. In the AMD study, the *Leptospirillum*, Group II genome is relatively homogeneous, with an average of 0.08% polymorphic nucleotides genome-wide. On the other hand, the unusual *Ferroplasma*, Type II genome is much more polymorphic, at 2.2%. This might be contributed to a higher mutation rate that likely results from the *Ferroplasma* population's high degree of recombination. Genome mosaicism as seen in the *Ferroplasma* population does not exist in the *E. persephone* metagenome [Figure 3.6], so this is not a fair comparison.

The *E. persephone* population is not a single genotype, but is virtually monoclonal (0.29% heterogeneity is very low). Nussbaumer *et al.* (2006) found that at the initiation of symbiosis, a number of bacteria infect the epidermal tissue of settled juveniles concomitantly. When considered with regards to the Nussbaumer *et al.* results, we cannot discern whether any stabilizing selection occurs within the trophosome, because we have no data on the population structure of the symbiont before or during infection.

The high diversity of organisms with similarity to *E. persephone* despite a single 16S type [Figure 3.1] is likely a result of the scarcity of sequenced bacterial relatives from reducing or hydrothermal areas. The closest sequenced organism at the time of this study is *Thiomicrospira crunogena*, which is only 86% identical at the 16S level. Accordingly, the genes found in the symbiont metagenome code for proteins that might have novel functional properties compared to their sequenced homologs. The diversity of ORF orthologs was verified using DarkHorse (Podell and Gaasterland, submitted), which takes into account all top BLAST hits of "equivalent" value (equivalency is defined as 0-40% of the bit score, by the user). DarkHorse then makes a phylogeny of those hits and places the symbiont's ORF within the phylogeny. Based on the phylogeny, a lineage probability score (LPI) is calculated for

each ORF within the genome. The higher the LPI, the closer the phylogeny is to what would be expected for the symbiont. By DarkHorse analysis, no single genus in GenBank is responsible for more than 10% of BLAST hits with e values $<e^{-05}$ and coverage of > 60bp. LPI frequency distributions usually have a peak near 0.9, whereas the *E. persephone* peak is at about 0.6 - 0.7 [Figure 3.7]. This is another indication that the *Riftia* symbiont has very few close relatives with sequences available in GenBank.

Though gamma proteobacteria represent the majority of sequenced microbial genomes, a high percentage of the *E. persephone* ORFs are more closely related to organisms outside of the gamma proteobacteria. The reason for this is made clear when these hits are categorized by function. When ORFs that hit to gamma proteobacteria are graphed according to their function, those that fall within the "metabolism" category are consistently approximately 12-14% of the hits. The vast majority of genes function in "cellular/molecular processes" (housekeeping genes). When compared to organisms outside of the gamma proteobacteria, however, 31-43% of the ORFs that hit to each order are involved in carbon, nitrogen or sulfur metabolism. These organisms look less similar to *E. persephone* with respect to housekeeping processes. These insights show that the core metabolic framework of the symbiont places it within the gamma proteobacteria, but regarding redox processes, it is similar to a diversity of organisms. The most likely reason for this is that most of the gamma proteobacterial species sequenced do not represent the diversity of their phylogenetic group. The majority of organisms with sequenced genomes within this group are pathogenic and are not found in diverse ecosystems, whereas the cultured representatives of other groups are of interest because of their metabolic capabilities. It is very likely that the redox metabolism of *E. persephone* would show closer similarity to the gamma proteobacteria if more ecologically diverse members had sequenced genomes.

The presence of a single symbiotic species within the *Riftia pachyptila* trophosome is not surprising when considering that the association has been studied for many years now without the identification of a secondary symbiont. The diversity of organisms that the ORFs

within the genome relate to is unexpected, however. A lot of this diversity may be contributed to the truncation of the ORFs. Perhaps if all ORFs were full length the symbiont would be more genomically similar to other gamma proteobacteria. The DarkHorse program compensates for such artifactual hits by taking into account all significant top BLAST hits, and annotating the gene based on the most common lineage among all proteins in the genome, that occurs within those orthologs (Podell and Gaasterland, submitted). Moreover, the consistent functional distribution of orthologs from orders within, and from orders outside of the gamma proteobacteria demonstrates that the orthologs from GenBank are not random. It is likely that with the release of more environmental, and specifically hydrothermal gamma proteobacterial genomes *E. persephone* will not seem so genomically distinct. The percentage of hypothetical genes in the genome, however, is about equivalent to the content in other genomes (10.3%), perhaps indicating that the pathways present are not novel, but rather present in a diversity of bacteria.

To gain further insight, a metagenome to sequence would be that of the biofilm community that covers the cuticle of recently-settled juveniles. This is likely the origin of the symbiont population that infects the worm and comparison of the symbiont population at this stage to that in an adult worm will provide information on the specialization of the association. Since the settled juveniles at this stage are so difficult to recover, let alone manipulate, and the amount of DNA extracted is likely to be minimal, this study will not be possible for quite some time. A more feasible alternative would be to amplify the MLSA genes used in this study and relate their heterogeneity to that in the symbiont metagenome. Exploration of the presence and specificity of any stabilizing selection directed by the host is appealing. In a well-studies bioluminescence symbiosis, the *Euprymna scolopes* squid expels *V. fischeri* symbiont mutants incapable of light production (Visick *et al*. 2000). Perhaps a similar mechanism is occurring within the *Riftia* association, where symbionts that do not produce carbon efficiently are lysed by the host.

The smaller, related hydrothermal vestimentiferans *Oasisia alvinae, Tevnia jerichonana*, *Escarpia* spp. and *Ridgeia piscesae* often contain phylogenetically identical symbiont species, based on 16S sequences (Di Meo *et al.* 2001). Experiments indicate, however, that the species vary genomically (Di Meo *et al*. 2000; Edwards and Nelson 1991). Genomic subtractive hybridization of these related symbionts with the *Riftia* symbiont might provide insight into adaptations within hosts and the specificity of the association. It would also be interesting to compare energy conversion and carbon fixation mechanisms to these smaller species.

The diversity of genes in the genome of *E. persephone* optimize this organism for a dual lifestyle: it can live in both the variable and toxic hydrothermal realm and the stable and nourishing trophosome environment. The organism must sense and respond to its different environments and change physiology accordingly. The metabolic flexibility of the symbiont is the subject of Chapter 4.

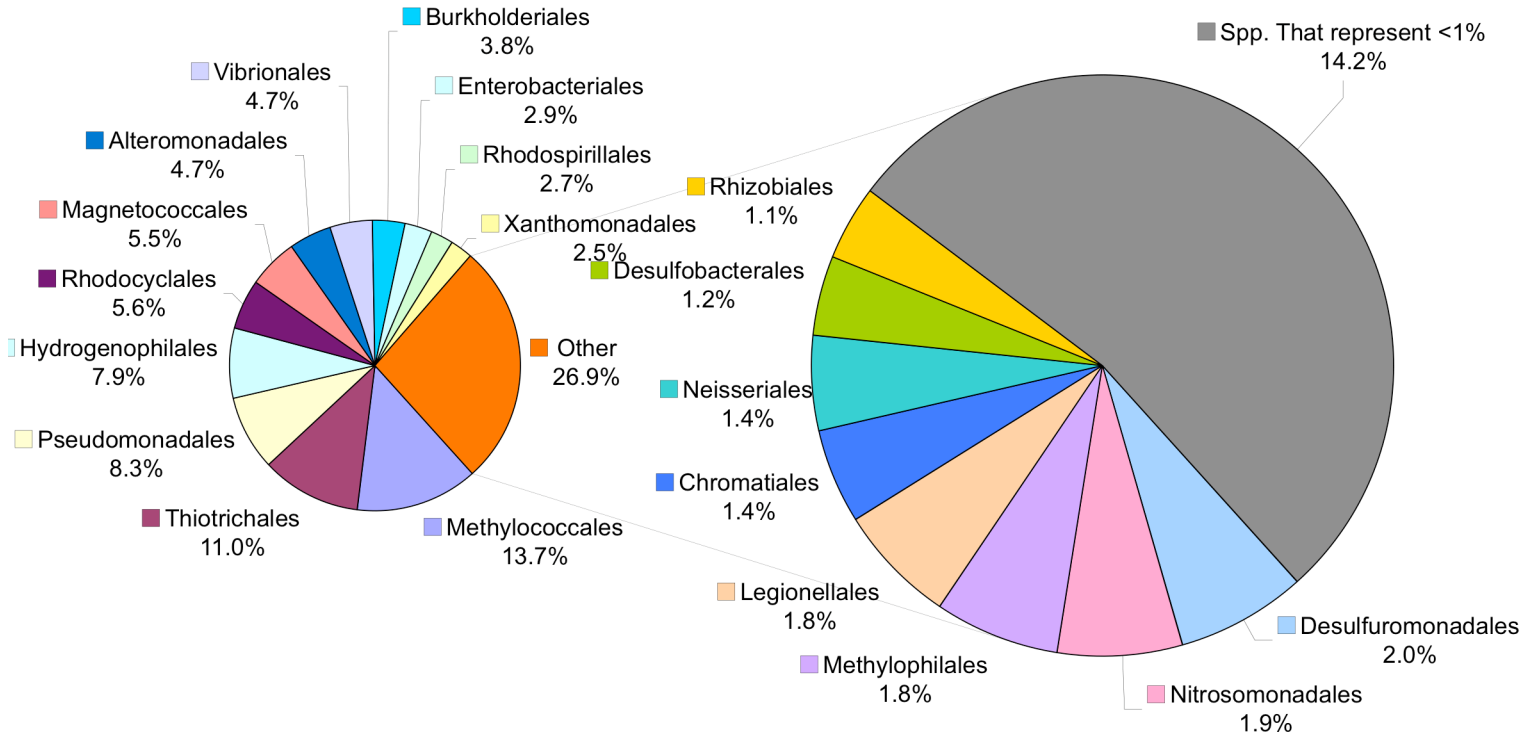


Figure 3.1 Closest relatives of *E. persephone*, identified by highest ORF homology within GenBank. Please note that this pie chart categorizes organisms by order, not species. Right pie is the component of "Other" in left pie. The metagenome resembles a large variety of organisms, not merely gamma proteobacteria.

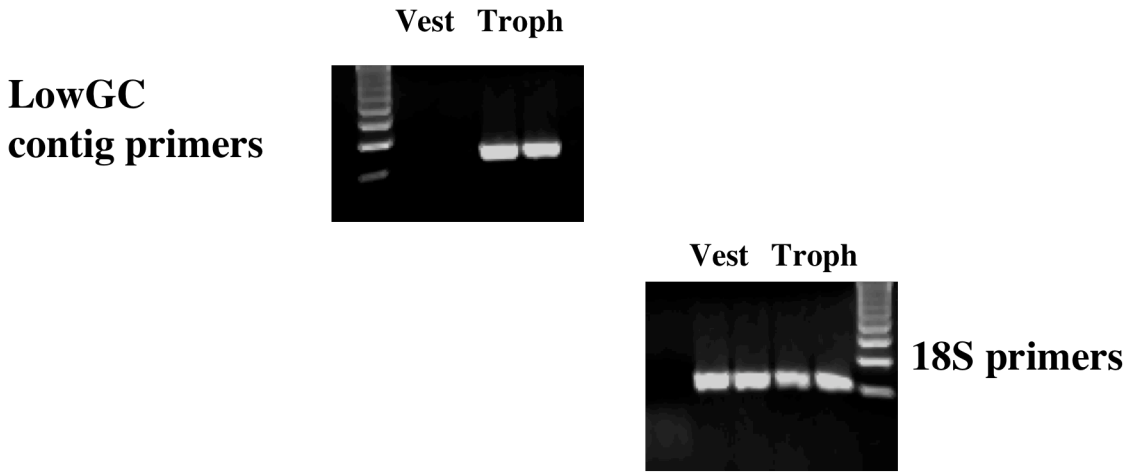


Figure 3.3 Amplification of low GC material from trophosome DNA but not vestimentum DNA. Bottom band in both marker lanes is 200bp, and bands increase by 200bp thereafter. First gel shows amplification using primers designed from the 10kb low GC contig. Amplification of vestimentum was negative, though amplification was positive with host-specific 18S rRNA primers (second gel).

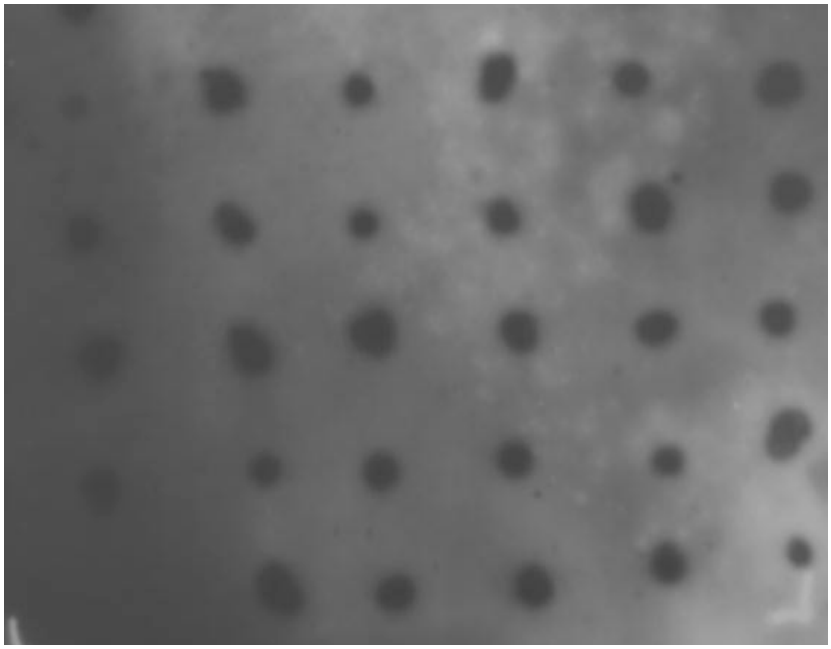


Figure 3.4 Filter of colonies from 16S library, hybridized with *E. persephone*-specific 16S probe. All inserts were the symbiont 16S sequences, or were products of unknown origin that were not 16S genes of other species.

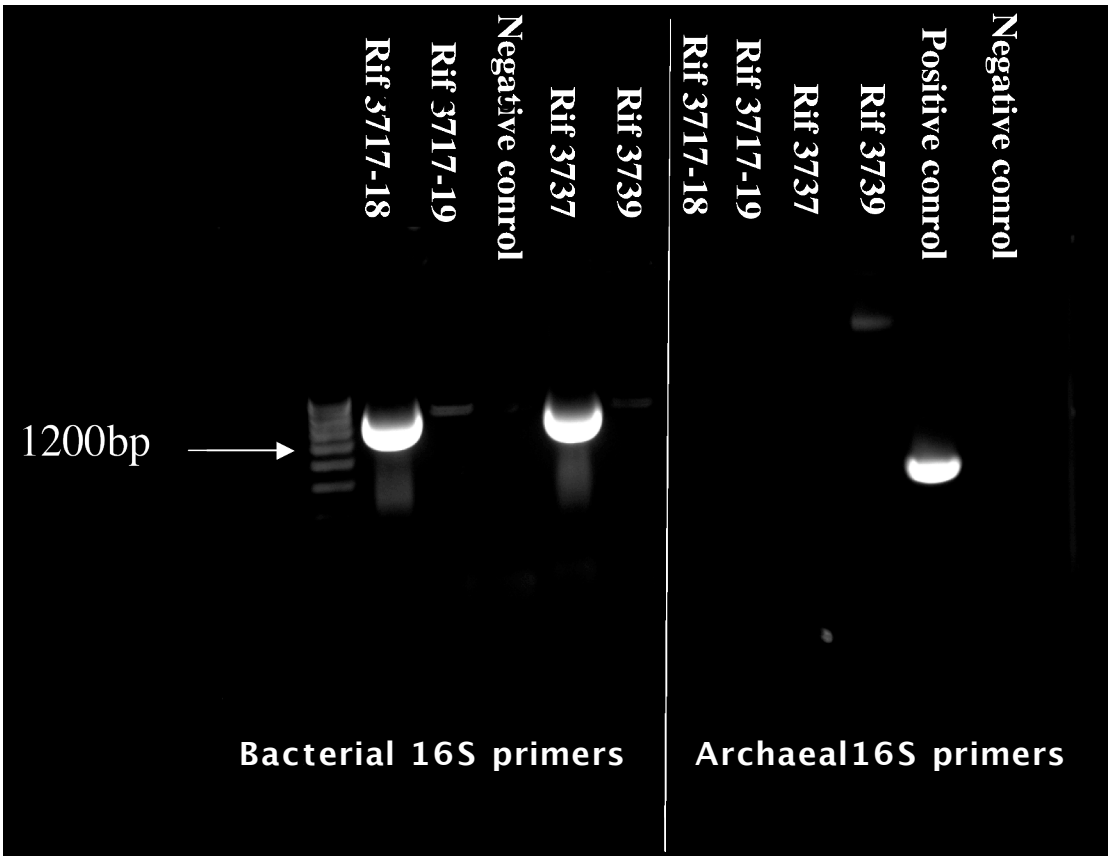


Figure 3.5 Agarose gel of 16S rRNA PCR products from trophosome tissue DNA. Rif 3717-18 and -19 are DNA samples from whole trophosome tissue, while Rif 3737 and 3739 are DNA samples extracted from purified symbionts after Percoll gradient centrifugation. Amplifications on the left of the gel are from reactions with bacterial primers, while the right are using archaeal primers. *Halobacterium* sp. was used as the archaeal positive control.

