Intra- and Inter-population diversity of the *Gammaproteobacteria* *Endorifita persephone* in vestimentiferan tubeworms from the eastern Pacific.

by

Maëva Perez
Bachelor of Science, Université de Montréal, 2011

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the School of Earth and Ocean Sciences

Supervisory Committee

Intra- and Inter-population diversity of the *Gammaproteobacteria* *Endorifita persephone* in vestimentiferan tubeworms from the eastern Pacific.

by

Maëva Perez
Bachelor of Science, Université de Montréal, 2011

Supervisory Committee

Dr. S. Kim Juniper, School of Earth and Ocean Sciences

Supervisor

Dr. Diana Varela, School of Earth and Ocean Sciences

Departmental Member

Dr. Francis Nano, Department of Microbiology

Outside Member

Dr. Réal Roy, Department of Biology

Outside Member

Abstract

Supervisory Committee

Dr. S. Kim Juniper, School of Earth and Ocean Sciences
Supervisor

Dr. Diana Varela, School of Earth and Ocean Sciences
Departmental Member

Dr. Francis Nano, Department of Microbiology
Outside Member

Dr. Réal Roy, Department of Biology
Outside Member

Vestimentiferan tubeworms of the eastern Pacific Ocean are often keystone species in vent communities. These polychaetes are host to intracellular *Gammaproteobacteria* symbionts. In this association, the siboglinid worms supply their symbionts with the compounds necessary to chemosynthesis while the sulfide oxidizing bacteria provide their host with the organic molecules necessary for their metabolism. The adult worms lack a digestive system and are therefore completely dependent on their symbionts for their nutrition. Given the obligate nature of the association for the host, it is surprising that the symbionts are not transmitted from parents to offspring but are acquired *de novo* from the environment at each generation. In other known cases of horizontally acquired mutualism (*e.g. Rhizobium*-legumes, dinoflagellates-corals), obtaining symbionts from the environment benefit the hosts by allowing for a degree of partner choice. According to the partner choice hypothesis, tubeworms that associate with the best-adapted partner(s) to a specific range of habitat conditions are in turn better adapted to this environment. Of course, this hypothesis assumes that there is diversity within the symbiotic partners. Phylogenetic analyses on the other hand seemed to indicate that nearly all species of vent tubeworms of the eastern Pacific were associated with the same species of symbionts: Candidatus *Endoriftia persephone.* However, these studies focussed on a few molecular markers. In this thesis, I used *in situ* hybridization and next generation sequencing to characterize the symbiont diversity at the species and strain level, as well as within individual hosts and across host species. I found that the intra-host symbiont populations are likely composed of multiple strains or lineages of the same bacterial species, that the symbiont populations separated by mid-ocean ridge discontinuities are vicariant, and that other factors such as local environmental conditions or host specificity might participate in shaping the genetic make-up of these populations.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would first like to express my deepest gratitude to my supervisor Kim Juniper for taking me into his lab, offering me the opportunity to embark on board the RV Thompson to collect vestimentiferan worms at the Endeavour vents, participate to the 14th Deep Sea Biology Symposium in Aveiro, Portugal, and for his extreme patience and precious help along my three year journey and particularly during the redaction of this thesis. I am also enormously thankful to the members of my committee; to Dr. Diana Varela for being amazing with her teaching assistants, and for bringing me to consider my data from the point of view of the free-living symbionts, to Dr. Francis Nano for his positive feedback and suggestions for improving the quality of my data, and to Dr. Réal Roy for his teachings on microbial ecology and for encouraging me to study other symbiosis models. I am especially indebted to Dr. Nathalie Forget for her contribution in the second chapter of my thesis, for sharing her pyrosequence libraries which greatly added to my third chapter, and for her precious advice and kind encouragements as I prepared for the conference in Aveiro. To Sheryl Murdock I am incredibly grateful for her expertise and assistance with all of the laboratory procedures I had to perform, for training me with the software mothur, and for getting me organized and ready for my first mission at sea. I also wish to thank the other members of my lab, Jessica Nephin and Catherine Stevens, as well as Dr. Verena Tunnicliffe, Jackson Chu, Jonathan Rose and all the other members of Dr. Tunnicliffe's and Dr. Dower's laboratories for their technical contributions, important criticisms, moral support, and for amazing me with their own research.

This research would have not been possible without the financial support provided by the Natural Sciences and Engineering Research Council and the computing support provided by WestGrid and Compute Canada. For the latter, I wish to thank Belaid Moa for patiently introducing me to bash script and teaching me to use UVic's computing facility while treating me to French yogurt and snacks. The new world of bioinformatics is vast and not easy to navigate. Thus, I am much obliged to all the contributors to online forums such as seqanswers.com and biostars.org for providing support to the bioinformatic noob that I once was and to the sources of free and ludic programing language tutorials such as Rosalind.com and Coursera.com. Last but not least, I want to give thanks to my family in Marseille, my family in Québec, and my adoptive family in Victoria. The Bottrell's welcomed me as one of their own the minute I moved on the island, they helped me finding a job, gave me a roof, fed me and keep on feeding me every Sunday. I praise their older son for calling my bluff across the Ocean, teaching me English, and keeping me warm for the past five years and for tomorrow.

Dedication

Je dédie cette thèse à mes parents Joëlle et Jean-Luc qui m'ont appris à être émerveillée par la nature et l'univers, m'ont ouvert au monde, et m'ont toujours soutenu moralement et financièrement tandis que je partais vivre loin d'eux.

Chapter 1. Introduction

Vestimentifera is a paraphyletic group of deep-sea tubeworms belonging to the family Siboglinidae (Annelida: Polychaete). It is comprised of nineteen species found worldwide at hydrothermal vents along mid-ocean ridges, transform faults, subduction zones and hot spots, and at a few cold seeps around continental margins (Figure 1.1). Vestimentiferans are characterized by the absence of a digestive tract and are estimated to have branched from a common ancestor during the Cenozoic; about fifty million years ago (Halanych *et al.*, 1998). The key to their success reside in the development of symbioses with chemosynthetic bacteria that provide their hosts with organic compounds synthesized from inorganic constituents (Scott *et al.*, 1998, 1999). The symbionts derive metabolic energy from reducing substances present in fluids discharging from hydrothermal vents and cold seeps. Hydrothermal vents are sites where geothermally-heated seawater is expelled from porous oceanic crust. As a result of the chemical exchanges between seawater and mafic rocks under conditions of high pressure and temperature, the fluids that discharge at hydrothermal vents are enriched in reducing substances that are used by chemolithoautotrophic microbes as a source of energy and reducing power to drive CO_2 fixation into organic molecules via chemosynthesis. Cold seeps, found on continental margins, are also chemosynthetic ecosystems fuelled by the weak discharge of hydrogen sulphide- or hydrocarbon-containing fluids through seafloor sediments.

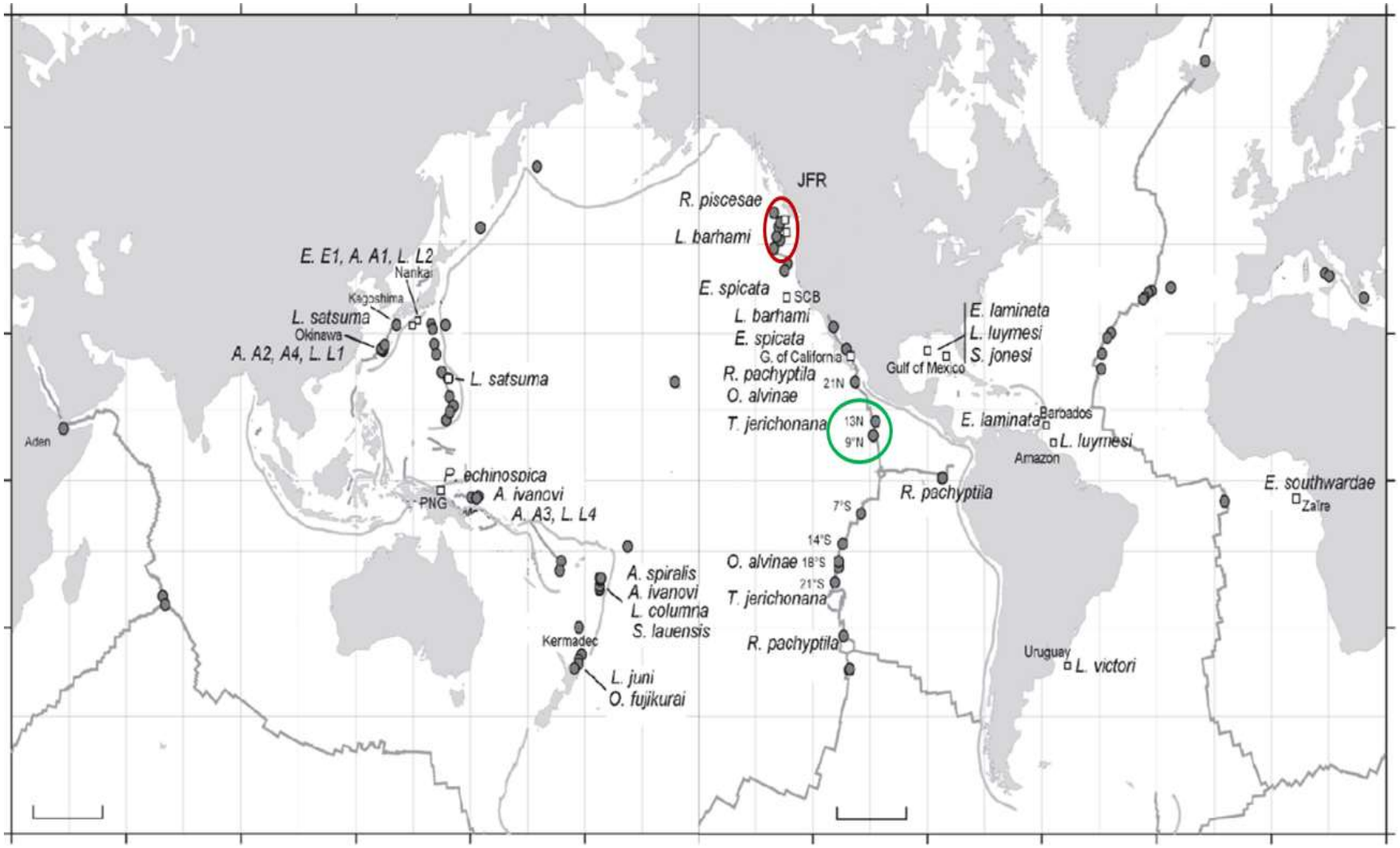

Figure 1.1 Geographic distribution of the 19 described species of Vestimentiferan tubeworms from hydrothermal vents (black circles) and cold seeps (white squares). This thesis covers symbionts in association with *R. piscesae* on the Juan de Fuca Ridge (circled in red) and symbionts in association with *R. pachyptila* and *T. jerichonana* from the 13°N and 9°N vents (circled in green) on the East Pacific Rise. From Bright and Lallier (2010). Scale=2000 km

Most species of vestimentiferan are found at vents along the mid-ocean ridges of the eastern Pacific Ocean (Figure 1.1). Their distribution is marked by discontinuities of mid-ocean spreading centre that have engendered a series of allopatric speciation events. For example, the sister species *Ridgeia piscesae* and *Oasisia alvinae* were estimated to have diverged from a common ancestral population following the interruption of the Farallon-Pacific Ridge about 28 million years ago (Chevaldonne *et al.*, 2002) (Figure 1.2). Later fragmentation of the Farallon plate into the Cocos and Nazca plates resulted in vicariant populations of the East Pacific Rise species (Plouviez *et al.*, 2009; Johnson *et al.*, 2006; Hurtado *et al.*, 2004).

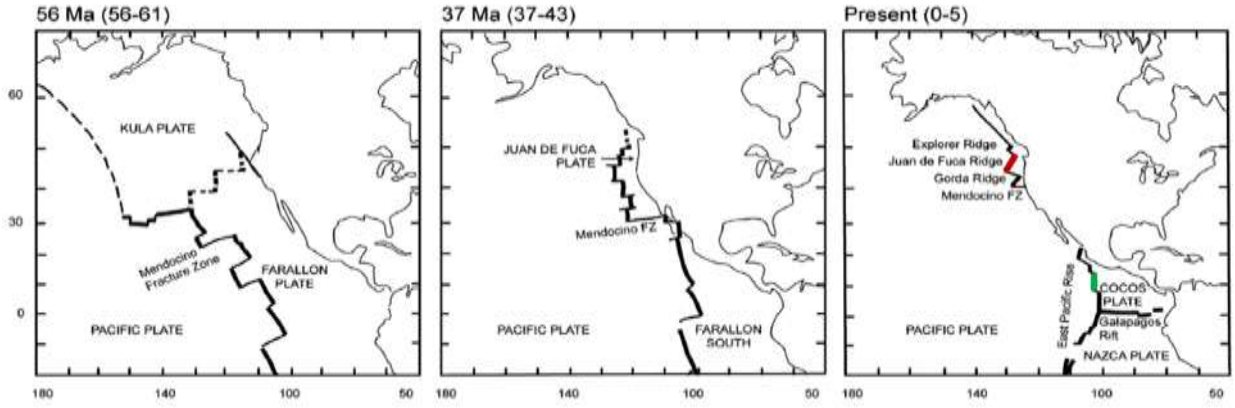

Figure 1.2 Geological history of the eastern Pacific mid-ocean ridge. Highlighted are the two regions studied in this thesis: the Juan de Fuca Ridge (red) and a section of the northern East Pacific Rise containing the 13°N and 9°N vent sites (green). From Vrijenhoek (2013).

Remarkably, up to six of these tubeworm species (*Riftia pachyptila*, *Tevnia jerichonana*, *Ridgeia piscesae*; plus possibly *Oasisia alvinae*, *Escarpia spicata*, and *Lamellibrachia* sp.) have all been found to host the same symbiont, the *Gammaproteobacteria* Candidatus *Endoriftia persephone* (short *Endoriftia*) (Di Meo *et al.*, 2000) (Figure 1.3). This contrasts with other endosymbioses in hydrothermal vent invertebrates such as vesicoymid clams which are characterized by host-symbiont cospeciation or bathymodiolin mussels which can harbor more than one symbiont phylotype (Vrijenhoek, 2010b).

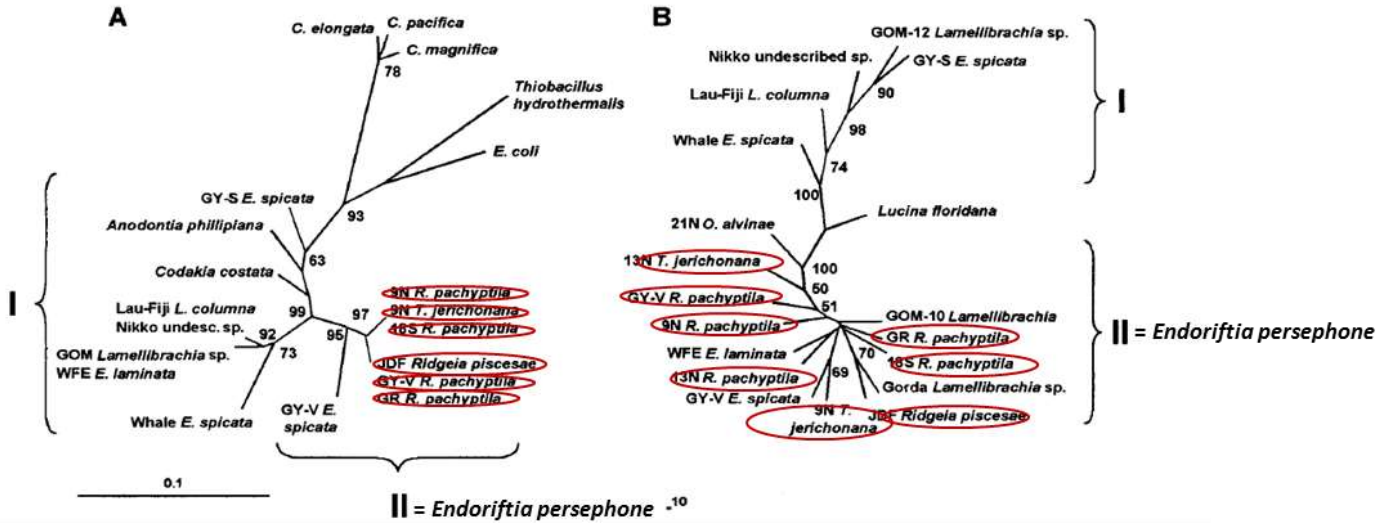

Figure 1.3 Phylogenetic trees of symbionts associated with different tubeworm species from the eastern Pacific Ocean based on A) the 23S rDNA and B) the internal transcribed spacer (ITS) sequences. The three species studied in this thesis are circled in red. From Di Meo *et al.* (2000).

1.1. The biology of the *Endoriftia*-vestimentifera holobiont

1.1.1. Trophosome structure and symbiostasis

The *Endoriftia* symbionts are hosted within specialized cells (bacteriocytes) contained within a host organ known as the trophosome that occupies most of the worm's coelomic cavity. In this mutualistic association, the worm supplies the bacteria with the inorganic compounds necessary for sulphide oxidation and CO_2 fixation: dioxygen, carbon dioxide and hydrogen sulphide. These substances diffuse across the gills into the blood of the animal and are then carried to the trophosome. In return, the endosymbionts provide the tubeworm with the organic molecules necessary for its metabolism and growth (Figure 1.4) either by excretion or by being directly digested (Felbeck and Jarchow, 1998; Bright *et al.*, 2000).

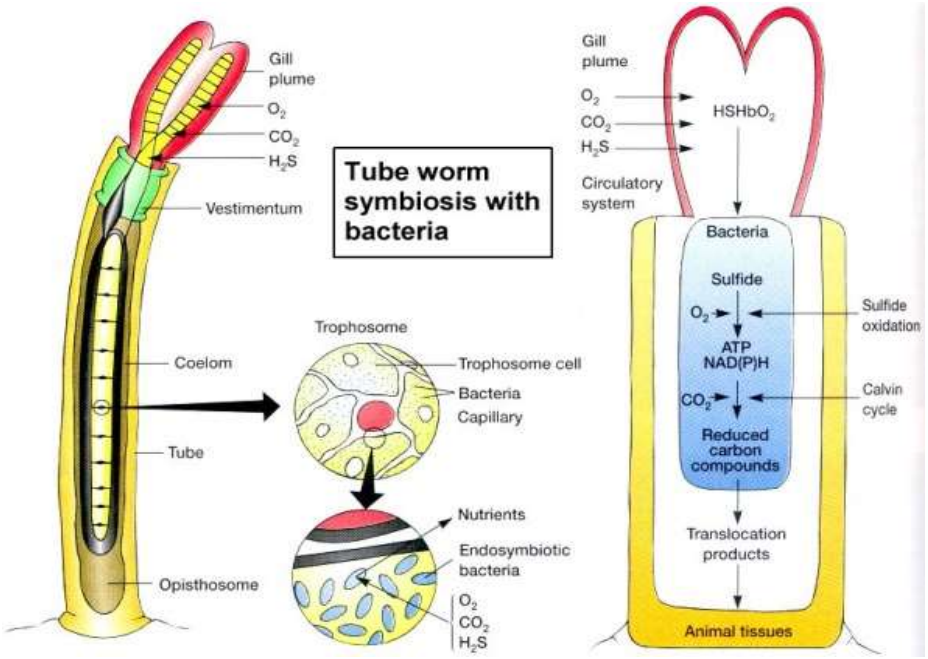

Figure 1.4 Schematic representation of the *Endoriftia*-vestimentifera holobiont metabolism. From Prescott *et al.* (2003).

In the tubeworms of the eastern Pacific, there is evidence that the proliferation of *Endoriftia* is highly controlled by the immune system of the worm (Pflugfelder

et al., 2009; Bunce, 2013; Klose *et al.*, 2016) and it has been suggested that the host and its symbionts are engaged in a continuous molecular dialogue involving Microbial Associated Molecular Patterns (MAMPs) and Pattern Recognition Receptors (PRRs) (Nyholm *et al.*, 2012).

1.1.2. Symbiont metabolism

Because the symbionts have yet to be successfully cultivated, most of what we know of their metabolism comes from experiments performed on freshly isolated and purified or preserved material; mostly from *Riftia pachyptila* endosymbionts (Figure 1.5).

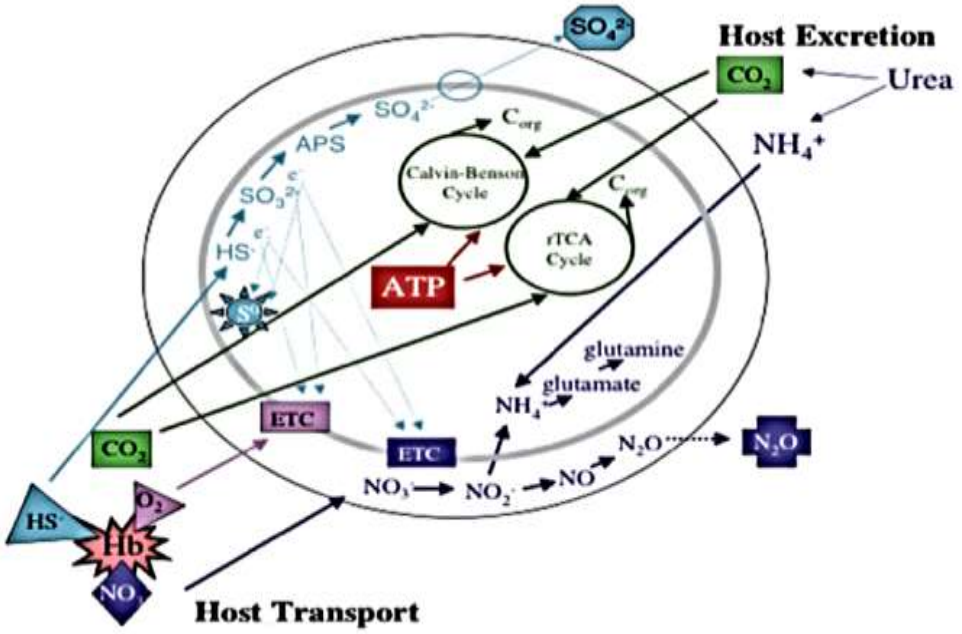

Figure 1.5 Metabolism of *Riftia pachyptila* endosymbionts. ETC: Electron transport chain. From Robidart *et al.* (2008).

These experiments have shown that the lithoautotrophic symbionts use the energy released during sulphide (HS^-) oxidation for fixing molecular carbon dioxide ($C0_2$) and unlike other symbiotic or free-living bacteria, they can only metabolise sulphide and not thiosulfate (Wilmot and Vetter, 1990). H_2S oxidation to elemental sulfur and sulfate ($S0_4^{2-}$) is metabolised through a pathway involving the following enzymes: the dissimilatory sulphite reductase

(DsrA), the APS reductase (AprA/AprB) and ATP-sulfurylase (SopT). These enzymes can represent up to 12% of the cytosolic proteome (Markert *et al.*, 2011). The stored elemental sulfur could also be used as an electron sink when oxygen is absent (Arndt *et al.*, 2001). To fix carbon dioxide, vent tubeworm endosymbionts can use two pathways: the Calvin-Benson-Bassham cycle (Elsaied *et al.*, 2002) and the reductive tricarboxylic acid (rTCA) cycle (Thiel *et al.*, 2012). Finally, the symbionts assimilate nitrogen from ammonia and recent studies have shown that they can perform nitrate respiration (Hentschel and Felbeck, 1993; Gardebrecht *et al.*, 2011; Liao *et al.*, 2013).

The sequencing of a near-complete genome of *Endoriftia* additionally revealed many genes involved in motility, chemotaxis, and defense mechanisms (Robidart *et al.*, 2008). These genes are probably expressed in the free-living form of the symbionts (see description of symbiont lifecycle in following section).

1.1.3. Life cycle

The symbiotic bacteria are horizontally transmitted, that is to say, acquired *de novo* from the surrounding environment at each generation. Indeed, they present no genomic reduction and no signs of coevolution with their host (McMullin *et al.*, 2003; Vrijenhoek, 2010a; Nelson and Fisher, 2000). Moreover, free-living symbionts have been found in basalts surrounding tubeworm aggregations (Harmer *et al.*, 2008) and a recent study demonstrated that intracellular symbionts can escape the tissues of a dead tubeworm host and potentially return to a free-living state (Klose *et al.*, 2015).

The symbiont worm hosts possess two separate sexes: male and female. Following an anisogamic reproduction via internal fertilization (Macdonald *et al.*, 2002; Southward and Coates, 1989), the embryos are released in the water. They can live a few weeks (about 38 days for *Riftia pachyptila*) in the water column and disperse passively via deep-ocean currents (Young *et al.*, 2008; Mullineaux *et al.*, 2002; Marsh *et al.*, 2001). When the trochophore larvae settle, they develop a mouth and digestive system (metatrochophore), and ingest bacteria and diatoms (Nussbaumer *et al.*, 2006). Shortly after, symbionts contact and penetrate the worm tissues through the skin and migrate to a region between the dorsal blood vessel and the foregut to form the proto-trophosome. As the metatrochophore larvae develop into adults, their digestive tract atrophies in favour of the trophosome that ends up occupying most of the space in the cœlomic cavity of the trunk (Nussbaumer *et al.*, 2006). The vestimentiferan adults thus become completely dependent on their bacteria for nutrition (Figure 1.6).

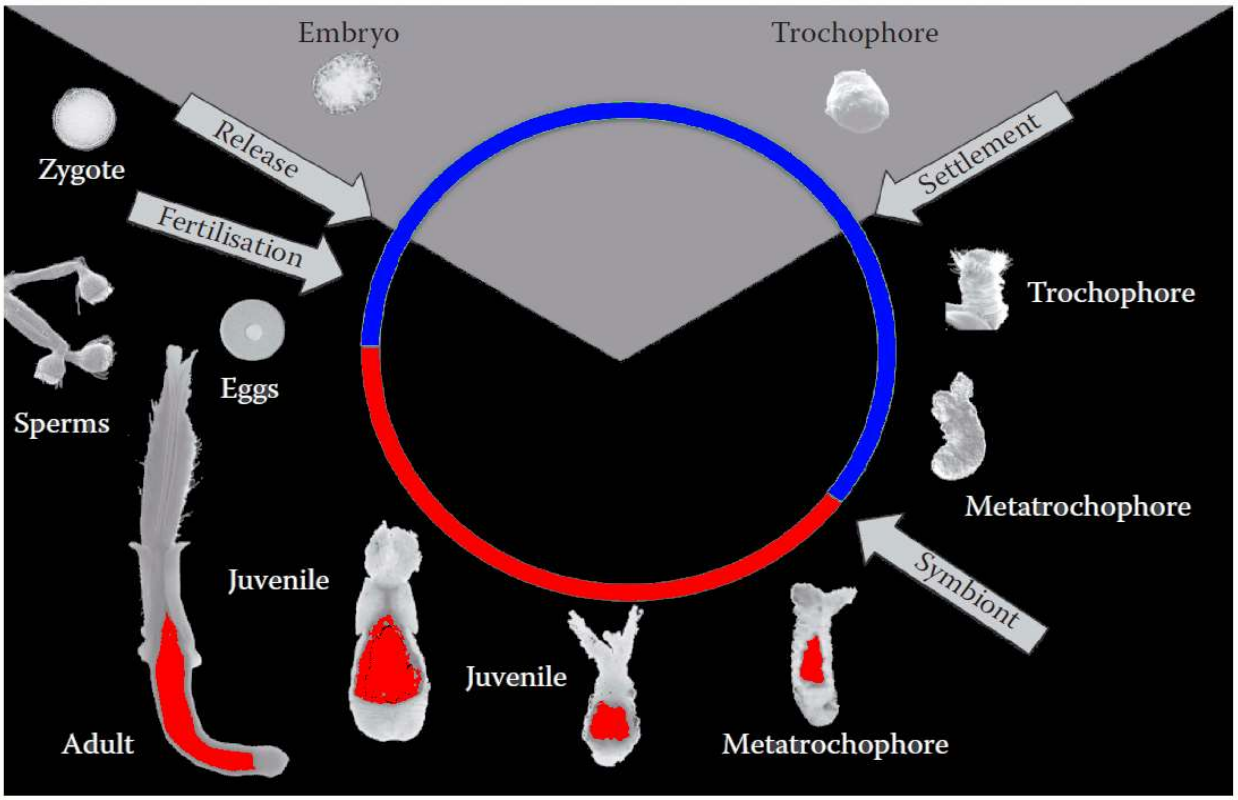

Figure 1.6 Life cycle and symbionts acquisition of *Riftia pachyptila*. Blue : aposymbiotic phase; Red symbiotic phase. Adapted from Bright and Lallier (2010).

1.2. General problematic: the enigma of horizontally acquired mutualism

If symbionts are so essential for the survival of the adult worms, why aren't they vertically transmitted? Why does each generation of worms risk not finding their symbiotic partner?

This contradiction underscores questions about the evolutionary origins, and advantages of horizontally acquired mutualism. It has been suggested that mutualistic symbiosis can evolve without vertical transmission in cases where: (1) vertical transmission involves a high cost for the host, (2) the symbionts suffer direct negative consequences if they exploit the host too intensively (Tit-For-Tat strategy), (3) the dispersal of both host offspring and symbionts is local (pseudo-vertical transmission), and (4) it facilitates spatial and temporal adaptation of the host by allowing a degree of partner choice (Genkai-Kato and Yamamura, 1999; Wilkinson and Sherratt, 2001; Sachs *et al.*, 2011).

Testing the first and second hypotheses would necessitate experimental manipulations along with measures of both symbiont and host fitness. To date, neither the symbionts nor the hosts have successfully been kept alive for more than a few days in the laboratory. The Klose *et al.* (2015) finding that the symbionts can (and do) escape their dead hosts supports the third hypothesis but evidence that released symbionts enrich or even contribute to local free-living populations has yet to be obtained. Finally, the last hypothesis seems particularly relevant at vents because the physico-chemical conditions can vary considerably both in space (close to or away from venting) and time (fluctuations in hydrothermal discharge).

1.2.1. Partner choice for eastern Pacific vent tubeworms?

The horizontal acquisition of endosymbionts could facilitate the worms' adaptation to its environment if it meant they could acquire the best-adapted partner(s) to a specific range of habitat conditions (Wilkinson and Sherratt, 2001). Because of their extreme physico-chemical conditions we expect hydrothermal vents to be environments of high selective pressure. Since bacteria reproduce asexually and are prone to horizontal transfer, a free-living population of potential symbionts should consistently reach maximal fitness faster than their eventual eukaryote hosts. Since the symbionts are transferred horizontally, tubeworm larvae should therefore acquire symbionts from a free-living pool that is best adapted to the environmental conditions in which the worms have settled. The worms could also actively select certain bacterial phenotypes depending on specific compatibility and environmentally-driven metabolic needs. Figure 1.7 illustrates this idea in the case of one (A) or several symbionts (B).

1.2.1. Study questions

To investigate the potential for partner choice in facilitating vestimentiferan adaptation in variable environmental conditions, we first need to we first need to confirm that symbionts within host populations are genetically diverse enough to suggest the existence of multiple species, strains or lineages that could have metabolic differences (Figure 1.8).

In this context, my thesis focusses on two main questions; how diverse are the symbionts within the trophosome, and how diverse are they at large geographical scales and across host species?

In the second and third chapters, I explored symbiont diversity at the species and strain level, respectively within individual host worms. In Chapter 4, I aimed at characterizing inter-specific symbiont diversity by comparing the genetic diversity across five symbiont populations associated with three different species of worms.

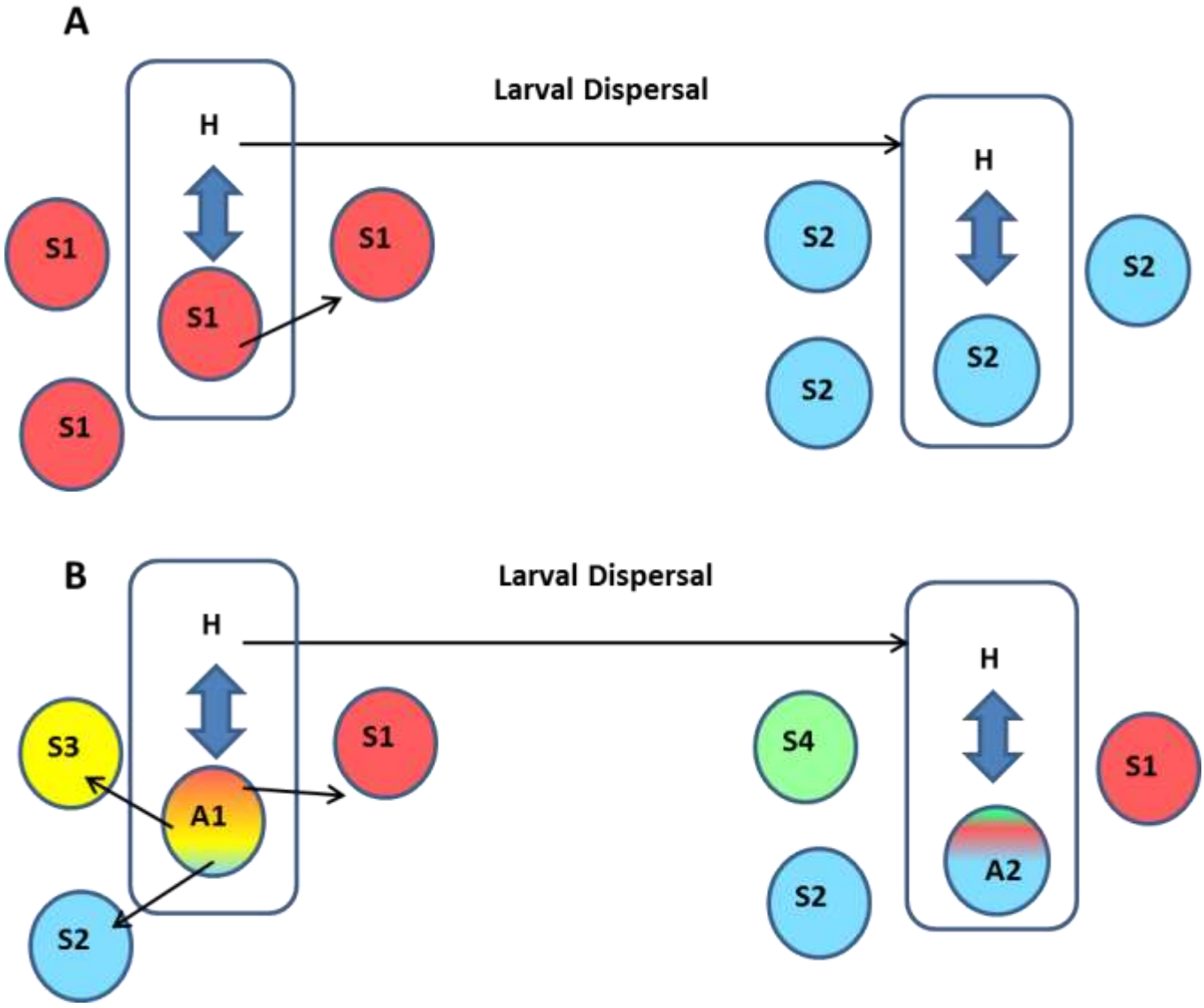

Figure 1.7 Schematic summary of the partner choice hypothesis in the case of A) one symbiotic partner or B) multiple partners. H=host, S=symbionts, A=assemblage of symbionts. The double arrows inside the host show the exchange of benefits between host and symbionts. The arrows extending outside the host show the transmission of host and symbionts to the next generation. The different colors associated with the symbionts in B illustrate that not only the quality but also the relative abundances of each symbiont are prone to change. On the left and the right, are two different environments with different selective pressures so that the free living S1 and S2 (or S1, 2, 3 and S1, 2, 4) are two populations of free-living symbionts that have reached maximal fitness with different sets of alleles and/or different allele frequencies. By associating with symbionts from the surrounding environment, a tubeworm host can associate with the 'best'/fittest partner(s).

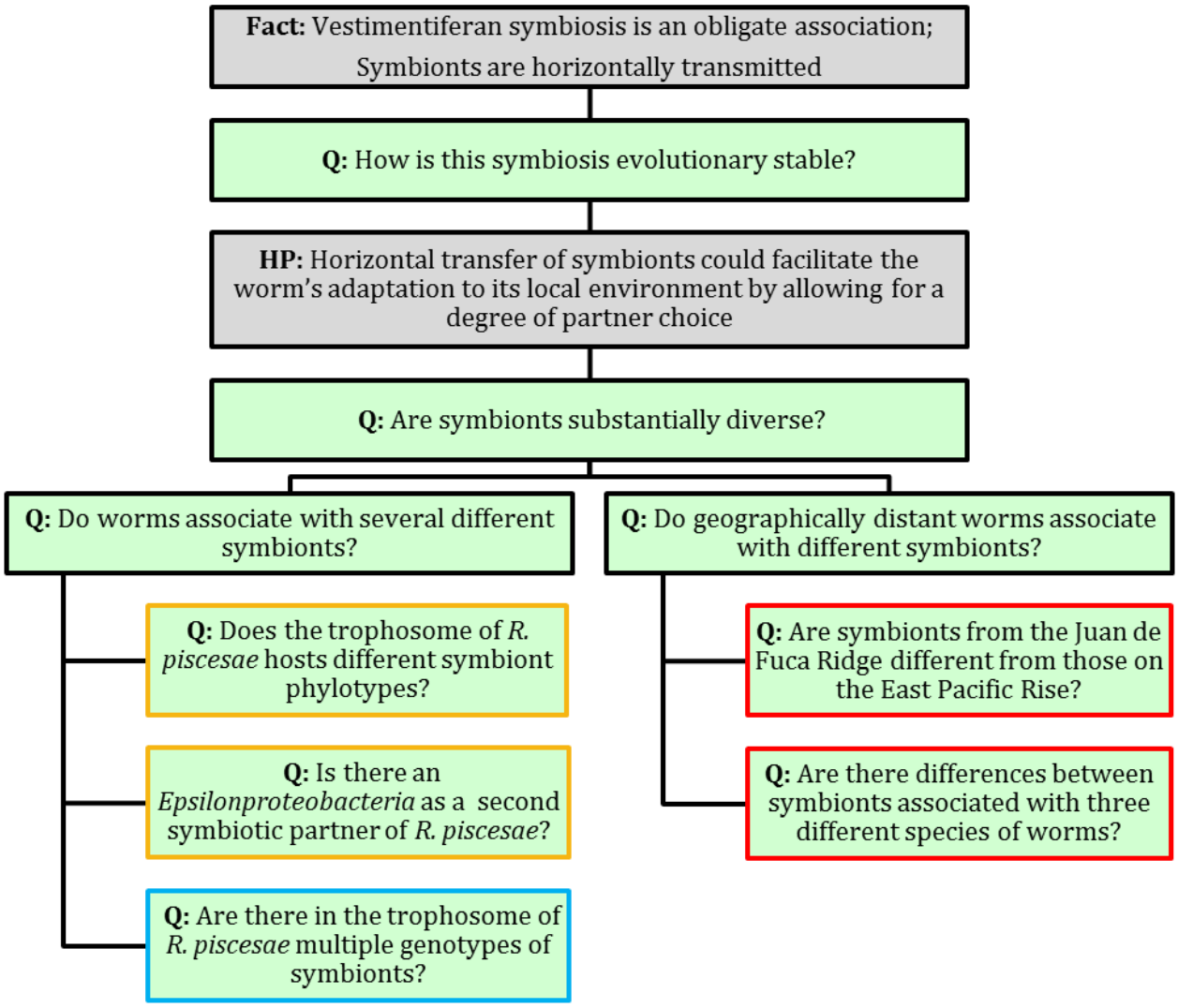

Figure 1.8 General problematic and study questions. Questions pertaining to Chapters 2, 3, and 4 are outlined in orange, blue and red, respectively.

1.2.2. Our model host species for investigating intra-individual diversity: *R. piscesae*

The species *Ridgeia piscesae* is found at hydrothermal vents of the North East Pacific Ocean (Explorer Ridge, Juan de Fuca Ridge, and Gorda Ridge). *Ridgeia piscesae* tolerates a wider range of environmental conditions than any other vestimentiferan known to date (Bright and Lallier, 2010). Its depth distribution extends from 1550m to 3220m depth (Young *et al.*, 2008).

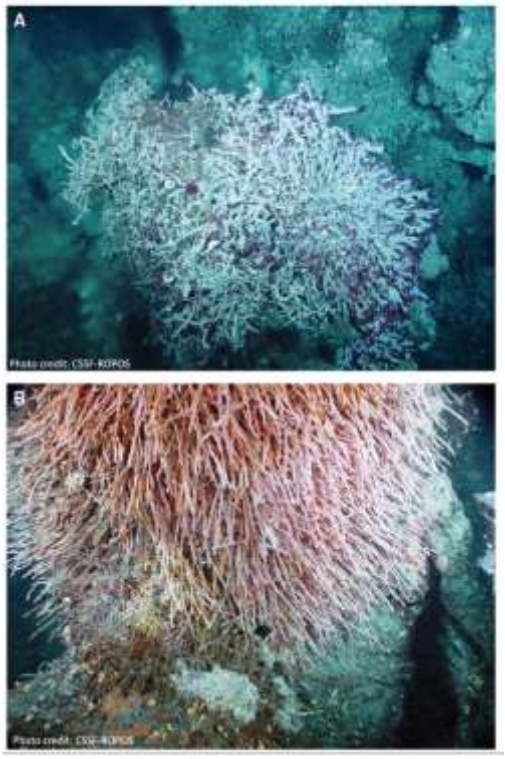

Figure 1.9 *Ridgeia piscesae* phenotypic plasticity; A) Short-fat morphotype in High-Flow environments, B) Long-skinny morphotype in Low-Flow environment. Note hydrothermal fluid causing shimmering in A. From Forget and Juniper (2013).

Temperatures within a tubeworm aggregation can vary from ambient (2°C) seawater to 30°C (Carney *et al.*, 2007), with up to 20°C difference between the base and the gill level of the worms (Urcuyo *et al.*, 2003). Sulphide concentrations at their branchial plumes range from <0.1µM to 200µM (Carney *et al.*, 2002; Urcuyo *et al.*, 2003; Brand *et al.*, 2007).

This resilience is associated with a great degree of phenotypic plasticity that has only been observed in this species. These phenotypic differences resulted in *R.piscesae* phenotypes previously being described as distinct species (Southward *et al.*, 1995; Malakhov *et al.*, 1996). Phenotypes range from short individuals (up to 20cm) with wide white tubes and well-developed, bright red gills (Short-fat morphotype, Figure 1.9, A) in habitats of intense hydrothermal fluid discharge (High-Flow), to long (>1m) thin worms with rusty-coloured narrow tubes and reduced branchial plumes (Long-skinny morphotype, Figure 1.9 B) in habitats of weak hydrothermal discharge (Low-Flow) (Forget and Juniper, 2013).

Comparative studies have documented different life strategies associated with each morphotype. Low-Flow worms tend to have slow growth, low mortality and low reproductive potential and body condition while High-Flow worms are

short-lived with high reproductive potential and body condition (Urcuyo *et al.*, 2003, 2007, 1998; Macdonald *et al.*, 2002; Tunnicliffe *et al.*, 2014). The two morphotypes also exhibit metabolic differences. The relative concentrations of thiotaurine, a potential intracellular transporter of sulphide, and extracellular hemoglobin, and globin gene expression have been found to be consistently higher in Low-Flow worms (Brand *et al.*, 2007; Carney *et al.*, 2007). On the other hand, it has been suggested that aggregations in Low-Flow environments are able to take up hydrogen sulphide via a "rootball" structure at their bases rather than through their gills (Urcuyo *et al.*, 2003, 2007).

1.3. Study questions and methods: Chapter 2

Questions: Does the trophosome of *R. piscesae* host multiple phylotypes of symbionts?

Is there an Epsilonproteobacteria as a second symbiotic partner?

Methods: CARD-FISH, pyrosequencing, phylogenetic analyses via a bioinformatic pipeline.

1.3.1. How multiple symbiosis could explain *R. piscesae* success

Although a significant amount of work has been carried on the phenotypic and metabolic versatility of *Ridgeia piscesae* (Urcuyo *et al.*, 1998, 2003, 2007; Carney *et al.*, 2007; Nyholm *et al.*, 2008, 2012; Brand *et al.*, 2007), few studies have considered the symbionts from the point of view of their contribution to the worm's success in this broad range of physico-chemical conditions (deBurgh *et al.*, 1989; Chao *et al.*, 2007).

In other symbioses between bacteria and deep-sea metazoans, the host adaptation to environmental variability can involve symbiosis with more than one group of bacteria, broadening the metabolic potential of the host. For example, the gutless marine worm *Olavius algarvensis* that inhabits the oxic-anoxic interfaces in Mediterranean Sea sediments harbours four different

extracellular symbionts under its cuticle: two *Deltaproteobacteria* and two *Gammaproteobacteria*. Metagenomic, metaproteomic and metabolomic studies have shown that the worm can acquire organic carbon via hydrogen (δ-symbionts) or reduced sulfur compound oxidation, or by heterotrophy (γ-symbionts). The symbionts can also recycle the worm's waste ultimately replacing its excretory organ (nephridium) (Woyke *et al.*, 2006; Kleiner *et al.*, 2012b). Another example is the hydrothermal vent mussel *Bathymodiolus puteoserpentis* that was found hosting two *Gammaproteobacteria* endosymbionts within its gill tissues: one sulphide-oxidizing and one methane-oxidizing. The relative abundance of each partner depended on vent fluid chemistry: methanotrophs were more abundant in habitats with high methane concentrations in venting fluids (Duperron *et al.*, 2006).

1.3.2. Potential for multiple symbionts in *R. piscesae*

While the roles of the different symbionts remain unclear in most cases, symbioses of metazoans with more than one group of bacteria have been described in several other invertebrates collected from deep-sea sediment (Edgcomb *et al.*, 2011), cold seep (Duperron *et al.*, 2005, 2008, 2009; Fujiwara *et al.*, 2001; Kimura *et al.*, 2003), and hydrothermal vent habitats (Petersen *et al.*, 2010; Grzymski *et al.*, 2008). More recently, Zimmermann *et al.* (2014) were the first to report two closely related but distinct phylotypes of *Gammaproteobacteria* in *Lamellibrachia anaximandri*, a vestimentiferan from the Mediterranean Sea. Their study does not however address whether the two bacteria had different metabolisms.

Given what has been observed in other hydrothermal vent invertebrates, and more recently in the Mediterranean tubeworm (Zimmermann *et al.*, 2014), it is reasonable to propose that the highly plastic species *Ridgeia piscesae* also hosts multiple symbionts. Chao *et al.* (2007) reported *Ridgeia piscesae* to host up to five operational taxonomic units (OTUs) belonging to the *Gammaproteobacteria*, *Alphaproteobacteria*, and Cytophaga-Flavobacterium-Bacteroidetes groups, based on terminal-restriction fragment length polymorphisms (tRFLP). Nathalie Forget, a former PhD. student in our laboratory, recovered *Gamma-* and *Epsilon-*proteobacteria from pyrosequencing data of trophosome homogenates (Forget *et al.*, 2014). Because the large taxonomic divergence between *R. piscesae*'s putative multiple symbionts could suggest different metabolic processes, we proposed the hypothesis that *R. piscesae* had an *Epsilonproteobacteria* as a second symbiotic partner.

1.3.3. Chapter structure and contributions

The results of Dr. Forget's pyrosequencing analyses of *R. piscesae* symbionts along with my CARD-FISH assays were jointly published in Marine Ecology in an article entitled "Molecular study of bacterial diversity within the trophosome of the vestimentiferan tubeworm *Ridgeia piscesae*" (Forget *et al.*, 2014). This article constitutes the first part of the second chapter of my thesis. The second part of this chapter presents a supplemental dataset of pyrosequences data from trophosome homogenates that I collected in 2013. These samples were preprocessed differently in order to reduce the risk of contamination, then analysed using the same bioinformatic pipeline as in Forget *et al.* (2015). Biases introduced by the bioinformatic pipeline are discussed.

1.3.4. Methods

To test whether an *Epsilonproteobacteria* symbiont was present in the trophosome of *R. piscesae*, I used a combination of histological (CARD-FISH) and high throughput sequencing (Pyrosequencing) analyses. These two methods are briefly described below.

1.3.4.1. CARD-FISH

Catalysis Reporter Deposition Fluorescent *In situ* Hybridization (CARD-FISH) is a technique of selective histological fluorescent staining based on the specific hybridization between a DNA probe and its complementary rRNA sequence in the target cells' ribosomes. Figure 1.10 depicts the principle of CARD-FISH staining. The DNA probe consists of a short (~50 bp), discriminatory 16S or 23S rDNA sequence with a horseradish peroxidase enzyme attached (HRP-probe). The horseradish peroxidase catalyses the accretion of fluorescein tyramine which results in an amplified fluorescent signal.

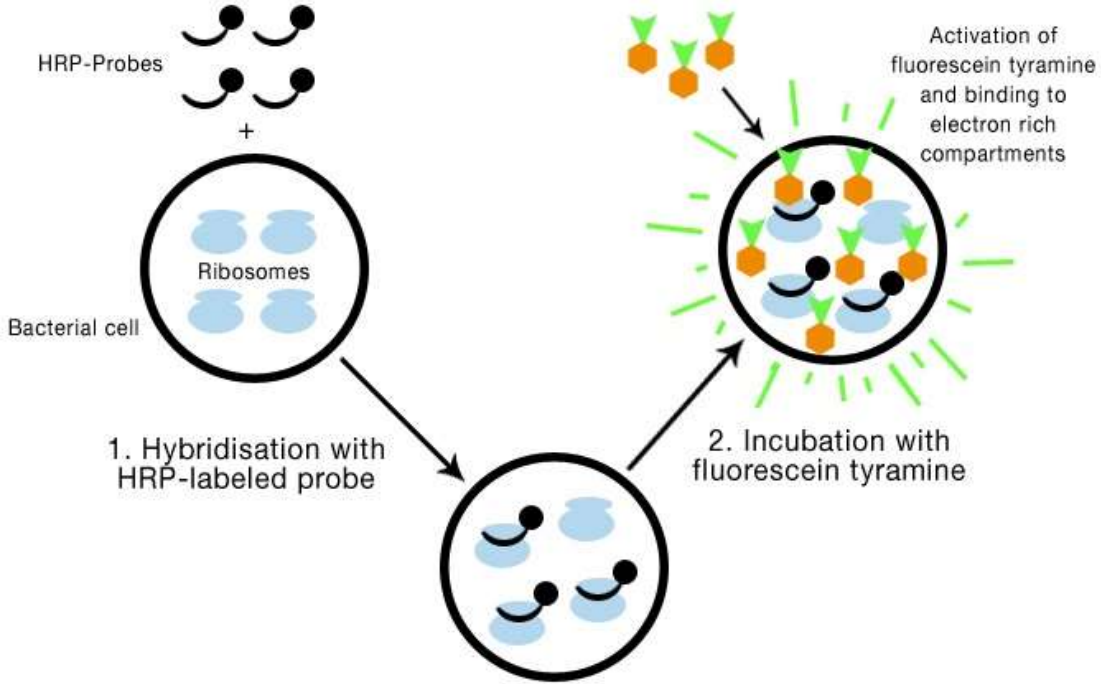

Figure 1.10 Catalysis Reporter Deposition Fluorescent *In situ* Hybridization (CARD-FISH). From http://www.arb-silva.de/fish-probes/fish-protocols/.

