Using Bioinformatics Tools to Explore Symbiotic Associations Between Marine Invertebrates and Microbes

by

Oluchi L. Aroh

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 5, 2023

Keywords: Bioinformatics, Genomics, Metagenome, Microbes, Symbiosis, Marine invertebrates

Copyright 2023 by Oluchi L. Aroh

Approved by

Kenneth Halanych, Co-chair, Executive Director, Center for Marine Science
Mark Liles, Co-chair, Professor, Department of Biological Sciences
Leslie Goertzen, Assistant Professor, Department of Biological Sciences
Neha Potnis, Assistant Professor, Department of Entomology & Plant Pathology

Abstract

The relationships and interdependence of microbes and marine invertebrates play a critical role in maintaining the marine ecosystem and are also pointers to environmental health and biodiversity. Some of these symbiotic associations have been shown to impact the morphology, behavior, and development of both the host and the symbionts. My dissertation expands on these studies by exploring the symbiosis between marine invertebrates and their microbial counterparts at various levels. My first project focuses on analyzing transposable elements characterized by long terminal retrotransposons (LTR retrotransposons) in the genome of the annelid *tubeworm Lamellibrachia luymesi*. These elements are integrated into the host genome, influencing the host's evolution and potentially gene function. During this project, I developed a bioinformatics pipeline that can be used to characterize these elements in non-model organisms in order to provide insight into their evolution. In the second project, a novel *Mycoplasma* symbiont was identified in the genome of *Gorgonocephalus chilensis*. This study provides insight into the metabolic capabilities and potential role of this novel *Mycoplasma* symbiont and its evolutionary placement. Surprisingly, we found the *Mycoplasma* symbiont to have a very broad geographic range spanning the Pacific Ocean. The last project focuses on the variations in the microbial composition of farm-raised and wild oysters, highlighting these variations and exploring their functional capabilities. By examining these variations, we gained insight into the presence of opportunistic pathogens in these communities and explored the functional capabilities of the microbial communities. Overall, my dissertation aims to contribute to the understanding of the intricate relationships between marine invertebrates and their microbial symbionts, enhance our knowledge of the marine ecosystem and contribute to the preservation of environmental health and biodiversity.

Acknowledgments

I owe my successes to several individuals who have played an integral role in shaping my academic and personal growth. First and foremost, I am deeply grateful to the almighty God for blessing me with the gift of life and good health, which has enabled me to pursue my academic goals with dedication and resilience.

I want to express my sincere appreciation to my wonderful supervisor, Dr. Kenneth Halanych, whose guidance, support, and encouragement have been invaluable throughout my graduate studies. Dr. Halanych's unwavering belief in my potential, and his commitment to providing me with the necessary resources and mentorship, has been instrumental in helping me reach this milestone. Furthermore, I would like to specifically thank Dr. Mark Liles, who went from being a member of my committee to being my interim advisor, offering crucial guidance and feedback, I am deeply grateful. To other members of my committee, Dr. Neha Potnis and Dr. Leslie Geortzen, I am grateful for your insightful feedback and academic expertise that has contributed immensely to this achievement. I extend my heartfelt gratitude to Dr. Scott McLeroy for his role as my university reader.

I'd also like to thank my colleagues and lab mates for their unwavering support, especially Dr. Candace Grimes, whose assistance was invaluable during the final chapter of my dissertation. To the other members of Halanych lab and DBS graduate students who have supported me in various ways throughout my academic journey, I am deeply grateful and say a resounding THANK YOU!

To my amazing husband, Tagbo, and our daughter, Jachimma, I am forever grateful for your support, encouragement, understanding, kindness, and patience, I love you both. To my parents, parents-in-law, siblings, siblings-in-law, and my friends, thank you for your prayers and support.

Table of Contents

List of Tables

List of Figures

Chapter I. Background

Marine invertebrates, a diverse group of organisms that inhabit the world's oceans, play a critical role in maintaining the health and function of marine ecosystems. The study of marine invertebrates is crucial as they serve as key indicators of environmental health and contributes significantly to biodiversity. Marine organisms surviving in diverse environments depend to a large extent, or completely, on symbiotic microbes [1]. These symbiotic relationships are continuously formed and have co-evolved with a long-term history that is known to impact morphology, behavior, development, metabolism, and even evolution of both the host and the symbiont [2]. Additionally, symbioses between marine invertebrates and microbes underscore the health of marine ecosystems, especially the most threatened ecosystems [3].

Symbiotic associations between marine invertebrates and microbes have been found in various ecosystems ranging from coral reefs in shallow coastal waters to hydrothermal vents and cold seeps in the deep sea [4], and in various organisms including corals [5], sponges [6], and mollusks [7–9]. These studies have revealed some of the profoundly important symbiotic roles microbes play in the lives of their host and the marine ecosystem at large. Ecosystem engineers, such as corals and hydrothermal vents tubeworms create habitats and nutrient resources that are crucial to the foundation of their ecosystem. These organisms engage in mutualistic nutritional symbioses with microbes and such relationships enable these hosts utilize resources or substrates otherwise unavailable to them [10]. An example of such relationships can be found in the Hawaiian bob tail squid which lives in mutualistic symbiosis with the bioluminescent bacteria *Aliivibrio fischeri*. The host supplies the bacteria with a solution of sugars and amino acids and in return, the bacteria provide bioluminescence to aid predator avoidance [9]. Another example can be seen in the vestimentiferan tube worm *Lamellibrachia luymesi*, which lacks a digestive tract and hosts sulfide-oxidizing, horizontally-transmitted bacterial symbionts for nutrition and growth [1112].

Furthermore, in the face of global climate change, understanding the interaction between marine invertebrates and symbiotic microbes in a changing environment can help predict whether symbiosis will allow marine life to cope with future threats to the biosphere [13]. Given the significance of symbiosis between marine invertebrates and microbes, my dissertation aims to contribute to a better understanding of these relationships and to underscore their roles and

importance. Herein, three projects which highlight the relationships between marine invertebrates and microbial symbionts are explored.

Project 1: **Characterizing LTR Retrotransposons in *Lamellibrachia luymesi* – published in BMC Genomics -** BMC Genomics 22, 466 (2021).

Long terminal repeat retrotransposons (LTR retrotransposons) are transposable elements characterized by long terminal repeats (LTRs) flanking an internal coding region. These elements are considered symbionts as they integrate into the host genome, co-evolving with the host, and influencing the host evolution, function, and regulation of genes [14]. Moreover, these elements are likely to play a relevant role in adaptation in their host due to their ability to generate mutations and their capacity to be responsive and susceptible to environmental changes and to colonize new ecological niches [15]. Importantly, LTR retrotransposons serve as a model for the study of retroviruses [16], because both are structurally similar and phylogenetically related [17]. In this project [18], I developed a bioinformatics pipeline to explore and characterize LTR retrotransposons present in the genome of *Lamellibrachia luymesi*, to augment understanding of the potential function and structure of these elements in non-model organisms. Furthermore, I explored the evolutionary history of LTR retrotransposons in *Lamellibrachia luymesi* by estimating if their insertion is due to a recent or ancient event.

Project 2: Genomic characterization of a novel, widely distributed *Mycoplasma* Species "*Candidatus* Mycoplasma mahonii" associated with the brittlestar *Gorgonocephalus chilensis* – Accepted pending minor revisions.

Despite rapid growth in research on host-associated microbes from individual microbial symbionts to host-associated taxa, little is known about their interactions with the vast majority of marine host species, hence there is a need for more studies to shed light on these relationships. This project seeks to explore the symbiotic relationships between a microbial symbiont and a marine invertebrate to provide more insight into the genomic mechanisms used to maintain such a relationship. In this study, we examined the symbiotic relationships between microbes and the filter-feeding basket star *Gorgonocephalus chilensis*. Through our analysis, we unveiled a 796kb *Mycoplasma* symbiont associated with *Gorgonocephalus chilensis*. The name "*Candidatus* Mycoplasma mahonii" was proposed for this novel species. Our study explored the metabolic

capabilities and potential roles of this symbiont within the host organism and investigated if this novel species occurred in other basket stars. Additionally, by utilizing 16S rRNA and multilocus phylogenetic analyses, we also explored the degree of relatedness between *Ca*. M. mahonii and other *Mycoplasma* spp, providing insights into its evolutionary context and geographic range.

Project 3: Comparative Analysis of the Microbial Composition of Farm-raised and Wild Oysters.

Oyster aquaculture is a crucial agricultural sector due to the different roles they play in their ecosystem including improving water quality and reef formation which in turn creates habitat for other marine organisms [19]. Additionally, oysters are typically eaten raw by humans and hence serve as a vector for various human pathogens. In this project, we analyzed farm-raised and wild oyster populations in close proximity to assess variations in their microbial populations. This work also highlights the critical role played by these microbes in oyster populations and their surrounding ecosystem.

Summarily, these projects explore the symbiosis between marine invertebrates and their microbial counterparts at different levels. It also contributes to a better understanding of the community structure and function of marine invertebrate-associated microbes and the interaction between marine invertebrates and their symbiotic microbes in the ever-changing marine environment.

Chapter II. Genome-wide characterization of LTR retrotransposons in the non-model deep-sea annelid *Lamellibrachia luymesi*

Introduction

Retrotransposons are transposable elements that replicate via an RNA intermediate [20]. They often make up a substantial fraction of the host genome in which they reside, occupying more than 40% of the human genome [21] and more than 50% of the maize genome[22] . Retrotransposons play a role in genome evolution [14] and can ultimately impact gene expression. However, our understanding of phylogenetic diversity of retrotransposons and their role in genome evolution is largely based on model organisms such as *Drosophila melanogaster*, *Caenorhabditis elegans*, *Danio rerio*, *Mus musculus*, *Bombyx mori*, etc. Animals living in marine environments and the deep-sea have been particularly underrepresented in transposable elements studies. For this reason, we explored the genome of the deep-sea tubeworm *Lamellibrachia luymesi* (Siboglinidae, Annelida) [11] which employs chemoautotrophic endosymbionts to inhabit hydrocarbon seeps in the Gulf of Mexico.

Long terminal repeat retrotransposons (LTR retrotransposons) are transposable elements that are characterized by having long terminal repeats (LTRs) flanking an internal coding region. LTR retrotransposons usually serve as a model for the study of retroviruses [16], because both are structurally similar and phylogenetically related [17]. The main distinguishing characteristic is the presence of an envelope (env) gene in retroviruses which is absent in LTR retrotransposons. LTR retrotransposons are classified into three super families (Copia, Gypsy, and Bel-pao), which differ in the arrangement of the protein domains encoded within the pol gene [23]. The two most common LTR retrotransposon super-families – Copia and Gypsy, are found in almost all eukaryotic lineages sampled to date [24]. These superfamilies display different distribution, abundance, and diversity based on the element type and the host taxon been considered [25].

LTR retrotransposons (Fig. 2.1) includes long terminal repeats flanking elements that range from a few hundred bases to more than 5kb and usually start with 5'TG-3' and ends with 5'-CA3', a target site duplication (TSD) of 4-6bp, a polypurine tract (PPT), a primer binding site (PBS) and also gag and pol genes between the two LTRs [26,27]. The gag gene encodes a structural protein that is essential for assembly of viral-like particles while the pol gene encodes four proteins domains including a protease (PR) which cleaves the Pol polyprotein, a ribonuclease H (RH) which cleaves

the RNA in the DNA-RNA hybrid, a reverse transcriptase (RT) that copies retrotransposons RNA into cDNA and an integrase (INT) which integrates the cDNA into the genome. Occasionally, an additional open reading frame (aORF) may be downstream or upstream of the gag-pol gene, in sense or antisense orientation [28,29]. Those located in the sense orientation encode proteins with certain structural and functional similarities to the env domain of retroviruses, and hence are sometimes called env-like domains [30,31]. The env domain encodes for protein that is responsible for binding the cellular receptor and facilitates the early steps in the virus-cell interaction and drives the fusion of viral and host cellular membrane [32]. In contrast, function of the aORF located in the antisense orientation is not clearly known, however, studies carried out so far suggests that they may be playing a regulatory role in retrotransposition [31,33,34].

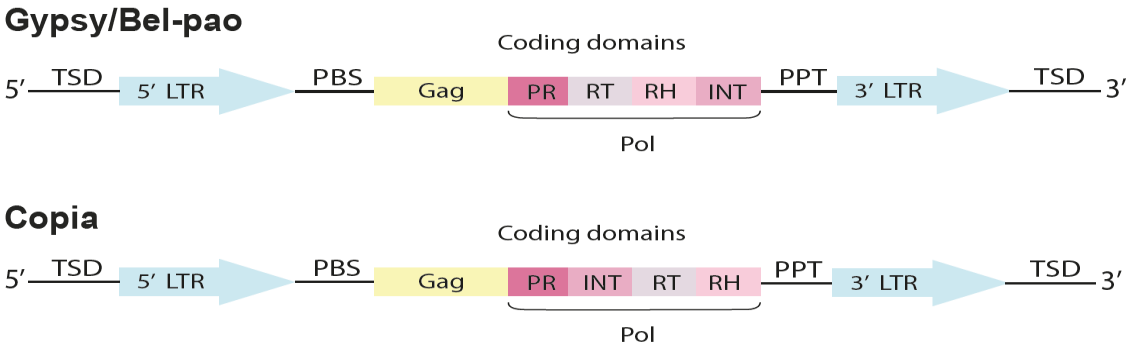

Figure 2.1. Structure of a LTR retrotransposon. Gag - group-specific antigen gene; TSD- target site duplication; PR - aspartic protease gene; RT - reverse transcriptase gene; RH - ribonuclease-H gene; INT- integrase gene; PBS - primer binding site; PPT - polypurine tract. LTR retrotransposon structure was generated using Adobe Illustrator.

In previous reports, retroelements have been identified in marine organisms including sea urchins [35], corals endosymbionts [36] and crustaceans [37]. However, to the best of our knowledge, there has been minimal effort to characterize the LTR retrotransposons present in deep-sea (>200m) animals or in annelids. Available studies [11,38,12] tend to only consider transposable elements in context of their role in genome composition rather than detailed assessment of the elements and their evolution. Of particular interest, Li et al. assessed *Lamellibrachia luymesi* van der Land & Norrevang 1975; a deep-sea annelid. *L. luymesi* is a vestimentiferan tubeworm that forms bush-like aggregations at hydrocarbon seeps in the Gulf of Mexico. These animals lack a digestive tract and hosts sulfide-oxidizing, horizontally-transmitted bacterial symbionts for nutrition and growth[11394041]. Their result showed that 2.52% of the genome consisted of LTR retroelements. However, the goal of the analysis was to see how much of the genome's DNA was derived from repetitive elements using

RepeatModeler [42] and RepeatMasker [43]. Their approach included altered copies such as truncated elements or solo LTR's to gain a comprehensive view of *L. luymesi's* genome composition rather than an exploration of the LTR retroelements biology. In the current study, we further characterized and classified LTR retrotransposons present in the genome of *Lamellibrachia luymesi* to shed light on the representation of LTR retrotransposon superfamilies, as well as augment understanding of the potential function and structure of intact elements. In addition, we also estimated insertion times of these elements to understand if they are due to recent or ancient events.

We hypothesized the possible presence of unknown LTR-retrotransposon families in marine organisms or unsampled animal lineages. This work represents an important step towards the characterization of LTR retrotransposons in marine systems (70% of the biosphere) and in unexplored animal lineages (e.g., annelids).

Methods

Genomic Sequence

Assembled whole genomic sequence of the siboglinid annelid *Lamellibrachia luymesi* generated by Li et al. ([11];WGS project - SDWI01, Bio project number - PRJNA516467 and Bio sample number - SAMN10789628) was accessed from NCBI [44]. Li et al. conducted a scaffold-level assembly of the genome using Illumina paired-end and mate-pair and sequence data. The total sequence length is 688MB with an overall BUSCO genome completeness of 95%.

