



Gene Flow among Candidatus Endoriftia persephone Symbionts of Riftia pachyptila across Hydrothermal Vent Systems in the Guaymas Basin

Citation

Rudawsky, Sarah Elizabeth. 2024. Gene Flow among Candidatus Endoriftia persephone Symbionts of Riftia pachyptila across Hydrothermal Vent Systems in the Guaymas Basin. Master's thesis, Harvard University Division of Continuing Education.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37378204>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.

Please share how this access benefits you. [Submit a story](#)

Gene Flow among *Candidatus* Endoriftia persephone Symbionts of *Riftia pachytila* across
Hydrothermal Vent Systems in the Guaymas Basin

Sarah E. Rudawsky

A Thesis in the Field of Biology
for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2024

Abstract

Deep-sea hydrothermal vents, which are commonly found along mid-ocean ridge systems and spreading centers in the deep ocean, are home to many endemic species of animals. While much of the deep sea is limited in primary production and readily available nutrients, mutualism between animal hosts and chemolithoautotrophic symbiotic bacteria is prevalent in hydrothermal vent systems, resulting in ecosystems which are brimming with life. Among the endemic species which can be found at hydrothermal vents is *Riftia pachyptila*, a siboglinid tubeworm species, which harbors the chemolithoautotrophic symbiotic bacterium, *Candidatus Endoriftia persephone*.

Hydrothermal vents represent island-like habitats that are often hundreds to thousands of kilometers apart, and it is still poorly understood how vent populations are connected across these vast geographic distances. Understanding the genetic connectivity among organisms leads to insights about population tolerance to change, including anthropogenic activities, such as extractive deep-sea mining. Although previous research has explored the level of dispersal and genetic connectivity among animal species at hydrothermal vent systems, little is known about the movement and gene flow of symbionts across these systems. Further, this is the first study of genetic connectivity of *Ca. E. persephone* populations across hydrothermal vent systems in the Guaymas Basin.

This study examined the level of gene flow across populations of *Ca. E. persephone* between distinct hydrothermal vent sites of the Guaymas Basin, Gulf of California, Mexico. Genetic diversity of *Ca. E. persephone* was investigated through

metagenomic sequencing of the symbionts' genomic DNA (gDNA). Symbiont populations were recovered from sixty-one *R. pachyptila* specimens collected from five hydrothermal vent sites across the Northern Trough and Southern Trough regions of the Guaymas Basin. Downstream population genomic analyses included examination of genetic variants and assessment of gene presence and absence. This study reports evidence of a high degree of *Ca. E. persephone* population connectivity across the Guaymas Basin. While further research is required to fully understand the drivers for this genetic homogeneity, these results do infer population stability and robustness.

Author's Biographical Sketch

Sarah Rudawsky is a passionate scientist with a diverse background in research. Currently, Sarah works at Foundation Medicine, Inc., where she develops assays for comprehensive genomic profiling of cancer patient tissues. Her current work involves next-generation sequencing (NGS), which has allowed her to gain a strong background in genetics and genomics, and an appreciation for the diverse applications of these tools across scientific fields.

Prior to moving to Boston, MA in 2015 and transitioning to work in NGS, Sarah worked at the Ohio Agricultural Research and Development Center (OARDC) at The Ohio State University (OSU) in agricultural and entomology research. Her undergraduate thesis focused on the ecological impacts of invasive emerald ash borer beetles (EAB), *Agrilus planipennis*, in southeastern Michigan through studying the impacts of EAB induced forest habitat disturbance on spider species abundance, biodiversity, and distribution.

Through her research with the Girguis Laboratory, Sarah has the opportunity to merge her interests in NGS, ecology, and population dynamics through the study of gene flow of the *Riftia pachyptila* endosymbiont, *Candidatus Endoriftia persephone*, across hydrothermal vent sites. This research is also allowing Sarah to pursue her interests in marine biology and ecology. Through this work, Sarah hopes to contribute to the scientific community through insights into the connectivity of these delicate systems.

Dedication

To my family, who grew by one this year.

Acknowledgments

I'd like to acknowledge and sincerely thank my co-thesis directors, Dr. Peter Girguis and Dr. Jessica Mitchell of Harvard University. Their generous support and guidance throughout this process have been vital for my ability to execute my thesis and grow as a scientist. I am additionally profoundly grateful to Dr. Corinna Breusing of the University of Rhode Island for her collaboration and mentorship throughout this project. Her expert advice and insights have been invaluable and are truly appreciated.

I'd also like to extend my gratitude to the Girguis Lab for sharing their time and space with me, especially Jennifer Delaney and Jennifer Thomson for their help in all practical matters in the lab. Additionally, I am appreciative to Yunha Hwang for her advice with navigating the research cluster.

I am thankful to Foundation Medicine Inc., for the generous investment of resources in support of this degree. Lastly, I'd like to thank my parents and family for their love and continued support. I would especially like to thank my husband, Nathan Keck, for his unwavering faith in me and encouragement as I returned to school to pursue something new.

Table of Contents

Author’s Biographical Sketch.....	v
Dedication.....	vi
Acknowledgments.....	vii
List of Tables	x
List of Figures	xi
Chapter I. Introduction.....	1
Hydrothermal Vents and the Deep Sea.....	1
The Guaymas Basin	2
Introduction to <i>Riftia pachyptila</i> and <i>Candidatus</i> Endoriftia persephone	
Symbionts	3
Gene Flow, Dispersal, and Previous Research	7
Study Aims.....	11
Definition of Terms.....	12
Chapter II. Materials and Methods	16
Sample Collection.....	16
<i>Riftia pachyptila</i> Dissection.....	21
DNA Extraction & Purification, Library Preparation, and Sequencing	23
Population Genomics	25
Analysis Pipeline and Statistical Analysis.....	25
NGS Analysis Pipeline	25

Symbiont Genetic Diversity between Hydrothermal Vent Sites	27
Chapter III. Results	28
Sample Collection, Dissection, and DNA Extraction.....	28
Metagenomic Sequencing Results	33
Inter-site Genetic Diversity.....	39
Impact of Host Physiology and Environment on Symbiont Genetic Diversity	51
Chapter IV. Discussion	56
Considerations for High Degree of Gene Flow in the Guaymas Basin	57
Potential for Deep-Sea Mining to adversely impact the <i>Candidatus</i> Endoriftia persephone Symbiont Communities in the Guaymas Basin.....	59
Research Limitations	61
Future Research Directions.....	61
Appendix 1. DNA Extraction Method Validation: DNA QC.....	63
Appendix 2. Sample E-gel and PicoGreen Quality Control Results from Psomagen, MACQCREPORT_V171016.....	67
Appendix 3. Sequencing Pilot Study	72
Appendix 4. Sequencing Coverage Calculations.....	74
Appendix 5. Sequencing Raw Data Report	76
Appendix 6. Uncharacterized Genes with Variable Presence and Absence across Hydrothermal Vent Sites.....	78
Appendix 7. PCoA of Nucleotide Counts and Haplotypes across Host Size, Intra-site Investigation.....	85
References.....	90

List of Tables

Table 1. Sample Collection Summary.	18
Table 2. Sample Condition, Physiology, and Lab Metrics.	30
Table 3. Population Divergence Across Hydrothermal Vent Sites.....	43
Table 4. <i>Candidatus</i> Endoriftia persephone Gene Annotations for Known Genes across Hydrothermal Vent Sites, Genes Universally Present Removed.....	45
Table 5. Environmental Conditions by Cruise and Dive ID.....	52
Table 6. NanoDrop Absorbance Ratios.	64
Table 7. Host to Symbiont DNA Proportions.....	73
Table 8. Required Output for Host and Symbiont Metagenome Sequencing.....	75
Table 9. Sequencing Raw Data Metrics.....	76
Table 10. <i>Candidatus</i> Endoriftia persephone Gene Annotations for Uncharacterized Genes across Hydrothermal Vent Sites, Genes Universally Present Removed.....	78

List of Figures

Figure 1. 2021 Guaymas Cruise Collection.....	17
Figure 2. <i>Riftia pachyptila</i> collected at 2032 meters with the ROV <i>SuBastian</i> during FK190211 Research Cruise Dive D233 (© Schmidt Ocean Institute).	18
Figure 3. 2019 Guaymas Cruise Collection.....	19
Figure 4. 2021 and 2019 Guaymas Basin Collection Sites.....	20
Figure 5. <i>R. pachyptila</i> Trunk Cross-Section Illustration depicting: Trophosome (A), Gonads (B), Blood Vessels (C), and Coelom or Body Wall Tissue (D).	23
Figure 6. Oneway Analysis of Sample Storage Condition vs. Read Count.....	34
Figure 7. Oneway Analysis of Sample Storage Condition vs. Q30 (%).	35
Figure 8. Oneway Analysis of Host Size vs. Read Count.	36
Figure 9. Oneway Analysis of Host Size vs. Q30 (%).	37
Figure 10. Q30 (%) for Samples Removed from Downstream Analysis vs. Samples Included in Population Genomics Analysis.....	38
Figure 11. PCoA of Nucleotide Counts across Hydrothermal Vent Sites (i.e., Dive ID).40	
Figure 12. PCoA of Haplotypes across Hydrothermal Vent Sites (i.e., Dive ID).	41
Figure 13. Heatmap of Present and Absent Genes across Hydrothermal Vent Sites by Sample.....	42
Figure 14. Heatmap of Genes Present and Absent across Hydrothermal Vent Sites by Sample, Universally Present Genes Removed.....	44
Figure. 15. PCoA of Nucleotide Counts across Collection Areas.	51
Figure 16. PCoA of Haplotypes across Collection Areas.....	53

Figure 17. PCoA of Nucleotide Counts by Host Size.....54

Figure 18. PCoA of Haplotype by Host Size.....55

Figure 19. Agilent Genomic DNA ScreenTape Analysis.....65

Figure 20. Agilent RNA ScreenTape Analysis.....66

Figure 21. PCoA of Nucleotide Counts across Host Size for Dive ID J2-1390.....86

Figure 22. PCoA of Haplotypes across Host Size for Dive ID J2-1390.....87

Figure 23. PCoA of Nucleotide Counts across Host Size for Dive ID J2-1392.....88

Figure 24. PCoA of Haplotypes across Host Size for Dive ID J2-1392.....89

Chapter I.

Introduction

Deep-sea hydrothermal vents are ecosystems with a significant amount of primary production and many endemic species (Van Dover et al., 2018). These vents play an important role in supplying nutrients in the oligotrophic depths of the pelagic sea (Le et al., 2017). Many of the animals which inhabit hydrothermal vents are dependent on chemolithoautotrophic bacterial symbionts, which utilize chemicals released in vent effluent for nutrient synthesis through processes such as sulfur oxidation and carbon fixation (Cavanaugh et al., 1981; Cavanaugh, 1983; Cavanaugh, 1994). The chemolithoautotrophic bacterium, *Candidatus Endoriftia persephone* (*Ca. E. persephone*), is a keystone microbial species and endosymbiont of the giant siboglinid tubeworm species, *Riftia pachyptila* (Perez et al., 2021). Together, this association supports remarkably fast growth rates and primary production in these ecosystems (Childress and Girguis, 2010).

Hydrothermal Vents and the Deep Sea

While most of the deep sea is a stable low-energy environment, hydrothermal vent systems are an exception. Hydrothermal vent effluents are anoxic, and contain reduced chemical compounds such as sulfide, hydrogen, and methane; as well as the potentially toxic heavy metals copper, cadmium, and lead (Dick, 2019). Vent effluent temperatures are typically between 250 - 350°C (though some vents can reach temperatures of approximately 500°C), which is in sharp contrast to temperatures of 2° -

4°C in the oxygen rich waters which immediately surround vents (López-García et al., 2002; Dick, 2019). The rapid mixing of extremely hot vent effluent and cold sea water at hydrothermal vent sites results in mineral deposits that cause formation of physical structures and create an ecosystem that is high in potential energy from redox reactions. The diversity in effluents and deposits results in a range of hydrothermal vent habitats, including iron sulfide rich “black smokers”, mineral dense “white smokers”, and both alkaline and carbonate vents. Hydrothermal vents are found globally along mid-ocean ridges, back-arc spreading centers, as well as concurrently with volcanoes and seamounts (Dick, 2019).

Hydrothermal vents provide profoundly important services which impact the water column and beyond. Geochemical output from hydrothermal vents and symbiosis with chemolithoautotrophic bacteria support abundant and diverse local communities of animals which vary regionally and globally (Beinart et al., 2012; Vic et al., 2018; Dick, 2019). Chemolithoautotrophs are responsible for primary production in hydrothermal vent environments by providing usable nutrients to their host through oxidation of reduced chemical compounds that yield energy for carbon fixation (Cavanaugh et al., 1981; Cavanaugh, 1994; Beinart et al., 2012; Dick, 2019). Further, geochemical outputs from hydrothermal vents influence oceanic heat and chemical budgets (Vic et al., 2018); as well as provide a source of reduced chemical compounds such as iron and manganese throughout the water column (Dick et al., 2013).

The Guaymas Basin

The Guaymas Basin is a rift basin and relatively young spreading center located centrally in the Gulf of California, Mexico (Rona, 1984; Aragón-Arreola et al., 2005).

The Guaymas Basin is approximately two-hundred and forty kilometers long and sixty kilometers wide (Geilert et al., 2018). This basin is organized into the Northern Trough and Southern Trough regions, separated by the central Guaymas Transform Fault. The hydrothermal vents of the Guaymas Basin in both the Northern and Southern Trough regions are located along the ridge axis (Geilert et al., 2018). While the hydrothermal vents in the southern Guaymas Basin have been studied extensively, since first reported by Lonsdale and Becker (1985), the hydrothermal vent systems of the northern Guaymas Basin were more recently discovered by Berndt et al. (2016).

The Guaymas Basin is rich in sediment deposits, reaching hundreds of meters in depth, resulting from a productive water column, vent deposits, and erosion of organic matter from the Mexican coast (Geilert et al., 2018). As a result of these sediments, Guaymas vent effluents are rich in methane and other organic compounds. Additionally, Guaymas vent effluents contain a helium isotope signature, indicating that these effluents are the result of contact with mid-ocean ridge basalt (Berndt et al., 2016; Geilert et al., 2018). These hydrothermal vents also release sulfide, as typical of basalt-associated vent systems (Rimskaya-Korsakova et al., 2021). While physical structures do vary between hydrothermal vent sites of the southern and northern Guaymas Basin (Ondréas et al., 2018; Teske et al., 2021), the geochemical composition and endmember temperature of vent effluents from the Northern and Southern Troughs of the Guaymas Basin are similar (Geilert et al., 2018).

Introduction to *Riftia pachyptila* and *Candidatus Endoriftia persephone* Symbionts

Among the unique species which inhabit hydrothermal vent systems is the giant siboglinid tubeworm, *Riftia pachyptila*, a charismatic species which can grow to two

meters in length (Jones, 1981). *Riftia pachyptila* was first collected by Corliss and Ballard (1977) through their exploratory dives in the HOV *Alvin*, at the Galápagos Rift and first described by Jones (1981). This species is distributed in the Pacific Ocean from approximately 27°N, 110°W to 32°S, 110°W, and has been found at bathymetric depths of approximately 1900 m – 3000 m (Vrijenhoek, 2010; Karaseva et al., 2016). Specimens have been observed at hydrothermal vents of the Galápagos Rift and East Pacific Rise (EPR) spreading centers (Cavanaugh et al., 1981), as well as the Guaymas Basin, Gulf of California, Mexico. Tubeworm species serve as a foundation for community structure in hydrothermal vent environments through their formation of dense clusters that provide habitat for other species (Sato and Sasaki, 2021).

Riftia pachyptila are dioecious organisms with distinctive physiology adapted for their environment and relationship with a chemolithoautotrophic endosymbiont. This species lacks a mouth and gut (Cavanaugh et al., 1981; Jones, 1981); and the body of *R. pachyptila* consists of a tentacular plume, a collar like vestimentum, the trunk, and posterior opisthosome (Jones, 1981). Additionally, *R. pachyptila* is housed within a chitinous tube, into which it is capable of retreating completely when disturbed (Tunnicliffe et al., 1989). The lamellae-lined tentacular plumes of *R. pachyptila* are situated anteriorly, supported by the obturaculum, and are the sites of reduced sulfur uptake and gas exchange (Jones, 1981; Tunnicliffe et al., 1989; Rimskaya-Korsakova et al., 2021). The trunk consists of the vascularized coelomic cavity, which encloses gonads, as well as bacteria hosting trophosome tissue organized in lobules (Cavanaugh et al., 1981; Jones, 1981). The trophosome extends along the length of the worm's trunk. The segmented opisthosome serves to anchor the worm and secretes the chitinous tube

(Miyamoto et al., 2014). *R. pachyptila* possesses phenotypic plasticity and allometric growth patterns which appear to be independent of genotypic differences in *R. pachyptila* populations (Black et al., 1994; Rinskaya-Korsakova et al., 2021) and may be the result of epigenetic mechanisms that regulate gene expression in response to environmental influences.

Riftia pachyptila are completely dependent on chemoautotrophic symbiosis with the sulfur-oxidizing symbiotic bacterium, *Candidatus Endoriftia persephone*, which provides usable nutrients and energy in exchange for a stable environment within its host (Cavanaugh et al., 1981; Rinskaya-Korsakova et al., 2021). Chemoautotrophic symbiosis, in which an animal is reliant on their sulphur-oxidizing symbiotic bacteria as its sole source of nutrition, was first discovered by Cavanaugh et. al., (1981) and Felbeck et. al., (1981) and fundamentally changed how we view biology.

Much of *R. pachyptila*'s physiology is specially adapted for and devoted to providing a stable environment for their endosymbiotic bacteria through: A) extensive vascularization of trophosome tissue, B) specialized hemoglobin that can bind both oxygen and sulfide, and C) an abundance of carbonic anhydrase that concentrates dissolved inorganic carbon (DIC), the substrates needed for autotrophy (Cavanaugh et al., 1981; Arp and Childress, 1983; Arp et al., 1987). Additionally, *R. pachyptila* protects its endosymbiont from oxidative damage (Hinzke et al., 2019). These endosymbiotic bacteria further benefit from an environment within the host that is devoid of competition from free-living bacteria, since *Ca. E. persephone* are obtained horizontally from the environment only at the larval stage of the *R. pachyptila* lifecycle (Nussbaumer et al., 2006; Polzin et al., 2019; Sato and Sasaki, 2021).

Another benefit to the symbiont, and foundation of this relationship, is that the host provides access to the substrates needed for carbon fixation via chemoautotrophy (Cavanaugh et al., 1981). Sulfide, carbon dioxide, and oxygen are taken up through the lamellae in tentacular crowns of *R. pachyptila*, then enter the bloodstream and trophosome, a dedicated organ which hosts the symbiotic *Ca. E. persephone* (Rimskaya-Korsakova et al., 2021). The sole function of the trophosome is to house these sulphur-oxidizing symbionts.

The lifecycle, growth, and morphology of *R. pachyptila* appear to be highly dependent on both environment and initial acquisition of the *Ca. E. persephone* endosymbiont. The lifecycle of tubeworms consists of free-living larval and sessile adult stages (Nussbaumer et al., 2006; Sato and Sasaki, 2021). Nussbaumer et al. (2006) proposed that post settlement of *R. pachyptila* larvae on substrate, the skin of the larvae is infected with symbiotic *Ca. E. persephone* bacteria that subsequently colonize the mesoderm. The colonization of larvae with symbiotic bacteria initiates the process of transition from juvenile to adult, including the loss of larval digestive system, apoptosis of host cells, and formation of the trophosome (Nussbaumer et al., 2006).

Ca. E. persephone are Gram-negative bacteria of the class Gammaproteobacteria. Metagenomics indicate that the *Ca. E. persephone* are mixotrophs capable of adapting to either host associated or free-living life stages, through maintenance of genes for both heterotrophic and host-associated metabolic pathways (Robidart et al., 2008). Additionally, metagenomic analysis reveals a notable portion of the *Ca. E. persephone* genome is dedicated to chemotaxis mechanisms which may allow free-living *Ca. E. persephone* to identify and reach suitable substrates or hosts. These genes include those

involved in motility, such as a functional flagellum, and chemoreception (Robidart et al., 2008; Bright and Bulgheresi, 2010; De Oliveira et al., 2022).

While *R. pachyptila* is specifically associated with *Ca. E. persephone*, this endosymbiont associates with many different species of tubeworms and little is still understood about strain or subpopulation diversity across these associations. Evidence supports symbiotic relationships between *Ca. E. persephone* and the vent tube worm species *Ridgeia piscesae*, *Escarpia spicata*, *Tevnia jerichonana*, and *Oasisia alvinae*, in addition to a relationship with *R. pachyptila* (Perez and Juniper, 2016).

Gene Flow, Dispersal, and Previous Research

Riftia pachyptila is a monospecific genus that is found at deep sea hydrothermal vents in the Eastern Pacific Ocean. Their northernmost range is ~27° N in the Guaymas Basin to ~32°S in the southern EPR, near Easter Island (Vrijenhoek, 2010). This means that *R. pachyptila* populations span approximately forty degrees latitude, or about five-thousand kilometers, a distance that may be markedly greater as vents are not entirely linearly distributed (Karaseva et al., 2016). In addition, vents that are found along the mid-ocean ridge are not contiguous as they are broken up by transverse faults (Young et al., 2008). As such, the connectivity between the *R. pachyptila* populations found at hydrothermal vents varies among sites, with good spatial and geochemical connectivity between some sites and very poor connectivity between others (Vrijenhoek, 2010).

To better understand the connectivity between populations, previous work has investigated the level of genetic connectivity between *R. pachyptila* found at EPR vents and reported that gene flow exists following a “stepping-stone model” of dispersal across

hydrothermal vent sites (Black et al., 1994; Coykendall et al., 2011). The stepping-stone model indicates that closer populations share more genetic traits than distant populations, in contrast to modes of dispersal over great distances, such as the “island model” of dispersal (Vrijenhoek, 1997). In the Guaymas Basin, collection of late forms of polychaete larva at 100 - 200 m above the bottom of the Southern Trough supports that some degree of host dispersal does occur between hydrothermal vent sites in this region (Wiebe et al., 1988); however, the approximate thirty-eight-day lifespan of *R. pachyptila* larvae (Marsh et al., 2001) may be a limiting factor to dispersal across vent sites and influence mode of dispersal for this organism.

In contrast to the host, very little is known about the modes of dispersal and genetic subdivision of the symbiont. Additionally, dispersal of *R. pachyptila* and *Ca. E. persephone* are likely uncoupled since the symbiont is horizontally transmitted to the host at the larval stage after the larva has settled on substrate (Nussbaumer et al., 2006). The selective pressure for genetic differentiation deviates between symbionts which are vertically or horizontally transmitted to their host. Vertical transmission, in which symbionts are passed from parent to offspring, provides opportunity for genetic drift and co-speciation of symbiont and host (Stewart and Cavanaugh, 2005). Further, in cases of vertical transmission, symbiont dispersal would be coupled with movement of host larvae. In contrast, horizontally transmitted symbionts are obtained through the environment with each new generation of host (Bright and Bulgheresi, 2010).

While no data on genetic connectivity of *Ca. E. persephone* populations associated with *R. pachyptila* currently exists prior to this study, several studies have examined dispersal, gene flow, and population structure of this symbiont associated with

other host species. Previous research on *Ca. E. persephone* populations across tubeworm hosts and hydrothermal vent regions support genetic differentiation between symbiont associated with *R. piscesae* in the Juan de Fuca Ridge (JdFR) compared to *R. pachyptila* and *T. jerichonana* associated symbiont from the EPR (Perez and Juniper, 2016). The JdFR and EPR regions were isolated by tectonic activity approximately 30 million years ago, as such Perez and Juniper (2016) theorize that deviations in *Ca. E. persephone* population structure across these regions can likely be attributed to genetic drift. No significant genetic deviations were observed across *Ca. E. persephone* populations from distinct sites within the EPR; however, there is evidence of symbiont subpopulations associated with *R. pachyptila* and *T. jerichonana* within the EPR (Perez and Juniper, 2016). Deviation in genetic structure between *Ca. E. persephone* populations associated with *R. pachyptila* and *T. jerichonana* has previously been demonstrated by Meo et al. (2000) as well.

In a 2021 study, Perez et al. utilize the CRISPR array to investigate population structure of *Ca. E. persephone* associated with *R. piscesae* in the JdFR. They observed divergence in symbiont genetic populations associated with circulation patterns which limit connectivity across regional rifts; however, low local population diversity was observed within regions. The polymorphisms observed by Perez et al. (2021) in the CRISPR array were related to spacer deletions, no new or unique spacers were observed at the leading end of the CRISPR array in *Ca. E. persephone*. Thus, the authors suggest that the immune response in *Ca. E. persephone* may have been lost over time; and for *Ca. E. persephone*, CRISPR array analysis may be more meaningful for observing evolutionary trends across millions of years rather than examining recent and current

gene flow across populations. Perez et al. (2021) additionally investigated SNPs in the housekeeping genes *lpxA*, *pleD* and *tufB*, and observed overall less genetic diversity through these SNPs; however, their analysis supported the patterns of population structure observed with CRISPR array analysis.

The study of virus dispersal across distinct hydrothermal vent sites can provide further insights into microbial movements in the deep ocean. Investigation of viral population genomics in the Caribbean Sea and Axial Seamount in the Pacific Ocean revealed low genetic connectivity across hydrothermal vent fields, resulting in strains of virus endemic to specific hydrothermal vent sites (Thomas et al., 2021).

Microbial and animal dispersal in the deep ocean has proven to be highly variable and dependent on geographical, geochemical, and biological factors. Previous research has shed light on broader evolutionary trends of *Ca. E. persephone* across regions; however, ongoing gene flow across sites is still poorly understood. Further, investigations of *Ca. E. persephone* dispersal have previously been conducted in the JdFR and EPR; and no dedicated population genomics studies of *R. pachyptila* have previously been conducted. As such, it would be relevant to investigate symbiont genetic connectivity in the Guaymas Basin, where influences of geographical barriers and deep-sea circulation vary from the EPR and JdFR. It would additionally be pertinent to examine the level of genetic connectivity between populations of *Ca. E. persephone* associated with *R. pachyptila*, as understanding of strain diversity between *Ca. E. persephone* endosymbionts associated with different hosts is still poorly understood.

Study Aims

Through this investigation, I aim to understand the level of connectivity and gene flow among *Ca. E. persephone* populations across distinct hydrothermal vent sites in the Guaymas Basin. I hypothesize that currents in the Gulf of California will aid in the dispersal of free-living or host-liberated *Ca. E. persephone* across vent sites (Vic et al., 2018); thus, contributing to gene flow across *Ca. E. persephone* populations. Further, I expect that the level of genetic connectivity will be dependent upon distance between vent sites according to a “stepping-stone model”, as previously described by Black et al. (1994); however, I contend that the “island-model” of dispersal may be relevant for free-living symbionts and some degree of genetic connectivity will be present across distant populations. If this hypothesis is not supported, potential barriers to gene flow across vent sites may include geographic barriers, distance, patterns of deep-sea currents (Perez et al., 2021), and localized adaptation to geochemical structure of discrete sites (Beinart et al., 2012).

This study will make novel contributions to understanding the genetic connectivity and dispersal of a keystone chemolithoautotrophic endosymbiont, *Ca. E. persephone*, in the Guaymas Basin hydrothermal vent system. This study will additionally utilize modern genomic techniques of metagenomic whole genome sequencing and SNP analysis to provide a better understanding about *Ca. E. persephone* genetic diversity. The findings of this study will lead to further understanding of hydrothermal vent ecology and symbiont population structure in the Guaymas Basin, which can lead to translational insights about the ecology of other vent systems.