1.3.4.2. Pyrosequencing

Pyrosequencing is a "sequencing by synthesis" technique allowing for independent sequencing of individual strands of DNA (Figure 1.11). Because of this, pyrosequencing permits ultra-deep sequencing of bacterial populations with minimal preprocessing steps that bring biases. Unfortunately, this method limits the length of individual reads of DNA to about 500 bp, so targeted loci must first be amplified using Polymerase Chain Reaction (PCR). On the 454/Roche sequencing platform (platform used for all pyrosequencing analyses in this thesis), the amplified sequences are loaded onto microbeads, one sequence per bead. Using emulsion PCR, the sequences are independently amplified on each bead in order to increase the electromagnetic signal in the subsequent step. Finally, the beads are loaded into wells (one per well) aligned with electromagnetic receptors that detect the light emission patterns emitted during the sequencing by synthesis step.

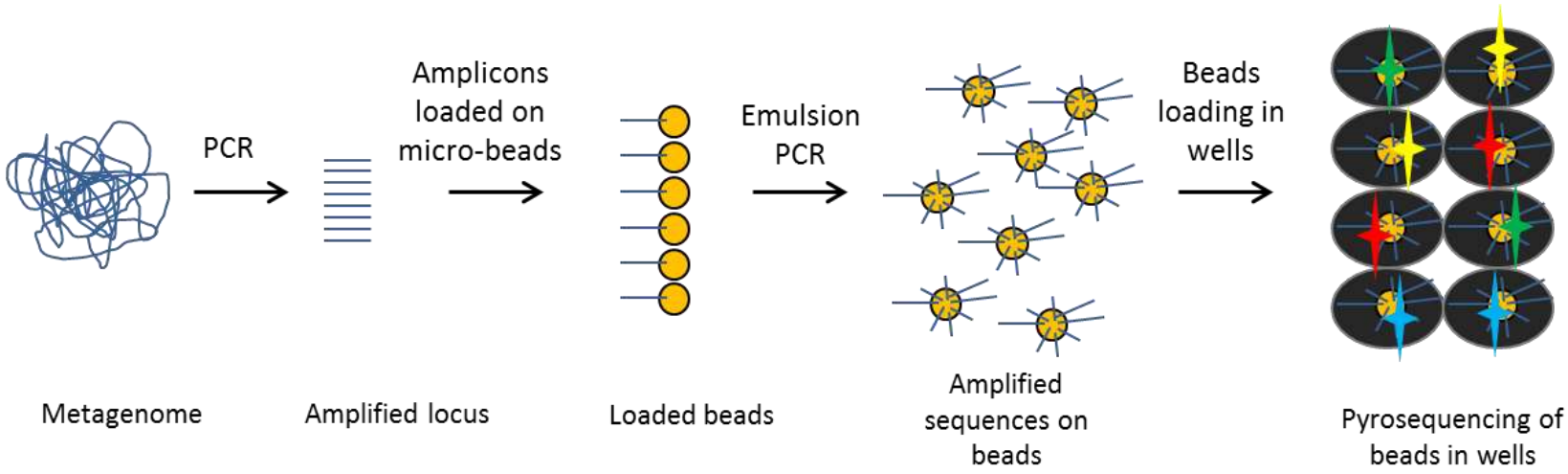

Figure 1.11 Simplified 454/Roche pyrosequencing workflow. See https://youtu.be/rsJoG-AulNE for a great animation of how 454/Roche sequencing works.

1.4. Study questions and methods: Chapter 3

Question: Does the trophosome of *R. piscesae* contain multiple genotypes of symbionts?

Methods: Whole genome shotgun sequencing, pyrosequencing, genetic variant detection, CRISPR typing.

In the pyrosequencing data from the 2013 worms, I found only *Gammaproteobacteria*. Interestingly, the bioinformatic pipeline I used seemed unable to resolve the taxonomy of the symbionts past the Class level; when aligned together, the symbiont sequences seemed to cluster in two phylogenetic groups that did not match the taxonomic affiliations outputted by my pipeline. This result could be interpreted as evidence that the trophosome of *R. piscesae* was not monoclonal but host to several genotypes or strains of *Gammaproteobacteria* symbionts.

To test this hypothesis, I used two different approaches: CRISPR typing and detection of genetic variants (for description of terms and concepts pertaining chapter 3, see Table B.1 as well as Figure B.1 in Appendix B; p. B.1).

1.4.1. CRISPR-typing

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) locus constitutes the adaptive immune system of Prokaryotes. It is composed of the Cas operon and the CRISPR array. The Cas operon contains genes responsible for editing the CRISPR array as well as the anti-viral function described below. The CRISPR array consists of short sequences (~40 bp) that are complements to sequences in phage nucleic acids (CRISPR spacers) in between short sequences of palindromic repeats (CRISPR repeats) that are species specific (Kunin *et al.*, 2007) (Figure 1.12). When a cell is infected by a new virus, the Cas operon is activated. Some of the Cas proteins can then copy a sample of the virus' genetic material and insert it into the CRISPR array in between repeats. If this cellular

lineage encounters the same virus again, other Cas proteins can copy the particular CRISPR spacer and produce a small RNA sequence that, together with Cas proteins, will bind to the virus nucleic acids, leading to their degradation and altering the phage virulence (Sorek *et al.*, 2008). Because CRISPR spacers accumulate in the CRISPR array, they constitute a historical record of the viral encounters of a particular lineage and thus can be used for short term strain typing. In the third chapter, I use this principle to detect multiple genotypes within two symbiont populations by determining whether or not there are 'chromosomic rearrangements' between the spacers of a particular CRISPR array.

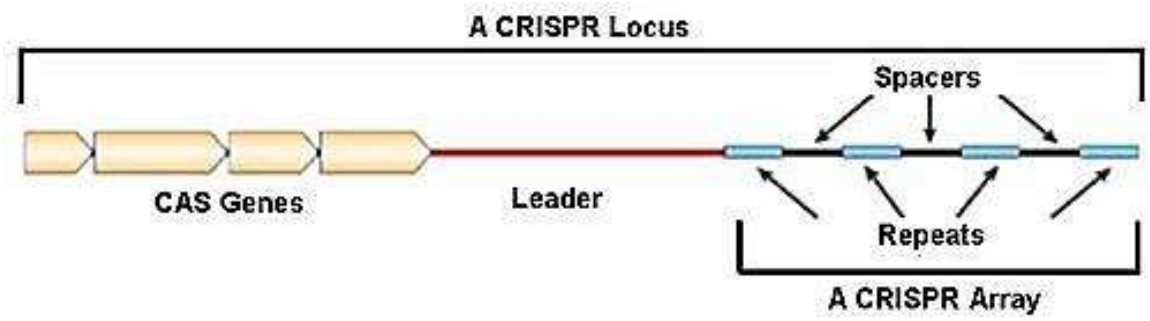

Figure 1.12 Model of a CRISPR locus. From Sorek *et al.*, 2008

1.4.2. Detection of genetic variants

Most of the genetic diversity is generated through mutations that result in insertions, deletions and substitutions of (most often) single nucleotides. These genetic variants can be detected by aligning mutated sequences to a reference sequence and detecting positions with heterogeneous nucleotide bases. Yet, variant calling methods are still somewhat unreliable and important disparities exists between algorithms (Mielczarek and Szyda, 2015; Huang *et al.*, 2015; Cheng *et al.*, 2014; Yu *et al.*, 2012; O'Rawe *et al.*, 2013).

In the third chapter, I attempted to detect genotypic variants (variants from distinct bacterial lineage) in two whole genome shotgun metagenomes (see description of shotgun sequencing in Section 1.5.1; p. 24). To do so, I used a bioinformatic pipeline that minimizes the number of false-positive variants and used different variant calling algorithms. To detect genetic polymorphism in pyrosequencing data, I developed a whole new bioinformatic pipeline that corrects for the contamination from non-symbiont DNA and the standard sequencing errors of the 454/Roche technology.

1.5. Study questions and methods: Chapter 4

Questions: Are the symbionts from the Juan de Fuca Ridge different from those on the East Pacific Rise? Are there differences between symbionts associated with three different species of worms?

Model: Eastern pacific tubeworm symbionts

Methods: Whole genome shotgun sequencing, genome-wide comparisons.

Four near-complete genomes of *Endorifita* in association with two tubeworm species from the East Pacific Rise *Riftia pachyptila* (3 genomes) and *Tevnia jerichonana* (1 genome), have previously been published by Robidart *et al.* (2008) and Gardebrecht *et al.* (2011). These fragmented genome assemblies were each reconstructed from shotgun sequences of symbiont genetic material found in the trophosome of individual worms. They thus represent consensus genomes of potentially heterogeneous intra-individual populations of symbionts. In Chapter 4, I used bioinformatic tools to reconstruct near-complete genomes for two *Endoriftia* populations from one, and 5 individual *R. piscesae* worms on the Juan de Fuca Ridge, respectively. These consensus genomes were assembled from whole genome shotgun sequences generated from DNA extract of worms' trunk tissues. A general description of whole genome shotgun sequencing and a description of the bioinformatic assembly are provided below.

1.5.1. Whole genome shotgun sequencing

The whole genome shotgun sequencing workflow is comprised of three mains steps (Figure 1.13). First, the extracted DNA is randomly broken down in smaller fragments of a given size. This can be done physically using high frequency acoustic waves, or chemically with enzymes. Then the short reads are sequenced using high throughput sequencing technology such as Illumina HiSeq, Illumina MiSeq or 454/Roche. *Ridgeia* symbiont whole genome shotgun sequences were obtained with Illumina HiSeq 2000 and MiSeq sequencers. A good animation of Illumina sequencing technology resulting in paired-end reads can be found here: https://youtu.be/womKfikWlxM. Finally, partially overlapping reads are assembled *in silico* using the algorithm described in the next section.

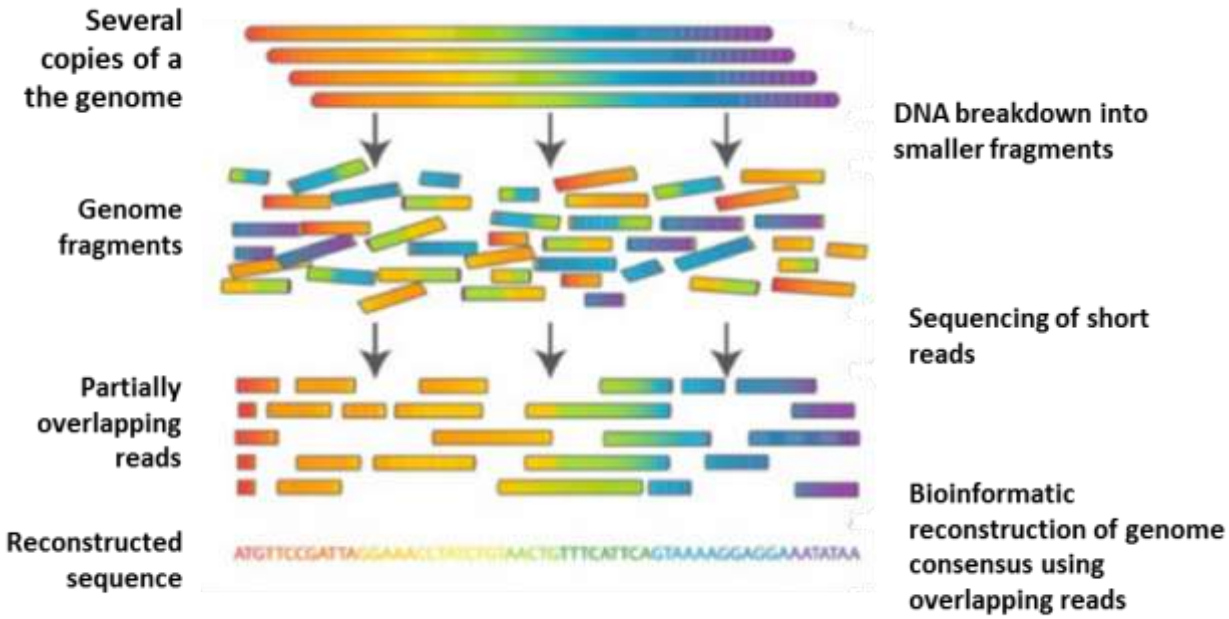

Figure 1.13 Schematic workflow of whole genome shotgun sequencing. From Commins *et al.*, 2009.

1.5.1. Genome assembly

Reconstructing a sequence from short overlapping reads is like solving a puzzle with partially overlapping pieces. Such puzzles can be solved computationally in three mains steps (Figure 1.14). First, overlaps are identified. Then, these overlaps are graphically represented as nodes connected by the reads into what is called a de Bruijn graph. Reassembling the sequence is then equivalent to finding the path connecting the overlaps that only goes through each read once.

While the work flow presented in Figure 1.14 is true in principle, modern assembly algorithms do not directly find overlaps between reads but work by decomposing each sequenced read into a series of smaller sequences (k-mers) of size k that are overlapping by k-1 nucleotides. These k-1 overlaps are then used as nodes connected in de Bruijn graphs in which edges are represented by k-mers (Compeau *et al.*, 2011). For more details, a very thorough and pedagogical explanation of how a modern assembly algorithm works can be found in the Educational videos associated with Chapter 3 of the book Bioinformatics Algorithms: An Active Learning Approach (Compeau and Pevzner, 2014) available at https://www.youtube.com/playlist?list=PLQ-85lQlPqFNGdaeGpV8dPEeSm3AChb6L.

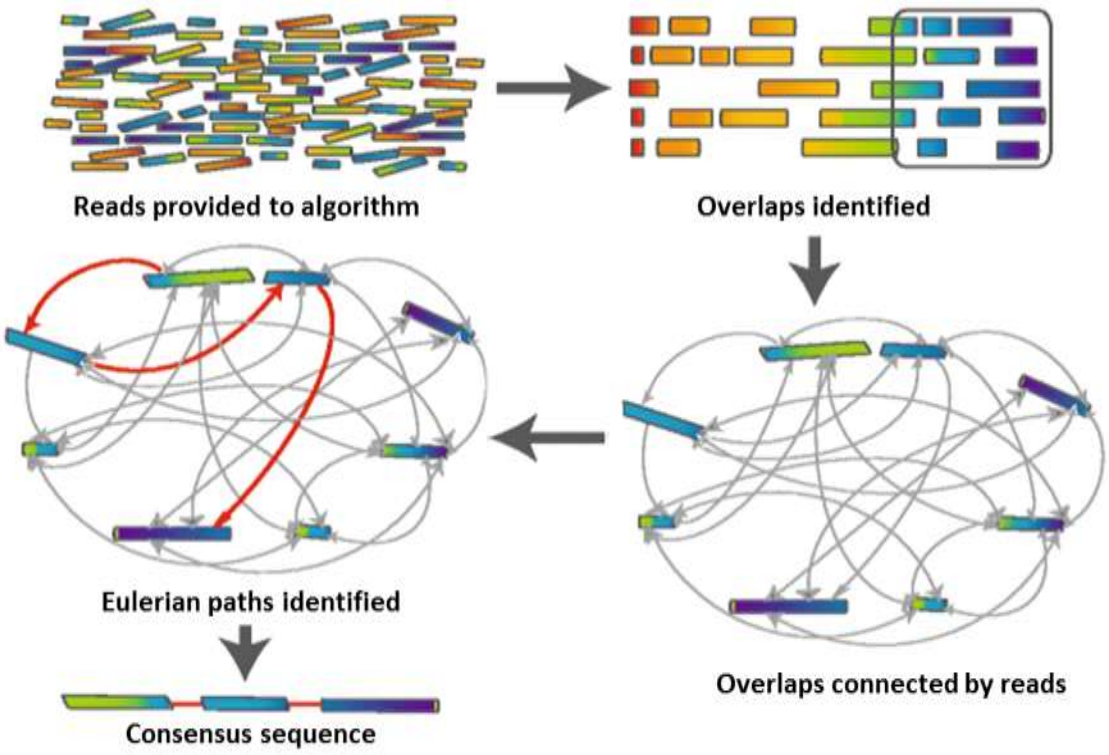

Figure 1.14 The overlapping puzzle: deBruijn graphs and Eulerian paths. Adapted from Commins *et al.*, 2009

1.6. Summary of study questions

Chapter 2: Does the trophosome of *R. piscesae* host multiple phylotypes of symbionts?

Is there an Epsilonproteobacteria as a second symbiotic partner?

Chapter 3: Does the trophosome of *R. piscesae* contain multiple genotypes of symbionts?

Chapter 4: Are the symbionts from the Juan de Fuca Ridge different from those on the East Pacific Rise? Are there differences between symbionts associated with three different species of worms?

See also Figure 1.8.

1.7. Thesis structure

This thesis is organized as a collection of articles. For clarity, the references pertaining to each article are all placed together into a unique bibliography section. On the electronic version, Figures, Tables, and cross-references possess hyperlinks for easy referencing.

Chapter 2. Investigating the possibility of *Epsilonproteobacteria* as a second endosymbiotic partner

This chapter is composed of two parts. Part one is the article entitled "Molecular study of bacterial diversity within the trophosome of the vestimentiferan tubeworm *Ridgeia piscesae*" co-written with Dr. Nathalie Forget and published in Marine Ecology:

Forget NL, Perez M, Juniper SK. (2014). Molecular study of bacterial diversity within the trophosome of the vestimentiferan tubeworm *Ridgeia piscesae*. *Marine Ecology* **36**: 35–44.

In this article, Dr. Forget performed the pyrosequence libraries analyses of the samples collected in 2010 and 2011 while I performed the CARD-FISH assays.

In Part two, I provide the pyrosequence data for the samples that I collected in 2013 (which are discussed but not shown in the published article), and discuss the different biases associated with the bioinformatic pipeline used in the article.

2.1. PART ONE: Molecular study of bacterial diversity within the trophosome of the vestimentiferan tubeworm *Ridgeia piscesae*

Abstract

A large proportion of the faunal biomass in hydrothermal vent ecosystems relies on symbiotic relationships with bacteria as a source of nutrition. While multiple symbioses have been observed in diverse vent hosts, siboglinid tubeworms have been thought to harbour a single endosymbiont phylotype affiliated to the *Gammaproteobacteria*. In the case of the Northeast Pacific vestimentiferan *Ridgeia piscesae*, two previous studies suggested the presence of more than one symbiont. The possibility of multiple, and possibly habitat specific, symbionts in *R. piscesae* provided a potential explanation for the tubeworm's broad ecological niche, compared to other hydrothermal vent siboglinids. This study further explored the diversity of trophosome bacteria in *R. piscesae* using two methodological approaches not yet applied to this symbiosis. We carried 454-pyrosequencing on trophosome samples from 46 individual worms and used catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) to verify the presence of the major groups detected in the pyrotag data. Both methods yielded inconsistent and sometimes contradictory results between sampling sites, and neither provided irrefutable evidence for the presence of symbionts other than the expected *Gammaproteobacteria*. We therefore conclude that the other adaptive mechanisms must be considered to explain the broad physico-chemical niche occupied by the different growth forms of *R. piscesae*.

2.1.1. Introduction

Hydrothermal vent environments host highly productive faunal communities relying on primary production by chemosynthetic microorganisms (Corliss *et al.*, 1979; Jannasch and Wirsen, 1979; Karl *et al.*, 1980). While free-living microbial communities represent an important source of organic carbon for suspension- and deposit-feeders, the bulk of the faunal biomass at most vent sites is supported by associations with symbiotic chemolithoautotrophic bacteria (Cavanaugh, 1994; Watsuji *et al.*, 2012; Ponsard *et al.*, 2013). At Eastern Pacific vents, symbioses are dominated by large populations of siboglinid tubeworms. These gutless polychaetes host symbionts in an organ known as the trophosome (Cavanaugh, 1994; Cavanaugh *et al.*, 1981; Felbeck, 1981). Most studies have detected a single, specific endosymbiont that is common to this group of worms (Edwards and Nelson, 1991; Feldman *et al.*, 1997; Black *et al.*, 1997). In contrast, other symbiont-bearing invertebrates known from deep-sea reducing habitats (vents, cold seeps and whale and wood falls), such as mytilid mussels (Distel *et al.*, 1995; Fiala-Médioni *et al.*, 2002), alvinocarid shrimp (Zbinden *et al.*, 2010; Petersen *et al.*, 2010) and provannid snails (Suzuki *et al.*, 2005; Urakawa *et al.*, 2005), host phylogenetically and metabolically diverse chemosynthetic partners. Investigation of the phylogenetic position of siboglinid symbionts has revealed two clusters corresponding to either cold seep or vent hosted organisms (Di Meo *et al.*, 2000), between which sequence divergence is around 4.3% on the 16S rRNA gene (Vrijenhoek, 2010a).

There is some evidence for a more diverse trophosomal symbiotic assemblage in the northeast Pacific siboglinid tubeworm *Ridgeia piscesae*. Early ultrastructural studies of *R. piscesae* trophosomes suggested that similar symbionts are found across worm morphotypes, but that two morphologically distinct bacteria could occur within a single host (deBurgh *et al.*, 1989). More recently, using terminal-restriction fragment length polymorphism (t-RFLP),(Chao *et al.*, 2007) detected the expected *Gammaproteobacteria*l phylotype plus two novel phylotypes from

the same class, together with one *Alphaproteobacteria*, and one *Bacteroidetes*. The goal of the present study was to pursue these investigations and explore the diversity of the bacteria within *R. piscesae*'s trophosome using pyrosequencing and catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH). Pyrosequencing has not previously been used for screening endosymbiont diversity in vestimentiferans and is therefore considered exploratory.

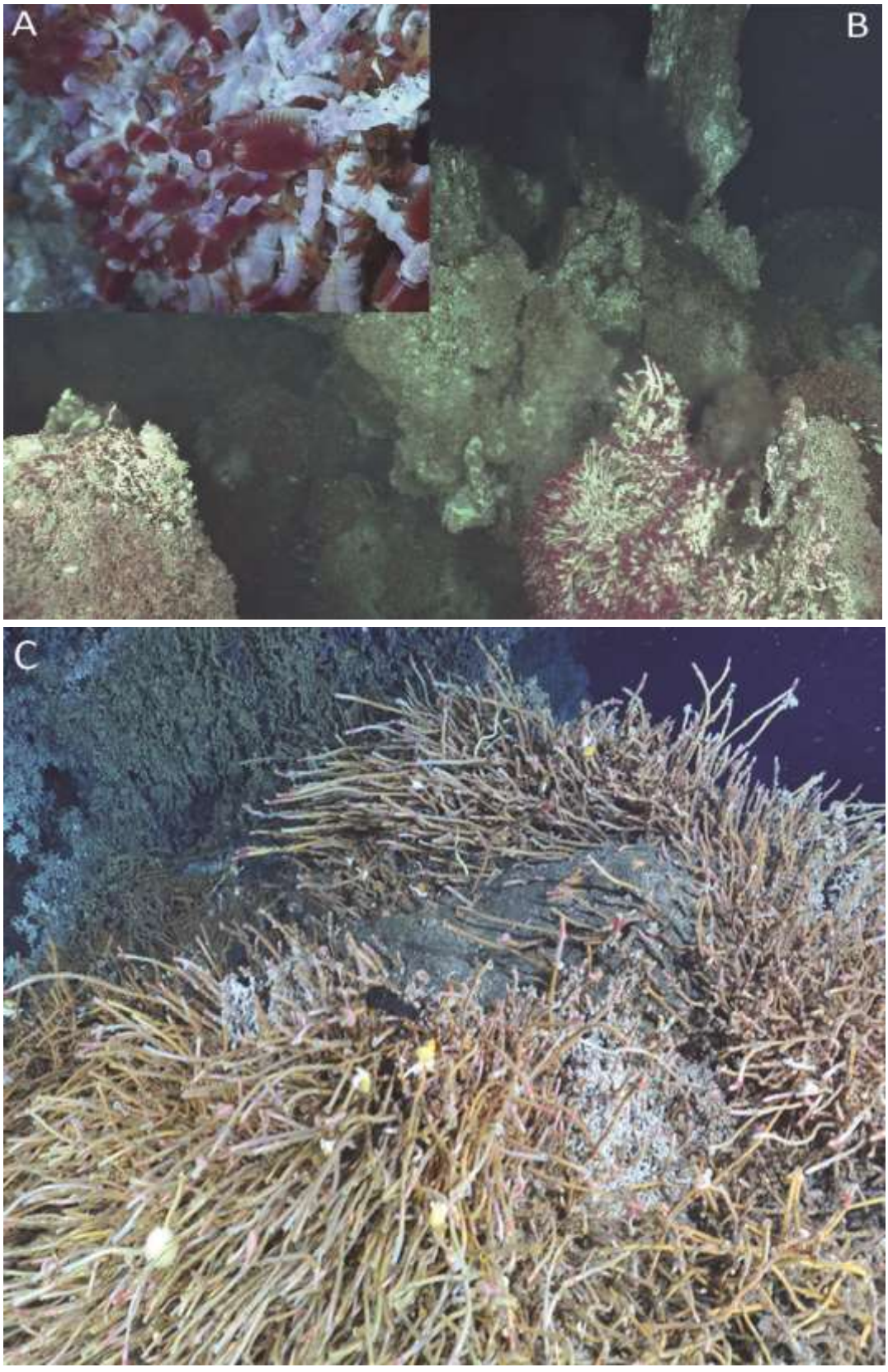

Figure 2.1 Examples of typical sampling sites. **A) Aggregation of the "short-fat" morphotype of *R. piscesae*. B) Zoom out showing a black smoker in the surrounding area. C) Habitat of the '"long-skinny" morphotype of *R. piscesae*.** Here, no shimmering is visible.

2.1.2. Material and Methods

2.1.2.1. Sample collection

Samples of *R. piscesae* were collected during three separate research expeditions in July 2010 onboard the R/V Atlantis, using the submersible *Alvin*, in July 2011 onboard the R/V Thomas G. Thompson, using the remotely-operated vehicle ROPOS, and in June 2013 onboard the R/V Thomas G. Thompson using an Oceaneering Millenium Plus remotely-operated vehicle. In 2010 and 2011, individuals of the two most extreme morphotypes of the tubeworm, known as the "short-fat" and the "long-skinny" morphotypes (Figure 2.1), were sampled from two vent sites on Axial Volcano and four in the Main Endeavour vent field (Table 2.1). Smaller samples of the two morphotypes were collected from the Main Endeavour vent field in 2013. Samples were transported to the surface in sealed, separated bioboxes to prevent contamination between samples and from ambient seawater. Once on board, samples intended for pyrosequencing were pre-processed in a 5˚C cold room: the bodies of the worms were carefully removed from their tubes and cleaned with 70% (v/v) ethanol, individually packed and frozen at -80°C. For CARD-FISH, the tubes were removed and the bodies were cleaned as described previously. For the 2010 and 2011 individuals, the bodies were dissected and subsamples of tubeworm trophosome were fixed as described by Dubilier *et al.*(1995) For the 2013 samples, the bodies were fixed whole according to the previous protocol with some modification: the whole bodies were incubated in 4% paraformaldehyde/0.1M PBS for 18 hours at 4˚C. After three rinse in filtered water, they were gradually dehydrated and stored in 70% ethanol at 4˚C until sectioning. Some specimens were fixed without rinsing in order to assess potential epibiotic contamination.

Table 2.1 Description and location of sampling sites.

Sampling Site ID	Tubeworm Morphotype	Vent Site	Latitude	Longitude	Depth (m)	Max. temp (°C) at plume	Collection Date	Analysis Technique(s)	No. of individuals analyzed
LF10AV1b	Long-Skinny	Axial Volcano (Hollywood Flats 1)	45° 56.147' N	129° 58.888' W	1518	na	July-10	Pyrosequencing	5
HF10AV2b	Short-fat	Axial Volcano (Hollywood Flats 2)	45° 56.155' N	129° 58.893' W	1517	4.1	July-10	Pyrosequencing & CARD-FISH	5 & 3[c]
LF10AV2b	Long-Skinny	Axial Volcano (Hollywood Flats 2)	45° 56.156' N	129° 58.890' W	1517	2.0	July-10	Pyrosequencing & CARD-FISH	5 & 3[c]
LF10TPb	Long-Skinny	Main Endeavour (TP)	47° 56.971' N	129° 5.854' W	2197	2.4	July-10	Pyrosequencing	5
HF10HUb	Short-fat	Main Endeavour (Hulk)	47° 57.007' N	129° 5.824' W	2190	14.0	July-10	Pyrosequencing & CARD-FISH	5 & 3[c]
LF10HUb	Long-Skinny	Main Endeavour (Hulk)	47° 57.007' N	129° 5.825' W	2191	2.5	July-10	Pyrosequencing & CARD-FISH	5 & 3[c]
HF11GRb	Short-fat	Main Endeavour (Grotto)	47° 56.953' N	129° 5.903' W	2188	21.6	July-11	Pyrosequencing	5
HF11LBb	Short-fat	Main Endeavour (Lobo)	47° 56.965' N	129° 5.900' W	2191	12.2	July-11	Pyrosequencing	5

[c] CARD-FISH

34

2.1.2.2. 454 pyrosequence library construction

DNA was extracted from approximately 25 mg of tissue using the DNeasy Blood and Tissue Kit (Qiagen, Carlsbad, CA, USA), following the manufacturer's instructions, from a total of 40 tubeworm trophosomes (Table 2.1). DNA was purified and concentrated using the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions, and 20 µl of DNA with a concentration of 20 ng/µl or higher was sent to the *Plateforme d'Analyses Génomiques* (Institute of Integrative and Systems Biology, Laval University, Quebec City, QC, Canada). The hypervariable region V1-V3 of the bacterial SSU rRNA gene was amplified by PCR using Takara Ex Taq premix (Fisher Scientific, Toronto, ON, Canada). PCR reactions were performed in a final volume of 50 µl containing 25 µl of Premix Taq, 1 µM of each primer and sterile MiliQ H_2O to up to 50 µl (a list of the primers is available in *Supporting information at* http://onlinelibrary.wiley.com/doi/10.1111/maec.12169/abstract, and in Appendix A, List A.1; p. A.1). PCR conditions were as follows: after a denaturing step of 30 s at 98°C, samples were processed through 30 cycles consisting of 10 s at 98°C, 30 s at 55°C and 30 s at 72°C. A final extension step was performed at 72°C for 4 min 30 s. Following amplification, samples were purified using magnetic AMPure Beads (Beckman Coulter Genomics) to recover PCR amplicons, separating them from contaminants. Samples were adjusted to 100 µl with EB buffer (Qiagen), to which 63 µl of beads were added. Samples were mixed and incubated for 5 min at RT. Using a Magnetic Particle Concentrator (MPC), the beads were pelleted against the wall of the tube and supernatant was removed. The beads were washed twice with 500 µl of 70% ethanol and incubated for 30 s each time. Supernatant were removed and beads were allowed to air dry for 5 min. Tubes were removed from the MPC and 15 µl of EB buffer were added. Samples were vortexed to resuspend the beads. Finally, using the MPC, the beads were pelleted against the wall once more and supernatant were

transferred to a new clean tube. DNA concentrations in sample were quantified by Nanodrop and mixed in equal amounts. Pyrosequencing was performed using a 454 GS-FLX DNA Sequencer with the Titanium Chemistry (Roche) according to the procedure described by the manufacturer.

2.1.2.3. Pyrosequencing read analysis

All data processing and analyses were performed using the software program mothur (Schloss *et al.*, 2009). Raw pyrosequences were checked for different quality criteria. Reads with an average quality score below 27 (Kunin *et al.*, 2010), containing an error in the forward primer sequence at the beginning of the read (Sogin *et al.*, 2006), containing one or more ambiguous bases (Ns) (Sogin *et al.*, 2006; Huse *et al.*, 2010), or shorter than 250 base pairs (De León *et al.*, 2012) were eliminated. A set of unique reads was created and aligned against the SILVA-based bacterial reference alignment (Pruesse *et al.*, 2007) provided by mothur using the Needleman-Wunsch pairwise alignment algorithm (Needleman and Wunsch, 1970). A pre-clustering step was applied to group sequences differing by less than 2% (corresponding to 5 mismatches for a 250-base-pair sequence) (Huse *et al.*, 2010). Potential chimeras were identified using the program UCHIME (Edgar *et al.*, 2011) and removed from the data set. Because there was a large range between the minimum and maximum number of reads found in our samples, the three samples with the lowest numbers of reads (one sample from HF10AV2 with 807 reads, one sample from HF10HUb with 2294 reads, and one sample from LF10AV1b with 1583 reads) were eliminated from further analyses. Singletons (unique sequences) were also eliminated. Then, the number of reads across samples was standardized by subsampling, based on the lowest number of sequences (3593) found in any of the remaining 37 samples. The remaining sequences were classified using the

Silva template database with 1000 bootstrap iterations. The command 'phylotype' was used to generate a file listing the sequences affiliated to each taxon at the phylum, class, order, family, and genus levels. A shared file, which described the number of times each taxon was observed in all samples, was generated.

2.1.2.4. Nucleotide sequence accession numbers

The pyrosequence reads have been deposited in the NCBI Short Read Archive under accession number SRA058565.

2.1.2.5. Catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH)

Fixed tissues of individual worms from the 2010 and 2013 collections were embedded in paraffin and sectioned (5 μm thickness) onto glass slides. The pre-hybridization treatments were performed as previously described (Dubilier *et al.*, 1995) with the following modifications: sections were deparaffinised in CitriSolv (Fisher Scientific), a less toxic alternative to xylene, and the post-fixation step in 3.7% formaldehyde was omitted. CARD-FISH was carried out as described by Blazejak *et al.* (2005) with the following horseradish peroxidase (HPR)-labelled oligonucleotide probes: EPSY549, specific to the *Epsilonproteobacteria*, GAM42a, covering most *Gammaproteobacteria*, EUB338, targeting the domain *Bacteria* as a positive control, and NON338, a complementary negative control. Sequences for these general probes were obtained through the probeBase website (Loy *et al.*, 2007). For each probe, formamide concentration was optimized using a range of different concentrations in order to get the best signal with the highest formamide

concentration (most stringent conditions possible) (Table 2.2). The fluorescently labelled tyramides were prepared as described by Pernthaler *et al.* (2004) with the Alexa Fluor 488, 555 (Molecular Probes - Invitrogen). A few sections were hybridized without a probe to control for background fluorescence. For multiple hybridizations, the CARD-FISH protocol was repeated with the same sections with different probes and dyes as described in Blazejak *et al.* (2005). Slides were imaged under epifluorescence and confocal illumination, using a Leica Leitz DMRB fluorescent microscope or a Nikon C1 Plus confocal microscope.

Table 2.2 Oligonucleotide probe description.

Probe	Specificity	Sequence (5'-3')	Position[a]	[Formamide] (%, v/v)[c]	Reference
EUB338	Bacteria	GCTGCCTCCCGTAGGAGT	338-355	50-**55**-60	Amann *et al.* 1990
EPSY549	*Epsilonproteobacteria*	CAGTGATTCCGAGTAACG	549-566	50-**55**-60	Lin *et al.* 2006
GAM42a	*Gammaproteobacteria*	GCCTTCCCACATCGTTT	1027-1043[b]	50-**55**-60	Manz *et al.* 1992
NON338	Negative control	ACTCCTACGGGAGGCAGC	338-355	25-30-35-40-**55**	Widdel and Bak 1992; Wallner *et al.* 1993

[a] Position in the 16S rRNA of *Escherichia coli* unless indicated otherwise.

[b] Position in the 23S rRNA of *E. coli*.

[c] In hybridization buffer. Numbers in bold indicate the concentration used in this study.

[d] Non labelled probe.

2.1.3. Results

2.1.3.1. 454 pyrosequence library

For the 40 tubeworm trophosomes sampled in 2010 and 2011, a total of 645 009 reads were obtained through pyrosequencing. After quality filtering and removing the three less abundant samples, 513 860 high-quality pyrosequences remained, representing 3022 unique sequences. A total of 1825 singletons were eliminated from further analysis and standardization left 132 941 sequences of which 893 were unique. Eleven different phyla were detected but only two represented more than 1% of the sequence library: the *Proteobacteria* and *Bacteroidetes*. Within the *Proteobacteria*, which represented 97.0% of the sequences, *Gammaproteobacteria* were clearly the most abundant class, followed by *Epsilonproteobacteria, Deltaproteobacteria*, *Alphaproteobacteria*, and *Betaproteobacteria* (Figure 2.2). The other phyla detected, accounting for 0.1% of the sequence library, included *Actinobacteria*, *Firmicutes*, *Chloroflexi*, *Spirochaetes*, *Acidobacteria*, *Cyanobacteria*, *Verrucomicrobia*, and the candidate divisions BD1-5 and TM7.

Results were not consistent between locations. For the six sites sampled in 2010, sequences belong to classes other than the *Gammaproteobacteria* constituted 10-30% of the sequence libraries (Figure 2.2), while for the two sites sampled in 2011 (HF11GRb & HF11GBb in Figure 2.2) and the 2013 sites (data not shown), non-*Gammaproteobacteria* made negligible contributions to the sequence libraries.

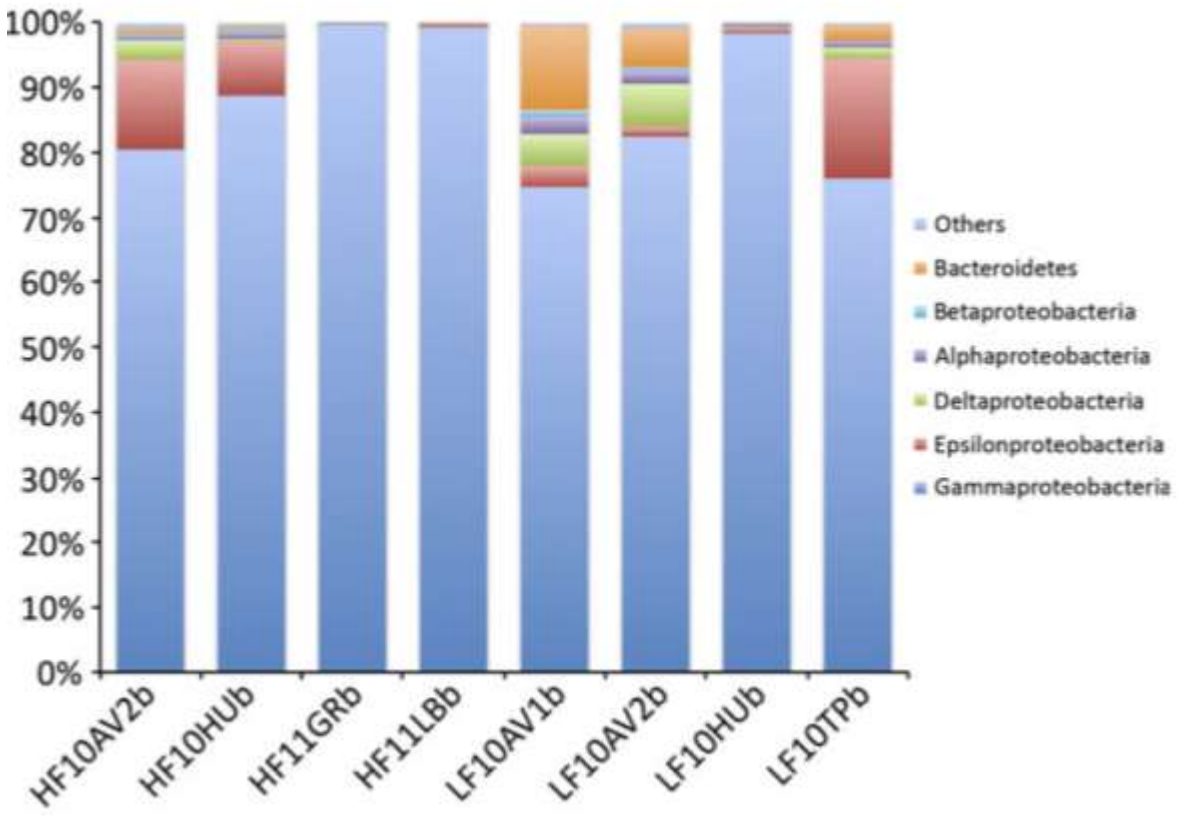

Figure 2.2 Relative abundance of the phyla accounting for > 1.0% of the pyrosequence library constructed from the trophosomes of 37 individuals of *R. piscesae*. Since the *Proteobacteria* accounted for 97.0%, this phylum was divided into the five classes detected.

2.1.3.2. Bacteria detection in *R. piscesae* trophosome

Multiple hybridizations GAM42a confirmed the dominant presence of members of *Gammaproteobacteria* within the trophosome of single *R. piscesae* in all individuals analyzed. Dual hybridization with EPSY549 and GAM42a, suggested a minority presence of dispersed *Epsilonproteobacteria* in the trophosome tissue in the 2010 samples (Figure 2.3 A), and an almost negligible presence in the 2013 samples (data not shown). To assess the binding specificity of EPSY549

the probe was hybridized in parallel with the general bacteria probe (EUB338) using tissue sections from 2010 samples (Figure 2.4 A,B). The high concentration of *Gammaproteobacteria* (Figure 2.3 B) made it difficult to assess if *Epsilonproteobacteria* were also detected with EUB338 (Figure 2.4 A). While the majority of the *Epsilonproteobacteria* appeared to co-localize with EUB338, there was some indication of non-specific binding (Figure 2.4 A,B, arrows). Hydridization with the NON338 probe yielded some very bright points, mostly co-localizing with nuclei of epithelial cells (Figure 2.4 C,D) while the remainder of the EPSY549 signal was localized in the central and median zone of the trophosome lobes (Figure 2.3 A and Figure 2.4 A). The non-specific signal was most likely the result of binding of the probe to other cellular components rather than mispairing with non- target sequences (Wallner *et al.*, 1993). Hybridization with another negative control (ECO1459 targeting the non-hydrothermal vent species *E.coli*) resulted in a similar, non-specific signal (data not shown).

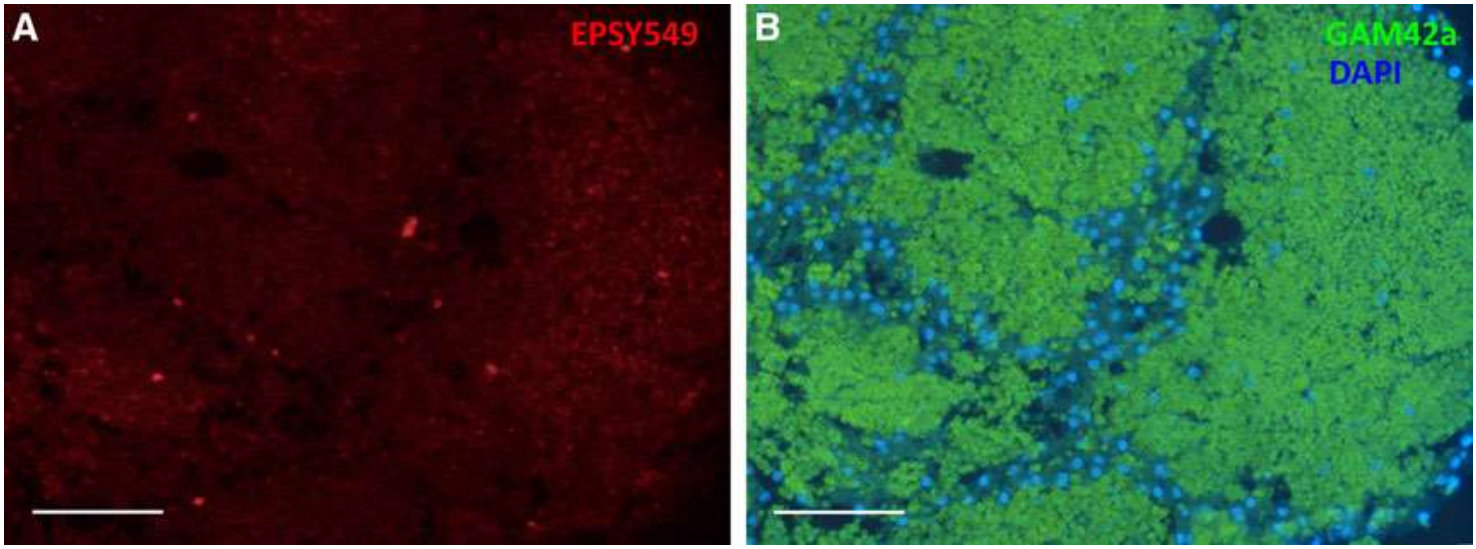

Figure 2.3 Double-probe catalysis reporter deposition fluorescent *in situ* hybridization of 5µm sections of *Ridgeia piscesae* dissected trophosomes, with A) EPSY549 (red), B) merged GAM42a (green) and DAPI (blue) signals. Scale bars =50µm. Image taken under Leica Leitz DMRB fluorescent microscope.

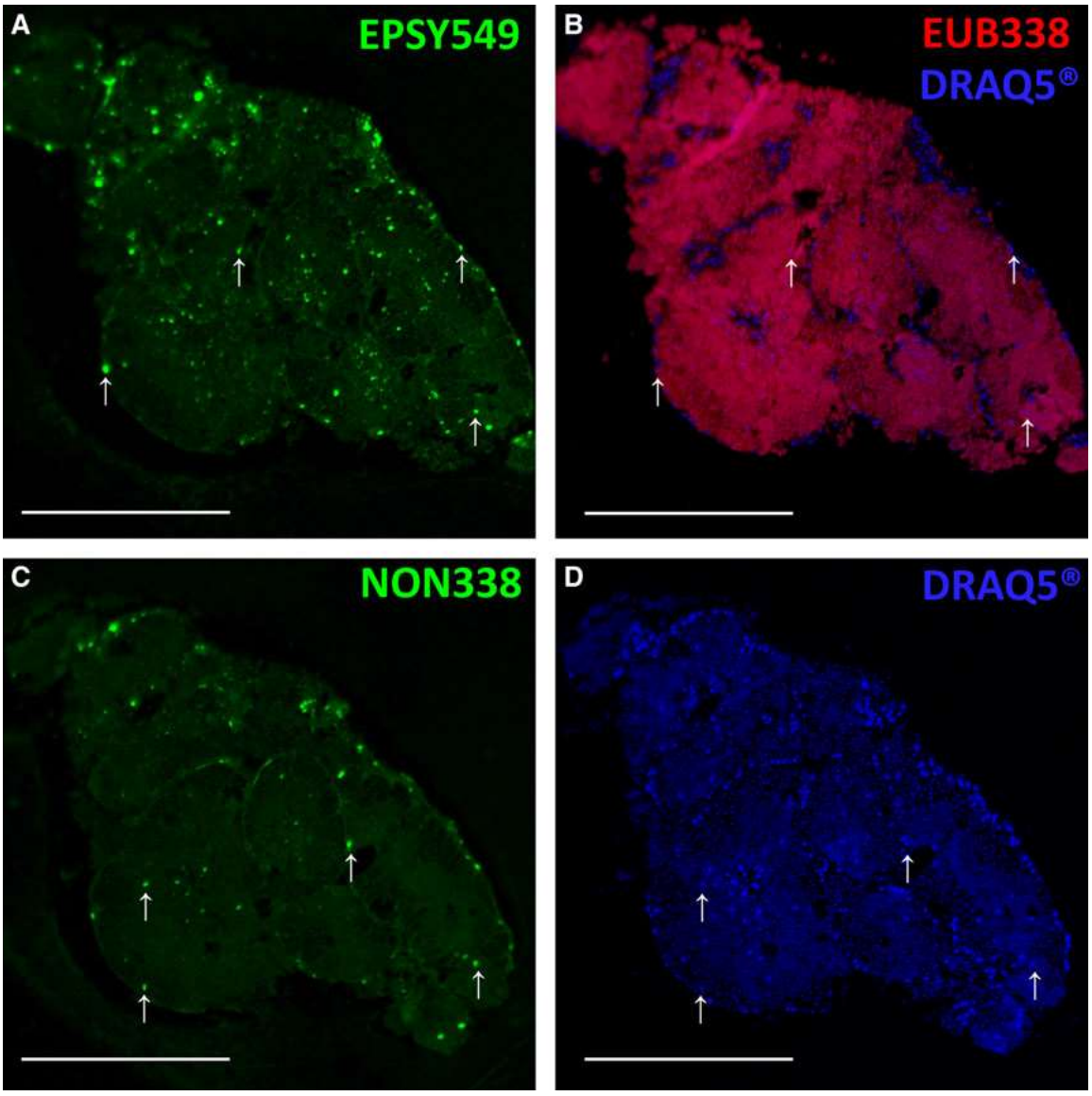

Figure 2.4 Double-probe catalysis reporter deposition fluorescent *in situ* hybridization of the same region of the dissected trophosome of an individual *Ridgeia piscesae*. A) EPSY549, B) merged EUB338 and DAPI signals, C) NON338, D) DAPI. Arrows represent some of the non-specific signal in A, B and C, D. Scale bars = 200µm. Image taken under a Nikon C1 Plus confocal microscope.