Identification of LTR retrotransposons

This study focused only on intact LTR retrotransposons, solo and nested insertions without coding domains were excluded from the analysis. We defined intact LTR retrotransposon as possessing two LTRs, at least one protein domain, and a pair of TSD (Target site duplication) regions.

The bioinformatics pipeline used to identify LTR retrotransposon candidates in the *L. luymesi* genome included two software tools for de-novo prediction of LTR retrotransposons, LTRharvest genometools v1.5.10 [45] and LTR_Finder v1.07 [46]. Both programs were run to provide a more thorough search for putative LTR elements and was based on previously published approaches [24]. In addition, LTRharvest tend to have greater sensitivity whereas LTR_Finder has a lower false-positive rate [47].

To prepare data for LTRharvest, genomic scaffolds were run through Suffixerator (also part of the genometools package) with default parameters to create an enhanced suffix file which is then scanned by LTRharvest. The following LTRharvest parameters were used to obtain LTR retrotransposon candidates with TGCA motifs '-minlentltr 100, -maxlenltr 7000, -mintsd 4, -maxtsd 6, -similar 85, -vic 10, -seed 20, -motif TGCA, -motifmis 1.' In contrast, to obtain LTR retrotransposon candidates without TGCA motifs, parameters were set to '-minlentltr 100, -maxlenltr 7000, -mintsd 4, -maxtsd 6, -similar 85, -vic 10, -seed 20'. These 2 approaches were taken to obtain a more robust putative LTR retroelements list from LTRharvest. Similarly, to obtain candidates with both TGCA and non-TGCA motifs the following parameters were used to run LTR_Finder '-D 15000, -d 1000, -l 100, -L 7000, -p 20, -C, -M 0.85'. In summary, parameters for both programs were set to minimum and maximum LTR length of 100bp and 7000bp respectively and at least 85% identity between two LTR regions.

LTR_retriever v2.8.5 [48] with default parameters was used to filter out false positives LTR candidates identified by LTRharvest and LTR_Finder. This downstream filtering was largely based on boundary mapping of LTRs, presence of TSDs, and presence of palindromic motifs. The palindromic motif library employed by LTR_retriever includes – TGCA, TGCT, TACA, TACT, TGGA, TATA, TGTA, and TCCA.

Classification of discovered LTR retrotransposons

Classification of LTR retrotransposons is dependent upon the presence and order of protein domains within the pol gene [26] (Fig. 2.1). LTR_retriever based the classification of LTR retrotransposons on identification of conserved protein domains of each LTR retrotransposon candidate using profile Hidden Markov Models (pHMMs) of LTR retrotransposon domains from Pfam database [49]. Elements returning ambiguous pHMMs matches were classified as unknown.

To refine classification, we employed the program TEsorter v1.2.5 [50] which translated nucleotide sequence of LTR retrotransposon candidates in all six frames and searched these sequences against HMM profiles obtained from existing mobile elements protein databases – specifically , REXdb [29] and Gyspsy database of mobile genetic elements [51]. For each domain of a sequence, only the best hit with highest score is retained. Classification into superfamilies and families were based on hits of the pol and gag genes to curated database. Elements lacking at least one domain were not classified.

To do this step, fasta sequences of LTR retrotransposon candidates were first extracted using the call_by_seq_list.pl script from LTR_retriever package. Obtained sequences were then input into TEsorter (parameters = '-db gydb, -st nucl and -p 10') for further classification.

Naming Conventions

To facilitate communication, naming conventions for LTR retrotransposons families and elements identified in this study were created. Gypsy families were designated as LGF (*Lamellibrachia* Gypsy Family), followed by a unique number (e.g., LGF1, LGF2 etc.), Copia families were designated as LCF (*Lamellibrachia* Copia Family), followed by a unique number (e.g., LCF1) while Bel-pao families were designated as LBF (*Lamellibrachia* Bel-pao Family), followed by a unique number (e.g., LBF1). For individual elements, identified LTR retrotransposons were designated as LLXY#, where LL denotes 2 letters representing *L. luymesi*, XY denotes the first two letters of the superfamily it belongs to and # denotes the element number (e.g., LLGY1 represents a Gypsy element).

Phylogenetic Analysis

Phylogenetic analysis was used to further validate family-level assignment of these elements and to access the evolutionary position of *L. luymesi* LTR retrotransposon candidates. For this purpose, amino acid sequences of INT, RT and RH domains were extracted from the LTR retrotransposon candidates following the guideline from TEsorter package. Gag and Protease (PR) sequences were excluded from analyses as they are known for their variability which prevents reliable alignments [52,53].

To infer phylogenetic trees, amino acid sequence of INT, RH and RT from other known organisms were obtained from the GYDB database and recent studies [54,55,56], and aligned using MAFFT v7.407 [57] to amino acid sequence of INT, RT and RH from LTR retrotransposons found in *L. luymesi* genome. Each of the 3 domains was analyzed separately and a combined analysis was not done due to difference in taxon sampling and the fact that the domains may have distinct evolutionary histories. Maximum likelihood with bootstrap analysis was employed to construct phylogenetic trees using IQtree v1.6.12 [58] with the following parameters '-bb 100000, -nt AUTO, --runs 5'. The substitution model employed by IQtree for the INT domain tree was LG+R7, the RT domain tree was LG+F+R6 while the RH domain tree was LG+R7. Phylogenetic trees were mid-point rooted, visualized and edited using Figtree v1.4.2 [59].

Estimation of Insertion time

Time since initial insertion of LTR retrotransposon candidates was estimated using scripts implemented in the LTR_retriever package. Insertion time were calculated as T=K/2μ, where K is the divergence rate measured by the Jukes-Cantor model with K = -3/4*ln(1-d*4/3) [60] and μ is the neutral mutation which is set at 1.3×10^{-8} mutations per bp per year [61].

Results

Identification and Classification of LTR-retrotransposon

A total of 223 intact LTR retrotransposons were identified in the 688Mb *L. luymesi* genome, by screening and adjustment of LTR candidates from LTRharvest and LTR_Finder using modules employed in LTR_retriever. Of the 223 intact LTR-retrotransposon identified by LTR_retriever, 51 were classified as unknown, 1 was classified as Copia while 171 were classified as Gypsy.

To further classify these elements, TEsorter was used to search their internal regions against Gypsy database (GYDB). Those matching at least one domain profile in GYDB were classified. All the 171 Gypsy and 1 Copia elements classified by LTR-retriever were also classified as Gypsy and Copia respectively in TEsorter. In addition, out of the 51 classified by LTR_retriever as unknown, 7 were classified as Gypsy, 2 were classified as Bel-pao while 1 was classified as Copia in TEsorter. The rest were not classified at all. Hence, in total, TEsorter classified 182 of the 223 intact LTR retrotransposons identified by LTR-retriever.

Further analyses were carried out on the remaining 41 elements not classified by TEsorter. This was accomplished by manually searching the internal region of these unclassified elements against PFAM [49] and Conserved Domains Database (CDD) [62] to identify domains present within their internal region. Results showed that 24 of the elements lacked domains matching any known profiles in the databases, 10 had domains that were unrelated to LTR retrotransposons (e.g., a transmembrane receptor, coagulation-inhibition site etc.), while the remaining 8 had only RT domains (Supplementary table 1). To further verify and classify these elements, we used REXdb-metazoan database option of TEsorter. We also performed a manual hmmscan search using GYDB hmm profiles. The REXdb- metazoan option classified these elements as LINEs (Long interspersed nuclear elements) while no match was found in the GYDB hmm profile scan. Due to the inability to accurately classify these 41 elements, they were excluded from further analysis.

Summary details of the 182 LTR retrotransposons used for downstream analysis, which includes 178 Gypsy, 2 Bel-pao and 2 Copia elements are shown in Table 2.1.

Table 2.1. Summary of LTR retrotransposons in *L. luymesi*

Superfamily	Structure	Total number	No. with all domains present	Average length of element (min-max)	Total length of elements in bp	Range of percentage LTR identity within Superfamily
Gypsy	Gag-PR-RT-RH-INT	178	30	5,123bp (1,389-8,866)	836,263	0.92% - 100%
Copia	Gag-PR-INT-RT-RH	2	0	3,453bp (2,037-4,869)	6,906	0.95% - 0.99%
Bel-pao	Gag-PR-RT-RH-INT	2	2	6,659bp (6,670-6,648)	13,318	0.92% - 0.99%
Total		182			856,487	

Structural Characterization

Of the 182 identified LTR retrotransposons, 32 elements had all domains (Gag and Pol – RT, INT, RH, PR) present with the remainder having at least one domain present. For Gypsy elements, 30 out of the 178 had a complete set of domains, both the Bel-pao elements had a complete set of domains and both Copia elements lacked a complete set of domains. Further analysis to describe the position of these elements in relation to coding elements showed that 26.4% of them overlap with coding elements, 46.2% were located >5kb of coding elements, 10.4% were located within 5-10kb and the remaining 17% were more than 10kb away from coding elements.

The target site duplication flanking ends of identified LTR retrotransposons ranged from 3-5bp in length, with a majority of them being 5bp in length. Palindromic motifs detected in the elements include TGCA, TACA, TATA, TCGT, TGAA, TGAC, TGAT, and TTAT, with 89% of the LTR-retrotransposons having TGCA motif. In addition, differences in the length of identified LTR-retrotransposons were substantial, ranging from 1,389bp-8,866bp while the length of the LTRs ranged from 103-1,468bp.

Estimation of Insertion Time

Insertion times of LTR retrotransposon elements in *L. luymesi* genome suggests most elements were inserted around 1.0 million years ago (MYA; Fig. 2.2). The oldest observed and complete inserted retrotransposon was a Gypsy element, inserted around 2MYA. Interestingly, 50 Gypsy elements showed a 100% LTR identity, suggesting that they very recently inserted into the genome. However, calculations of insertion times used a substitution rate of 1.3×10^{-8} substitution per bp per year, the LTR_retriever default based on the rice genome. Although these insertion time estimates for *L. luymesi* should be viewed with caution, decreasing the rate by two- or three-fold still suggests insertion times within the last few million years.

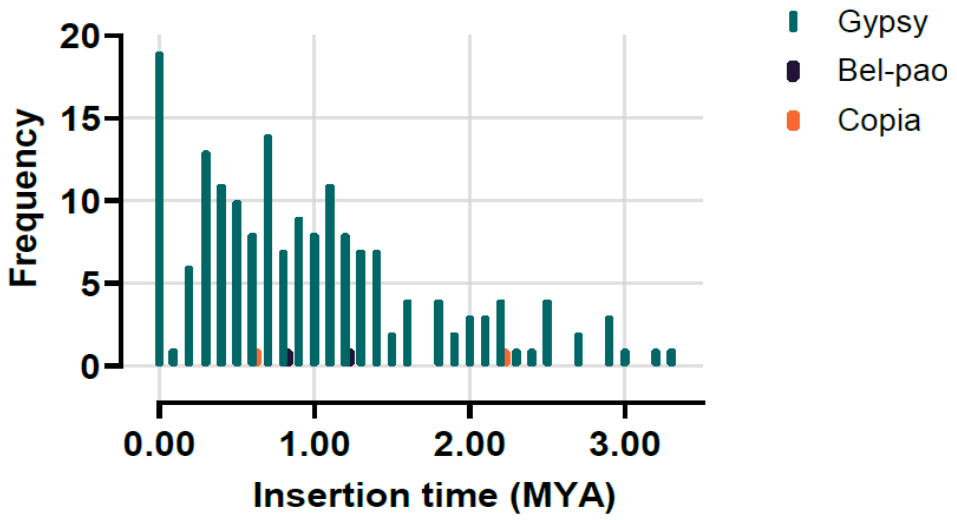

Figure 2.2. Insertion time distribution of intact LTR-RT in *L. luymesi* genome

Phylogenetic analysis of LTR-retrotransposons

Phylogenetic analysis corroborates assignments made by TEsorter. However, weak internodal support limited inferences about evolutionary relationships. Final family assignment was done by considering placements of elements with strong nodal support indicating monophyletic lineage representing gene families (Fig. 2.3 for RT domain, Fig 2.4 for RH domain, and Fig. 2.5 for INT domain). Due to issues of non-concordant evolutionary histories, domains were not combined into a single phylogenetic analysis. Naming conventions based on phylogenetic analyses are described in the Methods section.

For Gypsy elements, phylogenetic analysis of the RT, RH, and INT sequences showed that some elements fall into recognized families such as CSRN1 [63], Gmr1 [64] and Mag [65,66] while others

formed lineages distinct from previously recognized families. The 5 novel families were LGF2 (bootstrap value, bsv 100 in all the domain trees), LGF4 (bsv = 100, all domains), LGF7 (bsv = 94, 100, 91 in RH, RT and INT domain trees, respectively), LGF8 (bsv = 86, 93, 100 in RH, RT and INT domain trees) and LGF9 (bsv= 100, all domains). Other Gypsy elements fell within the Mag family (LGF5; bsv = 98, 100, 100 in RH, RT and INT domain trees), the Gmr1 family (LGF3; bsv = 95, 99, 100 in RH, RT and INT domain trees) and the CSRN1 family (LGF1; bsv = 99, 100, 100 in RH, RT and INT domain trees respectively). The LGF6 family was also inside the Mag family, but although this clade was monophyletic in the RH and INT trees (bsv= 74, 91 respectively), it was paraphyletic in the RT trees.

Mag elements (LGF5 and LGF6) which includes A, B and C clades where the most dominant with more than 70 elements. Elements in the 2 previously described families; CSRN1 (LGF1) and Gmr1 (LGF3), were fewer with less than 25 elements. The remaining novel families (LGF2 and LGF4) with strong bootstrap support had less than 15 elements. Three of the novel families (LFG8, LFG9 and LFG7) clustered within Mag elements, suggesting that they might be distinct lineage within the Mag radiation.

For the Copia elements, LLCO1 had all 3 domains used in tree building - RT, RH, and INT present while LLCO2 had only the RH domain (but still had GAG and PR domains not used in trees). Hence, LLCO2 was absent in INT and RT trees. In the RH tree, LLCO2 clustered within the GalEa family (LCF2) with a bootstrap value of 100. LLCO1 varied in position in the INT, RT, and RH domain tree (LCF1). In the INT and RT domain tree, this element fell within the pCetro and Hydra family respectively (bsv = 97 and 88, respectively), whereas LLCO1's position was unsupported in the RH trees (bsv = 58).

Both Bel-pao elements (LLBP1 and LBP2) clustered within Sinbad lineage, LBF1 (bsv = 94, 100, 98 in RH, RT and INT domain trees).