Finally, based on the findings of this study, I will discuss whether deep-sea mining could potentially have an adverse impact on distinct *Ca. E. persephone* populations and regional ecology. Understanding the extent of genetic connectivity between populations of symbionts across hydrothermal vent sites will provide critical evidence for gauging the impact of major extractive activities on adjacent and distant vent sites. It may be inferred that well-connected populations of *Ca. E. persephone* will be more resilient to the impacts of deep-sea mining, due to the known importance of genetic diversity on metapopulation robustness; however, localized extinction at hydrothermal vent sites may lead to overall loss of genetic diversity on a regional scale depending on the level or reach of gene flow across local populations (Black et al., 1994; Orcutt et al., 2020). As Van Dover et al. (2018) poignantly describe, active hydrothermal vent ecosystems are “‘Small Natural Features’ with ecological importance disproportionate to their size”. Developing a fuller understanding of vent system ecology is essential for protecting these habitats.

Definition of Terms

Allometry: Refers to a scaled relationship between body size and morphological traits during growth (Shingleton, 2010).

Annelid: Segmented worms of the Phylum Annelida. *Riftia pachyptila* are tube-dwelling annelids.

Anoxic: Severely reduced in or lacking oxygen.

Benthic: Refers to the sea floor, bottom of the ocean.

Chemolithoautotroph: Describes microbes which synthesize nutrients from chemicals derived from the environment, specifically the bedrock of the Earth (The

Cambrian Foundation, N.D.).

Clustered regularly interspaced short palindromic repeats (CRISPR) array: Genetic sequence possessed by many bacteria and majority of archaea species which allows for an adaptive immune response to phage infection. The CRISPR array is a highly modified, heritable gene locus which serves as a memory of past phage challenges.

Dioecious: Separation of sexes, reproductive organs are separated by male and female individuals.

Dissolved Inorganic Carbon (DIC): Describes the sum of carbon dioxide (CO_2), bicarbonate (HCO_3^-), and carbonate (CO_3^{2-}) in aqueous solution, such as sea water (Carlson et al., 2001).

Dissolved Organic Matter (DOM): A diverse mixture of hydrocarbon structures with attached functional groups, these structures can be aromatic or aliphatic and broadly range in size (Leenheer and Croué, 2003).

Epibiotic: Describes an organism which lives on the surface of another organism. Epibiotic organisms can often include bacteria or fungi.

Genetic connectivity: In this study, genetic connectivity is defined as shared genomic traits across populations. Genetic connectivity is established through population dispersal and rate of gene flow.

Geochemical: Refers to the composition of vent endmember fluids. In the context of this study, special focus is given to sulfide and oxygen concentration.

Haplotype: DNA polymorphisms (i.e., genetic variations) which are typically inherited together. These may be found on the same chromosome.

Horizontal transmission: The uptake of free-living symbiotic bacteria from the environment into the host. This may occur in germ cells, during development, or in larval/juvenile stages (Bright and Bulgheresi, 2010).

Human operated vehicle (HOV): Submersible which is operated by and carries human passengers.

in situ: Describes “on site”, such as observations or data collection.

Keystone species: An organism which is crucial for ecosystem function, and whose loss will lead to loss of biodiversity and disruption of ecosystem services (Mills et al., 1993).

Metagenomics: The investigation of nucleic acid sequences of organisms derived from a bulk source or sample.

Multiplexed: NGS approach of including multiple samples in a single assay through adhering sample specific molecular barcodes to each sample.

Next-generation sequencing (NGS): High-throughput sequencing technique based on massive-parallel sequencing of genomic fragments which are then aligned by overlapping sequences or regions of homology to generate the complete sequence. This technique can be utilized for whole genome sequencing or specific genomic targets.

Oligotrophic: Nutrient poor environment.

Pelagic zone: The open sea, all ocean outside of benthic zone (i.e., the water column).

Remotely Operated Vehicle (ROV): Unoccupied submersibles which can be controlled from a remote location. *Jason II* and *SuBastian* are ROVs used for operations in this study.

Siboglinid: Tube-dwelling polychaete worms of the family Siboglinidae.

Single Nucleotide Polymorphism (SNP): Single base pair variation in a genomic sequence. SNPs are usually benign and drive genetic diversity.

Trophosome: Symbiont hosting organ in tubeworms.

Vent effluent: Liquid and gaseous discharge from hydrothermal vents.

Chapter II.

Materials and Methods

This study investigates the genomic DNA (gDNA) of *Ca. E. persephone* endosymbionts collected from the host *R. pachyptila* through a metagenomics sequencing approach. *Riftia pachyptila* specimens were collected from hydrothermal vent sites in the Guaymas Basin, Gulf of California, an evolving rift basin north adjacent to the East Pacific Rise (EPR) (Horstmann et al., 2021). Symbiont gDNA was extracted and purified from *R. pachyptila* trophosome, and high-throughput whole genome sequencing was utilized to investigate *Ca. E. persephone* genomic diversity across hydrothermal vent sites.

Sample Collection

For this investigation, *R. pachyptila* were collected from hydrothermal vent sites of the Guaymas Basin. Collections took place during research cruise RR2107 on the research vessel, *Revelle*, from 13 November 2021 – 04 December 2021. *Riftia pachyptila* were collected at three distinct hydrothermal vent sites in the Northern Trough of the Guaymas Basin during dives with the remotely operated vehicle (ROV) *Jason II* (Figure 1).

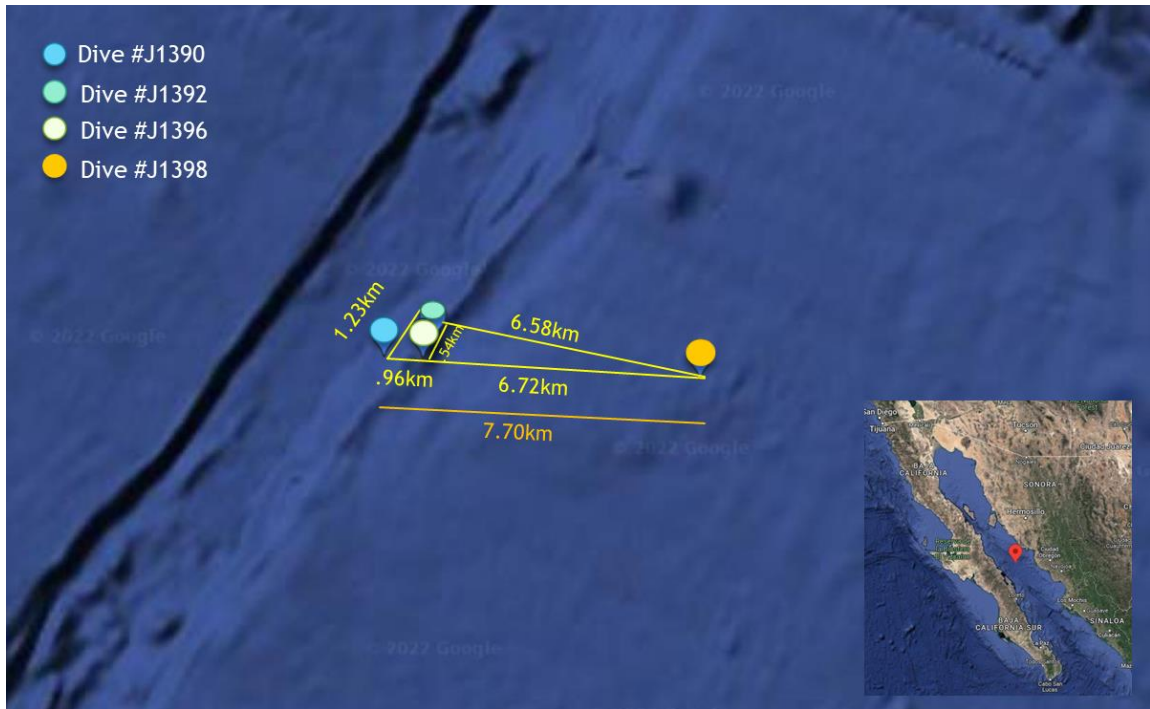


Figure 1. 2021 Guaymas Cruise Collection.

Collection sites from the 2021 RR2107 cruise in the Guaymas Basin. Dive #J2-1396 was not included in this investigation. Image created through Google Maps.

In addition to the collection of *R. pachyptila* during the 2021 RR2107 research cruise, previously collected specimens were utilized in this study to provide a more robust sample set across a broader range of the Guaymas Basin. Previous collections were conducted in 2019 on the research cruise FK190211 with the ROV *SuBastian* (Figure 2) from two distinct sites (Figure 3) in the Southern Trough of the Guaymas Basin. The 2021 and 2019 collection areas are approximately forty-three kilometers apart (Figure 4). A total of sixty-one samples of various sizes were included in this study (Table 1).



Figure 2. *Riftia pachyptila* collected at 2032 meters with the ROV *SuBastian* during FK190211 Research Cruise Dive D233 (© Schmidt Ocean Institute).

Use of an ROV provides the advantages of in-situ observations and data collection, ability to stay submerged longer than human operated vehicles (HOVs), and offers the opportunity for more scientists to participate in the dive.

Table 1. Sample Collection Summary.

Cruise Date	Cruise ID	Dive ID	Lat/ Longitude	Number of Samples
Guaymas 2021	<i>RR2107</i>	J2-1390	27.40923, -111.399083	17
Guaymas 2021	<i>RR2107</i>	J2-1392	27.41276, -111.3871717	23
Guaymas 2021	<i>RR2107</i>	J2-1398	27.40432, -111.3212689	8
Guaymas 2019	<i>FK190211</i>	231	27.01075, -111.406967	4
Guaymas 2019	<i>FK190211</i>	233	27.013717, -111.41105	9
				61

Number of samples collected across sites and collection years in Guaymas Basin. Samples from research cruise RR2107 were collected from the Northern Trough, while samples from research cruise FK190211 were collected from the Southern Trough region of the Guaymas Basin.

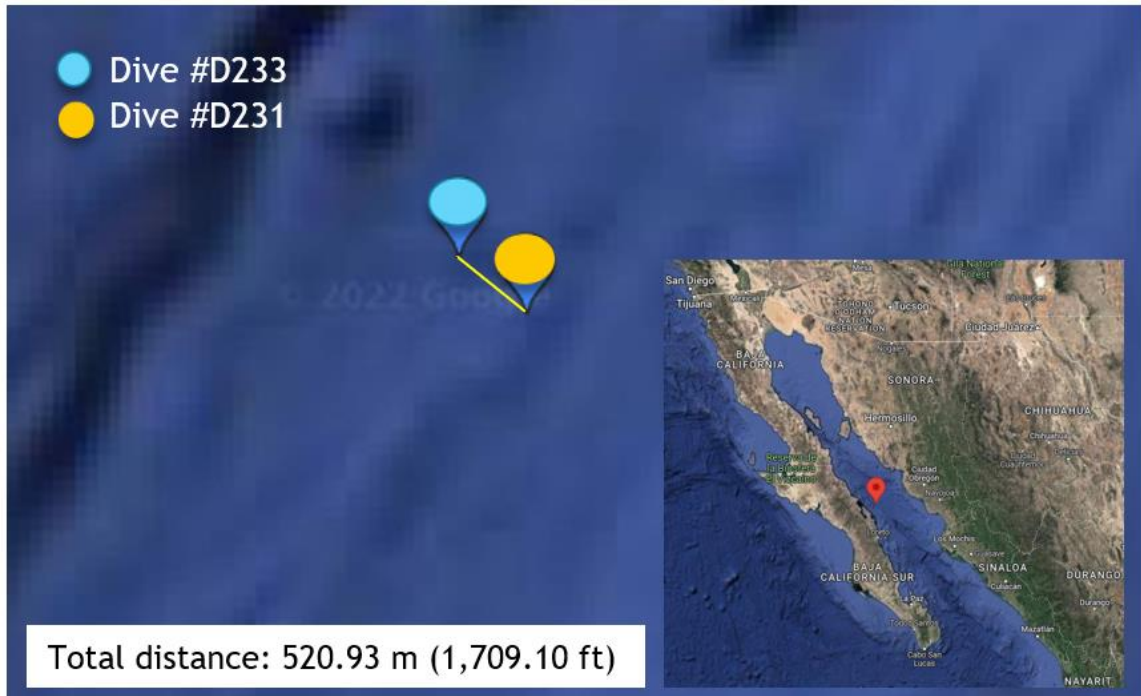


Figure 3. 2019 Guaymas Cruise Collection.

Collection sites from the 2019 FK190211 cruise in the Guaymas Basin. Image created through Google Maps.

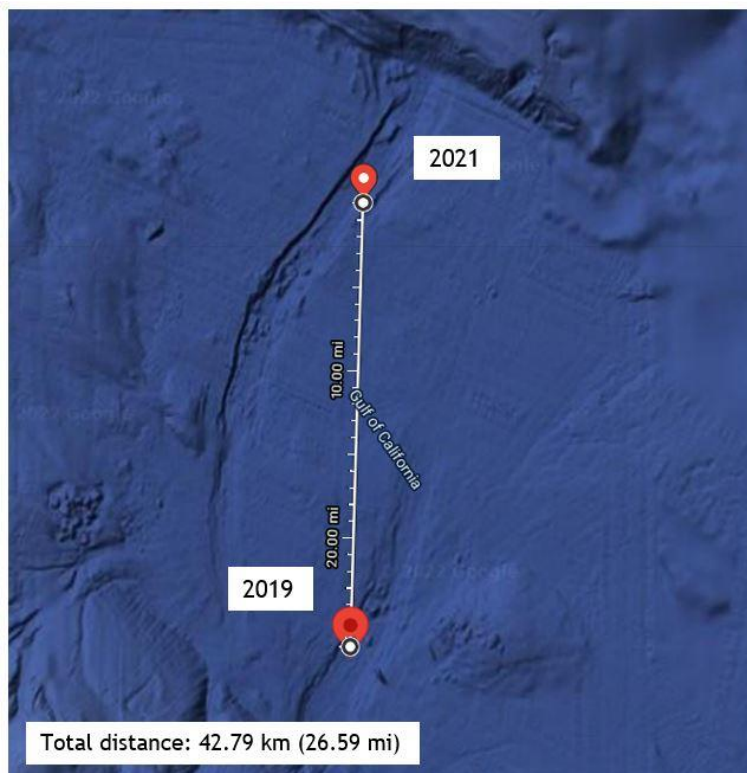


Figure 4. 2021 and 2019 Guaymas Basin Collection Sites.

Relative distance between collection areas for the 2021 RR2107 and 2019 FK190211 research cruises. Samples from 2021 were collected in the Northern Trough, while 2019 collection took place in the Southern Trough. Approximate distance between regions is forty-three kilometers. Image created through Google Maps.

All specimens were collected from active vent sites only. Post collection, *R. pachytila* were flash frozen in liquid nitrogen (N₂) and stored at -80°C until further processing for *Ca. E. persephone* DNA isolation. The trunk diameter of each specimen was measured and recorded to account for variability in *Ca. E. persephone* populations related to host size; specimens were defined as small (<15 mm), medium (15 – 25 mm), or large (>25 mm) based on mean trunk diameter across two measurements for each worm.

Riftia pachyptila Dissection

Post-collection, *R. pachyptila* specimens were frozen aboard ship and stored at -80°C until time of processing. Tubeworm handling and dissection was dependent on collection method and year. Prior to handling whole tubeworm specimens, dissection materials and trays were initially wiped down with 70% ethanol solution and allowed to dry. Dissection tools were submerged in 100% ethanol between use and flame treated prior to use. Whole worm specimens still in tubes were wiped down with 70% ethanol solution to decontaminate the surface from abundant external populations of bacteria (López-García et al., 2002); tube areas with visible bacterial mats (areas of raised gray blotches on the tube) were avoided.

A portion of worm samples had tubes removed onboard ship according to the method described by Perez et al. (2021): individuals were removed from tubes and treated with lysozyme and DNase to remove epibiotic contamination. Additionally, a subset of samples from RR2107 were dissected for trophosome collection aboard ship and remaining trunk tissue stored at -80°C for further processing (i.e., “Leftover Troph Prep” samples). Upon collection on deck, these worms were immediately placed in cold seawater, removed from tube, and externally sterilized with EtOH.

Using a fresh, sterile, and flame treated razor blade, a small section of trunk was cut from the whole worm and the tube peeled off, if present. Approximately 20 mg of trophosome was collected from the isolated section of the worm; blood vessels, gonads, and *Riftia* muscle tissue were avoided to prevent over-representation of *R. pachyptila* DNA in the collected sample (Figure 5). One to two replicates of trophosome were collected from each worm specimen; and mass - or mean mass in the case of duplicates -

was recorded for each trophosome sample. After sample dissection, excess blood and tissue material was cleaned from the dissection tray, and the tray was triple cleaned with 70% ethanol solution. Dissection tools were cleaned of blood and tissue with 70% ethanol solution, submerged in 100% ethanol, and flame treated between samples.

For a portion of the 2019 FK190211 worms, trophosomes of individual specimens were dissected from cleaned worms and stored in the DNA safe solution, RNAlater, to prevent nucleic acid degradation prior to DNA extraction. These samples were then stored at -80°C to further prevent degradation over time. The RNAlater stored trophosome samples were thawed on ice and collected from their storage tubes via dissection tools that had been 100% ethanol and flame sterilized. Approximately 20 mg of trophosome was collected and mass recorded for each sample.

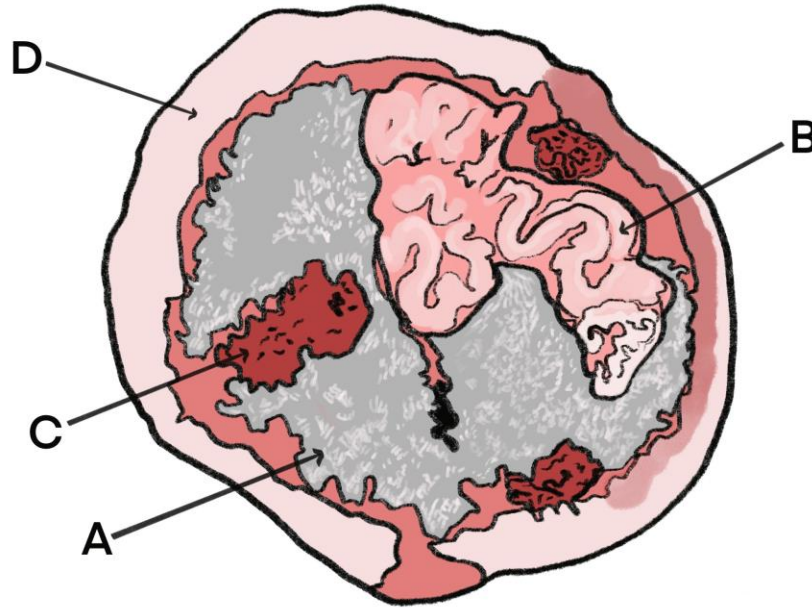


Figure 5. *R. pachyptila* Trunk Cross-Section Illustration depicting: Trophosome (A), Gonads (B), Blood Vessels (C), and Coelom or Body Wall Tissue (D).

Illustration based on R. pachyptila image taken during dissection for this study. Variability is present in structure across individual trunks and across specimens. Trophosome may vary in abundance and color, ranging from gray, brown, green, and yellow hues. Gonads were not present in all trunk sections sampled through the course of this study.

DNA Extraction & Purification, Library Preparation, and Sequencing

Genomic DNA extraction and purification was performed with use of the QIAGEN DNeasy® Blood & Tissue Kit (Cat No. 69504), following recommendations in the *DNeasy® Blood & Tissue Handbook* for DNA extraction from animal tissue, “Protocol: Purification of Total DNA from Animal Tissues (Spin-Column Protocol)”

(QIAGEN, 2020). Several customizations to the published dissection and purification recommendations used in this study follow. For tissue lysis, samples were placed in a heat block at 56°C with shaking at 500 RPM for 1.5 hours. During tissue lysis, samples were manually vortexed three to four times. To remove contaminating RNA from the samples, after the 1.5 hr lysis incubation samples were treated with 4 µl of 100 mg/ml RNase A (QIAGEN Cat No. 19101), thoroughly vortexed, then returned to incubation at 56°C with 500 RPM shaking for an additional 30 minutes. For DNA elution, 100 µl of Buffer AE was added to the column filter, incubated at room temperature for 2 minutes, then centrifuged at 6000 RCF for 1 minute. The elution step was then repeated for a total of 200 µl of eluted sample to optimize recovery of DNA from the spin column.

Samples were quantified by the Invitrogen™ Qubit™ Broad Range dsDNA Assay Kit (Cat No. Q33266) in duplicate with 5 µl of sample to reduce variability. To validate the extraction protocol, sample purity of a subset of samples was measured by a NanoDrop Spectrophotometer through which 260/230 and 260/280 absorbance values were assessed and recorded. The presence of RNA contamination was investigated for a subset of samples post RNA treatment; and extracted gDNA fragment quality was observed (Appendix 1).

Isolated DNA was sent to Psomagen for library preparation and multiplexed next-generation sequencing (NGS). At Psomagen, sample quality was assessed through Invitrogen E-gel and quantified by PicoGreen prior to moving forward with library preparation (Appendix 2). Whole genome multiplexed non-targeted libraries were prepared with the seqWell plexWell™ LP384 Kit (Cat No. PW384). Sequencing was done through the Illumina NovaSeq6000 platform: 150bp paired-end sequencing, S4 flow

cell, targeting 2.9 Gb per sample based on an initial pilot sequencing study (Appendices 3 and 4).

Population Genomics

Metagenomics were utilized to investigate *Ca. E. persephone* genomes from DNA coextracted from both host and symbiont. Non-targeted whole genome sequencing allows for a more comprehensive comparison of genomes across hydrothermal vent sites and increases sensitivity for detection of genetic diversity.

Genetic diversity was investigated as follows:

1. Symbiont genetic diversity between distinct hydrothermal vent sites.
2. Impact of host size and region on symbiont genetic diversity.

Analysis Pipeline and Statistical Analysis

NGS Analysis Pipeline

Population genomic analyses were approached through filtering symbiont and host metagenomic reads, metagenome reconstruction, *Ca. E. persephone* pangenome reconstruction, *Ca. E. persephone* variant calling, and analysis of gene content variation based on previous work by Breusing et al. (2022). For filtering of metagenomic reads, the FastQC tool was used for quality control and Trimmomatic was used for paired-end trimming (Bolger et al., 2014). Further filtering was done to remove contaminating sequences through read mappings with Bowtie2 against human (GRCh38) and PhiX reference genomes.

Post filtering of metagenomic reads, metagenomes of each individual tubeworm were assembled with metaSPADES (Nurk et al., 2017). Metagenome-assembled genomes (MAGs) of *Ca. E. persephone* were then binned from these metagenomes with Metabat2 (Kang et al., 2019), MaxBin2 (Wu et al., 2016), and metaWRAP (Uritskiy et al., 2018). DAS_Tool (Sieber et al., 2018) was used to identify the best *Ca. E. persephone* bin for each sample among results of these three different binners.

The final *Ca. E. Persephone* MAGs were annotated with Prokka (Seeman, 2014) and subsequently used for pangenome reconstruction with Panaroo (Tonkin et al., 2020). Symbiont variant calling was facilitated through mapping the metagenomic reads against the *Ca. E. persephone* pangenome with Bowtie2 in very sensitive mode (Langmead and Salzberg, 2012). Optical duplicates were removed with Picard tools; and indel realignment and base calibration were performed with LoFreq (Wilm et al., 2012). Read depth was normalized across samples by subsampling to the lowest number of aligned reads against the *Ca. E. persephone* pangenome. Variant calling analysis was performed with Freebayes using parameters for metagenomic data (Garrison and Marth, 2012). Finally, variant calls were filtered for strand bias, proximity to indel regions, base quality, and depth. Sites with greater than 25% missing data were removed by bcftools (Li et al., 2009). Haplotype and allele count information were generated with vcftools (Danecek et al., 2011) and the VariantsToTable tool from GATK. Gene content variation (i.e., gene presence and absence) among samples was determined with PanPhlAn (Beghini et al., 2021).

Symbiont Genetic Diversity between Hydrothermal Vent Sites

Downstream population genomic analysis was based on the methods outlined by Breusing et al. (2022). Population genetic structure was assessed through principal coordinate analysis (PCoA) based on Euclidean distances and Bray-Curtis dissimilarities of haplotype information and nucleotide counts, respectively, with the *ape* and *Stats* packages in R (Paradis and Schliep, 2019; R Core Team, 2022). The Cailliez correction was applied to correct for negative eigenvalues (Cailliez, 1983; Breusing et al., 2022). Samples were organized by dive number for analysis, corresponding to distinct hydrothermal vent sites. PCoA ordination plots were generated with *GGplot2* in R (Wickham, 2016).

Additional PCoA plots were generated to investigate the impact of host size and broader geographic region on symbiont population structure. SNP analysis was performed across samples through PCoA analysis of nucleotide count and haplotype for samples by host physiology (i.e., host size) and collection region using the same methods as above. SNP analysis relative to host size was additionally investigated within hydrothermal vent sites corresponding to dives J2-1390 and J2-1392, respectively.

Gene presence and absence data were assessed through the assembly of heatmaps and interpretation of clustering by dive number, or hydrothermal vent site. Heatmaps were generated with the use of the *ComplexHeatmap* package in R (Gu et al., 2016). Heatmaps were prepared with total gene presence and absence data across samples and with universally present genes removed for higher resolution. The degree of population overlap was further assessed through analysis of fixation index (F_{ST}) and pangenome fixation index (P_{ST}) (Picazo et al., 2019; Picazo et al., 2021).

Chapter III.

Results

Results from sample processing and DNA extractions and *Ca. E. persephone* genomic analysis are reported here. Sample processing and DNA extraction lab metric data provide context for downstream sample performance; and sample condition and physiology represent variables with the potential to impact symbiont population structure, as assessed through sequencing performance and genetic diversity. Distance between hydrothermal vent sites is investigated to understand the level of gene flow, or *Ca. E. persephone* population connectivity, across the Guaymas Basin.

Sample Collection, Dissection, and DNA Extraction

The condition and physiology (i.e., size range) of individual samples reported below allow for comparison of sequencing performance to storage condition and genetic diversity to size. Due to the horizontal transmission of *Ca. E. persephone* in the larval stage of the *R. pachytila* lifecycle, populations of *Ca. E. persephone* within the host are thought to be polyclonal, with one genotype dominating (Polzin et al., 2019). Therefore, the size of the *R. pachytila* host represents an important variable of time of *Ca. E. persephone* uptake since *R. pachytila* size corresponds to the worm's age.

Specimens ranged in size from small (<15 mm), medium (15 – 25 mm), and large (>25 mm), as defined by worm trunk diameter. A majority of samples were within the small size range, with twenty-two specimens <15 mm in trunk diameter. Twelve of the

remaining samples were medium sized and eleven specimens were large sized. Most samples used within this study were whole worm specimens with tube removed (n = 38) or tube intact (n = 8). The remainder of sample types were dissected trophosome preserved in RNAlater (n = 8) and leftover trophosome prep samples (n = 8).

Nonparametric analysis of variance shows a significant reduction in extracted DNA concentration for previously dissected and stored sample and trophosome (median = 46.7 ng/ μ l) relative to small (median = 85.65 ng/ μ l), medium (median = 108.75 ng/ μ l), and large (median = 110.50 ng/ μ l) sized intact worms (ChiSquare = 18.6561; DF = 3; p = 0.0003).

Table 2. Sample Condition, Physiology, and Lab Metrics.