Table 3.1 Genes chosen for multilocus sequence analysis and their corresponding heterogeneities. Some genes are not full open reading frames, since they were not available. Heterogeneity is calculated as the percent of nucleotides that are polymorphic over the range analyzed. Heterogeneity ranged from 0-1.49%, with an average of 0.29%.

rif_ID	Annotation	Annotation	AveRead Depth	Length (nt)	Heterog eneity
1114_002	Actin-like ATPase involved in cell morphogenesis	mreB	9	122	0.00%
1296_004	Chaperone protein dnaK (Heat shock protein 70)	dnaK	6	419	0.00%
1296_005	Chaperone protein dnaK (Heat shock protein 70)	dnaK	6	371	0.00%
2471_001	chromosomal DNA replication initiator DnaA	dnaA	35	1271	0.16%
2373_011	DNA-directed RNA polymerase beta chain	rpoB	10	1334	0.22%
1461_051	RecA/RadA recombinase	recA	11	90	0.00%
2169_002	Type IIA topoisomerase, A subunit	gyrA	25	788	0.40%
1451_001b	RecG-like helicase	recG	9	803	0.37%
2271_001	Ornithine carbamoyltransferase (OTCase)	argF	20	923	0.33%
2271_002	inner membrane protein	COG3103	20	677	0.00%
2271_003	Cell division inhibitor	sulA	27	605	0.33%
2271_005	LexA-SOS-response transcriptional repressor	lexA	15	635	0.31%
2176_003	hypothetical protein PJS6w01002739	xthA	9	715	0.30%
2344_007	COG: TolC Outer Membrane Protein	tolC	25	5466	0.40%
2373_001	Transcription antiterminator	nusG	12	353	0.30%
2382_012	4-alpha-glucanotransferase/amylomaltase/Disproportionating enzyme	malQ	26	532	0.75%
2388_006	membrane "protein," hemolysin III homolog	COG1272	8	659	0.00%
2393_002	organic radical activating enzyme	nrdG	30	491	0.41%
2410_002	dTDP-glucose "4,6-dehydratase"	rmlB	8	1010	1.49%
2427_008	Glutathionylspermidine synthase	gsp	25	560	0.18%
2433_001	Cytochrome c peroxidase	mauG	18	1106	0.53%
2444_010	ABC-type transport system involved in resistance to organic "solvents," ATPase component	ttg2a	16	806	0.50%
2446_011	Outer membrane protein	COG0729	28	1000	0.30%
2466_006	Uncharacterized component of anaerobic dehydrogenases	torD	9	863	0.50%
2176_003	hypothetical protein PJS6w01002739	xthA	9	715	0.30%
2470_017	COG0406: Fructose-2,6-bisphosphatase	sixA	22	650	0.00%
2470_024	dGTP triphosphohydrolase	dgt	46	866	0.23%
2470_025	COG3045: Uncharacterized protein conserved in bacteria	creA	38	485	0.00%
2471_010	Type IIA topoisomerase, B subunit	gyrB	24	2150	0.30%
2472_006	Adenylate kinase (ATP-AMP transphosphorylase) (AK)	adk	12	503	0.00%
				Average:	0.29%

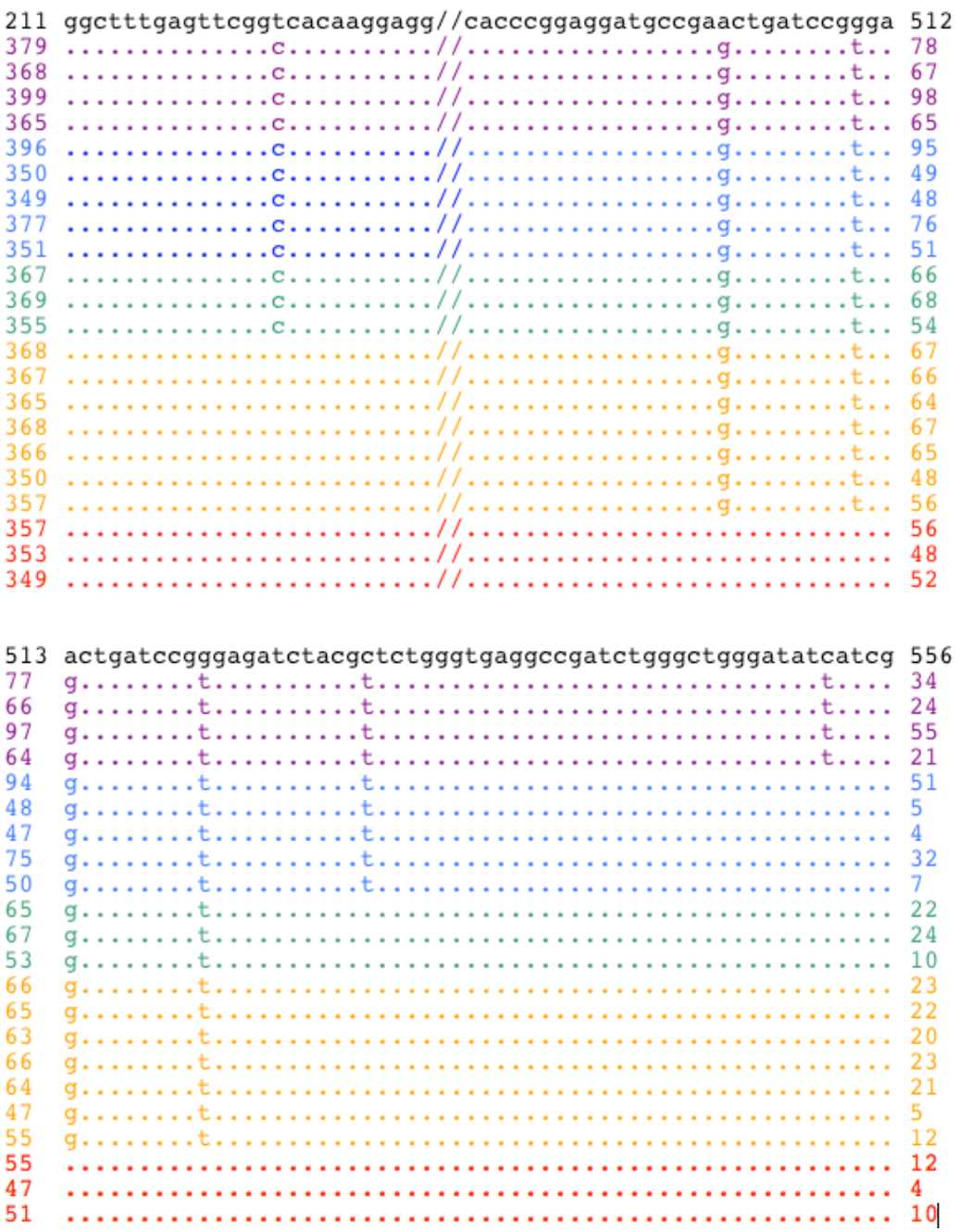


Figure 3.6 Example alignment of traces that compose the *fli*H gene. The consensus sequence is at the top in black, and the individual traces are below. Dots represent nucleotides that are the same as the consensus, letters represent SNPs. Each genotype is color coded. "//" signifies a break in the sequence; the intervening sequences are completely homologous

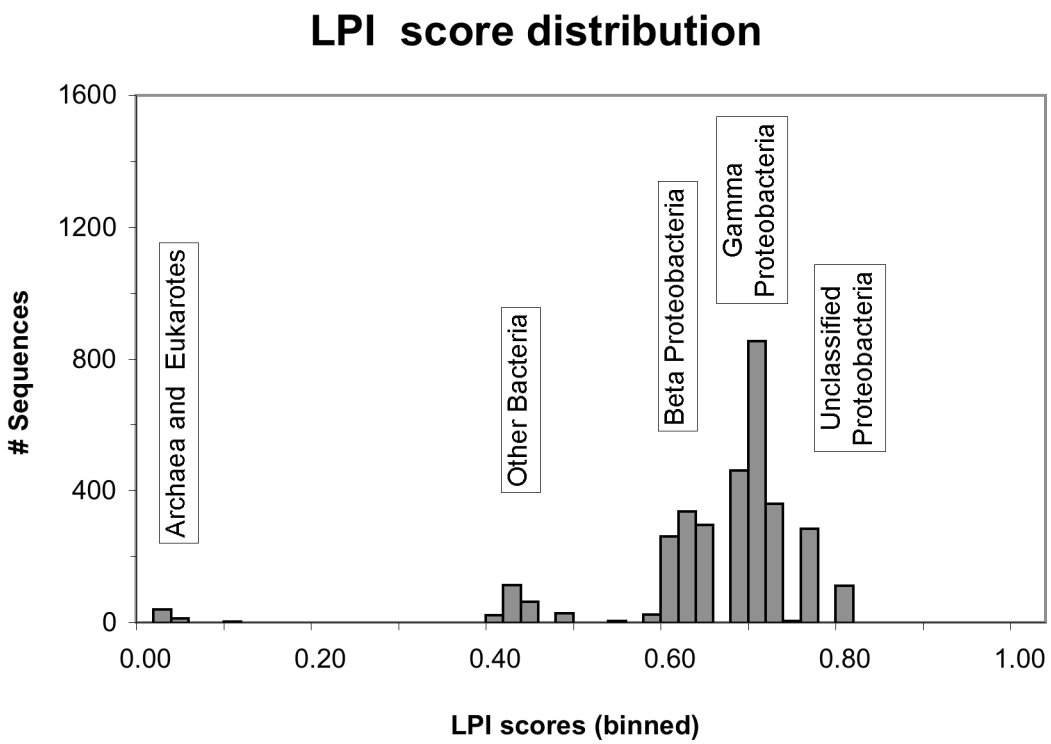


Figure 3.7 Lineage Probability Index (LPI) score distribution for the *E. persephone* metagenome. Higher LPI values indicate ORFs that have "expected" phylogenies, meaning phylogenies that place that ORF into a clade of ORFs that is similar to the clades from the remainder of the ORF phylogenies in the *E. persephone* metagenome.

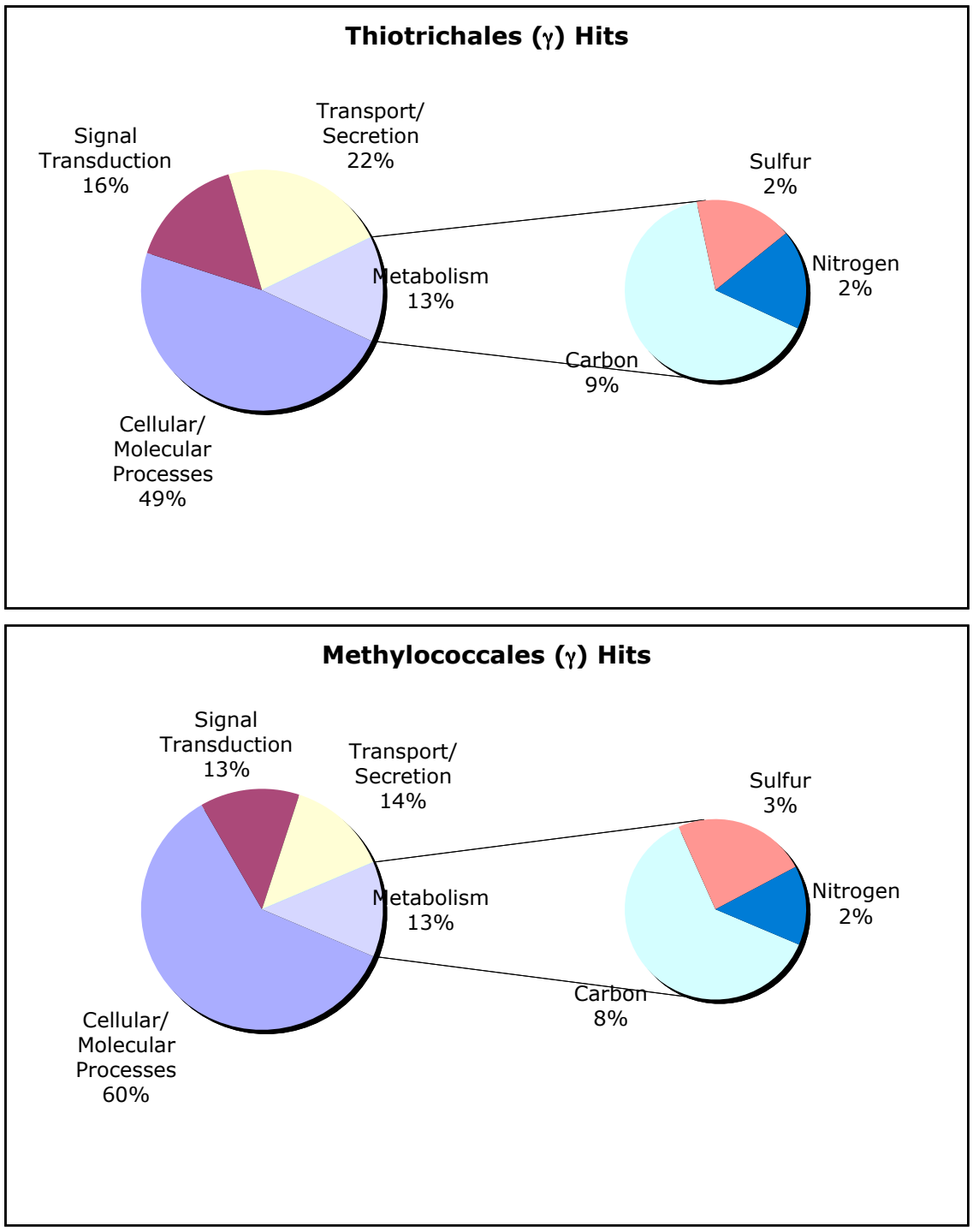


Figure 3.8 ORF "best hits" representative of gamma proteobacteria, allocated to each order and categorized by function. Titles of the graphs state the order that the graph represents. All such graphs made for gamma proteobacteria showed similar profiles, dominated by cellular/molecular processes with little similarity with respect to carbon, nitrogen and sulfur metabolism.

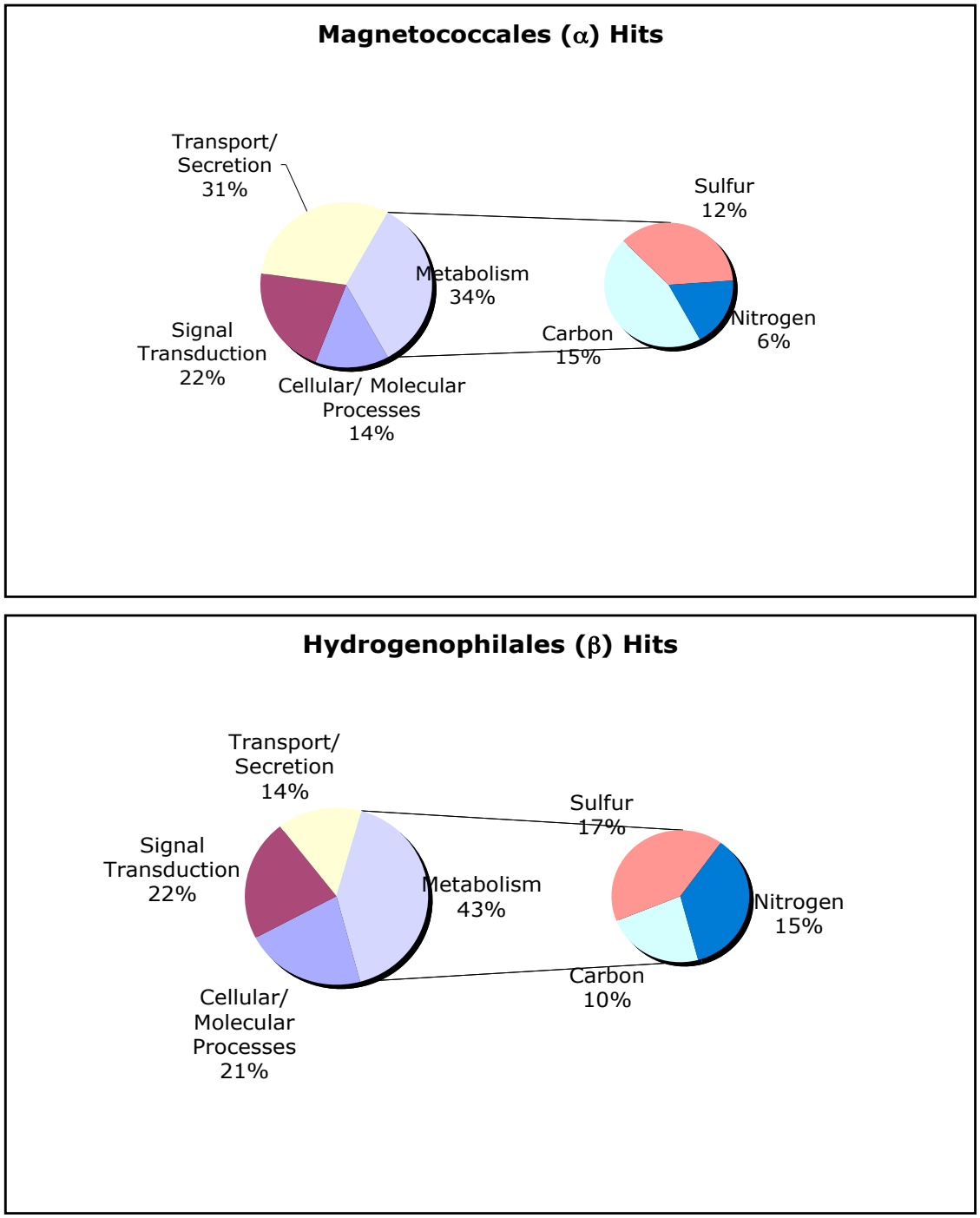


Figure 3.9 ORF "best hits" representative of non-gamma proteobacteria, allocated to each order and categorized by function. Titles of the graphs state the order that the graph represents. All such graphs made for non-gamma proteobacteria showed similar profiles, with relatively modest homology to cellular/molecular processes and a large contribution of ORFs involved in Carbon, Nitrogen and Sulfur metabolism.

ACKNOWLEDGMENTS

Thank you to Craig Young for the generous donation of time at sea (and Horst Felbeck for the invitation), and the captain and crew of the R/V Atlantis and DSR/V Alvin for collections and help during cruises. SymBio Corp. and Amersham Biosciences performed the sequencing and The Scripps Genome Center provided analyses. Sheila Podell was very helpful in discussions and analysis. Eric Allen guided the multilocus sequence analysis and I am indebted to Mark Hildebrand for the time he dedicated to advise me with molecular techniques. DNA sequencing (aside from the metagenome sequencing) was performed by the DNA Sequencing Shared Resource, UCSD Cancer Center, which is funded in part by NCI Cancer Center Support Grant #2 P30CA23100-18. Thank you to Doug Bartlett and Lena Palekar for the *Halobacterium* sp. sample.

REFERENCES

Abe, T., S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki and T. Ikemura (2003). Informatics for unveiling hidden genome signatures. *Genome Research* **13**: 693-702.

Acinas, S.G., R. Sarma-Rupavtarm, V. Klepac-Ceraj and M. Polz (2005). PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology* **71**: 8966-8969.

Amann, R. and W. Ludwig (2000). Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. *FEMS Microbiology Reviews* **24**: 555-565.

Auman, A.J., S. Stolyar, A.M. Costello, and M.E. Lidstrom (2000). Molecular characterization of methanotrophic isolates from freshwater lake sediment. *Applied and Environmental Microbiology* **66**: 5259-5266.

Baker, G.C., J.J. Smith and D.A. Cowan (2003). Review and re-analysis of domain-specific 16S primers. Journal of Microbiological Methods **55**: 541-555.