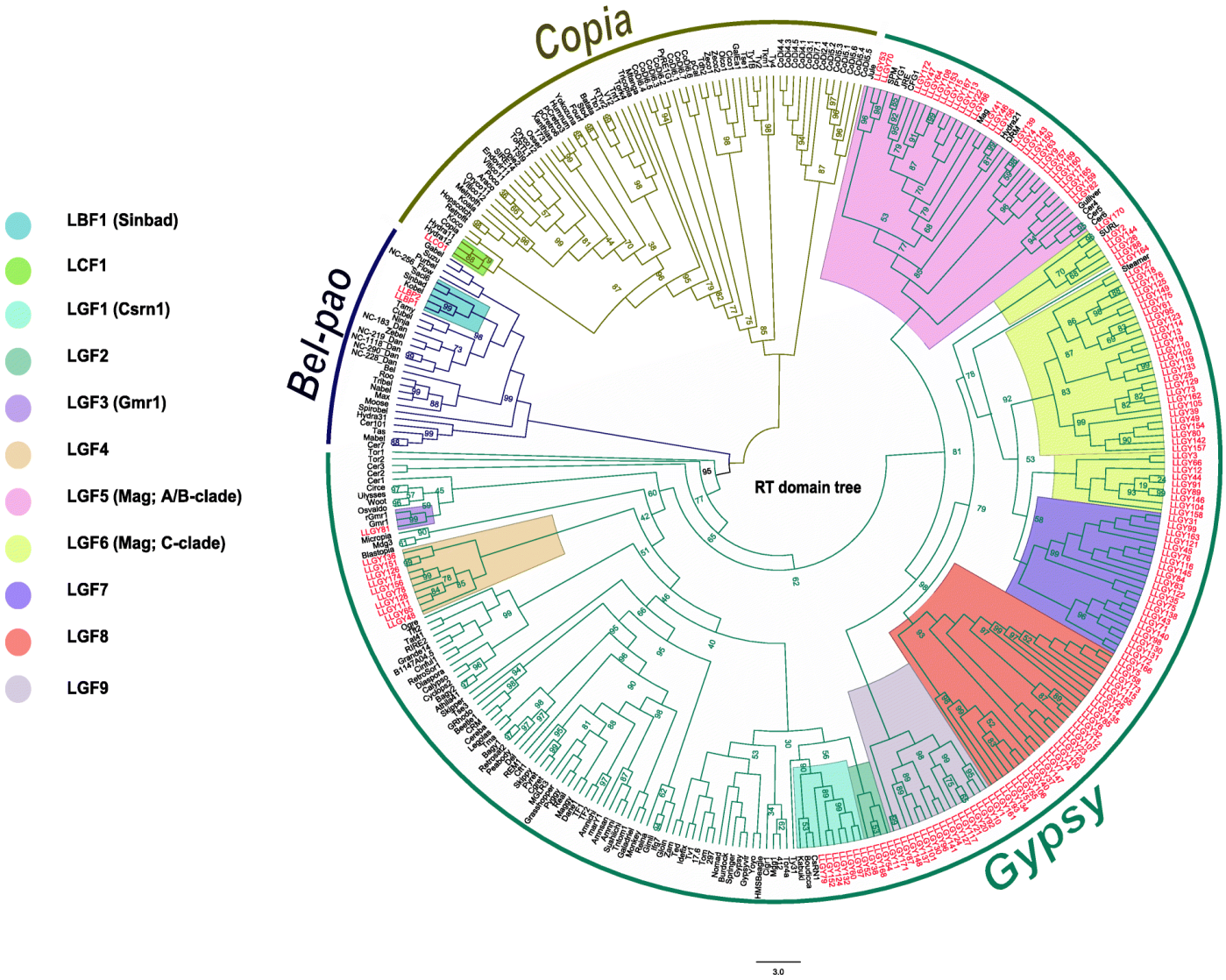

Figure 2.3. RT domain phylogenetic tree. RT phylogenetic tree was generated in IQtree with the LG + F + R6 model. Tree lines are color-coded according to the superfamily above it. Elements in red are elements identified in the genome of *L. luymesi*.

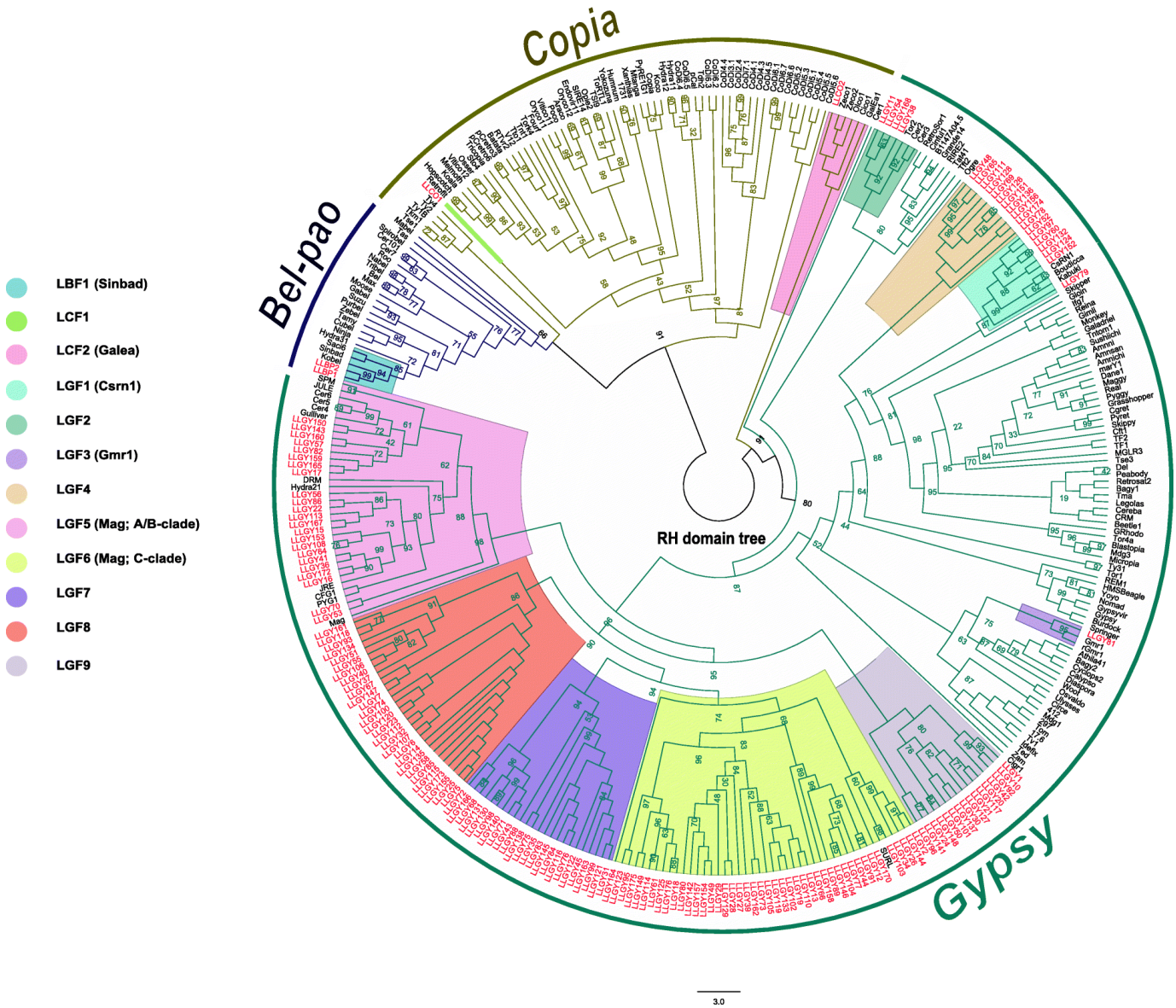

Figure 2.4. RnaseH domain phylogenetic tree. RnaseH phylogenetic tree was generated in IQtree with the LG + R7 model. Tree lines are color-coded according to the superfamily above it. Elements in red are elements identified in the genome of *L. luymesi*.

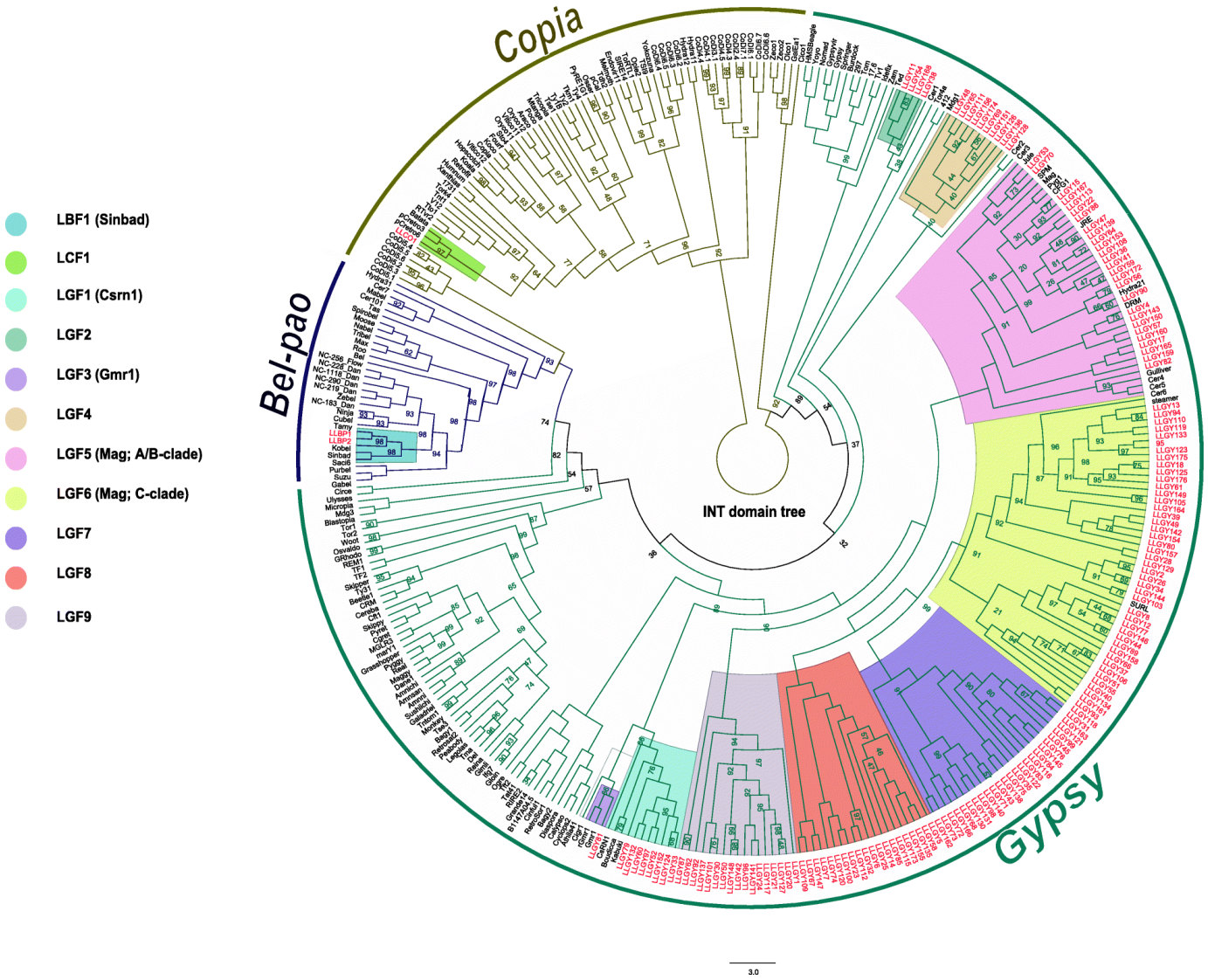

Figure 2.5. INT domain phylogenetic tree. INT phylogenetic tree was generated in IQtree with the LG + R7 model. Tree lines are color-coded according to the superfamily name above it. Elements in red are elements identified in the genome of *L. luymesi*.

Discussion

The deep-sea annelid *Lamellibrachia luymesi* genome contained at least 182 intact LTR retrotransposons which clustered into 12 families, 6 of which appear to be novel. All three known superfamilies of LTR retrotransposons – Gypsy, Copia and Bel-pao, were recovered, although several elements could not be classified in the existing families of these superfamilies.

Generally, LTR retrotransposons are known to be more abundant in plant genomes (e.g. > 50% in *Zea mays* genome; [22,67]) than in animal genomes (e.g. only 0.02% of the genome of C. gigas; [25]). In the genome sequencing study of *L. luymesi* done by Li et al., 2.52% of the genome were reported to be made up of LTR elements. Here, we expand this earlier effort to show that only ~0.1% of the genome is made up of intact LTR elements comprising mainly Gypsy representatives with a few Bel-pao and Copia elements. Importantly, many of these elements appear to represent families/clades new to science in addition to those that could not be classified. Our results, when compared to Li et al., indicates that most of the hits recovered by RepeatModeler and RepeatMasker are truncated, solo LTRs or nested LTR elements. However, a better understanding of LTR retrotransposon domains and a more robust database for LTR retrotransposon in non-model animals would likely allow a more accurate assessment as to the number, representation, and completeness of LTR retrotransposons in *L. luymesi*.

Comparative analysis done in eukaryotes such as crustaceans [37], fungi [24], *D. melanogaster* [68] and *B. mori* [69], show that Gypsy elements were the most abundant and with a high copy number. They are also the most diversified with numerous clades and families amongst the 3 superfamilies. Examination of LTR retrotransposons in *L. luymesi* genome corroborates these observations as 97% of the elements classified were Gypsy elements. According to our phylogenetic analysis, 3 previously described families including A-clade and C-clade of the Mag family, Gmr1 and CSRN1 were present in *L. luymesi*. Mag elements have been identified in diverse organisms such as *Caenorhabditis elegans* (roundworm, [70]), *Bombyx mori* (silkworm, [71]), *Anopheles gambiae* (mosquito, [66]) and *Xiphophorus maculate* (platyfish, [65]). In addition, a recent study shows that more than 290 Mag elements were identified in mollusc genomes [25]. Given their ubiquitous nature, Mag elements been the most common of the Gypsy elements found in *L. luymesi* is not surprising. Most of these Mag elements found are from Mag C-clade which includes SURL elements observed in marine echinoid species [35,72]). The LGF3 family in *L. luymesi* shared same lineage with the unusual Gmr1 clade. Gmr1 elements differ from other Gypsy LTR-retroelements in that the integrase domain

usually lie upstream of the reverse transcriptase domain, an arrangement mostly seen in Copia elements [64]. This clade includes elements that have been discovered in marine organisms such as the Atlantic cod *Gadus morhua* and the tunicate *Ciona intestinalis* [73,74]). In addition, the LGF1 family clustered within the CSRN1 clade, which was first described in a trematode [63] and is characterized by the elements Kabuki [75] CSRN1 [63], and Boudicca [76]. A recent study reports that CSRN1 clade is also represented in cephalopods [25]. *L. luymesi* also contained 5 novel families of Gypsy elements, making them the most diverse group of LTR retrotransposons in *L. luymesi*.

Copia elements appear to be less abundant in animal genomes than in plant genomes [67,37]. Here, only 2 intact Copia elements were identified in *L. luymesi*, consistent with these reports. Our phylogenetic analysis showed that these elements formed 2 distinct families, one previously described and one novel. The previously described family, GalEa, has been known to be one of the most predominant Copia retrotransposon as they are widely distributed among metazoans [25,77]. This element was the first Copia element found in crustaceans, specifically in a deep-sea squat lobster [77]. In a recent study [37], 29 out of 35 identified Copia elements from the deep-sea hydrothermal shrimp Rimicaris exoculata and other crustaceans belonged to the GalEa clade. Though, we only identified 2 Copia element in *L. luymesi*, one of them clustered within a clade found in marine metazoans, suggesting that this element may be common in marine environments. The other novel Copia element found herein did not cluster within any previously known families based on the RH domain tree (Fig. 2.4).

Recent studies of Bel-pao retrotransposons in metazoan genomes [55], including mollusc genomes [25] revealed that they are more abundant than Copia elements but lesser than Gypsy elements. In our case, an equivalent number of Copia and Bel-pao elements were found in *L. luymesi* genome. To date, seven Bel/pao families have been well described, namely, Bel, Pao , Sinbad, Suzu, Tas, Flow, and Dan [55]. A recent study further subdivides the Sinbad families into Sparrow and Surcourf [25]. In our study, the two Bel-pao elements clustered within the Sinbad family. Sinbad-like elements have been found in marine organism such as purple sea urchins, tunicates, pufferfish and the Atlantic salmon [78], making it a well described element in marine organisms.