Sample Name	Dive ID	Storage Condition	Worm Size Range	Dissection Mass (g)	Average Extraction Concentration (ng/ μ l)
B1_J1392_W1A_1	J2-1392	Whole Frozen Worm, Tube Intact	Large	21.4	59
B1_J1392_W1B_3	J2-1392	Whole Frozen Worm, Tube Intact	Large	30.3	111
B1_J1392_W2B_4	J2-1392	Whole Frozen Worm, Tube Intact	Large	24.4	138
B5_J1390_W1_5	J2-1390	Whole Frozen Worm, Tube Removed	Medium	18.9	100
B5_J1390_W2_6	J2-1390	Whole Frozen Worm, Tube Removed	Large	26.4	378
B6_J1390_W15_7	J2-1390	Whole Frozen Worm, Tube Removed	Small	25.2	47
B6_J1390_W1_8	J2-1390	Whole Frozen Worm, Tube Removed	Small	19.3	65
B6_J1390_W2_9	J2-1390	Whole Frozen Worm, Tube Removed	Small	25.2	81
B6_J1390_W3_10	J2-1390	Whole Frozen Worm, Tube Removed	Small	24.6	65
B6_J1390_W4_11	J2-1390	Whole Frozen Worm, Tube Removed	Small	19.2	42
B6_J1390_W5_12	J2-1390	Whole Frozen Worm, Tube Removed	Small	34.2	52
B6_J1390_W6_13	J2-1390	Whole Frozen Worm, Tube Removed	Small	25.6	67
B6_J1390_W7_14	J2-1390	Whole Frozen Worm, Tube Removed	Small	28.3	89
B6_J1390_W8_15	J2-1390	Whole Frozen Worm, Tube Removed	Small	31.7	130
B6_J1390_W9_16	J2-1390	Whole Frozen Worm, Tube Removed	Medium	38.1	106
B6_J1390_W10_17	J2-1390	Whole Frozen Worm, Tube Removed	Small	27.7	96
B6_J1390_W11_18	J2-1390	Whole Frozen Worm, Tube Removed	Small	22.1	46
B6_J1390_W12_19	J2-1390	Whole Frozen Worm, Tube Removed	Small	23.3	87
B6_J1390_W13_20	J2-1390	Whole Frozen Worm, Tube Removed	Small	30.4	113
B6_J1390_W14_21	J2-1390	Whole Frozen Worm, Tube Removed	Small	23.1	75
B3_J1392_W1_22	J2-1392	Whole Frozen Worm, Tube Removed	Large	13.0	238
B4_J1392_W1_23	J2-1392	Whole Frozen Worm, Tube Removed	Medium	29.2	89
B4_J1392_W2_24	J2-1392	Whole Frozen Worm, Tube Removed	Medium	32.1	87
B4_J1392_W3_25	J2-1392	Whole Frozen Worm, Tube Removed	Small	41.1	125
B4_J1392_W4_26	J2-1392	Whole Frozen Worm, Tube Removed	Medium	28.0	251

Sample Name	Dive ID	Storage Condition	Worm Size Range	Dissection Mass (g)	Average Extraction Concentration (ng/μl)
B4_J1392_W5_27	J2-1392	Whole Frozen Worm, Tube Removed	Medium	20.1	60
B4_J1392_W6_28	J2-1392	Whole Frozen Worm, Tube Removed	Small	26.8	75
B4_J1392_W7_29	J2-1392	Whole Frozen Worm, Tube Removed	Medium	21.2	76
B4_J1392_W8_30	J2-1392	Whole Frozen Worm, Tube Removed	Small	19.2	85
B4_J1392_W9_31	J2-1392	Whole Frozen Worm, Tube Removed	Medium	40.5	135
B4_J1392_W10_32	J2-1392	Whole Frozen Worm, Tube Removed	Small	32.3	115
B4_J1392_W11_33	J2-1392	Whole Frozen Worm, Tube Removed	Medium	25.4*	322
B4_J1392_W12_34	J2-1392	Whole Frozen Worm, Tube Removed	Small	27.8*	87
B4_J1392_W13_35	J2-1392	Whole Frozen Worm, Tube Removed	Small	31.8*	155
B4_J1392_W14_36	J2-1392	Whole Frozen Worm, Tube Removed	Small	23.7*	88
B4_J1392_W15_37	J2-1392	Whole Frozen Worm, Tube Removed	Small	28.5*	124
B2_J1392_W1A_38	J2-1392	Whole Frozen Worm, Tube Intact	Large	37.4*	188
B2_J1392_W1B_39	J2-1392	Whole Frozen Worm, Tube Intact	Large	24.7*	85
B2_J1392_W1C_40	J2-1392	Whole Frozen Worm, Tube Intact	Medium	17.6*	222
B2_J1392_W2C_41	J2-1392	Whole Frozen Worm, Tube Intact	Large	45.4*	113
G19_D233_W87_42	D233	Whole Frozen Worm, Tube Removed	Medium	43.7*	112
G19_D233_W89_43	D233	Whole Frozen Worm, Tube Removed	Large	36.4*	91
G19_D233_W86_44	D233	Whole Frozen Worm, Tube Removed	Large	29.4*	70
G19_D233_W88_45	D233	Whole Frozen Worm, Tube Removed	Large	39.5*	98
G19_D233_W90_46	D233	Whole Frozen Worm, Tube Removed	Medium	40.0*	238
G19_D231_W71_47	D231	Trophosome stored in RNAlater	N/A	61.1*	31
G19_D231_W72_48	D231	Trophosome stored in RNAlater	N/A	27.3*	30
G19_D231_W73_49	D231	Trophosome stored in RNAlater	N/A	34.1*	33
G19_D231_W74_50	D231	Trophosome stored in RNAlater	N/A	18.6*	24
G19_D233_W80_51	D233	Trophosome stored in RNAlater	N/A	45.9*	46
G19_D233_W81_52	D233	Trophosome stored in RNAlater	N/A	35.4*	31

Sample Name	Dive ID	Storage Condition	Worm Size Range	Dissection Mass (g)	Average Extraction Concentration (ng/ μ l)
G19_D233_W82_53	D233	Trophosome stored in RNAlater	N/A	33.4*	23
G19_D233_W83_54	D233	Trophosome stored in RNAlater	N/A	34.9*	60
B11_J1398_W1_55	J2-1398	Leftover Troph Prep	N/A	21.7*	61
B11_J1398_W2_56	J2-1398	Leftover Troph Prep	N/A	24.3*	122
B11_J1398_W3_57	J2-1398	Leftover Troph Prep	N/A	11.0*	53
B11_J1398_W4_58	J2-1398	Leftover Troph Prep	N/A	7.8*	82
B11_J1398_W5_59	J2-1398	Leftover Troph Prep	N/A	56.5*	147
B11_J1398_W6_60	J2-1398	Leftover Troph Prep	N/A	17.2*	37
B11_J1398_W7_61	J2-1398	Leftover Troph Prep	N/A	32.8*	123
B11_J1398_W8_62	J2-1398	Leftover Troph Prep	N/A	17.1*	47

*Description of sample type (i.e., sample storage conditions), size, dissection mass (g), and extraction yield (ng/ μ l). Size is provided for whole worm specimens in which the diameter could be measured. Size is defined as ‘Small’ (<15 mm), ‘Medium’ (15 – 25 mm), or ‘Large’ (>25 mm) based on observations of sample size range for this study. Fifty percent of samples were dissected in duplicate and combined equally by volume post DNA extraction for library prep and sequencing; average mass is reported for combined replicate samples. *Samples are a single dissection replicate.*

Metagenomic Sequencing Results

Raw sequencing data reported includes total reads, percent GC and AT content, and Q20 and Q30 phred quality scores. These data indicate adequate sequencing performance across all samples included in this study (Appendix 5). As such, the total sixty-one samples were included in downstream analyses; however, post Freebayes variant calling analysis, ten individuals were excluded due to high amounts of missing data (i.e., <75% of genetic sites detected in sample). Thus, fifty-one samples are included in the population genomics analysis that follows.

Sequencing performance was compared across study sites (i.e., dive number), storage condition, and host size by Kruskal-Wallis Test with Oneway ChiSquare Approximation to determine whether sample status and quality may impact downstream genomics analysis. The nonparametric Kruskal-Wallis Test was utilized, because read count and Q30 (%) sequencing metrics were determined to be non-normally distributed by the Shapiro-Wilk Goodness-of-Fit Test (read count: $W = 0.8752$; $p < 0.0001$) (Q30 (%): $W = 0.9515$; $p = 0.0170$).

Read count was significantly correlated with dive number, with greater read counts observed for southern Guaymas study sites (Dive IDs 231 and 233) (ChiSquare = 24.4342; DF = 4; $p < 0.0001$); however, equivalent sequencing quality was observed across study sites and dives as demonstrated by Q30 (%) scores (ChiSquare = 4.8482; DF = 4; $p = 0.3032$). Additionally, Kruskal-Wallis Test of supporting reads and sequencing quality across storage conditions revealed a significant correlation between sample storage condition and read count (ChiSquare = 16.0181; DF = 3; $p = 0.0011$) (Figure 6); yet there is no impact of storage condition on sample sequencing quality (ChiSquare =

2.2097; DF = 3; p= 0.5300) (Figure 7). It is likely that the storage condition or sample preparation associated with trophosome preserved in RNAlater samples influenced the greater read count observed in the southern Guaymas samples (Dive IDs 231 and 233), as these are the only samples associated with this advantageous storage condition.

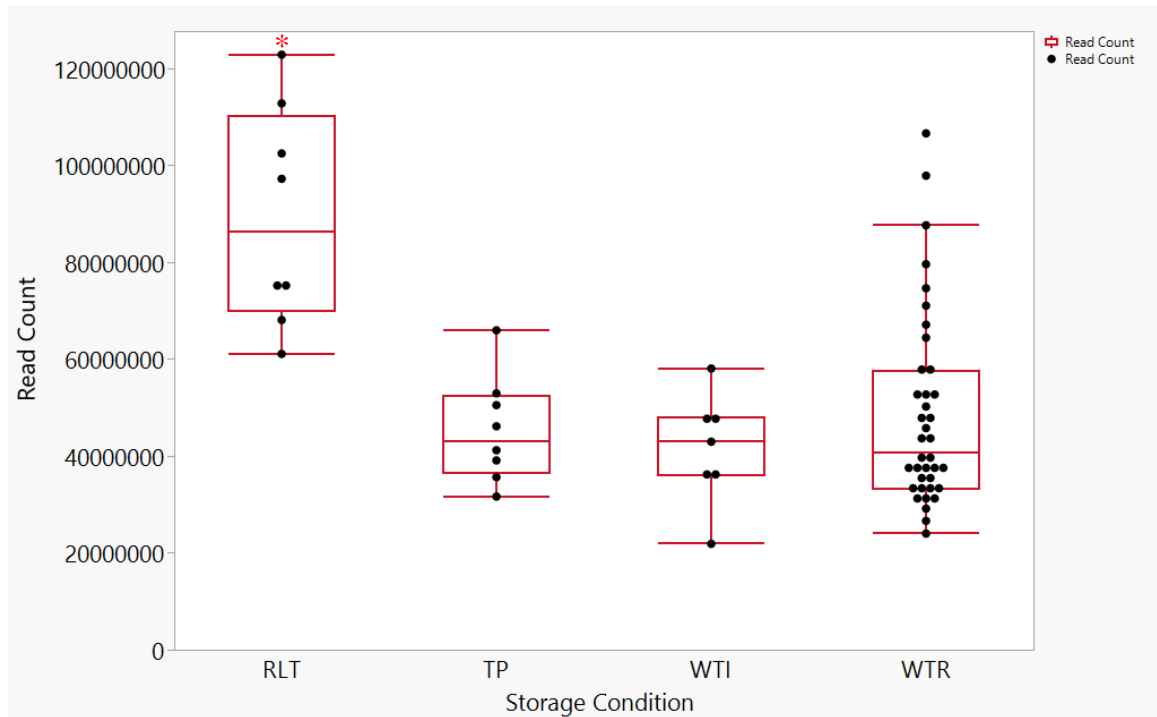


Figure 6. Oneway Analysis of Sample Storage Condition vs. Read Count.

RLT = RNAlater Treated Trophosome; TP = Leftover Trophosome Prep; WTI = Whole Frozen Worm, Tube Intact; WTR = Whole Frozen Worm, Tube Removed. Kruskal-Wallis analysis demonstrated that a significant increase in read count is observed for the RLT storage condition (ChiSquare = 16.0181; DF = 3; p = 0.0011) (). Quantiles for each variable are shown. Read count represents the total number of reads across reads 1 and 2 for paired-end sequencing.*

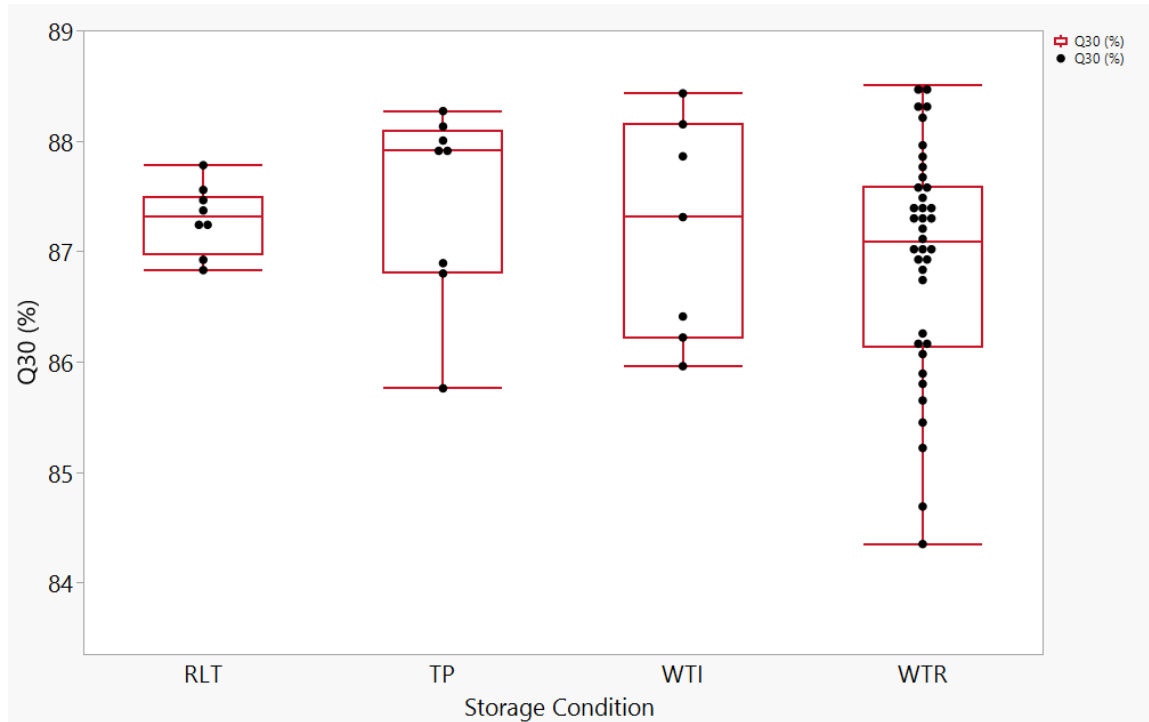


Figure 7. Oneway Analysis of Sample Storage Condition vs. Q30 (%).

RLT = RNAlater Treated Trophosome; TP = Leftover Trophosome Prep; WTI = Whole Frozen Worm, Tube Intact; WTR = Whole Frozen Worm, Tube Removed. There is no observed correlation between sample sequencing quality and sample storage condition by Kruskal-Wallis Test (ChiSquare = 2.2097; DF = 3; p= 0.5300). Quantiles for each variable are shown. Q30(%) = ratio of bases with a phred quality score of 30 or greater.

The correlation between host physiology, reported through size, and sequencing performance was additionally investigated through the Kruskal-Wallis Test. While read count for small samples was significantly lower than the count for *R. pachyptila* of medium and large size (ChiSquare = 9.9816; DF = 2; p= 0.0068) (Figure 8), there was no impact on sequencing quality as illustrated by Q30 (%) (ChiSquare = 0.2285; DF = 2; p= 0.8920) (Figure 9).

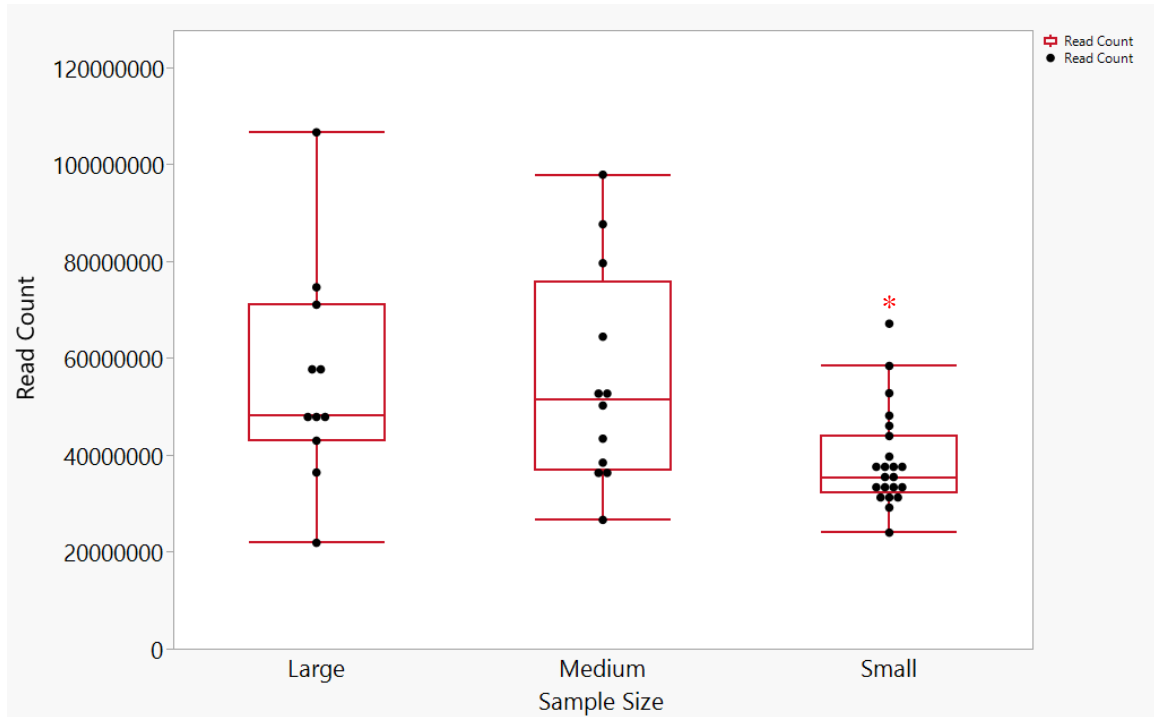


Figure 8. Oneway Analysis of Host Size vs. Read Count.

A significant decrease in read count by Kruskal-Wallis Test (ChiSquare = 9.9816; DF = 2; p= 0.0068) is observed for R. pachyptila of small size compared to large and medium host samples (). Quantiles for each variable are shown. Read count represents the total number of reads across reads 1 and 2 for paired-end sequencing.*

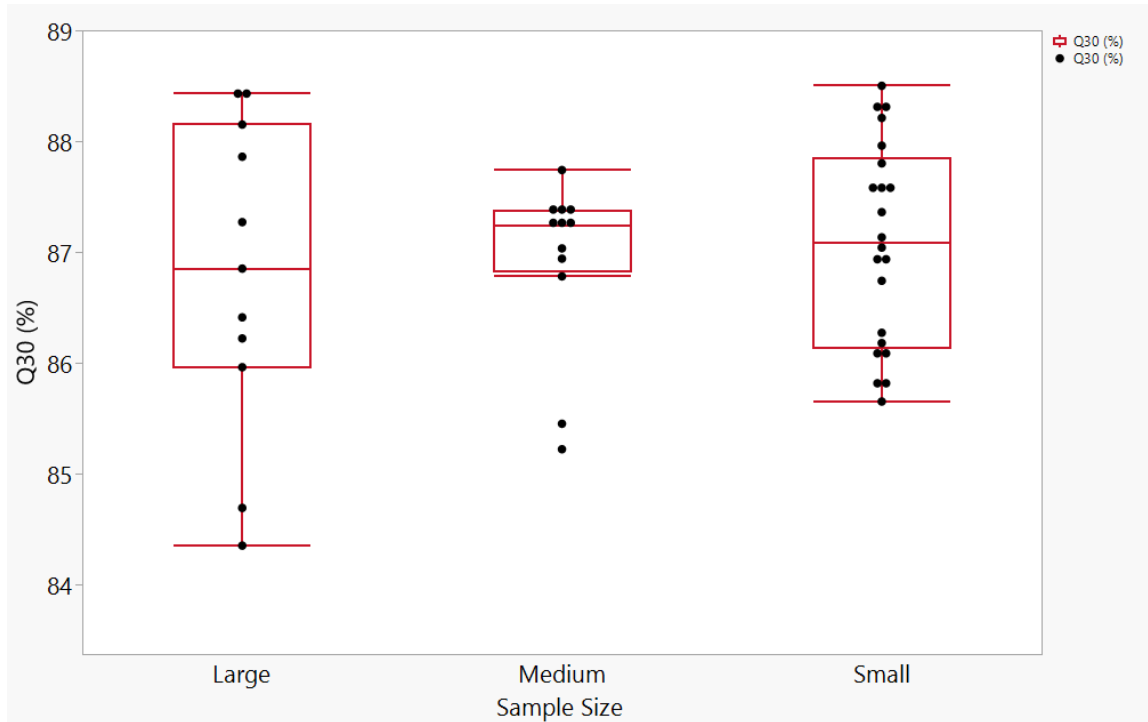


Figure 9. Oneway Analysis of Host Size vs. Q30 (%).

*There is no observed correlation between sample sequencing quality and *R. pachyptila* size by Kruskal-Wallis Test (ChiSquare = 0.2285; DF = 2; p = 0.8920). Quantiles for each variable are shown. Q30(%) = ratio of bases with a phred quality score of 30 or greater.*

The quality of the ten samples removed due to low genetic site detection was examined through read count and Q30 (%). There was no significant difference between read count for samples included and excluded from downstream genomics analysis due to low-confidence post Freebayes variant calling (Kruskal-Wallis Test: ChiSquare = 1.6034; DF = 1; p = 0.2054). For Q30 (%), samples excluded from analysis had significantly higher Q30 (%) values than those included in downstream analysis (Kruskal-Wallis Test: ChiSquare = 5.7423; DF = 1; p = 0.0166) (Figure 10). Additionally, Q30 (%) values were universally high across all samples (>84%). Further, samples excluded from analysis

came from a range of sites (J1392 (n=5), J1390 (n=3), J1398 (n=1), D231 (n=1)) and sample storage conditions (whole frozen worm, tube removed (n=8); RNAlater treated trophosome (n=1); leftover trophosome prep (n=1)). Thus, poor genetic site representation cannot be attributed to the sample conditions investigated here.

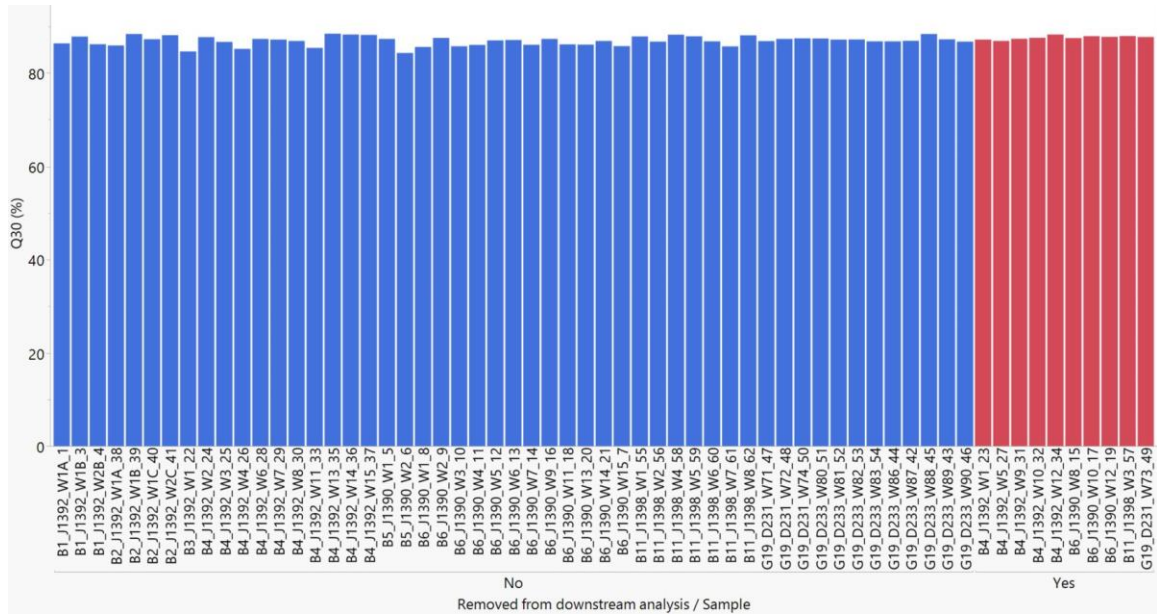


Figure 10. Q30 (%) for Samples Removed from Downstream Analysis vs. Samples Included in Population Genomics Analysis.

Q30 (%) for samples excluded from downstream population genomics analysis were significantly greater than those included in the investigation, as determined by Kruskal-Wallis Test (ChiSquare = 5.7423; DF = 1; p = 0.0166). The samples excluded from the population genomics analysis (red) had low confidence (<75% genetic site detection) post FreeBayes variant calling. These samples come from diverse sample sites and storage conditions. All samples had overall good sequencing performance, with Q30 (%) values >84%.

Inter-site Genetic Diversity

Inter-site genetic diversity investigates the impact of physical distance and geographical barriers on gene flow across distinct hydrothermal vent sites. This investigation considers trends across distinct sites, through variant analysis (i.e., comparison of nucleotide counts and haplotypes) of symbiont populations across sites, as well as gene presence and absence analysis. Variants of low confidence post Freebayes variant calling (i.e., detected in <25% of samples) are discarded from the following analyses. A total of fifty-one genetic variants were recovered.

PCoA of sample nucleotide counts (Figure 11) and sample haplotype (Figure 12) organized by dive ID, do not show a clear correlation of symbiont genomic variation with hydrothermal vent site. Instead, the data support high inter-individual variability within sites that exceeds any difference among sites.

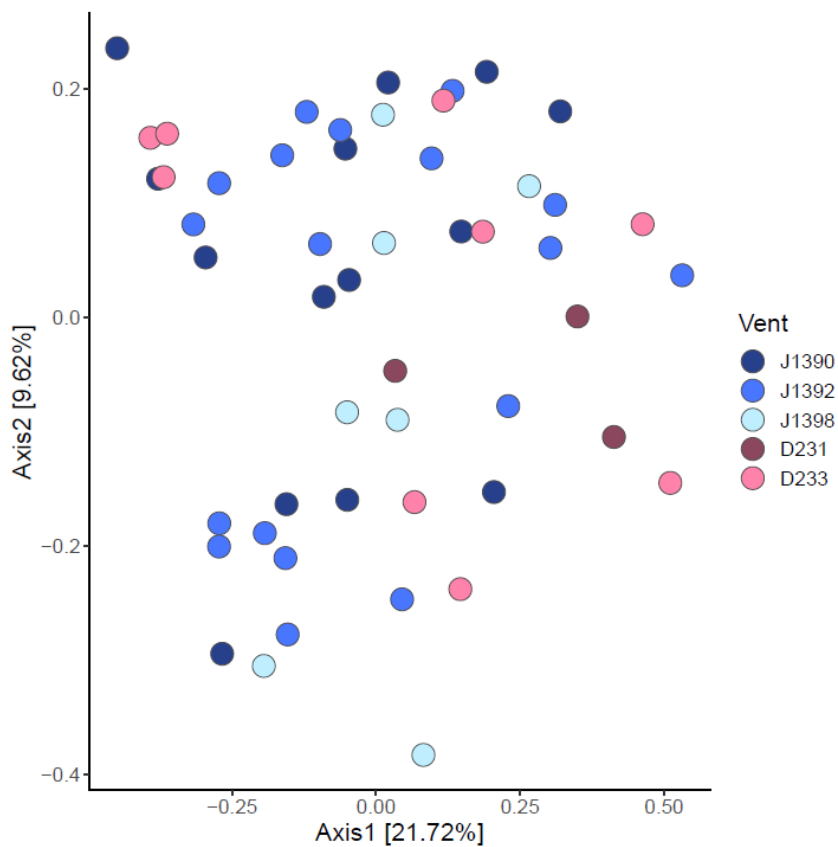


Figure 11. PCoA of Nucleotide Counts across Hydrothermal Vent Sides (i.e., Dive ID).