2.1.4. Discussion

Symbiotic associations between bacteria and eukaryotes are widespread in the biosphere and dominate faunal biomass in reducing habitats such as hydrothermal vents, cold seeps, and whale and wood falls, where the host bridges oxic and anoxic zones, facilitating access to oxidants and reductants for the chemoautotrophic symbionts (Cavanaugh *et al.*, 2006; Watsuji *et al.*, 2012; Ponsard *et al.*, 2013). While the association of vent organisms from many different families with multiple bacterial phylotypes is well documented (Distel *et al.*, 1995; Fiala-Médioni *et al.*, 2002; Suzuki *et al.*, 2005; Urakawa *et al.*, 2005; Petersen *et al.*, 2010), vent tubeworms have been thought to host only one symbiont phylotype shared across the entire group (Edwards and Nelson, 1991; Feldman *et al.*, 1997; Markert *et al.*, 2007; Robidart *et al.*, 2008). In the case of *R. piscesae*, two previous studies suggested the presence of more than one bacterial phylotype in the trophosome (deBurgh *et al.*, 1989; Chao *et al.*, 2007). The deBurgh *et al.* (1989)'s study was based entirely on morphological examination of bacterial cells within trophosome tissue, without molecular evidence that the different morphologies corresponded to phylogenetically distinct bacteria. Chao *et al.* (2007) provided molecular evidence, but the authors recognized that their conclusion, based solely on T-RFLP data, could have been influenced by contamination from bacteria associated with the worm's tube or the environment where they were collected (Chao *et al.*, 2007). In this study, we used 454 pyrosequencing and CARD-FISH to further explore the possibility of multiple endosymbionts in *R. piscesae*. The pyrosequencing approach has rarely been used to investigate endosymbiont diversity, and since the rare biosphere detected by this technique is more likely to contain artefacts (Huse *et al.*, 2010; Kunin *et al.*, 2010; Tedersoo *et al.*, 2010), we applied very strict quality filtering to reduce background contamination and a high number of replicates to improve data comparability (Zhou *et al.*, 2011). Pyrosequencing

results from a previous study of free-living bacteria associated with *R. piscesae* assemblages (Forget and Juniper, 2013) provide some background for interpretation of data presented here, particularly with respect to possible contamination. Nevertheless, the results obtained from pyrosequencing should be considered exploratory and the detection of rarer or unexpected bacterial lineages, *i.e.* other than *Gammaproteobacteria,* need be confirmed by PCR-independent methods such as CARD-FISH. This point is reinforced by the inconsistent pyrosequencing results obtained here for tubeworms collected from similar physico-chemical habitats. The latter observation rules out any simple explanation based on the worm's physiological requirements. The presence of non-symbiont sequences resulting from pyrosequencing artefacts or environmental contamination (external bacteria) represents a more parsimonious explanation for these inconsistencies.

The dominant presence of *Gammaproteobacteria* in *Ridgeia piscesae* trophosome tissue from both habitats was confirmed by CARD-FISH. This result is not surprising since the previously known siboglinid endosymbionts are members of the *Gammaproteobacteria*. The affiliation of the most abundant genera to the genus *Methylomicrobium* is doubtful. We used mothur and the Silva alignment to compare the classification of a SSU rRNA Gammaproteobacterial sequence identified as *R. piscesae* endosymbiont (accession number U77480 (Feldman *et al.*, 1997)). While the near-full length of the gene was identified as a member of the family *Sedimenticola*, which corresponds to the affiliation of *R. pachyptila*'s endosymbiont, Candidatus *Endoriftia persephone* (based on the EzTaxon-e server (Kim *et al.*, 2012)), the V1-V3 region of the sequence was classified as the genus *Methylomicrobium* from the family *Methylococcaceae*., Okubo *et al.* (2012) assessed phylogenetic drifts of pyrosequence read classification and suggested that assignment at the genus level is affected by read length. Such classification errors can lead to incorrect conclusions about the ecological role of the community investigated.

Members of the *Epsilonproteobacteria* have been found within the epibiotic community of the galatheid crab *Shinkaia crosnieri* (Watsuji *et al.*, 2012), the alvinocaridid shrimp *Rimicaris exoculata* (Zbinden *et al.*, 2008; Petersen *et al.*, 2010; Guri *et al.*, 2012), the alvinellid polychaetes *Paralvinella sulfincola*, *P. palmiformis* and *Alvinella pompejana* (Haddad *et al.*, 1995; Campbell *et al.*, 2013; Alain *et al.*, 2002; Pagé *et al.*, 2004), as well as the siboglinid polychaetes *R. pachyptila* and *R. piscesae* (López-García *et al.*, 2002; Kalanetra and Nelson, 2010). Epsilonproteobacterial endosymbionts have been previously detected in provannid gastropods from the genus *Alviniconcha* (Urakawa *et al.*, 2005), and in pectinodontid gastropods from the genus *Pectinodonta* (Zbinden *et al.*, 2010), but pyrosequencing results from our 2010 samples (30 individual worms) represent the first indication of their presence in vent siboglinids. The most abundant genus detected in our pyrosequence library, *Sulfurovum*, was also the most abundant group detected in the free-living bacterial communities associated with *R. piscesae* (Forget and Juniper, 2013). This result could indicate environmental contamination of the trophosome. The low relative abundance of epsilonproteobacterial pyrotags in the 2010 data and their absence in sequence data from 2011 and 2013 samples rule out a systematic presence and role in the tubeworm's nutrition, and reinforce the external contamination explanation. *Epsilonproteobacteria* have been shown to contribute to other invertebrate-prokaryote symbioses at hydrothermal vents. In the case of *R. exoculata*, the closest relative to their epsilonproteobacterial symbionts was also a member of the genus *Sulfurovum* (Petersen *et al.*, 2010). A trophic role was suggested for epsilonproteobacterial epibionts in the case of *S. crosnieri* and *R. exoculata* (Watsuji *et al.*, 2012; Ponsard *et al.*, 2013).

CARD-FISH revealed the expected dominance of *Gammaproteobacteria* in the trophosome tissue of all specimens plus a low-level presence of cells hybridizing to the *Epsilonproteobacteria* probe in the 2010 samples. The latter were diffusely distributed throughout the trophosome tissue sections, and the NON-

probe indicated some to be non-specific. Given the contradictory results obtained with the *Epsilonproteobacteria* probe in samples from the two years (and locations), and the fact that, when detected, putative *Epsilonproteobacteria* were very scarce and similar in abundance to points of non-specific hybridization, we could not definitely confirm the presence of *Epsilonproteobacteria* in the trophosome.

2.1.5. Conclusion

This study is the first to explore potential symbiotic diversity within a siboglinid tubeworm through 454-pyrosequencing. The surprisingly high diversity of taxonomic groups revealed by pyrosequencing must be handled carefully: the majority of the genera detected had very low frequency. Molecular methods based on rRNA genes can be sensitive to the relative abundance of organisms and biased toward those with higher gene copy numbers (Farrelly *et al.*, 1995; Crosby and Criddle, 2003; Hoshino *et al.*, 2008). The preponderance of *Gammaproteobacteria* within tubeworm trophosomes, and of one particular phylotype, could have limited the detection of less abundant groups/phylotypes in other studies. The next most abundant class of bacteria for which there was some pyrotag evidence was the *Epsilonproteobacteria*. However, 454-pyrosequencing results were inconsistent among the locations (and years) sampled in this study, with respect to the presence of *Epsilonproteobacteria*, and CARD-FISH results were even less supportive. We therefore conclude that there is currently no irrefutable evidence to support previous suggestions of that *R. piscesae* hosts multiple symbionts. It therefore appears to be necessary to consider other mechanisms that could permit *R. piscesae* and its symbionts to occupy the broad range of physico-chemical conditions that are exploited by the different growth forms of this tubeworm. For example, metagenomic analysis

and related profiles of gene expression may provide insight into how *R. piscesae* symbionts and their host interact with their external environment.

Acknowledgements

The authors would like to thank Dr. Bob Chow for access to his confocal microscope and for his expertise and valuable time training NLF and MP. We are also grateful to the team of Wax-it Histology Services for their collaboration during the preparation of the tissue sections, to Candice St. Germain for sharing her samples and for insightful discussions, and to Carol Doya and Emily Boulter for helping with the delicate dissections of worms on board the R/V Thomas G. Thompson. We thank the crews of the R/Vs Atlantis and Thomas G. Thompson as well as the pilots of the submersible *Alvin,* and the ROVs ROPOS and Oceaneering Millenium. This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to SKJ and a Canadian Healthy Oceans Network (NSERC Canada) grant to Dr. Verena Tunnicliffe. During this study, NLF benefitted from an NSERC graduate scholarship, a Montalbano Scholars Fellowship, a Dr. Arne Lane Graduate Fellowship, a Commander Peter Chance MASC Graduate Fellowship, an Alfred and Adriana Potvin Graduate Scholarship in Ocean Sciences, a W. Gordon Fields Memorial Fellowship, a Charles S. Humphrey Graduate Student Award, and a Maureen De Burgh Memorial Scholarship.

2.2. PART TWO; Supplement: Pyrosequences from the 2013 collection

2.2.1. Introduction

A subsample of the 2013 worms (used in CARD-FISH analysis) was used for pyrosequencing of symbiont 16S rRNA variable regions V1-V3 and V6-V8. To minimize potential contamination from external, this set of samples was treated differently from the 2010 samples.

Table 2.3 Set of samples collected on June 18th (M11 tag) and June 23rd (M16 tag) 2013 at the Main Endeavour Vent Fields. DNA concentration was measured by fluorometric quantification using PicoGreen®.

Sample ID	Tubeworm morphology	Vent site	Flow Regime	Lat	Long	Depth (m)	Dna concentration (ng/ul)	16S region
M1609	Short-fat	Grotto (High Tower)	High Flow	47° 56.9600 N	129° 05.9185 W	2195	138	Bact V1-V3 + Bact V6-V8
M1608	Short-fat	Grotto (High Tower)	High Flow	47° 56.9600 N	129° 05.9185 W	2195	45	Bact V1-V3
M1607	Short-fat	Grotto (High Tower)	Low Flow	47° 56.9600 N	129° 05.9185 W	2195	16	Bact V1-V3
M1112	Long-skinny	Grotto (High Tower)	Low Flow	47° 56.9663 N	129° 05.9174 W	2190	57	Bact V1-V3
M1110	Long-skinny	Grotto (High Tower)	Low Flow	47° 56.9663 N	129° 05.9174 W	2190	26	Bact V1-V3
M1106	Long-skinny	Grotto (High Tower)	Low Flow	47° 56.9663 N	129° 05.9174 W	2190	46	Bact V1-V3 + Bact V6-V8

2.2.2. Material and Methods

As previously described, all six tubeworm bodies were carefully removed from their tubes upon recovery to the ship, and rinsed with 70% v/v ETOH to remove potential epibiotic contamination. In 2010, tubeworms were then individually packed and frozen at -80°C at sea, and trophosomes were later dissected in the laboratory after thawing. In 2013 trophosomes were dissected on board to avoid the freezing and thawing cycle that could cause further contamination of their internal organs by environmental bacteria. The trophosomes dissected in 2013 underwent a more stringent decontamination treatment at sea, prior to freezing. As in Carney *et al.* (2002), the trophosomes were rinsed with sterile seawater several times and incubated in a TE buffer (10:1 tris HCl/EDTA) and lysozyme (1mg/ml) solution at room temperature for 30 min. Then, the tissues were suspended in DNase and $MgCl_2$ at 37°C for 5 min and washed several times with TE buffer (1:1 tris HCl/EDTA). Finally, the trophosomes were stored at -80°C in individual sterile microcentrifuge tubes and shipped to our laboratory for DNA extraction.

The construction of the 454-Pyrosequencing library and the pyrosequencing read analysis followed the same protocol in both 2010 and 2013 (Forget *et al.*, 2014).

2.2.1. Results

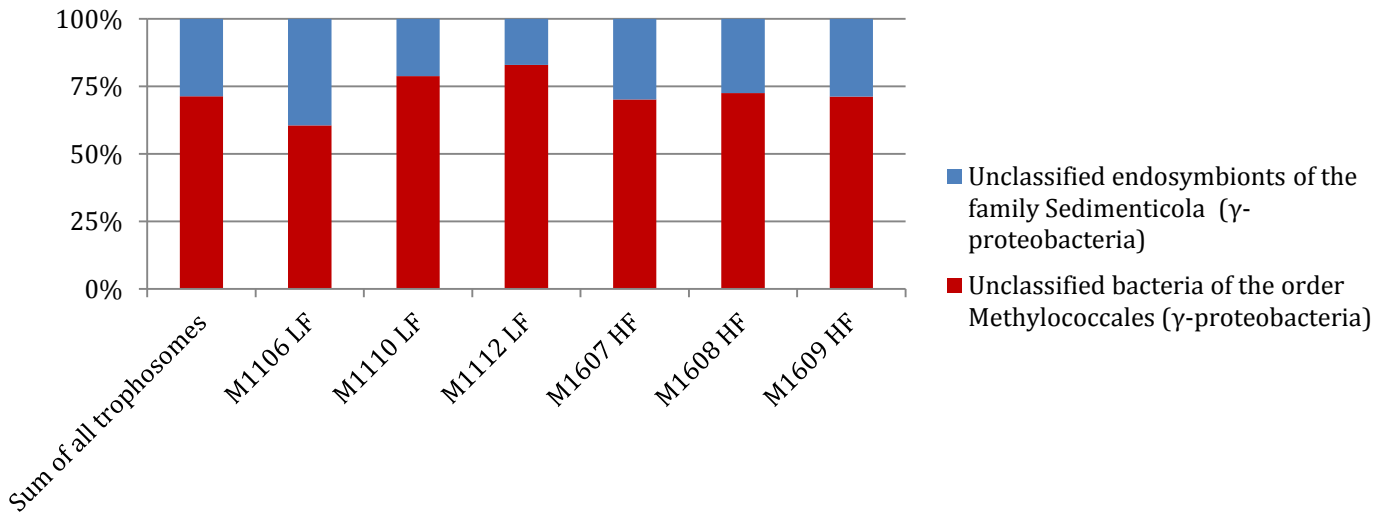

Figure 2.5 Relative abundances of the unique, preclustered pyrosequences of the trophosomes of six individual tubeworms. The 16S rRNA sequences (V1-V3 regions) were classified using the Silva gold database with 1000 bootstrap iterations. Note that only *Gammaproteobacteria* were detected.

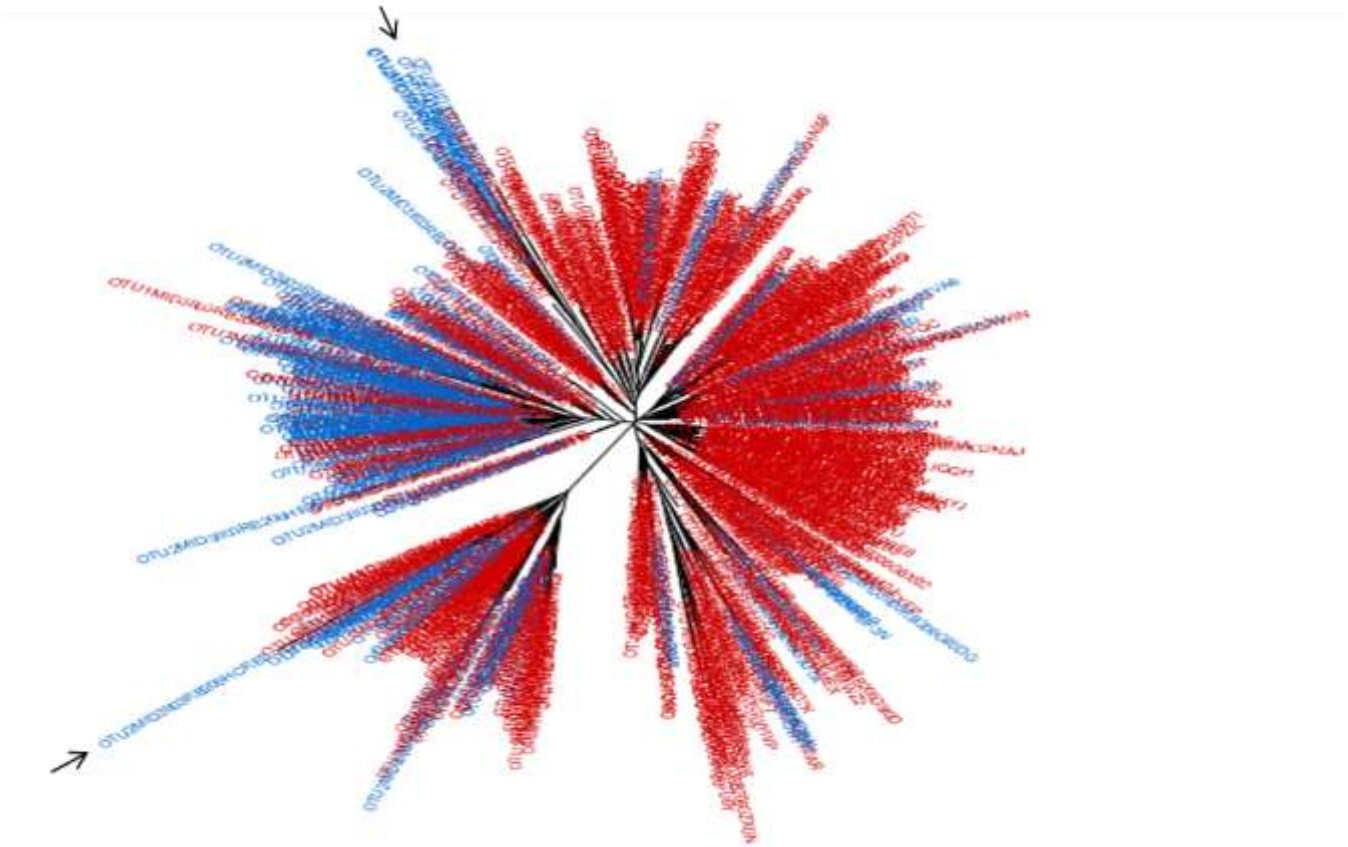

Figure 2.6 Neighbour joining tree constructed from the pairwise DNA distances between the unique, preclustered pyrosequences of M1106 (607 sequences). The 16S rRNA sequences (V1-V3 regions) sequences were aligned with Mafft (Katoh and Standley, 2013). Blue: sequences classified as Sedimenticola; Red: sequences classified as Methylococcales. The nucleodide identity between the two most divergent sequences (indicated by arrows) was >92%.

The pyrosequence libraries constructed from the individual trophosomes of the tubeworms collected in 2013 contained a lot less diversity than those from 2010. For both the V1-V3 and V6-V8 regions, all samples contained only *Gammaproteobacteria* sequences.

The taxonomy could not be resolved past the Class level. The classification of the pre-clustered sequences against the Silva template database resulted in two clades (Figure 2.5). These sequences were classified as bacteria of the Order Methylococcales (~70%) or as an unclassified endosymbiont of the family Sedimenticola (~30%) (Figure 2.5). However, the bootstrap values (1000 bootstraps) for both taxa were very low (~50%). In addition, the use of another database (RDP) or another region of the gene (V6-V8); for samples M1106 and 1609) led to different taxonomic affiliations within the *Gammaproteobacteria*. Finally, the dichotomy between the two taxa was not supported by the trees built from pairwise distances between the sequences (Figure 2.6).

2.2.2. Discussion

As mentioned in the Discussion, Section 2.1.4, pyrosequencing results should be interpreted with caution as this technique is prone to multiple biases (Quince *et al.*, 2009; Tedersoo *et al.*, 2010).

Some of these biases result in an underestimation of the true diversity of a bacterial community and are caused by the PCR steps necessary to amplify the 16SrRNA fragments from the bulk DNA prior to sequencing. PCR bias has been well documented (Suzuki and Giovannoni, 1996; Pinto and Raskin, 2012). The first source of biases comes from the variable affinity of primers to their target sequences due to differences in DNA structure at their target sites (Wagner *et al.*, 1994). In our case, however, preferential bias is not likely to explain the differences in diversity observed between the pyrosequence libraries of 2010 and 2013 because the exact same set of primers was used to amplify the

16SrRNA gene sequences in both datasets. Another source of bias is PCR drift (Wagner *et al.*, 1994). Because sequences are generated by a geometric process, uneven allelic frequencies caused by random sampling in the early cycles of the PCR or by the naturally low abundance of a given allele in the DNA extract are exacerbated. Alleles with the lowest frequencies are less amplified and lost in favor of dominant alleles. PCR drift caused by random sampling is more severe when the initial quantity of target DNA is low. While we did not know initial copy numbers of 16S rRNA gene in our extracts, the total DNA concentration of the 2010 extracts was not significantly different from the 2013 samples. The concentration averaged 56 ng/µl and ranged from 10 to 239 ng/µl in 2010 (data not shown) and averaged 55 ng/µl ranging from 16 to 138 ng/µl in 2013. Hence, the lower diversity of OTUs observed in the 2013 samples was more likely a consequence of the decontamination treatment applied before DNA extraction than a result of non-uniform PCR amplification.

454 sequencing is also prone to biases that inflate the abundance of one group over the other because several sequences can be produced from the same DNA fragment. Sequence duplication happens during the emulsion PCR when amplified DNA binds to empty beads and during sequencing when the chromatic signal from an amplicon-containing well 'bleeds' into an empty well (Gomez-Alvarez *et al.*, 2009). An effective way to control for the 454 biases and improve quantitative prediction is to remove duplicated sequences (Gomez-Alvarez *et al.*, 2009; Mariette *et al.*, 2011). Therefore, in both pyrosequencing analyses (2010 and 2013) all exact duplicates (completely identical sequences) were removed from our datasets prior to alignment, by using the function 'unique.seqs' in mothur.

Yet, in a highly homogeneous library (*e.g.* a library constructed from a monoclonal culture), one would expect the high abundance of identical genetic material to result in a high observed duplication rate. The duplication screening

might therefore result in an over representation of sequencing errors (*i.e.* insertions, deletions, ambiguous bases and chimeras) in the remaining data. These mutations will usually change the Hamming distance between two highly similar sequences and therefore result in an over estimation of the sample diversity (Tedersoo *et al.*, 2010). In their study, Gomez-Alvarez *et al*, (2009) found that between 11% and 35% of the pyrosequencing libraries from metagenomic samples of soil and seawater resulted from artificial duplication. Exact duplicates represented between 3% and 24% of their libraries. In the 2013 dataset, the exact duplication rate ranges from 29% to 71%, and it was over 90% in the 2010 dataset (Forget *et al.*, 2014), suggesting that the trophosome of *Ridgeia piscesae* is composed of symbionts that are highly homogeneous genetically. To reduce the effect of sequencing errors on our estimations of diversity, we screened for chimeras using the program Uchime (Edgar *et al.*, 2011) and we grouped together sequences with less than two percent divergence (*i.e.* Hamming distances inferior to 5 for 250bp sequences) prior to taxonomic affiliations using the 'pre.cluster' command in mothur. This step might artificially increase the differences between groups of sequences and might explain why the taxonomic affiliations are not corroborated by the phylogenetic tree.

To conclude, the pyrosequencing libraries constructed from the 2013 collections support the idea discussed in Forget *et al.* (2014) that the only symbionts of *Ridgeia piscesae* are *Gammaproteobacteria*. The 2013 pyrosequencing libraries seemed to be free of contamination, but the diversity analysis using mothur did not permit resolution of the genetic structure of *Ridgeia piscesae* trophosome flora; that is, distinguishing a monoclonal trophosome microflora from one composed of genotypically different bacteria.

In the next chapter, I will further investigate the genetic diversity of *Ridgeia* trophosome flora using a different approach and additional molecular data.

Chapter 3. Is the trophosome of *Ridgeia piscesae* monoclonal?

Abstract

The hydrothermal vent tubeworm *Ridgeia piscesae* relies on intracellular chemolithotrophic symbionts for its nutrition. Yet, little is known about symbiont diversity within and between individual worms. We report molecular evidence for multiple genotypes of very closely related symbionts within the trophosome of the *R. piscesae*. We examined the compositional and structural variations of CRISPR spacers as well as the distribution of genotypic variants (insertions, deletions, and substitutions) in whole genome shotgun sequences of symbionts from the trophosome of (1) a unique individual *R. piscesae* and (2) pooled sequences of five other tubeworms of the same species. Observed heterogeneity is unlikely to be the result of 'somatic' point or structural mutations of a monoclonal symbiont lineage. Finally we examined single nucleotide polymorphisms (SNPs) in pyrosequences of the highly variable regions V1 to V3 of the symbiont 16S rRNA gene across 31 individual hosts collected in 2010, 2011, and 2013 from two vent sites. Most of the identified SNPs were found in more than one individual, and two were found across the samples from the three collection years. Two of the identified SNPs were also present in metagenomic data generated from in high-throughput sequencing of trophosome material from an individual *R. piscesae*.

3.1. Introduction

Symbioses of metazoans with more than one group of chemolithoautotrophic bacteria have been described in invertebrates collected from cold seeps (Duperron *et al.*, 2005, 2009; Fujiwara *et al.*, 2001; Kimura *et al.*, 2003), and hydrothermal vent habitats (Duperron *et al.*, 2008; Petersen *et al.*, 2010; Grzymski *et al.*, 2008; Zimmermann *et al.*, 2014). While the role of the different symbionts remains unclear in most cases, some studies have successfully demonstrated that the different partners can allow the host to benefit from more than one type of energy metabolism and thereby colonize a broader range of reducing habitats (Woyke *et al.*, 2006; Kleiner *et al.*, 2012a; Duperron *et al.*, 2006). Siboglinid tubeworms colonize a broad range of habitats at hydrothermal vents in the eastern Pacific Ocean, yet up to six of these tubeworm species (*Riftia pachyptila, Tevnia jerichonana, Ridgeia piscesae*; plus possibly *Oasisia alvinae*, *Escarpia spicata*, and *Lamellibrachia sp*) have all been found to host the same symbiont, the *Gammaproteobacteria* Candidatus *Endoriftia persephone*, in their trophosome. *Ridgeia piscesae*, the vestimentiferan found at hydrothermal vents in the Northeast Pacific, has the broadest habitat range of any vestimentiferan known to date (Bright and Lallier, 2010). Several studies have explored the possibility that *R. piscesae* hosts multiple symbiont species, but none has yielded definitive evidence (Chao *et al.*, 2007; Forget *et al.*, 2014). The discovery of two distinct *Gammaproteobacteria* phylotypes in a hydrothermal vent tubeworm from the Mediterranean Sea (Zimmermann *et al.*, 2014) has renewed interest in possibilities for symbiont diversity in the Eastern Pacific vestimentifera. This study follows up on preliminary evidence (Perez, unpublished data) for intra-specific variability of the *Gammaproteobacteria*: Candidatus *Endoriftia persephone* found in the trophosome of *Ridgeia piscesae*.

The *Endoriftia* symbionts are acquired de novo from the environment by each generation of tubeworms. The symbionts penetrate the worm tissues through

the epidermis and migrate to a region between the dorsal blood vessel and the foregut to form the proto-trophosome. As the metatrochophore larvae develop into adults, their digestive tract atrophies in favour of the trophosome that ends up occupying most of the space in the cœlomic cavity of the trunk (Nussbaumer *et al.*, 2006). Robidart *et al.* (2008) who published the first draft genome assembly of Candidatus *Endoriftia persephone* noted that the intra-trophosome genetic diversity in the *R. pachyptila* host was very low (0.29% genome-wide nucleotide heterogeneity (Robidart, 2006)) but up to 20 *Endoriftia* cells were found in the undifferentiated mesoderm tissues of the host larvae suggesting the vestimentifera are infected by several bacterial cells and therefore could host multiple environmental genotypes (Nussbaumer *et al.*, 2006).

In addition to potential intra-specific variability of *Endoriftia* within individual tubeworms, comparative analyses of metagenomic assemblies of symbionts hosted by three different species of host (*Ridgeia*, *Riftia* and *Tevnia*) suggest that the genetic makeup of symbiont populations varies between different hydrothermal vents and between populations associated with different hosts (see Chapter 4). The latter finding, together with recent evidence that viable (and presumably reproducing) populations of *Endoriftia* can be released into the environment following the death of individual tubeworms (Klose *et al.*, 2015), suggest a mechanism for maintaining genotypic diversity in free-living populations of *Endoriftia*.

Improving our knowledge of genetic diversity in Candidatus *Endoriftia persephone* is important to resolving the question of host specificity, understanding dynamics and interconnectivity of symbiont populations, and explaining the success of this bacterium in the Eastern Pacific Ocean. This study aimed to determine whether the trophosome of *R. piscesae* is monoclonal or composed of multiple symbiont genotypes.

To detect the presence of multiple bacterial strains/lineages, we first evaluated the compositional and structural homogeneity of CRISPR spacers within the metagenome of a single *R. piscesae* individual (referred as Symb 1) and in the pooled metagenomic data from five other tubeworms of the same species (referred as Symb_pool). CRISPR spacers are acquired by bacteria upon viral infection and thus form a record of strains' viral histories and can be used for high resolution genotyping (Pourcel *et al.*, 2005).Using the same two metagenomes, we then detected genomic variants (substitutions (SNPs), insertions and deletions (indels)) in whole genome shotgun sequences (Illumina HiSeq and MiSeq). Finally, as two heterozygous positions were found in the highly variable regions of the symbiont 16S rRNA gene, variants were also called from pyrosequences of the V1-V3 and V6-V8 variable regions of 30 additional individuals from ten independent tubeworm aggregations collected in 2010, 2011, and 2013.

3.2. Material and Methods

3.2.1. Sample collection

Specimens of *Ridgeia piscesae* were collected from Axial Volcano and the Main Endeavour Field on the Juan de Fuca Ridge during cruises on board the R/V Atlantis (July 1010) and the R/V Thomas G. Thompson (July 2011 and June 2013). In 2010 and 2011, the worms were recovered to the ship in sealed bioboxes, but the bioboxes were unsealed in 2013. Once on board, individual worms were carefully removed from their tubes and those presenting no tissue damage were retained and rinsed with 70% v/v ETOH to remove potential contamination. These worms were then flash frozen at -80°C until further processing in our laboratory where the contents of their trunks (which include the trophosome) was extracted from the bodies and rinsed with 70% ETOH before DNA extraction. In 2013 however, the worm trunks were dissected on the ship and their contents treated according to Carney *et al.* (2002) in order to remove potential epibiotic contamination. The trophosomes were rinsed with sterile seawater several times and incubated in a TE buffer (10:1 tris HCl/EDTA) and lysozyme (1mg/ml) solution at room temperature for 30 min. Then, the tissues were suspended in DNase and MgCl2 at 37°C for 5 min and washed several times with TE buffer (1:1 tris HCl/EDTA). Individual tissues were stored at -80°C and transferred to our laboratory for later DNA extraction. In 2010, 2011, and 2013, DNA from each dissected trunk was extracted using the Qiagen DNEasy Blood and Tissue Kit.

3.2.2. High throughput sequencing of the whole genome and 16S rRNA gene

In order to detect polymorphism within individual worm's symbiotic populations, the trunk DNA extracts were sent to Génome Québec and to the

Plateforme d'Analyses Génomiques (Institute of Integrative and Systems Biology, Laval University, Quebec City, QC, Canada) for deep sequencing with Illumina and 454 technologies, respectively (

Table 3.**1**).

Whole genome shotgun sequences (WGS) of six trunk samples were obtained from the Illumina HiSeq 2000 and Miseq platforms. Reads from one of these samples (we will call this metagenome "Symb_1") were assembled to reconstruct the whole symbiont genome which we used as a reference for polymorphism detection. This assembly called '*Ridgeia* 1 symbionts' is published on Genbank under the accession number LDXT01 and a detailed description of its reconstruction can be found in the Material and Methods of Chapter 4 (Section 4.2.1; p. 93).

Data from the five other samples were pooled into the dataset "Symb_pool" in order to increase the depth of coverage because these samples contained less symbiont DNA.

Additionally, pyrosequences of the PCR-amplified hypervariable regions V1-V3 and V6-V8 of the bacterial 16S rRNA gene were obtained for 29 trunk samples using the 454/Roche platform. Details of the library construction and sequencing are available in the Forget *et al.* (2014) paper that forms the first part of Chapter 2.

Table 3.1 Samples used in this study. MEF: Main Endeavour Field, AV: Axial Volcano, ETOH: Ethanol. *Data pooled together into Symb_pool metagenome

Sample name	Year of collection	Site of collection	Worm morphotype	Sequencing method	Tissue treatment before DNA extraction
Symb_1	2010	MEF	Short-Fat	WGS (Illumina HiSeq)	70% ETOH rinse
HF10AV2b3*	2010	AV	Short-Fat	WGS (Illumina HiSeq + Miseq), reads pooled in Symb_pool	70% ETOH rinse
LF10AV2b3*	2010	AV	Long-Skinny	WGS (Illumina HiSeq + Miseq), reads pooled in Symb_pool	70% ETOH rinse
LF10AV2bNR*	2010	AV	Long-Skinny	WGS (Illumina HiSeq + Miseq), reads pooled in Symb_pool + Pyrosequencing (V1-V3)	70% ETOH rinse
HF10HUb4*	2010	MEF	Short-Fat	WGS (Illumina HiSeq + Miseq), reads pooled in Symb_pool	70% ETOH rinse
LF10HUb3*	2010	MEF	Long-Skinny	WGS (Illumina HiSeq + Miseq), reads pooled in Symb_pool + Pyrosequencing (V1-V3)	70% ETOH rinse
LF10TPb5	2010	MEF	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
LF10HUb5	2010	MEF	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
LF10HUb4	2010	MEF	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
LF10HUb2	2010	MEF	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
LF10HUb1	2010	MEF	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
LF10AV1bNR	2010	AV	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
LF10AV1b5	2010	AV	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
LF10AV1b3	2010	AV	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
LF10AV1b2	2010	AV	Long-Skinny	Pyrosequencing (V1-V3)	70% ETOH rinse
HF11LBb2	2011	MEF	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF11GRb5	2011	MEF	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF11GRb4	2011	MEF	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF10HUb5	2010	MEF	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF10HUb1	2010	MEF	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF10AV2bNR	2010	AV	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF10AV2b5	2010	AV	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF10AV2b4	2010	AV	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF10AV2b3	2010	AV	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF10AV2b2	2010	AV	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
HF10AV2b1	2010	AV	Short-Fat	Pyrosequencing (V1-V3)	70% ETOH rinse
M1106	2013	MEF	Long-Skinny	Pyrosequencing (V1-V3 + V6-V8)	As Carney *et al.*(2002)
M1110	2013	MEF	Long-Skinny	Pyrosequencing (V1-V3)	As Carney *et al.*(2002)
M1112	2013	MEF	Long-Skinny	Pyrosequencing (V1-V3)	As Carney *et al.*(2002)
M1607	2013	MEF	Short-Fat	Pyrosequencing (V1-V3)	As Carney *et al.*(2002)
M1608	2013	MEF	Short-Fat	Pyrosequencing (V1-V3)	As Carney *et al.*(2002)
M1609	2013	MEF	Short-Fat	Pyrosequencing (V1-V3 + V6-V8)	As Carney *et al.*(2002)

3.2.3. Detection of CRISPR spacers heterogeneity

Two CRISPR arrays were found in *R. piscesae* symbionts but only one was completely reconstructed (see Chapter 4, Sections 4.3.3.1 and 4.4.1.2); the CRISPR array located on the contig Ga0074115_104:48218-48978 (start-end positions) in '*Ridgeia* 1 symbionts' (NCBI's accession LDXT01). Thus, we focused on spacers inserted in this array.

CRISPR spacers were detected directly from the unassembled reads of Symb_1 and Symb_pool metagenomes using Crass (Skennerton *et al.*, 2013). The Crass algorithm initially detects reads with repeat sequences that are characteristic of CRISPR arrays. From these partially overlapping reads, adjacent CRISPR spacers are identified and graphically represented as linked nodes in a spacer graph (see Figure C.1 in Appendix C; p. C.1, for guidance on the interpretation of Crass spacer graphs).

In order to prevent false detection of spacers, raw reads were first trimmed with PrinSeq v.0.20.4 (Schmieder and Edwards, 2011b) to remove bad quality bases (qual<20) at their 5' and 3' end. To reduce computation time, Symb_1 and Symb_pool were mapped onto the CRISPR array bearing reference contig Ga0074115_104. For each metagenome, reads that mapped to the reference were then extracted. Finally, we ran Crass on these metagenome subsamples with the –K option conservatively set to 20, meaning that large overlaps were required to join spacers together.

3.2.4. Detection of genetic polymorphism

We used HaplotypeCaller from the Genome Analysis Tool Kit (GATK) (Van der Auwera *et al.*, 2002) and VarScan (Koboldt *et al.*, 2009) to detect variants in both WGS and pyrosequencing data. Samples were analysed independently

rather than batched for better sensitivity of the variant calling methods (Cheng *et al.*, 2014).

Before calling the variants, the sequences were preprocessed in order to 1) remove potential contaminants and bad quality sequences, 2) remove artefactual sequences, and 3) generate high quality alignments of the reads onto the reference genome sequences. Because the two types of sequences (WGS and pyrosequences) were generated by different technologies (Illumina and 454/Roche, respectively), they had fundamentally different characteristics and necessitated different preprocessing pipelines (Figure 3.1).

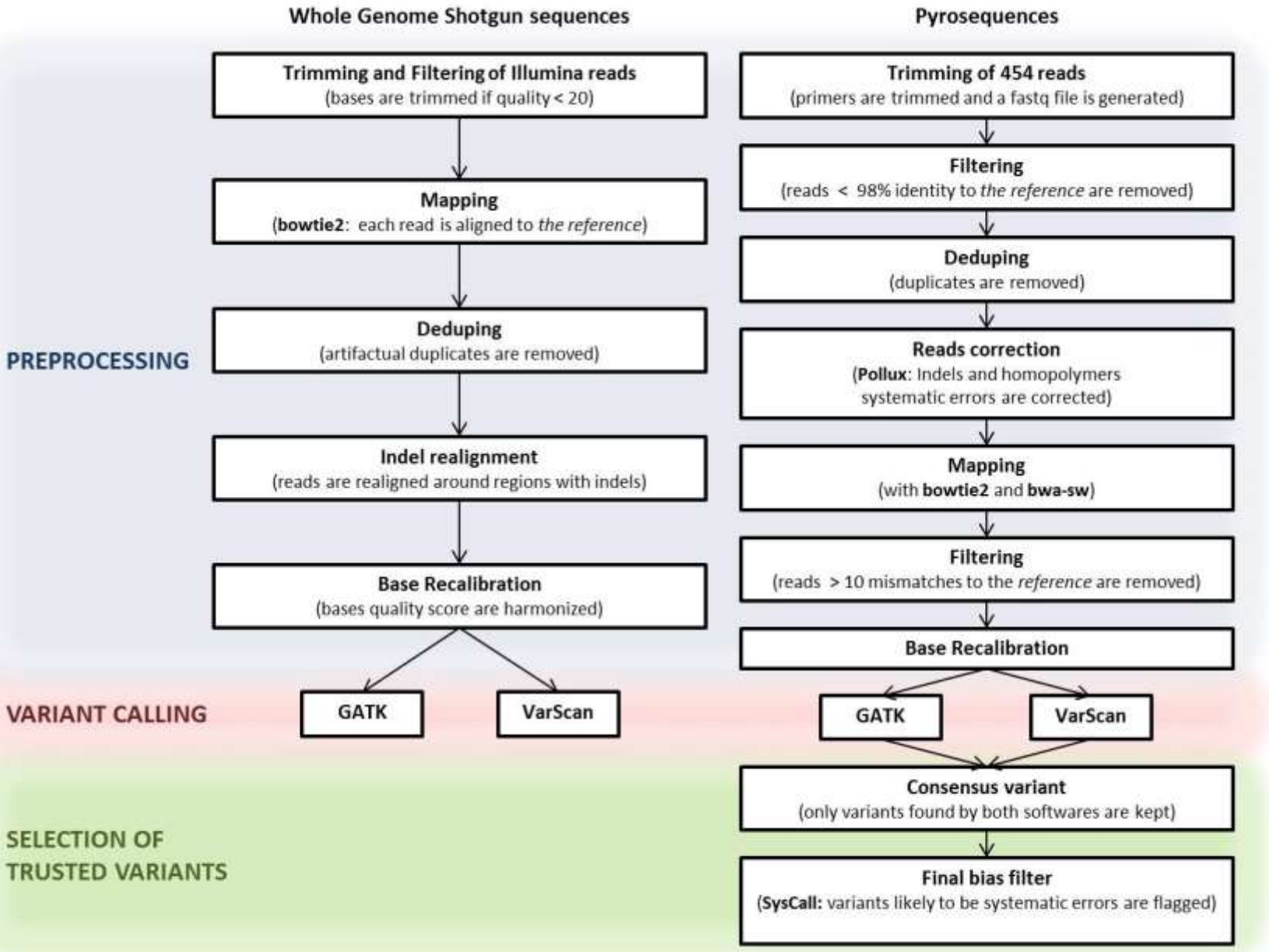

Figure 3.1 Variant calling pipelines for whole genome shotgun sequences and pyrosequences. See Section 3.2.4 for details and references.

3.2.4.1. Preprocessing of whole genome shotgun data (Illumina technology)

Quality-trimmed reads (see Section 0) were mapped onto the reference symbiont genome or onto the reference symbiont genome's 16S rRNA gene using the very-sensitive-local mode in bowtie2 v.2.2.4 (Langmead and Salzberg, 2012), one of the best performing tools for mapping Illumina-generated short reads (Wang *et al.*, 2013). A sorted bam file was generated with samtools (Li *et al.*, 2009). Using MarkDuplicate from Picard tools (http://picard.sourceforge.net), we flagged exact, 5', and optical duplicates that were, in the first two cases, biases generated during the emulsion PCR, and in the latter case, biases generated by leaking of the optical signal from a well to adjacent wells' photo-receptors. These reads were not considered during the variant calling step. Remaining mapped reads were realigned around indels with the GATK implemented tool IndelRealigner in order to correct for misalignments that lead to false detection of variants. Finally, base quality scores were recalibrated with GATK.

3.2.4.2. Preprocessing of pyrosequence data (454 technology)

454 sequences have different characteristics from Illumina sequences (see Figure C.3 in Appendix C; p. C.4), and different systematic biases. The most important one to consider here is the tendency of the 454 technology to poorly handle homopolymers. For example, a –CAAA- string can result in –CAAAA- or -CAA- (Kunin *et al.*, 2010). In addition to the homopolymer issue, samples from 2010 might be subject to more contamination due to different treatment before DNA extraction (as discussed in Chapter 2, Section 2.2.2; p. 53, Forget *et al.*,

2014). We therefore modified the preprocessing protocol to reduce both contamination biases and sequencing technology biases.

With the mothur command 'trim.seqs' (Schloss *et al.*, 2009), sequences in the fasta files and quality files were trimmed in parallel to remove the oligonucleotide primers (used during PCR) at their 5' end. Then, the fasta and quality files were combined into a fastq files using the command 'make.fastq' (Schloss *et al.*, 2009). We used DeconSeq (Schmieder and Edwards, 2011a) to remove any reads that had less than 98% identity to the reference (which corresponds to 10 variable positions in a 500bp alignments). The 98% threshold was chosen because it is high enough to exclude most of the non-symbiont 16S rRNA sequences but still include symbiont sequences with variable homopolymer lengths (bias subsequently corrected). Then, we removed exact duplicates using PrinSeq (Schmieder and Edwards, 2011b) and used Pollux (Marinier *et al.*, 2015) to correct for the systematic homopolymer errors of the 454 technology. Pollux corrects homopolymers in unaligned reads by generating k-mer frequency profiles and filtering out low frequency k-mers. Similarly to what Pollux creators found testing their tool on a real dataset, most of the corrections made here were additions or removals of one base in homopolymer strings (Marinier *et al.*, 2015). Following this, corrected sequences were mapped onto the reference using bowtie2 as well as bwa-sw (Li and Durbin, 2010) because bwa-sw has been found to perform better than bowtie2 on long reads (http://lh3lh3.users.sourceforge.net/alnROC.shtml). The very-sensitive-local mode was used with bowtie2, and bwa-sw was run with the default parameters. Additionally, we used bamtools' filter command (Barnett *et al.*, 2011) to discard reads with more than 10 mismatches to the reference 16S sequence and thus remove any remaining contaminant reads from the bam file. Finally, base quality scores were recalibrated as for the Illumina data.

3.2.4.3. Variant calling

Following the recommendation of Yu and Sun (2013), two independent variant calling software programs were used: HaplotypeCaller (GATK) and VarScan (Table 3.2). GATK was run with the default parameters. The parameters used for VarScan are presented in Table 3.3. The performance of the two has been compared many times with human genome data (Li and Homer, 2010; Cheng *et al.*, 2014; Pabinger *et al.*, 2014; Yi *et al.*, 2014; Mielczarek and Szyda, 2015; Huang *et al.*, 2015). While the probabilistic approach of GATK seems to perform better than VarScan, it might not be the best suited for non-diploid genomes.

Table 3.2 Comparison of the two variant caller algorithms used.

	HaplotypeCaller (GATK)	VarScan
Variant calling method	Probabilistic	Heuristic
Algorithm workflow	Finds regions of variation, breaks reads into kmers and reassemble them to identify different haplotypes	Reads from pileup file, discard reads that aligned with low identity or to multiple places, the best alignment for each read is screened for variations in the sequences, variants detected in multiple reads are combined into SNPs and indels
PRO	Statistical approach, may yield less false positives	Intuitive, modulable, output easily interpretable
CONS	Higher risk of false negative, output somewhat difficult to interpret	May be too simple, may yield more false positives

Table 3.3 VarScan parameters used in this study.

Option	Parameter used	Option description [1]
--min-coverage	10	Minimum read depth at a position to make a call
--min-reads2	2	Minimum supporting reads at a position to call variants
--min-avg-qual	20	Minimum base quality at a position to count a read
--min-var-freq	0.01	Minimum variant allele frequency threshold
--min-freq-for-hom	0.9	Minimum frequency to call homozygote
--p-value	0.05	Default p-value threshold for calling variants
--strand-filter	1 (on)	Ignore variants with >90% support on one strand

[1] from VarScan's User manual: http://varscan.sourceforge.net/using-varscan.html

3.2.4.4. Post processing steps

The results from both variant calling software packages were compared together with WGS data. Qualitative information about the variants (positions in genome, effect of substitutions on amino acids etc.) were obtained with snpEff (Cingolani *et al.*, 2012).

For the pyrosequencing data, we used the most conservative approach. Insertions and deletions were not considered and we called a SNP only if detected by both GATK and VarScan. In addition, we used a third bias filter on pyrosequences to flag potential false positives : SysCall (Meacham *et al.*, 2011). SysCall input files were generated with a custom python script. Statistical analyses and further data processing were performed in R (Ihaka and Gentleman, 1996). We used Hartigan's dip test to test for multimodality of variant frequency distribitions (Maechler, 2013), and Wilcoxon rank sum test to test if variant frequencies in Symb_1 were varied significantly from those in Symb_pool (Hollander *et al.*, 2013).

3.3. Results and Discussion: Evidences for multiple genotypes in *R. piscesae*

3.3.1. CRISPR spacer heterogeneity.

Figure 3.2 shows the spacer graphs generated by Crass from the unassembled reads of Symb_1 (left) and Symb_pool (right) metagenomes for the CRISPR array located on the contig Ga0074115_104:48218-48978 (start-end positions) in *Ridgeia* 1 symbiont genome (NCBI acquisition number LDXT01). These graphs represent the sequential organization of spacers in the CRISPR array. Spacers are represented as nodes and are linked to each-other by vectors indicating their relative arrangement in the 3'-5' direction. Sequences flanking the array such as leader sequences, next to which new spacers are integrated in arrays, are represented by diamond shaped nodes. A list of all nodes (flanker and spacers) is provided in Appendix C, List C.1; p. C.3.

A greater diversity of spacers was observed in the Symb_pool symbiont population compared to that of Symb_1 (14 and 46 unique spacers were found in the Symb_1 and Symb_pool symbiont populations, respectively). This could be associated with a greater diversity of symbiont lineages in Symb_pool which constitutes a priori a larger and more heterogeneous population; Symb_pool spacers came from symbionts associated with 5 individual worms from both the Main Endeavour Field and Axial Volcano (

Table 3.**1**).

Chromosomic rearrangements within the CRISPR array were observed in both metagenomes. In the Symb_pool metagenome, branching connections to spacers 184 and 208 were found for spacer 55 while in the Symb_1 metagenome, reads containing the spacer 440 could be linked to spacers 47 or 179. The apparent deletion of the spacers 47 and 37 in some of the reads seemed to corroborate with the abundance of mapped paired-end reads with abnormally large insert sizes (Figure 3.3). However, these alternative links were supported by a very small proportion of reads (Figure 3.2) suggesting one dominant symbiont strain. In addition, these alternative links could (though it is very unlikely) be the result chimeric binding of two independent DNA strands (Quail *et al.*, 2008).

11 spacers were shared between Symb_1 and Symb_pool (Figure 3.2, see also Figure C.2 in Appendix C; p. B.2). These spacers had the highest coverage in both metagenomes, indicating once more dominance of one or a few *Endoriftia* lineages.

Tracing back the shared spacers to the individual samples combined into Symb_pool revealed that their apparent dominance did not result from the overrepresentation of one single worm's microflora in the Symb_pool metagenome; the dominant spacers were found in two to five of the individual worm metagenomes (see Figure C.2 in Appendix C; p. C.2).

The CRISPR arrays of Symb_1 and Symb_pool seemed to be conserved at their 3' end (older spacers) but divergent at the 5' end (where new spacers are integrated). This could indicate lineages that have recently diverged (Pourcel *et al.*, 2005). For comparison, in *Tevnia pachyptila*, a cousin species of *R. piscesae* that inhabits vents on the East Pacific Rise and host *Endoriftia* symbionts that are predicted to have diverged from the *Ridgeia* ones about 28.5 Mya (See Chapter 4, Section 4.2.2.4; p. 100), a completely different set of spacers were found for this array (see Chapter 4, Section 4.4.1.2; p. 115).

Interestingly, the spacer 295 was detected at the 3' end of spacer 58 in Symb_pool but not in Symb_1 reads. However, this link was found in Symb_1 when reads were assembled into the consensus genome '*Ridgeia* 1 symbionts' (Figure 3.3). We suspect that some reads were lost during the reads-subsampling step and that those remaining did not sufficiently overlap spacer regions for Crass to recognize links between them. Additionally, we noticed that mutations accumulated in the repeats sequences located further on the 3' end of the CRISPR array. These degenerate repeats are typical of CRISPR arrays (Horvath *et al.*, 2008) and can explain why spacers detected by crass in the Symb_1 genome do not span the whole CRISPR array (Figure 3.3).