The distribution of inferred insertion times of LTR retrotransposons found in *L. luymesi* suggests that current retrotransposons are recent features in the genome of this organism (Fig. 2.2). Further analysis on the most recently transposed elements (less than 1 million years ago) showed that most of these elements had incomplete domains and are scattered across identified families.

However, they all had identical LTR's indicating that they are yet to accumulate mutations. This finding augments the fact that these elements are indeed recent in the genome of *L. luymesi*. A previous study of insertion time estimates has shown that some superfamilies of retrotransposon shows activity at different times in waves while others show activity to be linearly related to time [79], another study suggests difference in spatiality and directionality of insertions among species [80]. However, the insertion time estimates of LTR retrotransposons in *L. luymesi* indicates that Gypsy elements showed a steady activity over a long period of time (more than 3MYA). Unfortunately, we could not make the same inferences for Bel-pao and Copia elements given their limited number.

Understanding the timing of transposon activity is important because transposable elements have been known to impact gene expression, by either generating new gene copies or regulating gene activity [81]. As such, the timing of these events may offer clues as to when such animals experienced bursts of evolution. However, to infer the possible role of transposable elements more fully in the animal genomes, other types of retrotransposons such as non-LTR retrotransposons or other transposable elements needs to be identified and annotated in these organisms.

Lastly, *L. luymesi* belongs to a group of animals known as Lophotrochozoans [82], a large diverse group of animals including groups such as Brachiopoda, Nemertea, Annelida, Mollusca, Phoronida etc. whose genome has been understudied in retroelements study. This and other studies e.g. [25,54] provides a foundation of knowledge that can be built upon to understand the role of retrotransposons in non-model and marine animals.

Chapter III. Genomic characterization of a novel, widely distributed *Mycoplasma* Species "*Candidatus* Mycoplasma mahonii" associated with the brittlestar *Gorgonocephalus chilensis*

Introduction

Mycoplasma species are one of the smallest and simplest self-replicating organisms, with a very reduced genome size that can range from about 540 to 1300 Kb. Mycoplasmas possess the minimum set of genes essential for growth and replication and evolved from the *Bacillus*/*Clostridium* branch of Gram-positive eubacteria by reductive evolution [83,84]. These bacteria lack a cell wall, a feature responsible for their pleomorphism and their resistance to some antibiotics [84]

Mycoplasmas encompass over 100 species and usually live in close association with their plant or animal host(s) to fulfill nutritional requirements [85]. Pathogenic mycoplasmas are responsible for numerous respiratory and other infections including pneumonia, pelvic inflammatory disease, and urethritis in humans [86]. In addition, the existence of mycoplasmas has been successfully documented in marine organisms including fishes (where they mainly colonize the intestines, gills, liver, and kidney) [84,87,88], cold-water corals [89], lobster [90], octopus [91,92], abalone [93,94] and squid [92]. They have also been found to be associated with the microbiota of echinoderms such as the blue bat star *Patiria Pectnifera* [95]. In these studies, mycoplasmas were in a commensal association with their host organism except in salmonid [84] where a mutualistic association was observed.

In this study, we report the discovery of a *Mycoplasma* with a 796kb genome (CheckM completeness of 97.9%) in the tissue of *Gorgonocephalus chilensis*, a filter-feeding basket star. To promote understanding of *Mycoplasma* spp. diversity, and symbiosis with marine invertebrates, we explored the new species' genomic composition and inferred metabolic capabilities. Additionally, given the understudied environment from which this cold-water filter-feeding echinoderm was discovered, we explored the degree of relatedness between *Ca*. M. mahonii and other *Mycoplasma* spp. using 16S rRNA and multilocus phylogenetic analyses and whether this novel species occurred in other basket stars. To the best of our knowledge, this is the first genetic isolation and characterization of mycoplasmas in the *Gorgonocephalus* genus and the name *Candidatus* Mycoplasma mahonii is proposed for this novel species.

Materials and Methods

Genome sequencing

Genomic DNA of *Gorgonocephalus chilensis* (Ophiuroidea, Euryalida, Gorgonocephalidae) was extracted from an individual sampled in Argentinian waters during an ARSV Laurence M. Gould expedition (LMG-0605, May 2006, latitude; 54 º 49, longitude -60 º 16, depth 110m). Tissue was stored at -80 ºC and subsequently sent to the University of Arizona Genomics Institute where DNA was isolated and sequenced using PacBio CCS long read technology using the protocol outlined below.

High molecular weight (HMW) DNA was extracted from ground tissue in extraction buffer with Tris HCl buffer 0.1M pH 8.0, EDTA 0.1M pH8, SDS 1%, and Proteinase K in 50 ºC for 30 minutes. The mixture was spun down and the aqueous phase transferred to a new tube. Next, 5M Potassium acetate was added, precipitated on ice, and spun down. After centrifugation, the supernatant was gently extracted with 24:1 chloroform: isoamyl alcohol. The upper phase was transferred to a new tube and DNA precipitated with isopropanol. DNA was collected by centrifugation, washed with 70% ethanol, air dried, and dissolved thoroughly in 1x TE followed by RNAse treatment. The DNA purity was measured using Nanodrop, DNA concentration was measured with Qubit HS kit (Invitrogen), and DNA size was validated by the Femto Pulse System (Agilent). The extracted HMW DNA was sheared to appropriate size range (10-30 kb) using Megaruptor 3 (Diagenode). The sequencing library was constructed following manufacturers protocols using SMRTbell Express Template Prep kit 2.0. The final library was size selected on a Blue Pippin (Sage Science) using S1 marker with a 10-25 kb size selection. The recovered final library was quantified with Qubit HS kit (Invitrogen) and size checked on Femto Pulse System (Agilent). The sequencing library was prepared with PacBio Sequel II Sequencing kit 2.0 for HiFi library, loaded to 8M SMRT cells, and sequenced in CCS mode in the Sequel II instrument for 30 hours.

Raw reads were assembled using HiFiasm [96] (N50 =1057833bp, Genome size = 3.5GB). The assembled genome of *G. chilensis* was screened for bacterial 16S rRNA by BLASTn against the NCBI's 16S rRNA database [97]. Contigs that matched to 16S rRNA genes were assigned taxonomic labels using Kraken2 Silva database v.138 [98], with default parameters, and were further analyzed. For all software used herein, default parameters were employed unless otherwise noted.

Completeness of each matched contig was determined with CheckM [99], using default parameters. Contigs having completeness higher than 90% and contamination lower than 10% were considered a ‘’complete’’ metagenomic assembly. Taxonomic assignments of complete contigs were conducted using GTDB-Tk v1.6.0 [100], based on the Genome taxonomy database [101]. GTDB-Tk uses a combination of metrics, including average nucleotide identity to reference genomes in the NCBI Assembly database, placement in the GTDB reference tree, and the relative evolutionary divergence.

Genome annotation of novel *Mycoplasma*

For the complete bacterial genome, gene and subsystem annotation was conducted using RAST v2.0 [102]. RNAs were annotated using RNAmmer v1.2 [103]. The Kyoto Encyclopedia of Genes and Genomes (KEGG- BlastKOALA) v2.2 [104] was used to predict biological pathways present in the genome. Ori-Finder 1 [105,106] was used to identify the Origin of Replication (OriC). Clusters of orthologous genes (COGs) were annotated using eggnog-mapper v2.1.7 [107]. The genome was scanned for virulence factors using the search BLAST search tool of the Virulence Factor Database (VFDB) [108] (database used – core and full dataset protein sequence). CRISPRCasFinder v1.1.2 [109] was used to validate predicted CRISPR/CAS systems. The average nucleotide identity (ANI) value was calculated using the Ezbio ANI calculator tool [110], and the average amino acid identity (AAI) was calculated using the webserver available through the Georgia Institute of Technology [111]. The dDDH calculator from Type Strain Genome Server (TYGS) [112] was used to predict digital DNA:DNA hybridization (dDDH) values from intergenomic distances for *Ca*. M. mahonii and its most closely related type strains genome sequences as implied in the Genome to Genome distance calculator (GGDC) [113].

Screening for *Mycoplasma* from multiple *Gorgonocephalus* individuals

Fifteen *Gorgonocephalus* individuals were screened for *Mycoplasma*. Nine *G. chilensis* from Argentinian waters were obtained during the LMG-0605 cruise in 2006 and 6 *G. eucnemis* were obtained in 2014 near the University of Washington’s Friday Harbor laboratories in the Northeast Pacific; Table S1. All samples were collected by KMH by trawl and subsequently stored in a -80 ºC freezer. Total genomic DNA was extracted using the Qiagen DNeasy blood and tissue kit (Maryland, USA) following the manufacturer's instructions except that lysis was done overnight to ensure

complete digestion due to the calcium carbonate in the arm tissue. Agarose gel electrophoresis was used to verify the integrity of the isolated DNA.

A 716bp region of the 16S RNA gene was targeted for PCR amplification using two oligonucleotide primers (Forward- 5'-ACTCCTACGGGAGGCAGCAGTA-3'; Reverse 5'-TGCACCATCTGTCAYTCYGTTAACCTC-3') that were slightly modified from previously published *Mycoplasma* universal oligopeptide primers [114] to be more specific to *Ca*. M. mahonii. Thermocycling conditions included an initial denaturation at 98 ºC for 30 sec, followed by 35 cycles of denaturation at 98 ºC for 10 sec, annealing at 60 ºC – 62 ºC for 30 sec, extension at 72 ºC 30 sec, and a final extension at 72 ºC for 5 min. Negative controls were employed in PCRs, samples from Argentinian waters and Northeast Pacific samples were handled in the lab on different dates. Amplified PCR products were examined by 1% gel electrophoresis and purified using the Qiagen QIAquick purification kit (Maryland, USA). The purified template was Sanger sequenced by Genewiz (New Jersey, USA), and bidirectional reads were trimmed and verified using Geneious software (v 2021.2.2) [115].

Phylogenetic analysis

To determine the phylogenetic affiliation of the new organism, two sets of phylogenetic analyses were conducted. One analysis employed a single gene tree based on 16S rRNA data, allowing for greater taxon sampling. The second analyses employed a multilocus tree, albeit with fewer taxa, to overcome the bias that can be caused by single-gene analysis (e.g. lineage sorting and selection pressure) and to provide a more robust representation of genomic data.

All available 16S rRNA sequences from *Mycoplasma* type strain were obtained from Ezbiocloud database [116] and GenBank [117], additionally, Mycoplasmatales bacteria DT_67 and DT_68, as well as 2 Oyster Mollicutes MAGs, were added to sequences obtained from this study to reconstruct the phylogenetic history of the group. *Bacillus subtilis* was used as an outgroup to root the resultant trees based on current understanding of *Mycoplasma* evolutionary relationships [118,119]. Sequences were aligned using MAFFT v7.475 [57], with default parameters. Maximum likelihood analysis with bootstrap was employed to reconstruct phylogenetic relationships in IQtree v1.6.12 [58] using the following parameters '-bb 100000, -nt AUTO, --runs 5'. These parameters were employed for all IQtree phylogenetic analyses in this study. The substitution model used for the 16S rRNA gene-based phylogeny, GTR+F+R10, was selected as the best model by IQtree's ModelFinder.

For the multilocus phylogenetic analysis, a representative with a sequenced genome from each major clade present in the 16S rRNA tree was selected and its genome screened for the presence of five single-copy housekeeping genes – *recA*, *lepA*, *dnaK, ruvB*, and *gmk*. This gene choice was based on the available literature for *Mycoplasma* phylogeny [120–124] . DNA sequences of these 5 housekeeping genes were aligned using the MAFFT option in Geneious (Geneious Prime 2023.0.4, Java version 11.0.15+10, 64-bit) and then concatenated. The multilocus phylogenetic tree of the concatenated alignment was reconstructed by employing maximum likelihood analysis with bootstrap in IQtree v1.6.12. The substitution model used for the multilocus phylogeny, GTR+F+R6, was selected as the best model by IQtree's ModelFinder. Additionally, tree topologies for the individual genes were also inferred using a MAFFT v7.475 alignment and maximum likelihood in IQtree v1.6.12. A Bayesian Inference analysis was also run for both the 16S rRNA and multilocus tree using MrBayes [125]. The same set of sequences as mentioned earlier was used, and the GTR+I+G model was chosen through JModeLTest2 [126]. on CRIPES Science Gateway (https://www.phylo.org/index.php/ (accessed on April 2023)), and with 1000000 generations sampled every 500 generations. Burninfrac was set to 0.25. All phylogenetic trees were visualized and edited using Figtree v1.4.4 [59].

Results

Microbial identification and classification

BLAST results of the *G. chilensis* assembly against NCBI 16S rRNA genes revealed 4 contigs that contained bacterial 16S rRNA gene fragments. Three contigs were classified as *Mycoplasma* while one was unclassified by the Kraken2 silva database. The 3 classified contigs were 51 Kb, 73 Kb, and 796 Kb in size respectively, 16S rRNA gene percentage identities of the 769kb contig compared to the other contigs ranged from 75% - 88%. Previously described *Mycoplasma* spp. genome sizes range from 540 Kb to 1300 Kb, hence we hypothesized that the 51 Kb and 73 Kb contig were likely to be incomplete or fragments of *Mycoplasma* genomes. CheckM analyses confirmed this interpretation as completeness of 5.72%, 7.14%, and 97.93% were reported for the 73 Kb, 51 Kb, and 796 Kb contigs respectively. In addition, CheckM only reported 0.38% contamination for the 796 Kb contig. Hence, we considered the 796 Kb contig to represent a complete genome, and further downstream analyses were conducted only on this MAG.

GTDB-Tk robustly placed this complete MAG in the order Mycoplasmatales and family Metamycoplasmataceae (GTDB-tk RED value of 0.93; Table S2) and phylogenetic analysis shows it to be related to a marine clade of *Mycoplasma*. This identified novel *Mycoplasma* genome was designated as "*Candidatus* Mycoplasma mahonii" (formal description given below).

Genome annotation

General features of the genome

The novel *Ca*. M. mahonii genome consists of a single chromosome of 796,768 bp with a GC content of 30.1% (Fig. 3.1). The 16S, 23S, and 5S rRNA genes were present as single copies, with the 16S and 23S rRNA genes located in the same operon and the 5S rRNA gene in a separate genomic region. Thirty-one transfer RNAs (tRNA) were identified, and all standard amino acids were represented. RAST predicted a total of 780 protein-coding sequences (CDS) of which 406 CDS (52.1%) were assigned putative functions and 374 CDS (47.9%) were annotated as hypothetical proteins. Repeats comprised 6.9% of the genome. Average gene length of predicted CDS was 887bp. Among the predicted CDS, 397 CDS (50.8%) were classified into Clusters of Orthologous (COG) families comprising 18 functional categories with most genes belonging to the J class (Translational, ribosomal structure, and biogenesis). RAST and eggNOG (COG) annotation were similar to that seen in other *Mycoplasma* spp. genomes based on RAST subsystems and COG classifications (Fig. 3.2). Additionally, RAST subsystem category gene counts were compared between *Ca*. M. mahonii and closely related species.

To verify the functional abilities of the recovered MAG (i.e., the proposed genome), we examined a series of molecular and cellular pathways and structures, and below we describe select major systems in turn.