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. There is no distinct correlation between hydrothermal vent sites and nucleotide counts.

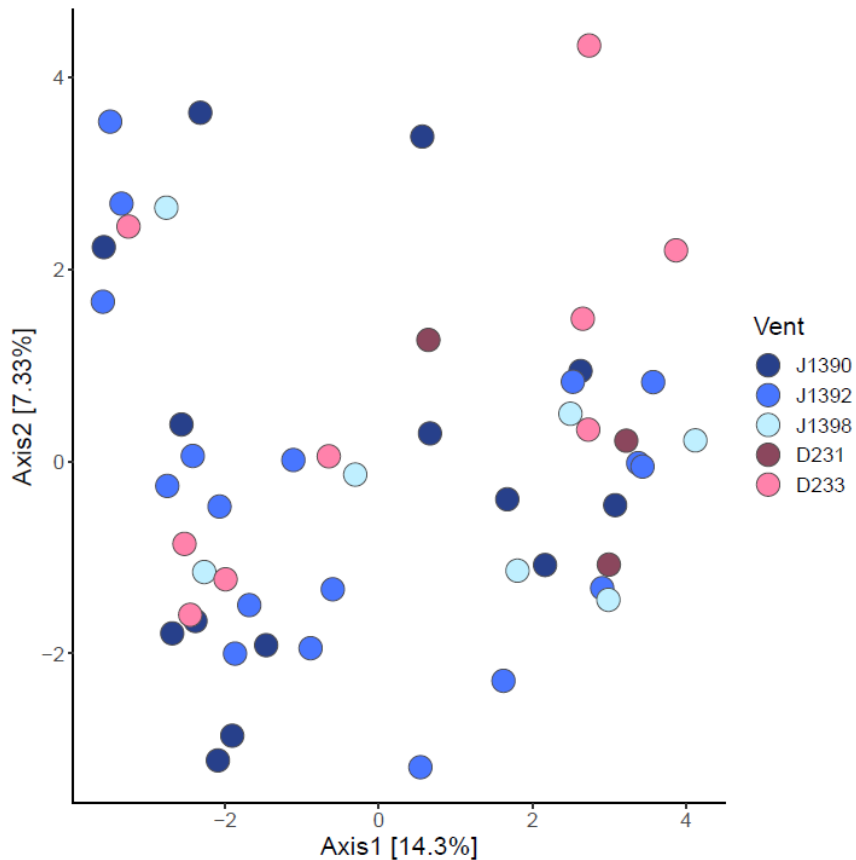


Figure 12. PCoA of Haplotypes across Hydrothermal Vent Sites (i.e., Dive ID).

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. There is no distinct correlation between hydrothermal vent sites and consensus genotype.

Investigation of gene presence and absence across distinct hydrothermal vent sites support the findings of the SNP analysis, with no clear evidence of genomic distinction between *Ca. E. persephone* populations across vent sites. Total gene presence and absence data for all samples included in the population genomics analysis (excluding those of low confidence) are visualized as a heatmap (Figure 13). It is apparent from these data that a majority of genes are present across all samples and vent sites.

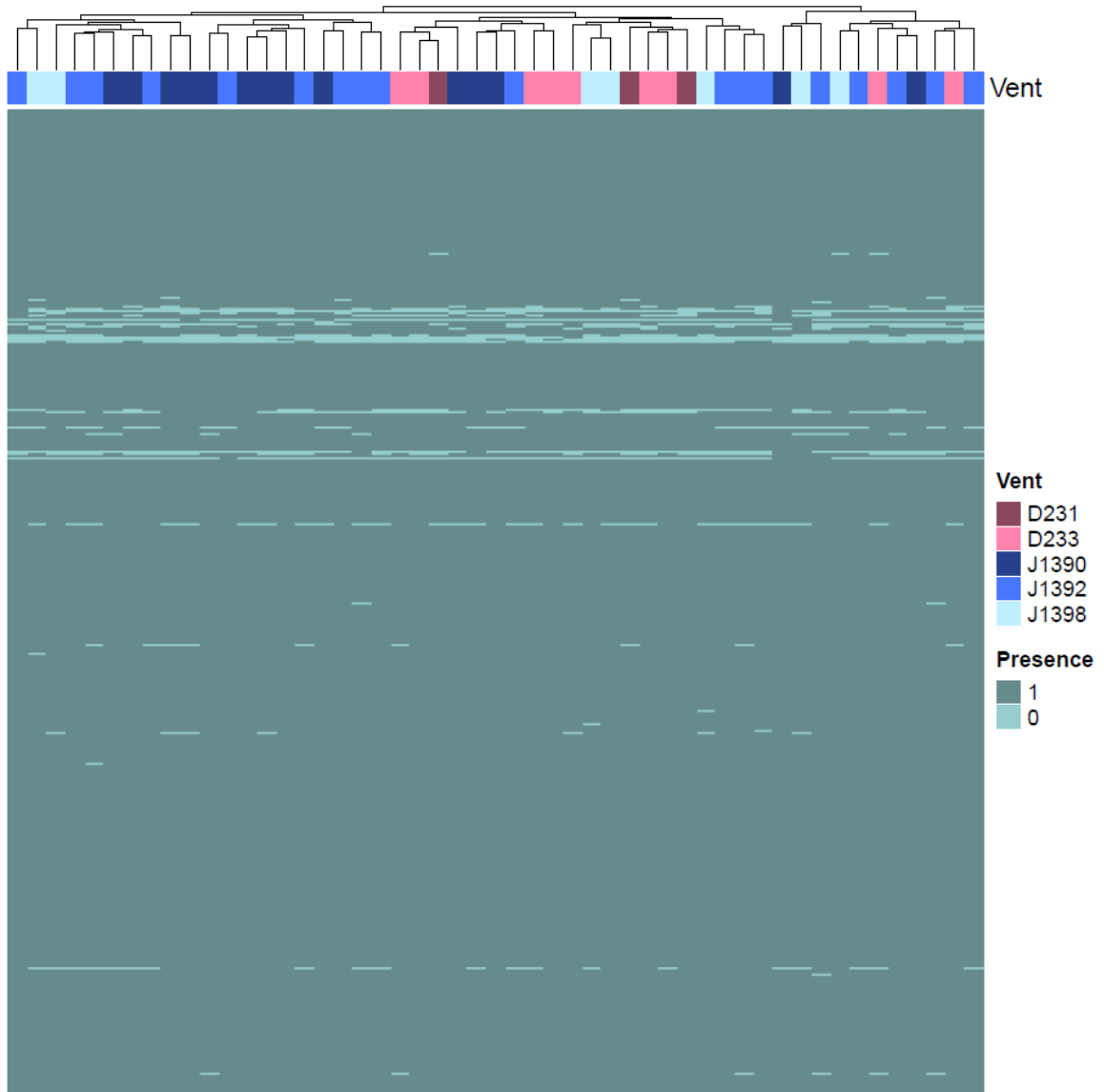


Figure 13. Heatmap of Present and Absent Genes across Hydrothermal Vent Sites by Sample.

Heatmap analysis of total gene presence and absence data across samples and hydrothermal vent sites. Samples and genes of low confidence post Freebayes variant calling are excluded from all population genomics analyses. These data support that a majority of genes are present in all samples across hydrothermal vent sites.

Removal of universally present genes allows for a higher resolution picture of the relationship between gene presence and absence and hydrothermal vent site (Figure 14). There is no distinction in gene presence and absence across vent sites when investigated by sample, as evidenced by lack of grouping by hydrothermal vent sites in heat map analysis and analyses of population divergence (F_{ST} and P_{ST}) (Table 3). Percent gene presence by hydrothermal vent sites and gene annotations for known *Ca. E. persephone* genes are reported in Table 4 below. The percent gene presence data in Table 4 further support that there are no fixed differences in genomic structure across populations. Uncharacterized genes included in the percent gene presence analysis are reported in Appendix 6.

Table 3. Population Divergence Across Hydrothermal Vent Sites.

Comparison	F_{ST}	P_{ST}
J1398_J1392	0.0129	0.0041
J1398_J1390	0.0248	0.0131
J1398_D231	0.0100	0.0445
J1398_D233	0.1152	0.0148
J1392_J1390	-0.0043	0.0066
J1392_D231	0.0600	0.0575
J1392_D233	0.0277	0.0214
J1390_D231	0.0877	0.0800
J1390_D233	0.0182	0.0368
D231_D233	0.1486	0.0223

F_{ST} and *P_{ST}* values across hydrothermal vent sites. Values of 1 indicate total population divergence, while values of 0 indicate complete overlap of populations. Practically applied, values >0.15 indicate a significant degree of population divergence (Frankham et al., 2002). There is no indication of significant population divergence across hydrothermal vent sites in the Guaymas Basin.

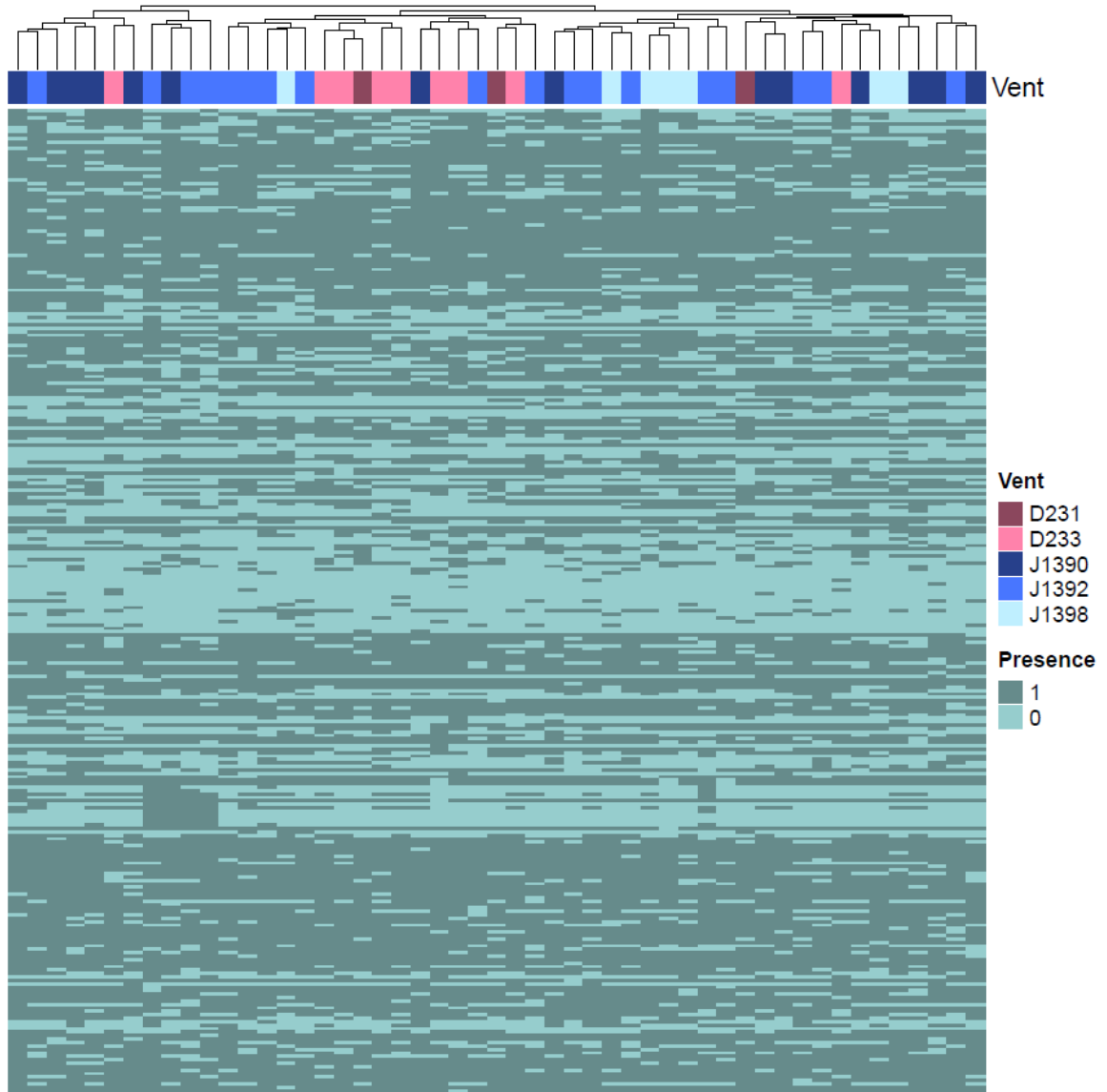


Figure 14. Heatmap of Genes Present and Absent across Hydrothermal Vent Sites by Sample, Universally Present Genes Removed.

Heatmap analysis of gene presence and absence data across samples and hydrothermal vent sites, with universally present genes removed. Samples and genes of low confidence are excluded from all population genomics analysis. These data support that no clear correlations in gene presence and absence are present across sites.

Table 4. *Candidatus* Endoriftia persephone Gene Annotations for Known Genes across Hydrothermal Vent Sites, Genes Universally Present Removed.

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
acpP_1	Acyl carrier protein	42.86	38.89	14.29	0.00	44.44
acpP_2	Acyl carrier protein	42.86	55.56	71.43	66.67	100.00
cas2_1	CRISPR-associated endoribonuclease Cas2	57.14	55.56	57.14	33.33	66.67
cas2_2	CRISPR-associated endoribonuclease Cas2	57.14	72.22	50.00	66.67	77.78
cas2b	CRISPR-associated endonuclease Cas2 2	85.71	83.33	78.57	100.00	88.89
cc4_4	Cytochrome c4	57.14	44.44	50.00	66.67	44.44
clpS	ATP-dependent Clp protease adapter protein ClpS	0.00	16.67	50.00	0.00	33.33
dgkA	Diacylglycerol kinase	100.00	100.00	92.86	66.67	88.89
dksA_2	RNA polymerase-binding transcription factor DksA	57.14	72.22	64.29	66.67	55.56
fdx_2	Ferredoxin	57.14	66.67	71.43	100.00	66.67
ftsL	Cell division protein FtsL	100.00	77.78	85.71	66.67	88.89
gatC	Glutamyl-tRNA(Gln) amidotransferase subunit C	85.71	100.00	100.00	66.67	88.89
groS	10 kDa chaperonin	100.00	72.22	71.43	100.00	100.00
group_1009	UniRef90_G2DEJ4 Glutaredoxin family protein	100.00	100.00	100.00	100.00	88.89
group_1019	UniRef90_G2DBL4 Cytochrome c, class I	85.71	50.00	71.43	66.67	44.44
group_1031	UniRef90_G2DCV9 HMA domain-containing protein	100.00	94.44	100.00	100.00	100.00
group_1034	UniRef90_A0A0T5YVH1 Cytochrome c	57.14	100.00	100.00	100.00	100.00
group_1052	UniRef90_D4TBZ2 ISxac2 transposase	100.00	61.11	42.86	66.67	77.78
group_1055	UniRef90_G2DBD5 UPF0235 protein Rfp1Sym_aq00240	100.00	100.00	92.86	100.00	100.00
group_1057	UniRef90_G2DFY6 Flagellar hook-length control protein	100.00	94.44	92.86	100.00	100.00
group_1067	UniRef90_G2D9K1 PepSY domain-containing protein	100.00	94.44	78.57	100.00	100.00
group_1071	UniRef90_A0A0T5ZAK6 Response regulatory domain-containing protein	100.00	94.44	92.86	100.00	100.00
group_1073	UniRef90_G2FIP8 Green heme protein	100.00	88.89	100.00	100.00	88.89

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1075	UniRef90_A0A3B0Y5K2 Glutaredoxin domain-containing protein	85.71	100.00	100.00	100.00	100.00
group_1078	UniRef90_A0A0T5YTT1 RNA recognition motif (A.k.a RRM, RBD, or RNP domain)	85.71	94.44	100.00	66.67	100.00
group_1087	UniRef90_A0A0T5Z8R4 Rho-binding antiterminator	100.00	100.00	92.86	100.00	100.00
group_1093	UniRef90_G2FDN0 Sensory box/GGDEF family protein	100.00	94.44	100.00	100.00	100.00
group_1095	UniRef90_A0A558D1M7 RNA-binding protein	71.43	77.78	28.57	66.67	66.67
group_1096	UniRef90_A0A084IQ19 Transposase family protein	0.00	16.67	35.71	66.67	44.44
group_1097	UniRef90_A0A0T5Z2E5 Coenzyme PQQ synthesis protein D (PqqD)	100.00	83.33	78.57	100.00	55.56
group_1100	UniRef90_G2FHG2 Regulatory protein, FmdB family	100.00	94.44	100.00	100.00	100.00
group_1102	UniRef90_G2DH70 TfoX_N domain-containing protein	57.14	61.11	71.43	33.33	44.44
group_1103	UniRef90_G2DDP8 DUF4212 domain-containing protein	57.14	55.56	57.14	33.33	33.33
group_1109	UniRef90_A0A0T5Z6Y7 IS66 C-terminal element	100.00	100.00	92.86	66.67	88.89
group_1110	UniRef90_G2FDI4 Protein RtcB	0.00	5.56	7.14	0.00	0.00
group_1111	UniRef90_A0A0T5YUD7 Iron-binding zinc finger CDGSH type	85.71	77.78	92.86	66.67	66.67
group_1112	UniRef90_G2FHH8 TPR repeat-containing protein	42.86	55.56	50.00	66.67	33.33
group_1122	UniRef90_G2FEJ5 Transglycosylase-associated protein	14.29	16.67	21.43	33.33	33.33
group_1129	UniRef90_G2DHZ2 Type I restriction-modification system, restriction subunit R	85.71	94.44	85.71	100.00	100.00
group_1132	UniRef90_G2FDA7 DUF772 domain-containing protein	0.00	5.56	14.29	0.00	22.22
group_1139	UniRef90_G2DGR3 SHOCT domain-containing protein	28.57	38.89	42.86	100.00	55.56
group_1142	UniRef90_B8KMI1 Integrase, catalytic region	14.29	11.11	21.43	66.67	55.56
group_1144	UniRef90_G2DCF4 Sulfurtransferase	100.00	83.33	100.00	100.00	77.78
group_1148	UniRef90_G2FG82 DUF3322 domain-containing protein	14.29	0.00	7.14	0.00	11.11
group_1149	UniRef90_A0A0T5YV80 Cbb3-type cytochrome oxidase component FixQ	85.71	100.00	85.71	66.67	88.89

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1150	Hypothetical protein UniRef90_A0A0T5Z1T6 Cytochrome oxidase maturation	14.29	0.00	0.00	0.00	0.00
group_1151	protein, cbb3-type	0.00	0.00	21.43	0.00	0.00
group_1165	UniRef90_A0A0T5Z276 Sulfur carrier protein ThiS UniRef90_A0A0T5Z3C5 Zn-binding Pro-Ala-Ala-Arg	14.29	0.00	0.00	66.67	0.00
group_1166	(PAAR) domain, involved in TypeVI secretion	85.71	88.89	85.71	100.00	77.78
group_1175	UniRef90_G2DCD4 zf-CHCC domain-containing protein	85.71	94.44	100.00	100.00	88.89
group_1177	UniRef90_G2FFD3 UPF0434 protein TévJSym_al00450	57.14	27.78	21.43	66.67	44.44
group_1181	UniRef90_G2DHQ9 Flagellar hook-associated protein 2	28.57	38.89	64.29	100.00	66.67
group_1204	UniRef90_A0A0T5Z0L0 Heme exporter protein D	0.00	0.00	14.29	0.00	11.11
group_1251	UniRef90_G2DFN7 Hydrolase_4 domain-containing protein	100.00	94.44	100.00	100.00	100.00
group_1272	UniRef90_G2FF44 FeoA domain-containing protein	100.00	94.44	92.86	100.00	100.00
group_1285	UniRef90_A0A0T5Z5N4 Probable Fe(2+)-trafficking protein	100.00	94.44	92.86	33.33	88.89
group_1297	UniRef90_G2FC76 Inner membrane protein CreD	85.71	94.44	100.00	100.00	100.00
group_1329	UniRef90_G2DAF0 SH3b domain-containing protein UniRef90_A0A0T5YYB5 PEP-CTERM protein-sorting	71.43	100.00	100.00	100.00	100.00
group_1362	domain	100.00	94.44	92.86	100.00	100.00
group_1433	UniRef90_G2FJ09 Type IV pilus assembly protein PilZ UniRef90_A0A0T5Z283 Response regulator receiver domain-	100.00	94.44	100.00	100.00	100.00
group_1439	containing protein	100.00	83.33	85.71	100.00	88.89
group_1584	UniRef90_A0A3Q3D3R7 Histone H3 UniRef90_A0A497TUC5 Reverse transcriptase domain-	28.57	16.67	7.14	0.00	0.00
group_1586	containing protein	14.29	11.11	21.43	0.00	0.00
group_1590	UniRef90_O18643 Histone H2B UniRef90_A0A497U3G8 Reverse transcriptase domain-	28.57	16.67	7.14	0.00	0.00
group_1673	containing protein	14.29	11.11	7.14	0.00	0.00
group_1674	UniRef90_A0A497TUC5 Reverse transcriptase domain-	14.29	11.11	7.14	0.00	0.00
	containing protein	14.29	11.11	7.14	0.00	0.00

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1675	UniRef90_A0A497U3G8 Reverse transcriptase domain-containing protein	14.29	11.11	7.14	0.00	0.00
group_1676	UniRef90_A0A497TUC5 Reverse transcriptase domain-containing protein	14.29	22.22	14.29	0.00	0.00
group_355	UniRef90_A0A0T5YZ71 Chemoreceptor zinc-binding domain-containing protein	42.86	50.00	42.86	33.33	44.44
group_71	UniRef90_G2DAR6 Transposase	100.00	77.78	85.71	66.67	77.78
group_902	UniRef90_A0A1D8K7V9 Transposase	85.71	100.00	100.00	100.00	100.00
group_939	UniRef90_A0A0T5Z2W7 SpoIIAA-like	100.00	88.89	100.00	100.00	100.00
group_941	UniRef90_G2D9H4 HTH cro/C1-type domain-containing protein	85.71	100.00	100.00	100.00	100.00
group_942	UniRef90_A0A0T5ZBJ7 Caspase domain-containing protein	85.71	88.89	100.00	100.00	88.89
group_944	UniRef90_A0A0T5YZQ0 Monoheme cytochrome SoxX (Sulfur oxidation)	100.00	94.44	100.00	100.00	100.00
group_956	UniRef90_G2DFB8 OmpA/MotB	71.43	50.00	28.57	33.33	66.67
group_960	UniRef90_G2DFB6 Transcriptional regulator, MerR family	100.00	94.44	100.00	100.00	100.00
group_968	UniRef90_A0A0T5YX77 PilZ domain	57.14	100.00	78.57	100.00	88.89
group_971	UniRef90_G2D9C0 Cobyrinic acid a,c-diamide synthase	100.00	66.67	78.57	100.00	88.89
group_981	UniRef90_G2DDL2 Putative sulfur globule protein SgpA	85.71	94.44	100.00	66.67	77.78
group_984	UniRef90_G2FB67 Response regulator NasT	100.00	88.89	85.71	0.00	100.00
grxC	Glutaredoxin 3	57.14	66.67	100.00	100.00	88.89
grxD	Glutaredoxin 4	100.00	88.89	92.86	66.67	100.00
hfq	RNA-binding protein Hfq	100.00	94.44	92.86	100.00	100.00
hpf_1	Ribosome hibernation promoting factor	100.00	94.44	100.00	100.00	100.00
hupB	DNA-binding protein HU-beta	100.00	72.22	92.86	66.67	44.44
hypC	Hydrogenase maturation factor HypC	100.00	83.33	85.71	100.00	100.00
ibaG	Acid stress protein IbaG	42.86	55.56	71.43	33.33	66.67
infA	Translation initiation factor IF-1	14.29	16.67	14.29	0.00	33.33

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
infC	Translation initiation factor IF-3	100.00	94.44	100.00	100.00	100.00
iscX	Protein IscX	28.57	5.56	7.14	0.00	22.22
lapA	Lipopolysaccharide assembly protein A	100.00	94.44	92.86	100.00	100.00
napD	Chaperone NapD	71.43	77.78	71.43	66.67	88.89
ndhI_1	NAD(P)H-quinone oxidoreductase subunit I,C chloroplastic	100.00	100.00	92.86	100.00	88.89
nqo7	NADH-quinone oxidoreductase subunit 7	100.00	88.89	100.00	100.00	88.89
oadG	oxaloacetate decarboxylase gamma chain	57.14	33.33	57.14	66.67	66.67
recX	Regulatory protein RecX	100.00	100.00	92.86	100.00	77.78
rplU	50S ribosomal protein L21	71.43	88.89	100.00	100.00	88.89
rpmB	50S ribosomal protein L28	85.71	66.67	71.43	100.00	66.67
rpmC	50S ribosomal protein L29	71.43	72.22	71.43	0.00	88.89
rpmD	50S ribosomal protein L30	28.57	5.56	57.14	0.00	0.00
rpmF	50S ribosomal protein L32	42.86	38.89	42.86	33.33	33.33
rpmG	50S ribosomal protein L33	0.00	5.56	0.00	0.00	0.00
rpmI	50S ribosomal protein L35	42.86	50.00	71.43	100.00	66.67
rpsL	30S ribosomal protein S12	100.00	94.44	100.00	100.00	100.00
rpsQ	30S ribosomal protein S17	85.71	88.89	100.00	100.00	100.00
rpsT	30S ribosomal protein S20	42.86	66.67	57.14	0.00	33.33
rpsU	30S ribosomal protein S21	100.00	66.67	57.14	100.00	77.78
slyX	Protein SlyX	85.71	94.44	85.71	66.67	77.78
tatA	Sec-independent protein translocase protein TatA	71.43	77.78	71.43	100.00	77.78
tusA_1	Sulfur carrier protein TusA	57.14	66.67	85.71	66.67	66.67
tusE_2	Sulfurtransferase TusE	100.00	83.33	92.86	100.00	88.89
tusE_3	Sulfurtransferase TusE	85.71	94.44	78.57	100.00	100.00
ubiK	Flavin prenyltransferase UbiX	100.00	100.00	100.00	66.67	88.89
xseB	Exodeoxyribonuclease 7 small subunit	85.71	77.78	85.71	100.00	100.00
ycgL	Protein YcgL	100.00	94.44	100.00	100.00	100.00
yefM	Antitoxin YefM	100.00	83.33	92.86	100.00	77.78

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
yffB	Protein YffB	100.00	100.00	92.86	100.00	100.00
ygbF	CRISPR-associated endoribonuclease Cas2	85.71	55.56	28.57	66.67	88.89
yqgF	Putative pre-16S rRNA nuclease	100.00	100.00	92.86	100.00	100.00
zapA	Cell division protein ZapA	100.00	100.00	100.00	100.00	88.89

Percent gene presence is reported across hydrothermal vent sites (i.e., dive IDs), with universally present genes removed. Only genes with known annotations are reported here, with additional uncharacterized genes for this analysis reported in Appendix 6.

Impact of Host Physiology and Environment on Symbiont Genetic Diversity

The impact of sample collection area (Table 5) and host physiology on *Ca. E. persephone* genomes are considered here. PCoA of nucleotide counts and consensus genotypes across collection areas did not indicate any clear association between symbiont populations and broader geographic region (Figure 15, 16).