Benson, D.A., D.J. Lipman, I. Karsch-Mizrachi, J. Ostell, B.A. Rapp and D.L. Wheeler (2000). GenBank. *Nucleic Acids Research* **28**: 15-18.

Bright, M. and A. Sorgo (2003). Ultrastructural reinvestigation of the trophosome in adults of *Riftia pachyptila* (Annelida, Siboglinidae). *Invertebrate Biology* **122**: 345-366.

Cary, S. C., W. Warren, E. Anderson and S. J. Giovannoni (1993). Identification and localization of bacterial endosymbionts in hydrothermal vent taxa with symbiont-specific polymerase chain reaction amplification and *in situ* hybridization techniques. *Molecular Marine Biology Biotechnology* **2**(1): 51-62.

Childress, J.J., C.R. Fisher, J.M. Brooks, M.C. Kennicutt, R. Bidigare and A.E. Anderson (1986). A methanotrophic marine molluscan symbiosis: mussels fueled by gas. *Science* **233**: 1306-1308.

Di Meo, C.A., Wilbur, A.E., W.E. Holben, R.A. Felman, R.C. Vrijenhoek and S.C. Cary (2000). Genetic variation among endosymbionts of widely distributed vestimentiferan tubeworms. *Applied and Environmental Microbiology* **66**: 651-658.

Dubilier, N., R. Amann, C. Erseus, G. Muyzer, S. Park, O. Giere and C.M. Cavanaugh (1999). Phylogenetic diversity of bacterial symbionts in the gutless marine oligochaete *Olavius loisae* (Annelida). *Marine Ecology Progress Series* **178**: 271-280.

Edwards, D.B. and D.C. Nelson (1991). DNA-DNA solution hybridization studies of the bacterial symbionts of hydrothermal vent tube worms (*Riftia pachyptila* and *Tevnia jerichonana*). *Applied and Environmental Microbiology* **57**: 1082-1088.

Feil E.J. (2004). Small change: Keeping pace with microevolution. *Nature Reviews Microbiology* **2**: 483-495.

Feldman, R. A. and M. B. Black (1997). Molecular phylogenetics of bacterial endosymbionts and their vestimentiferan hosts. *Molecular Marine Biology Biotechnology* **6**: 268-277.

Fisher, C.R. and J.J. Childress (1984). Substrate oxidation by the trophosome tissue from *Riftia pachyptila* Jones (Phylum Pogonophora). *Marine Biology Letters* **5**: 171-183.

Gevers, D., F.M. Cohan, J.G. Lawrence, B.G. Spratt, T. Coenye, E.J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F.L. Thompson and J. Swings (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology* **3**: 733-739.

Hugenholtz, P., B.M. Goebel and N.R. Pace (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology* **180**: 4765-4774.

Jennings R.M. and K.M. Halanych (2005). Mitochondrial genomes of *Clymenella torquata* (Malanidae) and *Riftia pachyptila* (Siboglinidae): Evidence for conserved gene order in Annelida. *Molecular Biology and Evolution* **22**: 210-222.

Nelson, D., J.B. Waterbury and H.W. Jannasch (1984). DNA base composition and genome size of the prokaryotic symbiont of *Riftia pachyptila* (Pogonophora). *FEMS Microbiology Letters* **24**: 267-271.

Nussbaumer, A.D. and M. Bright, unpublished observations.

Nussbaumer, A.D., C.R. Fisher and M. Bright (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* **441**: 345-348.

Podell, S. and T. Gaasterland, submitted. DarkHorse: A method for genome-wide prediction of horizontal gene transfer.

Sambrook, J. and D.W. Russel (2001). Molecular Cloning A laboratory manual. Cold Spring Harbor Press: New York, USA.

Suzuki, M.T. and S.J. Giovannoni (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology* **62**: 625-630.

Tyson, G.W., J. Chapman, P. Hugenholtz, E.A. Allen, R.J. Ram, P. M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar and J.F. Banfield (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.

Visick, K.L., J. Foster, J. Doino, M.J. McFall-Ngai and E.G. Ruby (2000). *Vibrio fischeri lux* genes play and important role in colonization and development of the host light organ. *Journal of Bacteriology* **182**: 4578-4586.

von Wintzingerode, F., U.B. Gobel, E. Stackebrandt (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews* **21**: 213-229.

Weisburg, W.G., S.M. Barns, D.A. Pelletier and D.J. Lane (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology* **73**: 697-703.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews* 51: 222-227.

Yanai, I., C.J. Camacho and C. DeLisi (2000). Predictions of gene family distributions in microbial genomes: Evolution by gene duplication and modification. *Physical Review Letters* **85**: 2641-2644.

CHAPTER 4

The Metabolic Versatility of *Endoriftia persephone* Revealed Through Metagenomics

ABSTRACT

The *Endoriftia persephone* metagenome corroborates experimental findings regarding the sulfur, carbon and nitrogen metabolism of the *Riftia pachyptila* symbiont. The genome indicates that the symbionts can use the reverse TCA cycle in addition to the Calvin-Benson Cycle for carbon fixation, and contains an unusual ATP citrate lyase as the key enzyme of this cycle. This possibility may explain the discrepancy between the carbon isotope values sampled from the tubeworm host and those expected based on the Calvin-Benson Cycle as the single mode of carbon fixation. Additionally, the role of the symbiont while in the free-living stage was investigated. The presence of all genes necessary for heterotrophic metabolism, a phosphotransferase system, and tripartite ATP-independent periplasmic (TRAP) type and ABC transporters indicate that the symbiont can live mixotrophically while in the hydrothermal realm. Surprisingly, the symbiont contains many genes necessary for using the host as a source of organic carbon, including hemolysins, Type III secretion systems, chitinases and collagenases. The metagenome has a large suite of signal transduction, defense (both biological and environmental) and chemotaxis mechanisms as well as secretion systems. The physiology of *E. persephone* is explored with respect to functionality while associated vs. free-living.

INTRODUCTION

The discovery of chemolithoautotrophic symbioses began with detection of ribulose bisphosphate carboxylase/oxygenase (RuBisCO) in the trophosome of the mouthless, gutless hydrothermal vent tubeworm *Riftia pachyptila* (Felbeck 1981). Since this finding, the carbon fixation and energy generation mechanisms of chemosynthetic symbioses has been the subject of various investigations and a wealth of information is now available with respect to *Riftia pachyptila* and *Endoriftia persephone* (Childress 1992). The activities of various enzymes have confirmed the pathways of energy generation and carbon fixation used by the symbionts. Sulfide oxidation is achieved *via* reverse sulfate reduction and the Calvin-Benson Cycle is responsible for carbon fixation (Childress 1988). Both oxygen and nitrate can serve as electron acceptors (Hentschel and Felbeck 1993), and it is thought that intracellular sulfur granules can serve as electron sinks under anoxic conditions (Arndt *et al.* 2001).

It is known that the symbionts use CO_2 as their inorganic carbon source and that they are not carbon limited (Childress *et al.* 1993). Large stores of bound sulfide are maintained in the vascular and coelomic fluids to maintain the symbionts' energy production when oxygenated sea water dominates (Childress *et al.* 1984; Arp and Childress 1983). Net autotrophy persists for more than 5 hours of exposure to sulfide-free sea water (Girguis and Childress 2006). Net autotrophy has been measured over the range of sulfide and oxygen concentrations found *in situ* in *Riftia* clumps, and the worms can be maintained for several months in good health in experimental aquaria (Girguis and Childress 2006; Childress pers. comm.).

Despite the sophisticated equipment developed to study this symbiosis and the advancements in knowledge of its physiology, many questions remain unanswered. One subject that has led to much speculation on carbon fixation pathways is that of the stable carbon isotope ratio of the association, which is consistently approximately -11‰. It has been shown that carbon fixation occurs *via* the Calvin-Benson Cycle, which generates much more

depleted carbon isotope values (ranging from –23 to -50‰). Many theories have been posited to explain this discrepancy (see Chapter 1), but experimental evidence exists to call each into question.

Another major query involves the versatility of the symbiont. Several studies support the theory that the symbiont has a free-living stage in the hydrothermal environment (Cary *et al.* 1993; Millikan *et al.* 1999; Nussbaumer *et al.* 2006). It has been detected using PCR and *in situ* hybridization on vastly different substrates, including *Riftia* tubes and distant basalt (Cavanaugh *et al.* 2005). The symbiont must therefore be highly adaptable to its environment. It is unknown whether it changes its physiology as a response to quorum sensing, like many symbionts and pathogens (Nyholm and McFall-Ngai 2004; Parsek and Greenberg 2000). There is similarly no information about the symbiont's metabolism in the free-living stage. Since they react with each other abiotically, access to both sulfide and an oxygen or nitrate is virtually impossible for a single cell without specialized structures (Larkin and Strohl 1983; McHatton *et al.* 1996). Remaining questions include: Does the symbiont continue to oxidize sulfide though the substrates needed have significant spatiotemporal gradients in the diffuse flow region (Le Bris *et al.* 2006)? How does the symbiont tolerate the high concentrations of heavy metals in the active flow region (Johnson *et al.* 1986)?

Sequencing the symbiont genome may answer some of these questions and certainly will lead to creation of more developed hypotheses. Since the symbiont is not yet in culture, the genome must be sequenced from the environment. This is a special case in which the tubeworm host maintains an essentially pure symbiont population within its trophosome, and the bacteria within grow to very high densities. Additionally, developed purification methods allow separation of host and symbiont tissue (Distel and Felbeck 1988), which can serve as samples for DNA extraction. The *Endoriftia persephone* metagenome will almost certainly prove valuable for all investigators studying hydrothermal vent biology and chemosynthesis in general.

MATERIALS AND METHODS

The *Riftia* symbiont metagenome.

Methods used in sequencing, assembly and annotation of the metagenome have been outlined elsewhere (Chapter 2).

Clusters of orthologous groups (COG) comparison.

The metagenome was compared with the COG database from the National Center for Biotechnology Information (NCBI), which consists of recurring sequences from specific genes that occur in sequenced genomes. The specific genes are sorted into functional categories. The output was imported into Excel and sorted by functional category. Percent composition of each functional group relative to the total number of COG hits was calculated for each organism involved in the analysis. These organisms include *Thiomicrospira crunogena* and *Methylococcus capsulatus* as phylogenetically similar organisms; the facultative pathogens *Yersinia pestis, Pseudomonas aeruginosa, Salmonella typhimurium,* and *Vibrio cholerae*; and the facultative plant symbiont species *Mesorhizobium loti, Sinorhizobium meliloti,* and *Agrobacterium tumefaciens*; and the bobtail squid symbiont *Vibrio fischeri.* COG comparisons were performed by calculating the contribution of each functional category and normalizing them to the *E. persephone* percentage.

RESULTS

A table of some enzymes identified in the metagenome is provided [Table 4.1]. In addition to the key enzymes of the Calvin Cycle, all enzymes necessary for both the oxidative and reductive TCA cycles are present, as are enzymes necessary for glycolysis, gluconeogenesis, glycogen synthesis and glycogenolysis. The metagenome information also verified the use of reverse sulfate reduction to oxidize sulfide and several rhodanese types and polysulfide reductases are present, as is a sulfur reductase. A dissimilatory nitrate reductase and several carbon, sulfur and nitrogen response systems were revealed.

DISCUSSION

The genome.

The sequences were manually curated and filtered to dispose of any possible contaminating host sequence, and very little contamination of any kind was encountered (Chapter 2 and 3). In a graph of percent guanine + cytosine (%GC) for all of the ORFs, there is a sharp peak at 60% [Figure 2.7], which is the expected GC content of the symbiont (Nelson *et al.* 1986). There is a sharp drop beyond 60%, while there is some noise at the lower end of the peak. This noise is due to the presence of horizontally transferred genetic material, with an average of 45% GC composition. The largest low GC contig sequence does not appear to be present in the host DNA (see Chapter 2) though it is present in the symbiont, and it is the second largest contig in the metagenome. It may therefore be part of the plasmid found in previous work (Nelson *et al.* 1986) since its possible presence in high copy number relative to the main chromosome would have resulted in its sequence overrepresentation, and therefore a large representative contig due to extensive sequence coverage. Interestingly, this contig contains important housekeeping genes including a chromosomal replication initiator protein (*dna*A), DNA polymerase III, beta chain (*dna*N), and DNA gyrase, beta subunit (*gyr*B), and a transposase. 18% of the DNA that makes up the longest 25 contigs is of low GC (< 50% average over the length of the contig), supporting the possibility of an extrachromosomal element that would be present in high copy number in the metagenome.

In the metagenome as a whole, few phage sequences were encountered in the annotations. Using the program DarkHorse (Podell and Gaasterland, submitted) confirmed that there is very little (< 0.1%) recognizable phage sequence in the *E. persephone* genome [Figure 3.7]. This may be an artifact, as phages from the hydrothermal environment have not been sequenced extensively. Comparison with the only available hydrothermal phage dataset

did not reveal any homologs with the symbiont genome (data not shown, Shellie Bench, Shannon Williamson and Eric Wommack, personal communication).

One gene shares 100% amino acid (>99% nucleotide) identity with a 621 basepair stretch of an ABC-type multidrug transporter, and another with 594 basepairs of a pantothenate vitamin synthetase from the bobtail squid symbiont *Vibrio fischeri*. A third gene is 95% identical to an N-acetylmannosamine-6-phosphate 2-epimerase (ManNAc-6-P) from *Streptococcus* sp., which is a precursor for sialic acid biosynthesis. Sialic acid is used by pathogens and symbionts to coat the cell surface in order to evade host immunity (Schauer 2000). Interestingly, the ManNAc-6-P epimerase is much more homologous to the pathogen types than to the symbiont types. The vast majority of genes are 30-70% identical at the amino acid level, to GenBank's nonredundant database (Benson *et al.* 2000), and only 5 of these genes are >95% identical. The high level of conservation to the *V. fischeri*-like genes may be a reflection of the necessity of these genes in these symbioses, the specific functions of which are unknown. All three of these ORFs have 35-42% GC content.

Archaeal genes make up approximately 0.7% and 1.1% of the genome analyzed by DarkHorse [Figure 3.7] and top BLAST hit, respectively. These genes encode proteins with a variety of functions and do not fall into a single functional category.

Physiology.

The stoichiometry of the reverse TCA (rTCA) cycle and the Calvin-Benson Cycle differ dramatically. Per cycle, the rTCA cycle fixes 2 CO_2 [Figure 4.1] with the use of 2 ATP, while the Calvin-Benson Cycle requires 18 ATP to fix 6 CO_2. Strictly based on cycling, therefore, the rTCA cycle is more energetically efficient (using 1 ATP vs. 3 ATP per carbon assimilated). Moreover, the enzyme RuBisCO of the Calvin-Benson Cycle is intrinsically inefficient, as the specificity for CO_2 is low; the oxygenase activity of RuBisCO substitutes O_2 for CO_2. Type II RuBisCO enzymes like that found in the symbiont have especially unspecific substrate binding affinities (Tabita 1988). The high levels of CO_2 in the trophosome decrease the degree to

which this mistaken substrate, oxygen, is utilized (Robinson and Cavanaugh 2003). The inherent inefficiency of this enzyme leads to speculation that it is the most abundant protein on Earth, since the organisms that use it keep it in abundance in the cell in order to increase carbon fixation activity (Tabita 1988). Likewise, some enzymes of the rTCA cycle are oxygen-sensitive (Hugler *et al.* 2003), but function under microaerophilic conditions. The TCA cycle is bidirectional, while the Calvin-Benson Cycle is irreversible. In contrast to their inefficiencies, both cycles' intermediates are regenerative (they are regenerated rather than exiting the pathway), ensuring the continued activity of each carbon fixation pathway.

Based on the comparative energetic efficiencies of the two carbon fixation cycles, it follows that the reductive TCA Cycle is favored when redox substrates are low and the symbiont must conserve energy. This theory is supported by comparative proteomics of sulfide-rich and sulfide-depleted trophosome tissue. In high sulfide conditions RuBisCO was present at four times higher concentrations than in sulfide-depleted trophosome, where rTCA cycle enzymes were increased up to twenty-four fold (Markert *et al.*, *in press*).

This theory is consistent with respect to a recent study of Girguis and Childress (2006), where net autotrophy was measured in *Riftia* specimens incubated in a wide variety of environmentally relevant sulfide and oxygen concentrations. It was shown that there is a small, consistent net increase in CO_2 uptake (to 5 μmol g wet weight^{-1} h^{-1}) over the range of 0 to 70 μM sulfide and 60 to 150 μM oxygen, but that when sulfide is increased to 111 μM (which approximates the highest recorded concentration in diffuse flow areas), there is a sharp increase in net CO_2 uptake, to 17 μmol g wet weight^{-1} h^{-1}. It may be that the two carbon fixation processes are alternating activities under each condition. Under low energy conditions, the symbionts are using a pathway with a lower overall carbon fixation rate. Alternatively, when sulfide concentrations are high and the energy yield *via* its oxidation is high, the symbionts can use the less efficient Calvin-Benson Cycle. If this is the case, it is curious that the inorganic carbon uptake is increased three-fold with use of the Calvin-Benson Cycle, considering the inefficiencies of RuBisCO. It is likely that a lower oxygen tension in the

presence of sulfide results in increased carboxylase activity of RuBisCO. The carbon turnover rate of the two cycles likely depends on the conditions, but it is entirely possible that despite RuBisCO's inefficiencies, the Calvin-Benson Cycle produces more organic carbon than the rTCA cycle when oxygen concentrations are low. This remains to be tested.

The inescapable ensuing question is: why does the symbiont use the Calvin-Benson Cycle at all? If another, more efficient mode of carbon fixation is available there should be no reason for the alternate cycle. I propose three possible scenarios: the enzyme-centric, the bacteriocentric and the host-centric views. The first is the enzyme-centric view. It is possible that conditions within the internal milieu of the bacterial cell cannot be replicated in the lab, and that under sulfidic conditions within the cell, the activity of RuBisCO increases. Increased activity in the presence of sulfide might be contributed to the low concentrations of oxygen that result. In lower oxygen conditions, carboxylation increases while oxygenation decreases. The enzyme-centric scenario is not likely to be the explanation, as proteomic studies indicate that the concentration of the enzyme rather than its activity is increased under high sulfur conditions.

The bacteriocentric view concerns the fact that the symbiont is thought to grow in size while in the trophosome to approximately 5 μm. Growth is then essentially arrested and dividing cells are rarely encountered (Bright and Sorgo 2003). In evolutionary terms, the "purpose" of a bacterium is to grow and replicate. In the trophosome, therefore, the bacteria's carbon requirements are low; the primary purpose of carbon fixation in this system is to provide for the host *via* translocation of organic carbon. Under these conditions, any energy gained is supplementary. If it is not used for biosynthesis or replication, it is superfluous from the bacteria's perspective. Under high sulfide conditions the symbiont can be wasteful and use the Calvin-Benson Cycle. If the cycle is of no benefit to the symbiont, then why does it remain part of its physiology? This question is beyond the scope of this thesis, but it is possible that without the increased organic carbon output that results from high sulfide

conditions, amount of organic carbon translocated by the symbionts is insufficient to sustain the host, resulting in death.