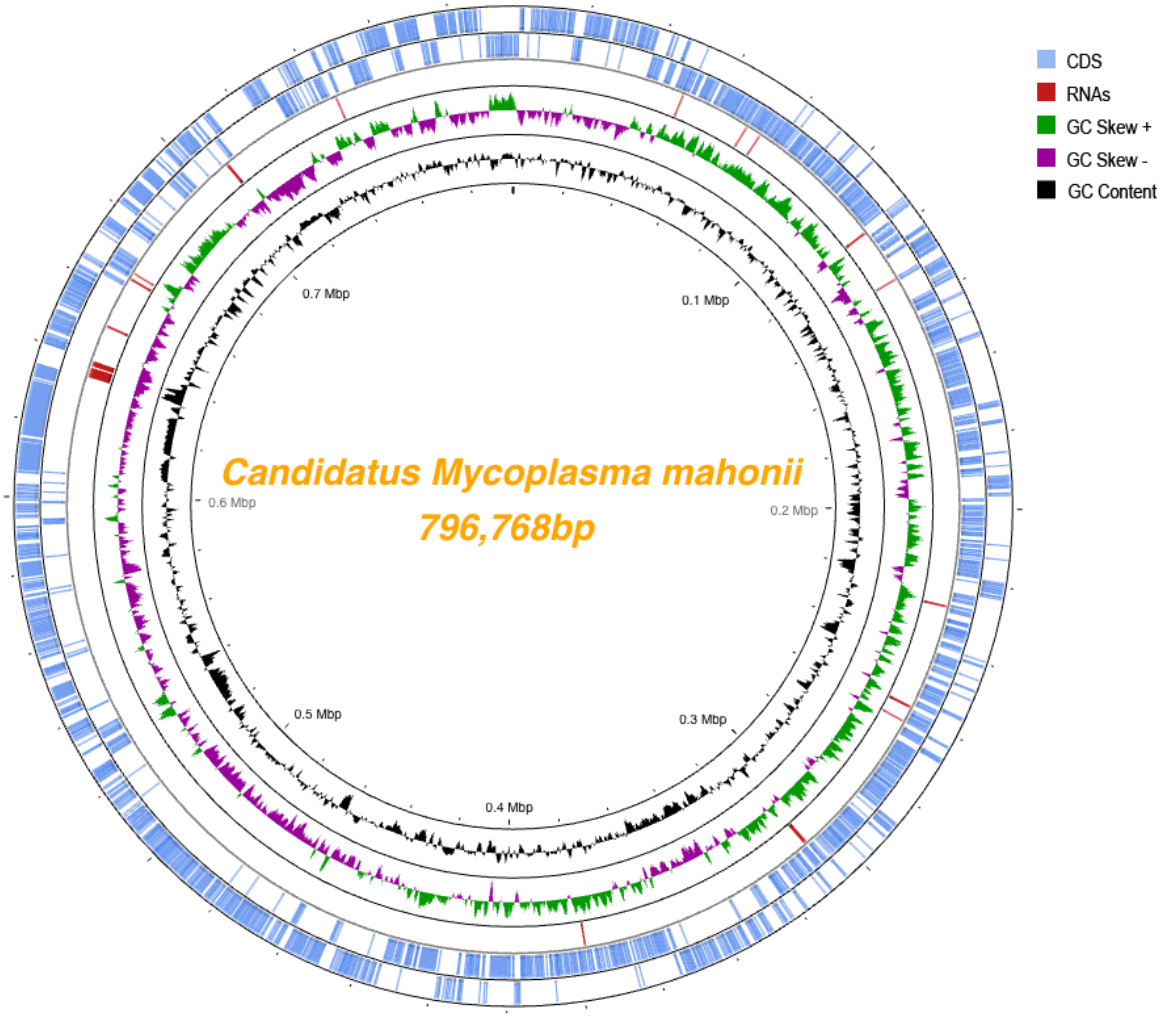

Figure 3.1. Chromosome atlas of *Candidatus* Mycoplasma mahonii. The scale is shown by the inner black circle. Starting with the outermost rings, the 1st and 2nd circles show predicted coding sequences on the minus and plus strands respectively. The 3rd circle represents RNAs including both tRNAs and rRNAs. The 4th circle represents GC skew (G−C)/(G+C) (green-above mean, purple-below mean; mean = 0.5) and the 5th circle represents mean-centered G+C content of the genome. The figure was generated using Proksee.

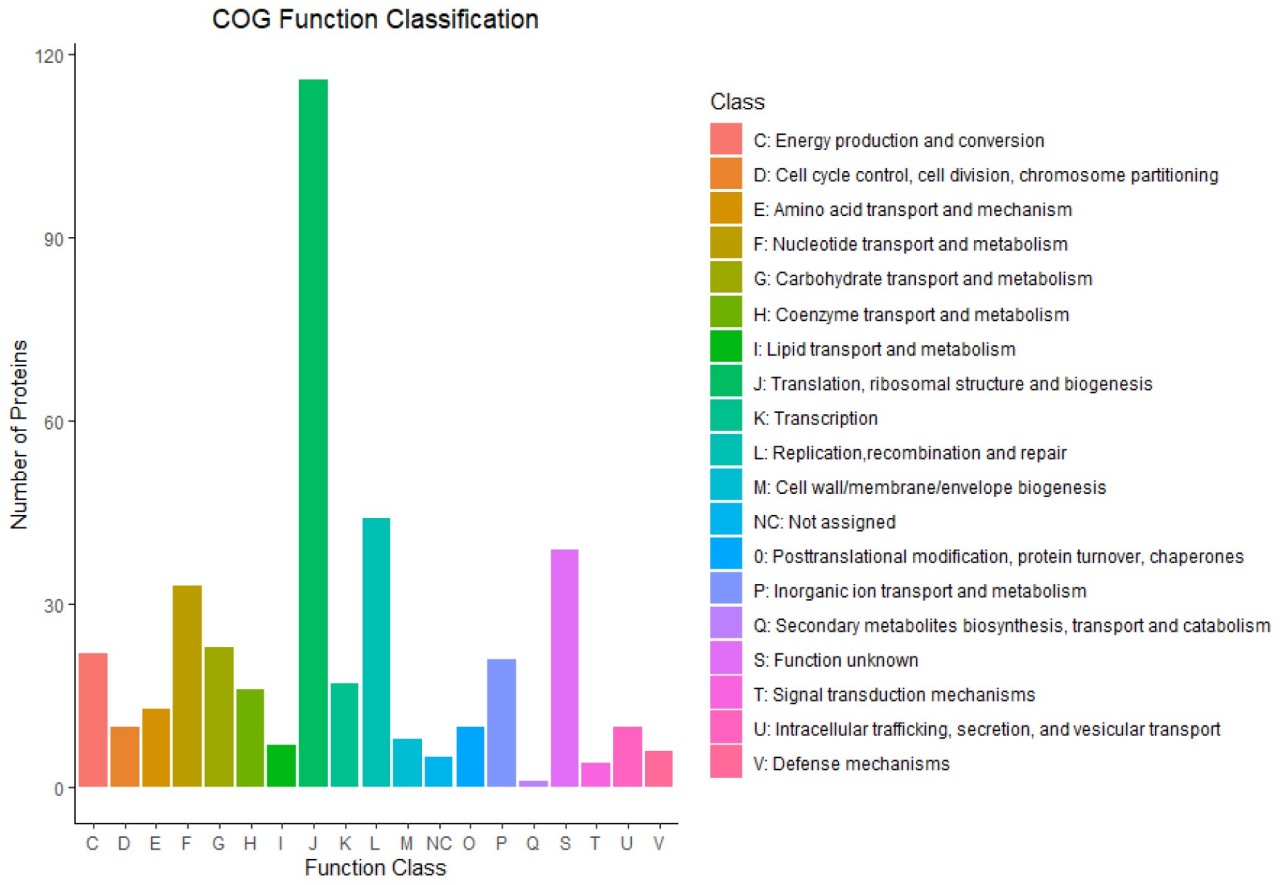

Figure 3.2. Clusters of orthologous group function classification. The Y-axis represents the number of protein/unigenes belonging to a particular category while the X-axis represents the COG categories.
Replication, transcription, and translation.

For DNA replication in Mycoplasma, the OriC usually contains *dnaA* boxes and is generally around the vicinity of the *dnaA* gene [127–130]. Using previously published dnaA box motif consensus sequence 5'-TTATCCACA-3' [127,131], and allowing one mismatch, two putative replication origins were found in the area surrounding *dnaA* gene using Ori-Finder. Both regions possessed the typical features of OriC in prokaryotes (i.e. *dnaA* box, *dnaA* gene vicinity, GC skew inversion) [132].

A total of 43 genes were predicted to be involved in replication, recombination, and repair. DNA repair appears to be mainly executed by nucleotide excision repair, SOS repair system, and recombination repair. No mismatch-repair system genes (MutHLS) were found.

For transcription, a total of 11 genes were predicted to be involved. Transcription termination and elongation are regulated by *Nus A*, *B*, *G*, and *GreA* genes in this organism. GreA prevents transcription arrest while the Nus proteins can induce transcription pausing or stimulate anti-

termination [133]. In addition, two transcription factors (*HigA* and *MraZ*) and one heat-inducible transcription repressor gene (*HrcA*) were identified.

Additionally, a total of 126 genes were predicted to be involved in translation, ribosomal structure, and biogenesis including 49 ribosomal proteins, 24 aminoacyl tRNA synthase genes, and 11 translation factors.

Secretion system and transporters

The transporter system of *Ca*. M. mahonii consists of 39 genes, which are mainly made up of the ATP-binding cassette (ABC) transporter system and the phosphotransferase (PTS) system (Table S8). For the PTS system, 4 genes that encode proteins required by the PTS system were present; pts1 which encodes Enzyme 1 (E1), *ptsH* which encodes Phosphocarrier protein *Hpr*, *fruA* which encodes a fructose-specific II component (EIIBC or EIIC) and *celA* which encodes a cellobiose-specific II component (EIIB).

Genes encoded by the ABC transport system include 12 ATP-binding proteins, 11 permease proteins, and 2 substrate-binding proteins. Three complete ABC-type transport systems including a phosphate transporter, a phosphonate transporter, and a general nucleoside transporter were present, while others such as oligopeptide transporter, energy-coupling factor transporter, saccharide/lipid transporter, etc. were incomplete. Other genes that were not part of the PTS system or the ABC transport system were associated with other transporters such as magnesium transporter, riboflavin transporter, potassium uptake system proteins, cytosine permease, and an adenine/guanine/hypoxanthine permease.

Genes such *as secA, secY, secD/F, secE, secG, ffh, yidC,* and *FtsY*, which make up the core proteins required for the Sec-SRP secretion pathway (SEC system) and two other genes associated with the Type II secretion system (comEB and comEC) were present as part of the organism's bacterial secretion system.

Metabolism

According to the annotation data, *Ca*. M. mahonii encode all needed enzymes in the Embden-Meyerhof-Parnas (EMP or glycolysis) pathway, arginine deaminase pathway, F1-ATP synthase, PRPP (phosphoribosyl-pyrophosphate) production, and pyruvate oxidation. Enzymes associated with lipid metabolism in this organism suggest that they utilize glycerol to generate phospholipids (cardiolipin), the only membrane component identified in this organism.

Genes involved in de novo nucleotide synthesis were lacking in *Ca*. M. mahonii; however, this bacterium was predicted to encode nucleotide salvage pathways genes such as *nrdA* and CTP synthase. *Ca*. M. mahonii lacks genes predicted to be required for different intermediate pathways such as the TCA cycle, citric acid cycle, and the oxidative phase of the pentose phosphate pathway. Partial pathways for CoA biosynthesis (specifically pantothenate to CoA), phospholipid biosynthesis, riboflavin/FMN/FAD biosynthesis, Tetrahydrofolate (THF) biosynthesis amongst others were present within the *Ca*. M. mahonii genome.

Defense System

Two bacterial defense systems to ward off phage infection, including a Type II Restriction - modification (R-M) system and a CRISPR-CAS system are present in the genome of *Ca*. M. mahonii. The Type II restriction system possessed only genes encoding methylase and methyltransferase enzymes but lacks the sequence-specific endonuclease which is usually responsible for recognizing and cleaving specific DNA sequences. No Type I or Type III R-M genes were found. On the other hand, RAST annotation (and CRISPR finder tools) identified 1 CRISPR array of length 9820bp, 148 spacers, and the direct repeat consensus GTTTAAGAATACACAAGAATGATACCACCCCAAAAC. Additionally, a thioredoxin reductase system which is predicted to provide defense against oxidative stress was also present.

Screening for Mycoplasmas in *Gorgonocephalus*

Nine out of 15 screened basket star samples (6 *G. chilensis* samples from Argentinian waters and 3 *G. eucnemis* samples from the Northeast Pacific) had a positive PCR result and these amplicons were Sanger sequenced. However, only 1 *G. eucnemis* sample had a good quality read and was the only *G. eucnemis* sample used in the phylogenetic analysis. These sequences have been deposited to NCBI under the accession numbers OP995472-OP995479.

Phylogenetic results

16S rRNA gene-based (Fig. 3.3) and multilocus (Fig. 3.4) phylogenetic analyses were used to validate the GTDB-Tk taxonomical assignment and provide higher resolution phylogenetic placement. The 16S rRNA gene analysis consisted of 129 sequences and the multilocus tree consisted of 61 terminals, including sequences from this study. Both analyses produced congruent results. *Ca*. M. mahonii was placed within the *Mycoplasma* genus, within a recently characterized

lineage of marine taxa, consisting of *M. marinum* (isolated from *Octopus vulgaris*) and *M. todarodis* 5H^T (isolated from the squid *Todarodes sagittatus*) [92] based on the multilocus tree. Additionally, 2 marine metagenomes, M. DT_67 and M. DT_68 which were isolated from ocean water sinking particles collected at abyssal depths in the Nort Pacific Subtropical Gyre were also part of the marine clade. The 16S rRNA topology placed *Ca*. M. mahonii with these same taxa plus *Candidatus* Mycoplasma corallicola (which was not included in the multilocus tree due to an incomplete genome). Furthermore, the *Ca*. M. mahonii 16S rRNA sequences from both Argentinian waters and North Pacific were ~99.7% identical and formed a monophyletic clade on the 16S rRNA gene-based phylogenetic tree. The Bayesian analysis also showed similar results. *Ca*. M. mahonii formed a monophyletic clade with *Ca*. M. corallicola, M. DT_67*, M. marinum*, and *M. todarodis* in the 16S rRNA-based phylogenetic tree and also formed a monophyletic clade with M. DT_67, M. DT_68, *M. marinum*, and *M. todarodis* in the multilocus tree. M. DT_68 lacks a 16S rRNA gene and hence was absent in the 16S rRNA phylogenetic tree.

Additionally, 16S sequence similarity analysis using TrueBac ID from Ezbiocloud [134] indicated that *Ca*. M. mahonii was most similar to *Ca*. M. corallicola (89.41%), *M. todarodis* (89.1%) and *M. marinum* (88.9%). TrueBac ID also indicated high similarity to *M. mobile* (88.9%). However, all the phylogenetic analyses conducted showed *M. mobile* as an unstable branch as its position varied on the different trees. Consequently, *M. mobile* was not further considered as a closely related species. Genomic features of *Candidatus*. M. mahonii and closely related taxa are shown in Table 3.1.

ANI similarity values for *Ca*. M. mahonii and its closely related species were 68.2% (M. marinum) and 67.4% (*M. todarodis*), while the AAI was 50.3% (*M. marinum*) and 49.25% (*M. todarodis*). A total of 422 orthologous gene groups are shared between *Ca*. M. mahonii, M. todarodis, and *M. marinum* (Fig. 3.5). In addition, the dDDH scores from TYGS pairwise comparison of *Ca*. M. mahonii against identified close strains were less than 70% (~21.5% - 27.9%) indicating that *Ca*. M. mahonii is indeed a novel species.