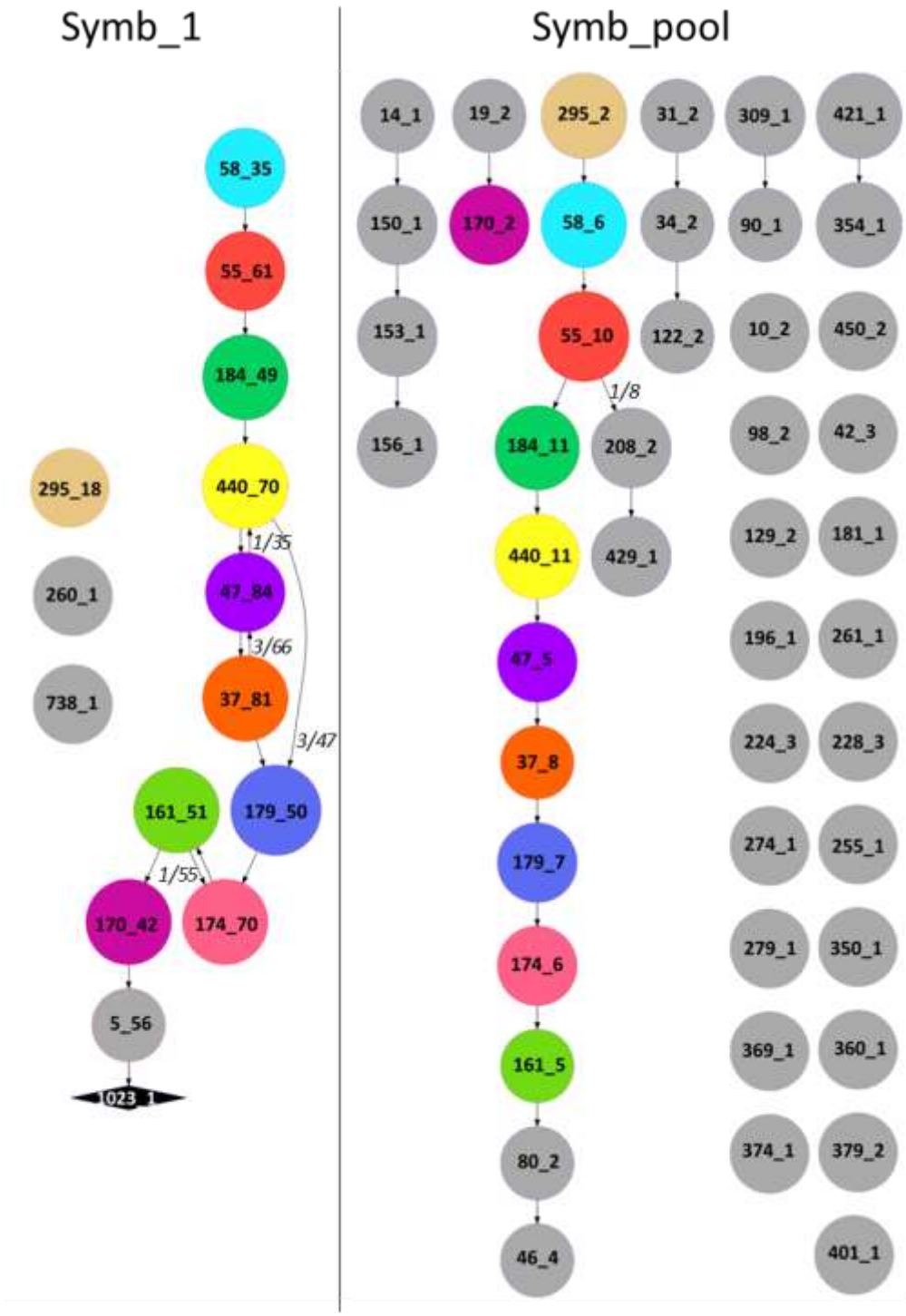

Figure 3.2 CRISPR spacers found in Symb_1 (left) and Symb_pool (right) for the CRISPR array located on the contig Ga0074115_104:48218-48978 (start-end positions) in *Ridgeia* 1 symbionts. The diamond shaped node represents the leader sequence on the 3' end of the CRISPR array while circular nodes represent CRISPR spacers. Nodes are labeled as follows: NodeID_coverage, with NodeID = randomly generated unique identification number. Identical spacers retrieved from the metagenomes of Symb_1 and Symb_pool are identified with matching colours. Unique spacers are in grey. Arrows indicate the 5'-3' order of spacers in between the CRISPR repeats CGGTTCATCCCCGCGGGTGCGGGGAACAC. Proportions of reads supporting alternative links are indicated in italic.

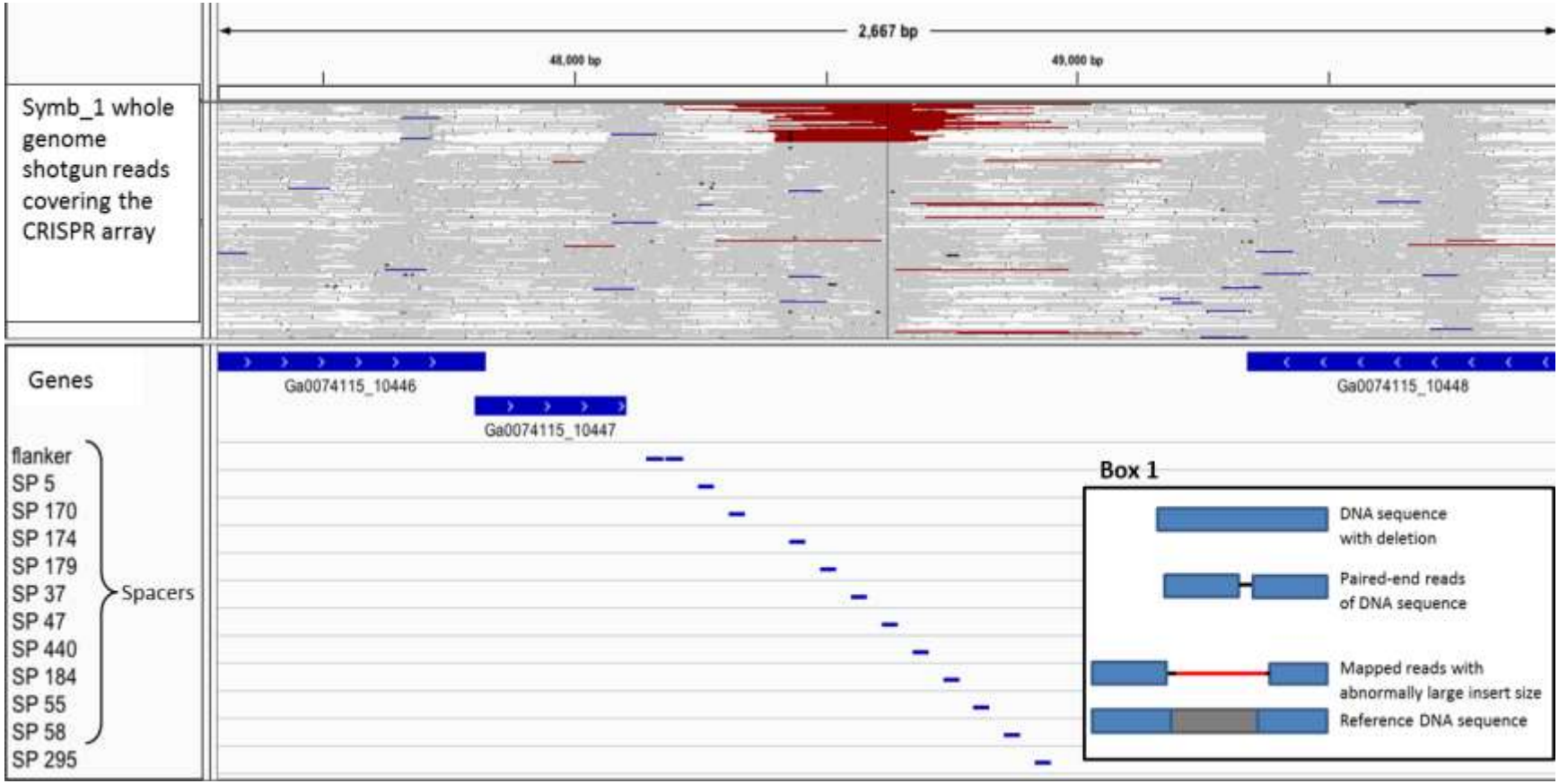

Figure 3.3 Unassembled read pairs from the Symb_1 metagenome mapped onto the reference contig Ga0074115_104 (*Ridgeia* 1 symbiont). Only the portion of the contig containing the CRISPR array is visible. Read pairs with abnormally large insert sizes (1% percentile) are coloured in red. To help interpretation, a schematic representation of a deletion is presented in Box 1.

3.3.2. Genetic variants in whole genome shotgun sequences

3.3.2.1. Variant frequency distributions in two symbiont metagenomes.

Whole genome shotgun sequences were obtained from two endosymbiont populations extracted from the hosting organ (trophosome) of one (Symb_1) and five (Symb_pool) *Ridgeia piscesae* individuals, respectively. Genetic variants (single nucleotide substitutions (SNPs), insertions and deletions) were detected in both metagenomes using two independent algorithms: VarScan and GATK (see Material and Methods).

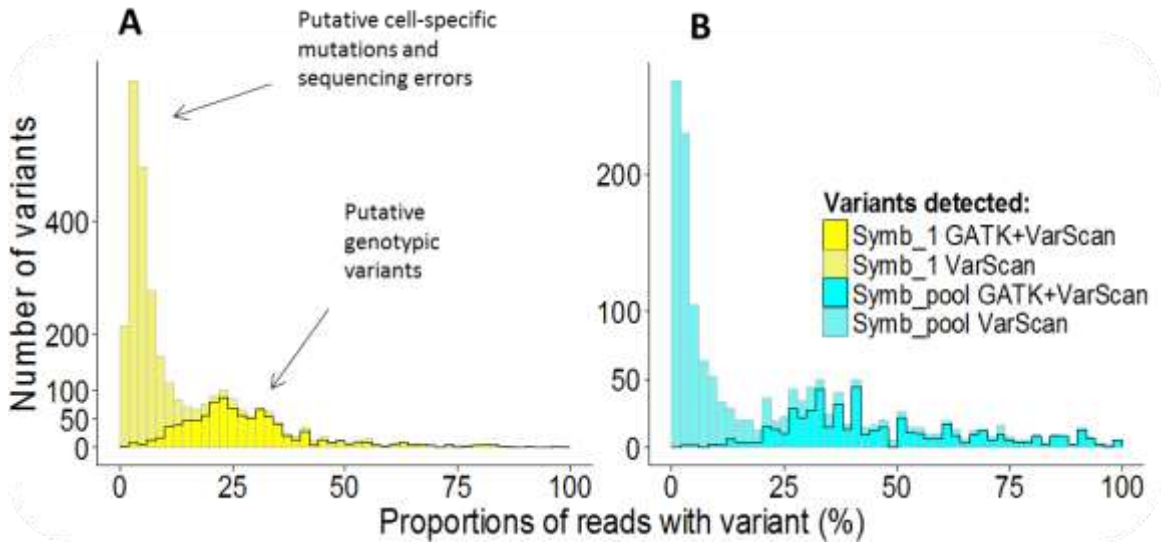

Figure 3.4 Frequency spectrum of variants in A) the Symb_1 and B) Symb_pool metagenomes.

GATK and VarScan produced notably different outputs corroborating the findings of Yu and Sun (2013) that the agreement between different variant calling algorithms is low. For the Symb_1 dataset, VarScan identified >2x more variants than GATK, and while the majority of variants identified by both methods were substitutions (SNPs) (SNP:Indel ratio = 8:1), more SNPs were identified by VarScan than by GATK (Figure 3.5).

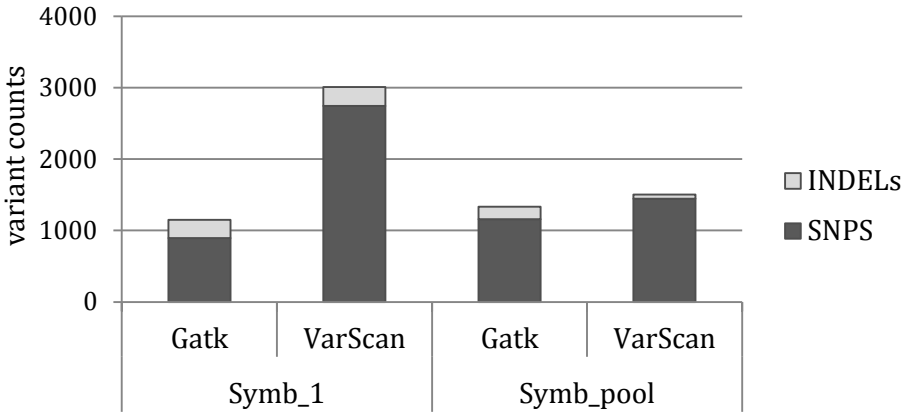

Figure 3.5 Comparisons of variants detected by VarScan and GATK in the metagenomes of Symb_1 and Symb_pool. Indel: Insertions and deletions, SNPs: Single nucleotide polymorphism; substitutions.

The VarScan and GATK methods both yielded variant frequency distributions that were multimodal (Hartigans' dip test p-value <0.02) for all datasets except for the Symb_1-VarScan dataset (Hartigans' dip test p value = 0.3) (Figure 3.4). In a monoclonal population of bacteria, variants result from point mutations in individual cells. Thus, the population metagenome would be composed of a majority of rare (low frequency/small proportion) variants. In the presence of multiple genotypes however, variants result from the genetic divergence between the reference genotype and the alternative genotypes. As a consequence, the frequencies of the variants are directly linked to the frequencies of their respective genotypes in the metagenomic population.

Hence, the multimodal distribution of variants in Symb_1 and Symb_pool suggested the presence of multiple genotypes.

Interestingly, the variants identified by VarScan in the metagenomic data Symb_1 and Symb_pool tended to be present in a much lower proportion of reads than the variants identified by GATK (Figure 3.4) indicating that VarScan might have detected somatic mutations and sequencing errors that were discarded by GATK. Somatic mutations and sequencing errors are more likely to be neutral due to lack of selective sweep while genotypic mutations should bear signs of purifying selection. Hence, artefactual variants should be uniformly distributed along the genome while genotypic variants should be under-represented in the coding regions which are typically under strong purifying selection (Yu *et al.*, 2015; Durbin *et al.*, 2010).

Variants called by both VarScan and GATK were slightly over-represented in intergenic regions (75 to 81% of variants were in coding sequences; the reference genome had 89% coding bases) but variants identified by VarScan only were even more over-represented in non-coding sequences (only 23 to 51% were in coding sequences) (Figure 3.6 A). However, these variants were localized on contigs that had lower gene densities.

Additionally, the SNPs identified by VarScan and GATK had a higher transition-transversion ratio (ts/tv) than those identified by VarScan only (ts/tv≈4.5 and ts/tv≈1.2, respectively) (Figure 3.6 B) and contained a lower proportion of non-synonymous substitutions (Figure 3.6 C). This bias towards transitions and synonymous substitutions is expected for genotypic variants but not for somatic mutations and sequencing errors that have a near-neutral ratio of transitions over transversions (ts/tv=0.5; for each nucleotide, one transition and two transversions are possible) (Wielgoss *et al.*, 2011; Dohm *et al.*, 2008)

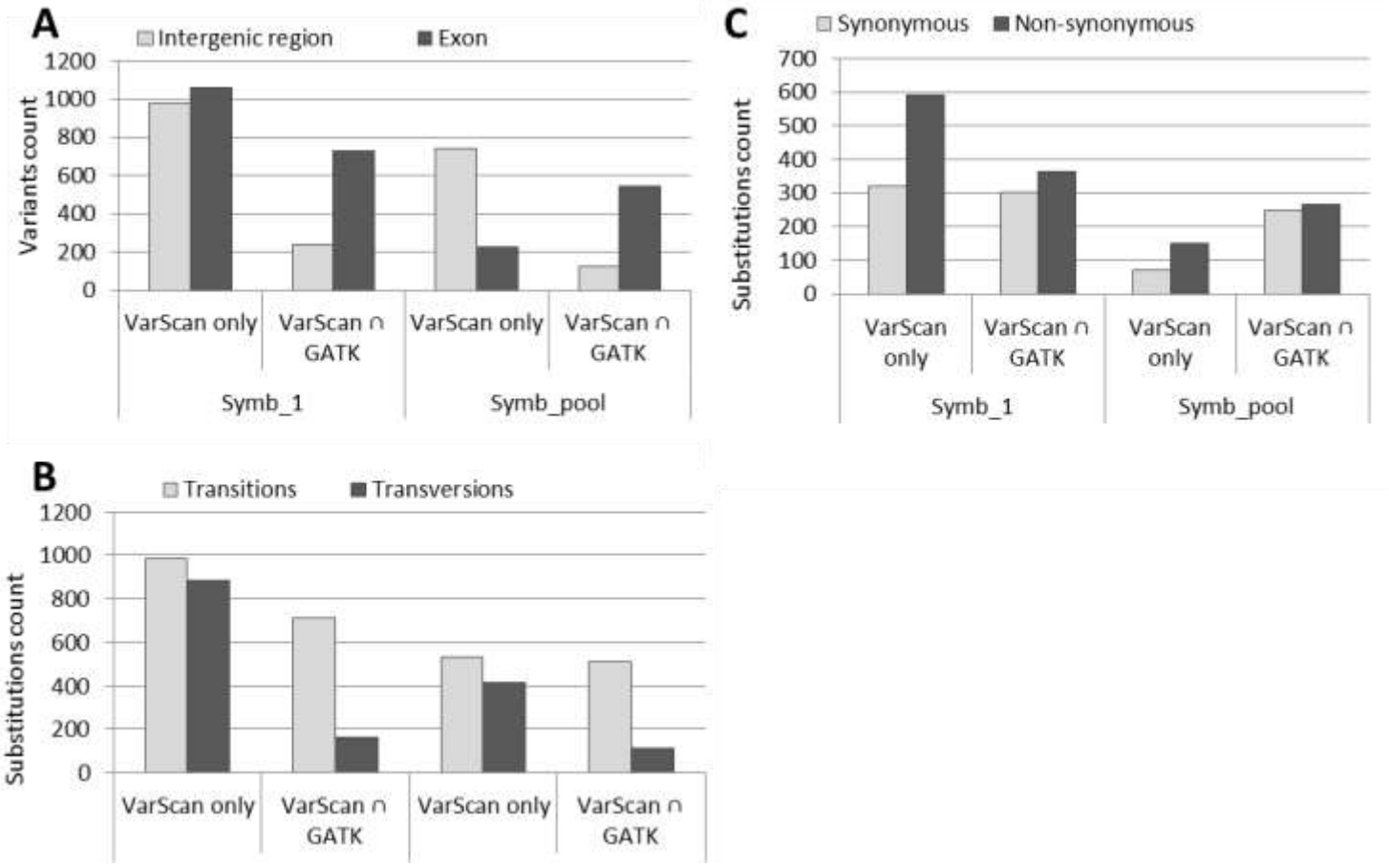

Figure 3.6 Comparisons of variants detected by VarScan only or both VarScan and GATK; A) variant positions in the genome (inside or outside coding regions), B) types of substitution (transition vs transversion), C) substitution effects on amino acid sequence.

3.3.2.2. Comparison of the two metagenomic symbiont populations (Symb_1 and Symb_pool).

Overall, two times fewer variants were detected in the Symb_pool metagenome because of the lower depth of coverage (Barrick and Lenski, 2009). Nevertheless, 10% of the variants detected by both VarScan and GATK were found in both the Symb_1 and Symb_pool metagenomes and this number increased to 17% for variants detected with either method (Figure 3.7). A considerable number of variants were shared between the two symbiont metagenomes. Because these two datasets represented independent samples, it is very unlikely that these variants were somatic.

In the Sym_1 metagenome, variants detected by the two variant calling algorithms were found in 27% of the reads on average while in the Symb_pool dataset, variants were present in a significantly larger proportions of reads (average frequency= 36%, Wilcoxon test p-value<0.001).

For the Symb_pool metagenome, 137 of the 495 variants detected by both pipelines were represented in more than 50% of the reads, meaning that for a little more than a quarter of the variant positions, the dominant allele in the symbiont population Symb_pool was different from the one in the reference symbiont genome which was assembled from the Symb_1 metagenome. This could be the result of (1) a change in the relative abundance of the symbiont genotypes in the Symb_pool relative to the Symb_1 population (54/137 variants were found at frequencies lower than 50% in Symb_1 (Figure 3.8)), or (2) the presence of one or several additional genotypes that were absent in the Symb_1 metagenome (65/137 variants were not found in the Symb_1 metagenome).

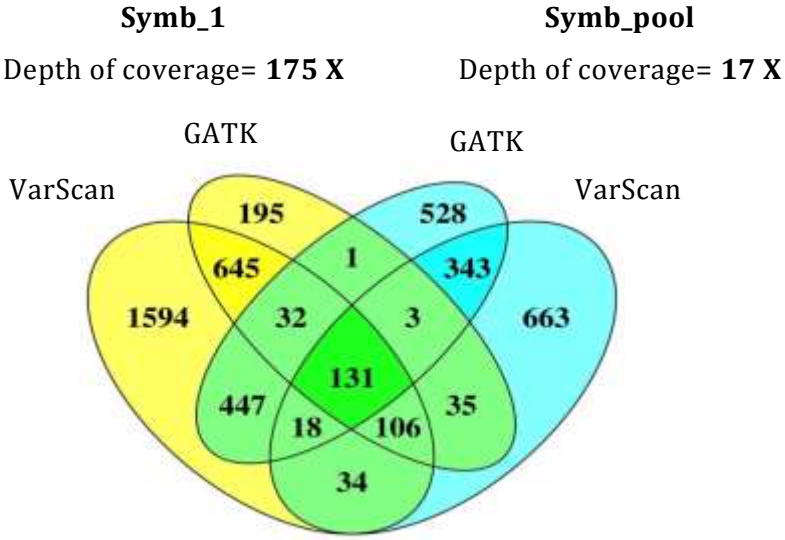

Figure 3.7 Variants detected in Symb_1 (yellow), Symb_pool (blue), and both (green). Note that fewer variants were detected in Symb_pool due to low depth of coverage (17X).

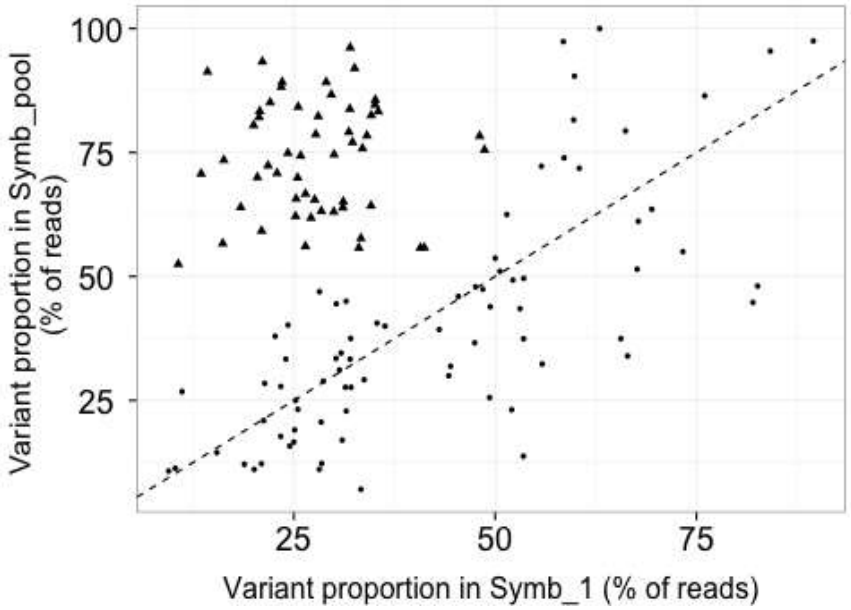

Figure 3.8 Correlation of variant frequencies in Symb_1 and Symb_pool. Dash-line: 1:1.Triangles: variants present in >50% of the reads in Symb_pool but <50% of the reads in Symb_1.

3.3.3. Genetic variants in pyrosequences of variable regions of the 16S rRNA gene

454 deep sequencing of the V1-V3 and V6-V8 regions of the 16S rRNA gene were obtained for the intracellular symbiont populations of 31 individual tubeworms (Figure 3.9). For these pyrosequences, reads were mapped with bwa-sw in addition to bowtie2. Variant calling was performed by both VarScan and GATK, but we conservatively retained only variants detected by both pipelines. In addition, we used SysCall to identify potential sequencing systematic errors. Because the SysCall algorithm has not been extensively tested on real data, SysCall flags should be considered with caution. Finally, only SNPs were considered because of the higher sequencing error rate on indels (Loman *et al.*, 2012).

Overall, Bwa-sw and Bowtie2 produced very similar results. The main difference between the two mappers was the absence of the SNP in position 514 in all MID-tag samples. Examination of the mapped reads revealed that this was due to bwa-sw hard clipping the reads at their 3' ends.

The V6-V8 region of the 16S rRNA gene was only sequenced for two of our samples. One of them had the same SNP found in the Symb_1 metagenome (position 1116), though flagged as erroneous by SysCall. In the V1-V3 region, the SNP at position 179 was found in similar abundances in both the metagenomic data of Symb_1 16S rRNA and in 7 out of 19 individuals from the Main Endeavour Field suggesting the presence of an alternative genotype in these samples. Additionally, this variant (along with the variant at position 514) was found again in a few Sanger sequences of the whole 16S rRNA gene (data not shown). This SNP was not found in the samples from Axial Volcano.

Other SNPs were also detected in similar abundance in several of the samples prior to the 2013 collection (MID- tag): SNPS at positions 195, 196, and 514. However, these SNPs were sometimes flagged as systematic error by SysCall

indicating that they might result from sequencing error. Furthermore, none of the SNPs called in the V1-V3 region were consistently linked to each other within the read sequences meaning that these SNPs could result from somatic compensatory mutations and/or that homologous recombination might happen between cells of different bacterial lineages within the worm host.

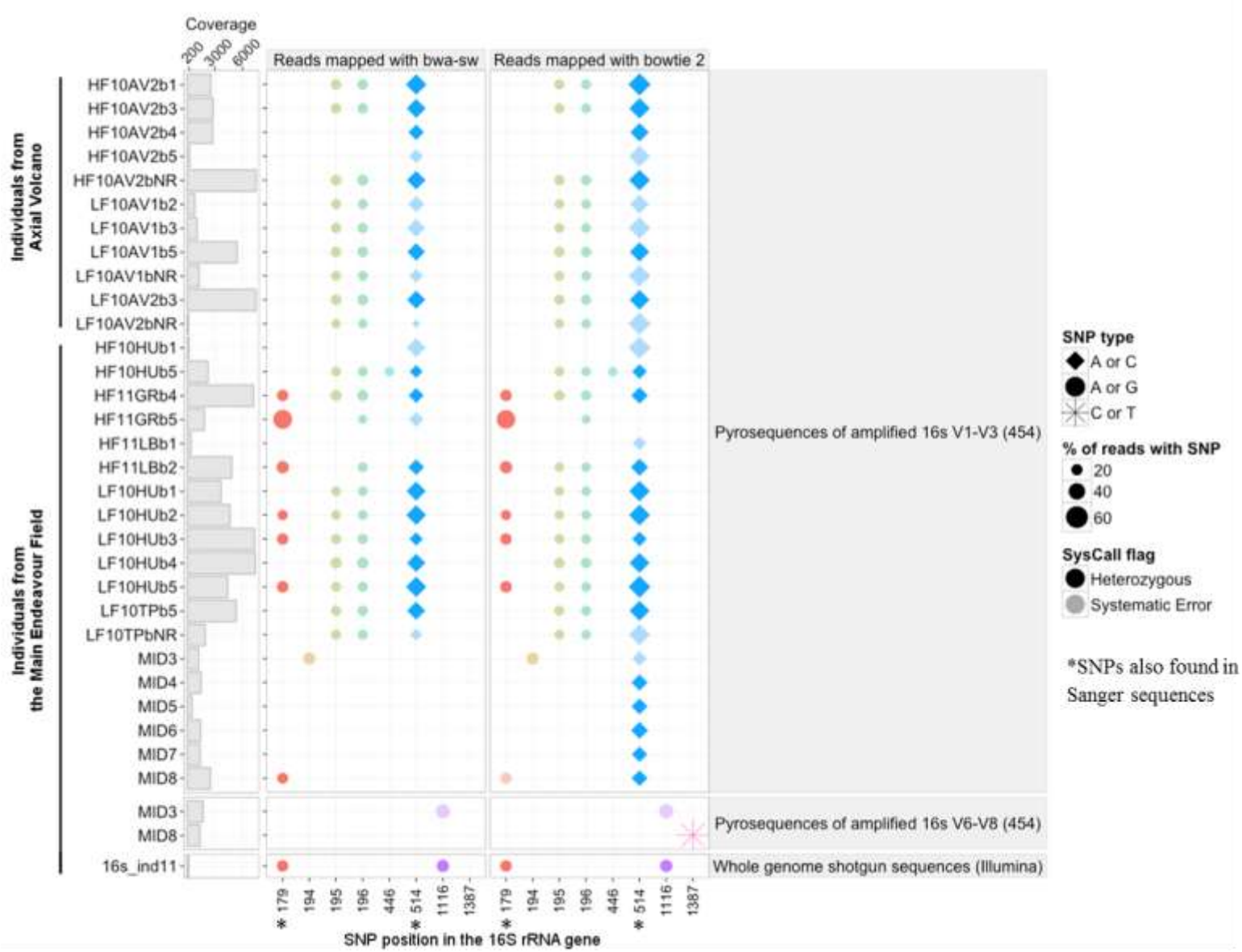

Figure 3.9 Single nucleotide polymorphism (SNPs) observed in the endosymbiont 16s rRNA genes from 31 *Ridgeia piscesae* tubeworms. Only SNPs present in >5% of the reads and detected by both GATK and VarScan variant callers are represented.

3.4. Conclusion and perspectives

Next generation sequencing technologies permit the genotyping of uncultivated microbial communities with unprecedented depth. Here, we used these deep-sequencing technologies to assess whether multiple genotypes of the *Gammaproteobacteria* Candidatus *Endoriftia persephone* were inhabiting the bacteriocytes of the tubeworm *Ridgeia piscesae.* Using conservative parameters, we found evidence for chromosomic rearrangement between CRISPR spacers of the two metagenomic symbionts populations (from one and 5 individuals). In a second line of evidence, the proportion and quality of some genetic variants found in these two metagenomes resembled those that would be issued from genetic polymorphism. Indeed, the putative genotypic variants seemed to be marked by purifying selection compared to the putative somatic variants, they were found in considerable abundance, and were often found in both metagenomes. Finally, the two SNPs in the 16S rRNA gene found in the symbiont metagenome Symb_1 from the Main Endeavour Field vents were also found in the pyrosequence libraries generated from other individual worms of this same hydrothermal vent site.

When considered independently, our individual lines of evidence do not permit definitive interpretation because of the many inestimable biases that can accumulate in the different analyses and contribute to both type I (*e.g.* false detection of links between CRISPR spacers due to chimeras, false detection of variant due to sequencing error or misalignment, compensatory mutations) and type II errors (*e.g.* missed detection of CRISPR spacer links due to lack of overlap between sequences, missed detection of variants due algorithm biases or low abundance). However, taken together, they constitute a more convincing body of proof that the trophosome of *R. piscesae* is not monoclonal but

composed of symbionts with different genotypes. Still, because of the small size of the reads, and the low genetic diversity observed in our metagenomes (1 variant every 1500 bp to 3500 bp depending on how conservatively we call genetic variants) we could not assess how many different genotypes were present and how divergent these genotypes might be. These questions will be more difficult to address given the possibility that *Endoriftia* populations might be partially sexual. Indeed, Stewart *et al.*(2009) suggested that the increased promiscuity between intracellular bacterial symbionts could favour homologous recombination thus increasing their genetic diversity and blurring lines between discrete bacterial lineages. Could the worm's trophosome act as diversity generators for free-living symbiont populations?

Our findings suggest that CRISPR arrays could be used for very fine scale genotyping of the symbionts and thus biogeographic distributions of different strains of *Endoriftia*. Such work has already been done for the pathogens *Yersinia pestis* (Pourcel *et al.*, 2005; Cui *et al.*, 2008), *Streptococcus agalactiae* (Lopez-Sanchez *et al.*, 2012), *Salmonella* sp. (Fabre *et al.*, 2012; Bachmann *et al.*, 2014), *Campylobacter jejuni* (Kovanen *et al.*, 2014), and *Escherichia coli* (Feng *et al.*, 2014) and for some free living microbial populations such as bloom-forming cyanobacteria (Kuno *et al.*, 2014) and *Sulfolobus islandicus* (Held *et al.*, 2010). These studies are often able to characterize diversity with a resolution higher than the strain level, raising questions about cell individuality (Tyson and Banfield, 2008). Further investigation of *Endoriftia* CRISPRs and coupling spacer diversity to genome-wide genetic diversity is required to confirm that CRISPRs can be used as proxies for detecting independent lineages. Understanding the mechanisms and dynamics of genetic diversity of these symbionts at the most basic level is important to understanding their evolution and predicting resilience of individual populations to changing environmental conditions.

Acknowledgments

We thank Nathalie Forget for sharing her pyrosequences libraries. We also thank the crews of the R/V Atlantis and R/V Thomas G. Thompson, the pilots of the submersibles Alvin and ROPOS and Oceaneering, Carol Doya, Emily Boulter, and Steven Hallam for their assistance during sample collections and sequencing of the metagenomes. This research was enabled from computing assistance provided by Belaid Moa, WestGrid (westgrid.ca) and Compute Canada/Calcul Canada (computecanada.ca). We also thank the members of Verena Tunnicliffe's lab, and all of the contributors to seqanswers.com and biostars.org.

This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Canadian Healthy Oceans Network (NSERC Canada) to S.K.J.

Chapter 4. Genome assembly for Candidatus *Endoriftia persephone* from Juan de Fuca Ridge tubeworm *Ridgeia piscesae* provides insight into symbiont population structure among three host species at eastern Pacific spreading centres.

This chapter has been submitted to Applied and Environmental Microbiology and is currently under review (manuscript # AEM00953-16).

Abstract

The symbiotic relationship between vestimentiferan tubeworms and their intracellular chemosynthetic bacteria is one of the more remarkable examples of adaptation to deep-sea hydrothermal vent environments. The tubeworm symbionts have never been cultured in the laboratory, but have been studied using molecular tools. Nucleotide sequences from the small subunit rRNA gene suggest that the intracellular symbionts of the eastern Pacific vent tubeworms *Oasisia alvinae*, *Riftia pachyptila, Tevnia jerichonana,* and *Ridgeia piscesae* belong to the same phylotype of *Gammaproteobacteria*; phylotype II. This bacterium, whose genome was first sequenced from *R. pachyptila* symbionts by Robidart *et al.* (2008), was named Candidatus *Endoriftia persephone.* In 2011, Gardebrecht *et al.* published high quality genomes of *R. pachyptila* and *T. jerichonana* endosymbionts and confirmed that these two tubeworm species from the East Pacific Rise share the same symbionts. Here, we present the first sequenced symbiont genome from the tubeworm *Ridgeia piscesae*, found at northeast Pacific vents. Two *Ridgeia* symbiont genomes were assembled from metagenomes of the symbiont-hosting organs of tubeworms from the Juan de Fuca Ridge (one and five individuals, respectively). We compared these assemblies to those of the sequenced *Riftia* and *Tevnia*'s symbionts. Pan-genome composition, genome wide comparisons of the nucleotide sequences, and

pairwise comparisons of 2313 orthologous genes indicated that *Endoriftia* symbionts are structured on large geographical scales but also at smaller scales and possibly through host specificity.

4.1. Introduction

A defining characteristic of hydrothermal vent ecosystems is the diversity and ubiquity of mutualistic partnerships between Metazoa (multicellular organisms) and chemolithoautotrophic Bacteria. Among these associations, one of the most remarkable is the well-studied, model symbiosis between the giant tubeworm *Riftia pachyptila* and its unique sulfide-oxidizing *Gammaproteobacteria* partner Candidatus *Endoriftia persephone* (Robidart *et al.*, 2008). These intracellular symbionts are hosted within the specialized cells (bacteriocytes) of an organ known as the trophosome that occupies most of the space in the cœlomic cavity of the animal's trunk. In this mutualistic association, the worm supplies the bacteria with the inorganic compounds necessary for coupling sulfide oxidation to CO_2 fixation: dioxygen, carbon dioxide and hydrogen sulfide (mostly HS^-). These substances diffuse across the gills into the blood of the animal and are then transported to the trophosome. In return, the endosymbionts provide the tubeworm with the organic molecules necessary for growth and metabolism either by excretion or by being directly digested (Felbeck and Jarchow, 1998; Bright *et al.*, 2000). The symbiotic bacteria are horizontally transmitted, that is to say, acquired de novo from the surrounding environment at each generation (Harmer *et al.*, 2008). The symbionts penetrate the worm tissues through the epidermis and migrate to a region between the dorsal blood vessel and the foregut to form the proto-trophosome. As the metatrochophore larvae develop into an adult, their digestive tract atrophies in favour of the trophosome (Nussbaumer *et al.*, 2006). The vestimentiferan adult thus becomes completely dependent on its bacteria for nutrition. For the symbionts however, this association seems facultative. Free-living *Endoriftia* symbionts have been detected in biofilms and seawater surrounding *R. pachyptila* aggregations (Harmer *et al.*, 2008) and recently, (Klose *et al.*, 2015) demonstrated that the

Riftia symbionts could return to their free-living stage upon death of the worm, thereby maintaining/sustaining environmental populations.

In addition to having a viable free-living stage, the symbionts exhibit low partner fidelity. *Endoriftia* is also associated with three to five other vent tubeworm species: *Tevnia jerichonana*, *Ridgeia piscesae*, *Oasisia alvinae*, and possibly *Escarpia spicata* and some *Lamellibrachia sp.* (Di Meo *et al.*, 2000), as evidenced by sequence analyses of the 16S rRNA gene marker along with the internal transcribed spacer (ITS) gene. This was somewhat surprising, given that these host species can be separated by thousands of kilometers of fragmented habitat and can colonize very different hydrothermal vent habitats (Bright and Lallier, 2010).

For example, *R. pachyptila* and *T. jerichonana* can both inhabit the same general vent locations in the East Pacific Rise (EPR) but thrive under contrasting venting conditions. *Tevnia* is typically found at sites of high hydrothermal discharge, characterized by low oxygen and high sulfide concentrations while *Riftia* flourishes in more diffuse flow, with higher oxygen and lower sulfide concentrations (Nees *et al.*, 2009). Further north, in the northeast Pacific Ocean, at the hydrothermal vents of Explorer Ridge, the Juan de Fuca Ridge (JdFR), and Gorda Ridge, the tubeworm species *Ridgeia piscesae* can be found in temperatures ranging from 2 to 30°C (Carney *et al.*, 2007; Urcuyo *et al.*, 2003) and sulfide concentrations at their branchial plumes ranging from <0.1μM to 200μM (Carney *et al.*, 2002; Urcuyo *et al.*, 2003; Brand *et al.*, 2007).

The symbionts' broad geographic distribution and the wide range of vent habitats occupied by their tubeworm hosts raises questions about the connectivity and the structure of *Endoriftia* populations. Limited data on *Endoriftia* populations indicate significant strain-level variations between symbionts from different geographical locations (Di Meo *et al.*, 2000) and, notably, that the symbionts associated with *Ridgeia* might belong to a different

population than those found in tubeworms from the EPR (Nelson and Fisher, 2000). However, these studies were based on comparison of only a few conserved genetic sequences.

Since the advent of accessible, high throughput sequencing, several *Endoriftia* draft genomes have been reconstructed from metagenomic sequences of *Riftia* and *Tevnia* trophosome extracts (Robidart *et al.*, 2008; Gardebrecht *et al.*, 2011). Assuming that the worm's trophosome is not monoclonal, an assembly essentially represents a consensus genome of the symbiont population inhabiting the host trophosome.

In this study, we sequenced and assembled consensus genomes representing the symbiont populations inhabiting the trophosomes of one and five individual *R. piscesae* worms, respectively. Upon confirmation that the *R. piscesae* symbionts indeed belonged to the same species as *Endoriftia*, we compared our genome assemblies to those previously published by Gardebrecht *et al.*, (2011) with the goals of characterizing *Endoriftia*'s pan genome and symbiont population structure in the different host species. For the latter, we undertook 1) genome wide comparisons of the nucleotide sequences of the core genome, 2) characterization of the composition of the accessory genomes, and 3) pairwise comparisons of 2313 putative orthologous genes.

These new genomic comparisons support the Nelson and Fisher (2000) hypothesis that the *Endoriftia* symbionts associated with *Ridgeia* belong to a different population than those on the EPR, and suggest that symbiont population structure may have habitat-specific or larger-scale spatial features within regions, and may be shaped by host selection.

4.2. Material and Methods

4.2.1. *Ridgeia* symbiont genome assembly

4.2.1.1. Samples collection

Specimens of *Ridgeia piscesae* were collected from Axial Volcano and the Main Endeavour Field, on the Juan de Fuca Ridge, during a remotely-operated vehicle (ROV) cruise on the R/V Thomas G. Thompson in July 2010. The worms were recovered to the ship in sealed bioboxes. Once on board, individual worms were carefully removed from their tubes and those show no visible tissue damage were rinsed with 70% v/v ETOH and flash frozen at -80°C until further processing. In our laboratory, the contents of the worms' trunks (that includes the trophosome) were dissected removed by dissection and rinsed with 70% ETOH before DNA extraction. Finally, the DNA from each dissected trunk was extracted using the Qiagen DNEasy Blood and Tissue Kit.

4.2.1.2. Whole Genome Sequencing and first data quality assessment

DNA extracts from six individual worms were sequenced and assembled at Genome Quebec (Table 4.1). Samples were prepared using standard protocols and sequenced on an Illumina HiSeq 2000 platform. A subset of these samples was also sequenced on the Illumina MiSeq platform. Raw reads were first assembled at Canada's Michael Smith Genome Sciences Centre using Abyss and preliminary quality assessments of the resulting scaffolds performed with Quast. All but one of the samples resulted in assemblies of poor quality (N50 < 1000). The very high quantity of scaffolds containing repetitive sequences, characteristic of eukaryote DNA, indicated significant contamination from worm genetic material. The highest quality sample (Ind 11; N50 = 61 758) had a bimodal distribution of GC% and a heterogeneous distribution of k-mer frequencies, evidence of a strong, distinct symbiont signal with fewer contaminating scaffolds than the other assemblies. This assembly was uploaded into the IMG-ER platform for gene calling and decontamination (Figure 4.1).

4.2.1.3. Decontamination

The Abyss assembly of Ind 11 was manually curated from potential contaminating scaffolds using a combination of reference dependant and independent methods according to JGI's Single Cell Data Decontamination guide. First, we considered scaffolds of lengths ≥ 1000 bp and used the IMG/ER embedded k-mer Frequency Analysis tool to generate a k-mer plot using the following parameters; window size=1000, fragment step=100, oligomer size=4 bp, minimum variation=10. Scaffolds out or extending out of the main scaffold cluster were tagged as contaminants if their respective gene counts were ≤ 1 (large intergenic sequences are not expected in bacterial genomes which have high gene density) or if they did not contained at least one gene phylogenetically affiliated to the genome of the close relative *Riftia pachyptila* symbiont (Robidart *et al.*, 2008). Contaminant scaffolds were then removed from the genome assembly. Finally, scaffolds smaller than 1000 bp with at least one gene phylogenetically affiliated to the genome of the *Riftia pachyptila* symbiont were added to the curated assembly. The resulting curated assembly GC% frequency distribution was unimodal with a mode at 60% of GC as for the *Riftia Tevnia* symbionts genomes.

4.2.1.4. De novo assembly

Raw paired-end reads were filtered with Prinseq (Schmieder and Edwards, 2011b), to remove nucleotides with a quality score inferior to 20 on both ends of each read. Ind 11 reads were then mapped onto the curated assembly using Bowtie2 (Langmead and Salzberg, 2012) and the mapped reads were extracted with Samtools (Li *et al.*, 2009).

Next, these 'good' reads were assembled with SPAdes (Bankevich *et al.*, 2012) using the following k-mer sizes: 26, 55, 67, 85, 89, 95 & 99. Scaffolding was performed with SSPACE-standard (Boetzer *et al.*, 2011) and resulted in a high quality assembly of the metagenome of *Ridgeia* symbionts. We will refer to this assembly as the *Ridgeia 1* symbiont.

Finally, we used the *Ridgeia 1* symbiont assembly to map and extract the symbiont reads from the Hiseq and Miseq data from the five other samples. These data were pooled to increase the depth of coverage and the subsequent assembly quality. The reads were assembled with SPAdes using the same parameters as for *Ridgeia 1* symbiont but with the additional k-mer size 127 to account for the longer reads generated by the MiSeq platform. This pooled assembly is referred to below as the *Ridgeia 2* symbiont.

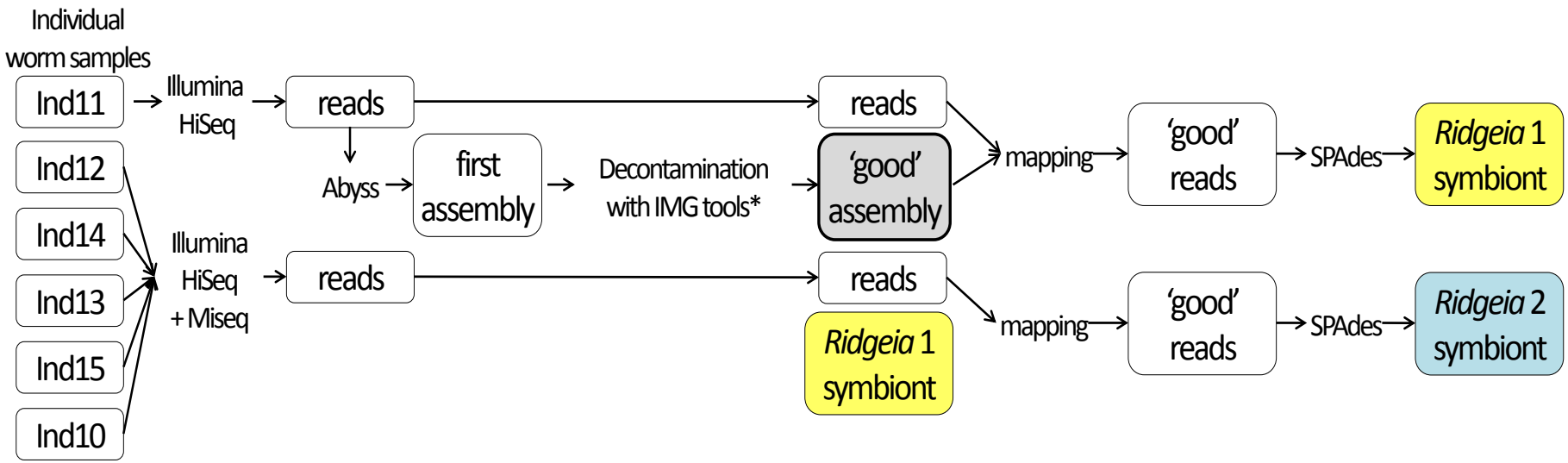

Figure 4.1 Graphical representation of the workflow for *Ridgeia*'s trunk sample metagenomes decontamination and *de novo* assembly. * IMG: Integrated Microbial Genomes System; sequences likely to be contaminants from the worm's DNA were found through tetramer frequency, gene content and phylogenetic affiliations of genes, and removed from the first assembly (see 4.2.1.3 Decontamination).

4.2.1.5. Gene annotations

Gene calling for the *Ridgeia 1* and *Ridgeia 2* symbiont assemblies was performed using the IMG-ER platform of the Joint Genome Institute. Both assemblies are published on the Joint Genome institute IMG with the submission IDs 65712 and 74895 as well as on GenBank under the accession numbers: LMXI00000000 and LDXT00000000, respectively. The versions described in this paper are LDXT01000000 and LMXI01000000.

Table 4.1 Metagenomic samples.

Sample name	Vent site	Flow regime Depth (m) plume/base max temperature (°C)	Sequencing platform	Symbiont reads * (% of total reads)	De novo assembly name
Ind 11	Main Endeavour Field (Hulk)	High-flow 2190 14.0/51.0	Hiseq	6 .94 M (33%)	*Ridgeia* 1 symbiont
Ind 13	Main Endeavour Field (Hulk)	High-flow 2190 14.0/51.0	Hiseq Miseq	0.18 M (<1%)	*Ridgeia* 2 symbiont
Ind 15	Main Endeavour Field (Hulk)	Low-flow 2191 2.5/2.5	Hiseq Miseq	0.11 M (<1%)	*Ridgeia* 2 symbiont
Ind 10	Axial Volcano (Hotspot 2)	Low-flow 1517 2.0/3.4	Hiseq	0.24 M (<1%)	*Ridgeia* 2 symbiont
Ind 12	Axial Volcano (Hotspot 2)	High-flow 1516 4.1/30.3	Hiseq Miseq	0.13 M (<1%)	*Ridgeia* 2 symbiont
Ind 14	Axial Volcano (Hotspot 2)	Low-flow 1517 2.0/3.4	Hiseq Miseq	0.29 M (<1%)	*Ridgeia* 2 symbiont

* based on alignment rates of reads mapped to *Endoriftia persephone* (from *Ridgeia* 1 assembly)

4.2.2. Genome-wide comparisons of *Ridgeia* symbionts with all other published Vestimentiferan genomes

4.2.2.1. Genome alignments

The genomes of the *Ridgeia 1* and *Ridgeia 2* symbionts were aligned with the closely related *Riftia 1*, *Riftia 2*, and *Tevnia* symbionts (Gardebrecht *et al.*, 2011) using Progressivemauve (Darling *et al.*, 2010). This anchored alignment algorithm finds so-called locally collinear blocks (LCB); genome segments that appear free of chromosomic rearrangement, and outputs the aligned sequences of each LCBs in XMFA multiple alignment format as well as a file containing the LCBs positions in each of the genomes (backbone file).