The third proposed situation is where the host has some control over the symbiont's metabolism. When sulfide is prevalent, the host and symbiont may be competing for limited oxygen as a terminal electron acceptor. The host may release a chemical cue as a result of hypoxia that is sensed by the large suite of signal transduction pathways in the symbiont. Such a cue would alert the symbiont that the chemical environment surrounding the bacteriocytes might become anoxic due to host respiration. The response regulator(s) in this pathway begin to transcribe nitrate reductase and the enzymes for the Calvin-Benson Cycle, which is more efficient in the absence of oxygen.

Tubeworm wet weight correlates well with isotopic signature; smaller worms (with higher surface: volume ratios for substrate uptake) have more negative carbon isotopic values (range from $\delta^{13}C$ of -9‰ in large vs. -16‰ in small organisms; Fisher *et al.* 1990). It is hypothesized that this correlation occurs because the symbionts in larger worms are carbon-limited, so their enzymes cannot discriminate against the heavier isotope because all the inorganic carbon available is fixed. In light of our new knowledge of the rTCA cycle in these organisms, we can now explain the carbon isotopic signature of -11‰ on average, which reflects the rTCA cycle enzymes' decreased discrimination (Londry *et al*. 2004). Why then, does it correlate with size? Based on the same theory that RuBisCO is used when redox substrates are abundant one can argue that the larger surface:volume ratio of juvenile worms allows the symbionts and host better access to available substrates from the environment. Since the symbionts have higher concentrations of redox substrates, they can fuel the energetically inefficient Calvin-Benson Cycle. These smaller worms are also presumably closer to the seafloor, which is the source of sulfide in diffuse flow environments. RuBisCO form II discriminates at approximately $\delta^{13}C$ of –19 to -23‰, while the smallest worms have isotope ratios of -15‰ (Fisher *et al.* 1990), so dependence on RuBisCO for organic carbon would not be obligate in smaller worms.

The rTCA cycle found in *E. persephone* differs from those described to date (Fuchs *et al*. 1980; Hugler *et al*. 2003). The first difference is that the cycle found in the symbiont begins with oxaloacetate formation *via* carboxylation of pyruvate rather than phosphoenolpyruvate [Figure 4.1]. The absence of a gene encoding the PEP carboxylase in the genome is substantiated by experimental assays, where activity of the enzyme is undetectable. Pyruvate carboxylase activity has been detected (Felbeck 1985). Moreover, another difference in the symbiont's rTCA cycle is the ATP citrate lyase.

The ATP citrate lyase (or ACL) was not identified in annotations of any of the versions of the metagenome. An ORF that is annotated as a succinyl CoA synthetase has low homology to eukaryotic ATP citrate lyases and is thought to be the gene coding for the symbiont's ACL. ATP citrate lyases were formed by gene duplication, diversification and fusion of the succinyl CoA synthetase (SCS), alpha subunit with citrate synthase (Fatland *et al.* 2002) [Figure 4.2]. The product is the beta subunit of ACL. The ACL alpha subunit, on the other hand, evolved from duplication and diversification of the succinyl CoA synthetase beta subunit. In most eukaryotes, the alpha and beta ACL subunits are fused into a single gene, but in prokaryotes using the rTCA cycle the two subunits remain separate (Fatland *et al.* 2002). Supporting the *E. persephone* SCS-type enzyme's identification as an ACL, a second gene with higher homology to a succinyl CoA synthetase (both alpha and beta chains) has been identified in the genome. Furthermore, *Magnetococcus* MC-1 has been found to have ACL activity and to depend on the rTCA cycle for carbon fixation, but the gene responsible for the citrate lyase activity was not identified in the genome annotation (Williams *et al.* 2006). The ACL candidate from *E. persephone* is 87% identical to a *Magnetococcus* MC-1 gene, and the MC-1 genome also contains a second, more likely succinyl CoA synthetase. Furthermore, ACL activity has been detected in symbiont extracts (Markert *et al. in press*) and the enzyme has been identified in proteomic gels. It therefore appears that either the ACL in *E. persephone* arose from a recent gene duplication event, or that the functional plasticity of the enzyme is very limited, resulting in very little divergence. Homologous enzymes are found in

Geobacter sulfurreducens, *Paracoccus denitrificans* and in other archaea [Figure 4.3]. As in other carbon fixation enzymes, it appears that this gene is widespread among different lineages within the prokaryotes (Hugler *et al.* 2003).

E. persephone has several TRAP-type C4-dicarboxylate transporters for the uptake of 4-carbon molecules such as malate and succinate. Phosphoenolpyruvate (PEP) carboxylase is an enzyme of the rTCA cycle, but has not yet been found in the genome nor detected by assays (Felbeck 1985), so it is likely absent. Carboxylation in the symbiosis may sometimes function as it does in C4 plants. Experiments with live *Riftia* show that a fraction of the label from radiolabelled inorganic carbon is incorporated into malate and succinate within the plume of the worm (Felbeck *et al.* 2004). These are also the two immediate products of radiolabelled carbon incubations with symbionts (Felbeck 1985). Malate and succinate are two intermediates of the rTCA cycle that can be transported by the TRAP-type dicarboxylate transporter to the symbiont. There they may enter the reverse TCA cycle anaplerotically, where subsequent carboxylation steps result in the majority of carbon fixation occurring in the trophosome tissue. It is therefore possible that at times carbon fixation is a combined effort between host and symbiont. Alternatively, a pyruvate carboxylase is present in the symbiont genome and its activity can substitute for the production of oxaloacetate for the rTCA cycle. The activity of the pyruvate carboxylase has been shown (Felbeck 1985). Both *Riftia* and the symbiont contain carbonic anhydrase for concentrating carbon dioxide (De Cian *et al*. 2003).

Researchers are now beginning to realize the significance of the reverse TCA cycle as a form of carbon fixation in prokaryotes. These studies have led to the discovery of a large number of fairly conserved ACLs particularly from the hydrothermal environment (Campbell and Cary 2004). The degree of diversity of ACLs requires further investigation prior to concluding that these enzymes are highly conserved among organisms. Campbell and Cary (2004) also found that ACL genes and 2-oxoglutarate oxidoreductase (OOR) genes are ubiquitous in hydrothermal samples, when detected by PCR, but RuBisCO was only weakly detected in two of the six samples tested. It is therefore not surprising that this cycle is also

present in *E. persephone*, a bacterium that spends much of its life history producing organic carbon for its host. The availability of both cycles ensures flexibility and optimization of organic carbon output in variable conditions.

A wide range of genes in the metagenome indicates that the symbiont can survive as a heterotroph and respond catabolically to various external carbon sources [Table 4.1, 4.2] including sugars and small organics, but not amino acids. Heterotrophic metabolism has been measured in symbiont purifications (Felbeck, personal communication), but little is known about this type of metabolism in the symbionts. The importance of heterotrophy in the symbiont metabolism is apparent by the large number of regulatory systems that respond to carbon compounds [see Table 4.2 for some examples]. Based on the vast array of genes involved in chemotaxis and the ability to respond to external carbon sources *via* catabolite regulation and the phosphotransferase (PTS) system, it is likely that *E. persephone* survives as a heterotroph while in the hydrothermal environment. Organic carbon is present at hydrothermal vents in concentrations that are sufficient to sustain many groups of hydrothermal microbial heterotrophs (Comita *et al*. 1984; Wirsen *et al.* 1993; Kormas *et al*. 2006, Jeanthon 2000, Ruby *et al*. 1981). When in the free-living form, the symbiont no longer has unlimited access to substrates for sulfide oxidation, which are rarely encountered concurrently. It is of benefit then, to possess this alternate mode of carbon and energy acquisition. Microbial cultures from hydrothermal vents are dominated by mixotrophs (Jeanthon 2000). This is likely biased due to the difficulties in culturing obligate chemolithotrophs, but in theory this is to be expected. Organic carbon is the most energetically favorable substrate available, and the symbiont would gain more energy by the use of heterotrophy over sulfide oxidation. Furthermore, while in their free-living stage, the symbiont does not have to expend energy creating organic carbon to fuel the astounding growth rate of its invertebrate host (Lutz *et al.* 1994). It can access both the carbon and energy it needs from organic carbon. It has recently been discovered that the symbiont is present in biofilms on the worm tubes and cuticle (Nussbaumer *et al.* 2006, Cavanaugh 2005).

In this environment, *E. persephone* can likely access substrates for heterotrophic metabolism and protect itself from toxic metals within the polysaccharide matrix. It is likely that outside of the trophosome the concentration of these metals is much higher. *E. persephone* has several metal efflux pumps to maintain function in the presence of such toxins. Outside of the host, encounters with oxidative radicals are not likely, and the symbiont is not required to pump protons to maintain alkaline pH to the extent that it does in the host's coelomic fluid (Goffredi *et al.* 1997) . This symbiosis is often addressed as a mutual benefit: *Riftia* can maintain extremely high growth rates by gaining organic carbon from the symbiont, and the symbiont has access to substrates that may otherwise be inaccessible and is maintained in a stable environment. It does not appear that *E. persephone* is at such an advantage, however, if it expends the majority of its energy to produce and excrete organic carbon for the host's metabolism. If the alternative is increased energetic gain by heterotrophy in an environment where that energy can be used for cell division, the more favorable environment might be outside of the trophosome.

Heterotrophy and signal transduction may also be involved in the symbiont physiology after the host perishes. It is currently assumed that *E. persephone* is abundant in the vent environment and symbionts are recruited from a new cohort every generation. Presumably, the symbionts associated with a host can no longer survive when their host dies. With the new knowledge of symbiont heterotrophy, it is plausible that the symbiont responds to a particular trigger after the host's death, at which time it must change physiology in order to subsist in the changing environment. Several genes from the genome may serve the symbiont in this mode, including hemolysins, signal transduction and chemotaxis mechanisms, flagella, and oxidative TCA cycle genes. The *Riftia* host is a substantial source of organic carbon and is likely consumed by heterotrophic microbes from the external environment once it dies. Why then, couldn't it serve as a carbon source for the symbiont? The symbiont has been detected on the cuticle and tube of the worm (Cavanaugh *et al.* 2005;

Nussbaumer *et al*. 2006) and the genome contains chitinases and collagenases that may function in

Sulfur, nitrogen and phosphate metabolism.

The *Riftia* symbiont is predicted to perform sulfide oxidation *via* a reverse sulfate reduction pathway, involving the enzymes APS reductase and ATP sulfurylase (Felbeck 1981) and both these enzymes are present in the metagenome. Several rhodanese types are also present, therefore the use of thiosulfate by the symbiont warrants further investigation. A sulfur reductase allows the use of elemental sulfur as an electron sink when oxygen is absent (Arndt *et al.* 2001).

The metagenome contains a dissimilatory nitrate reductase (Hentschel and Felbeck 1993), in order to respire nitrate as an alternate electron acceptor. It also has the ability to sense and respond to nitrite and nitrate, and to synthesize glutamate, glutamine and polyamines. Several phosphate ABC transporters are present, as well as phosphate sensors and starvation response genes.

Colonization.

It has been shown recently that the symbiont colonizes *Riftia* through the body wall of the juvenile tubeworm (Nussbaumer *et al.* 2006). Some host cells undergo apoptosis during this process, and it is unknown whether the host or the symbiont is responsible. Seven hemolysins and six metalloendopeptidases were found in the genome and may be involved in initiation of the symbiosis. Five LysR-type transcriptional regulators were also sequenced. Contig 2388 contains virulence-type genes possibly involved in infection, encoding a LysR regulator, a membrane hemolysin III homolog, a putative metalloendopeptidase, and a TolC-type outer membrane transporter. As expected from their predicted infection-associated status, these proteins were not detected in recent proteomics work with purified symbionts from an adult specimen (Markert *et al. in press*). A Type II secretion system is also present [Table 4.1].

Comparative genomics.

Clusters of Orthologous Genes (COGs) are used to classify genomic components by functional category. The COG database contains signature sequences of specific genes, and can be compared to genomes using BLAST. The symbiont's COG profiles were compared with *E. persephone*'s closest sequenced relatives *Thiomicrospira crunogena* (a sulfide oxidizing hydrothermal bacterium), *Methylococcus capsulatus* (a methanotrophic bacterium from reducing environments); the facultative pathogens *Yersinia pestis, Pseudomonas aeruginosa, Salmonella typhimurium,* and *Vibrio cholerae*; to the facultative plant symbiont species *Mesorhizobium loti, Sinorhizobium meliloti,* and *Agrobacterium tumefaciens*; and to the facultative bobtail squid symbiont *Vibrio fischeri*. Compared to these three groups of organisms, *E. persephone* COG profiles showed an enhanced ability for energy production and conversion, signal transduction and (surprisingly) defense [Figures 4.4 to 4.9]. Enhanced dedication to energy production is expected. The symbionts must provide energy for their own productivity as well as their host's, in the form of organic carbon and as the sole source of the worm's carbon. Most ORFs within this category concern carbon and redox metabolism. The metagenome contains a large number of cytochromes to aid in electron transfer, and 7.3% of the "Energy production and conversion" COG category consists of H^+ ATPases. Inhibition of these pumps has been shown to result in a decrease of the pH of the host's coelomic and vascular fluids. The symbionts therefore contribute to the maintenance of an alkaline pH in the host's coelomic and vascular fluids using these ATPases (Goffredi *et al.* 1997). The majority of these are of the "Archaeal/Vacuolar Type." It is likely that this classification is due to the relative paucity of hydrothermal bacterial genomes in the database, and that the pumps are unique due to adaptations to the pressurized and acidic conditions at vents. The bias towards sequenced archaeal genomes from hydrothermal vents may be reflected in the pumps' classification.

The relatively high percentage of signal transduction mechanisms in the genome is also the case in *Vibrio cholerae*, an organism that must sense and respond to both the variable open ocean environment and the presumably more hospitable host environment as

well. An organism that lives in both situations should contain an extensive metabolic repertoire; the symbiont requires an increased ability to sense its environment and respond to it physiologically since the hydrothermal environment has extensive and variable thermal and chemical gradients and differs substantially from the trophosome environment. Of the signal transduction mechanisms for which specificity was determined (that weren't broadly classified as histidine kinase/response regulators, based on annotation), 29.8% were sugar response mechanisms involved in heterotrophic metabolism, including TRAP-type permeases, catabolite repression and sugar phosphotransferase systems. 51.7% are involved in chemotaxis, and 13.3% in nitrogen response. This indicates that the symbiont is exceptionally responsive to variable conditions while outside of the trophosome. It is likely that signal transduction mechanisms relating to response to the host are present but were not identified because fewer are characterized. The number of orthologs is low and therefore the respective COGs are scarce in the database.

The increased percentage of defense-associated genes relative to pathogens was unexpected, as pathogens are expected to be required to fight host immunity more frequently than symbionts. However, the number may reflect environmental defense. 47% of the genes that fall into this category for *E. persephone* are cation/multidrug efflux pumps and may actually function to export toxins present in the hydrothermal environment. Indeed, when compared to *Methylococcus capsulatus*, another inhabitant of reducing environments, the percentages are comparable. If these genes are excluded from the analysis, the percentage of defense mechanisms is equivalent to the pathogens. A percentage equivalent to that of pathogens is nonetheless curious. This dedication to defense processes in a symbiont may be because response to the initial symbiont infection process likely involves immune-type reactions in *Riftia,* as has been found with the *Sitophilus zeamais* weevil symbiosis and the *Euprymna scolopes*/*V. fischeri* symbiosis initiation processes (Anselme *et al.* 2006; McFall-Ngai, personal communication). The symbiont in this situation must protect itself from host defenses and has evolved pathogen-type defense mechanisms accordingly. The percentage

of the genome dedicated to defense in *V. fischeri* is in fact also equivalent to that of *E. persephone* [Figure 4.10]. This leaves the question of environmental vs. host-immunity defense unanswered, and it is likely a combination of both mechanisms that is required for *E. persephone*'s survival.

Compared to the gamma proteobacterial pathogens and plant symbionts used in this study, and to the human gut microbiome (Gill *et al*. 2006), *E. persephone* dedicates less of its genome to nucleotide biosynthesis and metabolism [Figures 4.5, 4.7, 4.9]. The symbiont is responsible for the *de novo* synthesis of pyrimidine nucleotides for the host (Minic *et al.* 2001), and likely increases synthesis while associated by upregulation of the nucleotide biosynthetic genes present in the genome. Metabolic biosynthesis and transportation is altogether decreased in *E. persephone* with respect to these organisms, but not when compared to the C1 organisms *Methylococcus capsulatus* (a methane oxidizer) and *Thiomicrospira crunogena* (an autotroph). This may indicate that autotrophy is relatively important in the symbiont's physiology.

E. persephone has less of its genome dedicated to secondary metabolite biosynthesis than all bacterial genomes analyzed in this study aside from *V. fischeri* and *M. capsulatus*, which might be indicative of its sole colonization of the trophosome. It has a relatively larger motility component to its genome indicating the need for motility in its free-living stages. 30% of genes in this category are involved in chemotaxis, likely to aid the symbiont in the search for suitable substrates. *V. cholerae* has a similar genomic composition dedicated to chemotaxis [Figure 4.6], and this organism has 42 methyl-accepting chemotaxis proteins (Dziejman *et al*. 2002), while 22 were identified in the *E. persephone* genome.

E. persephone's relatively decreased genomic dedication to metabolism and transport of amino acids, carbohydrates, nucleic acids and coenzymes is a peculiar finding. If the symbiont has two modes of carbon fixation and has a large suite of genes dedicated to energy production and conversion, how does it transport these metabolites to the host? Autoradiographic examination (Bright *et al.* 2000) and studies with purified symbionts (Felbeck

and Jarchow 1997) indicate that translocation does in fact occur, and that it is the main mechanism of carbon transport to the host. In depth transporter analyses are currently being performed by Ian Paulsen and Quinghu Ru at The Institute for Genomic Research. It is possible that the symbionts have a higher percentage of export transporters, but that this is not reflected in the total transporter and metabolism component. There are also a large number of secretory systems, which are within the motility category due to their use in flagellar/pili synthesis.