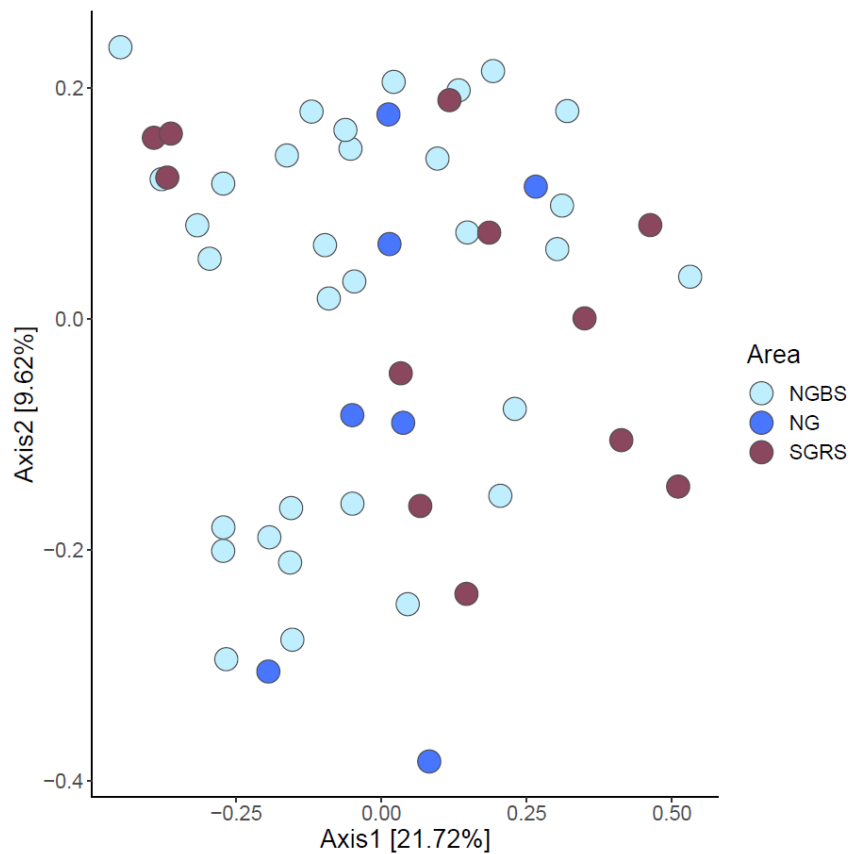


Figure. 15. PCoA of Nucleotide Counts across Collection Areas.

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. Collection areas are defined as North Guaymas, Black Smoker Chimneys 1 and 2 (NGBS); North Guaymas (NG); and South Guaymas, Ridge S (SGRS). There is no distinct correlation between collection area and symbiont nucleotide counts.

Table 5. Environmental Conditions by Cruise and Dive ID.

Collection Region	Collection Area	Cruise ID	Dive Vehicle	Dive ID	Collection Date	Lat/Long	Depth (m)
Guaymas Basin	North Guaymas Black Smoker, Chimneys 1 and 2	RR2107	ROV Jason II	J2-1390	19-Nov-2021	27.40923, -111.39908	1844
Guaymas Basin	North Guaymas Black Smoker, Chimneys 1 and 2	RR2107	ROV Jason II	J2-1392	21-Nov-2021	27.41276, -111.38717	1855
Guaymas Basin	North Guaymas	RR2107	ROV Jason II	J2-1398	28-Nov-2021	27.40432, -111.32127	2017
Guaymas Basin	South Guaymas, Ridge S, Rebecca's Roost	FK190211	ROV SuBastian	D231	27-Feb-2019	27.01075, -111.40697	2015
Guaymas Basin	South Guaymas, Ridge S, Big Pagoda	FK190211	ROV SuBastian	D233	1-Mar-2019	27.013717, -111.41105	2015

Collection and environmental metrics reported by cruise and dive ID.

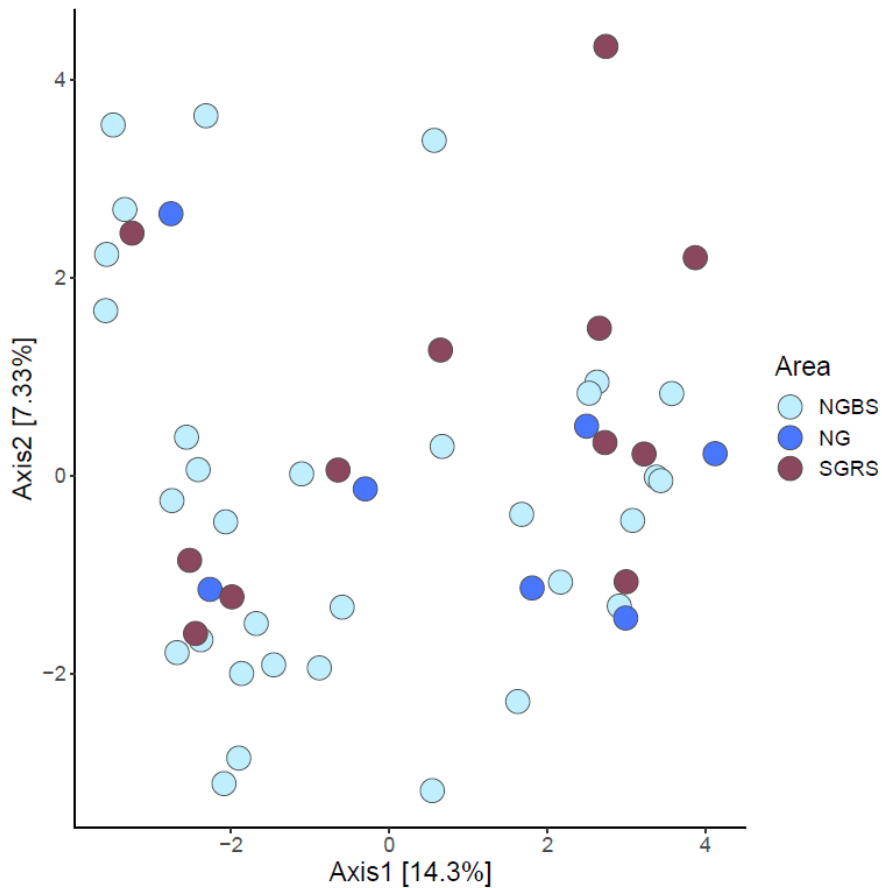


Figure 16. PCoA of Haplotypes across Collection Areas.

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. Collection areas are defined as North Guaymas, Black Smoker Chimneys 1 and 2 (NGBS); North Guaymas (NG); and South Guaymas, Ridge S (SGRS). There is no distinct correlation between collection area and symbiont genotype.

Analysis of impact of host size produced similar results by PCoA. Both PCoA of nucleotide counts (Figure 17) and haplotype (Figure 18) against *R. pachyptila* size revealed no significant correlations. These results are further supported by site specific investigations into the impact of host size on population genomic variation (Appendix 7).

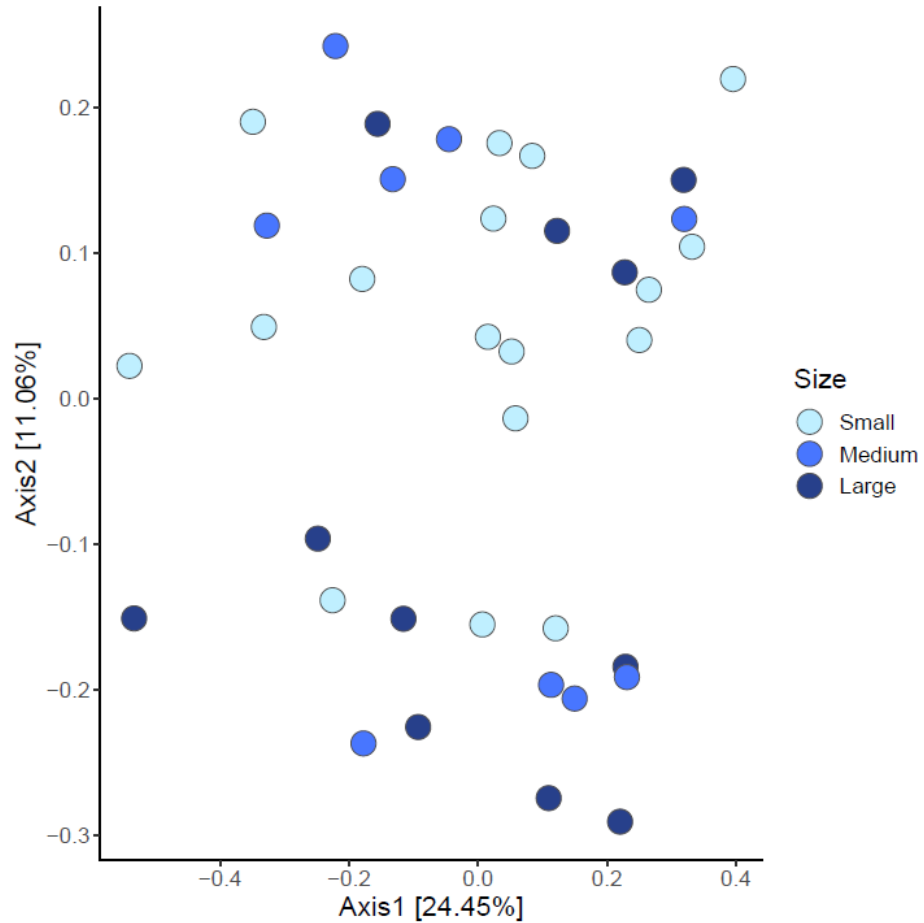


Figure 17. PCoA of Nucleotide Counts by Host Size.

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. Riftia pachyptila size is defined as 'Small' (<15 mm), 'Medium' (15 – 25 mm), or 'Large' (>25 mm). Samples with no known size information, due to storage conditions, are not included in this analysis. There is no distinct correlation between host size and symbiont nucleotide counts.

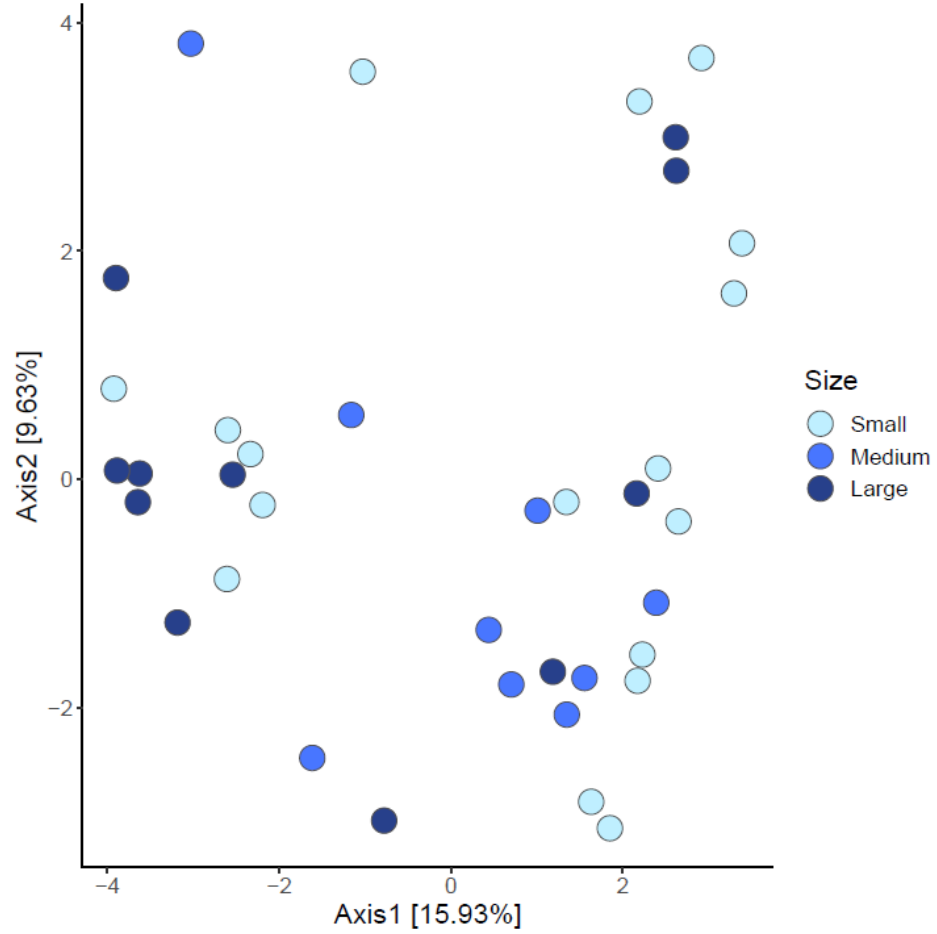


Figure 18. PCoA of Haplotype by Host Size.

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. Riftia pachyptila size is defined as 'Small' (<15 mm), 'Medium' (15 – 25 mm), or 'Large' (>25 mm). Samples with no known size information, due to storage conditions, are not included in this analysis. There is no distinct correlation between host size and symbiont genotype.

Chapter IV.

Discussion

This investigation aimed to examine the level of gene flow across populations of the symbiotic chemolithoautotrophic bacteria, *Ca. E. persephone*, in the Guaymas Basin. *Candidatus Endoriftia persephone* were investigated through the harvesting of *R. pachyptila*, across five distinct sites in the Northern and Southern Trough regions of the Guaymas Basin.

Principal co-ordinate analyses did not support any significant differences in *Ca. E. persephone* nucleotide counts or genotypes across hydrothermal vent sites. These results were supported by collection region as well, with no distinction in nucleotide counts or genotypes observed in samples collected from North Guaymas Black Smoker, North Guaymas, and South Guaymas Ridge S by PCoA. Further, fixation index analysis supported a significant degree of population overlap across all hydrothermal vent sites in the Guaymas Basin, as indicated through F_{ST} and P_{ST} values. These data support that gene flow does occur across hydrothermal vent sites, with no significant distinction between symbiont populations across all vent sites in the Guaymas Basin. Thus, it is possible that *Ca. E. persephone* dispersal does cover greater distances in the Guaymas Basin following an “island model” of dispersal; however, it may also be possible that this study region is too narrow to draw conclusions about modes of dispersal.

Multivariate analysis of *Ca. E. persephone* population genomic structure by host physiology or size did not support any correlation between host size and nucleotide count or haplotype for symbionts. This conclusion was further supported through intra-site investigation of the impact of host size on symbiont genetic structure, in which no

correlation was observed across hosts of small, medium, and large sizes from dives J2-1390 and J2-1392, respectively. Through these data, it may be concluded that host age does not impact symbiont population genomic structure for *R. pachyptila* associated *Ca. E. persephone* in the Guaymas Basin, at least among adult Riftia individuals.

Considerations for High Degree of Gene Flow in the Guaymas Basin

The low number of variants recovered across symbiont populations support observations of low genetic diversity in SNP analysis of the *Ca. E. persephone* housekeeping genes *lpxA*, *pleD* and *tufB* within regions of the EPR by Perez et al. (2021). Lack of significant geographic barriers in the study region of the Guaymas Basin may contribute to the high rate of gene flow between hydrothermal vent sites in this investigation. Moreover, the ability of *Ca. E. persephone* to adapt to both host-associated and free-living physiologies (Coykendall et al., 2011) could contribute to their ability to disperse over great distances. In all, this study supports that with adequate circulation and lack of geographic barriers, a distance of forty-three kilometers is not prohibitive to dispersal for *Ca. E. persephone* bacteria.

Previous research has shed light on the relatively high rate of *R. pachyptila* mortality in turbulent hydrothermal vent habitats (Tunnicliffe et al., 1989; Vrijenhoek, 2010; Klose et al., 2015). Tunnicliffe et al. (1989) recorded significant tubeworm mortality in the JdFR due to impact from sulfide deposits and predation. Further, low genetic diversity within and across populations of *R. pachyptila* has been attributed to frequent mortality and localized extinctions of *R. pachyptila* populations (Coykendall et al., 2011). For *R. pachyptila* populations, dispersal and genetic distinction have been

demonstrated to follow a stepping-stone model of dispersal (Black et al., 1994; Coykendall et al., 2011).

As *Ca. E. persephone* are presented with similar environmental challenges as their host, it may follow that these endosymbionts are also susceptible to low genetic diversity across distinct hydrothermal vent sites. Further, *Ca. E. persephone* demonstrate high metabolic diversity and retention of a host of genes suitable for various metabolic pathways, which may be advantageous for adapting to both host-associated and free-living lifestyles that demand genomic plasticity (Robidart et al., 2008). Additionally, the phenotypic fluidity of *Ca. E. persephone* is demonstrated through variation in morphology and function within the host, thus providing further rationale for the retention of a diverse suite of genes among *Ca. E. persephone* populations (Hinzke et al., 2021).

It can further be inferred that this symbiont genotype is well adapted to the range of environmental conditions in the Guaymas Basin over years of colonization, as no correlation was observed between host age and genetic divergence in *Ca. E. persephone* populations. Importantly, similar geochemical composition of vent effluent and endmember temperature across the Northern and Southern Troughs of the Guaymas Basin (Geilert et al., 2018) may not present adequate selective pressures for genomic divergence across these regions.

The dispersal of *Ca. E. persephone* endosymbionts post *R. pachyptila* death may pose an opportunity for contiguous genomic populations of symbiont across host ages. The horizontal transfer of *Ca. E. persephone* symbiont to *R. pachyptila* at the larval stage results in polyclonal populations of symbionts within the host (Polzin et al., 2019). These

host-specific symbiont populations are retained until an *R. pachyptila* individual becomes deceased and expels its internal endosymbiont population, thus maintaining a connection with the free-living endosymbiont population and posing the potential to colonize a juvenile organism (Nussbaumer et al., 2006; Klose et al., 2015; Polzin et al., 2019; Sato and Sasaki, 2022). This mode of symbiont uptake and release may provide an explanation for lack of *Ca. E. persephone* genetic diversity across adult host age and support an overall homogeneous and stable symbiont population in the Guaymas Basin. Potential larval selectivity of symbiont populations and subsequent sorting at the symbiont uptake stage may also drive low genetic diversity observed among adult hosts in this study.

Potential for Deep-Sea Mining to adversely impact the *Candidatus* Endoriftia persephone

Symbiont Communities in the Guaymas Basin

The changing global market has driven heightened interest in deep-sea mining (DSM) for minerals and metals, with projected demand to be two to three times greater by 2050 than current needs (Kung et al., 2021). Deep-sea sources of interest for extractive industries include cobalt-rich ferromanganese crusts, polymetallic nodules, and polymetallic sulfides (Van Dover et al., 2018; Kung et al., 2021). Hydrothermal vent ecosystems are especially vulnerable to DSM, due to the formation of rich polymetallic sulfide deposits containing desirable metals such as copper and zinc (Van Dover et al., 2018).

Kung et al. (2021) explain that governing bodies have been approaching regulation of DSM following existing structure for terrestrial mining projects; however, due to the unique properties of the deep-sea environments and hydrothermal vent systems, the authors pose that DSM warrants definition and regulation as a discrete

extractive industry. Further, existing data on deep sea and vent ecosystem structure and interactions across sites are not sufficient to accurately predict the impacts of deep-sea mining and inform regulation (Van Dover et al., 2018; Washburn et al., 2019; Kung et al., 2021); though it is noted that the likelihood of localized extinction of endemic species resulting from DSM projects is great (Van Dover et al., 2018).

As stated by Orcutt et al. (2020), the impacts of DSM on microbial ecology are poorly understood; however, the authors note that at hydrothermal vent sites, loss of microbial ecosystem function is predicted with local DSM. Thus, opportunities for investigating the ecology of hydrothermal vent systems are critical for understanding the potential impacts of DSM, informing proper regulatory framework, and protecting these unique and insurmountably important habitats.

As a source of massive polymetallic sulfide deposits and polymetallic nodules, the potential impact of DSM at hydrothermal vent sites in the Guaymas Basin must be considered (Orcutt et al., 2020; Kung et al., 2021). Since populations of *Ca. E. persephone* in the Guaymas Basin are genomically connected, it can be inferred that this gene flow amongst hydrothermal vent populations contributes to regional robustness of the *Ca. E. persephone* population. This level of connectivity may contribute to resiliency for *Ca. E. persephone* populations when challenged with disruptive activities, such as DSM; however, it is plausible that - if sufficiently expansive - DSM could lead to the eventual genetic isolation of populations of *Ca. E. persephone* and localized extinctions. The consequences of such genetic isolation remain unknown, but we can predict marked changes from the population connectivity seen amongst the current populations in the Guaymas basin.

Research Limitations

The interpretation of these data is limited by the availability of geochemical data across collection sites. Interpretation of genomic structure would benefit from vent effluent composition, temperature, and salinity data. Further, expanded sample size for the southern Guaymas Basin region would allow for more confidence in these results. Challenges to the collection of specimens and data at hydrothermal vent sites are not limited to this study alone and are representative of the limited scope typical of studies on these habitats and associated organismal communities (Dick, 2019). While these data are representative of the Guaymas Basin, distance between sites is relatively small with the maximum distance between northern and southern sites of approximately forty-three kilometers. Expanding this investigation to broader sites may provide more opportunity to investigate gene flow of *Ca. E. persephone* on a regional scale and to investigate modes of dispersal.

Future Research Directions

To understand the rate of gene flow more comprehensively across hydrothermal vent sites and systems, this study will be expanded to include EPR and Galápagos Rift hydrothermal vent systems. These data will provide a broader scale to measure population genetic similarity and distinction across vent sites for *Ca. E. persephone* endosymbionts of *R. pachyptila*. Additionally, metagenomic assembly of *R. pachyptila* genomes from coextracted samples would allow for the investigation of both symbiont and host gene flow across hydrothermal vent sites.

Transcriptome investigation may provide evidence for distinction not evident in variant analysis or gene presence and absence data alone. While genomic structure may

be similar across sites, levels of gene expression may vary. Similarly, investigation of epigenetic factors would provide insights into gene regulation across sites; thus, allowing for further understanding of similarities or differences in *Ca. E. persephone* populations across distinct hydrothermal vent sites.

Further analysis into the significance of present and absent genes across sites could lead to insights into the impact of environment on genomic structure. It would be especially beneficial to investigate the functions of the uncharacterized genes identified in the study. These findings may provide insights into why genetic homogeneity appears to exist across sites. Finally, future evaluations should more comprehensively capture geochemical metadata across collection sites to aid in the interpretation of population genomic structure.

Appendix 1.

DNA Extraction Method Validation: DNA QC

Sample purity and RNA contamination for a subset of samples were assessed to validate the DNA extraction method for this study. Sample purity was evaluated with 260/230 and 260/280 molecule absorbance ratios measured through the NanoDrop Spectrophotometer. DNA quality was observed through Agilent Genomic DNA ScreenTape Analysis (Cat Nos. 5067-5365 & 5067-5366) on the Agilent 4200 TapeStation System. The presence of RNA contamination was assessed on the Agilent 4200 TapeStation by Agilent RNA ScreenTape Analysis (Cat Nos. 5067-5576 & 5067-5577).

DNA is considered pure with a 260/280 absorbance ratio of ~1.8, while 260/230 absorbance ratios of 2.0-2.2 serve as a secondary measure of general nucleic acid purity (Thermo Fisher Scientific, N.D.). The average 260/280 value of 1.75 for this subset of samples supports high DNA purity; however, the average 260/230 value of 0.88 across samples indicates contamination (Table 6). Several sources of contamination can interfere with absorbance at 230 nm (e.g., phenols, Guanidine HCL, and carbohydrates), thus impacting 260/230 ratios. In this protocol residual Guanidine, which is present in both AL and AW1 buffers in the QIAGEN DNeasy® Blood & Tissue Kit, is the most likely source of contamination. Though this carry over contamination did cause concern, there did not appear to be any impact in library yield or sequencing performance in the pilot investigation; therefore, it was

determined that no additional pre-cleaning efforts were necessary prior to library preparation.

Table 6. NanoDrop Absorbance Ratios.

Sample ID	260/280	260/230
Worm S1_T2	1.76	0.90
Worm M1_T2	1.74	0.83
Worm L1_T2	1.75	0.90

260/280 absorbance ratios averaged 1.75 across the three replicates; 260/230 values averaged 0.88.

Agilent Genomic DNA ScreenTape Analysis supports the presence of high-quality gDNA post extraction (Figure 19). Despite efforts to remove RNA during the extraction protocol with RNase A (QIAGEN Cat No. 19101), RNA contamination was detected in samples via Agilent RNA ScreenTape Analysis (Figure 20). Though contaminating RNA can interfere with the efficiency of downstream library preparation processes, this nucleic acid material does not pose any issues for sample integrity.

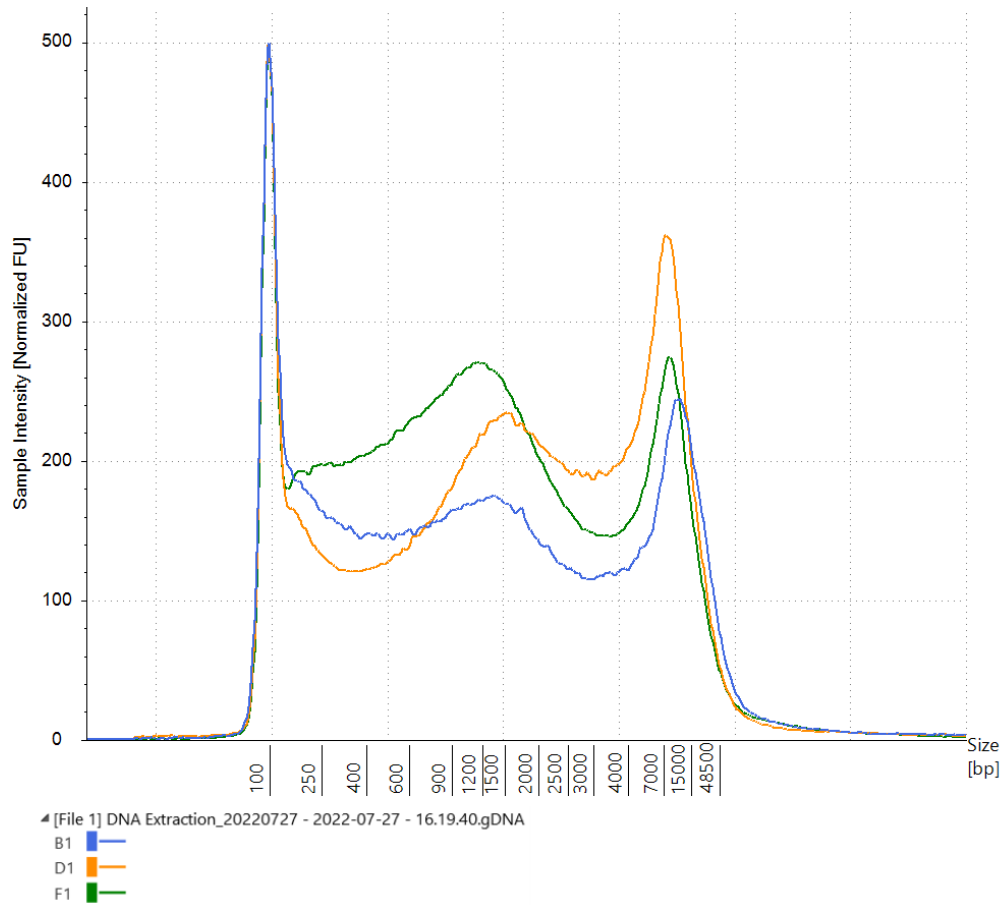


Figure 19. Agilent Genomic DNA ScreenTape Analysis.