4.2.2.2. Pan-Genome composition

The pan-genome composition was determined by the presence/absence and size of LCBs sequences in each genome. As collinear blocks have previously been found to be informative for phylogenomic analysis (Zhang and Lin, 2012), we further used the presence/absence of individual LCBs (>100bp) to compute Jaccard distances with the vegan package in R (Oksanen *et al.*, 2015), and built a neighbor-joining tree using the software Populations V1.2.32 (Langella, 2002). Bootstrap values were obtained from 100 bootstrap subsamples using the function 'boot.phylo' from the 'ape' package in R (Paradis *et al.*, 2004). Finally, a custom Python script was used to extract from GenBank files, the annotations of all genes within the LCBs of interest. These genes were further annotated through visual inspection against the Mauve-generated multiple genome alignment in order to record additional information such as representation of neighbouring LCBs across the assemblies, nucleotide conservation, etc.

4.2.2.3. Core genome nucleotide heterogeneity

Our analysis of core genome nucleotide heterogeneity for symbionts from the three tubeworm species also included the (Robidart *et al.*, 2008) assembly for the *Riftia* symbionts. This first published assembly was not used in subsequent, more in-depth analyses because of its lower quality, as explained below.

We used the command 'stripSubsetLCBs' to extract the large (>100bp) LCBs represented in all of the assemblies from the xmfa file. For each LCB, a fasta file was generated and the sequences were aligned with MAFFT (Katoh and Standley, 2013). Subsequently, all the resulting alignments were concatenated to form a single genome-wide alignment of 2 580 528 bp with 75 472 variable sites. Finally, we used SeaView (Gouy *et al.*, 2010) to calculate the pairwise genetic distances using the HKY model (Hasegawa *et al.*, 1985) and generate a 100-bootstrap neighbor-joining tree.

4.2.2.4. Pairwise comparison of homologous genes

A file containing a table of all the homologous protein-coding genes was obtained using the command Export Positional Homologs from MAUVE's menu and the following parameters: min identity=80, min coverage=50. This table was then curated to only keep the entries of genes present in all of the genomes. Subsequently, we extracted in fasta format the nucleotide and amino acid sequences of these genes from the respective nucleic acid and protein databases of the pan-genome coding sequences using the blastdbcmd of blast++ (Wang *et al.*, 2003). Then, we aligned the amino acid sequences and generated protein sequence identity matrices with Clustal-Omega (Sievers and Higgins, 2014). Subsequently, protein alignments were converted into codon-based nucleotide alignments with pal2nal (Suyama *et al.*, 2006). Finally, the nucleotide sequence identity matrices and the dN/dS (ratio of divergence at nonsynonymous and synonymous nucleotide substitution sites) ratios were calculated using Clustal-Omega (Sievers and Higgins, 2014) and paml's YN00 (Yang and Nielsen, 2000;

Yang, 2007), respectively. Genome-wide dN/dS ratios were generated from the concatenated codon-based alignments.

Mauve's transitive algorithm identifies positionally homologous sequences. In closely related genomes, these positional homologs are also orthologs but the algorithm could still mistakenly catch recently duplicated genes (paralogs). To prevent comparisons between paralogs, the proteins sequences of all homologs with nucleotide identities lower than 50% were reciprocally blasted against the five reference genomes. The homologous associations were then adjusted to include the true orthologs or removed from the dataset if orthologous sequences were missing in at least one genome. Fewer than a dozen homologous genes amongst the 2324 identified were thereby curated (excluded from further analyses). We expect that the remaining cases of paralogous associations would be limited to just a few extra genes and would not significantly affect our results.

4.3. Results

4.3.1. Metagenome assemblies of *Ridgeia* symbionts

Table 4.2 Overview of Vestimentiferan symbionts metagenomes.

Genomes	***Ridgeia* 1 symbiont[1]**	***Ridgeia* 2 symbiont[1]**	***Tevnia* symbiont[2]**	***Riftia* 1 symbiont[2]**	***Riftia* 2 symbiont[2]**	***Endoriftia persephone*[3]**
Genome size (Mbp)	3.44	3.42	3.64	3.48	3.71	3.20
No. contigs	97	693	184	197	414	2170
N50	83.9	7.6	92.7	28.4	24.6	1.9
Coverage	180 X	17 X	15 X	25 X	13 X	11 X
No. reads	7 436 749	993 690	212 833	467 070	205 880	130 000
GC%	58.9	58.9	58.2	58.2	58.2	57.9
No. genes	3188	3698	3277	3254	3566	6450
No. protein coding genes	3132	3641	3230	3209	3515	6414
No. rRNA	3	3	3	3	4	4
No. tRNA	47	43	44	42	47	32

1 This paper.

2 Gardebrecht *et al.*, 2011

3 Robidart *et al.* 2008

4.3.1.1. Assembly quality

The quality of a genome assembly depends on its completeness and coverage. These and other features of all available *Endoriftia* genome assemblies are compared in Table 4.2. Completeness can be estimated from the number and size distribution (N50) of contigs, while coverage, calculated as the average per-base sequencing depth, is a measure of the sampling effort. Higher sequencing depth results in higher sequence accuracy but also improves completeness of isolate genomes.

The *Ridgeia*1 symbiont assembly was of high quality (Table 4.2). It contained fewer and longer contigs than the *Riftia 1* symbiont assembly, and its coverage was seven to sixteen times greater than all previously published assemblies of *Endoriftia* (Robidart *et al.*, 2008; Gardebrecht *et al.*, 2011). Yet, even with such high sequencing depth, we were not able to close the genome. We suspect chromosome rearrangement within the symbiont population might be the cause of this fragmentation as it can create ambiguous links during the scaffolding step of the assembly.

The *Ridgeia*2 'pooled' symbiont genome was generally lower in quality than the Gardebrecht *et al.* (2008) assemblies but still notably superior in completeness and coverage to the Robidart *et al.* (2008) assembly (Table 2). Because of the overall lower quality and considerable differences in gene annotations, the Robidart *et al.* (2008) assembly was not used in our analyses.

4.3.1.2. Confirmation that *Ridgeia* symbionts are *Endoriftia*

The *Ridgeia* 1 and *Ridgeia* 2 16S, 23S rRNA, and ITS sequences were 100% identical to each other, and differed from the *Tevnia* and *Riftia* symbionts sequences by 1, 0, and 3 nucleotides, respectively. This is consistent with the

hypothesis that the same species of symbionts, *Endoriftia persephone*, is associated with *Riftia, Tevnia* and *Ridgeia*.

This is further supported by the fact that the majority of *Endoriftia* genes had homologs in the *Ridgeia* symbiont assemblies (Figure 4.4, see also Table B.1 in Appendix D; p. **Error! Bookmark not defined.**).

Like the symbionts associated with *Riftia* and *Tevnia*, the *Ridgeia* symbionts have a diverse metabolism and possess genes for sulfide oxidation, carbon fixation through the Calvin Benson Bassham and rTCA cycles, denitrification, motility and chemotaxis (see Table D.5 in Appendix D; p. D.20).

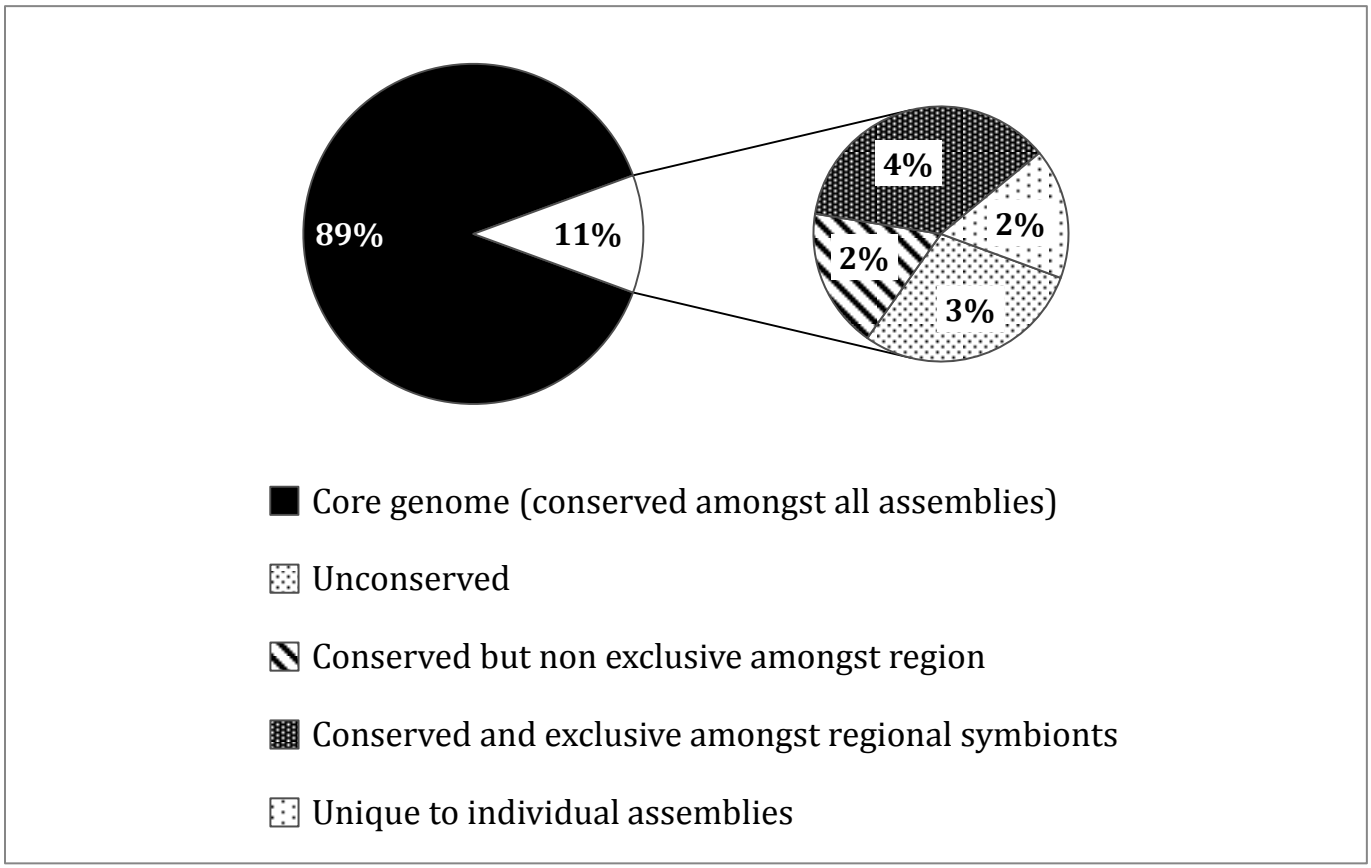

Figure 4.2 Pan-genome of Candidatus *Endoriftia persephone* based on the relative size of the Locally Collinear Blocks (LCBs) shared between five *Endoriftia* assemblies from two distinct geographical regions. *Ridgeia* 1 and *Ridgeia* 2 symbionts are from the Juan de Fuca Ridge (JdFR), *Tevnia*, *Riftia* 1 and *Riftia* 2 symbionts are from the East Pacific Rise (EPR). The five genome assemblies were aligned with Progressivemauve (Darling *et al.*, 2010).

4.3.2. *Endoriftia persephone*'s pan-genome composition

Figure 4.2 shows the composition of *Endoriftia persephone*'s pan-genome based on the nucleotide sequences of the five most recent *Endoriftia* assemblies. A core genome, representing 89% of *Endoriftia*'s pan-genome was shared across all of the assemblies of *Riftia*, *Tevnia* and *Ridgeia* symbionts. In addition, 4% of the pan-genome was region-specific, that is, only found in and shared among symbionts from the same geographical region; *i.e.* the JdFR symbionts associated with *R. piscesae*, and the EPR symbionts associated with *R. pachyptila* and *T. jerichonana* (Figure 4.2). Symbionts from the same geographical region shared up to 98% of their genomes.

Finally, we found that between 0.7 and 2.9% of the pan-genome was unique to the specific assemblies and was in part composed of contaminant sequences and/or exogenous genetic material recently acquired through horizontal transfer. This is supported by the fact that the GC content of the unique genome for some assemblies was notably different from that of the core genome. In *Tevnia* for example, the GC content of the unique genome (96 kbp) was 42% while it was 60% for the core genome.

The relatively large size of the unconserved genome (3% of the *Endoriftia* pan-genome) was likely the result of gaps in the genome assemblies and the small sample size. We expect that increasing data quality and the number of samples would reduce the relative importance of the unconserved genome in favor of the conserved pan-genome or the regional core genome.

4.3.3. Genes encoded in the accessory genome

4.3.3.1. Region-specific genome

The LCBs that were exclusive to the JdFR or EPR symbiont genomes both carried unique genes coding for transposases, integrases, and other phage associated proteins, as well as a few genes involved in cell wall/membrane/envelope biogenesis (see Table D.1 and Table D.2 in Appendix D; p. D.1).

Interestingly, two CRISPR/cas systems (Westra *et al.*, 2014) were found in all of the genome assemblies. The first was well conserved but the spacers were notably different in the two genomes for which the CRISPR locus was successfully assembled (*Ridgeia* 1 and *Tevnia*). In the second CRISPR/cas system, the cas operon was not conserved across symbionts from the JdFR and the EPR; half of the cas genes were not homologous (see Table D.1 and Table D.2 in Appendix D; p. D.1).

Finally, a 17 kbp scaffold with genes encoding for the type VI secretion system was uniquely found in the *Ridgeia* symbionts (see Table D.1 in Appendix D; p. D.2).

4.3.3.2. Vent site vs Species-specific genomes at the EPR

The LCBs unique to *Riftia* symbiont assemblies were limited to three contiguous scaffolds (see Table D.3 in Appendix D; p. D.13). The first was about 60 kbp in length and contained genes typically found in fertility factors, *i.e.* OmpA/MotB gene, tra genes, IS200 like transposases and two genes coding for nucleotide-binding proteins. The other two scaffolds were smaller (about 10kbp) and contained, respectively, six genes of the CRISPR-cas3 system, and four genes; two unannotated genes, one transcriptional regulator and a putative relaxase. Because of the incompleteness of the *Tevnia* symbiont assembly, we could not rule out the possibility that these differences resulted from a biased sampling of the *Tevnia* symbiont's metagenome. However, given the large size of the missing scaffolds, we would suggest that this was unlikely to have been the case.

Amongst the LCBs exclusively found in the assemblies of symbionts from 9°N, however, many seemed to have resulted from a poor sampling of the *Riftia* 1 symbiont's fragmented genome. They represented stretches of DNA of a few thousand bp, often located at the extremities of conserved contigs. In contrast, other unique sequences seemed to represent real chromosomic variation, tended to be larger (up to 16.3 kbp), were flanked by regions of low nucleotide conservation, and contained unique mobile elements, toxin/antitoxin and transcriptional regulators typically found in phage genomes (see Table D.4 in Appendix D; p. D.16).

4.3.4. Population structure of *E. persephone*

4.3.4.1. Cluster analyses

The first sequenced metagenome of *Endoriftia persephone* (Robidart *et al.* 2008) clustered apart from the more recent assemblies. This is probably a result of sequencing errors due to the overall lower quality of reads associated with the sequencing methods used at the time.

Ridgeia symbionts cluster apart from the EPR symbionts both in terms of nucleotide distance in the core genome, and in the composition of their accessory genome (Figure 4.3).

Within the EPR, the symbionts cluster by host species when considering the nucleotide sequences of the core genome while they cluster by vent site according to the composition of the accessory genome. Thus, at area of the EPR (near 9° N) for which there are symbiont genome assemblies for both *Riftia* and *Tevnia*, symbionts from these two host shared more exclusive LCBs than *Riftia* symbionts collected from different EPR areas (9° N and 13° N). Interestingly, the accessory genome exclusive to the 9°N vents was composed of shorter LCBs and

was overall slightly smaller than the symbiont genome exclusive to the *Riftia* host species (70 kbp and 80 kbp, respectively) (see Table D.3 and Table D.4 in Appendix D; p.D.1).

Finally, *Tevnia* symbionts seemed to be closer to those from *Ridgeia* (versus *Riftia*) in terms of nucleotide identity.

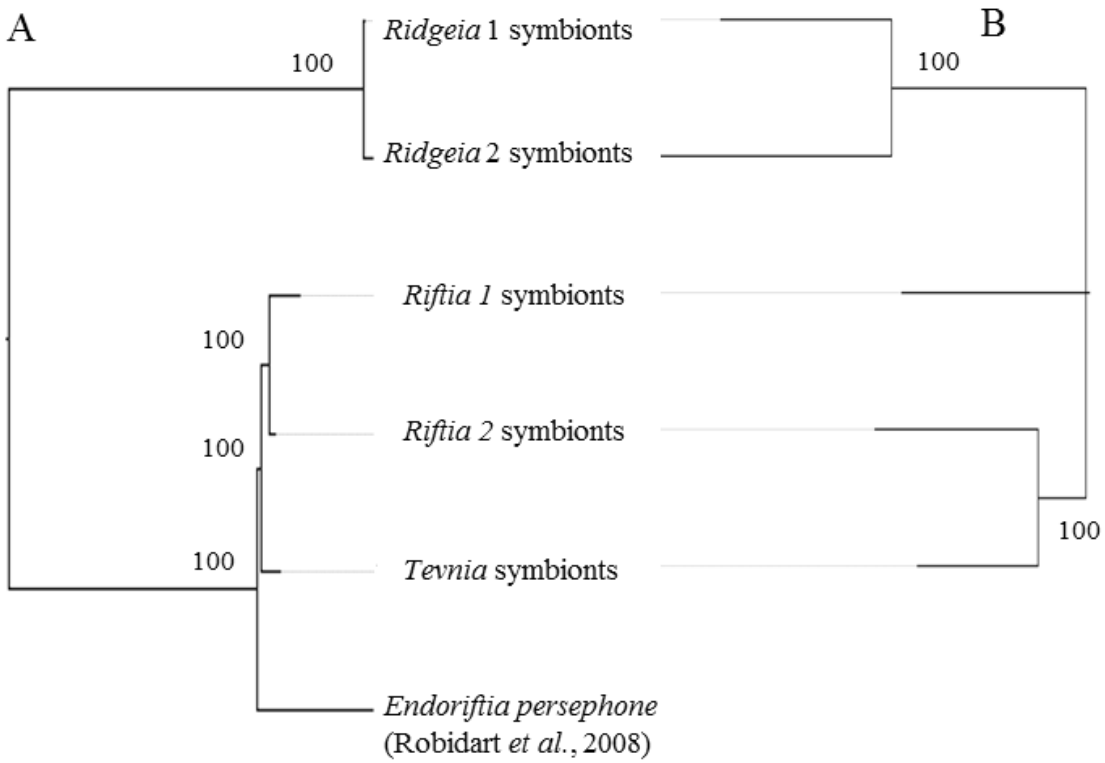

Figure 4.3 Neighbor-joining trees of Candidatus *Endoriftia persephone* based on A) the genetic distances (HKY model) between nucleotide sequences of the core genome, and B) the presence/absence of sequences of the accessory genome. A) The six assemblies were aligned with MAUVE and the Locally Collinear Blocks (LCBs) extracted. Of these, only the LCBs > 100 bp and represented in all assemblies were kept. The sequences within each LCB were aligned with MAAFT and concatenated to form a genome wide alignment of 2 580 528 bp containing 75 472 variable sites. B) The first assembly of *E. persephone* (Robidart *et al.*, 2008) was not included in this analysis, because of high genome fragmentation. Assemblies were aligned with MAUVE and the presence/absence of LCBs > 100 bp was used to generate a distance matrix (Jaccard index) from which a neighbor-joining tree was constructed using Population V1.2.32. Bootstrap values are indicated over the branches.

4.3.4.2. Comparisons of orthologous genes

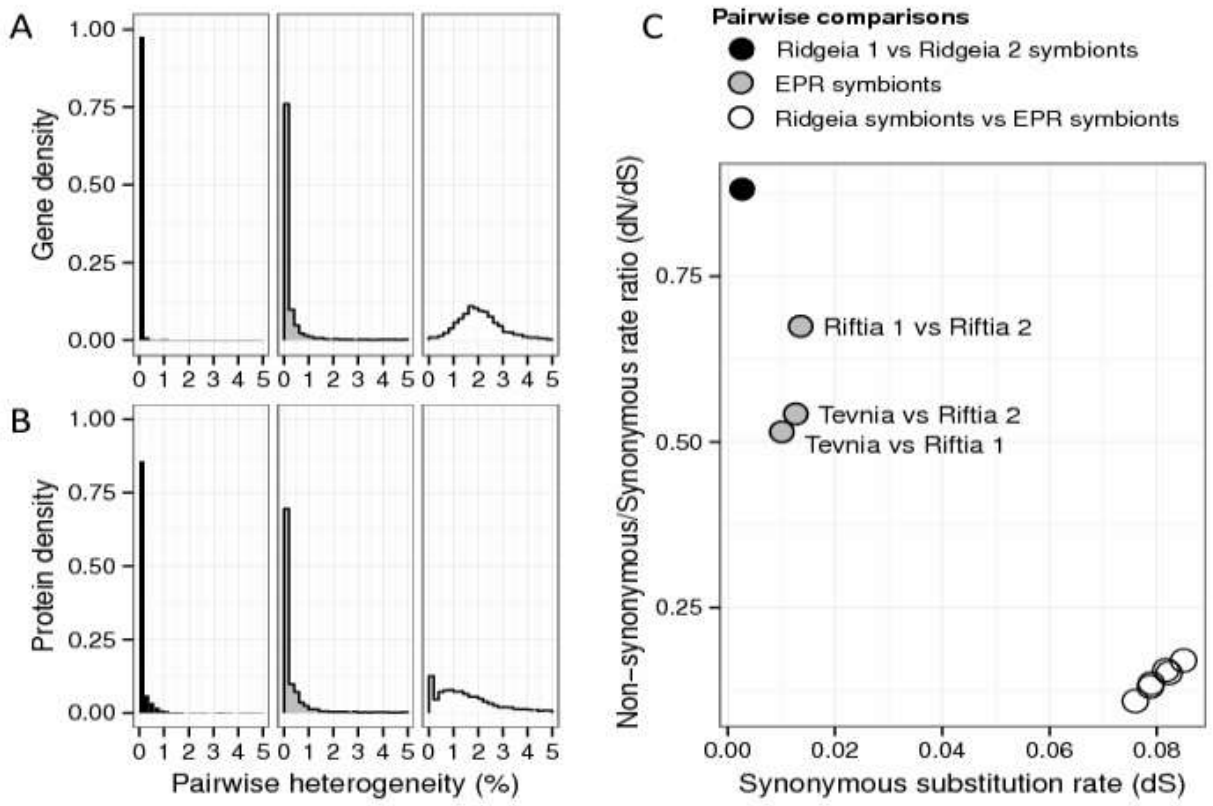

Figure 4.4 A) Distribution of heterogeneity between pairs of homologous genes based on nucleotide sequences and B) amino acid sequences. Only heterogeneities <5% are represented (>90% of data). **C) Negative correlation of the dN/dS ratio and divergence between individuals from different metapopulations based on the concatenated alignments of 2313 homologous gene sequences (1 926 255 bp).**

Nucleotide heterogeneity was inferior to 1% within the EPR tubeworm symbionts and the *Ridgeia* symbionts from the JdFR but it increased to around 2% when symbionts from the two regions were compared (Figure 4.4, A). Yet, many homologous proteins were highly conserved across the assemblies from the JdFR and the EPR indicating strong purifying selection acting on *Endoriftia* (Figure 4.4, B).

The *Ridgeia* symbionts appeared more homogeneous than those from the EPR. 89% of the genes in the two *Ridgeia* symbiont assemblies had identical nucleotide sequences compared to 54% on the EPR. These results were corroborated with the overall synonymous substitution rates (dS) and ratio of non-synonymous to synonymous substitution rates (dN/dS) between pairs of symbiont assemblies (Figure 4.4, C).

The synonymous substitution rate (dS) is the ratio of the number of synonymous substitutions over the number of synonymous sites. Because synonymous substitutions tend to be selectively neutral, they accumulate over time and thus can be used as a proxy for divergence between genomes (Ochman and Wilson, 1987; Wielgoss *et al.*, 2011). Assuming an allopatric divergence, we can then derive a molecular clock for the synonymous substitution rate (r) for the *Endoriftia* symbionts where r = dS/2T with dS the divergence observed between the vicariant populations at the synonymous sites and T the time of last contact between the East Pacific Rise and the northeast Pacific ridge systems. Following Vrijenhoek (2013), we used T= 28.5 Mya and obtained a substitution rate of 0.14% (± 0.01%) per Myrs.

This substitution rate is lower than rates observed for *E. coli* in culture (0.45%) (Ochman *et al.*, 1999) and for the host vestimentiferan tubeworms themselves (~0.2%) (Chevaldonne *et al.*, 2002). However, the latter were both based on comparisons of isolate genomes, whilst our rate was determined from comparisons of genome assemblies that resulted from the concatenation of

multiple symbionts with potentially multiple genotypes. Thus, we may have underestimated the genetic diversity within and across the symbiont populations, and therefore the rate of divergence between the two populations. Furthermore, the divergence between two populations depends on their respective reproduction rates and on the parameters that affect their respective genetic diversity over generations (*i.e.* their underlying biological mutation rate, effective size, and clonality (Shapiro *et al.*, 2009; Fraser *et al.*, 2007)). Current knowledge of doubling times and genetic diversity in these symbionts does not permit confident estimation of most of these parameters.

Our data can be used for an initial consideration of effective symbiont population sizes for the three host tubeworms considered here. The genome-wide dN/dS of *Endoriftia* symbionts falls into the upper range of what has been observed in other closely-related obligate symbionts (Kuo *et al.*, 2009). For closely related genomes, the dN/dS ratio is also intrinsically dependant on the time since divergence and the effective population size. More closely related lineages or lineages with smaller population sizes tend to have higher dN/dS ratios due to time-lag or delay in the curation of slightly deleterious mutations (Rocha *et al.*, 2006; Peterson and Masel, 2009). *Endoriftia* symbiont populations showed this pattern in that the dN/dS ratios were negatively correlated with the divergence between genome pairs. The highest divergence with the lowest dN/dS ratio was seen when comparing *Ridgeia* and EPR symbionts (dS~0.08) and the lowest divergence and the highest dN/dS ratio was between the two *Ridgeia* symbiont assemblies (Figure 4.4, C).

Interestingly, while the divergence between the EPR symbionts was quite similar (0.0101<dS <0.0136), the dN/dS ratio between the two *Riftia* assemblies was notably higher than for the other pairs. He *et al.* (2010) and (Luo *et al.*, 2014) made similar observations for the pathogen *Clostridium difficile* and the marine alphaproteobacteria *Roseobacter*, respectively. This suggests that the

symbionts in association with *Riftia* have a smaller effective population size than the overall EPR *Endoriftia* population, and thus, that the latter might be further structured either spatiotemporally, according to environmental conditions, or through host specificity.

4.4. Discussion

4.4.1. Divergence of Juan de Fuca Ridge (JdFR) and East Pacific Rise (EPR) symbionts

We used five high quality genome assemblies of *Endoriftia persephone* to analyse the structure of the *Endoriftia* population though pairwise comparisons of (1) the composition of the pan genome, (2) the nucleotide identity within the core genome, and (3) the synonymous and non-synonymous substitution rates for a large subsample of the core genome genes. Our results were consistent with those obtained from phylogenetic analyses based on 16S rRNA gene (Feldman *et al.*, 1997) and ITS sequences, and rep-PCR fingerprints (Di Meo *et al.*, 2000) and indicated that the population of *Endoriftia* symbionts in association with *Ridgeia piscesae* on the Juan de Fuca Ridge was distinct from the *Endoriftia* population on the East Pacific Rise that is associated with *Riftia pachyptila* and *Tevnia jerichonana*.

4.4.1.1. Allopatry

Comparisons of the composition of vent-associated macrofauna communities (Tunnicliffe, 1988) and the genetic structure of vestimentiferan worms (Chevaldonne *et al.*, 2002) provide evidence that the northeast Pacific and the EPR vent communities have been isolated by the development of discontinuities along the Pacific mid-ocean ridge caused by the tectonic fracturing of the Farallon plate about 30 million years ago (Tunnicliffe, 1988; Atwater and Stock, 1998). Similar dichotomies attributed to later plate fragmentation events were observed in populations of various invertebrate species spanning across multiple ridge systems in the northeast Pacific (Johnson *et al.*, 2006; Plouviez *et al.*, 2009; Hurtado *et al.*, 2004). It is therefore reasonable to assume that the *Endoriftia* populations were similarly affected by the emergence of these

geographical barriers. Our results indicate that the divergence of the JdFR and EPR symbionts was dominated by passive processes/genetic drift. On the one hand, the core genome was characterized by overall low dN/dS ratios and a conserved codon bias (data not shown) which suggests the same selective constraints acted on both populations. Additionally, when we compared the functional distribution of core genome genes with median values of dN/dS to that for genes with extreme values of dN/dS (5% highest dN/dS), no COG or KO categories appeared to be over-represented in the outliers (Chi-squared test of independence p-value>0.05). On the other hand, the accessory genome of each population of symbionts was composed of many mobile elements and selfish sequences as well as unique CRISPR spacers, all of which suggest two distinct histories of interactions that have independently modified the EPR and JdFR symbiont genomes.

4.4.1.2. Adaptations to viral predation

The presence of phage DNA as well as two to three (for *Ridgeia* 1) CRISPR operons can be seen as evidence that viruses are an important 'enemy' of free-living and/or intracellular *Endoriftia*, and that the symbionts genomes carry these markers of phage infections.

Although little considered until recently, there is accumulating evidence for a viable and presumably metabolically active free-living stage of *Endoriftia* (Harmer *et al.*, 2008; Klose *et al.*, 2015). Viruses are known to be abundant at deep-sea hydrothermal vents and a likely important source of mortality for free-living bacteria (Ortmann and Suttle, 2005). Alternatively, the trophosome might also be a favorable environment for the proliferation of phages among the dense and fast growing intracellular symbiont population.

The presence of different CRISPR spacers between the JdFR and the EPR symbiont populations could suggest the existence of different *Endoriftia*-specific viruses on these two mid-ocean ridges, although we have also found CRISPR-spacer variability within the symbiont population of a single worm (Chapter 3, Section 3.3.1).

4.4.1.3. Host adaptation

Some genes possibly involved in the symbiosis had relatively high dN/dS ratios (*e.g.* CheY chemotaxis protein, cell division protein DamX, outer membrane protein) but the divergence between the two populations was too small to detect the signature of positive selection (Kryazhimskiy and Plotkin, 2008). Nevertheless, the large scaffolds containing genes associated with the type VI secretion system, found only in *Ridgeia* symbionts, could be part of a mechanism of host adaptation. Indeed, the type VI secretion system can act as a virulence factor against eukaryotic cells or competing bacteria (Jani and Cotter, 2010; Coulthurst, 2013). It has also been found to be key in determining host specificity in *Rhizobium leguminosarum* (Bladergroen *et al.*, 2003). Other genes involved in cell wall/membrane biogenesis could be involved in the expression of microbial associated molecular patterns (MAMPs), hypothesized to be critical in mediating host-symbiont interactions (Nyholm *et al.*, 2012). Genomic and proteomic comparisons with a sympatric population of *Endoriftia* symbionts associated with a different host species (*e.g. Lamellibrachia* sp. (Di Meo *et al.*, 2000)) might tell us more about host specificity.

4.4.2. EPR symbionts are further structure into populations that might be relatively isolated spatially or temporally

Our results show little evidence for geographic differentiation of symbionts from the two sites on the EPR for which genome sequence data are available. Symbionts from 9°N were no more similar to each other than symbionts from 9°N and 13°N. In contrast, when symbionts from the two EPR host tubeworms were compared, the nucleotide sequences of symbionts hosted by the same host species were more homogeneous and had a higher dN/dS ratio, suggesting that *Riftia* symbionts formed a subpopulation within the EPR. Additionally, *Riftia* symbionts carried scaffolds with genes typically found in F-type conjugative plasmids. These genes have been speculated to play a role in the horizontal gene transfer (Gardebrecht *et al.*, 2011) and might allow for a higher degree of genetic exchange between *Riftia* symbionts thus keeping this population homogeneous.

While free-living symbionts probably can disperse on large scales and colonize new surfaces/vents independently of their hosts, small scale spatial or temporal variations in the environmental conditions could favor particular strains of symbionts resulting in population partitioning. This local increase in homogeneity might be exacerbated or maintained in presence of the tubeworm hosts through pseudo-vertical transfer of symbionts (Klose *et al.*, 2015).

Molecular mechanisms controlling host specificity might also exist but higher resolution of genetic diversity would be needed to clearly characterize variations in the symbionts accessory genome.

4.4.3. Conclusion: towards a better characterization of *Endoriftia* populations

This first characterization of *Endoriftia* symbionts at the population level found that *Endoriftia* symbionts are structured on large geographical scales but also at smaller scales and possibly through host specificity.

While the number and the quality of our samples were limited, we are confident that further population genetic studies, using rapidly advancing sequencing platforms, will provide further insight into the symbiont evolutionary history and adaptation to their hosts and environment.

We suggest that future studies focus on assessing the number and diversity of *Endoriftia* genotypes. To these ends, we propose that CRISPR spacers and extra-chromosomic genetic material may have the potential to be used for high resolution differentiation of populations of symbionts. For example, 'CRISPR typing' has been used for genotyping human bacterial pathogens (Fabre *et al.*, 2012; Cui *et al.*, 2008; Lopez-Sanchez *et al.*, 2012; Yin *et al.*, 2013; Kovanen *et al.*, 2014) and aquatic bacteria (Held *et al.*, 2010; Kuno *et al.*, 2012, 2014). In the meantime, sequencing the complete genome of individual *Endoriftia* cells would allow us to detect chromosomic variations.

Understanding the structure, dynamism and interconnectivity of *Endoriftia* populations is important to advancing our knowledge of the ecology and evolution of their host worms that are often keystone species in vent communities.

Acknowledgements

We thank the crews of the R/V Atlantis and R/V Thomas G. Thompson as well as the pilots of the submersibles Alvin and ROPOS, and Steven Hallam for their assistance in obtaining our metagenomic samples. This research was enabled in

part from computing assistance provided by Belaid Moa, WestGrid (westgrid.ca) and Compute Canada/Calcul Canada (computecanada.ca). We also thank Nathalie Forget, Steve Perlman, Sebastien Duperron, the members of Verena Tunnicliffe's lab, Lee Katz, Diana Varela, Real Roy, Francis Nano, and all of the contributors to seqanswers.com and biostars.org.

This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Canadian Healthy Oceans Network (NSERC Canada) to S.K.J.

Chapter 5. Conclusions and perspectives

5.1. Retrospective on the main problematic

In this thesis, I investigated the genetic diversity of the *Ridgeia piscesae* bacterial symbionts within and across individual hosts to begin examining a partner choice hypothesis, based on the wide variety of habitats occupied by the host species, and on partner choices known from symbioses in hydrothermal vent mussels, and evidence for multiple partners in a tubeworm from hydrothermal vents in the Mediterranean. The partner choice hypothesis maintains that the horizontal transmission of the symbionts allows the hosts to associate with bacterial partners that are best adapted to the local environmental conditions. These investigations took the form of two molecular studies of *Ridgeia* piscesae trophosome samples focussing on (1) the phylogenetic and (2) the genotypic diversity of *R. piscesae* intracellular bacteria, and (3) a study of the population structure of *Endoriftia* symbionts across two mid-ocean ridges and three different host species.

The results of these studies did not permit confirmation or rejection of partner choice participating as a stabilizing factor in the mutualistic relationship between *Endoriftia* and *Ridgeia piscesae*. I did not find multiple phylotypes and while there were multiple lines of evidence for genetic diversity in *Endoriftia,* I was unable to identify distinct lineages. I found that the genetic diversity of intracellular symbiont populations was possibly correlated with worms' settling environment but much more samples will be needed to confirm this hypothesis. Comparing Endoriftia from different worm hosts between different ridge systems revealed notable diversity in the accessory genome (but not in the core genome), but how or if the observed genetic diversity translates into metabolic diversity remains to be determine. Nonetheless, together, these studies have

revealed that *Endoriftia* is much more genetically diverse than previously thought, and provide a basis for follow-up research.

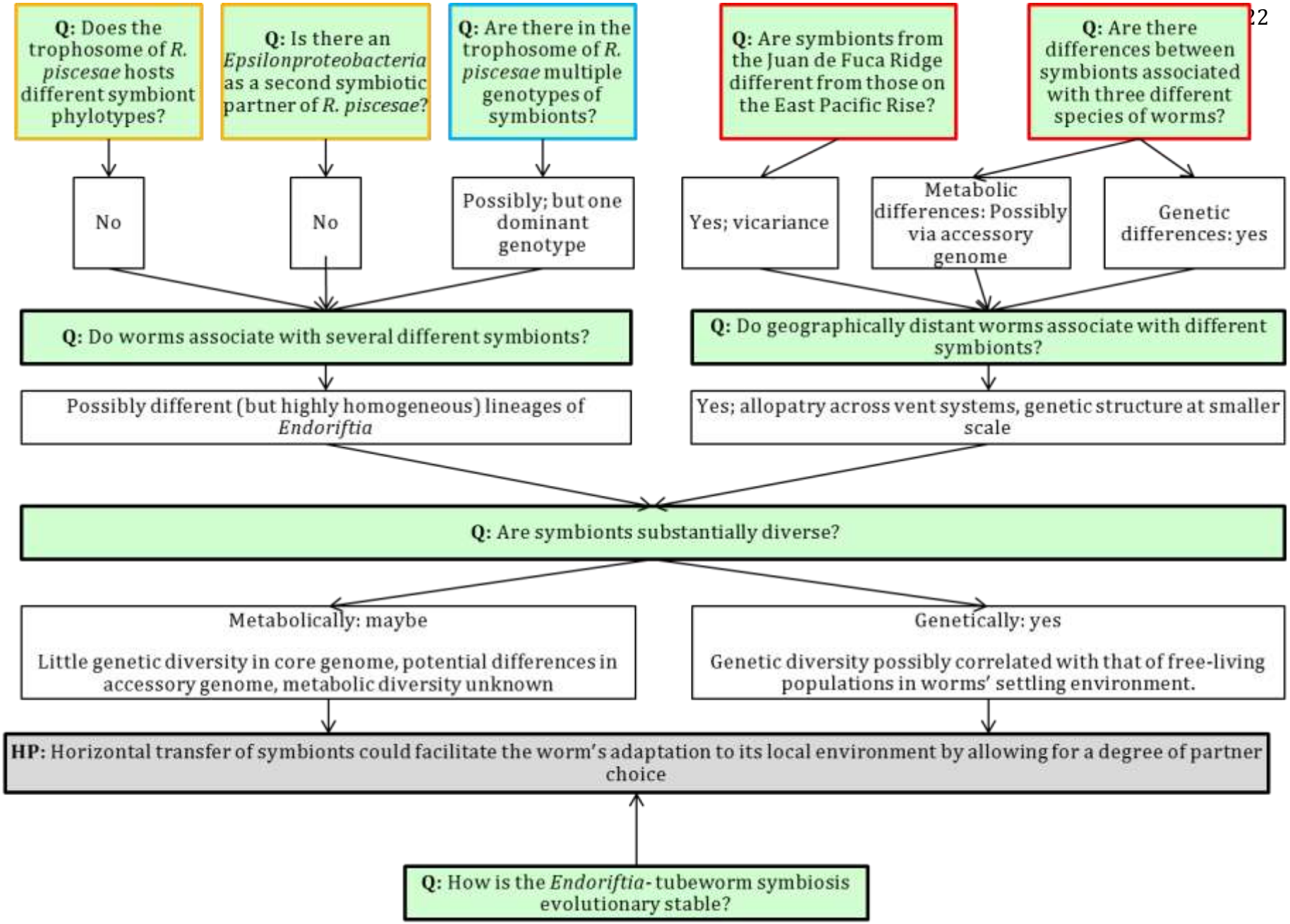

Figure 5.1 Retrospective on the general problematic. Questions pertaining to Chapters 2, 3, and 4 are outlined in orange, blue and red, respectively.

5.2. Summary and highlights of the three studies conducted

5.2.1. Study 1

In the first study (Chapter 2), Nathalie Forget and I used two methods that had not previously been applied to *R. piscesae* (CARD-FISH and 454 pyrosequence libraries) to analyze the genetic diversity of the trophosome of *R. piscesae* individuals. Pyrosequence libraries from trunk DNA extracts from 2010 revealed low diversity amongst individuals. The second most abundant phylotype putatively associated with the trophosome belonged to the Class *Epsilonproteobacteria* and represented only ~10 % of the sequences. *In situ* hybridisations with a probe specific to the *Epsilonproteobacteria* yielded no signal or only non-specific signals. On the other hand, hybridization with a probe targeting *Gammaproteobacteria* showed very high densities of intracellular bacterial cells. The overwhelming dominance of *Gammaproteobacteria* in *R. piscesae* trophosomes was corroborated by all of the pyrosequence data. Sequences from samples collected in 2013, which were treated to remove any potential epibiotic contamination, showed *Gammaproteobacteria* to be the sole bacteria within *R. piscesae*'s trophosome and that the phylogenetic affiliation of the symbionts could not be resolved further than the Class level using mothur's bioinformatic pipeline.

5.2.2. Study 2

The following study (presented in Chapter 3) attempted to further characterize a population of *Endoriftia* at the genotypic/strain level. This study asked the fundamental question of whether different lineages of symbionts coexist within the trophosome of *R. piscesae* or if the symbionts all belonged to a single clonal lineage. As proxies for genotypic diversity, I used (1) the heterogeneity of spacers in CRISPR arrays and (2) genetic variants in whole genome shotgun

sequences of two symbiont metagenomes (one from one individual and one from the pooled data of 5 other individuals), and (3) the single nucleotide polymorphisms found in the variable regions V1-V3 and V6-V8 of the 16S rRNA genes from the trophosome of an additional 31 individual worms. We found three lines of evidences supporting the existence of several strains on symbionts inhabiting the trophosome of *R. piscesae*; (1) heterogeneity within the CRISPR array, (2) multimodal distribution of genetic variants and variants holding signs of purifying selection, and (3) genetic polymorphism shared between independent intra-cellular populations of symbionts.

Considering the symbionts inhabiting the trophosome as a mixed population rather than a homogeneous colony has implications for interpreting analyses of data from symbiont genome assemblies. In fact, it implies that symbiont metagenomes presented in Chapter 4 do not represent individual bacterial genomes per se but are instead a consensus of more or less heterogeneous populations.

5.2.3. Study 3

Chapter 4 examined the role of geographic separation and host selection in the development of genetically distinct symbiont populations. For this purpose, I assembled the consensus genomes for two *Ridgeia* piscesae symbiont populations (from one and five individual hosts, respectively) found at hydrothermal vents of the Juan de Fuca Ridge. These two assemblies bring the total of near-complete genomes of *Endoriftia* to six. I then combined the two genome assemblies with those of symbionts associated with two siboglinid species from the East Pacific Rise *(Riftia pachyptila* and *Tevnia jerichonana* (Gardebrecht *et al.*, 2011)) to develop a first portrait of *Endoriftia*'s pan genome, and an initial assessment of symbiont population structure in the different host species. This study was the first to apply genome-wide comparisons of

Endoriftia assemblies in the context of population genetics and molecular evolution. These comparisons underline the importance of viruses and genetic drift in shaping the genetic makeup of the symbionts. Its findings suggest that as for vent animal species, mid-ocean ridge discontinuities in the eastern Pacific Ocean have resulted in allopatric divergence of symbiont populations on the Juan de Fuca Ridge and the East Pacific Rise (Figure 5.2). Furthermore, within a single ridge system, the symbiont populations are not panmictic but structured possibly according to environmental conditions or host specificity. Finally, the genome-wide comparisons revealed that the population-specific functional genes were likely encoded in the accessory genome and potentially plasmids.

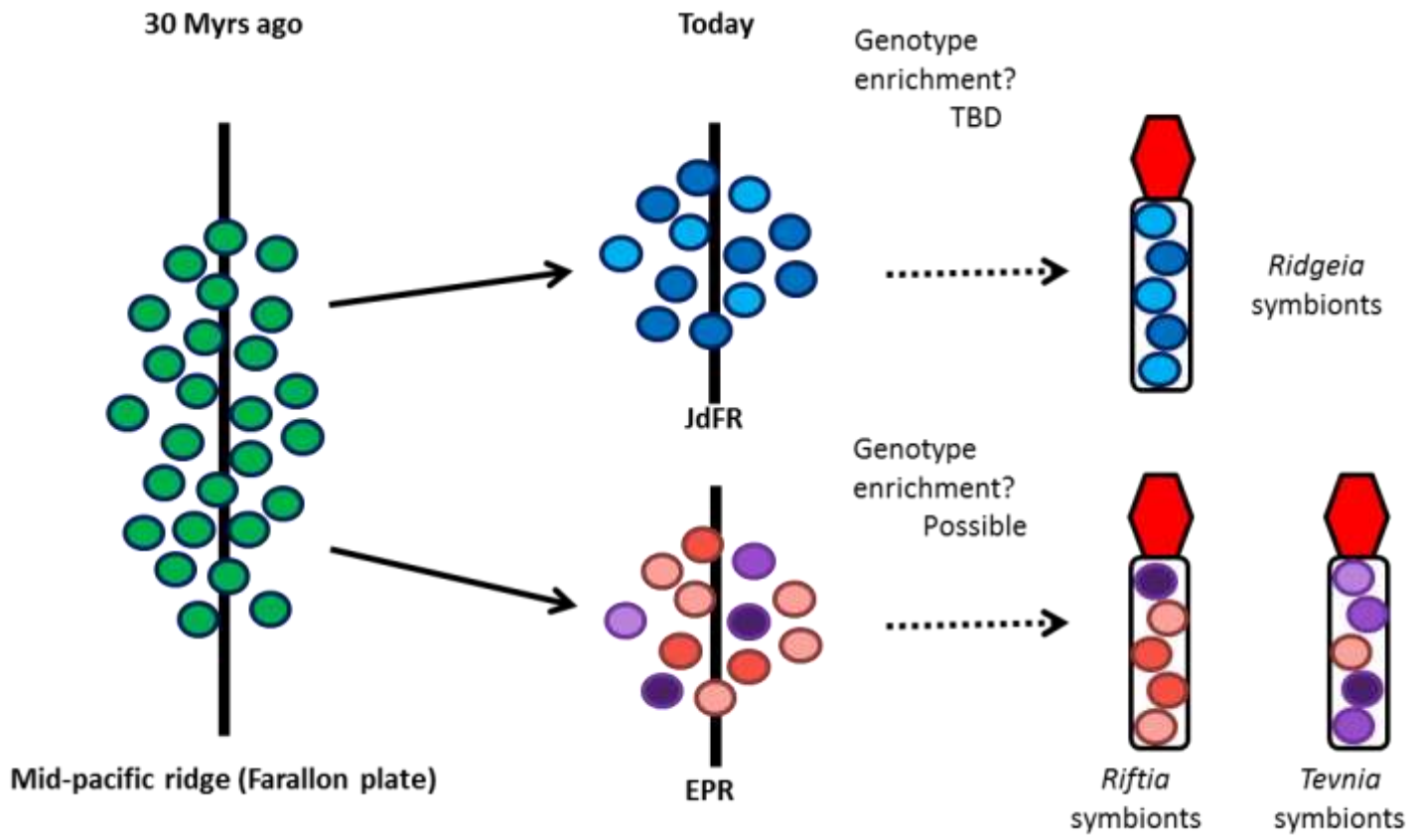

Figure 5.2 Schematic representation of Candidatus *Endoriftia persephone* vicariance leading to the population structure observed today. JFP: Juan de Fuca plate, EPR: East Pacific Rise. The EPR symbionts population might be structured temporally, spatially at a finer scale or species-specifically.

5.3. Remaining questions and leads to answer them

The work presented in this thesis was very much exploratory. It used bioinformatic pipelines not previously been applied to studying vestimentiferan symbiont diversity and these pipelines called for many recently developed algorithms and programs that have input-parameter-dependant outputs (e.g. mothur (2009), VarScan (2009), SPAdes (2012)) or have not been subject to extensive testing with real datasets (e.g. Crass (2013), Pollux (2015), SysCall (2011)). Furthermore, the degree of certainty for the findings described above (Section 5.2) was limited by small sample numbers in each study. Nevertheless, some of these findings offered new leads to answer fundamental questions about the specificity of the *Siboglinidae-Endoriftia* symbiosis and most revealed the need to look at the symbiosis within a population genetics frame by raising new questions about the diversity and distribution of these unique *Gammaproteobacteria* populations.

5.3.1. What drives the *Siboglinidae-Endoriftia* specificity?

The observed specificity between host and symbiont might seem surprising considering the environmental acquisition of the symbionts. Indeed, environmentally acquired symbioses are often described as opportunistic and facultative associations or cases of parasitism in which the parasite lost its virulence (Ewald, 1987). However, this view is being increasingly challenged as, researchers find examples of highly specific mutualistic symbiosis existing without vertical transmission (Sachs *et al.*, 2011; Wilkinson and Sherratt, 2001). Some of the most notable examples are the association between the bioluminescent *Gammaproteobacteria Vibrio fisheri* and the hawaian bobtail squid *Euprymna* scolopes (Nyholm and McFall-Ngai, 2004), the diazotrophic *Alphaproteobacteria Sinorhizobium meliloti* and the Barrel Medic *Medicago trunculata* (Jones *et al.*, 2007), and of course the dinoflagellate *Symbiodinium* sp. and their specific coral hosts (Baker, 2003; LaJeunesse *et al.*, 2004; Stat *et al.*,

2015). In the plant-rhizobium case, specificity is enabled by a continuous molecular dialogue between the two partners (Jones *et al.*, 2007). For the squid's bioluminescent bacteria however, specificity results from a stepwise screening process orchestrated by the host, that results in series of physical and chemical barriers that only the right partner can overcome (Nyholm and McFall-Ngai, 2004). Finally, for dinoflagellates-coral symbioses, associations from both mechanisms were observed; molecular recognition pre-phagocytosis (Rodriguez-Lanetty *et al.*, 2006) and host controlled winnowing (Dunn and Weis, 2009; Fay and Weber, 2012).