Table 3.1. Genomic features of *Candidatus*. M.mahonii and closely related taxa

	Candidatus Mycoplasma mahonii	*M. todarodis*	*M. marinum*
Genome completeness	97.9%	85%	97.1%
Genome size (bp)	796,768	1,007,879	1,171,149
Number of contigs	1	84	128
%GC	30.1%	30.95%	28.41%
Number of CDS	780	914	1003
Hypothetical genes	374	296	328
16S rRNA gene	1	1	1
23S rRNA gene	1	1	1
5S rRNA gene	1	2	2
Number of tRNAs	31	39	42

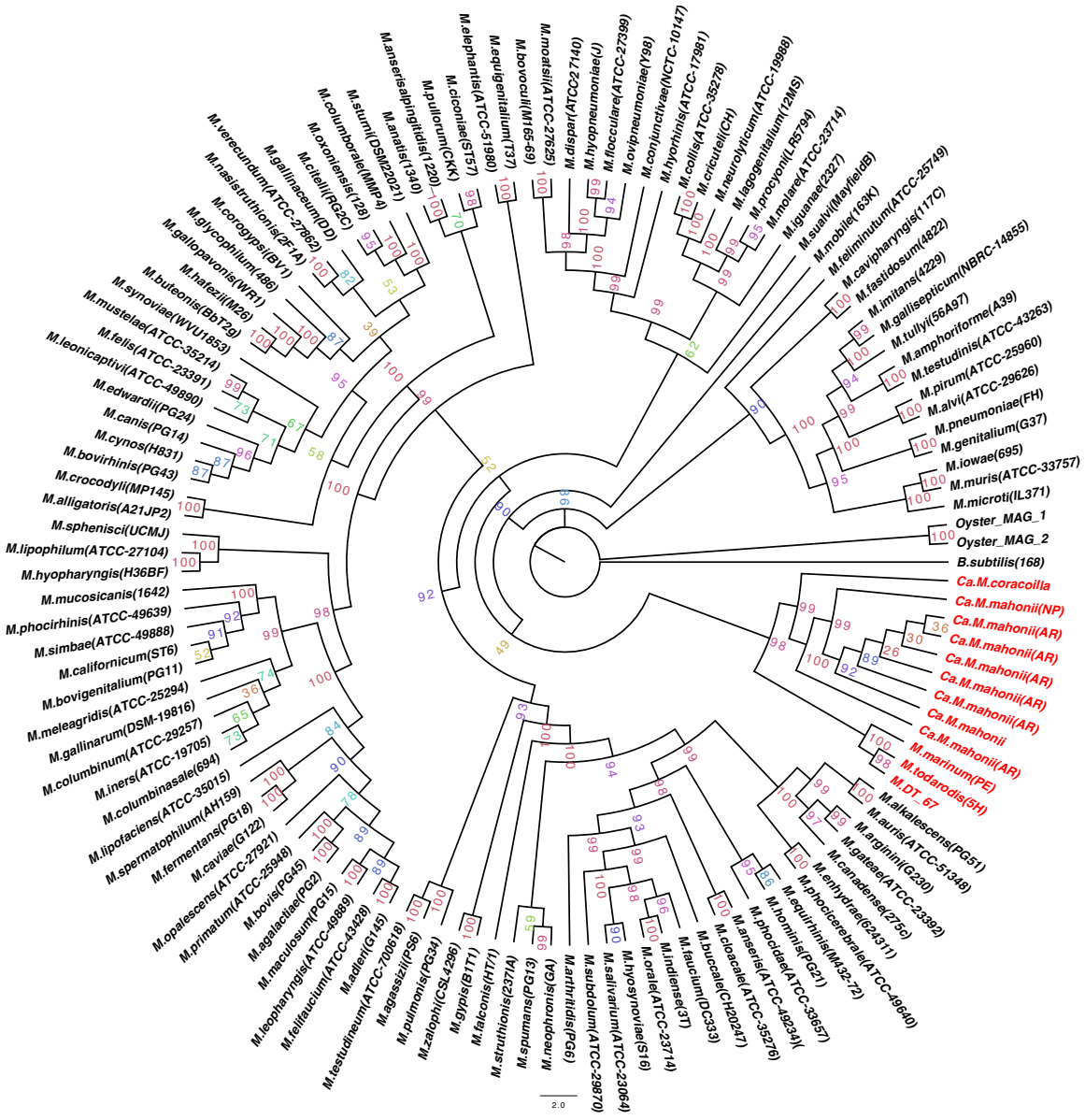

Figure 3.3. 16S rRNA maximum-likelihood phylogenetic tree. The 16S rRNA gene tree was reconstructed based on all available *Mycoplasma* spp. type strain available from the Ezbiocloud and GenBank databases, Mycoplasmatales bacteria DT_67 and DT_68, 2 Oyster Mollicutes Mag and *Ca*. M. mahonii. Bootstrap percentage values are shown on the tree. The tree was generated in IQtree with the GTR+F+R10 model. *Ca*. M. mahonii and other sequences making up the distinct marine clade are shaded red. AR – Argentinian waters samples, NP – North Pacific samples.

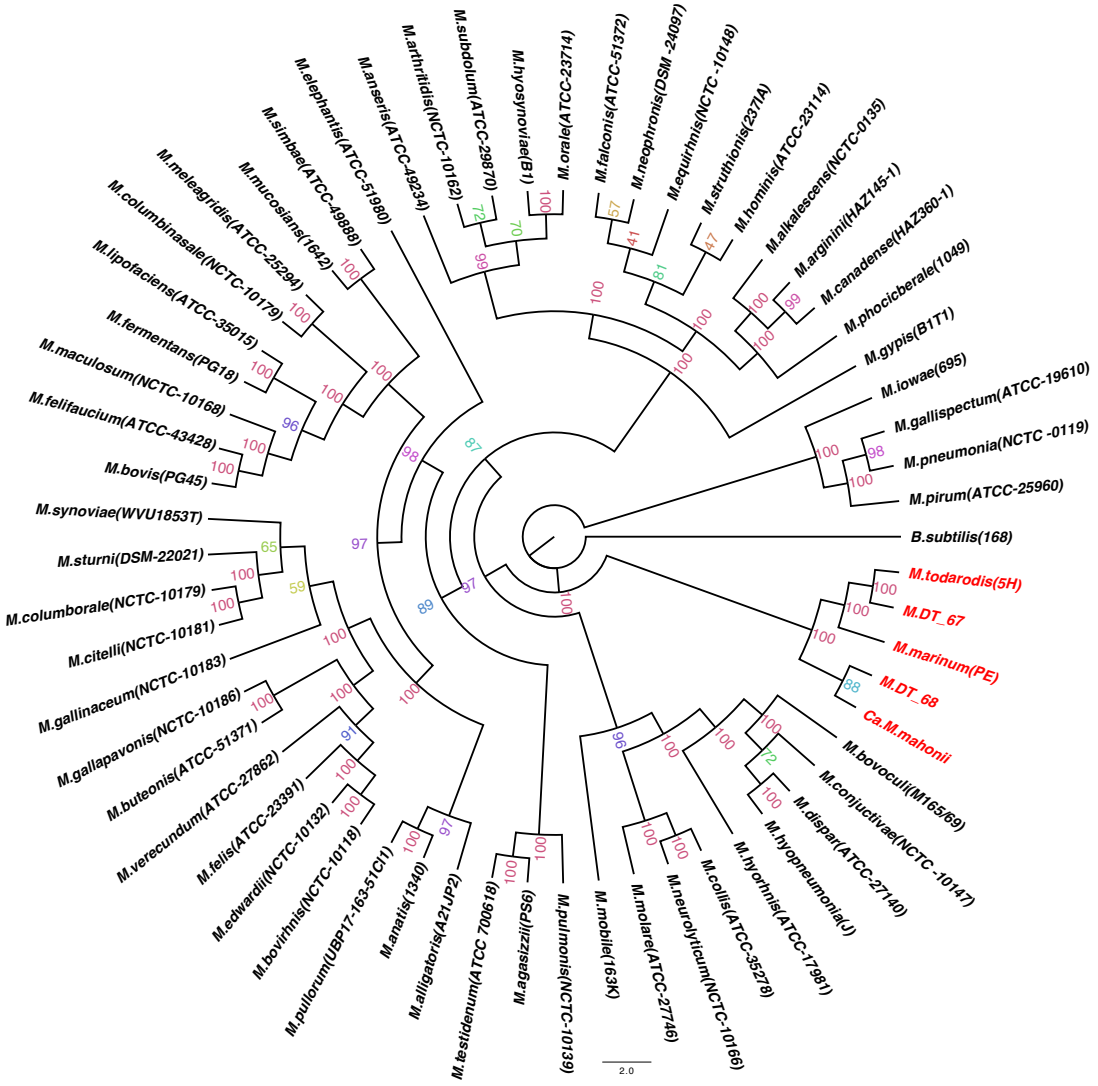

Figure 3.4. Multilocus maximum-likelihood phylogenetic tree. The multilocus phylogenetic tree was generated based on the concatenated sequence of five single-copy housekeeping genes – *recA, lepA, dnaK, ruvB,* and *gmk*. Bootstrap percentage values are shown on the tree. The tree was generated in IQtree with the GTR+F+R6 model. *Ca*. M. mahonii and other sequences making up the distinct marine clade are shaded red.

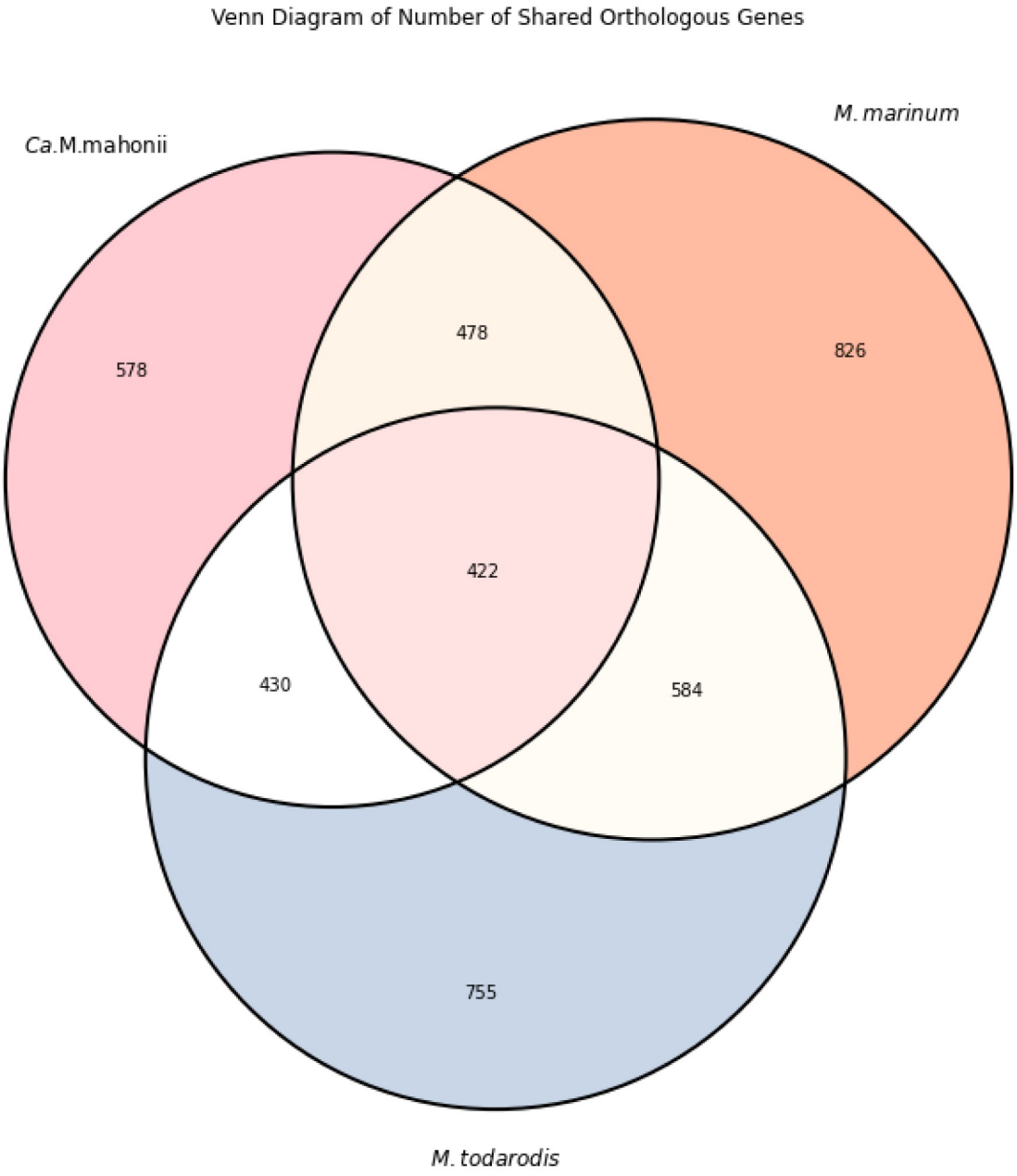

Figure 3.5. Venn diagram showing the number of shared orthologous genetic groups (OGs) between *Ca*. M. mahonii, *M. marinum* and *M. todarodis*.

Description of "*Candidatus* Mycoplasma mahonii"

The category "Candidatus" is used to describe prokaryotic entities for which information other than just a DNA sequence is available but lacks other characteristics required for description according to the International Code of Nomenclature of Bacteria [135]. The *Mycoplasma* genome described here represents a novel species of *Mycoplasma* and is currently the only representative of this candidate species. The species is designated "*Candidatus* Mycoplasma mahonii" (N.L. gen. masc. n. mahonii, of Mahon, named in honor of long-time Antarctic collaborator Andrew Mahon). This species was isolated from *Gorgonocephalus chilensis* (collected May 2006, latitude; -54^{o} 49, longitude -60^{o} 16, depth 110m) and assignment to "*Candidatus* Mycoplasma mahonii" is based on (i) the associated 16S rRNA gene sequence; accession number - OP995479), (ii) Similarity index score (ANI) of <95% to closest relatives [89,92], (iii) 97.9% genome completion (according to CheckM analysis), (iv) primer sequence complementary to a region of 16S rRNA- 5'-ACTCCTACGGGAGGCAGCAGTA-3'.

Genome size was ~796 Kb with a G+C content of 30.1%. Within the genome, single copies of rRNA genes and 31 tRNA genes were identified. KEGG-based analysis identified the presence of the following pathways Embden-Meyerhof-Parnas pathway, F1-type ATP Synthase, Acetate production from acetyl-CoA, Folate (vitamin B9) biosynthesis predicted from 7,8-dihydrofolate, nucleotide sugar biosynthesis pathway amongst others. Additionally, KEGG identified both PTS and ABC transport system genes. The genome sequence can be found under NCBI BioSample ID-SAMN32235174 and Accession ID – CP114583.

Discussion

A novel *Mycoplasma* species, *Candidatus* Mycoplasma mahonii associated with the basket sea star *Gorgonocephalus chilensis* inhabiting the Argentinian waters, was discovered. This taxon also occurs in *Gorgonocephalus* samples from the North Pacific. Phylogenetically, it is part of a recently characterized clade of non-free-living marine lineage of mollicutes that use marine invertebrate organisms as hosts.

The metagenomic assembly of *Ca*. M. mahonii consists of a single 796,768bp contig with a total of 780 predicted protein-coding sequences (CDS). Genomic features of *Ca*. M. mahonii are comparable to previously described *Mycoplasma* species in several respects: 1) The number of

predicted CDS in *Ca*. M. mahonii is comparable to other *Mycoplasma* spp., with 635 in *M. mobile* [136], 677 in *M. pneumonia* [137], and 742 in *M. gallisepticum* [130]; 2) The arrangement of rRNA genes is similar to those found in the genome of *M. mobile* [136] and *M. pulmonis* [138]; 3) Although the exact OriC could not be determined, *Ca*. M. mahonii possesses a tandem arrangement of *dnaA* and *dnaN* genes around the OriC as seen in other mycoplasmas [128,139,140]; however, in the case of *Ca*. M. mahonii, ribosomal protein L34 *(rmpH* gene) was also located upstream of the *dnaA* gene in an opposite direction; and 4) The Peptide Release Factor (RF-1) which recognizes the stop codons UAA and UAG and terminates translation was present in *Ca*. M. mahonii, although RF-2 which identifies the UGA stop codon was absent suggesting that this stop codon codes for a protein as seen in other mycoplasmas [141].