*Electropherogram overlay of three distinct samples used in DNA extraction validation. This figure supports the presence of high-quality gDNA in all three samples post DNA extraction. Extracted sample is comprised of both *R. pachyptila* and *Ca. E. persephone* gDNA.*

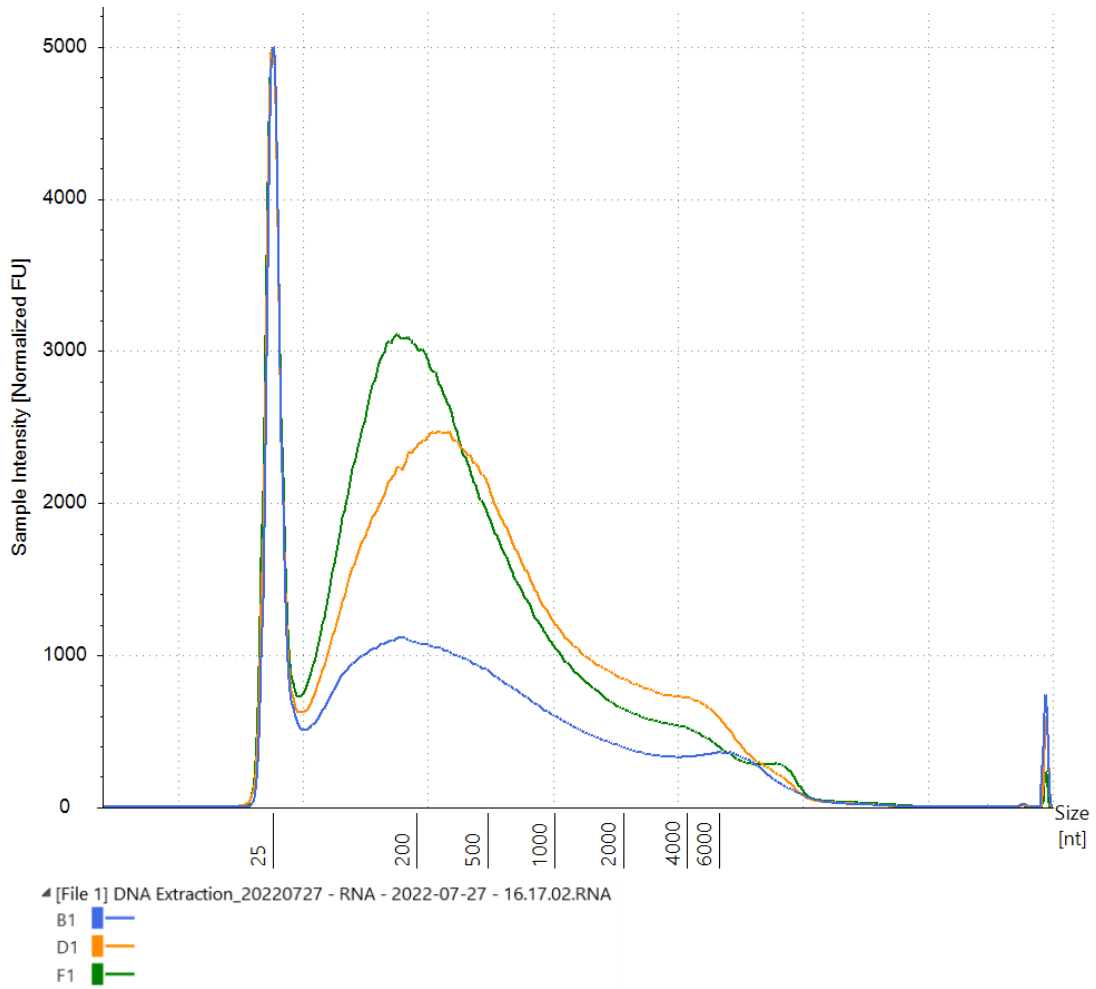


Figure 20. Agilent RNA ScreenTape Analysis.

Electropherogram overlay of three distinct samples used in DNA extraction validation. This figure confirms the presence of RNA contamination in all three extracted samples.

Appendix 2.

Sample E-gel and PicoGreen Quality Control Results from Psomagen,

MACQCREPORT_V171016

The original QC report from Psomagen is presented below. QC data reported here includes sample concentration measured through PicoGreen and projected mass, as well as 1% E-gel analysis for each sample. The E-gel analysis reports detection or absence of band (i.e., “Smear” or “No band”). In contrast to the Psomagen report of “No band” for a subset of samples, we determined that bands were visible for all samples present, but the low fluorescence samples may be degraded or lower quality.

Original Sample QC

General Information

Order Number	AN00010403	Name of Customer	Sarah Rudawsky	Date of Order	2022-09-02
--------------	------------	------------------	----------------	---------------	------------

Final QC Result of DNA sample(s)					
Arrival Date	Experiment Date	Sample count	Pass	Fail	Hold
2022-09-03	2022-09-06	61	0	0	61

Final QC Result of RNA sample(s)					
Arrival Date	Experiment Date	Sample count	Pass	Fail	Hold
N/A	N/A	N/A	N/A	N/A	N/A

The QC criteria are specified for requirements needed for a single run. Occasionally, we may encounter a shortage of sample volume or amount due to various reasons such as library construction failure or dried samples. In such case, we may notify the client and request for additional samples.

To avoid consequential delays, it is recommended to double the amount of sample, if possible.

* **Pass** : Samples automatically move forward to the next steps.

* **Hold** : A specific instruction should be given by the client for further processing.

PSOMAGEN, INC. does not proceed to the next step until we have received the client's confirmation.

* **Fail** : Samples have failed to meet all the criteria set and cannot proceed to the next step.

Sample(s) will be put on hold until further written notice from the client.

As 5 ul was taken from the sample (library) QC purposes, the indicated volume represents 5 ul less than the total volume received.

QC Result of DNA

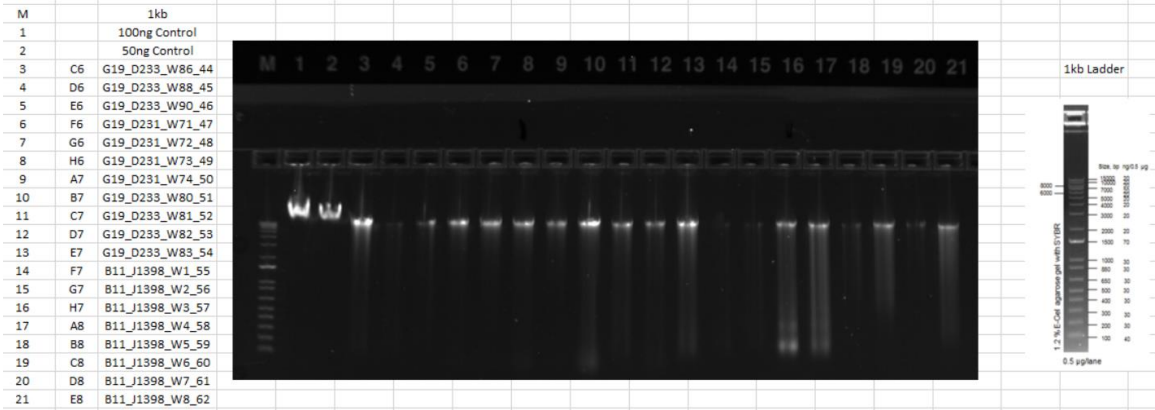
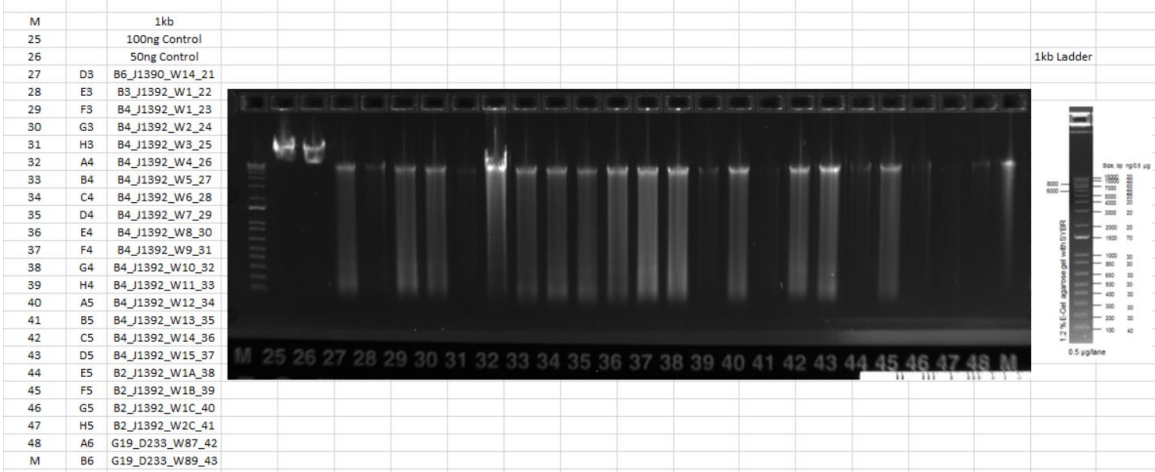
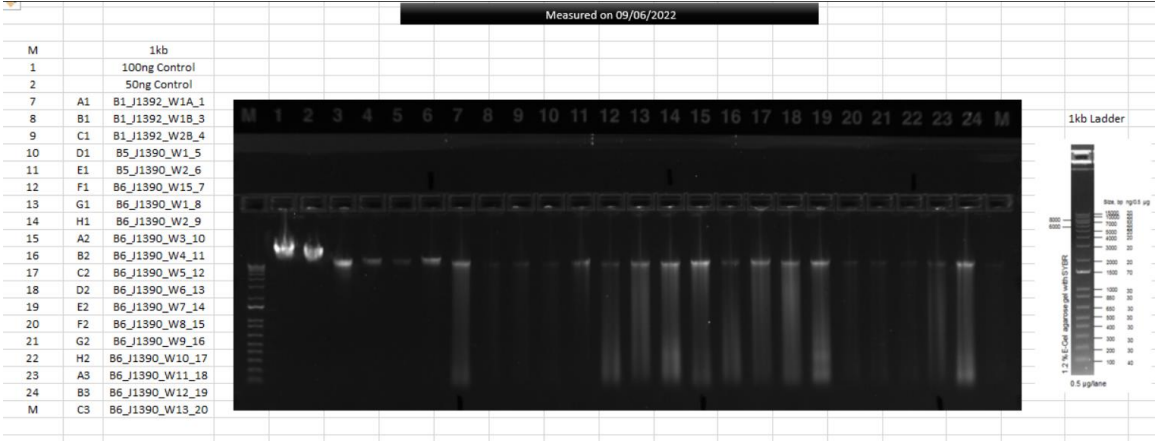
Arrival Date	2022-09-03	Experiment Date	2022-09-06	Tested by	Alice Seo
Comment					

#	Sample Name	WSID	Conc. (ng/ul)	Final Volume (ul)	Total Amount (ug)	Result*	
1	B1_J1392_W1A_1	ANW2209065004 AS0000053169	63.03	49	3.088	Hold	Smear
2	B1_J1392_W1B_3	ANW2209065004 AS0000053170	100.54	47	4.725	Hold	No band
3	B1_J1392_W2B_4	ANW2209065004 AS0000053171	156.74	48	7.524	Hold	No band
4	B5_J1390_W1_5	ANW2209065004 AS0000053172	102.28	45	4.603	Hold	No band
5	B5_J1390_W2_6	ANW2209065004 AS0000053173	327.88	35	11.476	Hold	Smear
6	B6_J1390_W15_7	ANW2209065004 AS0000053174	40.27	36	1.45	Hold	Smear
7	B6_J1390_W1_8	ANW2209065004 AS0000053175	73.49	38	2.793	Hold	Smear
8	B6_J1390_W2_9	ANW2209065004 AS0000053176	101.17	36	3.642	Hold	Smear
9	B6_J1390_W3_10	ANW2209065004 AS0000053177	57.68	24	1.384	Hold	Smear

#	Sample Name	WSID	Conc. (ng/ul)	Final Volume (ul)	Total Amount (ng)	Result*	
10	B6_J1390_W4_11	ANW220906S004 AS0000053178	36.68	47	1.724	Hold	Smear
11	B6_J1390_W5_12	ANW220906S004 AS0000053179	52.04	49	2.55	Hold	Smear
12	B6_J1390_W6_13	ANW220906S004 AS0000053180	73.49	47	3.454	Hold	Smear
13	B6_J1390_W7_14	ANW220906S004 AS0000053181	78.71	47	3.699	Hold	Smear
14	B6_J1390_W8_15	ANW220906S004 AS0000053182	161.73	47	7.601	Hold	Smear
15	B6_J1390_W9_16	ANW220906S004 AS0000053183	133.01	42	5.586	Hold	Smear
16	B6_J1390_W10_17	ANW220906S004 AS0000053184	112.32	48	5.391	Hold	Smear
17	B6_J1390_W11_18	ANW220906S004 AS0000053185	45.79	47	2.152	Hold	Smear
18	B6_J1390_W12_19	ANW220906S004 AS0000053186	82.54	48	3.962	Hold	Smear
19	B6_J1390_W13_20	ANW220906S004 AS0000053187	133.42	47	6.271	Hold	No band
20	B6_J1390_W14_21	ANW220906S004 AS0000053188	70.32	54	3.797	Hold	Smear
21	B3_J1392_W1_22	ANW220906S004 AS0000053189	314.48	46	14.466	Hold	Smear
22	B4_J1392_W1_23	ANW220906S004 AS0000053190	87.9	45	3.956	Hold	Smear
23	B4_J1392_W2_24	ANW220906S004 AS0000053191	69.19	45	3.114	Hold	Smear
24	B4_J1392_W3_25	ANW220906S004 AS0000053192	163.85	48	7.865	Hold	Smear
25	B4_J1392_W4_26	ANW220906S004 AS0000053193	246.34	46	11.332	Hold	Smear
26	B4_J1392_W5_27	ANW220906S004 AS0000053194	63.74	46	2.932	Hold	Smear
27	B4_J1392_W6_28	ANW220906S004 AS0000053195	58.58	47	2.753	Hold	Smear
28	B4_J1392_W7_29	ANW220906S004 AS0000053196	68.52	48	3.289	Hold	Smear
29	B4_J1392_W8_30	ANW220906S004 AS0000053197	82.82	46	3.81	Hold	Smear
30	B4_J1392_W9_31	ANW220906S004 AS0000053198	109.97	48	5.279	Hold	Smear
31	B4_J1392_W10_32	ANW220906S004 AS0000053199	103.48	50	5.174	Hold	Smear
32	B4_J1392_W11_33	ANW220906S004 AS0000053200	256.63	51	13.088	Hold	Smear
33	B4_J1392_W12_34	ANW220906S004 AS0000053201	75.16	49	3.683	Hold	Smear
34	B4_J1392_W13_35	ANW220906S004 AS0000053202	137.47	46	6.324	Hold	No band
35	B4_J1392_W14_36	ANW220906S004 AS0000053203	68.43	45	3.079	Hold	Smear
36	B4_J1392_W15_37	ANW220906S004 AS0000053204	97.54	46	4.487	Hold	Smear
37	B2_J1392_W1A_38	ANW220906S004 AS0000053205	320.6	54	17.312	Hold	No band
38	B2_J1392_W1B_39	ANW220906S004 AS0000053206	97.08	48	4.66	Hold	Smear
39	B2_J1392_W1C_40	ANW220906S004 AS0000053207	229.68	47	10.795	Hold	No band
40	B2_J1392_W2C_41	ANW220906S004 AS0000053208	119.47	53	6.332	Hold	No band
41	G19_D233_W87_42	ANW220906S004 AS0000053209	131.22	46	6.036	Hold	No band
42	G19_D233_W89_43	ANW220906S004 AS0000053210	82.41	52	4.285	Hold	Smear
43	G19_D233_W86_44	ANW220906S004 AS0000053211	79.65	49	3.903	Hold	Smear
44	G19_D233_W88_45	ANW220906S004 AS0000053212	129.72	49	6.356	Hold	No band
45	G19_D233_W90_46	ANW220906S004 AS0000053213	316.79	47	14.889	Hold	Smear

#	Sample Name	WSID	Conc. (ng/ul)	Final Volume (ul)	Total Amount (ug)	Result*	
46	G19_D231_W71_47	ANW220906S004 AS0000053214	32.55	49	1.595	Hold	Smear
47	G19_D231_W72_48	ANW220906S004 AS0000053215	34.76	48	1.668	Hold	Smear
48	G19_D231_W73_49	ANW220906S004 AS0000053216	30.01	49	1.47	Hold	Smear
49	G19_D231_W74_50	ANW220906S004 AS0000053217	30.96	49	1.517	Hold	Smear
50	G19_D233_W80_51	ANW220906S004 AS0000053218	64.68	46	2.975	Hold	Smear
51	G19_D233_W81_52	ANW220906S004 AS0000053219	32.42	47	1.524	Hold	Smear
52	G19_D233_W82_53	ANW220906S004 AS0000053220	26.61	47	1.251	Hold	Smear
53	G19_D233_W83_54	ANW220906S004 AS0000053221	78.33	47	3.682	Hold	No band
54	B11_J1398_W1_55	ANW220906S004 AS0000053222	52.47	47	2.466	Hold	No band
55	B11_J1398_W2_56	ANW220906S004 AS0000053223	124.91	48	5.996	Hold	Smear
56	B11_J1398_W3_57	ANW220906S004 AS0000053224	60.56	48	2.907	Hold	Smear
57	B11_J1398_W4_58	ANW220906S004 AS0000053225	75.27	49	3.688	Hold	Smear
58	B11_J1398_W5_59	ANW220906S004 AS0000053226	147.28	47	6.922	Hold	Smear
59	B11_J1398_W6_60	ANW220906S004 AS0000053227	41.49	49	2.033	Hold	Smear
60	B11_J1398_W7_61	ANW220906S004 AS0000053228	138.65	46	6.378	Hold	Smear
61	B11_J1398_W8_62	ANW220906S004 AS0000053229	53.57	47	2.518	Hold	Smear

Experiment Condition	1% E-Gel
----------------------	----------



Appendix 3.

Sequencing Pilot Study

This pilot study aimed to determine the required metagenomic sequencing output for adequate coverage of both *Ca. E. persephone* and *R. pachyptila* genomes through determining the ratio of host to symbiont DNA for a subset of samples. Adequate sequencing depth of the *Ca. E. persephone* genome requires at least 10x coverage. Further, a sequencing depth of 5x for the *R. pachyptila* genome was desired for suitable coverage of the host genome for future studies.

Samples were selected across three size ranges – small (<15 mm), medium (15 – 25 mm), and large (>25 mm) – based on trunk size to control for variability in symbiont population related to host size or age. Sample dissection and DNA extraction were conducted as described in the methods. Library preparation and Illumina MiSeq 500-cycle sequencing were executed at the Rhode Island IDeA Network of Biomedical Research Excellence (RI INBRE) Molecular Informatics Core. Library preparation was facilitated with IntegenX PrepX reagents. Post-sequencing, symbiont to host proportions were determined by comparing the metagenomic reads against both *Ca. E. persephone* and *R. pachyptila* reference genomes with BBSplit (Joint Genome Institute, N.D.).

This investigation resulted in symbiont proportions ranging from 52.19% - 73.81% across the three samples (Mean = 64.53%) (Table 7). Based on these data, we conservatively projected that required sequencing output should account for 50% symbiont DNA proportion.

Table 7. Host to Symbiont DNA Proportions.

Sample	Number Endoriftia reads	Number <i>Riftia</i> reads	Symbiont proportion (%)
Small	2,478,578	1,187,735	67.60
Medium	3,235,305	1,147,769	73.81
Large	2,309,072	2,115,639	52.19

Total reads for Ca. E. persephone and R. pachyptila. Metagenomic libraries were prepared for three R. pachyptila samples of different sizes. Mean symbiont proportion is 64.53%.

For a whole-genome sequencing approach, it was determined that 5,721,428,572 bases of output are required per sample. A total of 349,007,142,892 bases were estimated to be required across the sixty-one samples used in this study. Thus, it was calculated that a total of 350 Gb was needed for 5x coverage of the *R. pachyptila* genome, which allows for deep coverage of the *Ca. E. persephone* genome (Appendix 4).

Appendix 4.

Sequencing Coverage Calculations

Symbiont to host gDNA ratios for a subset of representative samples were determined through a pilot sequencing investigation. The Illumina Sequencing Coverage Calculator tool was used to determine required output for 5x host coverage (Illumina, 2022).

The following information was input into the Illumina Sequencing Coverage Calculator for DNA Applications:

- Application or product: Whole-Genome Sequencing
- Coverage (x) = 5
- Duplicates (%) = 2
- Genome or Region Size (i.e., host genome size) = 560.7 Mbases
- Read Length = 300 bp
- Production-Scale Sequencers = NovaSeq 6000

Output required (bases) from the tool, and scaled calculations to account for the full sample set and 50% symbiont DNA proportion are reported in Table 8, below. Using the Illumina Sequencing Coverage Calculator for DNA Applications, *Ca. E. persephone* genome size of 7.2 Mbases, and equivalent run information, it was determined that 7,346,939 bases are required for 1x symbiont coverage in sequencing. Therefore, based on the requirements reported below, 389x coverage of symbiont genome can be expected from 5x coverage of host whole-genome sequencing.

Table 8. Required Output for Host and Symbiont Metagenome Sequencing.

	Host	Symbiont
Output Required (bases)	2,860,714,286	5,721,428,572
Total Output (61 samples) (bases)	174,503,571,446	349,007,142,892
Total Output (61 samples) (Mb)	174503.57	349007.14
Total Output (61 samples) (Gb)	174.50	349.01

Required sequencing output (bases) from host as reported by the online Illumina Sequencing Coverage Calculator are scaled up to account for full sample set of sixty-one samples; and a 50% increase in required bases to accommodate 50% symbiont DNA. Total output is additionally reported in total Gb required.

Appendix 5.

Sequencing Raw Data Report

Raw data output from 150bp paired-end sequencing on the Illumina NovaSeq 6000 platform (Table 9). Library preparation and sequencing was performed at Psomagen and results below were provided in the project report from Psomagen.

Table 9. Sequencing Raw Data Metrics.

Sample ID	Total Read Bases (bp)	Total Read Count	GC (%)	AT (%)	Q20 (%)	Q30 (%)
B11_J1398_W1_55	4,776,871,712	31,634,912	52.78	47.22	94.60	87.89
B11_J1398_W2_56	7,990,942,952	52,920,152	46.29	53.71	93.80	86.80
B11_J1398_W3_57	5,379,207,390	35,623,890	56.40	43.60	94.73	87.99
B11_J1398_W4_58	5,902,593,624	39,090,024	53.21	46.79	94.80	88.27
B11_J1398_W5_59	6,033,582,198	39,957,498	53.67	46.33	94.65	87.93
B11_J1398_W6_60	7,625,397,018	50,499,318	52.18	47.82	93.95	86.85
B11_J1398_W7_61	9,958,390,204	65,949,604	44.48	55.52	93.14	85.76
B11_J1398_W8_62	6,965,610,370	46,129,870	53.03	46.97	94.78	88.13
B1_J1392_W1A_1	7,241,379,556	47,956,156	53.36	46.64	93.74	86.41
B1_J1392_W1B_3	6,484,178,580	42,941,580	56.09	43.91	94.66	87.86
B1_J1392_W2B_4	7,160,294,368	47,419,168	48.89	51.11	93.43	86.22
B2_J1392_W1A_38	8,766,050,112	58,053,312	45.76	54.24	93.31	85.96
B2_J1392_W1B_39	5,493,045,988	36,377,788	54.55	45.45	94.94	88.43
B2_J1392_W1C_40	5,435,786,184	35,998,584	47.98	52.02	94.19	87.31
B2_J1392_W2C_41	3,303,092,082	21,874,782	53.74	46.26	94.75	88.15
B3_J1392_W1_22	11,265,848,468	74,608,268	45.05	54.95	92.40	84.69
B4_J1392_W10_32	5,376,339,296	35,604,896	57.57	42.43	94.56	87.60
B4_J1392_W11_33	9,726,804,524	64,415,924	44.82	55.18	92.85	85.45
B4_J1392_W12_34	5,222,487,510	34,586,010	57.50	42.50	94.93	88.30
B4_J1392_W1_23	5,529,555,976	36,619,576	55.50	44.50	94.24	87.24
B4_J1392_W13_35	4,683,030,044	31,013,444	53.31	46.69	94.94	88.50
B4_J1392_W14_36	3,619,207,260	23,968,260	55.83	44.17	94.90	88.32
B4_J1392_W15_37	5,312,991,474	35,185,374	54.88	45.12	94.86	88.21
B4_J1392_W2_24	7,976,265,752	52,822,952	54.61	45.39	94.54	87.74
B4_J1392_W3_25	10,133,471,986	67,109,086	52.05	47.95	93.87	86.74
B4_J1392_W4_26	14,774,928,106	97,847,206	44.76	55.24	92.82	85.22
B4_J1392_W5_27	7,580,620,686	50,202,786	56.20	43.80	94.14	86.94

Sample ID	Total Read Bases (bp)	Total Read Count	GC (%)	AT (%)	Q20 (%)	Q30 (%)
B4_J1392_W6_28	5,564,794,846	36,852,946	56.65	43.35	94.40	87.36
B4_J1392_W7_29	6,549,118,848	43,371,648	56.06	43.94	94.27	87.24
B4_J1392_W8_30	7,055,160,014	46,722,914	54.03	45.97	94.04	86.93
B4_J1392_W9_31	5,713,081,980	37,834,980	56.37	43.63	94.43	87.41
B5_J1390_W1_5	7,930,409,770	52,519,270	54.36	45.64	94.31	87.38
B5_J1390_W2_6	16,095,299,890	106,591,390	43.27	56.73	92.15	84.35
B6_J1390_W10_17	6,695,850,078	44,343,378	57.58	42.42	94.75	87.96
B6_J1390_W11_18	4,398,299,914	29,127,814	48.57	51.43	93.44	86.21
B6_J1390_W12_19	4,951,651,796	32,792,396	57.70	42.30	94.68	87.80
B6_J1390_W13_20	4,463,293,636	29,558,236	48.51	51.49	93.39	86.15
B6_J1390_W14_21	5,032,149,896	33,325,496	55.71	44.29	94.11	86.94
B6_J1390_W15_7	5,002,120,526	33,126,626	48.31	51.69	93.15	85.83
B6_J1390_W1_8	6,631,351,032	43,916,232	49.86	50.14	93.13	85.65
B6_J1390_W2_9	5,535,293,372	36,657,572	56.0	44.0	94.52	87.58
B6_J1390_W3_10	7,964,810,288	52,747,088	49.58	50.42	93.27	85.80
B6_J1390_W4_11	4,507,283,862	29,849,562	52.54	47.46	93.36	86.07
B6_J1390_W5_12	8,815,850,818	58,383,118	52.53	47.47	94.07	87.04
B6_J1390_W6_13	5,752,931,484	38,098,884	55.13	44.87	94.19	87.13
B6_J1390_W7_14	4,945,994,128	32,754,928	51.72	48.28	93.41	86.10
B6_J1390_W8_15	5,686,513,228	37,659,028	57.22	42.78	94.52	87.56
B6_J1390_W9_16	4,016,660,098	26,600,398	57.35	42.65	94.38	87.36
G19_D231_W71_47	15,467,916,936	102,436,536	52.71	47.29	94.10	86.89
G19_D231_W72_48	17,028,200,238	112,769,538	51.86	48.14	94.40	87.37
G19_D231_W73_49	18,546,902,670	122,827,170	54.21	45.79	94.62	87.78
G19_D231_W74_50	14,678,551,450	97,208,950	52.17	47.83	94.48	87.50
G19_D233_W80_51	11,330,956,950	75,039,450	54.72	45.28	94.49	87.45
G19_D233_W81_52	11,372,864,584	75,316,984	54.67	45.33	94.35	87.22
G19_D233_W82_53	10,277,718,360	68,064,360	51.94	48.06	94.32	87.26
G19_D233_W83_54	9,220,510,584	61,062,984	48.15	51.85	93.92	86.83
G19_D233_W86_44	10,723,512,640	71,016,640	48.86	51.14	93.95	86.85
G19_D233_W87_42	13,231,393,724	87,625,124	48.05	51.95	94.01	86.96
G19_D233_W88_45	7,269,736,752	48,143,952	54.25	45.75	94.98	88.43
G19_D233_W89_43	8,650,670,408	57,289,208	49.35	50.65	94.20	87.27
G19_D233_W90_46	12,016,029,756	79,576,356	46.97	53.03	93.92	86.78

Total Read Bases represent the total number of bases sequenced. Total Reads are the total number of reads across read 1 and read 2 for paired-end sequencing. GC(%) = GC content; AT(%) = AT content; Q20(%) = ratio of bases with a phred quality score of 20 or greater; Q30(%) = ratio of bases with a phred quality score of 30 or greater.

Appendix 6.

Uncharacterized Genes with Variable Presence and Absence across Hydrothermal Vent Sites

Table 10. *Candidatus* Endoriftia persephone Gene Annotations for Uncharacterized Genes across Hydrothermal Vent Sites, Genes Universally Present Removed.