So what can we expect as a symbionts acquisition process of the eastern Pacific tubeworms? Using detailed histological analysis Nussbaumer *et al.* (2006) conducted an extensive study of larval infection by *Endoriftia* in *Riftia pachyptila*. Similarly to what was found in *Euprymna scolopes*, the authors observed that before the formation of a chitinous tube, the juvenile larvae possess a mucous coat in which different bacteria along with the *Endoriftia persephone* symbionts were found. However, as for the leguminous plants, selection for the symbionts likely happened before the initial infection. In fact, the authors demonstrated that the worm is colonized through the skin but only the symbiotic bacteria were detected infecting the worm's epithelial cells. This suggested that *Riftia*'s infection process was complex and most likely involved a finely tuned molecular dialog. This was supported by the finding of genes implicated in chemotaxis in the first sequenced genome of *Endoriftia* (Robidart *et al.*, 2008). Additionally, Nyholm *et al.* (2012) found in *Ridgeia piscesae* several pattern recognition receptors (*i.e.* special receptor proteins synthesized by the host that recognize molecules with distinct motifs produced by bacteria). The authors suggested these receptors not only play a key role during the initiation of the symbiosis but are also involved in the maintenance of the symbiont population throughout the life of the animal. Finally, genome-wide comparisons of symbionts in association with *Riftia*, *Tevnia* and *Ridgeia* species revealed that

symbionts might have different mechanisms of host adaptation (Chapter 4). Symbionts in association with *Ridgeia* had genes of the type VI secretion system that in *Rhizobium leguminosarum* participate in mediating its host pea plant specificity (Bladergroen *et al.*, 2003). Symbionts in association with *Riftia* may possess an F-like conjugative plasmid containing genes of the type IV secretion system (Gardebrecht *et al.*, 2011). This secretion system was found in other rhizobia to be implicated in in bacteria-host interaction (Hubber *et al.*, 2004).The F-like conjugative plasmid could also serve as to keep *Riftia* symbiont population genetically more homogeneous by allowing for recombination.

Building on my results, future studies interested in characterizing the molecular factors driving species specificity might find interest in answering the following questions: Are there plasmids or other accessory chromosomes? What is the composition and role of the F-like plasmid in symbionts associated with *Riftia*? What is the role of the type VI secretion system in symbionts associated with *Ridgeia*? Are the type VI secretion system genes expressed during the infection process? Where are these genes located?

5.3.2. How many strains are there and how different are they? Do they have functional differences?

Almost nothing is known of the genotypic diversity of *Endoriftia.* I was able to identify individual fragments of evidence for genotypic diversity within the trophosome of *Ridgeia piscesae* tubeworms, but the metagenomic and pyrosequences I had were too short to successfully link genotypic variants or spacers together into discrete genotypes and I was therefore unable to characterize specific lineages of symbionts or quantify disparity between them. Furthermore, it is not clear whether the genetic diversity of *Endoriftia* would translate into phenotypic diversity. In the *Symbiodinium microadriaticum* – Invertebrate model, Chang *et al.* (1983) found that different strains associated with a particular species of clam, sea anemone and scleractinian coral,

respectively, had different photoadaptative phenotypes. However, *Endoriftia* symbionts associated with *R. pachyptila* and *T. jerichonana* seemed to display little proteomic heterogeneity (Gardebrecht *et al.*, 2011). The genetic diversity I observed within the trophosome of *R. piscesae* was low (Chapter 3) but was only based on a few samples. I was not able to find positively selected genes amongst the *Endoriftia* assemblies but again I had very few samples to compare.

Further investigations of the symbiont metabolic diversity should replicate Gardebrecht *et al.*'s study with more samples and compare in parallel the genetics and proteomics of contrasting symbiont populations (*e.g.* symbionts from different hosts, different regions, and different venting conditions). The characterization of *Endoriftia* pan-genome also suggested that phenotypic variation could be encoded in extrachromosomal genome (*i.e.* plasmids).

5.3.3. How dynamic and how connected are the symbiont populations?

Even if the different symbiont populations do not have an ecologically significant metabolic diversity, resolving the symbiont diversity at the level of individual genotypes would still greatly improve our knowledge of the evolution and ecology of this *Gammaproteobacteria* species. For instance, single cell sequencing and quantification of linkage disequilibrium could allow us to quantify the rate of homologous recombination between individual cells (*i.e.* clonality) and thus help us define *Endoriftia* population structure on a temporal scale (Shapiro *et al.*, 2009; Smith *et al.*, 1993). For example, a mixture of recombination and clonal propagation results in maintaining low genetic diversity in some *Rhizobium* and *Symbiodinium* sp. populations (Provorov and Tikhonovich, 2015; Santos *et al.*, 2003). Alternatively, in vesicomyid clams, the increased promiscuity between different chemosynthetic bacterial lineages as they colonize the same individual host could favour recombination and result in

increasing the genotypic diversity of otherwise clonal obligate symbiont populations (Stewart *et al.*, 2009).

Assuming that intracellular symbiont populations are representative of free-living populations, we could characterize *Endoriftia* population structure on a biogeographic scale by sampling cells in association with various species of vestimentiferan hosts living in contrasting environmental conditions. Sampling of free-living symbionts remains to be achieved and would be important to test the validity of this assumption.

I propose that genotyping using CRISPRs which has already been applied for some pathogenic and free-living bacteria can be used for the siboglinid symbionts (Held *et al.*, 2010; Kuno *et al.*, 2014; Pourcel *et al.*, 2005). In *Endoriftia* genomes, two to three CRISPR arrays were found but only one was successfully reconstructed. This array is roughly 700 bp long and might easily be amplified from mixed DNA extracts by using its specific leader sequence and its conserved neighboring gene (coding for a phosphoribosylglycinamide formyltransferase) as promoters. These libraries could then be sequenced using high throughput sequencing technologies such as the latest 454/Roche platform that produces reads up to 1000 bp in length (GS FLX+ system; http://454.com/products/gs-flx-system/index.asp).

Endoriftia is the obligate symbiont of tubeworms that are keystone species at eastern Pacific vents. As our understanding of symbiont diversity and habitat related specificity increases, there may come a point where this knowledge will need to be considered in the development and updating of management plans for hydrothermal vents Marine Protected Areas (MPAs), such as the Endeavour MPA.

Bibliography

Alain K, Olagnon M, Desbruyères D, Pagé A, Barbier G, Juniper SK, *et al.* (2002). Phylogenetic characterization of the bacterial assemblage associated with mucous secretions of the hydrothermal vent polychaete Paralvinella palmiformis. *FEMS Microbiol Ecol* **42**: 463–476.

Arndt C, Gaill F, Felbeck H. (2001). Anaerobic sulfur metabolism in thiotrophic symbioses. *J Exp Biol* **204**: 741–750.

Atwater T, Stock J. (1998). Pacific-North America Plate Tectonics of the Neogene Southwestern United States: An Update. *Int Geol Rev* **40**: 375–402.

Bachmann NL, Petty NK, Ben Zakour NL, Szubert JM, Savill J, Beatson SA. (2014). Genome analysis and CRISPR typing of Salmonella enterica serovar Virchow. *BMC Genomics* **15**: 389.

Baker AC. (2003). Flexibility and Specificity in Coral-Algal Symbiosis: Diversity, Ecology, and Biogeography of Symbiodinium. *Annu Rev Ecol Evol Syst* **34**: 661–689.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**: 455–477.

Barnett D, Garrison E, Quinlan A, Strömberg M, Marth G. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* btr174.

Barrick JE, Lenski RE. (2009). Genome-wide Mutational Diversity in an Evolving Population of Escherichia coli. *Cold Spring Harb Symp Quant Biol* sqb.2009.74.018.

Black M, Halanych K, Maas P, Hoeh W, Hashimoto J, DesbruyeÁres D, *et al.* (1997). Molecular systematics of vestimentiferan tubeworms from hydrothermal vents and cold-water seeps. *Mar Biol* **130**: 141–149.

Bladergroen MR, Badelt K, Spaink HP. (2003). Infection-Blocking Genes of a Symbiotic Rhizobium leguminosarum Strain That Are Involved in Temperature-Dependent Protein Secretion. *Mol Plant Microbe Interact* **16**: 53–64.

Blazejak A, Erséus C, Amann R, Dubilier N. (2005). Coexistence of bacterial sulfide oxidizers, sulfate reducers, and spirochetes in a gutless worm (Oligochaeta) from the Peru margin. *Appl Environ Microbiol* **71**: 1553–1561.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**: 578–579.

Brand GL, Horak RV, Bris NL, Goffredi SK, Carney SL, Govenar B, *et al.* (2007). Hypotaurine and thiotaurine as indicators of sulfide exposure in bivalves and vestimentiferans from hydrothermal vents and cold seeps. *Mar Ecol* **28**: 208–218.

Bright M, Keckeis H, Fisher C. (2000). An autoradiographic examination of carbon fixation, transfer and utilization in the Riftia pachyptila symbiosis. *Mar Biol* **136**: 621–632.

Bright M, Lallier FH. (2010). The biology of vestimentiferan tubeworms. *Oceanogr Mar Biol Annu Rev* **48**: 213–266.

Bunce CM. (2013). Comparative Gene Analysis and Prediction of Innate Immunity and Apoptosis Machinery in Hydrothermal Vent Tubeworms. Master, University of Connecticut.

Campbell BJ, Polson SW, Zeigler Allen L, Williamson SJ, Lee CK, Wommack KE, *et al.* (2013). Diffuse flow environments within basalt- and sediment-based hydrothermal vent ecosystems harbor specialized microbial communities. *Front Microbiol* **4**. e-pub ahead of print, doi: 10.3389/fmicb.2013.00182.

Carney SL, Flores JF, Orobona KM, Butterfield DA, Fisher CR, Schaeffer SW. (2007). Environmental differences in hemoglobin gene expression in the hydrothermal vent tubeworm, Ridgeia piscesae. *Comp Biochem Physiol B Biochem Mol Biol* **146**: 326–337.

Carney SL, Peoples JR, Fisher CR, Schaeffer SW. (2002). AFLP analyses of genomic DNA reveal no differentiation between two phenotypes of the vestimentiferan tubeworm, Ridgeia piscesae. *Cah Biol Mar* **43**: 363–366.

Cavanaugh CM. (1994). Microbial Symbiosis: Patterns of Diversity in the Marine Environment. *Am Zool* **34**: 79–89.

Cavanaugh CM, Gardiner SL, Jones ML, Jannasch HW, Waterbury JB. (1981). Prokaryotic Cells in the Hydrothermal Vent Tube Worm Riftia pachyptila Jones: Possible Chemoautotrophic Symbionts. *Science* **213**: 340–342.

Cavanaugh CM, McKiness ZP, Newton ILG, Stewart FJ. (2006). Marine Chemosynthetic Symbioses. In: Dr MDP, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E (eds). *The Prokaryotes*. Springer New York, pp 475–507.

Chang SS, Prézelin BB, Trench RK. (1983). Mechanisms of photoadaptation in three strains of the symbiotic dinoflagellate Symbiodinium microadriaticum. *Mar Biol* **76**: 219–229.

Chao LS, Davis RE, Moyer CL. (2007). Characterization of bacterial community structure in vestimentiferan tubeworm Ridgeia piscesae trophosomes. *Mar Ecol* **28**: 72–85.

Cheng AY, Teo Y-Y, Ong RT-H. (2014). Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics* **30**: 1707–1713.

Chevaldonne P, Jollivet D, Desbruyeres D, Lutz R, Vrijenhoek R. (2002). Sister-species of eastern Pacific hydrothermal vent worms (Ampharetidae, Alvinellidae, Vestimentifera) provide new mitochondrial COI clock calibration. *CBM - Cah Biol Mar* **43**: 367–370.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, *et al.* (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**: 80–92.

Compeau P, Pevzner P. (2014). Bioinformatics algorithms: an active learning approach. Active Learning Publishers.

Compeau PEC, Pevzner PA, Tesler G. (2011). How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **29**: 987–991.

Corliss J, Dymond, J, Gordon, LI, Herzen, RPV, Ballard, RD, Green, K, *et al.* (1979). Submarine thermal springs on the Galapagos Rift. *Science* **203**: 107321083.

Coulthurst SJ. (2013). The Type VI secretion system – a widespread and versatile cell targeting system. *Res Microbiol* **164**: 640–654.

Crosby L, Criddle C. (2003). Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. *BioTechniques* **34**: 790–4, 796, 798 passim.

Cui Y, Li Y, Gorgé O, Platonov ME, Yan Y, Guo Z, *et al.* (2008). Insight into Microevolution of Yersinia pestis by Clustered Regularly Interspaced Short Palindromic Repeats. *PLoS ONE* **3**: e2652.

Darling AE, Mau B, Perna NT. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement Stajich JE (ed). *PLoS ONE* **5**: e11147.

De León KB, Ramsay BD, Fields MW. (2012). Quality-score refinement of SSU rRNA gene pyrosequencing differs across gene region for environmental samples. *Microb Ecol* **64**: 499–508.

deBurgh ME, Juniper SK, Singla CL. (1989). Bacterial symbiosis in Northeast Pacific Vestimentifera: a TEM study. *Mar Biol* **101**: 97–105.

Di Meo CA, Wilbur AE, Holben WE, Feldman RA, Vrijenhoek RC, Cary SC. (2000). Genetic variation among endosymbionts of widely distributed vestimentiferan tubeworms. *Appl Environ Microbiol* **66**: 651–658.

Distel DL, Lee HK, Cavanaugh CM. (1995). Intracellular coexistence of methano- and thioautotrophic bacteria in a hydrothermal vent mussel. *Proc Natl Acad Sci* **92**: 9598–9602.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**: e105–e105.

Dubilier N, Giere O, Distel DL, Cavanaugh CM. (1995). Characterization of chemoautotrophic bacterial symbionts in a gutless marine worm Oligochaeta, Annelida) by phylogenetic 16S rRNA sequence analysis and in situ hybridization. *Appl Environ Microbiol* **61**: 2346–2350.

Dunn SR, Weis VM. (2009). Apoptosis as a post-phagocytic winnowing mechanism in a coral–dinoflagellate mutualism. *Environ Microbiol* **11**: 268–276.

Duperron S, Bergin C, Zielinski F, Blazejak A, Pernthaler A, McKiness ZP, *et al.* (2006). A dual symbiosis shared by two mussel species, Bathymodiolus azoricus and Bathymodiolus puteoserpentis (Bivalvia: Mytilidae), from hydrothermal vents along the northern Mid-Atlantic Ridge. *Environ Microbiol* **8**: 1441–1447.

Duperron S, De Beer D, Zbinden M, Boetius A, Schipani V, Kahil N, *et al.* (2009). Molecular characterization of bacteria associated with the trophosome and the tube of Lamellibrachia sp., a siboglinid annelid from cold seeps in the eastern Mediterranean. *FEMS Microbiol Ecol* **69**: 395–409.

Duperron S, Halary S, Lorion J, Sibuet M, Gaill F. (2008). Unexpected co-occurrence of six bacterial symbionts in the gills of the cold seep mussel Idas sp.(Bivalvia: Mytilidae). *Environ Microbiol* **10**: 433–445.

Duperron S, Nadalig T, Caprais J-C, Sibuet M, Fiala-Médioni A, Amann R, *et al.* (2005). Dual symbiosis in a Bathymodiolus sp. mussel from a methane seep on the Gabon continental margin (Southeast Atlantic): 16S rRNA phylogeny and distribution of the symbionts in gills. *Appl Environ Microbiol* **71**: 1694–1700.

Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, *et al.* (2010). A map of human genome variation from population-scale sequencing. *PMC*. http://dspace.mit.edu/handle/1721.1/74035 (Accessed March 1, 2016).

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–2200.

Edgcomb VP, Leadbetter ER, Bourland W, Beaudoin D, Bernhard JM. (2011). Structured multiple endosymbiosis of bacteria and archaea in a ciliate from marine sulfidic sediments: a survival mechanism in low oxygen, sulfidic sediments? *Front Microbiol* **2**.

Edwards DB, Nelson DC. (1991). DNA-DNA solution hybridization studies of the bacterial symbionts of hydrothermal vent tube worms (Riftia pachyptila and Tevnia jerichonana). *Appl Environ Microbiol* **57**: 1082–1088.

Elsaied H, Kimura H, Naganuma T. (2002). Molecular characterization and endosymbiotic localization of the gene encoding D-ribulose 1, 5-bisphosphate carboxylase–oxygenase (RuBisCO) form II in the deep-sea vestimentiferan trophosome. *Microbiology* **148**: 1947–1957.

Ewald PW. (1987). Transmission Modes and Evolution of the Parasitism-Mutualism Continuuma. *Ann N Y Acad Sci* **503**: 295–306.

Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, *et al.* (2012). CRISPR Typing and Subtyping for Improved Laboratory Surveillance of Salmonella Infections. *PLoS ONE* **7**: e36995.

Farrelly V, Rainey FA, Stackebrandt E. (1995). Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl Environ Microbiol* **61**: 2798–2801.

Fay SA, Weber MX. (2012). The Occurrence of Mixed Infections of Symbiodinium (Dinoflagellata) within Individual Hosts. *J Phycol* **48**: 1306–1316.

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

Felbeck H, Jarchow J. (1998). Carbon release from purified chemoautotrophic bacterial symbionts of the hydrothermal vent tubeworm Riftia pachyptila. *Physiol Biochem Zool* **71**: 294–302.

Feldman R, Black M, Cary C, Lutz R, Vrijenhoek R. (1997). Molecular phylogenetics of bacterial endosymbionts and their vestimentiferan hosts. *Mol Mar Biol Biotechnol* **6**: 268.

Feng PCH, Delannoy S, Lacher DW, Santos LF dos, Beutin L, Fach P, *et al.* (2014). Genetic Diversity and Virulence Potential of Shiga Toxin-Producing Escherichia coli O113:H21 Strains Isolated from Clinical, Environmental, and Food Sources. *Appl Environ Microbiol* **80**: 4757–4763.

Fiala-Médioni A, McKiness Z, Dando P, Boulegue J, Mariotti A, Alayse-Danet A, *et al.* (2002). Ultrastructural, biochemical, and immunological characterization of two populations of the mytilid mussel Bathymodiolus azoricus from the Mid-Atlantic Ridge: evidence for a dual symbiosis. *Mar Biol* **141**: 1035–1043.

Forget NL, Juniper SK. (2013). Free-living bacterial communities associated with tubeworm (Ridgeia piscesae) aggregations in contrasting diffuse flow hydrothermal vent habitats at the Main Endeavour Field, Juan de Fuca Ridge. *MicrobiologyOpen*.

Forget NL, Perez M, Juniper SK. (2014). Molecular study of bacterial diversity within the trophosome of the vestimentiferan tubeworm Ridgeia piscesae. *Mar Ecol* **36**: 35–44.

Fraser C, Hanage WP, Spratt BG. (2007). Recombination and the Nature of Bacterial Speciation. *Science* **315**: 476–480.

Fujiwara Y, Kato C, Masui N, Fujikura K, Kojima S. (2001). Dual symbiosis in the cold-seep thyasirid clam Maorithyas hadalis from the hadal zone in the Japan Trench, western Pacific. *Mar Ecol Prog Ser* **214**: 151–159.

Gardebrecht A, Markert S, Sievert SM, Felbeck H, Thürmer A, Albrecht D, *et al.* (2011). Physiological homogeneity among the endosymbionts of Riftia pachyptila and Tevnia jerichonana revealed by proteogenomics. *ISME J* **6**: 766–776.

Genkai-Kato M, Yamamura N. (1999). Evolution of Mutualistic Symbiosis without Vertical Transmission. *Theor Popul Biol* **55**: 309–323.

Gomez-Alvarez V, Teal TK, Schmidt TM. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3**: 1314–1317.

Gouy M, Guindon S, Gascuel O. (2010). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol Biol Evol* **27**: 221–224.

Grzymski JJ, Murray AE, Campbell BJ, Kaplarevic M, Gao GR, Lee C, *et al.* (2008). Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. *Proc Natl Acad Sci* **105**: 17516–17521.

Guri M, Durand L, Cueff-Gauchard V, Zbinden M, Crassous P, Shillito B, *et al.* (2012). Acquisition of epibiotic bacteria along the life cycle of the hydrothermal shrimp Rimicaris exoculata. *ISME J* **6**: 597–609.

Haddad A, Camacho F, Durand P, Cary SC. (1995). Phylogenetic characterization of the epibiotic bacteria associated with the hydrothermal vent polychaete Alvinella pompejana. *Appl Environ Microbiol* **61**: 1679–1687.

Halanych KM, Lutz R, Vrijenhoek RC. (1998). Evolutionary origins and age of vestimentiferan tube-worms. *Cah Biol Mar* **39**: 355–358.

Harmer TL, Rotjan RD, Nussbaumer AD, Bright M, Ng AW, DeChaine EG, *et al.* (2008). Free-living tube worm endosymbionts found at deep-sea vents. *Appl Environ Microbiol* **74**: 3895–3898.

Hasegawa M, Kishino H, Yano T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**: 160–174.

He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, *et al.* (2010). Evolutionary dynamics of Clostridium difficile over short and long time scales. *Proc Natl Acad Sci* **107**: 7527–7532.

Held NL, Herrera A, Cadillo-Quiroz H, Whitaker RJ. (2010). CRISPR Associated Diversity within a Population of Sulfolobus islandicus. *PLoS ONE* **5**: e12988.

Hentschel U, Felbeck H. (1993). Nitrate respiration in the hydrothermal vent tubeworm Riftia pachyptila.

Hollander M, Wolfe DA, Chicken E. (2013). Nonparametric statistical methods. John Wiley & Sons.

Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, *et al.* (2008). Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. *J Bacteriol* **190**: 1401–1412.

Hoshino T, Yilmaz LS, Noguera DR, Daims H, Wagner M. (2008). Quantification of target molecules needed to detect microorganisms by fluorescence in situ hybridization (FISH) and catalyzed reporter deposition-FISH. *Appl Environ Microbiol* **74**: 5068–5077.

Huang HW, Mullikin JC, Hansen NF. (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics* **16**: 235.

Hubber A, Vergunst AC, Sullivan JT, Hooykaas PJJ, Ronson CW. (2004). Symbiotic phenotypes and translocated effector proteins of the Mesorhizobium loti strain R7A VirB/D4 type IV secretion system. *Mol Microbiol* **54**: 561–574.

Hurtado LA, Lutz RA, Vrijenhoek R c. (2004). Distinct patterns of genetic differentiation among annelids of eastern Pacific hydrothermal vents. *Mol Ecol* **13**: 2603–2615.

Huse SM, Welch DM, Morrison HG, Sogin ML. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**: 1889–1898.

Ihaka R, Gentleman R. (1996). R: A Language for Data Analysis and Graphics. *J Comput Graph Stat* **5**: 299–314.

Jani AJ, Cotter PA. (2010). Type VI Secretion: Not Just for Pathogenesis Anymore. *Cell Host Microbe* **8**: 2–6.

Jannasch HW, Wirsen CO. (1979). Chemosynthetic primary production at East Pacific sea floor spreading centers. *Bioscience* **29**: 592–598.

Johnson SB, Young CR, Jones WJ, Warén A, Vrijenhoek RC. (2006). Migration, Isolation, and Speciation of Hydrothermal Vent Limpets (Gastropoda; Lepetodrilidae) Across the Blanco Transform Fault. *Biol Bull* **210**: 140–157.

Jones KM, Kobayashi H, Davies BW, Taga ME, Walker GC. (2007). How rhizobial symbionts invade plants: the Sinorhizobium-Medicago model. *Nat Rev Microbiol* **5**: 619–633.

Kalanetra KM, Nelson DC. (2010). Vacuolate-attached filaments: highly productive Ridgeia piscesae epibionts at the Juan de Fuca hydrothermal vents. *Mar Biol* **157**: 791–800.

Karl DM, Wirsen CO, Jannasch HW. (1980). Deep-sea primary production at the Galapagos hydrothermal vents. *Sci States* **207**. http://www.osti.gov/scitech/biblio/5341819 (Accessed March 15, 2016).

Katoh K, Standley DM. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**: 772–780.

Kim O-S, Cho Y-J, Lee K, Yoon S-H, Kim M, Na H, *et al.* (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol* **62**: 716–721.

Kimura H, Higashide Y, Naganuma T. (2003). Endosymbiotic microflora of the vestimentiferan tubeworm (Lamellibrachia sp.) from a bathyal cold seep. *Mar Biotechnol* **5**: 593–603.

Kleiner M, Petersen JM, Dubilier N. (2012a). Convergent and divergent evolution of metabolism in sulfur-oxidizing symbionts and the role of horizontal gene transfer. *Curr Opin Microbiol*.

Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, *et al.* (2012b). Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc Natl Acad Sci* **109**: E1173–E1182.

Klose J, Aistleitner K, Horn M, Krenn L, Dirsch V, Zehl M, *et al.* (2016). Trophosome of the Deep-Sea Tubeworm Riftia pachyptila Inhibits Bacterial Growth: e0146446. *PLoS One* **11**. e-pub ahead of print, doi: http://dx.doi.org/10.1371/journal.pone.0146446.

Klose J, Polz MF, Wagner M, Schimak MP, Gollner S, Bright M. (2015). Endosymbionts escape dead hydrothermal vent tubeworms to enrich the free-living population. *Proc Natl Acad Sci* **112**: 11300–11305.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, *et al.* (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**: 2283–2285.

Kovanen S m., Kivistö R i., Rossi M, Hänninen M-L. (2014). A combination of MLST and CRISPR typing reveals dominant Campylobacter jejuni types in organically farmed laying hens. *J Appl Microbiol* **117**: 249–257.

Kryazhimskiy S, Plotkin JB. (2008). The Population Genetics of dN/dS. *PLoS Genet* **4**: e1000304.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118–123.

Kunin V, Sorek R, Hugenholtz P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**: R61.

Kuno S, Sako Y, Yoshida T. (2014). Diversification of CRISPR within coexisting genotypes in a natural population of the bloom-forming cyanobacterium Microcystis aeruginosa. *Microbiology* **160**: 903–916.

Kuno S, Yoshida T, Kaneko T, Sako Y. (2012). Intricate Interactions between the Bloom-Forming Cyanobacterium Microcystis aeruginosa and Foreign Genetic Elements, Revealed by Diversified Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Signatures. *Appl Environ Microbiol* **78**: 5353–5360.

Kuo C-H, Moran NA, Ochman H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19**: 1450–1454.

LaJeunesse TC, Thornhill DJ, Cox EF, Stanton FG, Fitt WK, Schmidt GW. (2004). High diversity and host specificity observed among symbiotic dinoflagellates in reef coral communities from Hawaii. *Coral Reefs* **23**: 596–603.

Langella O. (2002). POPULATIONS 1.2. 28. Population genetic software (individuals or populations distances, phylogenetic trees). *CNRS Fr*.

Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Li H, Durbin R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl* **26**: 589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li H, Homer N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* **11**: 473–483.

Liao L, Wankel SD, Wu M, Cavanaugh CM, Girguis PR. (2013). Characterizing the plasticity of nitrogen metabolism by the host and symbionts of the hydrothermal vent chemoautotrophic symbioses Ridgeia piscesae. *Mol Ecol*.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, *et al.* (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* **30**: 434–439.

López-García P, Gaill F, Moreira D. (2002). Wide bacterial diversity associated with tubes of the vent worm Riftia pachyptila. *Environ Microbiol* **4**: 204–215.

Lopez-Sanchez M-J, Sauvage E, Da Cunha V, Clermont D, Ratsima Hariniaina E, Gonzalez-Zorn B, *et al.* (2012). The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. *Mol Microbiol* **85**: 1057–1071.

Loy A, Maixner F, Wagner M, Horn M. (2007). probeBase—an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Res* **35**: D800–D804.

Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. (2014). Comparing effective population sizes of dominant marine alphaproteobacteria lineages. *Environ Microbiol Rep* **6**: 167–172.

Macdonald IR, Tunnicliffe V, Southward EC. (2002). Detection of sperm transfer and synchronous fertilization in Ridgeia piscesae at Endeavour Segment, Juan de Fuca Ridge. *Cah Biol Mar* **43**: 395–398.

Maechler M. (2013). Package 'diptest'. R package version 0.75-5. *Retrieved HttpzlCRAN R-Proj Orgpackage Diptest*.

Malakhov V, Popelyaev I, Galkin S. (1996). Microscopic anatomy of Ridgeia phaeophiale Jones, 1985 (Pogonophora, Vestimentifera) and the problem of the position of Vestimentifera in the system of the animal kingdom: III. Rudimentary digestive system, trophosome, and blood vascular system. *Russ J Mar Biol CC Biol MORIA* **22**: 125–136.

Mariette J, Noirot C, Klopp C. (2011). Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Res Notes* **4**: 149.

Marinier E, Brown DG, McConkey BJ. (2015). Pollux: platform independent error correction of single and mixed genomes. *BMC Bioinformatics* **16**: 10.

Markert S, Arndt C, Felbeck H, Becher D, Sievert SM, Hügler M, *et al.* (2007). Physiological Proteomics of the Uncultured Endosymbiont of Riftia pachyptila. *Science* **315**: 247–250.

Markert S, Gardebrecht A, Felbeck H, Sievert SM, Klose J, Becher D, *et al.* (2011). Status quo in physiological proteomics of the uncultured Riftia pachyptila endosymbiont. *Proteomics* **11**: 3106–3117.

Marsh AG, Mullineaux LS, Young CM, Manahan DT. (2001). Larval dispersal potential of the tubeworm Riftia pachyptila at deep-sea hydrothermal vents. *Nature* **411**: 77–80.

McMullin ER, Hourdez S, Schaeffer SW, Fisher CR. (2003). Phylogeny and biogeography of deep sea vestimentiferan tubeworms and their bacterial symbionts. *Symbiosis* **34**: 1–41.

Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**: 451.

Mielczarek M, Szyda J. (2015). Review of alignment and SNP calling algorithms for next-generation sequencing data. *J Appl Genet* 1–9.

Mullineaux L, Speer K, Thurnherr A, Maltrud M, Vangriesheim A. (2002). Implications of cross-axis flow for larval dispersal along mid-ocean ridges. *CBM-Cah Biol Mar* **43**: 281–284.

Needleman SB, Wunsch CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453.

Nees HA, Lutz RA, Shank TM, Luther III GW. (2009). Pre- and post-eruption diffuse flow variability among tubeworm habitats at 9°50′ north on the East Pacific Rise. *Deep Sea Res Part II Top Stud Oceanogr* **56**: 1607–1615.

Nelson K, Fisher C. (2000). Absence of cospeciation in deep-sea vestimentiferan tube worms and their bacterial endosymbionts. *Symbiosis* **28**: 1–15.

Nussbaumer AD, Fisher CR, Bright M. (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* **441**: 345–348.

Nyholm SV, McFall-Ngai M. (2004). The winnowing: establishing the squid–Vibrio symbiosis. *Nat Rev Microbiol* **2**: 632–642.

Nyholm SV, Robidart J, Girguis PR. (2008). Coupling metabolite flux to transcriptomics: insights into the molecular mechanisms underlying primary productivity by the hydrothermal vent tubeworm Ridgeia piscesae. *Biol Bull* **214**: 255–265.

Nyholm SV, Song P, Dang J, Bunce C, Girguis PR. (2012). Expression and Putative Function of Innate Immunity Genes under in situ Conditions in the Symbiotic Hydrothermal Vent Tubeworm Ridgeia piscesae. *PloS One* **7**: e38267.

Ochman H, Elwyn S, Moran NA. (1999). Calibrating bacterial evolution. *Proc Natl Acad Sci* **96**: 12638–12643.

Ochman H, Wilson AC. (1987). Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. *J Mol Evol* **26**: 74–86.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara R, *et al.* (2015). Package 'vegan'.

Okubo T, Ikeda S, Yamashita A, Terasawa K, Minamisawa K. (2012). Pyrosequence read length of 16S rRNA gene affects phylogenetic assignment of plant-associated bacteria. *Microbes Environ* **27**: 204–208.

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, *et al.* (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* **5**: 28.

Ortmann AC, Suttle CA. (2005). High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. *Deep Sea Res Part Oceanogr Res Pap* **52**: 1515–1527.

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, *et al.* (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* **15**: 256–278.

Pagé A, Juniper SK, Olagnon M, Alain K, Desrosiers G, Querellou J, *et al.* (2004). Microbial diversity associated with a Paralvinella sulfincola tube and the adjacent substratum on an active deep-sea vent chimney. *Geobiology* **2**: 225–238.

Paradis E, Claude J, Strimmer K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

Pernthaler A, Pernthaler J, Amann R, Kowalchuk GA, de Bruijn FJ, Head IM, *et al.* (2004). Sensitive multi-color fluorescence in situ hybridization for the identification of environmental microorganisms. *Mol Microb Ecol Man Vol 1 2* 711–725.

Petersen JM, Ramette A, Lott C, Cambon-Bonavita M, Zbinden M, Dubilier N. (2010). Dual symbiosis of the vent shrimp Rimicaris exoculata with

filamentous gamma-and epsilonproteobacteria at four Mid-Atlantic Ridge hydrothermal vent fields. *Environ Microbiol* **12**: 2204–2218.

Peterson GI, Masel J. (2009). Quantitative Prediction of Molecular Clock and Ka/Ks at Short Timescales. *Mol Biol Evol* **26**: 2595–2603.

Pflugfelder B, Cary SC, Bright M. (2009). Dynamics of cell proliferation and apoptosis reflect different life strategies in hydrothermal vent and cold seep vestimentiferan tubeworms. *Cell Tissue Res* **337**: 149–165.

Pinto AJ, Raskin L. (2012). PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. *PLoS ONE* **7**: e43093.

Plouviez S, Shank TM, Faure B, Daguin-Thiebaut C, Viard F, Lallier FH, *et al.* (2009). Comparative phylogeography among hydrothermal vent species along the East Pacific Rise reveals vicariant processes and population expansion in the South. *Mol Ecol* **18**: 3903–3917.

Ponsard J, Cambon-Bonavita M-A, Zbinden M, Lepoint G, Joassin A, Corbari L, *et al.* (2013). Inorganic carbon fixation by chemosynthetic ectosymbionts and nutritional transfers to the hydrothermal vent host-shrimp Rimicaris exoculata. *ISME J* **7**: 96–109.

Pourcel C, Salvignol G, Vergnaud G. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**: 653–663.

Provorov NA, Tikhonovich IA. (2015). Genetic and molecular basis of symbiotic adaptations. *Biol Bull Rev* **4**: 443–456.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, *et al.* (2008). A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**: 1005–1010.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, *et al.* (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.

Robidart JC. (2006). Metagenomics of the Riftia pachyptila symbiont. University of California, San Diego. http://gradworks.umi.com/32/37/3237569.html (Accessed April 19, 2016).

Robidart JC, Bench SR, Feldman RA, Novoradovsky A, Podell SB, Gaasterland T, *et al.* (2008). Metabolic versatility of the Riftia pachyptila endosymbiont revealed through metagenomics. *Environ Microbiol* **10**: 727–737.

Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, *et al.* (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* **239**: 226–235.

Rodriguez-Lanetty M, Wood-Charlson EM, Hollingsworth LL, Krupp DA, Weis VM. (2006). Temporal and spatial infection dynamics indicate recognition events in the early hours of a dinoflagellate/coral symbiosis. *Mar Biol* **149**: 713–719.

Sachs JL, Essenberg CJ, Turcotte MM. (2011). New paradigms for the evolution of beneficial infections. *Trends Ecol Evol* **26**: 202–209.

Santos SR, Gutiérrez-Rodríguez C, Lasker HR, Coffroth MA. (2003). Symbiodinium sp. associations in the gorgonian Pseudopterogorgia elisabethae in the Bahamas: high levels of genetic variability and population structure in symbiotic dinoflagellates. *Mar Biol* **143**: 111–120.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microbiol* **75**: 7537–7541.

Schmieder R, Edwards R. (2011a). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS One* **6**: e17288.

Schmieder R, Edwards R. (2011b). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863–864.

Scott K, Bright M, Fisher C. (1998). The burden of independence: Inorganic carbon utilization strategies of the sulphur chemoautotrophic hydrothermal vent isolate Thiomicrospira crunogena and the symbionts of hydrothermal vent and cold seep vestimentiferans. *Cah Biol Mar* **39**: 379–381.

Scott K, Bright M, Macko S, Fisher C. (1999). Carbon dioxide use by chemoautotrophic endosymbionts of hydrothermal vent vestimentiferans: affinities for carbon dioxide, absence of carboxysomes, and δ13C values. *Mar Biol* **135**: 25–34.

Shapiro BJ, David LA, Friedman J, Alm EJ. (2009). Looking for Darwin's footprints in the microbial world. *Trends Microbiol* **17**: 196–204.

Sievers F, Higgins D. (2014). Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. In: Russell DJ (ed) Methods in Molecular Biology. *Multiple Sequence Alignment Methods*. Humana Press, pp 105–116.

Skennerton CT, Imelfort M, Tyson GW. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res* **41**: e105–e105.

Smith JM, Smith NH, O'Rourke M, Spratt BG. (1993). How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**: 4384–4388.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci* **103**: 12115–12120.

Sorek R, Kunin V, Hugenholtz P. (2008). CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**: 181–186.

Southward EC, Coates KA. (1989). Sperm masses and sperm transfer in a vestimentiferan, Ridgeia piscesae Jones, 1985 (Pogonophora: Obturata). *Can J Zool* **67**: 2776–2781.

Southward EC, Tunnicliffe V, Black M. (1995). Revision of the species of Ridgeia from northeast Pacific hydrothermal vents, with a redescription of Ridgeia piscesae Jones (Pogonophora: Obturata= Vestimentifera). *Can J Zool* **73**: 282–295.

Stat M, Yost DM, Gates RD. (2015). Geographic structure and host specificity shape the community composition of symbiotic dinoflagellates in corals from the Northwestern Hawaiian Islands. *Coral Reefs* **34**: 1075–1086.

Stewart FJ, Young CR, Cavanaugh CM. (2009). Evidence for homologous recombination in intracellular chemosynthetic clam symbionts. *Mol Biol Evol* **26**: 1391–1404.

Suyama M, Torrents D, Bork P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609–W612.

Suzuki MT, Giovannoni SJ. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62**: 625–630.

Suzuki Y, Sasaki T, Suzuki M, Nogi Y, Miwa T, Takai K, *et al.* (2005). Novel chemoautotrophic endosymbiosis between a member of the Epsilonproteobacteria and the hydrothermal-vent gastropod Alviniconcha aff. hessleri (Gastropoda: Provannidae) from the Indian Ocean. *Appl Environ Microbiol* **71**: 5440–5450.

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, *et al.* (2010). 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytol* **188**: 291–301.

Thiel V, Hügler M, Blümel M, Baumann HI, Gärtner A, Schmaljohann R, *et al.* (2012). Widespread occurrence of two carbon fixation pathways in tubeworm endosymbionts: lessons from hydrothermal vent associated tubeworms from the Mediterranean Sea. *Front Microbiol* **3**.

Tunnicliffe V. (1988). Biogeography and Evolution of Hydrothermal-Vent Fauna in the Eastern Pacific Ocean. *Proc R Soc Lond B Biol Sci* **233**: 347–366.

Tunnicliffe V, Germain CS, Hilário A. (2014). Phenotypic Variation and Fitness in a Metapopulation of Tubeworms (Ridgeia piscesae Jones) at Hydrothermal Vents. *PLOS ONE* **9**: e110578.

Tyson GW, Banfield JF. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* **10**: 200–207.

Urakawa H, Dubilier N, Fujiwara Y, Cunningham DE, Kojima S, Stahl DA. (2005). Hydrothermal vent gastropods from the same family (Provannidae) harbour ε-and γ-proteobacterial endosymbionts. *Environ Microbiol* **7**: 750–754.

Urcuyo IA, Bergquist DC, MacDonald IR, VanHorn M, Fisher CR. (2007). Growth and longevity of the tubeworm Ridgeia piscesae in the variable diffuse flow habitats of the Juan de Fuca Ridge. *Mar Ecol Prog Ser* **344**: 143–157.

Urcuyo IA, Massoth GJ, Julian D, Fisher CR. (2003). Habitat, growth and physiological ecology of a basaltic community of Ridgeia piscesae from the Juan de Fuca Ridge. *Deep Sea Res Part Oceanogr Res Pap* **50**: 763–780.

Urcuyo I, Massoth G, MacDonald I, Fisher C. (1998). In situ growth of the vestimentiferan Ridgeia piscesae living in highly diffuse flow environments in the main Endeavour Segment of the Juan de Fuca Ridge. *Cah Biol Mar* **39**: 267–270.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, *et al.* (2002). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1110s43/abstract (Accessed September 30, 2015).

Vrijenhoek RC. (2010a). Genetic diversity and connectivity of deep-sea hydrothermal vent metapopulations. *Mol Ecol* **19**: 4391–4411.

Vrijenhoek RC. (2010b). Genetics and evolution of deep-sea chemosynthetic bacteria and their invertebrate hosts. In: *The Vent and Seep Biota*. Springer, pp 15–49.

Vrijenhoek RC. (2013). On the instability and evolutionary age of deep-sea chemosynthetic communities. *Deep Sea Res Part II Top Stud Oceanogr* **92**: 189–200.

Wagner A, Blackstone N, Cartwright P, Dick M, Misof B, Snow P, *et al.* (1994). Surveys of Gene Families Using Polymerase Chain Reaction: PCR Selection and PCR Drift. *Syst Biol* **43**: 250–261.

Wallner G, Amann R, Beisker W. (1993). Optimizing fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms. *Cytometry* **14**: 136–143.

Wang H, Ooi BC, Tan K-L, Ong T-H, Zhou L. (2003). BLAST++: BLASTing queries in batches. *Bioinformatics* **19**: 2323–2324.

Wang W, Xu F, Wang J. (2013). Assessment of Mapping and SNP-Detection Algorithms for Next-Generation Sequencing Data in Cancer Genomics. In: Wu W, Choudhry H (eds). *Next Generation Sequencing in Cancer Research*. Springer New York, pp 301–317.

Watsuji T, Nishizawa M, Morono Y, Hirayama H, Kawagucci S, Takahata N, *et al.* (2012). Cell-specific thioautotrophic productivity of epsilon-proteobacterial epibionts associated with Shinkaia crosnieri. *PloS One* **7**: e46282.

Westra ER, Buckling A, Fineran PC. (2014). CRISPR-Cas systems: beyond adaptive immunity. *Nat Rev Microbiol* **12**: 317–326.

Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, *et al.* (2011). Mutation Rate Inferred From Synonymous Substitutions in a

Long-Term Evolution Experiment With Escherichia coli. *G3 Genes Genomes Genet* **1**: 183–186.

Wilkinson DM, Sherratt TN. (2001). Horizontally acquired mutualisms, an unsolved problem in ecology? *Oikos* **92**: 377–384.

Wilmot DBJ, Vetter RD. (1990). The bacterial symbiont from the hydrothermal vent tubewormRiftia pachyptila is a sulfide specialist. *Mar Biol* **106**: 273–283.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, *et al.* (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**: 950–955.

Yang Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**: 1586–1591.

Yang Z, Nielsen R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Mol Biol Evol* **17**: 32–43.

Yi M, Zhao Y, Jia L, He M, Kebebew E, Stephens RM. (2014). Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Res* gku392.

Yin S, Jensen MA, Bai J, DebRoy C, Barrangou R, Dudley EG. (2013). The Evolutionary Divergence of Shiga Toxin-Producing Escherichia coli Is Reflected in Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Spacer Composition. *Appl Environ Microbiol* **79**: 5710–5720.

Young C, Fujio S, Vrijenhoek R. (2008). Directional dispersal between mid-ocean ridges: deep-ocean circulation and gene flow in Ridgeia piscesae. *Mol Ecol* **17**: 1718–1731.

Yu F, Lu J, Liu X, Gazave E, Chang D, Raj S, *et al.* (2015). Population Genomic Analysis of 962 Whole Genome Sequences of Humans Reveals Natural Selection in Non-Coding Regions. *PLOS ONE* **10**: e0121644.

Yu T, Li J, Yang Y, Qi L, Chen B, Zhao F, *et al.* (2012). Codon usage patterns and adaptive evolution of marine unicellular cyanobacteria Synechococcus and Prochlorococcus. *Mol Phylogenet Evol* **62**: 206–213.

Yu X, Sun S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* **14**: 274.

Zbinden M, Pailleret M, Ravaux J, Gaudron SM, Hoyoux C, Lambourdière J, *et al.* (2010). Bacterial communities associated with the wood-feeding gastropod Pectinodonta sp.(Patellogastropoda, Mollusca). *FEMS Microbiol Ecol* **74**: 450–463.

Zbinden M, Shillito B, Le Bris N, de Montlaur CDV, Roussel E, Guyot F, *et al.* (2008). New insigths on the metabolic diversity among the epibiotic microbial communitiy of the hydrothermal shrimp Rimicaris exoculata. *J Exp Mar Biol Ecol* **359**: 131–140.

Zhang Y, Lin K. (2012). A phylogenomic analysis of Escherichia coli / Shigella group: implications of genomic features associated with pathogenicity and ecological adaptation. *BMC Evol Biol* **12**: 174.

Zhou J, Wu L, Deng Y, Zhi X, Jiang Y-H, Tu Q, *et al.* (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* **5**: 1303–1313.

Zimmermann J, Lott C, Weber M, Ramette A, Bright M, Dubilier N, *et al.* (2014). Dual symbiosis with co-occurring sulfur-oxidizing symbionts in vestimentiferan tubeworms from a Mediterranean hydrothermal vent. *Environ Microbiol*.

Appendices

Appendix A Supplementary information for Chapter 2, Part ONE

List A1 List of primers used for pyrosequencing

To achieve the PCR amplifications of the bacterial SSU rRNA hypervariable regions V1-V3, a general reverse primer (R519) combined with B primer (Roche) was used in combination with a unique to tagged forward primer (F63-targeted) combined with A primer (Roche):

519R CCTATCCCCTGTGTGCCTTGGCAGTCTCAGGWATTACCGCGGCKGCTG
F63-1 CCATCTCATCCCTGCGTGTCTCCGACTCAGACGAGTGCGTCAGGCCTAACACATGCAAGTC
F63-2 CCATCTCATCCCTGCGTGTCTCCGACTCAGACGCTCGACACAGGCCTAACACATGCAAGTC
F63-3 CCATCTCATCCCTGCGTGTCTCCGACTCAGAGACGCACTCCAGGCCTAACACATGCAAGTC
F63-4 CCATCTCATCCCTGCGTGTCTCCGACTCAGAGCACTGTAGCAGGCCTAACACATGCAAGTC
F63-5 CCATCTCATCCCTGCGTGTCTCCGACTCAGATCAGACACGCAGGCCTAACACATGCAAGTC
F63-6 CCATCTCATCCCTGCGTGTCTCCGACTCAGATATCGCGAGCAGGCCTAACACATGCAAGTC
F63-7 CCATCTCATCCCTGCGTGTCTCCGACTCAGCGTGTCTCTACAGGCCTAACACATGCAAGTC
F63-8 CCATCTCATCCCTGCGTGTCTCCGACTCAGCTCGCGTGTCCAGGCCTAACACATGCAAGTC
F63-9 CCATCTCATCCCTGCGTGTCTCCGACTCAGTCTCTATGCGCAGGCCTAACACATGCAAGTC
F63-10 CCATCTCATCCCTGCGTGTCTCCGACTCAGTGATACGTCTCAGGCCTAACACATGCAAGTC
F63-11 CCATCTCATCCCTGCGTGTCTCCGACTCAGCATAGTAGTGCAGGCCTAACACATGCAAGTC
F63-12 CCATCTCATCCCTGCGTGTCTCCGACTCAGCGAGAGATACCAGGCCTAACACATGCAAGTC
F63-13 CCATCTCATCCCTGCGTGTCTCCGACTCAGATACGACGTACAGGCCTAACACATGCAAGTC
F63-14 CCATCTCATCCCTGCGTGTCTCCGACTCAGTCACGTACTACAGGCCTAACACATGCAAGTC
F63-15 CCATCTCATCCCTGCGTGTCTCCGACTCAGCGTCTAGTACCAGGCCTAACACATGCAAGTC
F63-16 CCATCTCATCCCTGCGTGTCTCCGACTCAGTCTACGTAGCCAGGCCTAACACATGCAAGTC
F63-17 CCATCTCATCCCTGCGTGTCTCCGACTCAGTGTACTACTCCAGGCCTAACACATGCAAGTC
F63-18 CCATCTCATCCCTGCGTGTCTCCGACTCAGACGACTACAGCAGGCCTAACACATGCAAGTC
F63-19 CCATCTCATCCCTGCGTGTCTCCGACTCAGCGTAGACTAGCAGGCCTAACACATGCAAGTC
F63-20 CCATCTCATCCCTGCGTGTCTCCGACTCAGTACGAGTATGCAGGCCTAACACATGCAAGTC
F63-21 CCATCTCATCCCTGCGTGTCTCCGACTCAGTACTCTCGTGCAGGCCTAACACATGCAAGTC
F63-22 CCATCTCATCCCTGCGTGTCTCCGACTCAGTAGAGACGAGCAGGCCTAACACCATGCAAGTC
F63-23 CCATCTCATCCCTGCGTGTCTCCGACTCAGTCGTCGCTCGCAGGCCTAACACATGCAAGTC
F63-24 CCATCTCATCCCTGCGTGTCTCCGACTCAGACATACGCGTCAGGCCTAACACATGCAAGTC

F63-25 CCATCTCATCCCTGCGTGTCTCCGACTCAGACGCGAGTATCAGGCCTAACACATGCAAGTC
F63-26 CCATCTCATCCCTGCGTGTCTCCGACTCAGACTACTATGTCAGGCCTAACACATGCAAGTC
F63-27 CCATCTCATCCCTGCGTGTCTCCGACTCAGACTGTACAGTCAGGCCTAACACATGCAAGTC
F63-28 CCATCTCATCCCTGCGTGTCTCCGACTCAGAGACTATACTCAGGCCTAACACATGCAAGTC
F63-29 CCATCTCATCCCTGCGTGTCTCCGACTCAGAGCGTCGTCTCAGGCCTAACACATGCAAGTC
F63-30 CCATCTCATCCCTGCGTGTCTCCGACTCAGAGTACGCTATCAGGCCTAACACATGCAAGTC
F63-31 CCATCTCATCCCTGCGTGTCTCCGACTCAGATAGAGTACTCAGGCCTAACACATGCAAGTC
F63-32 CCATCTCATCCCTGCGTGTCTCCGACTCAGCACGCTACGTCAGGCCTAACACATGCAAGTC
F63-33 CCATCTCATCCCTGCGTGTCTCCGACTCAGCAGTAGACGTCAGGCCTAACACATGCAAGTC
F63-34 CCATCTCATCCCTGCGTGTCTCCGACTCAGCGACGTGACTCAGGCCTAACACATGCAAGTC
F63-35 CCATCTCATCCCTGCGTGTCTCCGACTCAGTACACACACTCAGGCCTAACACATGCAAGTC
F63-36 CCATCTCATCCCTGCGTGTCTCCGACTCAGTACACGTGATCAGGCCTAACACATGCAAGTC
F63-37 CCATCTCATCCCTGCGTGTCTCCGACTCAGTACAGATCGTCAGGCCTAACACATGCAAGTC
F63-38 CCATCTCATCCCTGCGTGTCTCCGACTCAGTACGCTGTCTCAGGCCTAACACATGCAAGTC
F63-39 CCATCTCATCCCTGCGTGTCTCCGACTCAGTAGTGTAGATCAGGCCTAACACATGCAAGTC
F63-40 CCATCTCATCCCTGCGTGTCTCCGACTCAGTCGATCACGTCAGGCCTAACACATGCAAGTC
F63-41 CCATCTCATCCCTGCGTGTCTCCGACTCAGTCGCACTAGTCAGGCCTAACACATGCAAGTC
F63-42 CCATCTCATCCCTGCGTGTCTCCGACTCAGTCTAGCGACTCAGGCCTAACACATGCAAGTC
F63-43 CCATCTCATCCCTGCGTGTCTCCGACTCAGTCTATACTATCAGGCCTAACACATGCAAGTC

Appendix B Glossaries for Chapters 3 and 4

Table B.1 Concepts and vocabulary pertaining to Chapters 3 and 4.