Future studies will examine the number of orthologs between the *Riftia* symbiont and some related invertebrate symbionts (of the bobtail squid *Euprymna scolopes* and the oligochaete *Olavius algarvensis*) and oceanic pathogens (*Vibrio cholerae* and *Vibrio vulnificus*). These comparisons will aid in examination of whether this heterotrophic bacterium, which has been shown to infect its host through the epidermis (unusual for a bacterium) (Nussbaumer *et al.* 2006), contains a larger component of pathogen-like or symbiont-like genes. This will provide insight into the transitions from pathogens to symbionts that have occurred in many symbioses (Anselme *et al.* 2006). Moreover, enumeration and identification of such orthologs between *E. persephone*, *Methylococcus capsulatus* and *Thiomicrospira crunogena* will reveal why the *E. persephone* genome is so much larger. Is there a certain suite of genes contributing to the larger size of the genome?

Analysis of the symbiont metagenome has solved the long debated paradox involving the host carbon isotope ratio by enlightening investigators on the dominant mode of carbon fixation in this organism. This metagenome has possibly revealed a new type of ATP citrate lyase involved in carbon fixation. Additionally, it has led to speculation into queries involving the symbiont physiology and interactions with its environment while associated and free-living. Resultant hypotheses can now be tested with the added advantage of knowledge of the symbiont's metabolic repertoire. The genome may not be closed, but in its present form it suffices for a physiological overview and in depth comparative analyses with related genera.

Table 4.1 Genes indicative of carbon metabolic pathways and possible translocation of organic compounds to the host. Open reading frames are categorized by host. Italicized ORFs are clustered with functional genes on the contig identified, but may not necessarily be involved in the same function.

TCA Cycle	
rif_0162_002	isocitrate dehydrogenase, NADP-dependent, monomeric-type
rif_2276_012	Malate dehydrogenase, NAD or NADP
rif_1135_001	Malate dehydrogenase, NAD-dependent
rif_0821_001	succinyl CoA ligase, beta subunit
rif_2004_006	Succinyl-CoA synthetase, alpha subunit
rif_1830_005	AconitaseA
rif_1087_001	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases
rif_1393_007	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide acyltransferase
rif_1393_001	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase
Reductive TCA	
rif_0726_002	2-oxoglutarate synthase subunit KorA (2-ketoglutarate oxidoreductase
rif_1727_001	2-oxoglutarate synthase subunit KorA (2-ketoglutarate oxidoreductase
rif_2123_002	succinate dehydrogenase/ fumarate reductase
rif_0260_002	putative pyruvate carboxylase/oxaloacetate decarboxylase, alpha subunit
rif_0677_004	ATP citrate lyase, large subunit
rif_0938_002	ATP citrate lyase, large subunit
Oxidative TCA	
rif_1673_002	2-oxoglutarate dehydrogenase E1 component
rif_1673_004	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide acyltransferase
rif_1673_007	Fumarase B (fumarate hydratase class I), anaerobic isozyme
rif_2258_001	Citrate synthase
rif_1368_001	2-oxoglutarate dehydrogenase (lipoamide) E1 component
Glycolysis/Gluconeogenesis	
rif_2227_005	3-phosphoglycerate kinase
rif_2227_001	Pyruvate kinase
rif_2227_009	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)
rif_1178_003	Fructose-bisphosphate aldolase, Tagatose-bisphosphate aldolase
rif_2454_004	Phosphoglyceromutase
rif_0663_002	Enolase (2-phosphoglycerate dehydratase)
rif_2345_001	6-phosphofructokinase
rif_2426_012	Glucose-6-phosphate isomerase
rif_3051_001	Triosephosphate isomerase
Calvin-Benson Cycle	
rif_0288_001	Ribulose,1-5,bisphosphate carboxylase oxygenase form II large subunit
rif_2240_002	Phosphoenolpyruvate synthase/pyruvate phosphate dikinase
rif_2217_001	Phosphoribulokinase (Phosphopentokinase) (PRKASE)
rif_1747_041	Pentose-5-phosphate-3-epimerase
rif_2227_001	Pyruvate kinase
rif_2227_005	3-phosphoglycerate kinase
rif_2227_010	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)
rif_2227_012	Transketolase
Translocation	
rif_1581_001	Type II secretory pathway, component PulF
rif_1581_003	Type II secretory pathway, prepilin signal peptidase PulO
rif_1651_003	Type II secretory pathway, component PulD
rif_1958_003	General secretion pathway protein, putative
rif_1587_001	Type II secretory pathway, component PulM
rif_1799_001	Type II secretory pathway, component ExeA
rif_2053_002	Type II secretory pathway, component PulD
rif_1979_003	Type II secretory pathway, component PulK
rif_1979_005	Type II secretory pathway, component PulJ
rif_1799_006	*Response regulator for gln (sensor glnL) (nitrogen regulator I)*
rif_2103_001	*Imidazoleglycerol-phosphate synthase*
rif_2103_004	*Phosphoribosyl-AMP cyclohydrolase*
rif_2103_005	*Phosphoribosyl-ATP pyrophosphohydrolase*
rif_2103_040	Sec-independent protein secretion pathway component
rif_2103_041	Sec-independent protein secretion pathway component
rif_2103_042	Sec-independent protein secretion pathway component TatC

Table 4.2 Lists of genes possibly involved in heterotrophy. Open reading frames are categorized by function. Italicized ORFs are clustered with functional genes on the contig identified, but may not necessarily be involved in the same function. The clustering of a cation/multidrug specific efflux pump with a mannitol-specific PTS system component may indicate that the symbiont uses both concurrently. This would be possible in the free-living stage if the symbiont lives heterotrophically outside of the host.

Regulation	
rif_0300_002	Phosphotransferase system IIC components, glucose/maltose
rif_0708_001	Transcriptional regulator (Crp family, AMP-binding protein)
rif_1263_002	Pyruvate kinase II, glucose stimulated
rif_1277_001	Phosphoenolpyruvate-protein kinase (PTS system EI component)
rif_1730_004	putative PTS system mannose-specific, factor IIC
rif_0693_003	possible TrapT family, dctP subunit, C4-dicarboxylate periplasm
rif_2043_005	Serine kinase of the HPr protein, regulates carbohydrate metabolism
rif_2043_040	Phosphotransferase system, HPr-related proteins
rif_2043_041	Phosphotransferase system, mannose/fructose-specific component IIA
rif_2043_042	Phosphotransferase system mannitol/fructose-specific IIA domain (Ntr-type)
rif_2233_001	Phosphotransferase system, mannitol-specific IIBC comp
rif_2233_002	*Cation/multidrug efflux pump*
rif_1525_001	Methyl-accepting chemotaxis citrate transducer
rif_1074_002	Phosphoenolpyruvate-protein phosphotransferase ptsP
rif_1277_001	Phosphoenolpyruvate-protein kinase (PTS system EI component)
rif_1074_001	(Di)nucleoside polyphosphate hydrolase (Ap4A pyrophosphatase)
rif_1074_002	Phosphoenolpyruvate-protein phosphotransferase ptsP
rif_1788_001	PEP:sugar phosphotransferase system enzyme I; signal transduction protein
rif_0331_002	PEP:sugar phosphotransferase system enzyme I; signal transduction protein

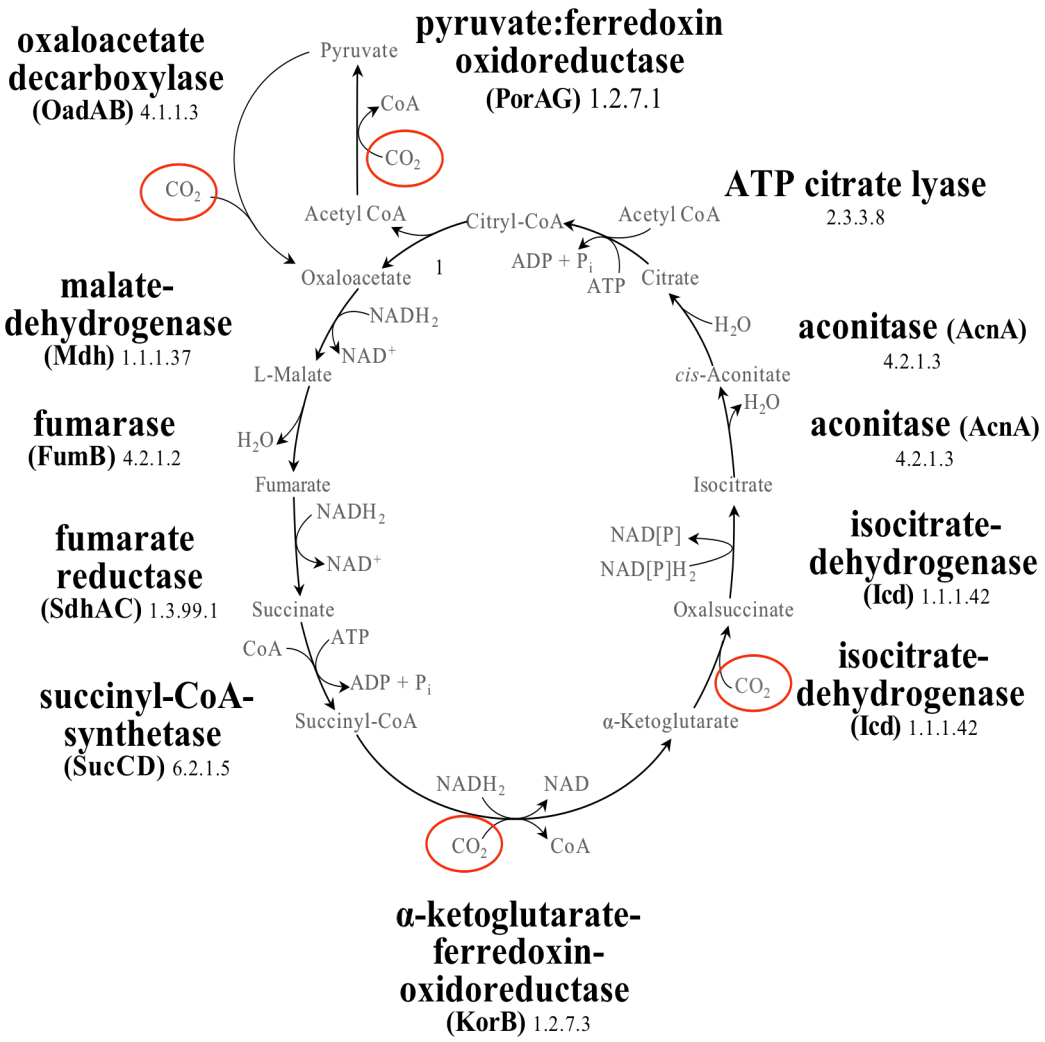
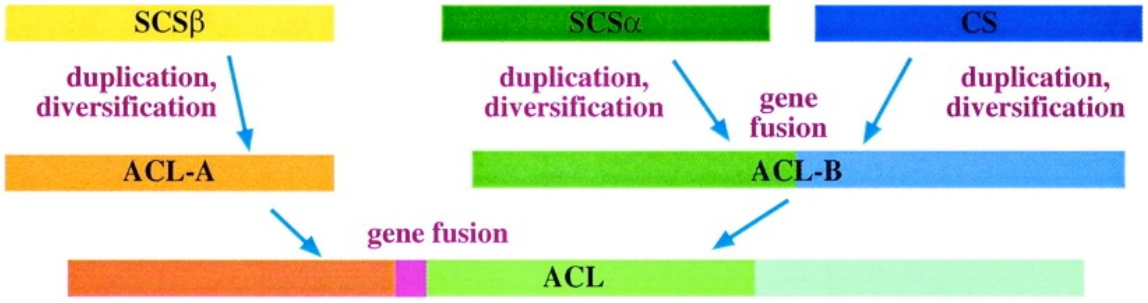


Figure 4.1 The reductive TCA cycle. Cycling in the oxidative direction occurs in heterotrophic metabolism. Reductive cycling results in carbon fixation. Key enzymes of the oxidative TCA Cycle include citrate synthase (in place of ATP citrate lyase), α-ketoglutarate dehydrogenase (in place of α-ketoglutarate oxidoreductase) and pyruvate carboxylase or oxaloacetate decarboxylase generate oxaloacetate from pyruvate, using 1 CO_2 (and 1 ATP in the case of pyruvate carboxylase; this happens instead of the pyruvate dehydrogenase reaction). Another difference between the two pathways is that pyruvate:ferredoxin oxidoreductase converts acetyl-CoA to pyruvate, consuming one CO_2. Figure modified from Markert *et al. in press.*



Fatland et al. 2002

Figure 4.2 The generation of ATP citrate lyase by duplication and diversification of succinyl CoA synthetase alpha and beta subunits and citrate synthetase. The final step only occurs in some eukaryotes. SCS = succinyl CoA synthetase, ACL = ATP citrate lyase.

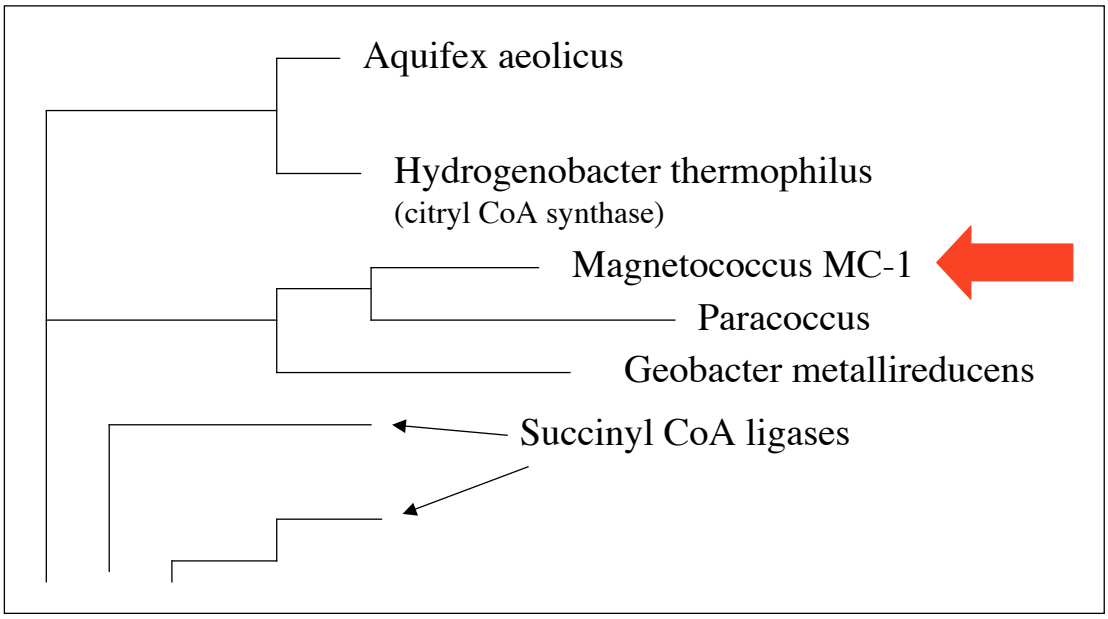


Figure 4.3 Phylogenetic tree of genes annotated as succinyl CoA synthetases. The top two major branches are likely to be ATP citrate lyases rather than succinyl CoA synthetases. The arrow indicates the Magnetococcus version, which is 87% identical to the *E. persephone* enzyme, at the amino acid level. *Aquifex* and *Hydrogenobacter* are archaea, while *Magnetococcus* is an epsilon and *Geobacter* a delta proteobacterium. This lack of phylogenetic grouping of organisms containing this enzyme is also seen with several CO_2 fixation enzymes (Hugler *et al.* 2003). This is a portion of a phylogenetic tree containing all succinyl CoA synthetases from the Seed database (http://theseed.uchicago.edu/FIG/index.cgi).

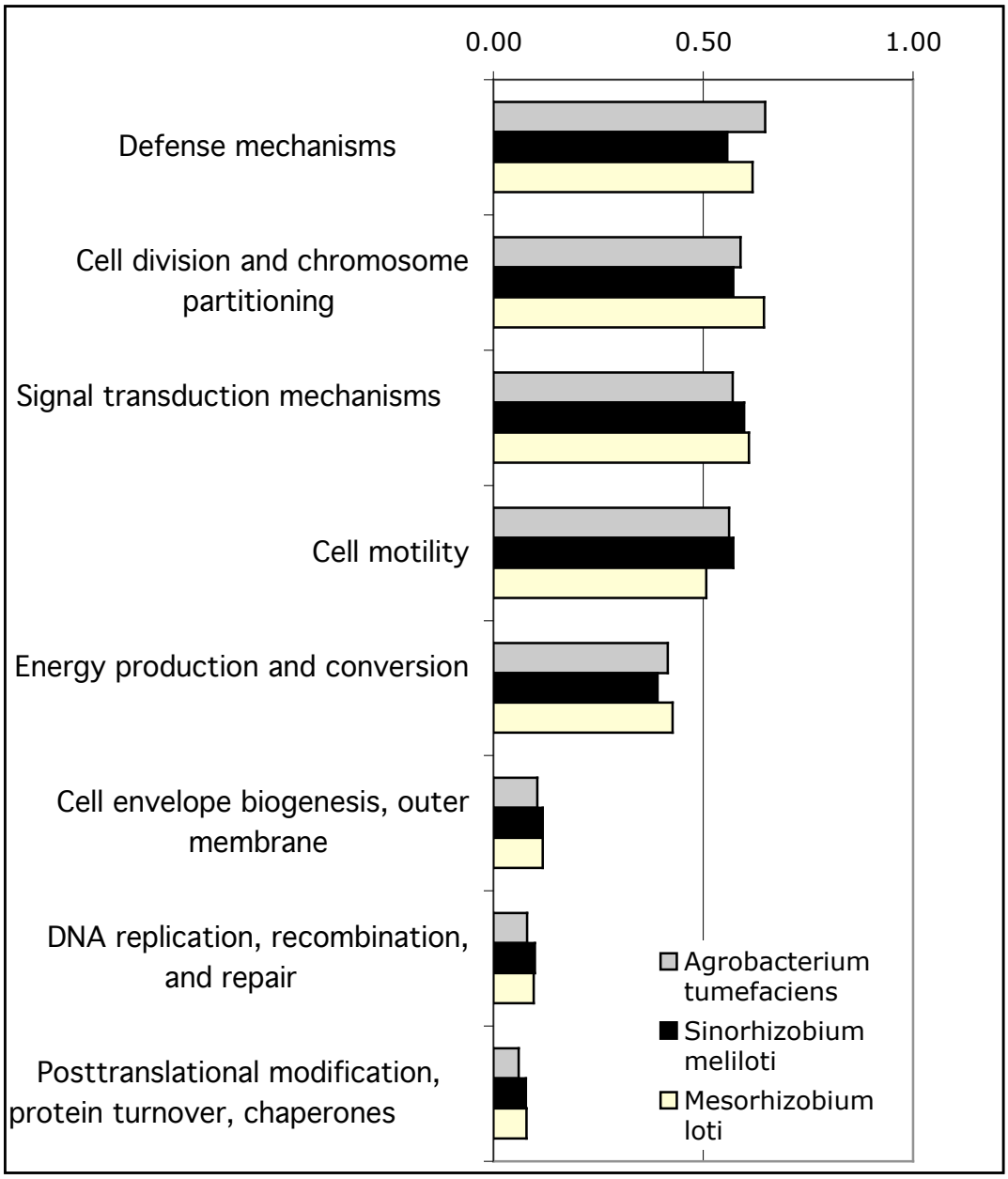


Figure 4.4 COG profiles showing categories for which *E. persephone* is genomically enhanced relative to plant symbionts. Percent composition from each genome is normalized to the percent composition in the *E. persephone* genome to calculate the degree of enhancement. Everything to the right of the line signifies enhancement in the *E. persephone* metagenome.