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1015	UniRef90_G2DCF5 Uncharacterized protein	100.00	100.00	92.86	100.00	100.00
group_1017	UniRef90_G2DHF4 Uncharacterized protein	100.00	94.44	100.00	100.00	100.00
group_1023	UniRef90_G2FBJ1 Uncharacterized protein	71.43	66.67	92.86	66.67	77.78
group_1033	UniRef90_A0A0T5YVR2 Uncharacterized protein	100.00	77.78	42.86	33.33	33.33
group_1036	UniRef90_G2DDN6 Uncharacterized protein	71.43	38.89	64.29	66.67	66.67
group_1037	UniRef90_G2DDB2 Uncharacterized protein	85.71	100.00	85.71	100.00	100.00
group_1042	UniRef90_A0A2V9BKJ7 Uncharacterized protein	85.71	50.00	35.71	100.00	55.56
group_1043	Hypothetical protein	42.86	22.22	57.14	33.33	33.33
group_1044	UniRef90_UPI0001699869 hypothetical protein	100.00	72.22	71.43	100.00	77.78
group_1046	UniRef90_G2D8Y2 Uncharacterized protein	100.00	94.44	100.00	100.00	100.00
group_1048	Hypothetical protein	100.00	100.00	92.86	100.00	100.00
group_1049	UniRef90_A0A0T5YYY2 Uncharacterized protein	85.71	72.22	71.43	66.67	88.89
group_1058	UniRef90_G2DBQ7 Uncharacterized protein	100.00	100.00	100.00	100.00	88.89
group_1060	UniRef90_A0A0T5YUD4 Uncharacterized protein	100.00	94.44	100.00	100.00	100.00
group_1062	Hypothetical protein	100.00	100.00	100.00	100.00	77.78
group_1068	UniRef90_UPI001110552B hypothetical protein	100.00	88.89	100.00	100.00	100.00
group_1074	UniRef90_G2FD95 Uncharacterized protein	100.00	100.00	92.86	100.00	100.00
group_1076	UniRef90_G2DCV1 Uncharacterized protein	57.14	33.33	14.29	33.33	77.78
group_1079	UniRef90_G2FHG3 Uncharacterized protein	100.00	100.00	92.86	100.00	100.00

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1084	UniRef90_G2FD88 Uncharacterized protein	100.00	94.44	100.00	100.00	100.00
group_1086	UniRef90_G2FE50 Uncharacterized protein	100.00	88.89	92.86	100.00	66.67
group_1091	UniRef90_A0A539EBA8 Uncharacterized protein	71.43	77.78	85.71	100.00	88.89
group_1092	UniRef90_G2DBP0 Uncharacterized protein	71.43	88.89	71.43	100.00	100.00
group_1099	UniRef90_G2DH82 Uncharacterized protein	100.00	83.33	85.71	100.00	100.00
group_1101	UniRef90_G2FDC0 Uncharacterized protein	42.86	38.89	50.00	100.00	77.78
group_1104	UniRef90_A0A0T5Z7J8 Uncharacterized protein	0.00	5.56	7.14	33.33	11.11
group_1106	UniRef90_G2DFR5 Uncharacterized protein	28.57	44.44	50.00	66.67	44.44
group_1108	Hypothetical protein	0.00	5.56	14.29	0.00	11.11
group_1113	UniRef90_G2FFK4 Uncharacterized protein	85.71	83.33	100.00	66.67	66.67
group_1114	UniRef90_G2FE86 Uncharacterized protein	100.00	100.00	100.00	66.67	100.00
group_1115	UniRef90_A0A0T5YV24 Uncharacterized protein	71.43	77.78	92.86	100.00	88.89
group_1116	UniRef90_G2FGZ6 Uncharacterized protein	57.14	72.22	71.43	66.67	33.33
group_1117	UniRef90_G2DFQ2 Uncharacterized protein	85.71	72.22	78.57	66.67	100.00
group_1119	UniRef90_G2DDN9 Uncharacterized protein	28.57	44.44	28.57	33.33	44.44
group_1120	UniRef90_G2FDE0 Uncharacterized protein	57.14	77.78	42.86	66.67	55.56
group_1121	UniRef90_G2FDY1 Uncharacterized protein	57.14	88.89	100.00	66.67	77.78
group_1123	UniRef90_G2DH85 Uncharacterized protein	85.71	66.67	92.86	100.00	77.78
group_1124	UniRef90_A0A0T5YSV8 Uncharacterized protein	71.43	83.33	78.57	33.33	44.44
group_1125	UniRef90_G2D9E4 Uncharacterized protein	85.71	94.44	100.00	100.00	100.00
group_1126	UniRef90_G2DC89 Uncharacterized protein	85.71	88.89	92.86	100.00	77.78
group_1127	Hypothetical protein	28.57	0.00	7.14	0.00	22.22
group_1128	UniRef90_A0A0T5Z8J3 Uncharacterized protein	100.00	66.67	92.86	33.33	77.78
group_1131	UniRef90_A0A426W631 Uncharacterized protein	85.71	77.78	78.57	33.33	100.00
group_1133	UniRef90_G2FDE1 Uncharacterized protein	28.57	11.11	0.00	66.67	11.11
group_1135	UniRef90_I3Y972 Uncharacterized protein	42.86	33.33	35.71	0.00	11.11
group_1136	Hypothetical protein	100.00	88.89	92.86	100.00	100.00
group_1137	UniRef90_G2DAK5 Uncharacterized protein	14.29	0.00	21.43	0.00	0.00

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1138	UniRef90_G2FGW1 Uncharacterized protein	14.29	11.11	7.14	0.00	0.00
group_1140	UniRef90_A0A0T5Z417 Uncharacterized protein	85.71	55.56	85.71	100.00	66.67
group_1141	UniRef90_G2FE49 Uncharacterized protein	100.00	100.00	100.00	100.00	88.89
group_1143	UniRef90_G2DDQ4 Uncharacterized protein	100.00	100.00	92.86	100.00	100.00
group_1147	UniRef90_A0A0T5ZBT4 Uncharacterized protein	57.14	27.78	50.00	66.67	22.22
group_1150	Hypothetical protein	14.29	0.00	0.00	0.00	0.00
group_1152	UniRef90_G2D913 Uncharacterized protein	28.57	22.22	35.71	0.00	55.56
group_1153	UniRef90_G2DGT1 Uncharacterized protein	71.43	66.67	92.86	100.00	88.89
group_1154	UniRef90_G2FCM1 Uncharacterized protein	71.43	83.33	85.71	100.00	55.56
group_1155	UniRef90_G2DH39 Uncharacterized protein	0.00	5.56	7.14	0.00	0.00
group_1156	UniRef90_G2D974 Uncharacterized protein	85.71	88.89	92.86	66.67	66.67
group_1157	UniRef90_G2DHZ6 Uncharacterized protein	100.00	88.89	100.00	66.67	88.89
group_1158	UniRef90_G2FIN4 Uncharacterized protein	0.00	11.11	7.14	0.00	11.11
group_1159	UniRef90_G2FGD5 Uncharacterized protein	42.86	50.00	71.43	66.67	77.78
group_1160	UniRef90_A0A0T5Z698 Uncharacterized protein	57.14	33.33	14.29	66.67	22.22
group_1161	UniRef90_G2FGK7 Uncharacterized protein	85.71	100.00	100.00	66.67	77.78
group_1162	UniRef90_G2DH78 Uncharacterized protein	42.86	22.22	42.86	100.00	33.33
group_1163	UniRef90_A0A0T5ZBQ6 Uncharacterized protein	71.43	88.89	85.71	66.67	55.56
group_1167	UniRef90_G2FE11 Uncharacterized protein	85.71	55.56	64.29	66.67	55.56
group_1168	UniRef90_G2DFZ9 Uncharacterized protein	0.00	38.89	35.71	33.33	22.22
group_1169	UniRef90_A0A0T5YXQ4 Uncharacterized protein	0.00	11.11	0.00	0.00	11.11
group_1171	UniRef90_G2FCA3 Uncharacterized protein	0.00	11.11	14.29	33.33	11.11
group_1172	UniRef90_G2FD92 Uncharacterized protein	85.71	100.00	100.00	100.00	100.00
group_1173	UniRef90_A0A0T5Z514 Uncharacterized protein	42.86	61.11	85.71	33.33	55.56
group_1174	UniRef90_A0A0T5Z354 Uncharacterized protein	0.00	0.00	7.14	0.00	0.00
group_1176	UniRef90_G2DFD9 Uncharacterized protein	71.43	83.33	50.00	66.67	44.44
group_1179	UniRef90_A0A0T5Z2B5 Uncharacterized protein	14.29	5.56	14.29	0.00	0.00
group_1180	UniRef90_D5BY90 Uncharacterized protein	14.29	27.78	7.14	0.00	22.22

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1182	UniRef90_G2DDV4 Uncharacterized protein	100.00	88.89	78.57	33.33	66.67
group_1183	UniRef90_G2FG57 Uncharacterized protein	0.00	11.11	14.29	33.33	11.11
group_1185	UniRef90_G2DH77 Uncharacterized protein	28.57	11.11	14.29	66.67	11.11
group_1186	Hypothetical protein	28.57	16.67	14.29	100.00	33.33
group_1187	UniRef90_A0A4U0YKL5 Uncharacterized protein	14.29	38.89	50.00	66.67	44.44
group_1188	UniRef90_A0A0T5Z965 Uncharacterized protein	71.43	16.67	7.14	0.00	11.11
group_1189	UniRef90_A0A0T5Z8X7 Uncharacterized protein	28.57	22.22	14.29	0.00	33.33
group_1190	UniRef90_G2DAB3 Uncharacterized protein	14.29	0.00	21.43	0.00	0.00
group_1191	UniRef90_A0A0T5YSQ4 Uncharacterized protein	0.00	0.00	0.00	0.00	11.11
group_1192	UniRef90_G2FAV9 Uncharacterized protein	0.00	0.00	7.14	0.00	11.11
group_1194	UniRef90_G2DAG6 Uncharacterized protein	14.29	0.00	0.00	0.00	0.00
group_1196	UniRef90_UPI0002FE5D78 hypothetical protein	0.00	0.00	0.00	0.00	11.11
group_1197	UniRef90_G2DG89 Uncharacterized protein	0.00	0.00	7.14	0.00	22.22
group_1198	UniRef90_G2FBR3 Uncharacterized protein	14.29	5.56	0.00	0.00	11.11
group_1202	UniRef90_G2DCI2 Uncharacterized protein	0.00	0.00	0.00	0.00	11.11
group_1203	UniRef90_A0A0T5Z5W6 Uncharacterized protein	28.57	22.22	21.43	0.00	0.00
group_1208	Hypothetical protein	0.00	5.56	7.14	0.00	0.00
group_1209	UniRef90_A0A0T5Z8J3 Uncharacterized protein	42.86	38.89	57.14	0.00	77.78
group_1214	Hypothetical protein	42.86	11.11	21.43	0.00	0.00
group_1215	UniRef90_G2D9N8 Uncharacterized protein	0.00	5.56	7.14	0.00	11.11
group_1216	UniRef90_G2DGS0 Uncharacterized protein	0.00	5.56	0.00	33.33	11.11
group_1217	UniRef90_G2FB59 Uncharacterized protein	14.29	38.89	7.14	0.00	0.00
group_1219	Hypothetical protein	0.00	0.00	0.00	0.00	11.11
group_1223	Hypothetical protein	14.29	0.00	0.00	33.33	0.00
group_1257	UniRef90_A0A0T5Z1V0 Uncharacterized protein	85.71	66.67	78.57	66.67	88.89
group_1262	UniRef90_A0A0T5Z6A0 Uncharacterized protein	85.71	72.22	92.86	100.00	77.78
group_1265	UniRef90_UPI0003FF14FD hypothetical protein	71.43	77.78	57.14	33.33	55.56
group_1322	UniRef90_G2DC94 Uncharacterized protein	85.71	100.00	100.00	100.00	100.00

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1324	Hypothetical protein	14.29	33.33	42.86	100.00	66.67
group_1327	UniRef90_G2D956 Uncharacterized protein	100.00	83.33	100.00	100.00	88.89
group_1328	UniRef90_A0A0T5YYK6 Uncharacterized protein	100.00	100.00	92.86	100.00	88.89
group_1336	UniRef90_UPI00111095B0 hypothetical protein	28.57	5.56	14.29	0.00	0.00
group_1449	UniRef90_G2DDZ9 Uncharacterized protein	42.86	55.56	85.71	0.00	22.22
group_1456	UniRef90_G2FJB9 Uncharacterized protein	42.86	50.00	42.86	0.00	33.33
group_1459	UniRef90_G2FFL8 Uncharacterized protein	28.57	0.00	7.14	0.00	0.00
group_1465	UniRef90_G2FJY0 Uncharacterized protein	100.00	100.00	92.86	33.33	100.00
group_1472	UniRef90_A0A0T5YZU0 Uncharacterized protein	71.43	83.33	100.00	100.00	100.00
group_1473	UniRef90_G2DG40 Uncharacterized protein	85.71	77.78	78.57	66.67	77.78
group_1477	UniRef90_A0A0T5YUT6 Uncharacterized protein	100.00	94.44	100.00	100.00	100.00
group_1483	UniRef90_G2DEC6 Uncharacterized protein	71.43	83.33	92.86	100.00	88.89
group_1493	UniRef90_G2FFK6 Uncharacterized protein	0.00	0.00	7.14	33.33	55.56
group_1497	UniRef90_G2FFK7 Uncharacterized protein	0.00	11.11	14.29	0.00	11.11
group_1498	UniRef90_G2DES3 Uncharacterized protein	85.71	94.44	100.00	100.00	77.78
group_1500	UniRef90_A0A0T5YUM7 Uncharacterized protein	57.14	22.22	57.14	100.00	100.00
group_1501	UniRef90_G2DH44 Uncharacterized protein	0.00	5.56	7.14	0.00	33.33
group_1503	UniRef90_G2DFK8 Uncharacterized protein	85.71	83.33	78.57	66.67	66.67
group_1504	UniRef90_G2DGL6 Uncharacterized protein	100.00	72.22	92.86	66.67	77.78
group_1505	UniRef90_G2FJX7 Uncharacterized protein	100.00	83.33	85.71	100.00	100.00
group_1506	UniRef90_A0A0T5Z0P1 Uncharacterized protein	0.00	0.00	0.00	0.00	11.11
group_1508	UniRef90_A0A0T5Z271 Uncharacterized protein	71.43	94.44	100.00	66.67	77.78
group_1509	UniRef90_G2FFL0 Uncharacterized protein	71.43	61.11	64.29	100.00	77.78
group_1510	UniRef90_G2DF17 Uncharacterized protein	28.57	33.33	7.14	33.33	11.11
group_1511	UniRef90_G2DGM1 Uncharacterized protein	0.00	22.22	14.29	0.00	22.22
group_1512	UniRef90_A0A0T5YZJ2 Uncharacterized protein	57.14	33.33	35.71	33.33	22.22
group_1519	UniRef90_G2FJ00 Uncharacterized protein	71.43	22.22	21.43	0.00	22.22
group_1520	UniRef90_A0A0T5YY82 Uncharacterized protein	71.43	77.78	85.71	100.00	100.00

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_1521	UniRef90_A0A0T5YY66 Uncharacterized protein	42.86	5.56	0.00	0.00	11.11
group_1563	UniRef90_G2FJ08 Uncharacterized protein	85.71	94.44	78.57	100.00	100.00
group_1579	UniRef90_A0A401TIH0 Uncharacterized protein	100.00	94.44	92.86	100.00	88.89
group_1580	Hypothetical protein	100.00	94.44	92.86	100.00	88.89
group_1581	Hypothetical protein	14.29	5.56	14.29	0.00	0.00
group_1582	Hypothetical protein	28.57	11.11	7.14	0.00	0.00
group_1583	Hypothetical protein	42.86	44.44	57.14	0.00	55.56
group_1585	Hypothetical protein	100.00	100.00	100.00	100.00	88.89
group_1588	Hypothetical protein	71.43	38.89	57.14	0.00	55.56
group_1677	Hypothetical protein	100.00	100.00	92.86	100.00	100.00
group_216	UniRef90_G2DAQ5 Uncharacterized protein	0.00	5.56	14.29	33.33	0.00
group_588	Hypothetical protein	100.00	88.89	100.00	100.00	100.00
group_710	UniRef90_UPI001112004F hypothetical protein	100.00	94.44	100.00	100.00	100.00
group_735	UniRef90_G2D9J8 Uncharacterized protein	85.71	100.00	100.00	100.00	100.00
group_817	UniRef90_G2D957 Uncharacterized protein	100.00	100.00	92.86	100.00	100.00
group_855	UniRef90_G2DD11 Uncharacterized protein	100.00	94.44	85.71	100.00	88.89
group_883	UniRef90_G2DGU7 Uncharacterized protein	100.00	94.44	92.86	100.00	100.00
group_893	UniRef90_UPI00016986E1 hypothetical protein	100.00	77.78	35.71	100.00	77.78
group_908	UniRef90_G2DCI6 Uncharacterized protein	100.00	100.00	92.86	100.00	100.00
group_919	UniRef90_A0A419DAW6 Uncharacterized protein	71.43	77.78	78.57	66.67	66.67
group_920	UniRef90_UPI000587CB90 hypothetical protein	100.00	94.44	92.86	100.00	88.89
group_924	UniRef90_A0A0T5Z5Y6 Uncharacterized protein	85.71	72.22	71.43	66.67	88.89
group_930	UniRef90_A0A0T5YWF3 Uncharacterized protein	100.00	88.89	92.86	100.00	100.00
group_959	UniRef90_A0A0T5YWL6 Uncharacterized protein	85.71	100.00	100.00	100.00	100.00
group_964	UniRef90_A0A0T5Z1A9 Uncharacterized protein	71.43	61.11	57.14	100.00	55.56
group_967	Hypothetical protein	85.71	77.78	100.00	100.00	100.00
group_975	UniRef90_A0A0T5Z409 Uncharacterized protein	42.86	83.33	71.43	66.67	77.78
group_979	UniRef90_A0A0T5ZBY5 Uncharacterized protein	100.00	94.44	92.86	100.00	100.00

Gene	Annotation	J1398 (%)	J1392 (%)	J1390 (%)	D231 (%)	D233 (%)
group_982	UniRef90_A0A0T5Z4S5 Uncharacterized protein	85.71	100.00	100.00	100.00	100.00
group_990	UniRef90_G2FEE2 Uncharacterized protein	100.00	100.00	92.86	100.00	100.00
group_992	UniRef90_A0A1A6FHU6 Uncharacterized protein	71.43	38.89	57.14	66.67	100.00
group_997	UniRef90_A0A0T5Z6W5 Uncharacterized protein	85.71	66.67	64.29	100.00	100.00

Percent gene presence is reported across hydrothermal vent sites (i.e., dive IDs), with universally present genes removed. Genes with uncharacterized annotations are reported here

Appendix 7.

PCoA of Nucleotide Counts and Haplotypes across Host Size, Intra-site Investigation

An intra-site specific investigation of host size versus symbiont genomic structure was done to remove influence of site-to-site variability. Hydrothermal vent sites J2-1390 and J2-1392 from the North Guaymas Black Smoker area were investigated, because these data sets were complete for *R. pachyptila* size ranges. PCoA of nucleotide count (Figure 21) and haplotype (Figure 22) by host size for Dive J2-1390 did not indicate any clear correlation between host size and genomic structure. Similarly, there is no observed correlation for PCoA between nucleotide count (Figure 23) and haplotype (Figure 24) by host size for samples collected from Dive J2-1392. These results support aggregate observations of no clear correlation between genomic structure and host physiology by PCoA.

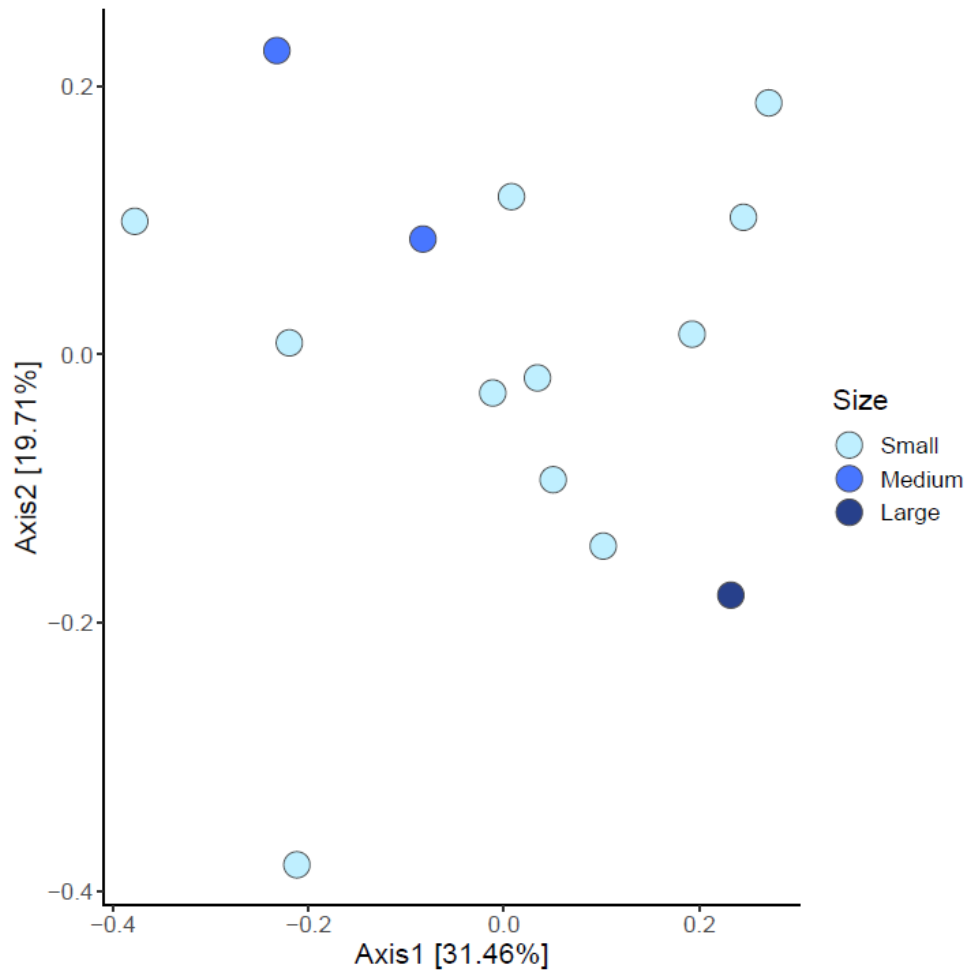


Figure 21. PCoA of Nucleotide Counts across Host Size for Dive ID J2-1390.

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. Riftia pachyptila size is defined as ‘Small’ (<15 mm), ‘Medium’ (15 – 25 mm), or ‘Large’ (>25 mm). There is no distinct correlation between host size and nucleotide count for samples collected in dive J2-1390.

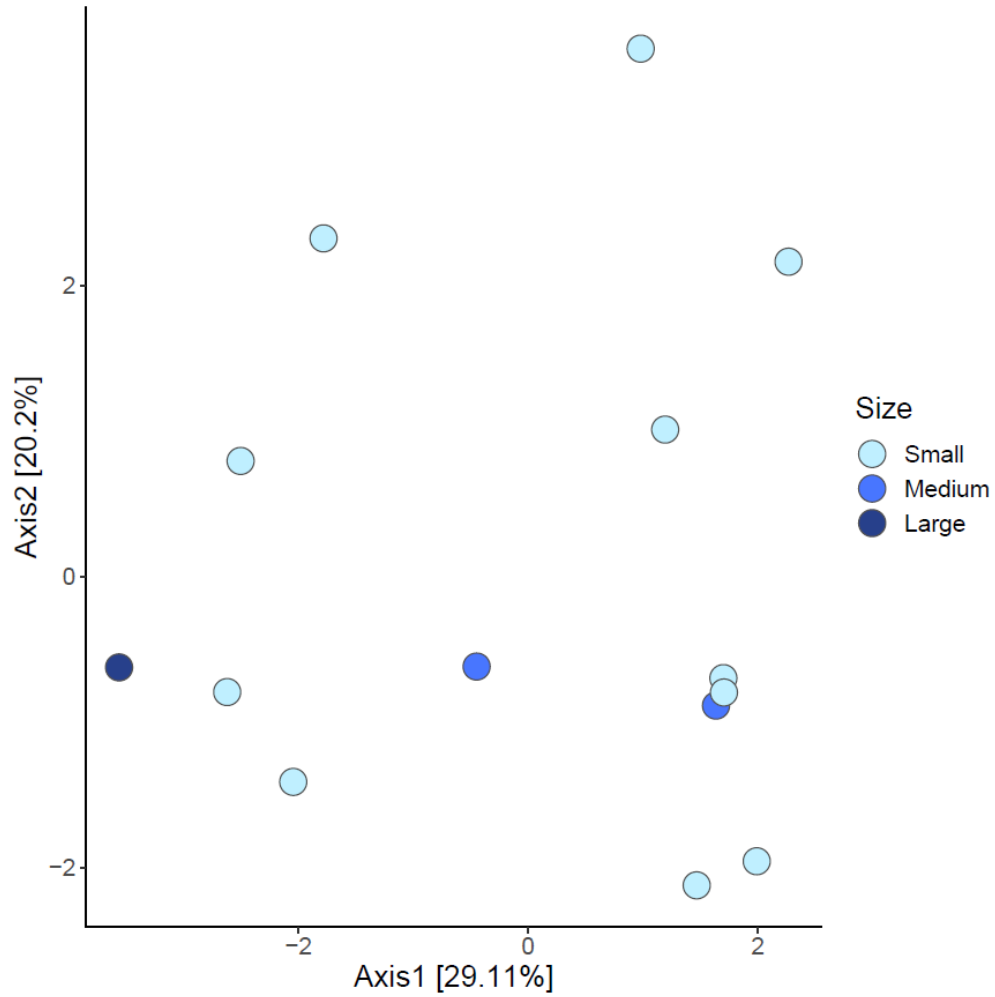


Figure 22. PCoA of Haplotypes across Host Size for Dive ID J2-1390.

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. Riftia pachyptila size is defined as 'Small' (<15 mm), 'Medium' (15 – 25 mm), or 'Large' (>25 mm). There is no distinct correlation between host size and genotype for samples collected in dive J2-1390.

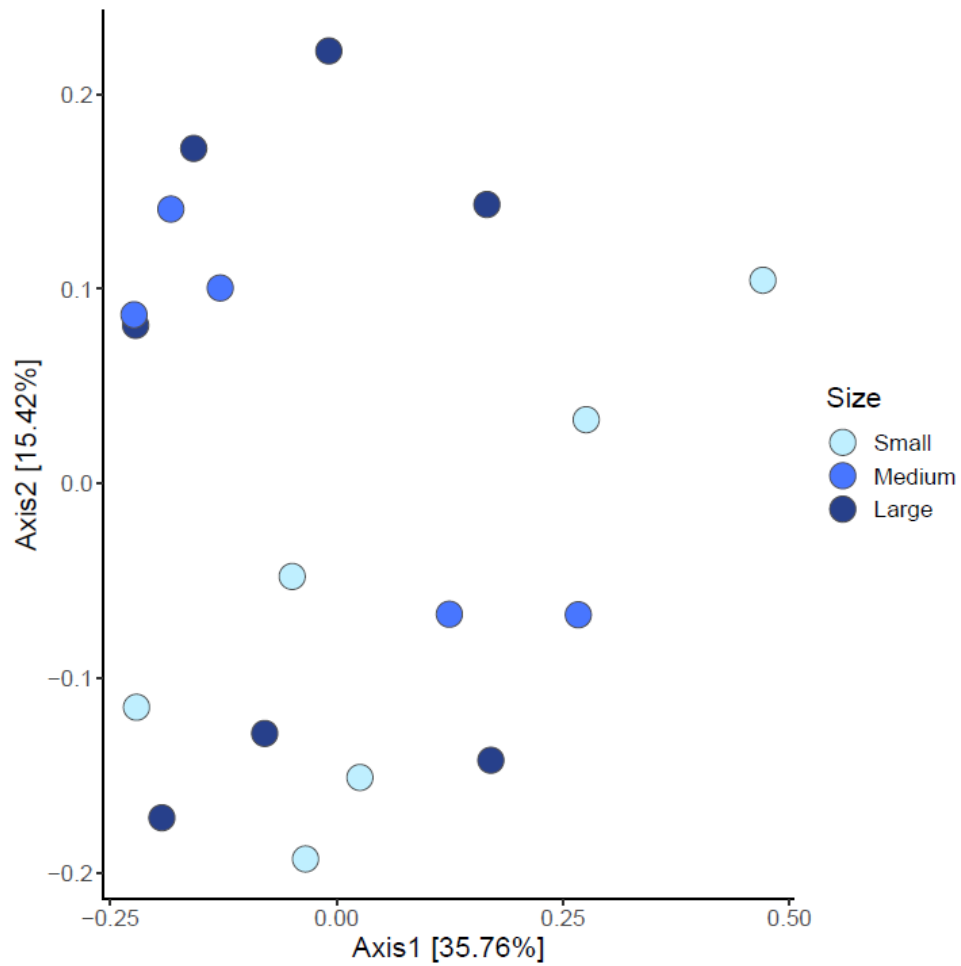


Figure 23. PCoA of Nucleotide Counts across Host Size for Dive ID J2-1392.