	Term	Definition
General	Reads	Short (100-500 bp) sequences of DNA obtained from high throughput sequencing technology
	Coverage (depth of)	Number of reads representing a given nucleotide
	Contigs	Contiguous sequences
	Scaffolds	Collection of contigs separated by gaps of known length
	Assembly	Genome assembled from reads. Usually a collection of scaffolds/reads
	CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats. Part of the adaptive immune system of Prokaryotes
Chapter 3	Genetic variant	Insertion, deletion, or substitution in the DNA sequence leading to polymorphism
	Structural variant	Chromosome rearrangement
	Mapping	Aligning reads to a reference sequence
	Calling variants	Detecting genetic variants
Chapter 4	Pan-genome	Union of all representative genomes (strains) in a clade which can have large variation in gene content
	Core genome	Part of the pan-genome that is shared amongst all strains
	Accessory genome	Part of the pan-genome that is not shared amongst all strains but unique to some
	LCBs	Locally Collinear Blocks. Genome segments that appear free of chromosomic rearrangement
	dN/dS	Ratio of divergence at nonsynonymous and synonymous nucleotide substitution sites. Relative measure of selection and genetic drift. Calculated as: $\frac{non\text{-}synonymous\ substitutions/non\text{-}synonymous\ sites\ (dN)}{synonymous\ substitutions/synonymous\ sites\ (dS)}$

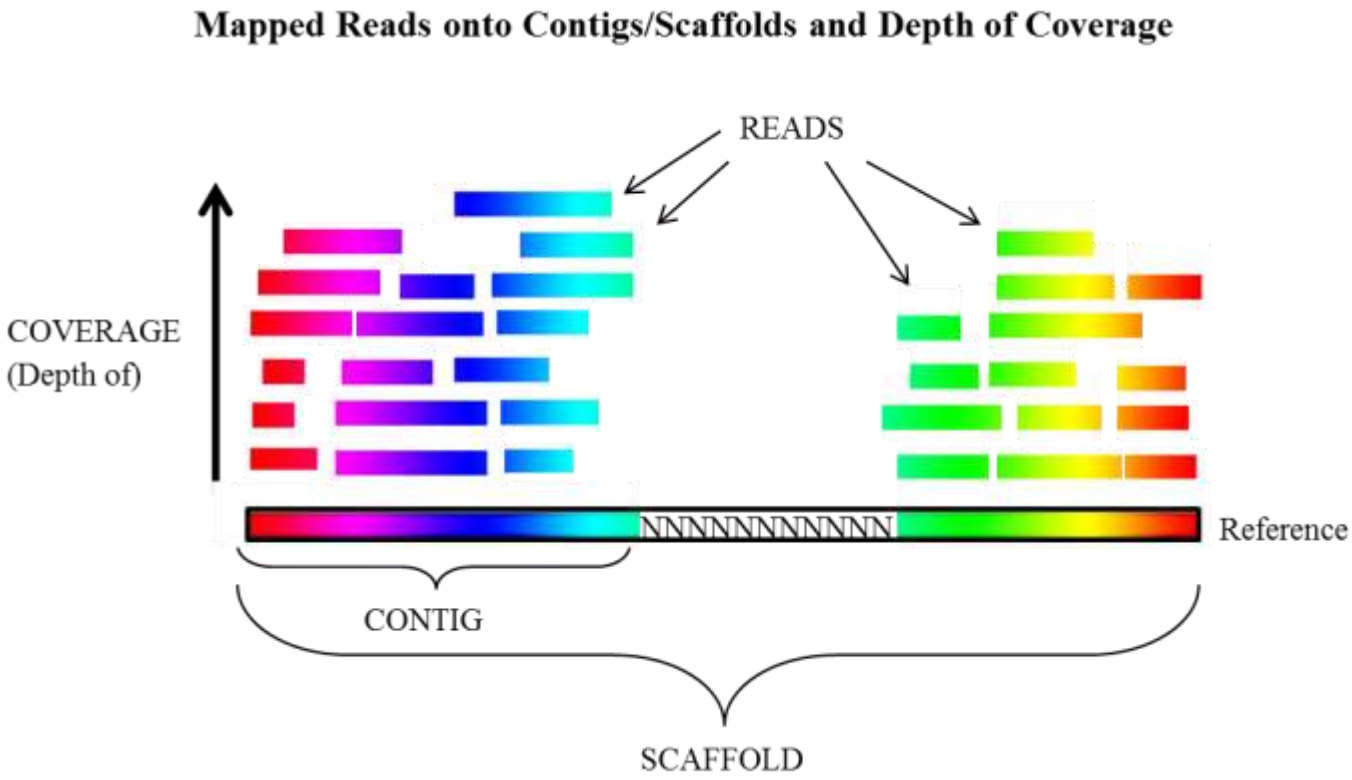

Figure B.1 Graphical glossary representing mapped reads onto a scaffold. See Table B.1 for definitions.

Appendix C Supplementary information for Chapter 3

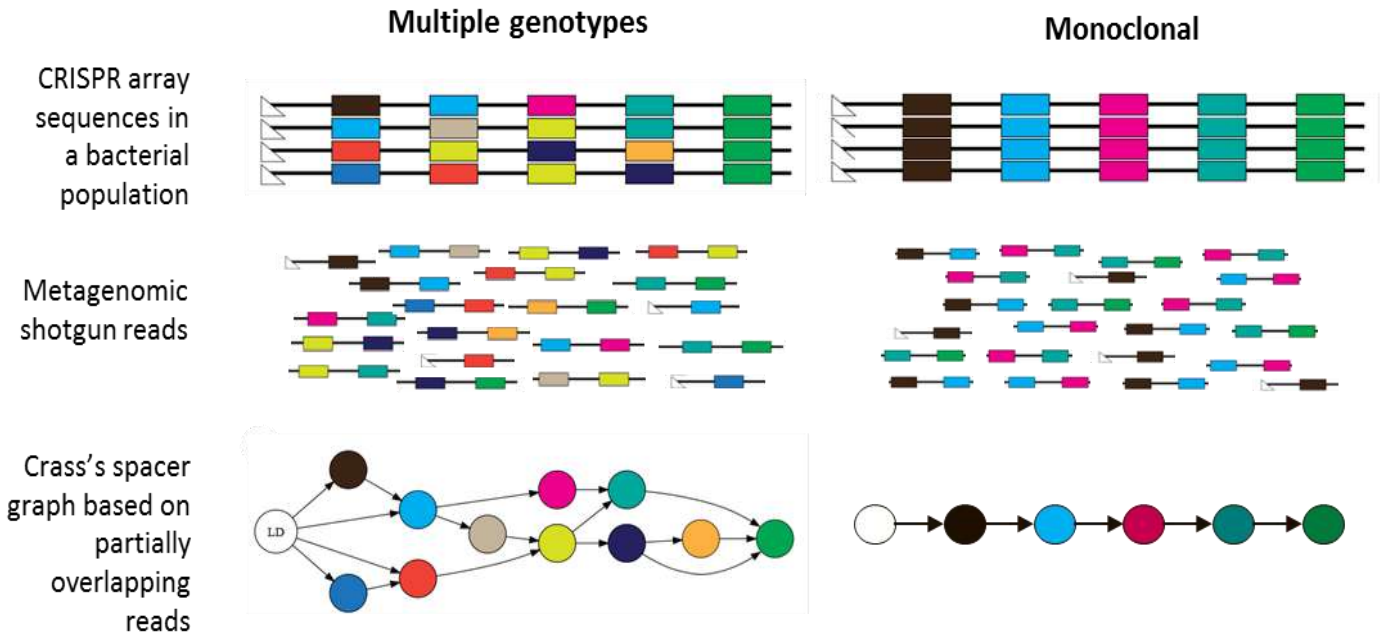

Figure C.1 CRISPR spacer typing with Crass; how to interpret spacer graphs. Crass algorithm works similarly to a genome assembly algorithm. It creates a map using spacers as nodes and finds link between spacers from the partially overlapping reads. Spacer graph that shows branching might indicate chromosomic rearrangement on spacers in the CRISPR array. Adapted from Skennerton *et al.*, 2013.

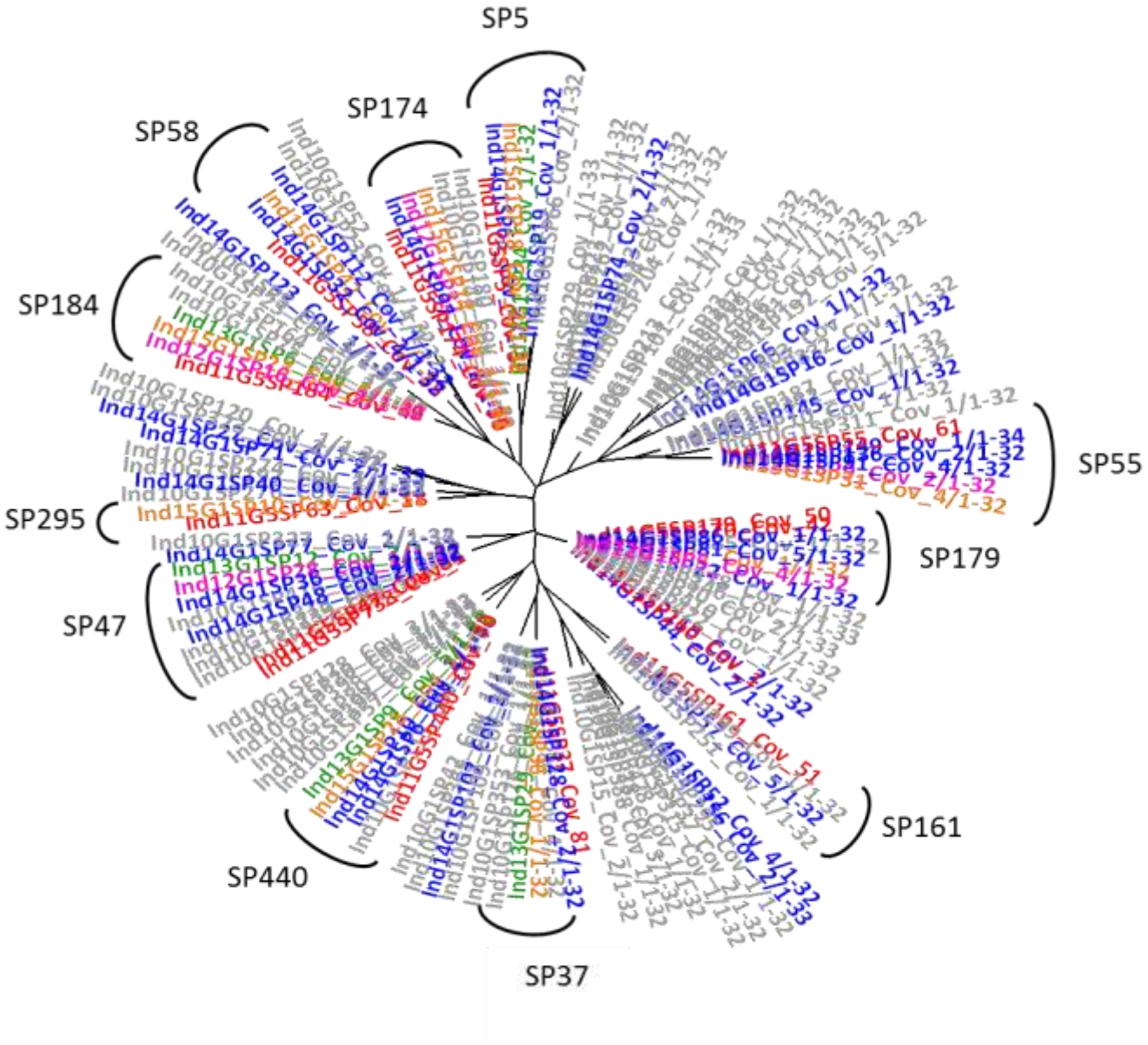

Figure C.2 Neighbor-joining tree based on the CRISPR sequences found in the symbiont metagenomes from 6 individual worms. Red: Ind11 (Symb_1); Grey: Ind10; Pink: Ind12; Green: Ind13; Blue: Ind14; Orange: Ind15. The symbiont metagenomes from Ind10, Ind12, Ind13 Ind14, and Ind15 were pooled together into Symb_pool. Spacers (SP) found in both Symb_1 and Symb_pool are represented outside the clades. See List C.1 for list of spacers.

List C.1 CRISPR spacers found in Symb_1 and Symb_pool

> Spacer 5
GTGGATGATCGCCATCGTCAGCACCACCAGCC
> Spacer 170
ACAAAAAAGCCGAAAGGCAAGCAGTTAATGTT
> Spacer 161
CCGCCAGGAATCGGTAACTATCTCCGGGGTGG
> Spacer 174
AGTACCAGGGCAAAGCCAATCTTGGCAATTGC
> Spacer 179
GACCCAACCCCGTAGCCAGCACGAAACGACCA
> Spacer 37
TTCTAAGCCGTCGGATATTTGCGGCCCGGAAT
> Spacer 47
CACATGGATCATCCGCCAGAACGGAAAGAATG
> Spacer 440
CAATAACGCCTTTGTGGGCGCGGAATATTGAC
> Spacer 184
ATCGCACCTATGTCCGTCGACTGGCCGAGTGC
> Spacer 55
TTACGGCGTCAGAACACTGGCGCTCATTGATG
> Spacer 58
GGCGATTATCAAGGGATACGAGGCAGTCCAAA
> Spacer 295
AGTTATTACGCGAGGCAATCAGATCGTAAGT
> Spacer 260
GACCCAACGCCGTAGCCAGCACGAAACGACCA
> Spacer 738
CACATGGATAATCCGCCAGAACGGAAAGAAGG
> Spacer 401
GGCGATTATCAAGGGATATGAGGCAGTCCAAA
> Spacer 379
GTTACACGCTGCTAGCCCCGCGTTTGAAAAG
> Spacer 360
TTTGTCAACCCGAGCTTCAGCACGTCTTTATAC
> Spacer 350
GACCCCGAGCCTGGTGTCGATCTGCTAGATGT
> Spacer 255
GCGATAAATACTACCCTCGTCCGGCGTGCTAT
> Spacer 228
CCCAATCGCAAGCAGCATTGCAATATCAGATA
> Spacer 261
ACGCCGCGCCACGCGTTGCCAGAATAACAAAAAA
> Spacer 181
TACGGCGCACTCTACTGACGAGGATATTCAGG
> Spacer 42
TTTTGTCGCGCTACGGGCCGTCCCATATGGGG
> Spacer 450
ATTTTTATTTATTTTTATCAGCAAAACACAGC
> Spacer 354
GTCTGCACCTGAATTCGCCCCTTTTTGGGTTT
> Spacer 421
ATCGGCTTCAGTGTATACGACACGTGCCCAGG
> Spacer 309
GATCGTCTATCGATTTAGGCTTTTTCTGGAAA
> Spacer 90
ACGCACTCGGTCGATACTACGTAGTAATTGCT
> Spacer 10
AGCAGAAAACGGGAGAAGAGCCGGTCTTTTGG
> Spacer 98
TATTGGAGATAACTCAATCGCACCTCCCATGTG
> Spacer 129
AATAAACAAAAATGGAGCGCGCTCTCCGGTCA
> Spacer 196
CGTATACACTAAAGCCGATTGGATATGAGCGG
> Spacer 224
TGACCCTGTCACGTCACGCTCCGTAATATAGA
> Spacer 274
TTTTCAAGAGGAAATTATCACTCTGCGTTTGA
> Spacer 279
CAAATGGAAAACCATGCCGGAAAAAATGTTCA
> Spacer 369
TTTGAGAAATGGGCCAAAAAGGCTGAGGAGAAG
> Spacer 374
CATTAAACGCTGCTAGCACCGCGTTTGAACAG
> Spacer 122
ATCTCAGGGATGCCAATCTCCGGCGTGCTAAT
> Spacer 34
CCCGCTATCAAGGCATTCAGCCAGGCGGCGTA
> Spacer 31
GCACTAACTATATCGATGGGCAGCGACCGTCCT
> Spacer 429
ATAATTGAGATAGATGGGGAGACAGGCGAGAT
> Spacer 208
AGTAAGTGATGCTCATTTTGTCGTAAAAAAAG
> Spacer 19
CCACCGCTCGCGATCCAGTCCATGCCCAAACA
> Spacer 156
TTACTGAGGTGCAGAAAAAAACATTAGAGCAG
> Spacer 153
CGTCGGGTGCAGCTGATTCCCGTCAAAAACTC
> Spacer 150
TCTAGCCTGCTCCCCCGGTTTCGGGCCTGTCT
> Spacer 14
GACGCCGCTCAGGCTGCGGGTCATGAGGCTGG
> Spacer 80
TCAAACGACAAAAGAAGTGGCGTGTTGTAGAG
> Spacer 46
TCAAGCGCAGCGCCGGTCAGGATCACGGTCCC
>Flanker 1023
TCCGATCGGCCAAAAAGATCGGTAGAACTCAACACGGTGATT
TTTCTTATCTCTATCAAATAGATAGAATAAGA

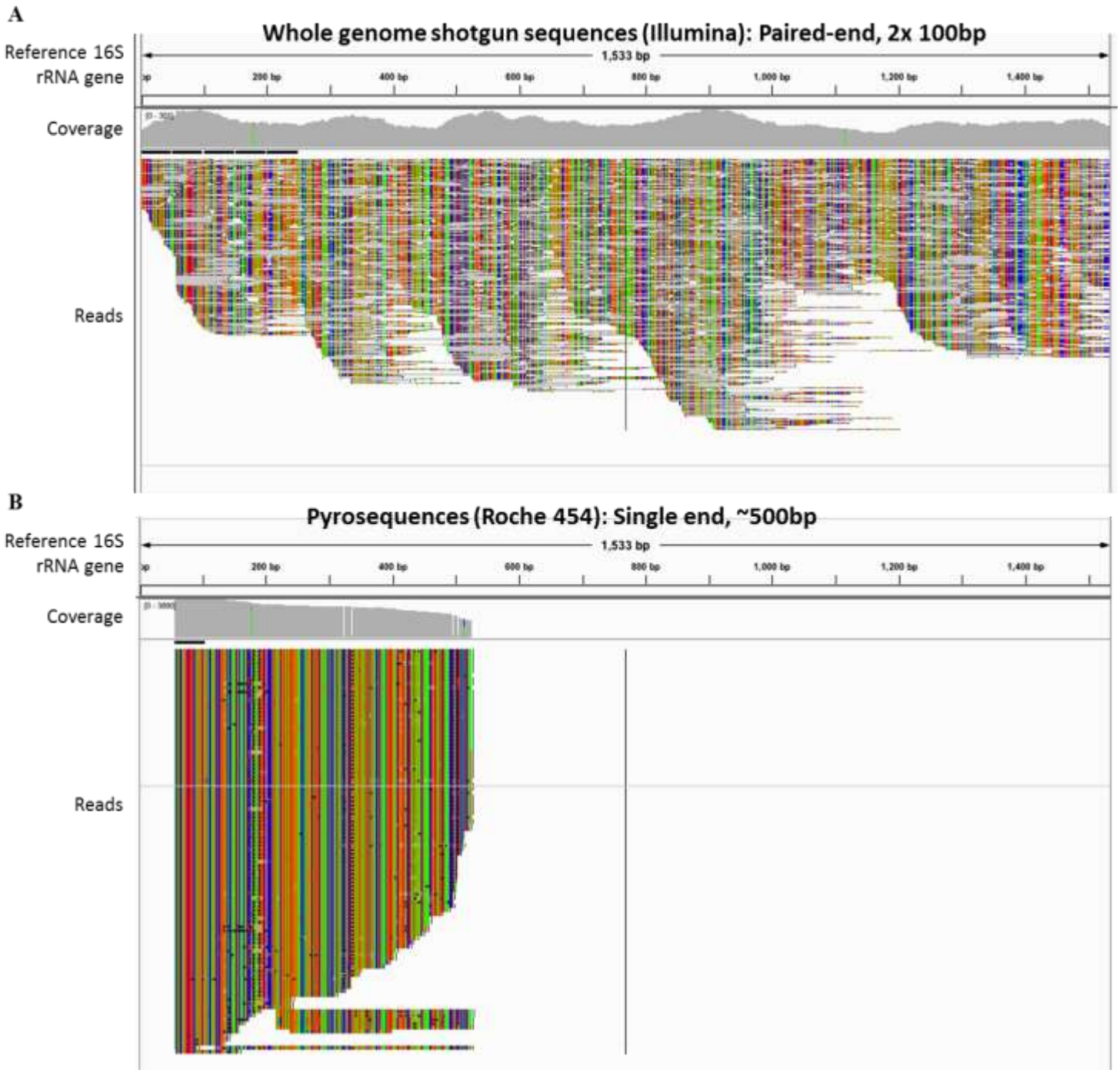

Figure C.3 Whole genome shotgun reads (Illumina technology) vs pyrosequence reads (Roche 454 technology). Illumina reads are a pair of 100bp sequences. Roche/454 reads are a singled end and all start at the same position.

Appendix D Supplementary information for Chapter 4

For shortcuts to the tables, click on the following links:

Back to Appendix D

Table D.1 Accessory genome exclusive to *Ridgeia* symbionts (*Ridgeia* 1 and *Ridgeia* 2 symbiont genome assemblies). Total size = 95 456 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Ridgeia* 1 symbionts	Locus-tag in *Ridgeia* 2 symbionts	Product
Exclusive sequence 1 Ga0074115_146 17457 0 - 17457		Ga0074115_1461	Ga0076813_11512	Curli production assembly/transport component CsgG
		Ga0074115_1462	Ga0076813_11513	Zn-binding Pro-Ala-Ala-Arg (PAAR) domain-containing protein, involved in TypeVI secretion
		Ga0074115_1463	Ga0076813_11514	Methyltransferase domain-containing protein
		Ga0074115_1464	Ga0076813_14753	Protein of unknown function (DUF1795)
		Ga0074115_1465	Ga0076813_14752	Hypothetical protein
		Ga0074115_1466	Ga0076813_14751; Ga0076813_16801	Type VI secretion protein, EvpB/VC_A0108 family/Type VI secretion protein, VC_A0107 family
		Ga0074115_1467	Ga0076813_16802	Type VI secretion system secreted protein VgrG
		Ga0074115_1468	Ga0076813_16803	Hypothetical protein
		Ga0074115_1469	Ga0076813_16804	Type VI secretion system protein ImpA
		Ga0074115_14610	Ga0076813_16805	Type VI secretion system protein ImpB
		Ga0074115_14611	Ga0076813_13723	Type VI secretion system protein ImpC
		Ga0074115_14612	Ga0076813_13722	Type VI secretion system secreted protein Hcp
		Ga0074115_14613	Ga0076813_13721; Ga0076813_14484	ImpE protein
		Ga0074115_14614	Ga0076813_14483	Type VI secretion system protein ImpF
		Ga0074115_14615	Ga0076813_14482	Type VI secretion system protein ImpG
		Ga0074115_14616	Ga0076813_14481	Type VI secretion system protein ImpH
Exclusive sequence 2 Ga0074115_125 11590 33173 - 44763	✓	Ga0074115_12524	Ga0076813_10858	WxcM-like, C-terminal
		Ga0074115_12525	Ga0076813_10859	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase
		Ga0074115_12526	Ga0076813_108510	Hypothetical protein
		Ga0074115_12527	Ga0076813_108511	Hydroxymethylglutaryl-CoA synthase
		Ga0074115_12528	Ga0076813_108512	Acyl carrier protein
		Ga0074115_12529	Ga0076813_108513	dTDP-4-amino-4,6-dideoxygalactose transaminase
		Ga0074115_12530	Ga0076813_108514	Glycosyl transferase family 2
		Ga0074115_12531	Ga0076813_108515	Lipopolysaccharide transport system permease protein
		Ga0074115_12532	Ga0076813_108516	ABC-2 type transport system ATP-binding protein
		Ga0074115_12533	Ga0076813_108517	Glycosyl transferases group 1
		Ga0074115_12534	Ga0076813_108518	Hypothetical protein
		Ga0074115_12535	Ga0076813_108519	Glycosyltransferase involved in cell wall bisynthesis
		Ga0074115_12536	Ga0076813_108520	Hypothetical protein
Exclusive sequence 3 Ga0074115_157 8235 0 - 8235		Ga0074115_1571	Ga0076813_10168	Hypothetical protein
		Ga0074115_1572	Ga0076813_10167	Hypothetical protein
		Ga0074115_1573	Ga0076813_10166	Hypothetical protein
		Ga0074115_1574	Ga0076813_10165	Hypothetical protein
		Ga0074115_1575	Ga0076813_10164	TerY-C metal binding domain
		Ga0074115_1576	Ga0076813_10163	Hypothetical protein
		Ga0074115_1577	Ga0076813_10162	Mobile mystery protein A
		Ga0074115_1578	Ga0076813_10162; Ga0076813_16633	Mobile mystery protein B
		Ga0074115_1579	Ga0076813_16632	Fic/DOC family protein
		Ga0074115_15710	Ga0076813_14733;	Hypothetical protein

Back to Appendix D

Table D.1 Accessory genome exclusive to *Ridgeia* symbionts (*Ridgeia* 1 and *Ridgeia* 2 symbiont genome assemblies). Total size = 95 456 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Ridgeia* 1 symbionts	Locus-tag in *Ridgeia* 2 symbionts	Product
			Ga0076813_16631	
		Ga0074115_15711	Ga0076813_14732	Predicted nuclease of the RNAse H fold, HicB family
		Ga0074115_15712	Ga0076813_14731	Protein of unknown function (DUF3732)
		Ga0074115_15713	n.a.	Hypothetical protein
Exclusive sequence 4		n.a.	Ga0076813_142210	Transposase
Ga0074115_147		Ga0074115_1471	Ga0076813_14229	Protein of unknown function DUF91
8311		Ga0074115_1472	Ga0076813_14228	Hypothetical protein
0 - 8311		Ga0074115_1473	Ga0076813_14227	Relaxase/Mobilisation nuclease domain-containing protein
		Ga0074115_1474	Ga0076813_14226	Hypothetical protein
		Ga0074115_1475	Ga0076813_14225	Hypothetical protein
		Ga0074115_1476	Ga0076813_14224	Hypothetical protein
		Ga0074115_1477	Ga0076813_14223	Hypothetical protein
		Ga0074115_1478	Ga0076813_14222	tRNA-splicing ligase RtcB
Exclusive sequence 5	✓	Ga0074115_105110	Ga0076813_13394	Hypothetical protein
Ga0074115_105		Ga0074115_105111	Ga0076813_11781;	Methyltransferase domain-containing protein
7528			Ga0076813_14331;	
121389 - 128917			Ga0076813_13762	
		Ga0074115_105112	Ga0076813_13761;	Hypothetical protein
			Ga0076813_12832	
		Ga0074115_105113	Ga0076813_12831	Hypothetical protein
Exclusive sequence 6		Ga0074115_13218	Ga0076813_14266	Carbohydrate binding domain-containing protein
Ga0074115_132		Ga0074115_13219	Ga0076813_14265	Hypothetical protein
7011		Ga0074115_13220	Ga0076813_14264	Hypothetical protein
25125 - 32136		Ga0074115_13221	Ga0076813_14263	GntR family transcriptional regulator
		Ga0074115_13222	Ga0076813_14262	Tripartite-type tricarboxylate transporter, receptor component TctC
		Ga0074115_13223	Ga0076813_14261;	Tripartite tricarboxylate transporter TctB family protein
			Ga0076813_13331	
Exclusive sequence 7	✓	Ga0074115_1094	Ga0076813_16722	CRISPR-associated protein, Csd2 family
Ga0074115_109		Ga0074115_1095	Ga0076813_16721;	CRISPR-associated protein Csd1
4694			Ga0076813_10881;	
1942 - 6636			Ga0076813_14464	
		Ga0074115_1096	Ga0076813_14463	CRISPR-associated protein, Cas5d family
		Ga0074115_1098	Ga0076813_14462	Hypothetical protein
		Ga0074115_1097	Ga0076813_14461	Hypothetical protein
Exclusive sequence 8		Ga0074115_1411;	Ga0076813_10531	Hypothetical protein
Ga0074115_141		Ga0074115_1412		
4127		Ga0074115_1413	Ga0076813_10532	Hypothetical protein
0 - 4127				
Exclusive sequence 9	✓	Ga0074115_103137	Ga0076813_16812	Type I secretion system ABC transporter, HlyB family
Ga0074115_103		Ga0074115_103138	Ga0076813_16811;	Hemolysin D
2816			Ga0076813_14381	
141098 - 143914		Ga0074115_103139	Ga0076813_14382	Hypothetical protein
Exclusive sequence 10		Ga0074115_11438	Ga0076813_127316	Hypothetical protein

Back to Appendix D

Table D.1 Accessory genome exclusive to *Ridgeia* symbionts (*Ridgeia* 1 and *Ridgeia* 2 symbiont genome assemblies). Total size = 95 456 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Ridgeia* 1 symbionts	Locus-tag in *Ridgeia* 2 symbionts	Product
Ga0074115_114 2482 43812 - 46294		Ga0074115_11439 n.a. Ga0074115_11440 Ga0074115_11441 Ga0074115_11442 Ga0074115_11443	Ga0076813_127315 Ga0076813_127314 Ga0076813_127313 Ga0076813_127312 Ga0076813_127310 Ga0076813_12739	Hypothetical protein Hypothetical protein Integrase core domain-containing protein HTH-like domain-containing protein Hypothetical protein Transposase
Exclusive sequence 11 Ga0074115_170 2271 0 - 2271		Ga0074115_1701 Ga0074115_1702 Ga0074115_1703 Ga0074115_1704	n.a. Ga0076813_13551 Ga0076813_13552 n.a.	Cation diffusion facilitator family transporter AcrB/AcrD/AcrF family protein AcrB/AcrD/AcrF family protein Hypothetical protein
Exclusive sequence 12 Ga0074115_132 2223 3349 – 5572		Ga0074115_1641	Ga0076813_14711	PKD repeat-containing protein
Exclusive sequence 13 Ga0074115_161 2011 3372 – 5383	✓	Ga0074115_1614 Ga0074115_1615	Ga0076813_13865 Ga0076813_13866	Transposase Transposase IS66 family protein
Exclusive sequence 14 Ga0074115_159 1926 0 – 1926		Ga0074115_1591	Ga0076813_14181	Secreted protein containing bacterial Ig-like domain and vWFA domain
Exclusive sequence 15 Ga0074115_123 1831 335 – 2166	✓	Ga0074115_1232 Ga0074115_1233	Ga0076813_14196 Ga0076813_14195	TIGR02452 family protein Hypothetical protein
Exclusive sequence 16 Ga0074115_140 1736 0 – 1736		n.a. Ga0074115_10577 Ga0074115_10578	Ga0076813_14051 Ga0076813_14053 Ga0076813_14052	Transcriptional regulatory protein, C terminal Carbon-nitrogen hydrolase ubiE/COQ5 methyltransferase family protein
Exclusive sequence 17 Ga0074115_132 1726 0 - 1726		n.a. Ga0074115_1321	Ga0076813_15291 Ga0076813_15292	Transposase Hypothetical protein
Exclusive sequence 18 Ga0074115_180 1396 0 – 1396		Ga0074115_1801 Ga0074115_1802	Ga0076813_11412 Ga0076813_11411	Spherulation-specific family 4 Hypothetical protein
Exclusive sequence 19 Ga0074115_102 1154 129891 – 131045	✓	Ga0074115_102124 Ga0074115_102125	Ga0076813_169035 Ga0076813_169034	Putative transposase Putative transposase

Back to Appendix D

Table D.1 Accessory genome exclusive to *Ridgeia* symbionts (*Ridgeia* 1 and *Ridgeia* 2 symbiont genome assemblies). Total size = 95 456 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Ridgeia* 1 symbionts	Locus-tag in *Ridgeia* 2 symbionts	Product
Exclusive sequence 20 Ga0074115_165 903 2388 – 3291		Ga0074115_1654	Ga0076813_10561	Hypothetical protein
Exclusive sequence 21 Ga0074115_126 856 11285 – 12141		Ga0074115_1262 Ga0074115_1263	Ga0076813_11072 Ga0076813_11402	Hypothetical protein Hypothetical protein
Exclusive sequence 22 Ga0074115_143 852 15512 – 16364		Ga0074115_14312 Ga0074115_14313	Ga0076813_15414 Ga0076813_15415	Nitrate/nitrite transport system substrate-binding protein Hypothetical protein
Exclusive sequence 23 Ga0074115_156 628 6057 – 6685	✓	Ga0074115_1567	Ga0076813_13514	Aminoacyl-tRNA editing domain-containing protein
Exclusive sequence 24 Ga0074115_156 591 4884 – 5475	✓	Ga0074115_1566	Ga0076813_13515	Hypothetical protein
Exclusive sequence 25 Ga0074115_118 586 61784 – 62370		Ga0074115_11853	Ga0076813_16776	Chemoreceptor zinc-binding domain-containing protein
Exclusive sequence 26 Ga0074115_112 515 0 - 515		Ga0074115_1121	Ga0076813_16143	Hypothetical protein

Back to Appendix D

Table D.2 Accessory genome exclusive to the East Pacific Rise symbionts (*Tevnia*, *Rifita* 1, and *Rifita* 2 symbiont genome assemblies). Total size = 141 335 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig Sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Tevnia* symbionts	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
Exclusive sequence 1 AFZB01000005 33791 69865 - 103656		TevJSym_ae00610; TevJSym_ae00620	Rifp1Sym_dq00030	Rifp2Sym_fc00010	Anaerobic dimethyl sulfoxide reductase subunit A
		n.a.	n.a.	Rifp2Sym_fc00020	Hypothetical protein
		TevJSym_ae00630	n.a.	Rifp2Sym_fc00030	Hypothetical protein
		TevJSym_ae00640	Rifp1Sym_dq00020	Rifp2Sym_fc00040	Transcriptional regulator, ArsR family
		TevJSym_ae00650	Rifp1Sym_gl00020	Rifp2Sym_fc00050	Fatty acid hydroxylase
		TevJSym_ae00660	Rifp1Sym_gl00040	Rifp2Sym_fc00060	Uncharacterized membrane protein YdjX, TVP38/TMEM64 family, SNARE-associated domain
		TevJSym_ae00670	Rifp1Sym_gl00060	Rifp2Sym_fc00070	Hypothetical protein
		TevJSym_ae00680	Rifp1Sym_gl00080	Rifp2Sym_fc00080	Uncharacterized protein involved in oxidation of intracellular sulfur
		TevJSym_ae00690	n.a.	n.a.	Adenylate cyclase
		TevJSym_ae00700	n.a.	Rifp2Sym_cv00010	Cyclic nucleotide binding protein
		TevJSym_ae00710	Rifp1Sym_ew00090	Rifp2Sym_cv00020	Hypothetical protein
		TevJSym_ae00720	Rifp1Sym_ew00080	Rifp2Sym_cv00030	Glyoxylase, beta-lactamase superfamily II
		TevJSym_ae00730	Rifp1Sym_ew00070; Rifp1Sym_ew00060; Rifp1Sym_ew00050	Rifp2Sym_cv00040	NADPH-dependent glutamate synthase beta chain
		TevJSym_ae00740	Rifp1Sym_ew00040	Rifp2Sym_cv00050	Pyridine nucleotide-disulphide oxidoreductase
		TevJSym_ae00750	Rifp1Sym_ew00030	Rifp2Sym_cv00060	Hypothetical protein
		TevJSym_ae00760	Rifp1Sym_ew00020	Rifp2Sym_cv00070; Rifp2Sym_cv00080	Hydroxylamine dehydrogenase
		n.a.	Rifp1Sym_ew00010	n.a.	Hypothetical protein
		n.a.	n.a.	Rifp2Sym_cv00090	Hypothetical protein
		TevJSym_ae00770	n.a.	Rifp2Sym_cv00100	Hypothetical protein
		TevJSym_ae00780	n.a.	n.a.	Hypothetical protein
		TevJSym_ae00790	Rifp1Sym_bf00360	Rifp2Sym_cv00110	D-alanyl-lipoteichoic acid acyltransferase DltB, MBOAT superfamily
		TevJSym_ae00800	Rifp1Sym_bf00350	Rifp2Sym_cv00120	Hypothetical protein

Back to Appendix D

Table D.2 Accessory genome exclusive to the East Pacific Rise symbionts (*Tevnia*, *Rifita* 1, and *Rifita* 2 symbiont genome assemblies). Total size = 141 335 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig Sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Tevnia* symbionts	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
		TevJSym_ae00810	Rifp1Sym_bf00340	Rifp2Sym_cv00130	DNA-binding transcriptional response regulator, NtrC family, contains REC, AAA-type ATPase, and a Fis-type DNA-binding domains
		TevJSym_ae00820	Rifp1Sym_bf00330	Rifp2Sym_cv00140	Signal transduction histidine kinase
		TevJSym_ae00830	Rifp1Sym_bf00320	Rifp2Sym_cv00150	ParE-like toxin of type II toxin-antitoxin system
		TevJSym_ae00840	n.a.	Rifp2Sym_cv00160	ParD-like antitoxin of type II toxin-antitoxin system
		TevJSym_ae00850	Rifp1Sym_bf00310	Rifp2Sym_cv00170	Cytochrome b
		TevJSym_ae00860	Rifp1Sym_bf00300	Rifp2Sym_cv00180	Protein of unknown function (DUF1924)
		TevJSym_ae00870	Rifp1Sym_bf00290	Rifp2Sym_cv00190	Dihaem cytochrome c
		TevJSym_ae00880	Rifp1Sym_bf00280	Rifp2Sym_bo00320	His Kinase A (phospho-acceptor) domain-containing protein
		TevJSym_ae00890	Rifp1Sym_bf00270	Rifp2Sym_bo00310	Two-component system, NtrC family, response regulator AtoC
		TevJSym_ae00900	Rifp1Sym_bf00260	Rifp2Sym_bo00300	Hypothetical protein
		TevJSym_ae00910	Rifp1Sym_bf00250	Rifp2Sym_bo00290	Major Facilitator Superfamily protein
		TevJSym_ae00920	n.a.	Rifp2Sym_bo00280	Cytochrome C'
		TevJSym_ae00930	Rifp1Sym_bf00240	Rifp2Sym_bo00270	Glyceraldehyde-3-phosphate dehydrogenase 3
Exclusive sequence 2 AFZB01000004 24661 114255 - 138916	✓	TevJSym_ad01110	Rifp1Sym_dc00020	n.a.	
		TevJSym_ad01120	Rifp1Sym_cn00190	Rifp2Sym_az00350; Rifp2Sym_az00340; Rifp2Sym_az00330; Rifp2Sym_az00320; Rifp2Sym_az00310	ATP-dependent helicase HrpA
		TevJSym_ad01130	Rifp1Sym_cn00180	Rifp2Sym_az00300	SEC-C motif-containing protein
		TevJSym_ad01140	Rifp1Sym_cn00170	Rifp2Sym_az00290	Protein of unknown function (DUF4124)
		TevJSym_ad01150	Rifp1Sym_cn00160	Rifp2Sym_az00280	Glycosyltransferase involved in cell wall bisynthesis
		TevJSym_ad01160	Rifp1Sym_cn00150	Rifp2Sym_az00270	PilZ domain-containing protein
		TevJSym_ad01170	Rifp1Sym_cn00140	Rifp2Sym_az00260	Protein of unknown function (DUF1631)

Back to Appendix D

Table D.2 Accessory genome exclusive to the East Pacific Rise symbionts (*Tevnia*, *Rifita* 1, and *Rifita* 2 symbiont genome assemblies). Total size = 141 335 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig Sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Tevnia* symbionts	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
		TevJSym_ad01180	Rifp1Sym_cn00130	Rifp2Sym_az00250	Hypothetical protein
		TevJSym_ad01190	Rifp1Sym_cn00120	Rifp2Sym_az00240	HD-like signal output (HDOD) domain, no enzymatic activity
		TevJSym_ad01200	Rifp1Sym_cn00110	Rifp2Sym_az00230	Sulfate adenylyltransferase
		TevJSym_ad01210	Rifp1Sym_cn00100	Rifp2Sym_az00220	Aconitase
		TevJSym_ad01220	Rifp1Sym_cn00090	Rifp2Sym_az00210	Hypothetical protein
		TevJSym_ad01230	Rifp1Sym_cn00080	Rifp2Sym_az00180; Rifp2Sym_az00190; Rifp2Sym_az00200	Short chain dehydrogenase reductase Sdr
		n.a.	Rifp1Sym_cn00070	Rifp2Sym_az00170	Integral membrane protein
		n.a.	Rifp1Sym_cn00060	Rifp2Sym_az00160	Integral membrane protein
		TevJSym_ad01240	Rifp1Sym_cn00050	Rifp2Sym_az00150	Putative membrane protein
Exclusive sequence 3 AFZB01000005 22442 0 - 22442		TevJSym_ae00010	Rifp1Sym_el00010	Rifp2Sym_et00150	Proteic killer suppression protein
		TevJSym_ae00020	Rifp1Sym_el00020	Rifp2Sym_et00140	Addiction module antidote protein, HigA family
		TevJSym_ae00030	Rifp1Sym_el00030; Rifp1Sym_el00040; Rifp1Sym_el00050	Rifp2Sym_et00130; Rifp2Sym_et00120	Toprim domain-containing protein
		TevJSym_ae00040	Rifp1Sym_el00060	Rifp2Sym_et00110	Hypothetical protein
		n.a.	Rifp1Sym_el00070	Rifp2Sym_et00100	Spermidine synthase
		TevJSym_ae00050	Rifp1Sym_el00080	Rifp2Sym_et00090	Phage integrase family protein
		TevJSym_ae00060	Rifp1Sym_el00090	Rifp2Sym_et00080	Hypothetical protein
		TevJSym_ae00070	Rifp1Sym_el00100	Rifp2Sym_et00070	Hypothetical protein
		TevJSym_ae00080; TevJSym_ae00090	Rifp1Sym_el00110	Rifp2Sym_et00060	Virulence-associated protein I
		TevJSym_ae00100	Rifp1Sym_el00120	Rifp2Sym_et00050	Hypothetical protein
		TevJSym_ae00110	Rifp1Sym_el00130	Rifp2Sym_et00040	Bacteriophage phi gp55-like protein
		TevJSym_ae00120	Rifp1Sym_el00140	Rifp2Sym_et00030	Hypothetical protein
		TevJSym_ae00130	Rifp1Sym_el00150	Rifp2Sym_et00020	Hypothetical protein
		TevJSym_ae00140	Rifp1Sym_el00160	Rifp2Sym_et00010	Hypothetical protein
		TevJSym_ae00150	Rifp1Sym_dx00010	n.a.	Hypothetical protein
		TevJSym_ae00160	Rifp1Sym_dx00020	Rifp2Sym_ee00010	Hypothetical protein
		TevJSym_ae00170	Rifp1Sym_dx00030	n.a.	Hypothetical protein

Back to Appendix D

Table D.2 Accessory genome exclusive to the East Pacific Rise symbionts (*Tevnia*, *Rifita* 1, and *Rifita* 2 symbiont genome assemblies). Total size = 141 335 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig Sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Tevnia* symbionts	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
		TevJSym_ae00180	Rifp1Sym_dx00040; Rifp1Sym_dx00050; Rifp1Sym_dx00060	Rifp2Sym_ee00020; Rifp2Sym_ee00030	Hypothetical protein
		TevJSym_ae00190	Rifp1Sym_dx00070	Rifp2Sym_ee00040	Hypothetical protein
		TevJSym_ae00200	Rifp1Sym_dx00080; Rifp1Sym_dx00090; Rifp1Sym_dx00100	Rifp2Sym_ee00050; Rifp2Sym_ee00060; Rifp2Sym_ee00070	Phage tape measure protein
		n.a.	n.a.	Rifp2Sym_ee00080	Hypothetical protein
		n.a.	Rifp1Sym_dx00110	n.a.	Hypothetical protein
		TevJSym_ae00210	Rifp1Sym_dx00120	Rifp2Sym_ee00090	Hypothetical protein
		TevJSym_ae00220	Rifp1Sym_dx00130	Rifp2Sym_ee00100	Hypothetical protein
		TevJSym_ae00230	Rifp1Sym_dx00140	Rifp2Sym_ee00110	Hypothetical protein
Exclusive sequence 4 AFZB01000004 9533 17066 - 26599	✓	TevJSym_ad00160	Rifp1Sym_co00130	Rifp2Sym_cu00010	Hypothetical protein
		TevJSym_ad00170	Rifp1Sym_co00140	Rifp2Sym_fm00020	Protein of unknown function (DUF323)
		TevJSym_ad00180	Rifp1Sym_co00150; Rifp1Sym_co00160	Rifp2Sym_fm00010	Sulfatase-modifying factor enzyme 1
		TevJSym_ad00190	Rifp1Sym_co00160	n.a.	Hypothetical protein
		TevJSym_ad00200	Rifp1Sym_co00170	Rifp2Sym_fu00090; Rifp2Sym_fu00080	Protein phosphatase
		TevJSym_ad00210	Rifp1Sym_co00180	Rifp2Sym_fu00070	Protein phosphatase ImpM
Exclusive sequence 5 AFZB01000008 7596 63197 - 70793		n.a.	Rifp1Sym_ev00020	n.a.	Aspartate aminotransferase
		TevJSym_ah00640	Rifp1Sym_ev00030	Rifp2Sym_ii00010	N-6 DNA Methylase
		TevJSym_ah00650	Rifp1Sym_ev00040	Rifp2Sym_ii00020; Rifp2Sym_ii00030; Rifp2Sym_mm00020; Rifp2Sym_mm00010	Type I restriction enzyme M protein
		TevJSym_ah00660	Rifp1Sym_ev00050	n.a.	DEAD/DEAH box helicase
Exclusive sequence 6 AFZB01000004 7306 0 - 7306	✓	TevJSym_ad00010	Rifp1Sym_ez00030	Rifp2Sym_cu00190	Putative glycosyltransferase
		TevJSym_ad00020	Rifp1Sym_ez00020	Rifp2Sym_cu00180	Asparagine synthase (glutamine-hydrolysing)
		TevJSym_ad00030	Rifp1Sym_ez00010	Rifp2Sym_cu00170	Putative glycosyltransferase
		TevJSym_ad00040	n.a.	Rifp2Sym_cu00160	Glycosyltransferase, family 2
Exclusive sequence 7 AFZB01000038	✓	TevJSym_bl00040	Rifp1Sym_ep00040	Rifp2Sym_gt00010	CRISPR-associated protein Cas7/Csd2, subtype I-C/DVULG

Back to Appendix D

Table D.2 Accessory genome exclusive to the East Pacific Rise symbionts (*Tevnia*, *Rifita* 1, and *Rifita* 2 symbiont genome assemblies). Total size = 141 335 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig Sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Tevnia* symbionts	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
6555 2424 - 8979		TevJSym_bl00050	Rifp1Sym_ep00050	Rifp2Sym_gt00020	CRISPR-associated protein Csd1
		TevJSym_bl00060	Rifp1Sym_ep00060	Rifp2Sym_gt00030	CRISPR-associated protein Cas5d
		TevJSym_bl00070	Rifp1Sym_ep00070	Rifp2Sym_gt00040	CRISPR-associated endonuclease/helicase Cas3
Exclusive sequence 8 AFZB01000002 5745 158129 - 163874	✓	TevJSym_ab01580; TevJSym_ab01590	Rifp1Sym_bb00150	Rifp2Sym_ga00040	Putative transmembrane protein
		TevJSym_ab01600	Rifp1Sym_bb00140	Rifp2Sym_dh00160	Endoglucanase A
		TevJSym_ab01610	Rifp1Sym_bb00130	Rifp2Sym_dh00150	Putative transcriptional regulator
		n.a.	n.a.	Rifp2Sym_dh00140	Hypothetical protein
		TevJSym_ab01620	Rifp1Sym_bb00120	Rifp2Sym_dh00130	Hypothetical protein
		TevJSym_ab01630	Rifp1Sym_bb00110	Rifp2Sym_dh00120	Transposase insF for insertion sequence IS3
		TevJSym_ab01640	Rifp1Sym_bb00100	Rifp2Sym_dh00110	Integrase catalytic region
Exclusive sequence 9 AFZB01000037 3882 17256 - 21138	✓	TevJSym_bk00210	Rifp1Sym_cv00080	Rifp2Sym_bj00120	Adenosylhomocysteinase
		n.a.	n.a.	Rifp2Sym_bj00110	Hypothetical protein
		TevJSym_bk00220	Rifp1Sym_cv00090	Rifp2Sym_bj00100	Cobalt-zinc-cadmium efflux system protein
Exclusive sequence 10 AFZB01000045 3514 12768 - 16282	✓	TevJSym_bs00140	Rifp1Sym_gd00030	Rifp2Sym_da00120	Lipopolysaccharide transport system permease protein
		TevJSym_bs00150	Rifp1Sym_gd00020	Rifp2Sym_da00130	Lipopolysaccharide transport system ATP-binding protein
Exclusive sequence 11 AFZB01000064 2784 0 - 2784		TevJSym_cl00010	n.a.	Rifp2Sym_bz00010	Sodium/proton antiporter, NhaA family
		TevJSym_cl00020	Rifp1Sym_gw00010	Rifp2Sym_bz00020	Hypothetical protein
Exclusive sequence 12 AFZB01000018 2422 32751 - 35173		TevJSym_ar00310	Rifp1Sym_bh00100	n.a.	Hypothetical protein
		TevJSym_ar00320	n.a.	Rifp2Sym_ab00280	Hypothetical protein
Exclusive sequence 13 AFZB01000005 1918 119131 - 121049		TevJSym_ae01170	Rifp1Sym_bf00030	Rifp2Sym_bo00030	Hypothetical protein