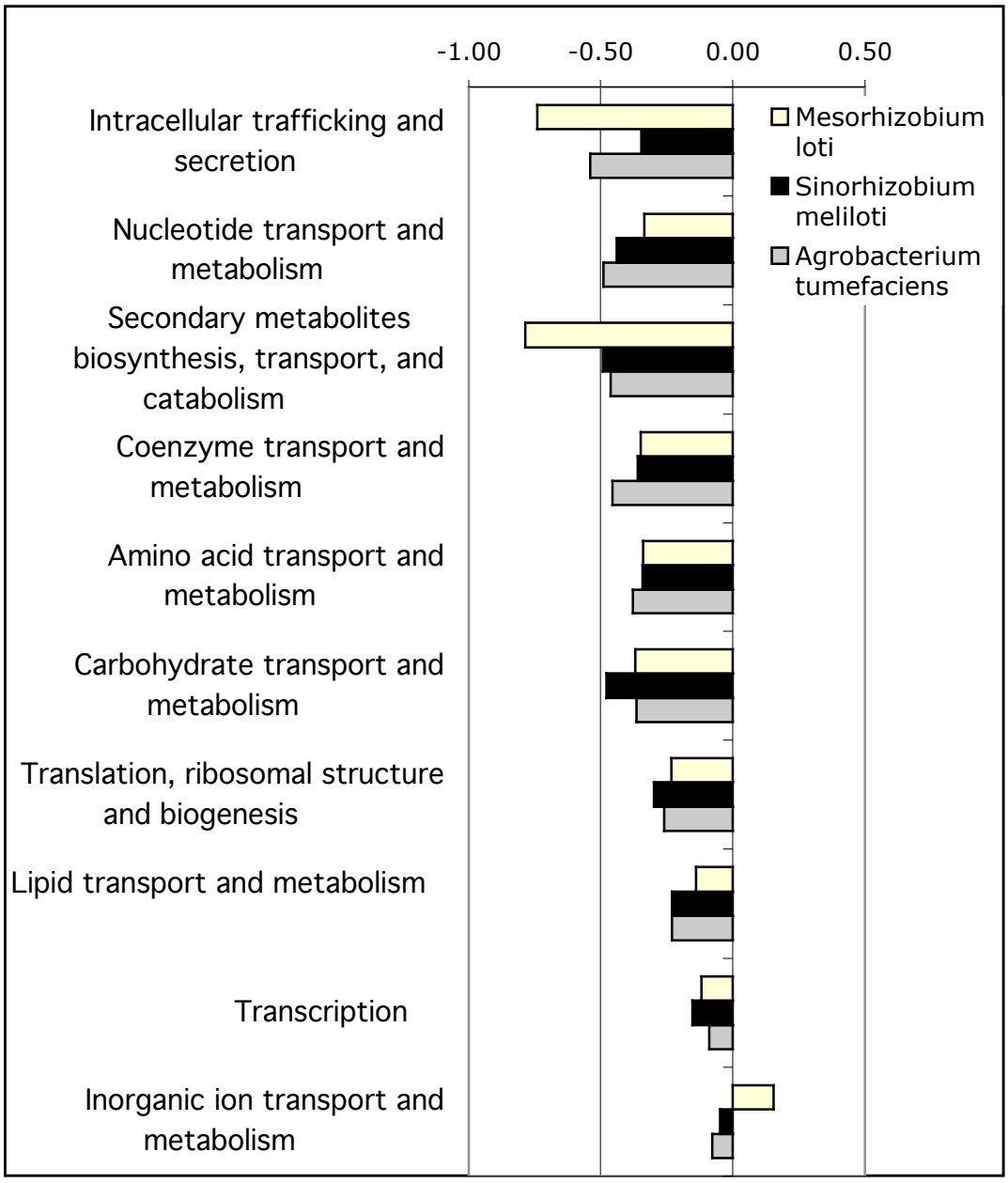


Figure 4.5 COG profiles showing categories for which *E. persephone* is genomically underrepresented relative to plant symbionts. Percent composition from each genome is normalized to the percent composition in the *E. persephone* genome to calculate the degree of genomic depletion. Everything to the left of the line signifies depletion in the *E. persephone* metagenome.

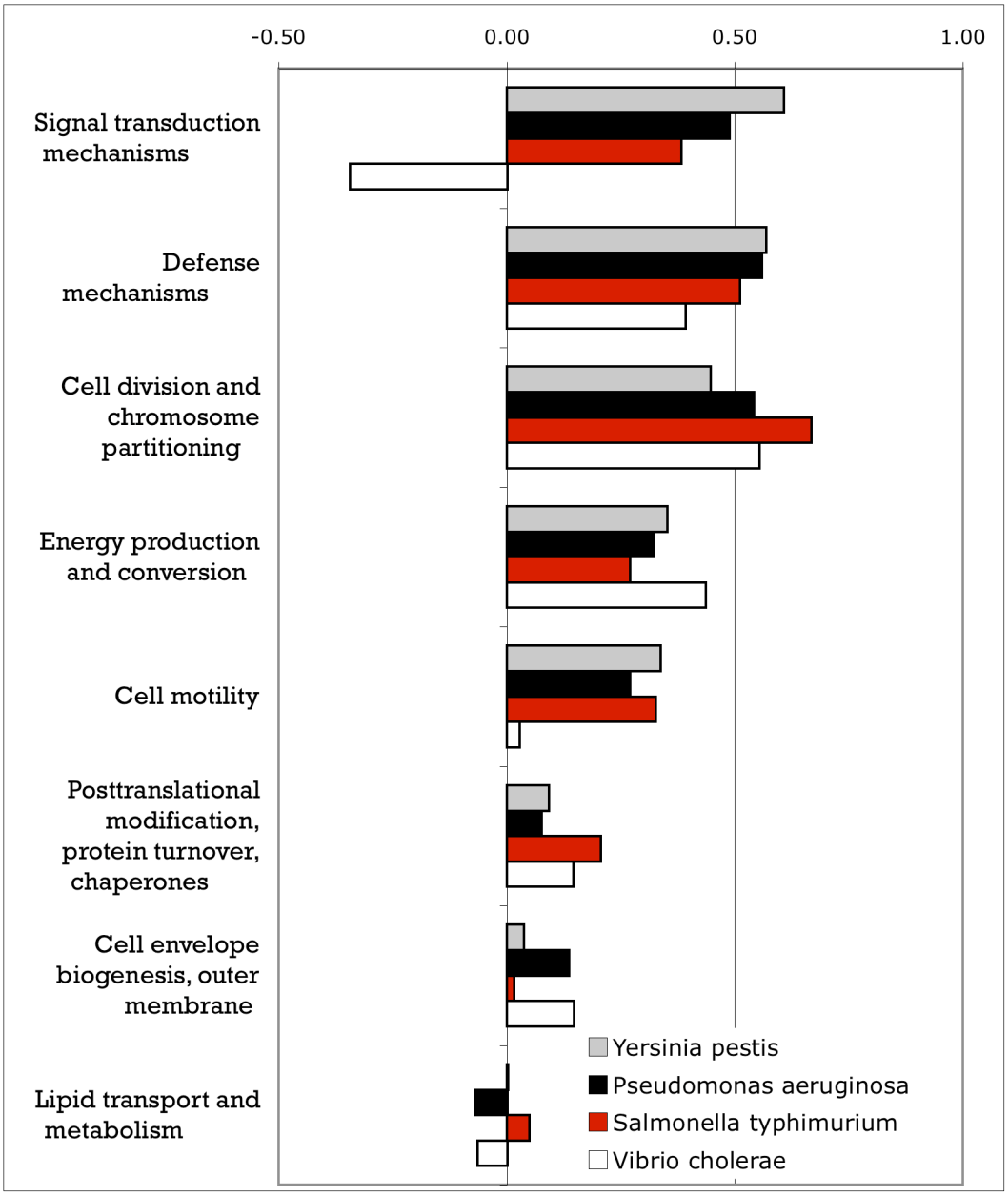


Figure 4.6 COG profiles showing categories for which *E. persephone* is genomically enhanced relative to gamma proteobacterial pathogens. Percent composition from each genome is normalized to the percent composition in the *E. persephone* genome to calculate the degree of genomic enhancement.

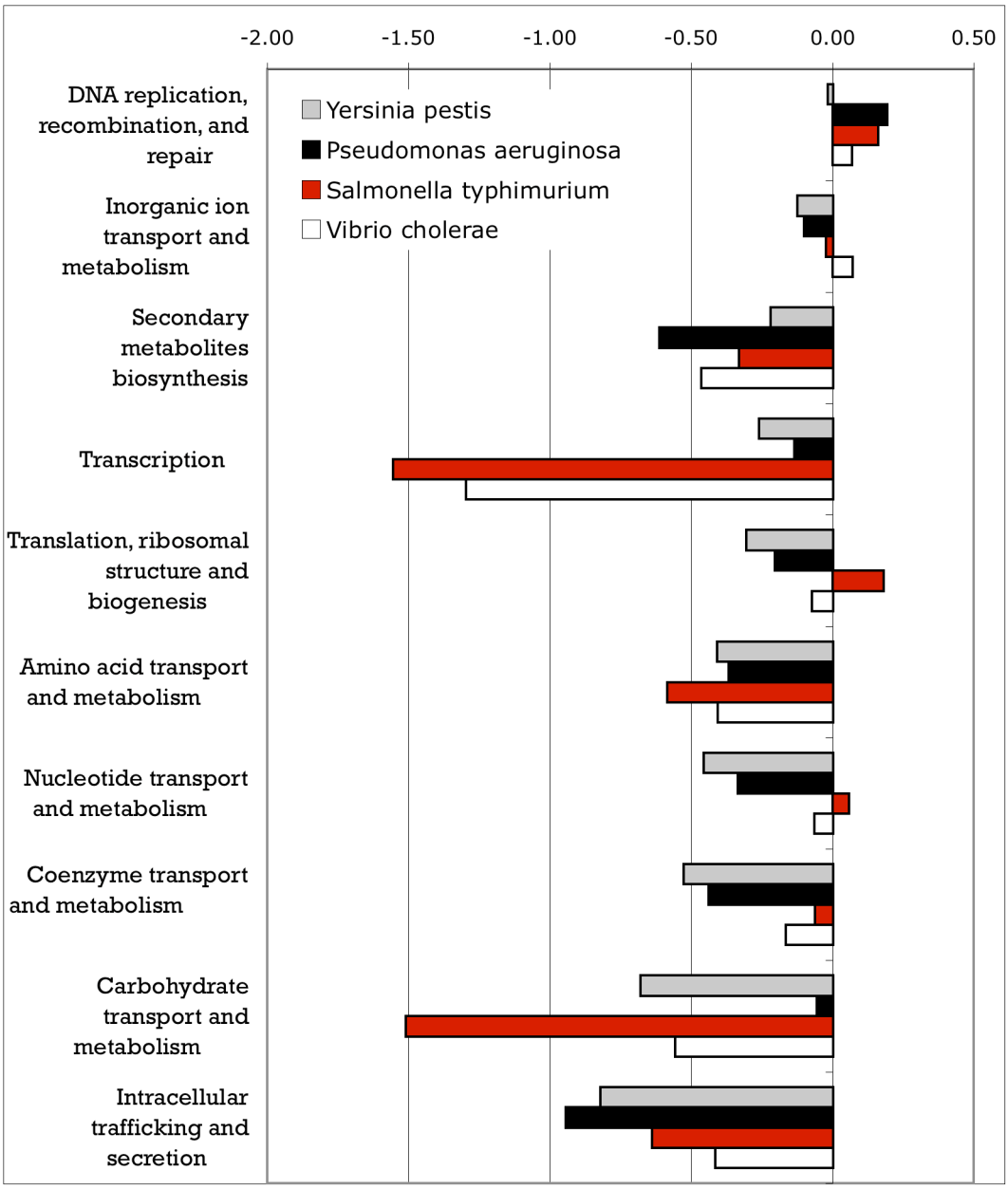


Figure 4.7 COG profiles showing categories for which *E. persephone* is genomically underrepresented relative to gamma proteobacterial pathogens. Percent composition from each genome is normalized to the percent composition in the *E. persephone* genome to calculate the degree of genomic depletion.

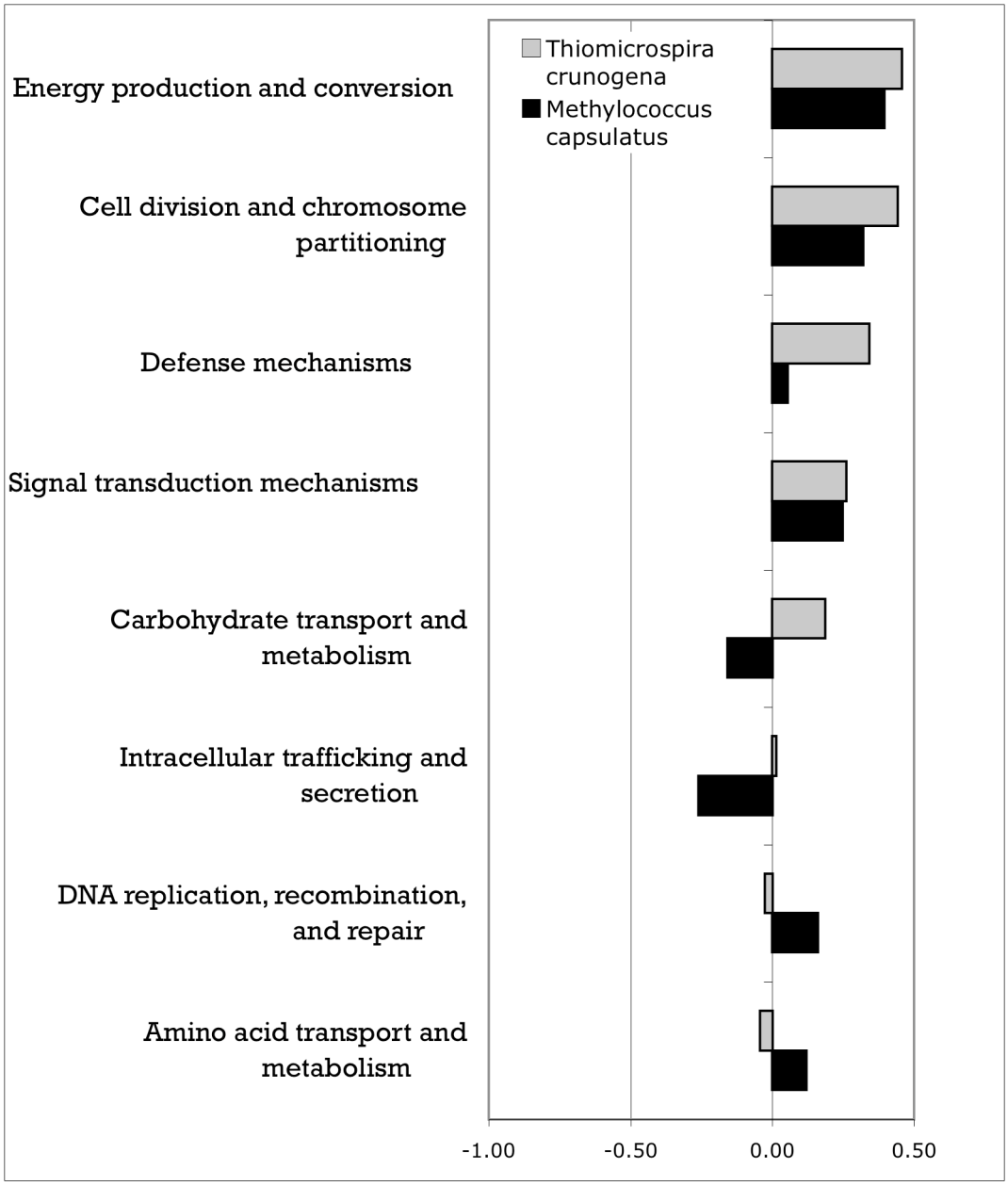


Figure 4.8 COG profiles showing categories for which *E. persephone* is genomically enhanced relative to its most closely related relatives with sequenced genomes. Percent composition from each genome is normalized to the percent composition in the *E. persephone* genome to calculate the degree of genomic enhancement.