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. Riftia pachyptila size is defined as ‘Small’ (<15 mm), ‘Medium’ (15 – 25 mm), or ‘Large’ (>25 mm). There is no distinct correlation between host size and nucleotide count for samples collected in dive J2-1392.

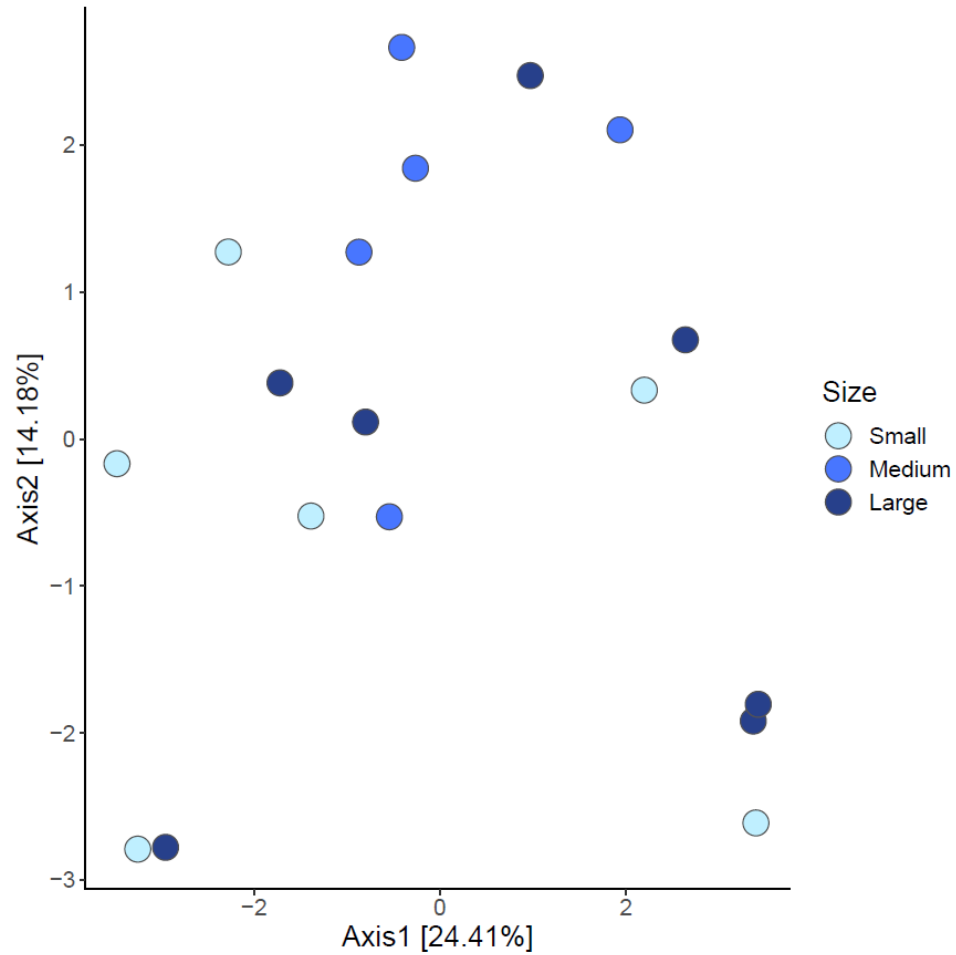


Figure 24. PCoA of Haplotypes across Host Size for Dive ID J2-1392.

PCoA based on Euclidean distance and Bray-Curtis dissimilarity with Cailliez correction. Riftia pachyptila size is defined as ‘Small’ (<15 mm), ‘Medium’ (15 – 25 mm), or ‘Large’ (>25 mm). There is no distinct correlation between host size and haplotype for samples collected in dive J2-1392.

References

- Aragón-Arreola, M., Morandi, M., Martín-Barajas, A., Delgado-Argote, L., & González-Fernández, A. (2005). Structure of the rift basins in the central Gulf of California: Kinematic implications for oblique rifting. *Tectonophysics*, *409*(1–4), 19–38. <https://doi.org/10.1016/j.tecto.2005.08.002>
- Arp, A. J., & Childress, J. J. (1983). Sulfide Binding by the Blood of the Hydrothermal Vent Tube Worm *Riftia pachyptila*. *Science*, *219*(4582), 295–297. <https://doi.org/10.1126/science.219.4582.295>
- Arp, A. J., Childress, J. J., & Vetter, R. D. (1987). The Sulphide-binding protein in the blood of the Vestimentiferan Tube-worm, *Riftia pachyptila*, is the Extracellular Haemoglobin. *Journal of Experimental Biology*, *128*, 139–158. <https://doi.org/10.1242/jeb.128.1.139>
- Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *ELife*, *10*, e65088. <https://doi.org/10.7554/eLife.65088>
- Beinart, R. A., Sanders, J. G., Faure, B., Sylva, S. P., Lee, R. W., Becker, E. L., Gartman, A., Luther, G. W., Seewald, J. S., Fisher, C. R., & Girguis, P. R. (2012). Evidence for the role of endosymbionts in regional-scale habitat partitioning by hydrothermal vent symbioses. *Proceedings of the National Academy of Sciences*, *109*(47), E3241–E3250. <https://doi.org/10.1073/pnas.1202690109>
- Berndt, C., Hensen, C., Mortera-Gutierrez, C., Sarkar, S., Geilert, S., Schmidt, M., Liebetrau, V., Kipfer, R., Scholz, F., Doll, M., Muff, S., Karstens, J., Planke, S., Petersen, S., Böttner, C., Chi, W.-C., Moser, M., Behrendt, R., Fiskal, A., ... Lizarralde, D. (2016). Rifting under steam—How rift magmatism triggers methane venting from sedimentary basins. *Geology*, *44*(9), 767–770. <https://doi.org/10.1130/G38049.1>
- Berta, A., Sumich, J. L., & Kovacs, K. M. (2015). Chapter 14 - Population Structure and Dynamics. In: A. Berta, J.L. Sumich, & K.M. Kovacs (Eds.), *Marine Mammals: Third Addition* (pp. 533–595). Academic Press.
- Black, M. B., Lutz, R. A., & Vrijenhoek, R. C. (1994). Gene flow among vestimentiferan tube worm (*Riftia pachyptila*) populations from hydrothermal vents of the eastern Pacific. *Marine Biology*, *120*, 33–39. <https://doi.org/10.1007/BF00381939>

- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Breusing, C., Genetti, M., Russell, S. L., Corbett-Detig, R. B., & Beinart, R. A. (2022). Horizontal transmission enables flexible associations with locally adapted symbiont strains in deep-sea hydrothermal vent symbioses. *Proceedings of the National Academy of Sciences*, *119*(14), 11. <https://doi.org/10.1073/pnas.2115608119>
- Bright, M., & Bulgheresi, S. (2010). A complex journey: Transmission of microbial symbionts. *Nature Reviews Microbiology*, *8*(3), 218–230. <https://doi.org/10.1038/nrmicro2262>
- Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika*, *48*, 305–30. <https://doi.org/10.1007/BF02294026>
- The Cambrian Foundation. (N.D.). *Chemolithoautotrophic bacteria*. <http://cambrianfoundation.org/chemolithoautotrophic-bacteria/>
- Carlson, C. A., Bates, N. R., Hansell, D. A., & Steinberg, D. K. (2001). Carbon Cycle. In: J.H. Steele (Ed.), *Encyclopedia of Ocean Sciences* (pp. 390-400). Academic Press.
- Cavanaugh, C. M. (1983). Symbiotic chemoautotrophic bacteria in marine invertebrates from sulphide-rich habitats. *Nature*, *302*(5903), 58–61. <https://doi.org/10.1038/302058a0>
- Cavanaugh, C. M., Gardiner, S. L., Jones, M. L., Jannasch, H. W., & Waterbury, J. B. (1981). Prokaryotic Cells in the Hydrothermal Vent Tube Worm *Riftia pachyptila* Jones: Possible Chemoautotrophic Symbionts. *Science*, *213*(4505), 340–342. <https://doi.org/10.1126/science.213.4505.340>
- Cavanaugh, C. M. (1994). Microbial Symbiosis: Patterns of Diversity in the Marine Environment. *American Zoologist*, *34*(1), 79–89. <https://doi.org/10.1093/icb/34.1.79>
- Childress, J. J., & Girguis, P. R. (2011). The metabolic demands of endosymbiotic chemoautotrophic metabolism on host physiological capacities. *Journal of Experimental Biology*, *214*(2), 312–325. <https://doi.org/10.1242/jeb.049023>
- Corliss, J.B., & Ballard, R.D. (1977). Oasis of Life in the Cold Abyss. *National Geographic*, *152*(4), 144.
- Coykendall, D. K., Johnson, S. B., Karl, S. A., Lutz, R. A., & Vrijenhoek, R. C. (2011). Genetic diversity and demographic instability in *Riftia pachyptila* tubeworms from eastern Pacific hydrothermal vents. *BMC Evolutionary Biology*, *11*(1), 96. <https://doi.org/10.1186/1471-2148-11-96>

- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Dean, W., Pride, C., & Thunell, R. (2004). Geochemical cycles in sediments deposited on the slopes of the Guaymas and Carmen Basins of the Gulf of California over the last 180 years. *Quaternary Science Reviews*, 23(16–17), 1817–1833. <https://doi.org/10.1016/j.quascirev.2004.03.010>
- De Oliveira, A. L., Srivastava, A., Espada-Hinojosa, S., & Bright, M. (2022). The complete and closed genome of the facultative generalist *Candidatus* Endoriftia persephone from deep-sea hydrothermal vents. *Molecular Ecology Resources*, 00, 1-18. <https://doi.org/10.1111/1755-0998.13668>
- Dick, G. J., Anantharaman, K., Baker, B. J., Li, M., Reed, D. C., & Sheik, C. S. (2013). The microbiology of deep-sea hydrothermal vent plumes: Ecological and biogeographic linkages to seafloor and water column habitats. *Frontiers in Microbiology*, 4(124), 1-16. <https://doi.org/10.3389/fmicb.2013.00124>
- Dick, G. J. (2019). The microbiomes of deep-sea hydrothermal vents: Distributed globally, shaped locally. *Nature Reviews Microbiology*, 17, 271-283. <https://doi.org/10.1038/s41579-019-0160-2>
- Felbeck, H., Childress, J. J., & Somero, G. N. (1981). Calvin-Benson cycle and sulphide oxidation enzymes in animals from sulphide-rich habitats. *Nature*, 293, 291-293. <https://doi.org/10.1038/293291a0>
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl), W29–W37. <https://doi.org/10.1093/nar/gkr367>
- Frankham, R., Ballou, J. D., & Briscoe, D. A. (2002). *Introduction to Conservation Genetics*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511808999>
- Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing (arXiv:1207.3907). 1-9. <https://doi.org/10.48550/arXiv.1207.3907>
- Geilert, S., Hensen, C., Schmidt, M., Liebetrau, V., Scholz, F., Doll, M., Deng, L., Fiskal, A., Lever, M. A., Su, C. C., Schloemer, S., Sarkar, S., Thiel, V., & Berndt, C. (2018). On the formation of hydrothermal vents and cold seeps in the Guaymas Basin, Gulf of California. *Biogeosciences*, 15(18), 5715–5731. <https://doi.org/10.5194/bg-15-5715-2018>

- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics (Oxford, England)*, 32(18), 2847–2849. <https://doi.org/10.1093/bioinformatics/btw313>
- Hinzke, T., Kleiner, M., Breusing, C., Felbeck, H., Häsler, R., Sievert, S. M., Schlüter, R., Rosenstiel, P., Reusch, T. B. H., Schweder, T., & Markert, S. (2019). Host-Microbe Interactions in the Chemosynthetic *Riftia pachyptila* Symbiosis. *American Society for Microbiology mBio*, 10(6), 20. <https://doi.org/10.1128/mbio.02243-19>
- Hinzke, T., Kleiner, M., Meister, M., Schlüter, R., Hentschker, C., Pané-Farré, J., Hildebrandt, P., Felbeck, H., Sievert, S. M., Bonn, F., Völker, U., Becher, D., Schweder, T., & Markert, S. (2021). Bacterial symbiont subpopulations have different roles in a deep-sea symbiosis. *ELife*, 10, e58371. <https://doi.org/10.7554/eLife.58371>
- Horstmann, E., Tomonaga, Y., Brennwald, M. S., Schmidt, M., Liebetrau, V., & Kipfer, R. (2021). Noble gases in sediment pore water yield insights into hydrothermal fluid transport in the northern Guaymas Basin. *Marine Geology*, 434, 106419. <https://doi.org/10.1016/j.margeo.2021.106419>
- Horvath, P., & Barrangou, R. (2010). CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science*, 327, 167-170. <https://doi.org/10.1126/science.1179555>
- Illumina. (2022). Sequence Coverage Calculator [Online Tool]. https://support.illumina.com/downloads/sequencing_coverage_calculator.html
- Joint Genome Institute (N.D.). *BBMap Guide*. <https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbmap-guide/>
- Jones, M. L. (1981). *Riftia pachyptila*: Observations on the Vestimentiferan Worm from the Galapagos Rift. *Science*, 213(4505), 333–336. <https://doi.org/10.1126/science.213.4505.333>
- Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7, e7359. <https://doi.org/10.7717/peerj.7359>
- Karaseva, N. P., Rimskaya-Korsakova, N. N., Galkin, S. V., & Malakhov, V. V. (2016). Taxonomy, geographical and bathymetric distribution of vestimentiferan tubeworms (Annelida, Siboglinidae). *Biology Bulletin*, 43(9), 937–969. <https://doi.org/10.1134/S1062359016090132>
- Karginov, F. V., & Hannon, G. J. (2010). The CRISPR System: Small RNA-Guided Defense in Bacteria and Archaea. *Molecular Cell*, 37, 7-19.

- Klose, J., Polz, M. F., Wagner, M., Schimak, M. P., Gollner, S., & Bright, M. (2015). Endosymbionts escape dead hydrothermal vent tubeworms to enrich the free-living population. *Proceedings of the National Academy of Sciences*, *112*(36), 11300-11305. <https://doi.org/10.1073/pnas.1501160112>
- Kung, A., Svobodova, K., Lèbre, E., Valenta, R., Kemp, D., & Owen, J. R. (2021). Governing deep sea mining in the face of uncertainty. *Journal of Environmental Management*, *279*, 111593. <https://doi.org/10.1016/j.jenvman.2020.111593>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
- Le, J. T., Levin, L. A., & Carson, R. T. (2017). Incorporating ecosystem services into environmental management of deep-seabed mining. *Deep Sea Research Part II: Topical Studies in Oceanography*, *137*, 486–503. <https://doi.org/10.1016/j.dsr2.2016.08.007>
- Leenheer, J. A., & Croué, J. P. (2003). Peer Reviewed: Characterizing Dissolved Aquatic Organic Matter. *Environmental Science & Technology*, *37*(1) 9-26. <https://doi.org/10.1021/es032333c>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lonsdale, P., & Becker, K. (1985). Hydrothermal plumes, hot springs, and conductive heat flow in the Southern Trough of Guaymas Basin. *Earth and Planetary Science Letters*, *73*, 211-225.
- López-García, P., Gaill, F., & Moreira, D. (2002). Wide bacterial diversity associated with tubes of the vent worm *Riftia pachyptila*. *Environmental Microbiology*, *4*(4), 204–215. <https://doi.org/10.1046/j.1462-2920.2002.00286.x>
- Mallawaarachchi, V., Wickramarachchi, A., & Lin, Y. (2020). GraphBin: Refined binning of metagenomic contigs using assembly graphs. *Bioinformatics*, *36*(11), 3307–3313. <https://doi.org/10.1093/bioinformatics/btaa180>
- Marsh, A. G., Mullineaux, L. S., Young, C. M., & Manahan, D. T. (2001). Larval dispersal potential of the tubeworm *Riftia pachyptila* at deep-sea hydrothermal vents. *Nature*, *411*(6833), 77–80. <https://doi.org/10.1038/35075063>
- Meo, C. A. D., Wilbur, A. E., Holben, W. E., Feldman, R. A., Vrijenhoek, R. C., & Cary, S. C. (2000). Genetic Variation among Endosymbionts of Widely Distributed Vestimentiferan Tubeworms. *Applied and Environmental Microbiology*, *66*(2), 651-658. <https://doi.org/10.1128/AEM.66.2.651-658.2000>

- Mills, L. S., Soule, M. E., & Doak, D. F. (1993). The Keystone-Species Concept in Ecology and Conservation: Management and Policy Must Explicitly Consider the Complexity of Interactions in Natural Systems. *American Institute of Biological Sciences*, 43(4), 219-224. <https://doi.org/10.2307/1312122>
- Miyamoto, N., Shinozaki, A., & Fujiwara, Y. (2014). Segment Regeneration in the Vestimentiferan Tubeworm, *Lamellibrachia satsuma*. *Zoological Science*, 31(8), 535-541. <https://doi.org/10.2108/zs130259>
- Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, 27(5), 824–834. <https://doi.org/10.1101/gr.213959.116>
- Nussbaumer, A. D., Fisher, C. R., & Bright, M. (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature*, 441(7091), 345–348. <https://doi.org/10.1038/nature04793>
- Ondréas, H., Scalabrin, C., Fouquet, Y., & Godfroy, A. (2018). Recent high-resolution mapping of Guaymas hydrothermal fields (Southern Trough). *BSGF - Earth Sciences Bulletin*, 189(1), 6. <https://doi.org/10.1051/bsgf/2018005>
- Orcutt, B. N., Bradley, J. A., Brazelton, W. J., Estes, E. R., Goordial, J. M., Huber, J. A., Jones, R. M., Mahmoudi, N., Marlow, J. J., Murdock, S., & Pachiadaki, M. (2020). Impacts of deep-sea mining on microbial ecosystem services. *Limnology and Oceanography*, 65(7), 1489–1510. <https://doi.org/10.1002/lno.11403>
- Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, 35(3), 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- Perez, M., Angers, B., Young, C. R., & Juniper, S. K. (2021). Shining light on a deep-sea bacterial symbiont population structure with CRISPR. *Microbial Genomics*, 7(8). <https://doi.org/10.1099/mgen.0.000625>
- Perez, M., & Juniper, S. K. (2016). Insights into Symbiont Population Structure among Three Vestimentiferan Tubeworm Host Species at Eastern Pacific Spreading Centers. *Applied and Environmental Microbiology*, 82(17), 5197–5205. <https://doi.org/10.1128/AEM.00953-16>
- Picazo, D. R., Dagan, T., Ansoerge, R., Petersen, J. M., Dubilier, N., & Kupczok, A. (2019). Horizontally transmitted symbiont populations in deep-sea mussels are genetically isolated. *The ISME Journal*, 13, 2954-2968. <https://doi.org/10.1038/s41396-019-0475-z>
- Picazo, D. R., Werner, A., & Dagan, T. (2022). Pangenome Evolution in Environmentally Transmitted Symbionts of Deep-Sea Mussels Is Governed by Vertical Inheritance. *Genome Biology and Evolution*, 14(7), 1-19. <https://doi.org/10.1093/gbe/evac098>

- Polzin, J., Arevalo, P., Nussbaumer, T., Polz, M. F., & Bright, M. (2019). Polyclonal symbiont populations in hydrothermal vent tubeworms and the environment. *Proceedings of the Royal Society B: Biological Sciences*, 286(20181281), 1-10. <https://doi.org/10.1098/rspb.2018.1281>
- QIAGEN (2020). *DNeasy® Blood & Tissue Handbook*. <https://www.qiagen.com/ie/resources/download.aspx?id=68f29296-5a9f-40fa-8b3d-1c148d0b3030&lang=en>
- R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://R-project.org/>.
- Rhun, A. L., Escalera-Maurer, A., Bratovic, M., & Charpentier, E. (2019). CRISPR-Cas in *Streptococcus pyogenes*. *RNA Biology*, 16(4), 380-389.
- Ricotta, C., & Podani, J. (2017). On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecological Complexity*, 31, 201–205. <https://doi.org/10.1016/j.ecocom.2017.07.003>
- Rimskaya-Korsakova, N., Fontaneto, D., Galkin, S., Malakhov, V., & Martínez, A. (2021). Geochemistry drives the allometric growth of the hydrothermal vent tubeworm *Riftia pachyptila* (Annelida: Siboglinidae). *Zoological Journal of the Linnean Society*, 193(1), 281–294. <https://doi.org/10.1093/zoolinnean/zlaa148>
- Robidart, J. C., Bench, S. R., Feldman, R. A., Novoradovsky, A., Podell, S. B., Gaasterland, T., Allen, E. E., & Felbeck, H. (2008). Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. *Environmental Microbiology*, 10(3), 727–737. <https://doi.org/10.1111/j.1462-2920.2007.01496.x>
- Rona, P. A. (1984). Hydrothermal mineralization at seafloor spreading centers. *Earth-Science Reviews*, 20(1), 1–104. [https://doi.org/10.1016/0012-8252\(84\)90080-1](https://doi.org/10.1016/0012-8252(84)90080-1)
- Sato, M., & Sasaki, A. (2021). Evolution and Maintenance of Mutualism between Tubeworms and Sulfur-Oxidizing Bacteria. *The American Naturalist*, 197(3), 351–365. <https://doi.org/10.1086/712780>
- Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>
- Schmidt Ocean. (2019, March 1). *ROV SuBastian Dive 233 - Guaymas Basin: Quetzalcóatl* [Video]. https://www.youtube.com/watch?v=H12vKv_Utlk
- Shingleton, A. (2010) Allometry: The Study of Biological Scaling. *Nature Education Knowledge* 3(10), 2.
- Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a

- dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3(7), 836–843. <https://doi.org/10.1038/s41564-018-0171-1>
- Stewart, F. J., & Cavanaugh, C. M. (2005). Symbiosis of Thioautotrophic Bacteria with *Riftia pachyptila*. *Progress in Molecular and Subcellular Biology*, 41, 197–225. https://doi.org/10.1007/3-540-28221-1_10
- Teske, A., Wegener, G., Chanton, J. P., White, D., MacGregor, B. Hoer, D., de Beer, D., Zhuang, G., Saxton, M. A., Joye, S. B., Lizarralde, D., Soule, S. A., & Emil Ruff, S. (2021). Microbial Communities Under Distinct Thermal and Geochemical Regimes in Axial and Off-Axis Sediments of Guaymas Basin. *Frontiers in Microbiology*, 12, 1-23. <https://10.3389/fmicb.2021.633649>
- Thermo Fisher Scientific (N.D.). *T042-Technical Bulletin, NanoDrop Spectrophotometers: 260/280 and 260/230 Ratios* (Rev 03/9). https://medicine.yale.edu/keck/dna/protocols/tube/t042-nanodrop-spectrophotometers-nucleic-acid-purity-ratios_407666_284_7035_v1.pdf
- Thomas, E., Anderson, R. E., Li, V., Rogan, L. J., & Huber, J. A. (2021). Diverse Viruses in Deep-Sea Hydrothermal Vent Fluids Have Restricted Dispersal across Ocean Basins. *mSystems*, 6(3), 1-18. <https://doi.org/10.1128/msystems.00068-21>
- Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, 21(1), 180. <https://doi.org/10.1186/s13059-020-02090-4>
- Tunnicliffe, V., Garrett, J. F., & Johnson, H. P. (1990). Physical and biological factors affecting the behaviour and mortality of hydrothermal vent tubeworms (vestimentiferans). *Deep Sea Research Part A. Oceanographic Research Papers*, 37(1), 103–125. [https://doi.org/10.1016/0198-0149\(90\)90031-P](https://doi.org/10.1016/0198-0149(90)90031-P)
- Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, 6(1), 158. <https://doi.org/10.1186/s40168-018-0541-1>
- Van Dover, C. L., Arnaud-Haond, S., Gianni, M., Helmreich, S., Huber, J. A., Jaeckel, A. L., Metaxas, A., Pendleton, L. H., Petersen, S., Ramirez-Llodra, E., Steinberg, P. E., Tunnicliffe, V., & Yamamoto, H. (2018). Scientific rationale and international obligations for protection of active hydrothermal vent ecosystems from deep-sea mining. *Marine Policy*, 90, 20–28. <https://doi.org/10.1016/j.marpol.2018.01.020>
- Vic, C., Gula, J., Rouillet, G., & Pradillon, F. (2018). Dispersion of deep-sea hydrothermal vent effluents and larvae by submesoscale and tidal currents. *Deep Sea Research Part I: Oceanographic Research Papers*, 133, 1–18. <https://doi.org/10.1016/j.dsr.2018.01.001>

- Vrijenhoek, R. C. (1997). Gene Flow and Genetic Diversity in Naturally Fragmented Metapopulations of Deep-Sea Hydrothermal Vent Animals. *Journal of Heredity*, 88(4), 285-293. <https://doi.org/10.1093/oxfordjournals.jhered.a023106>
- Vrijenhoek, R. C. (2010) Genetic diversity and connectivity of deep-sea hydrothermal vent metapopulations. *Molecular Ecology*, 19, 4391–4411. <https://doi.org/10.1111/j.1365-294X.2010.04789>
- Washburn, T. W., Turner, P. J., Durden, J. M., Jones, D. O. B., Weaver, P., & Van Dover, C. L. (2019). Ecological risk assessment for deep-sea mining. *Ocean & Coastal Management*, 176, 24–39. <https://doi.org/10.1016/j.ocecoaman.2019.04.014>
- Wickham, H. (2016). Chapter 9 – Data Analysis. In: R. Gentleman, K. Hornik, & G. Parmigiani (Eds.), *Use R!: ggplot2 Elegant Graphics for Data Analysis: Second Edition* (pp. 189-201). Springer.
- Wiebe, P. H., Copley, N., Van Dover, C., Tamse, A., & Manrique, F. (1988). Deep-water zooplankton of the Guaymas basin hydrothermal vent field. *Deep Sea Research Part A. Oceanographic Research Papers*, 35(6), 985-1013. [https://doi.org/10.1016/0198-0149\(88\)90072-6](https://doi.org/10.1016/0198-0149(88)90072-6)
- Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., Khor, C. C., Petric, R., Hibberd, M. L., & Nagarajan, N. (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, 40(22), 11189–11201. <https://doi.org/10.1093/nar/gks918>
- Wu, Y. W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4), 605–607. <https://doi.org/10.1093/bioinformatics/btv638>
- Young, C. R., Fujio, S., & Vrijenhoek, C. (2008). Directional dispersal between mid-ocean ridges: deep-ocean circulation and gene flow in *Ridgeia piscesae*. *Molecular Ecology*, 17, 1718-1731. <https://doi.org/10.1111/j.1365-294X.2008.03609.x>