Back to Appendix D

Table D.2 Accessory genome exclusive to the East Pacific Rise symbionts (*Tevnia*, *Rifita* 1, and *Rifita* 2 symbiont genome assemblies). Total size = 141 335 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig Sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Tevnia* symbionts	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
Exclusive sequence 13 AFZB01000005 1918 119131 - 121049		TevJSym_ae01180; TevJSym_ae01190	Rifp1Sym_bf00020	Rifp2Sym_bo00020	Cytochrome C, class I
Exclusive sequence 14 AFZB01000005 1748 63277 - 65025		TevJSym_ae00560	Rifp1Sym_dq00090	Rifp2Sym_ep00040	Transcriptional regulatory protewin ZraR
Exclusive sequence 15 AFZB01000077 1739 0 - 1739		TevJSym_cx00010	Rifp1Sym_db00080	Rifp2Sym_jp00010	Transposase
Exclusive sequence 16 AFZB01000033 1378 27912 - 29290		TevJSym_bg00220	Rifp1Sym_bk00090	Rifp2Sym_bh00120	KAP P-loop domain containing protein
		TevJSym_bg00230	Rifp1Sym_bk00080	Rifp2Sym_bh00110	KAP family P-loop domain-containing protein
		TevJSym_bg00240	Rifp1Sym_bk00070	Rifp2Sym_bh00100	Hypothetical protein
Exclusive sequence 17 AFZB01000037 1027 10832 - 11859	✓	TevJSym_bk00130	Rifp1Sym_cv00020	Rifp2Sym_bj00190	Integrase core domain-containing protein
Exclusive sequence 18 AFZB01000016 984 79511 - 80495		TevJSym_ap00810	Rifp1Sym_fn00070	Rifp2Sym_fv00020	Putative transposase

Back to Appendix D

Table D.2 Accessory genome exclusive to the East Pacific Rise symbionts (*Tevnia*, *Rifita* 1, and *Rifita* 2 symbiont genome assemblies). Total size = 141 335 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence Reference contig Sequence length Start - End positions (in reference contig)	Low flanking conservation	Locus-tag in *Tevnia* symbionts	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
Exclusive sequence 19 AFZB01000029 392 3174 - 3566		TevJSym_bc00050	Rifp1Sym_al00320	Rifp2Sym_aw00250	Transposase DDE domain-containing protein

Back to Appendix D

Table D.3 Accessory genome found in *Riftia* symbionts (*Rifita* 1 and *Rifita* 2 symbiont genome assemblies). Total size = 77 303 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called; n.d. = Not Detected; sequence absent from the assembly. **(Continued)**

Exclusive sequence	Reference contig	Sequence length	Start - End (in reference contig)	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
Exclusive sequence 1	AFOC01000021	39573	0 - 39573	Rifp1Sym_au00010	Rifp2Sym_gm00020	Type II restriction/modification system, DNA methylase subunit YeeA
				Rifp1Sym_au00020	Rifp2Sym_gm00010	Protein of unknown function (DUF4263)
				Rifp1Sym_au00030	Rifp2Sym_hf00010	TIGR02687 family protein
				Rifp1Sym_au00040	Rifp2Sym_hf00020	Transposase IS200 like
				Rifp1Sym_au00050	n.a.	Transposase IS200 like
				Rifp1Sym_au00060	Rifp2Sym_hf00030; Rifp2Sym_hv00040	ATP-dependent Lon protease
				Rifp1Sym_au00070	Rifp2Sym_hv00020	Hypothetical protein
				Rifp1Sym_au00080	n.a.	Uncharacterized protein YydD, contains DUF2326 domain
				Rifp1Sym_au00090	n.a.	HicB family protein
				Rifp1Sym_au00100	n.a.	Hypothetical protein
				Rifp1Sym_au00110	n.a.	Conjugal transfer mating pair stabilization protein TraG
				Rifp1Sym_au00120	n.a.	Conjugative transfer pilus assembly protein TraH
				Rifp1Sym_au00130	n.a.	Conjugal transfer pilus assembly protein TraF
				Rifp1Sym_au00140	n.a.	Type-F conjugative transfer system mating-pair stabilization protein TraN
				Rifp1Sym_au00150	n.a.	Conjugal transfer pilus assembly protein TrbC
				Rifp1Sym_au00160	n.a.	Conjugal transfer pilus assembly protein TraU
				Rifp1Sym_au00170	n.a.	Patatin-like phospholipase
				Rifp1Sym_au00180	n.a.	Conjugal transfer pilin signal peptidase TrbI
				Rifp1Sym_au00190	n.a.	Type-F conjugative transfer system protein (TrbI_Ftype)
				Rifp1Sym_au00200	n.a.	Conjugal transfer ATP-binding protein TraC
				Rifp1Sym_au00210	n.a.	Type IV conjugative transfer system lipoprotein TraV
				Rifp1Sym_au00220	n.a.	Thiol:disulfide interchange protein DsbC
				Rifp1Sym_au00230	n.a.	Conjugal transfer pilus assembly protein TraB
				Rifp1Sym_au00240	n.a.	Conjugal transfer pilus assembly protein TraK
				Rifp1Sym_au00250	n.a.	Conjugal transfer pilus assembly protein TraE

D.13

Back to Appendix D

Table D.3 Accessory genome found in *Riftia* symbionts (*Rifita* 1 and *Rifita* 2 symbiont genome assemblies). Total size = 77 303 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called; n.d. = Not Detected; sequence absent from the assembly. **(Continued)**

Exclusive sequence	Reference contig	Sequence length	Start - End (in reference contig)	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
				Rifp1Sym_au00260	Rifp2Sym_ew00060	Conjugal transfer pilus assembly protein TraA
				Rifp1Sym_au00270	Rifp2Sym_ew00050	Outer membrane protein OmpA
	AFOC01000086	15845	0 - 15845	Rifp1Sym_dg00010	Rifp2Sym_ew00020	ABC-2 type transporter
				Rifp1Sym_dg00020	Rifp2Sym_ew00010	Hypothetical protein
				Rifp1Sym_dg00030	Rifp2Sym_kv00030	Putative cysteine desulfurase
				Rifp1Sym_dg00040	Rifp2Sym_kv00020	Hypothetical protein
				Rifp1Sym_dg00050	Rifp2Sym_kv00010	Hypothetical protein
				Rifp1Sym_dg00060	Rifp2Sym_kj00010	Hypothetical protein
				Rifp1Sym_dg00070	Rifp2Sym_kj00020	Hypothetical protein
				Rifp1Sym_dg00080	n.d.	Hypothetical protein
				Rifp1Sym_dg00090	n.a.	Ribosomal large subunit pseudouridine synthase C
				Rifp1Sym_dg00100	Rifp2Sym_lw00010	Hypothetical protein
				Rifp1Sym_dg00110	n.a.	Hypothetical protein
				Rifp1Sym_dg00120	Rifp2Sym_me00010	Hypothetical protein
				Rifp1Sym_dg00130	n.a.	Putative plasmid stabilization system protein
				Rifp1Sym_dg00140	n.d.	Hypothetical protein
				Rifp1Sym_dg00150	n.d.	Hypothetical protein
				Rifp1Sym_dg00160	n.a.	Hypothetical protein
	AFOC010000153	5054	0 - 5054	Rifp1Sym_fv00010	Rifp2Sym_ja00010	Hypothetical protein
				Rifp1Sym_fv00020	Rifp2Sym_ja00020	Hypothetical protein
				Rifp1Sym_fv00030	Rifp2Sym_kh00010	Hypothetical protein
				Rifp1Sym_fv00040	n.a.	YbaK/prolyl-tRNA synthetase associated region
Exclusive sequence 2	AFOC010000112	9453	0 - 9453	Rifp1Sym_eg00010	Rifp2Sym_jz00010	CRISPR-associated protein Cas1
				Rifp1Sym_eg00020	n.a.	CRISPR system Cascade subunit CasE
				Rifp1Sym_eg00030	n.a.	CRISPR system Cascade subunit CasD
				Rifp1Sym_eg00040	Rifp2Sym_hp00010	CRISPR system Cascade subunit CasC
				Rifp1Sym_eg00050	Rifp2Sym_hp00030	CRISPR system Cascade subunit CasA

D.14

Back to Appendix D

Table D.3 Accessory genome found in *Riftia* symbionts (*Rifita* 1 and *Rifita* 2 symbiont genome assemblies). Total size = 77 303 bp; n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called; n.d. = Not Detected; sequence absent from the assembly. **(Continued)**

Exclusive sequence	Reference contig	Sequence length	Start - End (in reference contig)	Locus-tag in *Riftia* 1 symbionts	Locus-tag in *Riftia* 2 symbionts	Product
				Rifp1Sym_eg00060	Rifp2Sym_mb00010	CRISPR-associated endonuclease/helicase Cas3
Exclusive sequence 3	AFOC010000148	5239	0 - 5239	Rifp1Sym_fq00010	Rifp2Sym_fa00040	Protein of unknown function (DUF1788)
				Rifp1Sym_fq00020	Rifp2Sym_fa00030; Rifp2Sym_fa00020	Hypothetical protein
				Rifp1Sym_fx00010	n.a.	Transcriptional regulator, AlpA family
				Rifp1Sym_fx00020	n.a.	Integrating conjugative element relaxase, PFL_4751 family
	contig00270	2139	0 - 2139	n.d.	Rifp2Sym_iy00020	Conjugative coupling factor TraD, SXT/TOL

D.15

Back to Appendix D

Table D.4 Accessory genome found in 9°N symbionts (*Tevnia* and *Rifita* 2 symbiont genome assemblies but not in *Riftia* 1 symbionts). Total size = 75 723 bp;n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence	Reference contig name	sequence length	Start – End (in reference contig)	Low Flanking conservation	Locus-tag in *Tevnia* symbionts	Locus-tag in *Riftia* 2 symbionts	Product
Exclusive sequence 1	AFZB01000005	16398	22436 - 38834		TevJSym_ae00240	Rifp2Sym_ee00120	Phage protein, HK97 gp10 family
					TevJSym_ae00250	n.a.	Hypothetical protein
					TevJSym_ae00260	n.a.	Hypothetical protein
					TevJSym_ae00270	n.a.	Hypothetical protein
					TevJSym_ae00280	n.a.	Serine-rich adhesin-like protein for platelets precursor
					TevJSym_ae00290	Rifp2Sym_er00020	Hypothetical protein
					TevJSym_ae00300	Rifp2Sym_er00030	Peptidase family M23
					TevJSym_ae00310	Rifp2Sym_er00040	Hypothetical protein
					TevJSym_ae00320	Rifp2Sym_er00050	Hypothetical protein
					TevJSym_ae00330	Rifp2Sym_er00060	Mu-like prophage I protein
					TevJSym_ae00340	Rifp2Sym_er00080; Rifp2Sym_er00070	Antitoxin YefM
					TevJSym_ae00350	Rifp2Sym_er00090	Hypothetical protein
					TevJSym_ae00360	Rifp2Sym_er00110; Rifp2Sym_er00100	Hypothetical protein
					TevJSym_ae00370	Rifp2Sym_bw00170	Hypothetical protein
					TevJSym_ae00380	Rifp2Sym_bw00160	Hypothetical protein
					TevJSym_ae00390	Rifp2Sym_bw00150	Hypothetical protein
					TevJSym_ae00400	Rifp2Sym_bw00140	Hypothetical protein
Exclusive sequence 2	AFZB01000002	2860	126914 - 129774	✓	TevJSym_ab01270	Rifp2Sym_bw00010	Hypothetical protein
					TevJSym_ab01280	n.a.	Hypothetical protein

D.16

Back to Appendix D

Table D.4 Accessory genome found in 9°N symbionts (*Tevnia* and *Rifita* 2 symbiont genome assemblies but not in *Riftia* 1 symbionts). Total size = 75 723 bp;n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

					TevJSym_ab01290	n.a.	Vault protein inter-alpha-trypsin domain-containing protein
					TevJSym_ab01300	n.a.	Transposase InsO and inactivated derivatives
Exclusive sequence 3	AFZB01000050	9786	0 - 9786		TevJSym_bx00010	n.a.	ATPase
					TevJSym_bx00020	n.a.	Putative ATPase involved in DNA repair
					TevJSym_bx00030	n.a.	Hypothetical protein
Exclusive sequence 4	AFZB01000057	6657	0 - 6657		TevJSym_ce00010	n.a.	T/G mismatch-specific endonuclease
					TevJSym_ce00020	n.a.	DNA (cytosine-5)-methyltransferase 1
					TevJSym_ce00030	n.a.	KamA family protein
Exclusive sequence 5	AFZB01000036	6411	26656 - 33067	✓	TevJSym_bj00220	Rifp2Sym_ht00010	Hemolysin-type calcium-binding repeat-containing protein
					TevJSym_bj00230	Rifp2Sym_ag00010	Hemolysin D
					TevJSym_bj00240	Rifp2Sym_jc00010	Leukotoxin translocation ATP-biding protein/toxin secretion ABC transporter ATP-binding
Exclusive sequence 6	AFZB01000022	5730	0 - 5730		TevJSym_av00010	Rifp2Sym_he00010	Integron integrase
Exclusive sequence 7	AFZB01000006	1847	18500 - 20347	✓	TevJSym_af00170	Rifp2Sym_de00140	Membrane-bound lytic murein transglycosylase F
					TevJSym_af00180	Rifp2Sym_de00150	Phospholipase A1
					TevJSym_af00190	Rifp2Sym_de00160	tRNA(adenine34) deaminase
					TevJSym_af00200	n.a.	GMP synthase (glutamine-hydrolysing)
Exclusive sequence 8	AFZB01000083	1367	0 - 1367		n.a.	Rifp2Sym_km00020	Group II intron, maturase-specific domain

D.17

Back to Appendix D

Table D.4 Accessory genome found in 9°N symbionts (*Tevnia* and *Rifita* 2 symbiont genome assemblies but not in *Riftia* 1 symbionts). Total size = 75 723 bp;n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

				n.a.	Rifp2Sym_km00020	IS3 family transposase, orfB
Exclusive sequence 9	AFZB01000002	5754	173150 - 178904	TevJSym_ab01730	Rifp2Sym_jq00010	ATP-dependent RNA helicase RhlE
				TevJSym_ab01740	n.a.	Iron complex outermembrane recepter protein
Exclusive sequence 10	AFZB01000045	3703	0 - 3703	TevJSym_bs00010	Rifp2Sym_da00010	Integrase/recombinase XerD
				TevJSym_bs00020	Rifp2Sym_da00020	Resolvase, N terminal domain
Exclusive sequence 11	AFZB01000035	3521	1360 - 4881	TevJSym_bi00030	Rifp2Sym_hr00010	Peptide methionine sulfoxide reductase msrA/msrB
				TevJSym_bi00040	Rifp2Sym_hr00020	Hypothetical protein
				TevJSym_bi00050	Rifp2Sym_hr00030; Rifp2Sym_fg00010	Ribonucleoside-triphosphate reductase class III catalytic subunit
Exclusive sequence 12	AFZB01000068	3381	0 - 3381	TevJSym_cp00010	n.a.	Tertatricopeptide TPR_2 repeat protein
Exclusive sequence 13	AFZB01000005	2120	79186 - 81306	TevJSym_ae00700	Rifp2Sym_cv00010	MMPL family protein
Exclusive sequence 14	AFZB01000079	1665	0 - 1665	TevJSym_cz00010	Rifp2Sym_jr00010	Filamentation induced by cAMP protein Fic
Exclusive sequence 15	AFZB01000064	1541	0 - 1541	TevJSym_cl00010	Rifp2Sym_ic00010	Sodium/proton antiporter, NhaA family
Exclusive sequence 16	AFZB01000064	1487	2936 - 4423	TevJSym_cl00030	n.a.	Ribosomal subunit interface protein
				TevJSym_cl00040	Rifp2Sym_ko00010; Rifp2Sym_ko00020; Rifp2Sym_ko00030	LysR family transcriptional regulator, transcriptional activator of nhaA

D.18

Back to Appendix D

Table D.4 Accessory genome found in 9°N symbionts (*Tevnia* and *Rifita* 2 symbiont genome assemblies but not in *Riftia* 1 symbionts). Total size = 75 723 bp;n.a. = Not Annotated; presence of sequence or part of sequence but no genes were called. **(Continued)**

Exclusive sequence 17	AFZB01000001	1211	135783 - 136993	✓	TevJSym_aa01480	Rifp2Sym_bk00140	Nitrate/nitrite transport system substrate-binding protein
					TevJSym_aa01490	Rifp2Sym_bk00150	ANTAR domain-containing protein
Exclusive sequence 18	AFZB01000016	284	12994 - 13278		TevJSym_ap00160	Rifp2Sym_is00010	LTXXQ motif family protein

D.19

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
1. Sulfur metabolism		aprA	Adenylylsulfate reductase, alpha subunit	Ga0074115_10324	Ga0076813_14621	TevJSym_ag00880	Rifp1Sym_ej00110	Rifp2Sym_bz00020 ; Rifp2Sym_bz00030
		aprB	Adenylylsulfate reductase, beta subunit	Ga0074115_10323	Ga0076813_14622	TevJSym_ag00890	Rifp1Sym_ej00100	Rifp2Sym_bz00040
		dsrA	Dissimilatory sulfite reductase, alpha subunit	Ga0074115_12121	Ga0076813_15802	TevJSym_aw00210	Rifp1Sym_am00340	Rifp2Sym_aa00430 ; Rifp2Sym_aa00440
		dsrB	Dissimilatory sulfite reductase, beta subunit	Ga0074115_12120	Ga0076813_15803	TevJSym_aw00190	Rifp1Sym_am00350	Rifp2Sym_aa00410
		sopT	ATP sulfurylase	n.d.	n.d.	TevJSym_ad01200	Rifp1Sym_cn00110	Rifp2Sym_az00240
2. Carbon metabolism	2.1. Calvin Benson Bassham cycle	cbbM	Ribulose 1,5 bisphosphate carboxylase/oxygenase	Ga0074115_14812	Ga0076813_14651	TevJSym_aj00630	Rifp1Sym_at00130	Rifp2Sym_bi00210
		gapA	Glyceraldehyde 3 phosphate dehydrogenase	Ga0074115_12924	Ga0076813_16395	TevJSym_an00400	Rifp1Sym_dp00070	Rifp2Sym_ar00120
		pgk	Phosphoglycerate kinase	Ga0074115_12925	Ga0076813_1392	TevJSym_an00390	Rifp1Sym_dp00060	Rifp2Sym_ar00110
		prkB	Phosphoribulokinase	Ga0074115_1484	Ga0076813_15275	TevJSym_aj00540	Rifp1Sym_at00030	Rifp2Sym_bf00040 ; Rifp2Sym_bf00050
		aclA	ATP citrate lyase, alpha subunit	Ga0074115_13717	Ga0076813_13921	TevJSym_az00170	Rifp1Sym_bt00020	Rifp2Sym_br00010 ; Rifp2Sym_aa000101

D.20

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
	2.2. TCA cycle	aclB	ATP citrate lyase, beta subunit	Ga0074115_13718	Ga0076813_14412	TevJSym_az00180	Rifp1Sym_bt00010	Rifp2Sym_br00020
		acnA	Aconitate hydratase/aconitase A	Ga0074115_13715	Ga0076813_13923	TevJSym_az00150	Rifp1Sym_bt00040	Rifp2Sym_aa00030
		icd	Isocitrate dehydrogenase [NADP]	Ga0074115_13716	Ga0076813_13922	TevJSym_az00160	Rifp1Sym_bt00030	Rifp2Sym_aa00020
		korA	2 oxoglutarate oxidoreductase, alpha subunit 2	Ga0074115_13621	Ga0076813_14941	TevJSym_az00290	Rifp1Sym_dm00080	Rifp2Sym_br00130
				Ga0074115_102185	Ga0076813_169112	TevJSym_bb00050	Rifp1Sym_aa00580	Rifp2Sym_an00050
				Ga0074115_11633	Ga0076813_13526	TevJSym_ar00630	Rifp1Sym_bp00050	Rifp2Sym_ab00620
		korB	2 oxoglutarate oxidoreductase, beta subunit 2	Ga0074115_13622	Ga0076813_14942	TevJSym_az00280	Rifp1Sym_dm00070	Rifp2Sym_br00120
				Ga0074115_102184	Ga0076813_169111	TevJSym_bb00040	Rifp1Sym_aa00590	Rifp2Sym_an00040
				Ga0074115_11632	Ga0076813_13525	TevJSym_ar00620	Rifp1Sym_bp00060	Rifp2Sym_ab00610
		korD	2 oxoglutarate oxidoreductase, delta subunit	Ga0074115_13620	Ga0076813_15152	TevJSym_az00300	Rifp1Sym_dm00090	Rifp2Sym_br00140
		korG	2 oxoglutarate oxidoreductase, gamma subunit	Ga0074115_13623	Ga0076813_14943	TevJSym_az00270	Rifp1Sym_dm00060	Rifp2Sym_br00110
		gltA2	citrate synthase I	Ga0074115_1551	Ga0076813_11841	TevJSym_bc00420	Rifp1Sym_ax00310	Rifp2Sym_aq00120
		maeB	malate dehydrogenase	Ga0074115_1552	Ga0076813_11901	TevJSym_bc00430	Rifp1Sym_ax00320 ; Rifp1Sym_ax00330	Rifp2Sym_aq001101 ; Rifp2Sym_aq00100
		sdhA	Succinate dehydrogenase, flavoprotein subunit2	Ga0074115_13138	Ga0076813_15895	TevJSym_ah01000	Rifp1Sym_dd00140	Rifp2Sym_ez00030

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
				Ga0074115_1555	Ga0076813_14861	TevJSym_cj00010	Rifp1Sym_ax00360	Rifp2Sym_aq000701 ; Rifp2Sym_aq00060
		sdhB	succinate dehydrogenase iron sulfur protein2	Ga0074115_13139	Ga0076813_15896	TevJSym_ah01010	Rifp1Sym_dd00150	Rifp2Sym_ez00020
				Ga0074115_1556	Ga0076813_14862	TevJSym_cj00020	Rifp1Sym_ax00370	Rifp2Sym_aq00050
		sdhC	succinate dehydrogenase membrane anchor 2	Ga0074115_13137	Ga0076813_15894	TevJSym_ah00990	Rifp1Sym_dd00130	Rifp2Sym_ez00040
				Ga0074115_1554	Ga0076813_15132	TevJSym_bc00450	Rifp1Sym_ax00350	Rifp2Sym_aq00080
		sdhD	succinate dehydrogenase cyt b subunit	Ga0074115_1553	Ga0076813_11902	TevJSym_bc00440	Rifp1Sym_ax00340	Rifp2Sym_aq00090
3. Nitrogen metabolism		narG	Respiratory nitrate reductase, alpha chain	Ga0074115_10332	Ga0076813_150210	TevJSym_ag00800	Rifp1Sym_ak00460	Rifp2Sym_bt00080
		narH	Respiratory nitrate reductase, beta chain	Ga0074115_10333	Ga0076813_15028	TevJSym_ag00790	Rifp1Sym_ak00450	Rifp2Sym_bt00090
		narJ	Respiratory nitrate reductase, delta chain	Ga0074115_10334	Ga0076813_15027	TevJSym_ag00780	Rifp1Sym_ak00440	Rifp2Sym_bt00100
		narI	Respiratory nitrate reductase, gamma chain	Ga0074115_10335	Ga0076813_15026	TevJSym_ag00770	Rifp1Sym_ak00430	Rifp2Sym_bt00110
		napD	Periplasmic nitrate reductase, subunit NapD	Ga0074115_1444	Ga0076813_15362	TevJSym_al00660	Rifp1Sym_ao00250	Rifp2Sym_bd00210
		napA	Periplasmic nitrate reductase, subunit NapA	Ga0074115_1443	Ga0076813_15361	TevJSym_al00650	Rifp1Sym_ao00240 ; Rifp1Sym_ao00230	Rifp2Sym_bd00220

D.22

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
		napG	Periplasmic nitrate reductase, subunit NapG	Ga0074115_1442	n.d.	TevJSym_al00640	Rifp1Sym_ao00220	Rifp2Sym_bd00230
		napH	Periplasmic nitrate reductase, subunit NapH	Ga0074115_1441	n.d.	TevJSym_al00630	Rifp1Sym_ao00210	Rifp2Sym_bd00240
		napB	Periplasmic nitrate reductase, subunit NapB	Ga0074115_1762	Ga0076813_1782	TevJSym_al00620	Rifp1Sym_ao00200	Rifp2Sym_bd00240
		napC	Periplasmic nitrate reductase, subunit NapC	Ga0074115_12753	Ga0076813_1783	TevJSym_al00610	Rifp1Sym_ao00190	Rifp2Sym_bd00260
		nirN	Cytochrome cd1/nitrite reductase, subunit NirN	Ga0074115_10211	Ga0076813_13753	TevJSym_bo00090	Rifp1Sym_cp00140	Rifp2Sym_db00140
		nirJ	Cytochrome cd1/nitrite reductase, subunit NirJ	Ga0074115_10212	Ga0076813_12089	TevJSym_bo00100	Rifp1Sym_cp00120 ; Rifp1Sym_cp00130	Rifp2Sym_db00110
		nirH	Cytochrome cd1/nitrite reductase, subunit NirH	Ga0074115_10213	Ga0076813_12088	TevJSym_bo00110	Rifp1Sym_cp00110	Rifp2Sym_db00100
		nirG	Cytochrome cd1/nitrite reductase, subunit NirG	Ga0074115_10215	Ga0076813_12086	TevJSym_bo00130	Rifp1Sym_cp00090	Rifp2Sym_db00080
		nirL	Cytochrome cd1/nitrite reductase, subunit NirL	Ga0074115_10216	Ga0076813_12085	TevJSym_bo00140	Rifp1Sym_cp00080	Rifp2Sym_db00070
		nirD	Cytochrome cd1/nitrite reductase, subunit NirD	Ga0074115_10217	Ga0076813_12084	TevJSym_bo00150	Rifp1Sym_cp00070	Rifp2Sym_db00060
		nirF	Cytochrome cd1/nitrite reductase,	Ga0074115_10218	Ga0076813_12083	TevJSym_bo00160	Rifp1Sym_cp00060	Rifp2Sym_db00050

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
			subunit NirF					
		nirC	Cytochrome cd1/nitrite reductase, subunit NirC	Ga0074115_10219	Ga0076813_12082	TevJSym_bo00170	Rifp1Sym_cp00050	Rifp2Sym_db00040
		nirT	Cytochrome cd1/nitrite reductase, subunit NirT	Ga0074115_10221	Ga0076813_10286	TevJSym_bo00190	Rifp1Sym_cp00030	Rifp2Sym_db00010
		nirS	Cytochrome cd1/nitrite reductase, subunit NirS	Ga0074115_10222	Ga0076813_10285	TevJSym_bo00200	Rifp1Sym_cp00020	Rifp2Sym_gn00030
		norB	Nitric oxide reductase subunit B	Ga0074115_10224	Ga0076813_10283	TevJSym_cv00020	n. a.	Rifp2Sym_jh00010
		norC	Nitric oxide reductase subunit C	Ga0074115_10225	Ga0076813_10282	TevJSym_cv00010	n. a.	Rifp2Sym_jh00030
		nosZ	Nitrous oxide reductase	Ga0074115_11648	Ga0076813_16471	TevJSym_ac00130	Rifp1Sym_ag00460	Rifp2Sym_au00130 ; Rifp2Sym_au00140
4. Storage compounds		cphA	Cyanophycin synthase	Ga0074115_1582	Ga0076813_12881	TevJSym_ay00040	Rifp1Sym_ah00040	Rifp2Sym_at00040
5. Oxidative stress response		ahpC	Alkyl hydroperoxide reductase	Ga0074115_1298	Ga0076813_1752	TevJSym_ah00580	Rifp1Sym_ai00550	Rifp2Sym_af00050
		sodB	Superoxide dismutase [Fe]	Ga0074115_10415	Ga0076813_192	TevJSym_an00790	Rifp1Sym_eb00070	Rifp2Sym_dw00080
6. Secretion systems	6.1. Secretion across the inner	secA	Preprotein translocase subunit SecA	Ga0074115_10789	Ga0076813_142814	TevJSym_aj00210	Rifp1Sym_bc00080	Rifp2Sym_ax00280

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
	membrane							
		secB	Preprotein translocase subunit SecB	Ga0074115_10285	Ga0076813_14246	TevJSym_at00230	Rifp1Sym_cl00120	Rifp2Sym_as00170
		secD	Preprotein translocase subunit SecD	Ga0074115_1295	Ga0076813_13813	TevJSym_an00610	Rifp1Sym_cf00160	Rifp2Sym_ar00340
		secE	Preprotein translocase subunit SecE	Ga0074115_12212	n.d.	TevJSym_bu00110	Rifp1Sym_di00120	Rifp2Sym_dg00010
		secF	Preprotein translocase subunit SecF	Ga0074115_1296	Ga0076813_13814	TevJSym_an00600	Rifp1Sym_cf00150	Rifp2Sym_ar00330
		secG	Preprotein translocase subunit SecG	n.d.	n.d.	TevJSym_ap00640	Rifp1Sym_bl00190	Rifp2Sym_eh00010
		secY	Preprotein translocase subunit SecY	Ga0074115_1496	Ga0076813_11654	TevJSym_aa00220	Rifp1Sym_bd00210	Rifp2Sym_ce00130
		tatA	Twin arginine translocation protein TatA	Ga0074115_10885	Ga0076813_150318	TevJSym_bh00280	Rifp1Sym_cq00210	Rifp2Sym_cb00170
		tatB	Twin arginine translocation protein TatB	Ga0074115_10884	Ga0076813_150319	TevJSym_bh00290	Rifp1Sym_cq00220	Rifp2Sym_cb00180
		tatC	Twin arginine translocation protein TatC	Ga0074115_10883	Ga0076813_150320	TevJSym_bh00300	Rifp1Sym_cq00230	Rifp2Sym_cb00190
	6.3. Secretion across the outer membrane	exeA	General secretion pathway protein A 2	Ga0074115_11430	Ga0076813_15682	TevJSym_ab01170	Rifp1Sym_aa00370	Rifp2Sym_an00280
				Ga0074115_1393	Ga0076813_11442	TevJSym_bg00130	Rifp1Sym_bk00170	Rifp2Sym_bh00210
				Ga0074115_1331	Ga0076813_11735	TevJSym_bb00270	Rifp1Sym_aa00370	Rifp2Sym_an00280

D.25

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
							Rifp1Sym_cw00170	
				Ga0074115_11330	Ga0076813_168710	TevJSym_ai00590	Rifp1Sym_cw00160	Rifp2Sym_cf00120
				Ga0074115_11430	Ga0076813_15682	TevJSym_ab01170	Rifp1Sym_ee00070	Rifp2Sym_dq00020
		xpsD	General secretion pathway protein D	Ga0074115_12041	Ga0076813_12179	TevJSym_ay00270	Rifp1Sym_ah00230	Rifp2Sym_do00140
		outE	General secretion pathway protein E	Ga0074115_1392	Ga0076813_11452	TevJSym_bg00140	Rifp1Sym_bk00160	Rifp2Sym_bh00200
		xcpR	General secretion pathway protein E	Ga0074115_12421	Ga0076813_15101	TevJSym_av00370	Rifp1Sym_ci00020	Rifp2Sym_be00150
		xcpS	General secretion pathway protein F	Ga0074115_12420	Ga0076813_14619	TevJSym_av00380	Rifp1Sym_ci00030	Rifp2Sym_be00160
				Ga0074115_1391	Ga0076813_11451	TevJSym_bg00150	Rifp1Sym_bk00150	Rifp2Sym_bh00190
	6.4. Type VI secretion system		Zn-binding Pro-Ala-Ala-Arg (PAAR) domain, involved in TypeVI secretion	Ga0074115_1462	Ga0076813_11513	n.d.	n.d.	n.d.
			putative component of the type VI protein secretion system	Ga0074115_1466	Ga0076813_14751 ; Ga0076813_16801	n.d.	n.d.	n.d.
		ImpA	type VI secretion-associated protein, ImpA family	Ga0074115_1469	Ga0076813_16804	n.d.	n.d.	n.d.
		ImpB	type VI secretion protein, VC_A0107 family	Ga0074115_14610	Ga0076813_16805	n.d.	n.d.	n.d.
		ImpC	type VI secretion protein, EvpB/VC_A0108 family	Ga0074115_14611	Ga0076813_13723	n.d.	n.d.	n.d.

D.26

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
		Hcp	Type VI protein secretion system component Hcp (secreted cytotoxin)	Ga0074115_14612	Ga0076813_13722	n.d.	n.d.	n.d.
		ImpF	type VI secretion system lysozyme-like protein	Ga0074115_14614	Ga0076813_14483	n.d.	n.d.	n.d.
		ImpG	type VI secretion protein, VC_A0110 family	Ga0074115_14615	Ga0076813_14482	n.d.	n.d.	n.d.
		ImpH	type VI secretion protein, VC_A0111 family	Ga0074115_14616	Ga0076813_14481	n.d.	n.d.	n.d.
7. Horizontal gene transfer	7.1. F Type conjugative plasmid	traG	IncF plasmid conjugative transfer protein TraG	n.d.	n.d.	n.d.	Rifp1Sym_au00110	n.a.
		traH	IncF plasmid conjugative transfer pilus assembly protein	n.d.	n.d.	n.d.	Rifp1Sym_au00120	n.a.
		traF	IncF plasmid conjugative transfer pilus assembly protein	n.d.	n.d.	n.d.	Rifp1Sym_au00130	n.a.
		traN	IncF plasmid conjugative transfer protein TraN	n.d.	n.d.	n.d.	Rifp1Sym_au00140	n.a.
		trbC	IncF plasmid conjugative transfer protein TrbC	n.d.	n.d.	n.d.	Rifp1Sym_au00150	n.a.
		traU	IncF plasmid conjugative transfer	n.d.	n.d.	n.d.	Rifp1Sym_au00160	n.a.

D.27

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
			pilus assembly protein					
		pat17	Patatin like phospholipase A2	n.d.	n.d.	n.d.	Rifp1Sym_au00170	n.a.
		traC	IncF plasmid conjugative transfer pilus assembly protein	n.d.	n.d.	n.d.	Rifp1Sym_au00200	n.a.
		traV	Conjugative transfer protein TraV	n.d.	n.d.	n.d.	Rifp1Sym_au00210	n.a.
		traB	IncF plasmid conjugative transfer pilus assembly protein	n.d.	n.d.	n.d.	Rifp1Sym_au00230	n.a.
		traK	IncF plasmid conjugative transfer pilus assembly protein	n.d.	n.d.	n.d.	Rifp1Sym_au00240	n.a.
		traE	Plasmid like conjugative transfer protein TraE	n.d.	n.d.	n.d.	Rifp1Sym_au00250	n.a.
7. Horizontal gene transfer	7.2. Transposases		IS5Y transposase	n.d.	n.d.	TevJSym_bc00030	Rifp1Sym_al00300	Rifp2Sym_aw00230
			Transposase IS66	Ga0074115_1614	Ga0076813_13865	TevJSym_el00010	n.d.	Rifp2Sym_mp00010
				Ga0074115_1615	Ga0076813_13866	n.a	n.a	n.a
			Transposase	Ga0074115_102126	Ga0076813_169033	TevJSym_aq00210	Rifp1Sym_ae00170	Rifp2Sym_bm00130
				Ga0074115_10991	Ga0076813_1302	TevJSym_ar00260	Rifp1Sym_bh00140	Rifp2Sym_ab00230
				Ga0074115_1654	Ga0076813_10561	TevJSym_ck00010	Rifp1Sym_db00090	Rifp2Sym_gs00010
				Ga0074115_12166	Ga0076813_16881	n.d.	n.d.	n.d.
				Ga0074115_11443	Ga0076813_12739	n.d.	n.d.	n.d.
				Ga0074115_1652	Ga0076813_10563	n.d.	n.d.	n.d.

D.28

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
				Ga0074115_102124	Ga0076813_169035	n.d.	n.d.	n.d.
				Ga0074115_1651	Ga0076813_10564	n.d.	n.d.	n.d.
				Ga0074115_102125	Ga0076813_169034	n.d.	n.d.	n.d.
			IS5 transposase	n.d	n.d	TevJSym_bc00040	Rifp1Sym_al00310	Rifp2Sym_aw00240
			Transposase insF for insertion sequence IS3	n.a	n.a	TevJSym_ab01630	Rifp1Sym_bb00110	Rifp2Sym_dh00120
			Transposase for IS1668	n.d	n.d	TevJSym_bf00240	Rifp1Sym_bg00180	Rifp2Sym_bl00030
			Transposase, IS4 family	n.a	n.a	TevJSym_cx00010	Rifp1Sym_db00080	Rifp2Sym_jp00010
8. Possibly involved in host infection	8.1. Adherence by type IV pili/fimbriae	fimT	Type IV fimbrial biogenesis protein FimT	Ga0074115_11736	Ga0076813_159923	TevJSym_am00340	Rifp1Sym_aq00380	Rifp2Sym_ak00170
				Ga0074115_12515	Ga0076813_14637	TevJSym_bs00040	Rifp1Sym_et00070	Rifp2Sym_da00030
		fimV	Type IV pilus assembly protein FimV	Ga0074115_10978	Ga0076813_15952	TevJSym_ar00130 ; TevJSym_ar00120	Rifp1Sym_dy00010 ; Rifp1Sym_dy00020	Rifp2Sym_ab00080
		pilB	Type IV fimbrial assembly protein PilB	Ga0074115_11359	Ga0076813_139111	TevJSym_ai00290	Rifp1Sym_az00200	Rifp2Sym_ba00260
				Ga0074115_11426	Ga0076813_14567	TevJSym_ab01140	Rifp1Sym_ee00030	Rifp2Sym_dq00060
		pilE	Type IV pilus biogenesis protein PilE	Ga0074115_11742	Ga0076813_159917	TevJSym_am00410	Rifp1Sym_an00070	Rifp2Sym_ak00100
		pilG	twitching motility protein PilG	Ga0074115_14215	Ga0076813_11134	TevJSym_bn00040	Rifp1Sym_cb00050	Rifp2Sym_cj00200
		pilH	twitching motility protein PilH	Ga0074115_14214	Ga0076813_11965	TevJSym_bn00030	Rifp1Sym_cb00040	Rifp2Sym_cj00210
		pilM	Type IV pilus biogenesis protein PilM	Ga0074115_102194	Ga0076813_12694	TevJSym_bb00140	Rifp1Sym_aa00490	Rifp2Sym_an00140

D.29

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
		pilN	Type IV pilus biogenesis protein PilN	Ga0074115_102195	Ga0076813_12695	TevJSym_bb00150	Rifp1Sym_aa00480	Rifp2Sym_an00150
				Ga0074115_12033	Ga0076813_113526	TevJSym_ay00340	Rifp1Sym_ah00310	Rifp2Sym_do00050
		pilQ	Type IV pilus biogenesis protein PilQ	Ga0074115_102198	Ga0076813_12698	TevJSym_bb00180	Rifp1Sym_aa00450	Rifp2Sym_an00180
		pilR	Type IV fimbriae expression regulatory protein PilR	Ga0074115_11363	Ga0076813_10808	TevJSym_ai00250	Rifp1Sym_az00160	Rifp2Sym_ba00220
		pilS	Sensor protein PilS	Ga0074115_11364	Ga0076813_10807	TevJSym_ai00240	Rifp1Sym_az00150	Rifp2Sym_ba00210
		pilT	Twitching mobility protein PilT	Ga0074115_11726	Ga0076813_12482	TevJSym_am00210	Rifp1Sym_aq00280	Rifp2Sym_ak00290
		pilV	Type IV pilus modification protein PilV	Ga0074115_11737	Ga0076813_159922	TevJSym_am00350	Rifp1Sym_aq00390	Rifp2Sym_ak00160
		pilY	Type IV fimbrial biogenesis protein PilY	Ga0074115_11740	Ga0076813_159924	TevJSym_am00390	Rifp1Sym_an00030 ; Rifp1Sym_an00040 ; Rifp1Sym_an00050	Rifp2Sym_ak00130 ; Rifp2Sym_ak00120
		pilZ	Type IV pilus assembly PilZ	Ga0074115_11818	Ga0076813_12504	TevJSym_bc00220	Rifp1Sym_ax00100	Rifp2Sym_aq00310
		TadD	Flp pilus assembly protein TadD, contains TPR repeats	Ga0074115_12218	Ga0076813_16271	TevJSym_bt00010	Rifp1Sym_fu00030	Rifp2Sym_ct00190
		pilO	Tfp pilus assembly protein PilO	Ga0074115_102196	Ga0076813_12696	TevJSym_bb00160	Rifp1Sym_aa00470	Rifp2Sym_an00160
		pilX	Tfp pilus assembly protein PilX	Ga0074115_11739	Ga0076813_159920	TevJSym_am00380	Rifp1Sym_an00020	Rifp2Sym_ak00140
		pilP	Tfp pilus assembly	Ga0074115_102197	Ga0076813_12697	TevJSym_bb00170	Rifp1Sym_aa00460	Rifp2Sym_an00170

D.30

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
			protein PilP					
	8.2. Adherence by flagellum	mshL	Pilus (MSHA Type) biogenesis protein MshL	Ga0074115_1395	Ga0076813_13261	TevJSym_bg00100	Rifp1Sym_bk00190	Rifp2Sym_bh00230
		flaG	Flagellin protein FlaG	Ga0074115_1433	Ga0076813_12133	TevJSym_aa01600	Rifp1Sym_fp00070	Rifp2Sym_bk00240
		fleN	Flagellar synthesis regulator FleN	Ga0074115_12148	Ga0076813_12351	TevJSym_aw00490	Rifp1Sym_am00080	Rifp2Sym_aa00690
		flgA	Flagellar basal body P ring formation protein FlgA	Ga0074115_13818	Ga0076813_14258	TevJSym_aa01200	Rifp1Sym_bw00270	Rifp2Sym_aj00140
		fldG	Flagellar hook protein FlgG	Ga0074115_13812	Ga0076813_14252	TevJSym_aa01270	Rifp1Sym_bw00200	Rifp2Sym_aj00080
		flgE	Flagellar hook protein FlgE	Ga0074115_13814	Ga0076813_14254	TevJSym_aa01250	Rifp1Sym_bw00220	Rifp2Sym_aj00100
		flgF	Flagellar basal body rod protein FlgF	Ga0074115_13813	Ga0076813_14253	TevJSym_aa01260	Rifp1Sym_bw00210	Rifp2Sym_aj00090
		flgH	Flagellar L ring protein FlgH	Ga0074115_13811	Ga0076813_14251	TevJSym_aa01280	Rifp1Sym_bw00190	Rifp2Sym_aj00070
		flhA	Flagellar biosynthesis protein FlhA	Ga0074115_12150	Ga0076813_16232	TevJSym_aw00510	Rifp1Sym_am00060	Rifp2Sym_aa00720
		fliA	RNA polymerase sigma factor for flagellar operon	Ga0074115_12147	Ga0076813_14524	TevJSym_aw00480	Rifp1Sym_am00090	Rifp2Sym_aa00680
		fliD	Flagellar hook associated protein 2	Ga0074115_1432	Ga0076813_12134	TevJSym_aa01610	Rifp1Sym_fp00060	Rifp2Sym_bk00250
		fliG	Flagellar motor switch protein FliG	Ga0074115_1197	Ga0076813_12727	TevJSym_ac00750	Rifp1Sym_cx00180	Rifp2Sym_cm00100
		fliJ	Flagellar protein FliJ	Ga0074115_1194	Ga0076813_12724	TevJSym_ac00720	Rifp1Sym_ck00160	Rifp2Sym_cm00060
		fliM	Flagellar motor switch	Ga0074115_12157	Ga0076813_12633	TevJSym_aw00580	n.d.	Rifp2Sym_aa00790

D.31

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
			protein FliM					
		flgB	Flagellar basal-body rod protein FlgB	Ga0074115_13817	Ga0076813_14257	TevJSym_aa01210	Rifp1Sym_bw00260	Rifp2Sym_aj00130
		FliL	Flagellar basal body-associated protein FliL	Ga0074115_1347	Ga0076813_155326	TevJSym_ah00540	Rifp1Sym_ai00510	Rifp2Sym_af00090
				Ga0074115_12158	Ga0076813_12632	TevJSym_aw00590	Rifp1Sym_ct00020	Rifp2Sym_aa00800
		FliO	flagellar biosynthetic protein FliO	Ga0074115_12155	Ga0076813_12635	TevJSym_aw00560	Rifp1Sym_am00010	Rifp2Sym_aa00770
		GldG	ABC-type uncharacterized transport system involved in gliding motility, auxiliary component	Ga0074115_1173	Ga0076813_12682	TevJSym_ai00040	Rifp1Sym_aq00050	Rifp2Sym_ba00010
		flgC	flagellar basal-body rod protein FlgC	Ga0074115_13816	Ga0076813_14256	TevJSym_aa01220	Rifp1Sym_bw00250	Rifp2Sym_aj00120
		FlgG	flagellar basal-body rod protein FlgG, Gram-negative bacteria	Ga0074115_13812	Ga0076813_14252	TevJSym_aa01270	Rifp1Sym_bw00200	Rifp2Sym_aj00080
		FlgI	Flagellar basal body P-ring protein FlgI	Ga0074115_13810	Ga0076813_13251	TevJSym_aa01290	Rifp1Sym_bw00180	Rifp2Sym_aj00060
		FlgL	flagellar hook-associated protein 3	Ga0074115_1387	Ga0076813_1246	TevJSym_aa01320	Rifp1Sym_bw00140	Rifp2Sym_aj00030
		FlhF	flagellar biosynthetic protein FlhF	Ga0074115_12149	Ga0076813_12352	TevJSym_aw00500	Rifp1Sym_am00070	Rifp2Sym_aa00710
		motA	Flagellar motor rotation protein motA	Ga0074115_12143	Ga0076813_1553	TevJSym_aw00430	Rifp1Sym_am00130	Rifp2Sym_aa00640
				Ga0074115_12447	Ga0076813_164510	TevJSym_av00100	Rifp1Sym_av00240	Rifp2Sym_ci00070
		motB	Flagellar motor	Ga0074115_10496	Ga0076813_13374	TevJSym_bi00240	Rifp1Sym_cj00080	Rifp2Sym_ai00430

D.32

Back to Appendix D

Table D.5 Genes of particular interest. Modified or complemented from Gardebrecht *et al.* (2011). n.d.: Not Detected; no sequences were found in this genome assembly; n.a.: Not Annotated; orthologous sequence was found but no genes were detected in this genome assembly. **(Continued)**

				Locus-tag in Genbank accessions				
Categorie	**Sub-Categorie**	**Gene**	**Function**	***Ridgeia* 1 symbionts**	***Ridgeia* 2 Symbionts**	***Tevnia* symbiont**	***Riftia* 1 symbiont**	***Riftia* 2 symbiont**
			rotation protein motB					
				Ga0074115_12448	Ga0076813_15591	TevJSym_av00090	Rifp1Sym_av00250	Rifp2Sym_ci00060
				Ga0074115_12142	Ga0076813_1552	TevJSym_aw00420	Rifp1Sym_am00140	Rifp2Sym_aa00630
	8.3. Cell attachment	FliK	Flagellar hook length control protein	Ga0074115_11949	Ga0076813_12961	TevJSym_ac01170	Rifp1Sym_da00090	Rifp2Sym_ay00290
		hlyIII	Hemolysin III like protein	Ga0074115_10985	Ga0076813_11011	TevJSym_ar00200	Rifp1Sym_bh00190	Rifp2Sym_ab00180
		TlyC	Putative Hemolysin C	Ga0074115_1751	Ga0076813_14012	TevJSym_bc00400	Rifp1Sym_ax00290	Rifp2Sym_aq00130
		hlyD	Type I secretion protein, HlyD family	Ga0074115_101107	Ga0076813_16091	TevJSym_ao00270	Rifp1Sym_ac00300	Rifp2Sym_dc00070
				Ga0074115_10993	Ga0076813_1304	TevJSym_ar00280	Rifp1Sym_bh00120	Rifp2Sym_ab00250
				Ga0074115_1373	Ga0076813_11203	TevJSym_az00030	Rifp1Sym_bt00160	Rifp2Sym_aa00180
		kamA	Lysine 2,3 aminomutase	Ga0074115_12934	Ga0074115_11543	TevJSym_an00300	Rifp1Sym_ey00040	Rifp2Sym_ar00010
		cvpA	Colicin V production protein	Ga0074115_12444	Ga0076813_16457	TevJSym_av00130	Rifp1Sym_av00210	Rifp2Sym_ci00100
			Outer membrane adhesin like protein	Ga0074115_13915	Ga0076813_12205	TevJSym_ae00280	Rifp1Sym_gp00020	Rifp2Sym_fz00070
				Ga0074115_11649	Ga0076813_16472	TevJSym_ac00140	Rifp1Sym_ag00450	Rifp2Sym_au00150
			Hyalin	Ga0074115_11167	Ga0076813_11761	TevJSym_bf00250	Rifp1Sym_bg00190	Rifp2Sym_bl00010 ; Rifp2Sym_bl00020
			Fibronectin Type III domain protein	Ga0074115_10729	Ga0076813_12454	TevJSym_au00400	Rifp1Sym_bo00250	Rifp2Sym_al001101 ; Rifp2Sym_al00120
			Cell wall associated biofilm protein	Ga0074115_1301	Ga0076813_12431	TevJSym_ay00010	Rifp1Sym_ah00020	Rifp2Sym_at000101 ; Rifp2Sym_at00020