Due to its reduced genome size, *Ca*. M. mahonii lacks genes involved in de-novo biosynthesis of nucleotide, lipids, co-factors, and intermediate energy metabolism pathways such as the TCA cycle, citric acid cycle, phosphate pathway, etc., imposing a host-dependent lifestyle on this organism. However, this bacterium is predicted to encode genes involved in the nucleotide salvage pathway such as *nrdA* which converts ribonucleotides to deoxyribonucleotides, and genes involved in nucleotide interconversion such as CTP synthase which converts UTP to CTP as well as permease for pyrimidine and purine transport which allows them to take up these molecules.

Most transport protein-encoding genes in *Ca*. M. mahonii are associated with ABC transport which transports a range of molecules such as peptides, lipids, phosphate, ions, iron, etc., and PTS transport system which transports extracellular sugars such as mannose, fructose, and cellobiose. The presence of these broad substrate transport systems compensates for the lack of various other transport systems such as GLUT (glucose transporters) and may allow the microorganism to obtain nutrients directly from its host rather than synthesizing de-novo, a trend common in mycoplasmas [142]. Additionally, a gene encoding *TrkA* which is responsible for potassium uptake and is necessary for intracellular survival in prokaryotes was predicted to be present in the genome of *Ca*. M. mahonii.

The lack of a complete TCA cycle, quinones, or cytochromes rules out the possibility of ATP generation through oxidative phosphorylation in this bacterium. Metabolic pathways present in *Ca*. M. mahonii suggest that they are glycolytic species that rely on energy generation through fermentation of sugars, ATP-synthase, pyruvate oxidation to acetate, and hydrolysis of arginine. Additionally, in the non-oxidative phase of the pentose phosphate pathway present in *Ca*. M.

mahonii, transaldolase (which catalyzes the transfer of a dihydroxyacetone group from donor compounds to aldehyde acceptor compounds) is absent. This reaction is presumably carried out by an unrecognized protein as the pentose pathway has been reported in other mycoplasmas to be incomplete but functional [136,143]. In the case of *Ca*. M. mahonii, this reaction is likely carried out by the non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN) enzyme, as the gene encoding this enzyme was predicted in the annotation. GAPN reduces NADP to NADPH and can maintain NADPH production in bacteria lacking some pentose phosphate enzyme [144].

The defense systems present in *Ca*. M. mahonii includes R-M Type II system, CRISPR/CAS system, and thioredoxin system. The CRISPR/CAS system and R-M system are a natural pathogenic adaptive immune system that protects prokaryotic organisms against invading nucleic acids most especially viruses [145,146]. The R-M system is present in almost all mollicutes sequenced so far [145] while the CRISPR/CAS system has been reported in some but not all mollicutes [147]. On the other hand, the thioredoxin system present in *Ca*. M. mahonii protects it from oxidative stress [148] and has been reported in some *Mycoplasma* species such as *M. suis* [140], *M. bovis* [149] and *M. capricolum* [148].

Virulence factors typically associated with Mycoplasma such as adhesins [83], ClpC ATPase [139] , variable surface lipoproteins (Vsps) [150,151], capsular polysaccharides [152], were absent in *Ca*. M. mahonii. Moreover, no virulence factor was detected using the BLAST search tool of the VFDB database, suggesting that *Ca*. M. mahonii is potentially a non-pathogenic *Mycoplasma* species. Interestingly, the absence of virulence factors was also observed in other members of the distinct marine clade of Mycoplasmas namely *M. marinum*, *M. todarodis*, and *Ca*. M. corallicola, suggesting that this monophyletic clade of mycoplasmas are commensals and potentially a natural part of its host microbiome.

Lastly, the high percent identities (~99.5%) between the 16S rRNA genes of *Ca*. M. mahonii from *Gorgonocephalus* chilensis found in Argentinian waters and the Northeast Pacific (*Gorgonocephalus eucnemis*), suggests that this species is broadly distributed and likely native to multiple *Gorgonocephalus* host species. The annotation of *Candidatus* Mycoplasma mahonii, conducted herein, is the first step to understanding the biology and potential pathogenicity of this bacterium. Future studies will expand on this knowledge by focusing on the metabolic pathway interplay between this species and its basket star host.

Chapter IV. Comparative Analysis of Microbial Composition of Farm-raised and Wild Oysters

Introduction

Crassostrea virginica, commonly known as eastern oyster, is a commercially important bivalve mollusc that inhabits estuarine and coastal environments. The microbial composition and overall health of oysters have been a long topic of interest because of their economic benefits and the ecological role they play in ecosystems. *C. virginica*, like other oysters, are filter feeders, hence, they improve water quality by removing particles from water columns to get food. They are also considered ecosystem engineers due to their ability to form reefs that serve a variety of beneficial functions such as carbon sequestration, protection against erosion, and creating habitats for other marine organisms [19]. Oyster production is a vital and growing agricultural sector in the United States with a farm gate value estimated to be $219 million in 2018 [153].

Due to their filter-feeding behavior, oysters usually interact significantly with living and non-living particles in their environments including microbes [154]. Microbial communities associated with oysters have been previously studied using both culture-dependent and independent methods [155156] [157,158]. These host-associated microbial communities are known to perform a variety of beneficial functions to the host such as providing nutrition, producing antimicrobials, influencing immune response, and reducing proliferation of detrimental microbes. On the other hand, bacteria such as *Vibrio*, *Salmonella* spp., and norovirus which can be found in oysters are of particular concern due to the health implications associated with their consumption. Pathogenic species such as *Vibrio vulnificus* and *Vibrio parahaemolyticus* are known to cause severe illness, or even death, when consumed [159,160] and norovirus is the leading cause of nonbacterial illness in shellfish consumers [161]. Oysters are able to concentrate microbes from contaminated water and are often eaten raw or lightly cooked, hence, they typically serve as vectors of these pathogens to humans.

Commercially, oysters can be wild-caught or grown as part of farming activities. Farmed oysters are frequently grown using suspended grow-out systems known as 'floating cages' [162]. As a result, farmed oysters can be cultivated near the surface of the estuary, whereas wild oysters usually grow on the benthos. Due to this difference, farm-raised and wild oysters experience different growth conditions such as differences in temperature, agitation, water-column height, UV radiation,

and handling [163]. Such variation in environmental parameters can lead to variation in microbial communities [154,158,164].

Understanding differences in microbial composition between wild and farm-raised oysters is not only crucial to ensure the safety of oyster consumption, but it is also integral to effective management strategies to sustain the oyster aquaculture industry. Here, we employed a metagenomic approach to study microbial communities present in farm-raised and wild oysters that are in close spatial proximity. We examine to what degree the variation in microbial diversity and functional pathways found in both wild and farm-raised (caged) oysters differ. Our approach overcomes the shortcoming of 16S rRNA-based approaches such as limited functional information and low taxonomic resolution[165], and provides a more general view of the microbial communities in these oysters.

Materials and Methods

Sample collection

Eight wild oysters and eleven three-year-old farmed oysters (*C. virginica*) were collected by hand from the University of North Carolina Wilmington shellfish lab. The farmed oysters were grown in floating cages on the surface of the estuary while the wild oysters were at the bottom of the water column. Samples were collected between November 12th to November 14th, 2021. Effort was taken to ensure that the dimension and general attributes (e.g., shape, fouling) were similar between the oysters sampled in the study.

Collected oysters were immediately taken to the laboratory where each individual oyster's outer shell was thoroughly rinsed with cold filtered saltwater to eliminate any visible sediments. Subsequently, the oysters were shucked aseptically, the fluids were drained using a sterile syringe, and a microbial sample was taken the inner surface of the top and bottom shell was taken using a sterile cotton swab. To maintain sample integrity, the collected samples were stored in a -80 ^{o}C freezer until further processing.

DNA isolation and sequencing

DNA from the swab samples was processed and isolated using the ZymoBiomics DNA Microprep kit following the manufacturer's protocol with slight modifications that included

incubating proteinase K at 50^0C for 2 hours, increasing wash centrifuge speed to 15,000 x g, and using the DNase/RNAse free water eluted DNA for downstream analysis.

Isolated DNA was then sent to Novogene for shotgun metagenomics library preparation and sequencing. Paired-end sequencing was performed on the Illumina NovaSeq platform (PE150). All sequences obtained have been deposited to National Center for Biotechnology Information (NCBI) Small Read Archive (SRA) under accession numbers (currently pending).

Preprocessing of sequence data

Raw sequencing reads obtained from the Illumina NovaSeq sequencing platform were processed using fastp [166]with default parameters. In this step, low-quality reads were removed (including reads containing adapters, polyG tail trimming, and reads with over 40% of bases having q-values ≤15). To focus on oyster microbiome, the quality-filtered reads were then mapped with bwa-mem2 [167] to the genomes of *C. virginica* (RefSeq assembly accession GCF_002022765.2) and the human genome (RefSeq assembly accession - GCF_000001405.40), available from NCBI. Reads unmapped to either of these genomes were retained for downstream analysis of microbiomes.

Bioinformatic analysis

In order to compare and contrast diversity and putative functional pathways of both wild and farm-raised oyster microbiomes, we employed a series of bioinformatic routines that focused either on assessing biodiversity, assigning functional pathways observed in the data, or exploring metagenome-assembled genomes (MAGs) (Fig. 4.1).

Figure 4.1. Bioinformatic Pipeline

Taxonomic profiling

Default parameters of Kraken2 [98] were used for taxonomic classification of quality-filtered reads and Bracken [168] was used to compute taxonomic abundance using taxonomy labels assigned by Kraken2. Resulting taxonomic profiles (count data) from individual samples were merged using Kraken-biom [169] and employed for further statistical analysis.

Prior to statistical analysis, taxonomic profiles were filtered, and only taxonomic features present in at least 11% of the samples (i.e., in at least 2 samples out of 19) were retained for subsequent analyses. Filtered data were normalized by rarefaction to account for differences in sequencing depth using phyloseq package [170]. Filtered data and rarefied data were used for alpha and beta diversity calculations in R studio using phyloseq [170] and vegan package [171]. Observed diversity and Alpha diversity metrics such as Shannon, and Chao1 metrics were used to evaluate bacterial diversity and richness. The observed metric measures the actual/observed diversity within a community, the Shannon index considers both the diversity and evenness of species in the community, while Chao1 estimator provides an estimate of the total number of species based on the presence of rare or uncommon species. The Kruskal-Wallis non-parametric test was employed to evaluate the significant effect of the environments on the diversity metrics.

To assess differences in diversity and taxonomic composition between farm-raised and wild oysters, a principal coordinate analysis (PCoA) based on Bray-Curtis dissimilarity distance was conducted. The statistical difference between the groups (environments) was evaluated using PERMANOVA and PERMDISP, which employed the adonis2 and betadisper functions, respectively, in the R vegan package [171]. PERMANOVA tests whether the centroids of all groups are equivalent by comparing the distances between samples within the same group to the distances between groups while PERMDISP evaluates whether the dispersion(variation) between samples differs from the dispersion between groups. Filtered and unrarefied taxonomic profiles were used to calculate relative abundance for the community composition using the phyloseq package, and results were visualized using ggplot2 [172].

Differential abundance analysis was performed to identify significantly abundant taxa between the farm-raised and wild oyster groups using DESeq2 [173]. DESeq2 function performs differential abundance analysis by: (1) estimation of size factors, (2) estimation of dispersions from the negative binomial likelihood for each feature; (3) fitting a negative binomial generalized linear model to each feature on a specified class and performing hypothesis testing using Wald test.

The results were obtained using the results function, considering only features with a Benjamin-Hochberg (BH) FDR-adjusted p-value of 0.01 (Adjusted FDR < 0.01).

All statistical analyses were done in R studio version 4.2.3 [174].

Functional profiling

Quality-filtered reads (excluding the 2 pruned samples) were profiled for potential functional content (UniRef90 gene-families [175] and MetaCyc metabolic pathways [176]) using HUMAnN3 [177] with default parameters. HUMAnN3 utility tools were used to prepare data for subsequent analyses. Briefly, humann_join_table and human_split_startified_table were used to merge the functional profile of individual samples and filter the merged profiles to contain only community-level data (unstratified table), respectively. The humann_regroup_table was used to convert the Uniref90 gene families into Kyoto Encyclopedia of Genes and Genomes (KEGG) ortholog (KO) [178] groups. Converted KO groups and MetaCyc metabolic pathway abundance data were renormalized from reads per kilobase (RPKs) to relative abundance using humann_renorm_table, for downstream analysis.

To identify differentially abundant KO groups and MetaCyc pathways, MaAsLin2 [179] was employed. MaAsLin2 fits a linear model to each feature's transformed abundance on a specified sample grouping, tests significance using a Wald test, and outputs BH FDR-corrected p-values. Default parameters ($p < 0.05$; adjusted FDR <0.25) were used except that minimum prevalence was set to 0.4, filtering the data to only test features with at least 40% non-zero values, thereby filtering out low abundance features.

Metagenome assembly, binning, and annotation

Quality-filtered reads were co-assembled using the default parameters of MetaSPAdes [180]. Assembly quality was accessed using default parameters of MetaQUAST [181], and contigs with less than 500bp were discarded.

Co-assembled scaffolds were binned using MetaBAT2 [182] and MaxBin [183]. For Metabat2, binning of the co-assembled contigs was performed with minimum contigs set to 2000bp, using read mappings performed with bwa-mem2, with the resulting SAM files converted to BAM format and sorted with SAMtools [184]. Default parameters of MaxBin2 were used for the analysis.

Resulting genomic bins from MetaBat2 and Maxbin were further refined using Dastool [185] to obtain more complete genomes with less contamination. Completeness and contamination of the reconstructed MAGs were accessed using CheckM [99] using the standard bacterial marker set. Bins having completeness higher than 70% and contamination lower than 10% were considered as good quality MAGs and used for downstream analysis.

Taxonomic assignments of MAGs were conducted using GTDB-Tk [100] based on the Genome taxonomy database. GTDB-Tk uses a combination of metrics, including average nucleotide identity to reference genomes in the NCBI Assembly database, placement in the GTDB reference tree, and the relative evolutionary divergence. Potential metabolic functions of MAGs were annotated using DRAM. DRAM utilizes a series of databases (UniRef90 [175], PFAM [49], dbCAN [186], Refseq viral (https://www.ncbi.nlm.nih.gov/genome/viruses/), VOGDB (https://vogdb.org) and MEROPS peptidase (https://www.ebi.ac.uk/merops/) to annotate genes and curate these annotations into functional categories.

Results

Sequence data preprocessing

Resulting paired-end sequences from swab samples of 8 wild oysters and 11 farm-raised oysters yielded approximately 56 to 92 million total reads per sample. Removing low-quality reads eliminated approximately 2% of the reads. Mapping to eliminate potential contamination from oyster and human genomes removed a significant proportion of the reads. Approximately 65% to 95% mapped to the oyster genome and 2%- 37% of the reads mapped to human genome. Consequently, decontamination steps reduced the total number of reads used for further downstream analysis to approximately 4 million to 54 million reads per sample.

Table 4.1. Sequence data preprocessing.