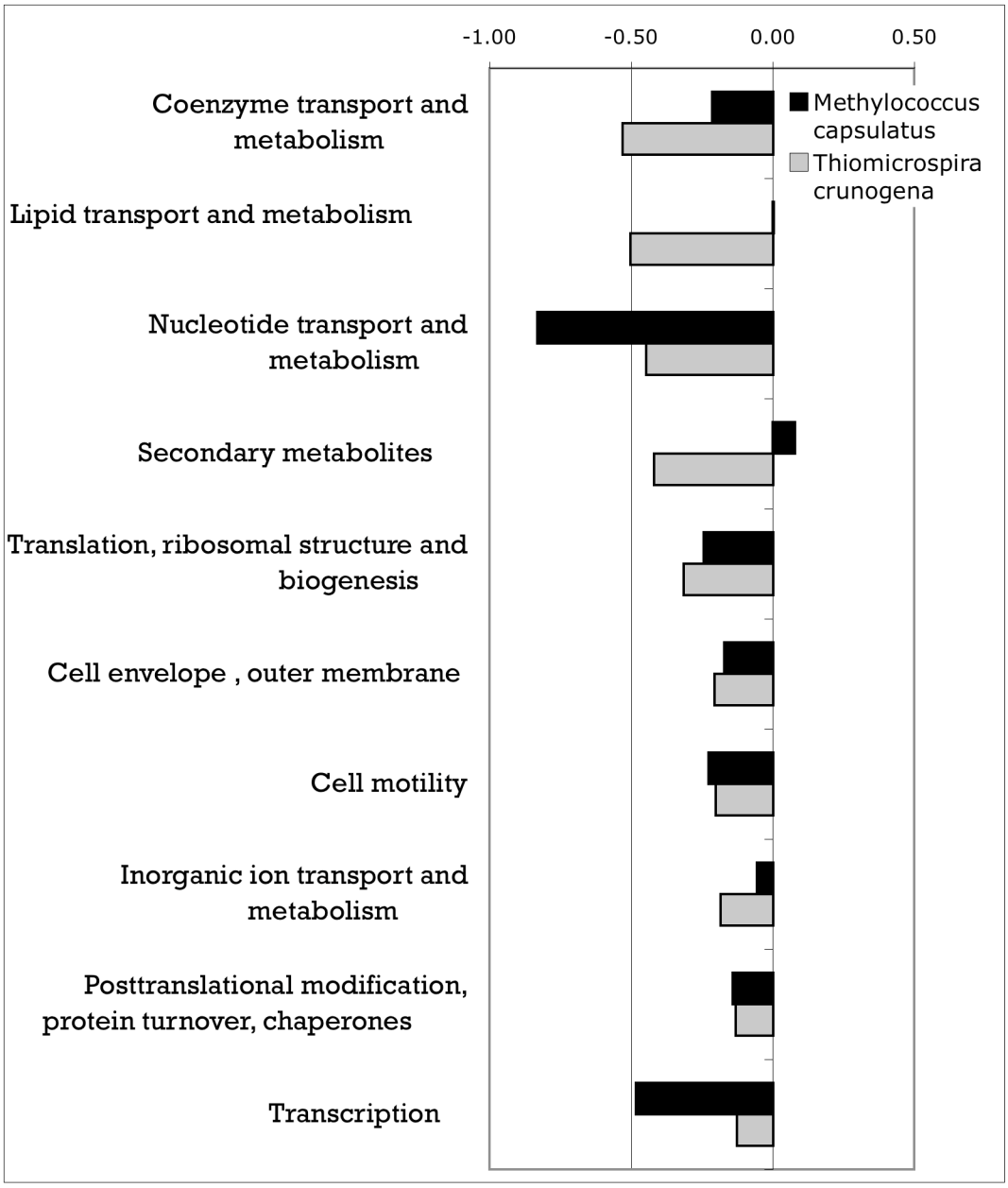


Figure 4.9 COG profiles showing categories for which *E. persephone* is genomically underrepresented relative to its closest relatives with sequenced genomes. Percent composition from each genome is normalized to the percent composition in the *E. persephone* genome to calculate the degree of genomic depletion.

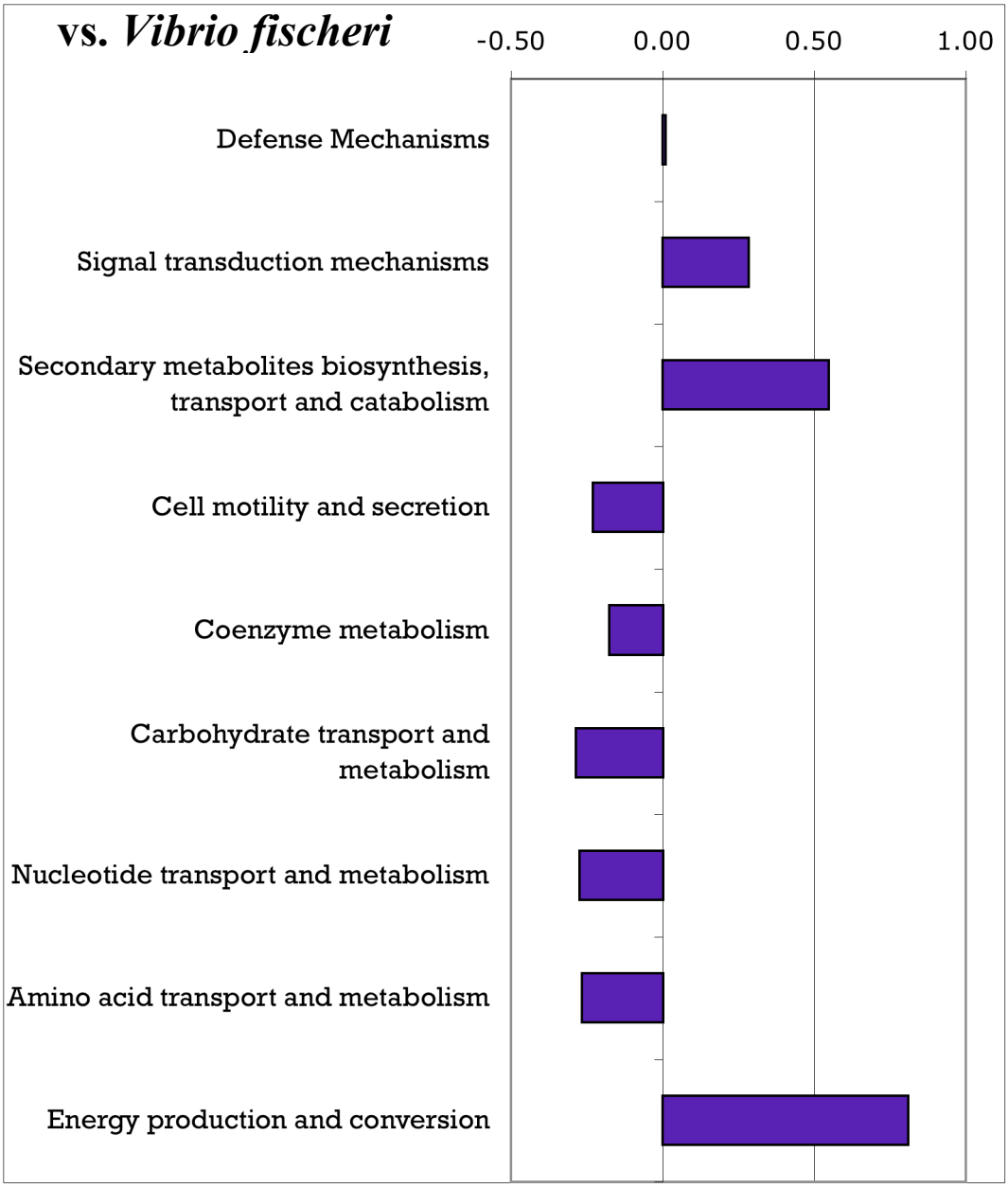


Figure 4.10 COG profiles showing comparison of *E. persephone* genomic component with that of *Vibrio fischeri*. Percent composition from each genome is normalized to the percent composition in the *E. persephone* genome to calculate the degree of genomic enhancement or depletion.

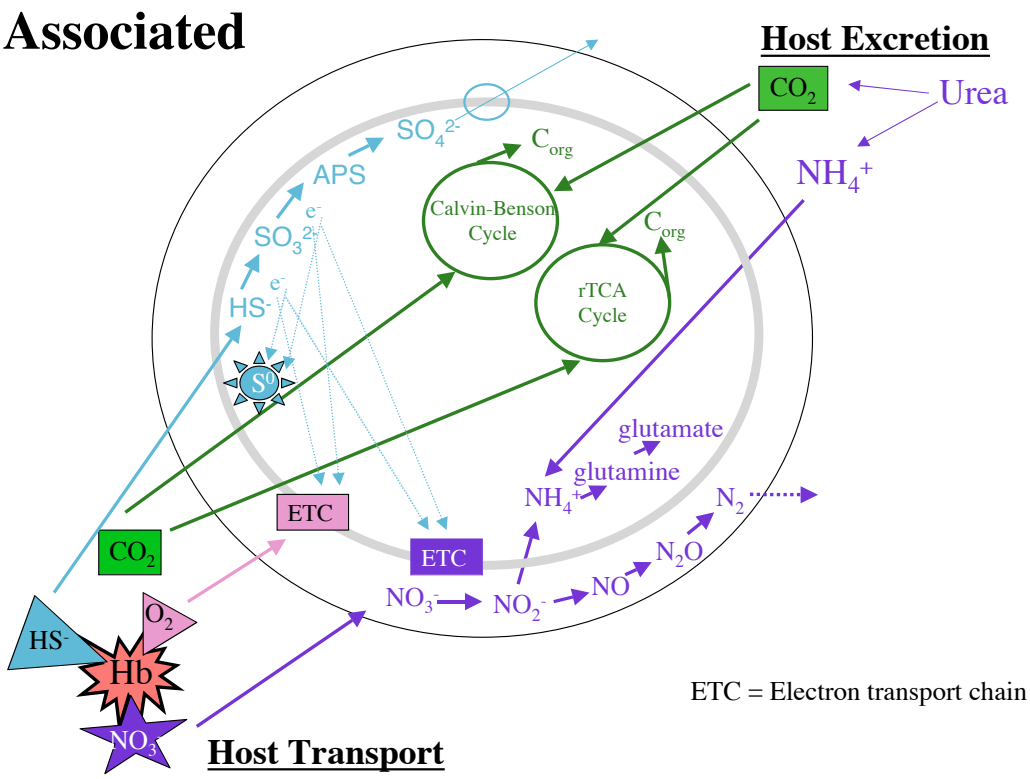


Figure 4.11 Physiology of *E. persephone* when associated with *R. pachyptila*. The symbiont can oxidize sulfide *via* reverse sulfate reduction and use oxygen, nitrate or sulfur as an electron sink. The energy generated from electron transport can fuel carbon dioxide fixation *via* the Calvin-Benson Cycle under high sulfide conditions, or the reverse TCA cycle when sulfide is low.

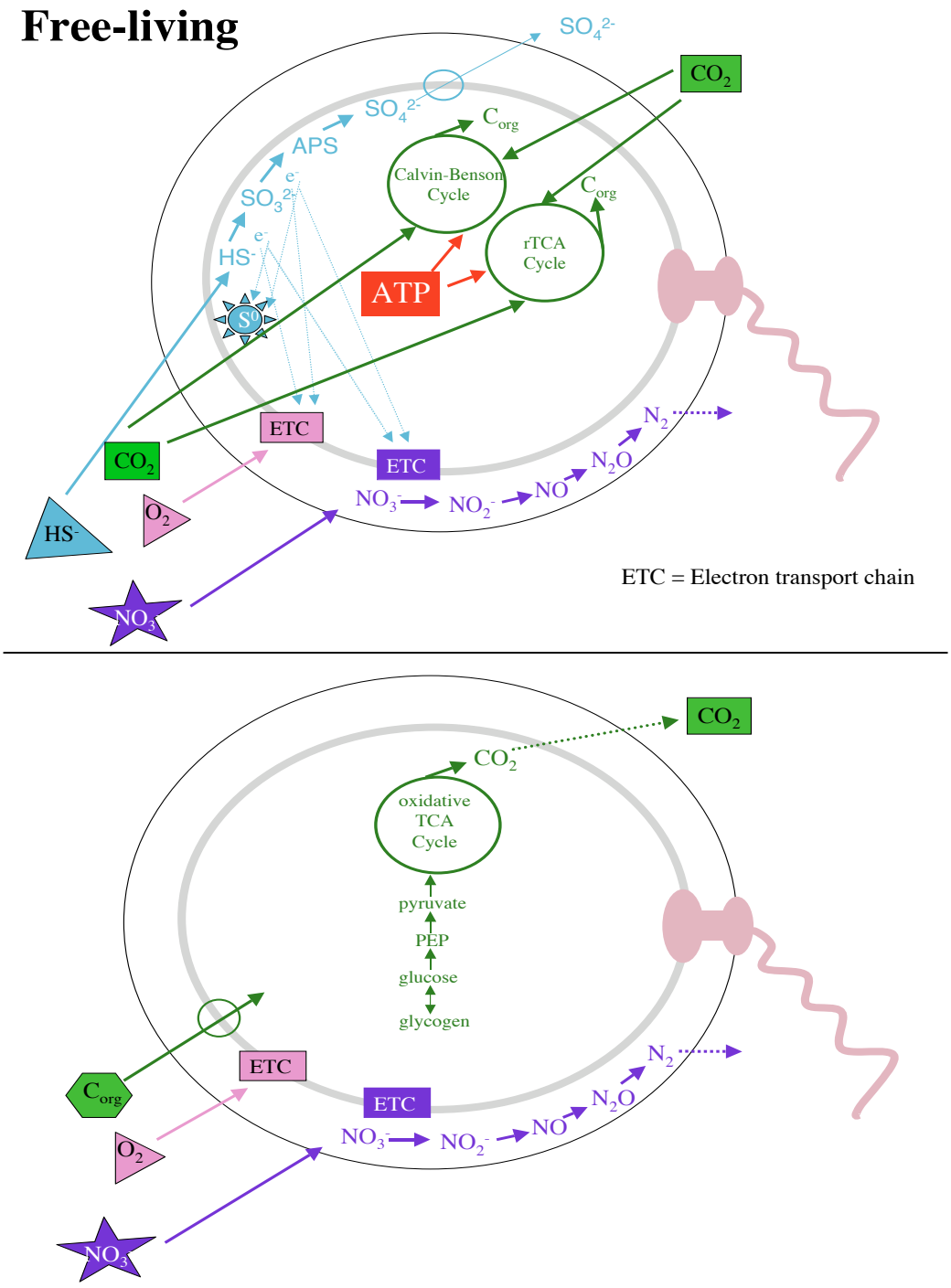


Figure 4.12 Physiology of *E. persephone* when free-living. The symbiont can live mixotrophically when free-living. Upper diagram shows autotrophy, lower shows heterotrophy. When organic carbon is available, it can gain both carbon and energy by heterotrophy and does not have to expend energy to fix carbon. The symbiont has a functional flagellum and several chemotaxis mechanisms.

ACKNOWLEDGMENTS

Thank you to Horst Felbeck for shared wisdom regarding the physiology of this symbiosis. Eric Allen advised with the comparative analysis. Thank you to Craig Young for the generous donation of time at sea (and Horst Felbeck for the invitation), and the captain and crew of the R/V Atlantis and DSR/V Alvin for collections and help during cruises. Thanks to Bianca Brahamsha, Doug Bartlett, Victor Nizet and Lihini Aluwihare for captivating discussions and suggestions for research directions. SymBio Corp. and Amersham Biosciences performed the sequencing and The Scripps Genome Center provided analyses. Sheila Podell was imperative in discussions and analysis.

REFERENCES

Anselme, C., A. Vallier, S. Balmand, M. Fauvarque and A. Heddi (2006). Host PGRP gene expression and bacterial release in endosymbiosis of the weevil *Sitophilus zeamais. Applied and Environmental Microbiology* **72**: 6766-6772.

Arndt, C., F. Gaill and H. Felbeck (2001). Anaerobic sulfur metabolism in thiotrophic symbioses. *Journal of Experimental Biology* **204**: 741-750.

Arp, A.J. and J.J. Childress (1983). Sulfide binding by the blood of the hydrothermal vent tubeworm *Riftia pachyptila. Science* **219**: 295-297.

Benson, D.A., D.J. Lipman, I. Karsch-Mizrachi, J. Ostell, B.A. Rapp and D.L. Wheeler (2000). GenBank. *Nucleic Acids Research* **28**: 15-18.

Bright, M., H. Keckelts and C. R. Fisher (2000). An autoradiographic examination of carbon fixation, transfer and utilization in the *Riftia pachyptila* symbiosis. *Marine Biology* **136**: 621-632.

Bright, M. and A. Sorgo (2003). Ultrastructural reinvestigation of the trophosome in adults of *Riftia pachyptila* (Annelida, Siboglinidae). *Invertebrate Biology* **122**: 345-366.

Campbell, B.J., and S.C. Cary (2004). Abundance of reverse tricarboxylic acid cycle genes in free-living microorganisms at deep-sea hydrothermal vents. *Applied and Environmental Microbiology* **70**: 6282-6289.

Cary, S. C., W. Warren, E. Anderson and S. J. Giovannoni (1993). Identification and localization of bacterial endosymbionts in hydrothermal vent taxa with symbiont-specific polymerase chain reaction amplification and *in situ* hybridization techniques. *Molecular Marine Biology Biotechnology* **2**(1): 51-62.

Cavanaugh, C.M. (1983). Symbiotic chemoautotrophic bacteria in marine invertebrates from sulfide-rich habitats. *Nature* **302**: 58-60.

Cavanaugh, C.M., T.L. Harmer, A.D. Nussbaumer, and M. Bright (2005). Environmental transmission in the hydrothermal vent Vestimentiferan – chemoautotroph symbiosis: Stalking the wild symbiont. 3rd International Symposium on Hydrothermal Vent and Seep Biology, La Jolla, USA.

Childress, J.J. (1988). Biology and chemistry of a deep-sea hydrothermal vent on the Galapagos Rift: the Rose Garden in 1985, and introduction. *Deep Sea Research* **35**: 1677-1680.

Childress, J.J. (1992). The biology of hydrothermal vent animals: physiology, biochemistry, and autotrophic symbioses. *Oceanography Marine Biology Annual Reviews* **30**: 337-441.

Childress, J.J., R.W. Lee, N.K. Sanders, H. Felbeck, D.R. Oros, A. Toulmond, D. Desbruyeres, M.C. Kennicutt and J.M. Brooks (1993). Inorganic carbon uptake in hydrothermal vent tubeworms facilitated by high environmental pCO_2. *Nature* **362**: 147-149.

Childress, J.J., A.J. Arp and C.R. Fisher (1984). Metabolic and blood characteristics of the hydrothermal vent tubeworm *Riftia pachyptila*. *Marine Biology* **83**: 109-124.

Comita, P.B., R.B. Gagosian and P.M. Williams (1984). Suspended particulate organic material from hydrothermal vent waters at 21ºN. *Nature* **307**: 450-453.

Distel, D.L. and H. Felbeck (1988). Pathways of inorganic carbon fixation in the endosymbiont-bearing lucinid clam *Lucinoma aequizonata*. Part 1. Purification and characterization of the endosymbiotic bacteria. *Journal of Experimental Zoology* **247**: 1-10.

Dziejman, M., E. Balon, D. Boyd, C.M. Fraser, J.F. Heidelberg and J.J. Mekalanos (2002). Comparative genomic analysis of *Vibrio cholerae*: Genes that correlate with cholera endemic and pandemic disease. *Proceedings of the National Academy of Sciences* **99**: 1556-1561.

Fatland, B.L., J. Ke, M.D. Anderson, W.I. Mentzen, L.W. Cui, C.C. Allred, J.L. Johnston, B.J. Nikolau and E.S. Wurtele (2002). Molecular characterization of a heteromeric ATP-citrate lyase that generates cytosolic acetyl-coenzyme A in *Arabidopsis*. *Plant Physiology* **130**: 741-756.

Felbeck H (1981). The chemoautotrophic potential of the hydrothermal vent tube worm, *Riftia pachyptila* Jones (Vestimentifera). *Science* **213**: 336-338.

Felbeck, H. (1985). CO_2 fixation in the hydrothermal vent tubeworm *Riftia pachyptila* (Jones). *Physiological Zoology* **58**: 272-281.

Felbeck, H. and J. Jarchow (1997). Carbon release from purified chemoautotrophic bacterial symbionts of the hydrothermal vent tubeworm *Riftia pachyptila*. *Physiological Zoology* **71**: 294-302.

Felbeck, H., C. Arndt, U. Hentschel and J.J. Childress (2004). Experimental application of vascular and coelomic catheterization to identify vascular transport mechanisms for inorganic carbon in the vent tubeworm, *Riftia pachyptila*. *Deep-Sea Research I* **51**: 401-411.

Fisher, C.R., M.C.I. Kennicutt and J.M. Brooks (1990). Stable carbon isotopic evidence for carbon limitation in hydrothermal vent vestimentiferans. *Science* **247**: 1094-1096.

Fuchs, G., E. Stupperich and R. Jaenchen (1980). Autotrophic CO_2 fixation in *Chlorobium limicola*. Evidence against the operation of the Calvin cycle in growing cells. *Archives of Microbiology* **128**: 56-63.

Gill, S.R., . Pop, R.T. DeBoy, P.B. Eckburg, P.J. Turnbaugh, B.S. Samuel, J.I. Gordon, D.A. Relman, C.M. Fraser-Liggett and K.E. Nelson (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312**: 1355-1359.

Girguis, P.R. and J.J. Childress (2006). Metabolite uptake, stoichiometry and chemoautotrophic function of the hydrothermal vent tubeworm *Riftia pachyptila*: responses to environmental variations in substrate concentrations and temperature. *Journal of Experimental Biology* **209**: 3516-3528.

Goffredi, S.K., J.J. Childress, N.T. Desaulniers, R.W. Lee, F.H. Lallier and D. Hammond (1997). Inorganic carbon acquisition by the hydrothermal vent tubeworm *Riftia pachyptila* depends upon high external pCO_2 and upon proton-equivalent ion transport by the worm. *Journal of Experimental Biology* **200**: 883-896.

Hentschel, U., and H. Felbeck (1993). Nitrate respiration in the hydrothermal vent tubeworm *Riftia pachyptila*. Nature **366**: 338-340.

Hugler, M., H. Huber, K.O. Stetter and G. Fuchs (2003). Autotrophic CO_2 fixation pathways in archaea (Crenarchaeota). *Archives Microbiology* **179**: 160-173.

Jeanthon, C. (2000). Molecular ecology of hydrothermal vent microbial communities. *Antonie van Leeuwenhoek* **77**: 117-133.

Jennings R.M. and K.M. Halanych (2005). Mitochondrial genomes of *Clymenella torquata* (Malanidae) and *Riftia pachyptila* (Siboglinidae): Evidence for conserved gene order in Annelida. *Molecular Biology and Evolution* **22**: 210-222.

Johnson, K.S., J.J. Childress, R.R. Hessler, A.C.M. Sakamoto and C.L. Beehler (1988). Chemical and biological interactions in the Rose Garden hydrothermal vent field. *Deep Sea Research Part A: Oceanography Research Papers* **35**: 1723-1744.

Johnson, K.S., C.L. Beehler, C.M. Sakamoto-Arnold and J.J. Childress (1986). In situ measurements of chemical distributions in a deep sea hydrothermal vent field. *Science* **231**: 1139-1141.

Kormas, K.A., M.K. Tivey, J. Von Damm and A. Teske (2006). Bacterial and archaeal phylotypes associated with distinct mineralogical layers of a white smoker spire from a deep-sea hydrothermal vent site (9ºN, East Pacific Rise). *Environmental Microbiology* **8**: 909-920.

Le Bris, N., B. Govenar, C. Le Gall and C.R. Fisher (2006). Variability of physico-chemical conditions in 9º50'N EPR diffuse flow vent habitats. *Marine Chemistry* **98**: 167-182.

Londry, K.L., L.L. Jahnke and D.J. Des Marais (2004). Stable carbon isotope ratios of lipid biomarkers of sulfate-reducing bacteria. *Applied and Environmental Microbiology* **70**: 745-751.

Luther, G.W., B.T. Glazer, L. Hohman, J.I. Popp, M. Taillefert, T.F. Rozan, P.J. Brendel, S.M. Theberge and D.B. Nuzzio (2001). Sulfur speciation monitored in situ with solid state gold amalgam voltammetric microelectrodes: polysulfides as a special case in sediments, microbial mats and hydrothermal vent waters. *Journal of Environmental Monitoring* **3**: 61-66.

Lutz, R.A., T.M. Shank, D.J. Fornari, R.M. Haymon, M.D. Lilley, K.L. Von Damm, D. Desbruyeres (1994). Rapid growth at deep-sea vents. *Nature* **371**: 663-664.

Markert, S. C. Arndt, H. Felbeck, D. Becher, S.M. Sievert, M. Hugler, D. Albrecht, J. Robidart, S. Bench, R.A. Feldman, M. Hecker and T. Schweder (*in press*). Approaching the uncultured endosymbiont of *Riftia pachyptila* by physiological proteomics. *Science.*

Millikan, D.S. H. Felbeck and J.S. Stein (1999). Identification and characterization of a flagellin gene from the endosymbiont of the hydrothermal vent tubeworm *Riftia pachyptila*. *Applied and Environmental Microbiology* **65**: 3129-3133.

Minic, Z. V. Simon, B. Penverne, F. Gaill and G. Herve (2001). Contribution of the bacterial endosymbiont to the biosynthesis of pyrimidine nucleotides in the deep-sea tube worm *Riftia pachyptila*. *Journal of Biological Chemistry* **276**: 23777-23784.

Nussbaumer, A.D., C.R. Fisher and M. Bright (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* **441**: 345-348.

Nyholm, S.V. and M.J. McFall-Ngai (2004). The winnowing: Establishing the squid-Vibrio symbiosis. *Nature Reviews Microbiology* **2**: 632-642.

Parsek, M.R. and E.P. Greenberg (2000). Acyl-homoserine lactone quorum sensing in Gram-negative bacteria: A signaling mechanism involved in associations with higher organisms. *Proceedings of the National Academy of Sciences* **97**: 8789-8793.

Podell, S. and T. Gaasterland, submitted. DarkHorse: A method for genome-wide prediction of horizontal gene transfer.

Robinson, J.J., K.M. Scott, S.T. Swanson, M.H. O'Leary, K. Horken, F.H. Tabita and C.M. Cavanaugh (2003). Kinetic isotope effect and characterization of form II RuBisCO from the chemoautotrophic endosymbionts of the hydrothermal vent tubeworm *Riftia pachyptila. Limnology and Oceanography* **48**: 48-54.

Ruby, E.G., C.O. Wirsen and H.W. Jannasch (1981). Chemolithotrophic sulfur-oxidizing bacteria from the Galapagos Rift hydrothermal vents. *Applied and Environmental Microbiology* **42**: 317-324.

Schauer, R. (2000). Achievements and challenges of sialic acid research. *Glycoconjugate Journal* **17**: 499.

Tabita, F.R. (1988). Molecular and cellular regulation of autotrophic carbon dioxide fixation in microorganisms. *Microbiological Reviews* **52**: 155-189.

Williams, T.J., C.L. Zhang, J.H. Scott and D. Bazylinski (2006). Evidence for autotrophy *via* the reverse tricarboxylic acid cycle in the marine magnetotactic coccus strain MC-1. *Applied and Environmental Microbiology* **72**: 1322-1329.

Wirsen, C.O., H.W. Jannasch and S.J. Molyneaux (1993). Chemosynthetic microbial activity at Mid-Atlantic Ridge hydrothermal vent sites. *Journal of Geophysical Research* **98**: 9693-9703.

Chapter 5

Conclusion:

The Utility of Metagenomics in Studies of Chemosynthetic Symbioses

CONCLUSIONS

Various molecular tools facilitate studies of biological organisms in remote environments. A wealth of knowledge can be gained by studies of the DNA, RNA and protein composition of an organism. In microbiology, the ever-increasing field of genomics has allowed investigators to access an organism's entire metabolic repertoire. The value of microbial genome sequences is two-fold: molecular biology on the system can be advanced with the design and use of specific primers, probes, antibodies and various other molecular materials, and the genome can be used to learn about the potential physiology of the organism to create new hypotheses to be tested experimentally.

Through the *E. persephone* metagenome project we have learned various strategies to manage fragmented environmental metagenomes. Specifically with regard to *E. persephone,* we found as expected (Distel et al. 1988; Edwards and Nelson 1991) that the symbiotic inhabitants in the trophosome are near to, if not truly a single species. We have also elucidated the enzymatic components of various biochemical pathways, and learned of a second carbon fixation pathway in the symbiont. With the use of alternate terminal electron acceptors, two possible carbon fixation pathways, as well as the possibility of heterotrophic metabolism, *Endoriftia persephone* is a truly versatile organism.

The *E. persephone* metagenome is not similar to any single bacterial species. Based on 16S ribosomal RNA phylogeny, there are few organisms with sequenced genomes that resemble *E. persephone*. There are also few sequenced genomes from hydrothermal vent bacteria. The description of the *E. persephone* genome as unique is not likely to be accurate in a few years from now, as new genomes are added to GenBank with increasing frequency. The unique nature of the symbiosis, however, will likely remain the case and may contribute to its eternal distinction.

The symbiont's clusters of orthologous groups (COG) profile is almost as distinctive as that of *Vibrio cholerae*. These two organisms live in a range of conditions, including the open ocean and the host environment. *V. cholerae* can infect various hosts from

phytoplankton to humans (Reidi and Klose 2002), while the *Riftia* symbiont infects several vestimentiferan species (Di Meo et al. 2000) but has not been found in other genera. *V. cholerae*'s wide host range is reflected in its increased genomic component dedicated to signal transduction and chemotaxis. It is likely that the hydrothermal origin of *E. persephone* is at least partly responsible for an augmented capacity for response to the environment relative to pathogens and symbionts in general.

It seems that the symbiont has the capacity to respond and adapt by signal transduction rather than quorum sensing. Its is very metabolically versatile and has been found in a variety of environments (Cavanaugh *et al.* 2005). It is not an obligate symbiont and so is not dependent on its host (Cary *et al*. 1993). It appears from the genome that this organism should be a good candidate for culture studies, and it is curious that the symbiont has not been cultured to date. Perhaps a heterotrophic requirement for unassociated growth has contributed to the past culturing failures.

The genomic comparison of the *Riftia* symbiont to the symbiotic bacteria of the related genera *Oasisia*, *Tevnia*, and *Ridgea* may reveal that these symbionts are metabolically diverse. Genomic variation in symbionts from these sister taxa does exist (Edwards and Nelson 1991; Di Meo *et al*. 2000), though the 16S signatures identify them as identical species. Genomic subtractive hybridization with purified symbionts from each host species can provide this information. Moreover, a subtractive hybridization study comparing symbionts from the same host species, but at separate hydrothermal sites might provide insight into adaptation to environmental conditions at different type of vents (e.g. fast-spreading vs. slow-spreading). There is a distinct advantage with these associations: collection of large quantities of biological material for experimentation is possible due to high concentrations of symbionts within the trophosome. This has proven useful in various types of metabolic studies and has allowed experimentation despite the lack of culture: the host acts as the culture vessel. Metagenomic analysis of this uncultured organism is only possible because of this factor. Studies of vestimentiferan species from non-vent habitats or with

unrelated symbionts is appealing, to verify our speculations on the unusual COG characteristics. The genome of the phylogenetically related organelle-type chemosynthetic symbiont of *Calyptogena magnifica* is currently being sequenced (Cavanaugh, personal communication). This is a more difficult metagenomic study, as the symbionts are not as abundant and they are localized within the gills, which are used to filter seawater and therefore contain "contaminating" non-symbiont microbes. This metagenome will likely be more fragmented than that of *E. persephone*, but will be useful for comparative analyses of genome reduction and specific adaptations to bivalve hosts.

Whether the same *Riftia* symbiont population present pre-eruption infects newly colonizing juveniles in a post-eruption hydrothermal system remains unknown. With the recent eruption of 9ºN on the East Pacific Rise (http://www.interridge.org), we will soon have the ability to answer this question.

The hypothesis of the symbiont's consumption of the host after the host perishes would be simple to verify or reject. After collection of a worm, it can be removed from its tube and the body wall can be sterilized. If that does not kill the worm, the worm must be killed and subsequently maintained in a semi-permeable bag in a pressure chamber. Once the worm is decomposed any live bacteria present can be identified.

The use of the reductive TCA (rTCA) cycle and the Calvin-Benson Cycle under high sulfide and oxygen conditions should be studied. In Girguis and Childress (2006), the graph of CO_2 uptake by the association shows two distinct peaks at different sulfide concentrations, and uptake is much higher when sulfide is abundant. This is when the Calvin-Benson Cycle is functioning, according to proteomics (Markert *et al. in press*). It is unusual that a pathway with an enzyme renowned for its inefficiency fixes carbon to a much greater degree than the rTCA cycle. These two cycles in a single organism allow for direct comparative kinetics. Incubation of the intact symbiosis in experimental aquaria with varied substrate concentrations will allow for concurrent analysis of host and symbiont activities. The exact high and low sulfide conditions compared in Markert *et al*. are unknown. The relative sulfide concentrations from

worms used in this study were qualitatively determined by relating to the concentrations of sulfur stores in the worms' trophosomes. Internal sulfur storage has been demonstrated to reflect sulfide conditions (Arndt *et al*. 2001). It is therefore necessary to revisit the exhaustive Girguis and Childress (2006) experiment to unambiguously contribute the large peak to the Calvin-Benson Cycle. Proteomics is the method of choice in this investigation, as proteins are not as labile as RNA and will be more likely to survive the purification process (Sambrook and Russel 2001). Analysis by tandem mass spectrometry will allow characterization of much more of the proteome than identification by 2-dimensional gel analysis coupled with mass spectrometry alone. It is likely that along with RuBisCO, various other enzymes are differentially expressed under various conditions, and total protein analysis would yield more information concerning physiological adaptation of the symbiont under different conditions.

The inability to culture *E. persephone* creates difficulties for biologists studying the *Riftia pachyptila* symbiosis. Easy access to a growing organism contributes to great gains in any biological field. It is thought that the lack of success with cultivation attempts stems from an irreversible physiological change that the symbionts underwent post-colonization. (Felbeck, personal communication). Perhaps the symbionts in the trophosome are in a viable but non-culturable state, leading to failed culture attempts. Now that investigators are aware that the symbiont is present in widely diverse conditions in the hydrothermal environment (Cavanaugh *et al*. 2005), perhaps environmental sampling (e.g. of worm tubes) rather than of trophosome will lead to successful cultivation. The genomic component of *Endoriftia persephone* strongly suggests the possibility of heterotrophic metabolism. Many microbes from hydrothermal vents are indeed mixotrophic (Jeanthon 2000) and in theory it is likely of great benefit to vent microbes to have the ability to change physiology according to available substrates. The concentrations of chemical substrates fluctuates greatly in the hydrothermal environment (Johnson et al. 1986) and adaptability is likely necessary for survival. This new information will aid in culturing approaches with the symbiont. If *E. persephone* does indeed live heterotrophically while outside of the host, addition of small concentrations of environmentally

relevant types of organic carbon to the media may provide the symbiont with a more appropriate substrate.

The metagenomic potential of the *Riftia* symbiont has led to much speculation regarding its physiology. Testable hypotheses have been created and the publication of the genome will undoubtedly prove a useful tool for all investigators studying this symbiosis.

REFERENCES

Arndt, C., F. Gaill and H. Felbeck (2001). Anaerobic sulfur metabolism in thiotrophic symbioses. *Journal of Experimental Biology* **204**: 741-750.

Cary, S. C., W. Warren, E. Anderson and S. J. Giovannoni (1993). Identification and localization of bacterial endosymbionts in hydrothermal vent taxa with symbiont-specific polymerase chain reaction amplification and *in situ* hybridization techniques. *Molecular Marine Biology Biotechnology* **2**(1): 51-62.

Cavanaugh, C.M., T.L. Harmer, A.D. Nussbaumer, and M. Bright (2005). Environmental transmission in the hydrothermal vent vestimentiferan – chemoautotroph symbiosis: Stalking the wild symbiont. 3rd International Symposium on Hydrothermal Vent and Seep Biology, La Jolla, USA.

Di Meo, C.A., Wilbur, A.E., W.E. Holben, R.A. Feldman, R.C. Vrijenhoek and S.C. Cary (2000). Genetic variation among endosymbionts of widely distributed vestimentiferan tubeworms. *Applied and Environmental Microbiology* **66**: 651-658.

Distel, D.L., D.J. Lane, G.J. Olsen, S.J. Giovanonni, B. Pace, N.R. Pace, D.A. Stahl and H. Felbeck (1988). Sulfur-oxidizing bacterial endosymbionts: analysis of phylogeny and specificity by 16S rRNA sequences. *Journal of Bacteriology* **170**: 2506-2510.

Edwards, D.B. and D.C. Nelson (1991). DNA-DNA solution hybridization studies of the bacterial symbionts of hydrothermal vent tube worms (*Riftia pachyptila* and *Tevnia jerichonana*). *Applied and Environmental Microbiology* **57**: 1082-1088.

Girguis, P.R. and J.J. Childress (2006). Metabolite uptake, stoichiometry and chemoautotrophic function of the hydrothermal vent tubeworm *Riftia pachyptila*: responses to environmental variations in substrate concentrations and temperature. *Journal of Experimental Biology* **209**: 3516-3528.

Jeanthon, C. (2000). Molecular ecology of hydrothermal vent microbial communities. *Antonie van Leeuwenhoek* **77**: 117-133.

Johnson, K.S., C.L. Beehler, C.M. Sakamoto-Arnold and J.J. Childress (1986). In situ measurements of chemical distributions in a deep sea hydrothermal vent field. Science **231**: 1139-1141.

Markert, S. C. Arndt, H. Felbeck, D. Becher, S.M. Sievert, M. Hugler, D. Albrecht, J. Robidart, S. Bench, R.A. Feldman, M. Hecker and T. Schweder (*in press*). Approaching the uncultured endosymbiont of *Riftia pachyptila* by physiological proteomics.

Reidi, J. and K. Klose (2002). *Vibrio cholerae* and cholera: out of the water and into the host. *FEMS Microbiology Reviews* **26**: 125-139.

Sambrook, J. and D.W. Russel (2001). Molecular Cloning A laboratory manual. Cold Spring Harbor Press: New York, USA.