Samples	Number of reads after quality filter (bp)	% mapped to oyster genome	Number of reads after oyster genome removal (bp)	% mapped to human genome	Number of reads after human genome removal (bp)
SW1	31957069	92.5%	2164014	22.1%	1684359
SW2	33960232	85.8%	4633615	7.1%	4303219
SW3	44706362	86.9%	5792089	8.8%	5277201
SW4	49046335	39.9%	29627901	5.7%	27925499
SW8	41773795	93.1%	2556106	8.8%	2327027
SW10	47294050	74.8%	11781885	10.7%	10506543
SW11	31186413	95.2%	1219050	11.3%	1079035
SW14	28343593	88.1%	3141970	10.7%	2802634
SW16	38434938	88.3%	4316554	11.7%	3806605
SW18	39768226	83.4%	6459451	1.7%	6337642
SW20	28861260	92.1%	2028630	6.8%	1891227
SW47	34801911	92.6%	2285503	8.3%	2094380
SW48	28256550	91.17%	2286943	5.85%	2151480
SW49	40227838	93.6%	2208195	8.66%	2016355
SW50	35520267	94.6%	1560200	4.57%	1487405
SW51	31435809	85.7%	4367181	7.57%	4034828
SW52	33936181	91.0%	2776693	36.51%	1777518
SW53B	41452803	91.3%	3369921	10.91%	3000637

Taxonomic profiling

The raw taxonomic profile obtained from Bracken2 contained 11,078 OTUs of which 9,063 were affiliated with the domain Bacteria, 1,586 were affiliated with viruses, 427 were affiliated with the domain Archaea, and two were affiliated with Eukarya.

Following removal of low-abundance features after initial data inspection, two samples were removed from the farm environment resulting in 9 farm-raised oyster samples and 8 wild oyster samples. Outlier analysis using oulier_multi [187] revealed that one sample was an outlier (distances exceed 0.2813 from the center 0.4439; dist = 0.79 and SD = 3.64). The other had a considerably lower count/abundance compared to the remaining samples, the sample was 43,258 while others ranged from 235,391 to 55,957,10. The resulting taxonomic profile, following these adjustments consisted of 9,387 OTUs including 8,793 affiliated with the domain Bacteria, 201 affiliated to the domain viruses, 391 affiliated to Archaea, and two affiliated with Eukarya. This filtered profile was used for subsequent taxonomic analysis.

To account for differences in sequencing depth across samples, filtered count data was normalized using rarefaction prior to conducting alpha and beta diversity analyses. Rarefaction resulted in the removal of only 7 OTUs from the count data, leaving a total of 9,380 OTUs (Fig. 4.1 and 4.2).

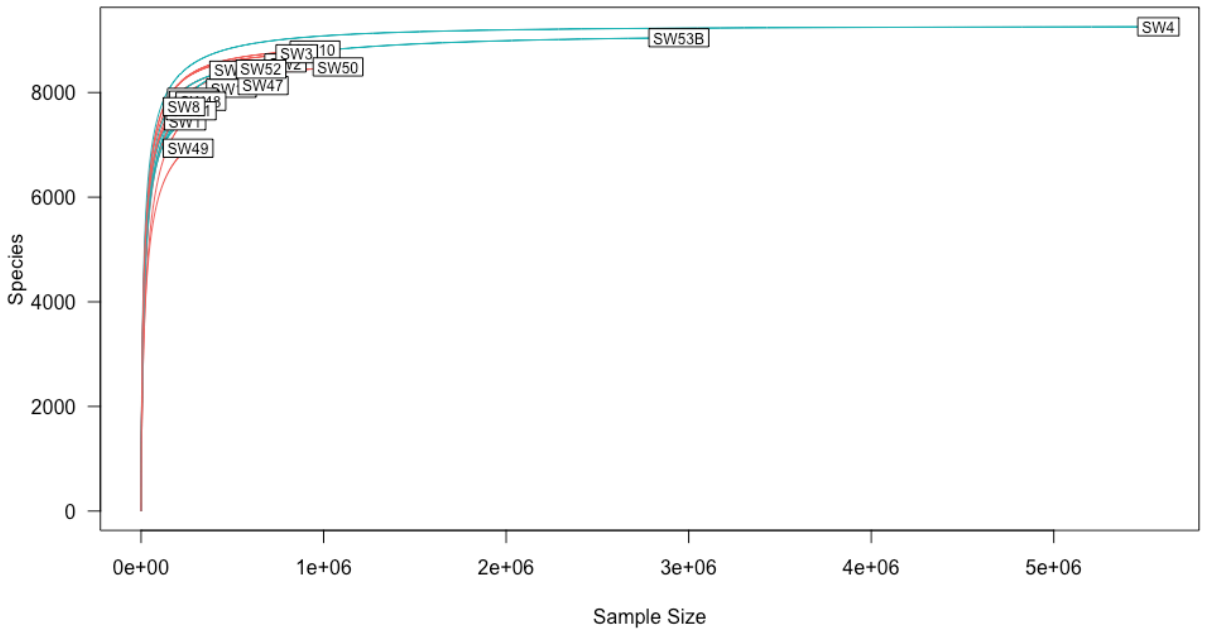

Figure 4.2. Species accumulation curve before

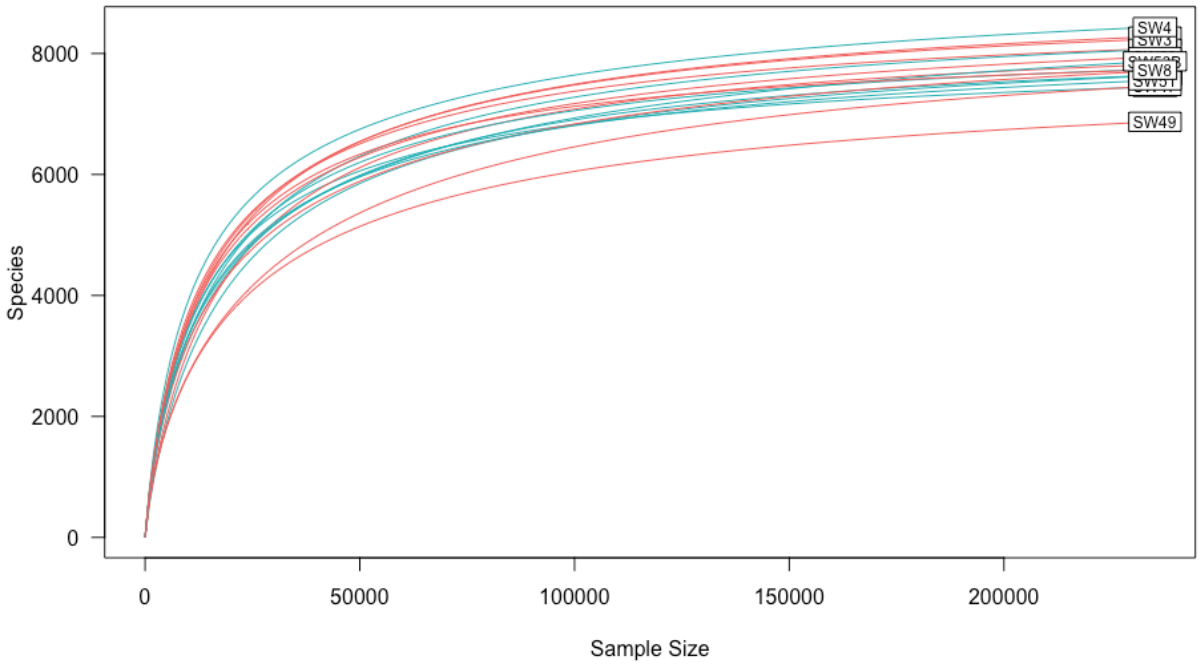

Figure 4.3. Rarefaction curve of all samples.

Alpha diversity was evaluated using the Shannon index and the Chao1 estimator respectively (Fig. 4.4). The Kruskal-Wallis test was used to compare differences in alpha indices between groups. Although both indices were higher in farm-raised oyster samples than wild oyster samples, differences were not statistically significant (Shannon, p-value = 0.74; Chao1, p-value= 0.24).

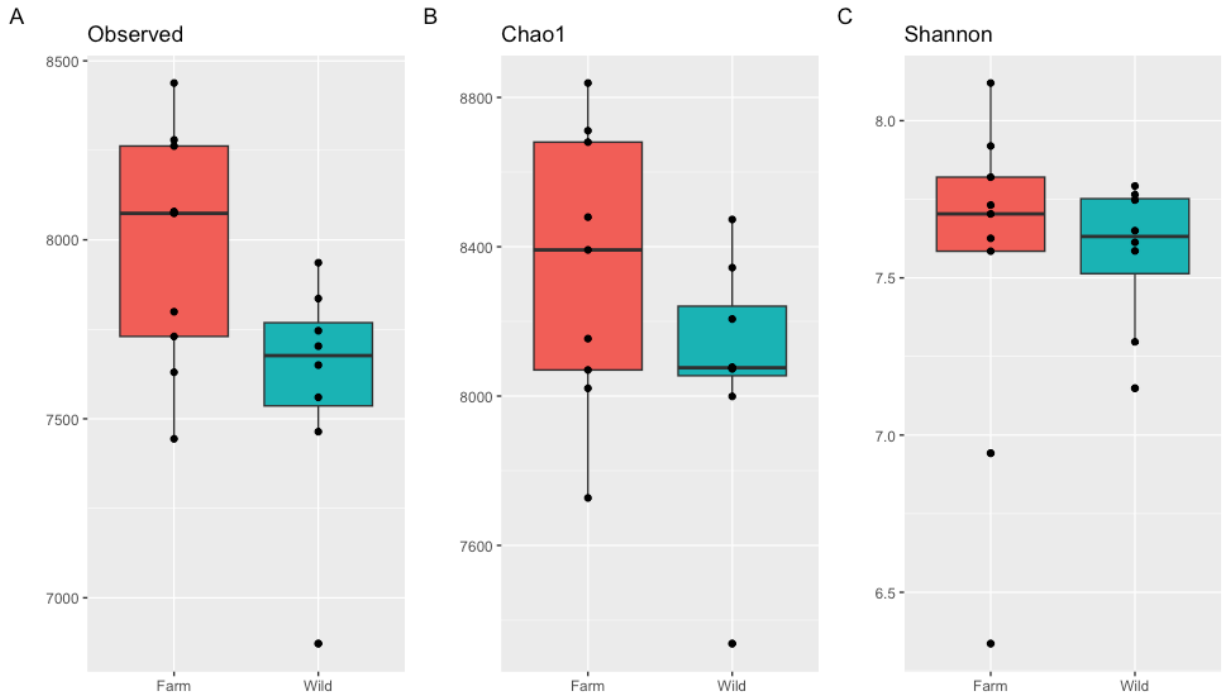

Figure 4.4. Alpha diversity plots (Shannon, p-value = 0.74; Chao1, p-value= 0.24; Observed, p-value = 0.07)

For beta-diversity analysis, the Bray-curtis distance was calculated using vegan, and the resulting matrix was visualized using PCOA (Fig. 4.5). The 2-dimensional PCOA plot showed 45.5% of the total variance between the samples. The first axis accounted for 30.1% of the variation while the second axis accounted for 15.4% of the variation. PERMANOVA revealed that centroids of the groups are not equivalent hence there is significant variation between the 2 groups (p-value = 0.001). Furthermore, PERMDISP results indicated that the dispersion (variation) observed is primarily driven by the differences between the groups rather than within-group variability (p-value = 0.15).

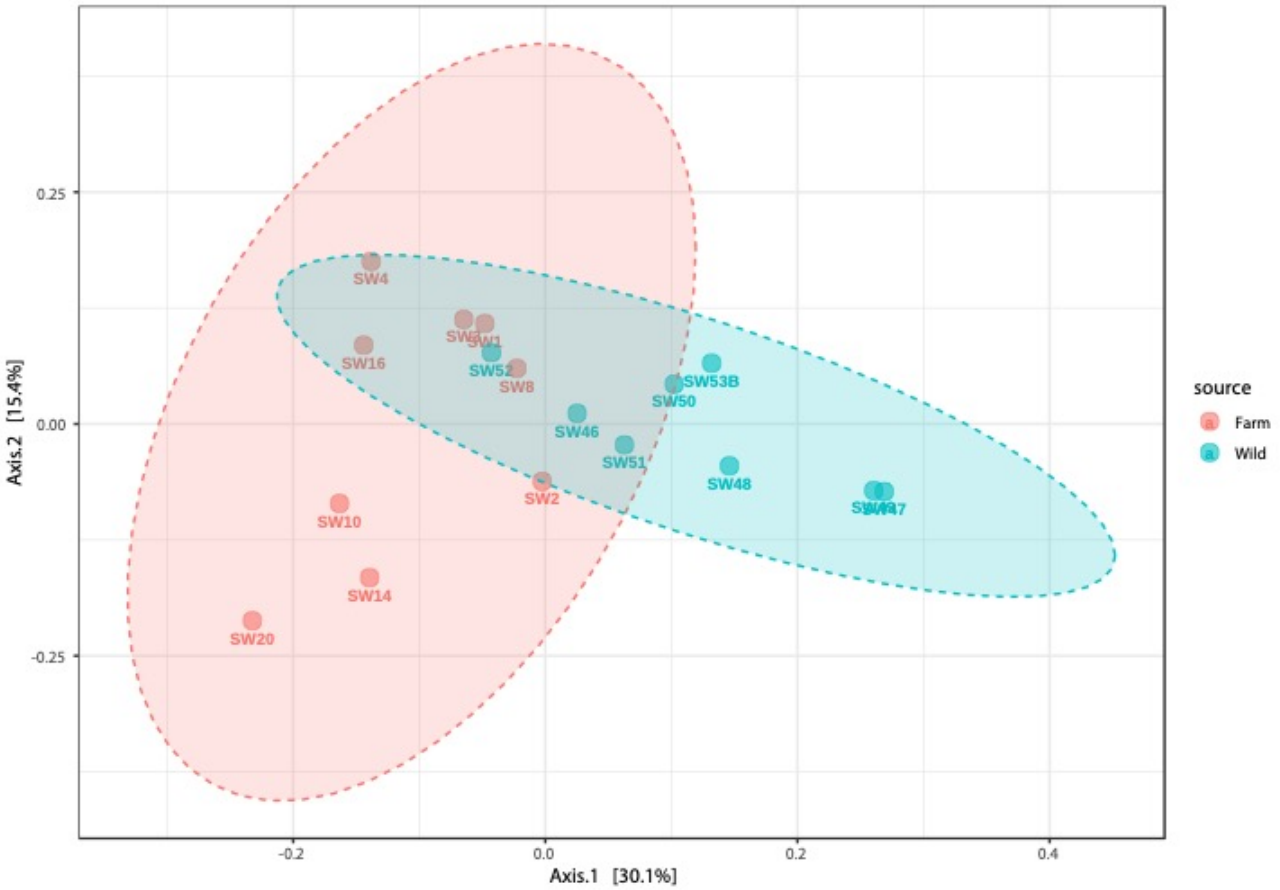

Figure 4.5. Beta diversity PCOA plot (PERMANOVA, p-value = 0.001; PERMDISP, p-value= 0.15)

Relative abundance analysis at the phylum level revealed that *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, *Pseudomonadota*, *Myxococcota*, and *Planctomycetes* were predominant in both environments (Fig. 4.6). At the genus level (Fig. 4.7), the plots showed that *Qipengyuania* and *Leisingera* were more abundant in the wild oyster samples while *Vibrio* and *Tenacibaculum* were

more prevalent in the farm-raised oyster samples. Other genera such as *Pseudomonas*, *Ruegeria*, and *Sphingomonas* were present in samples from both environments while *Flocullibacter*, *Aliiroseovariius*, *Erythrobacter*, *Prosthecochloris*, *Paracoccus*, and *Psychrobacter* were present in only a few samples from either wild or farm-raised oysters.

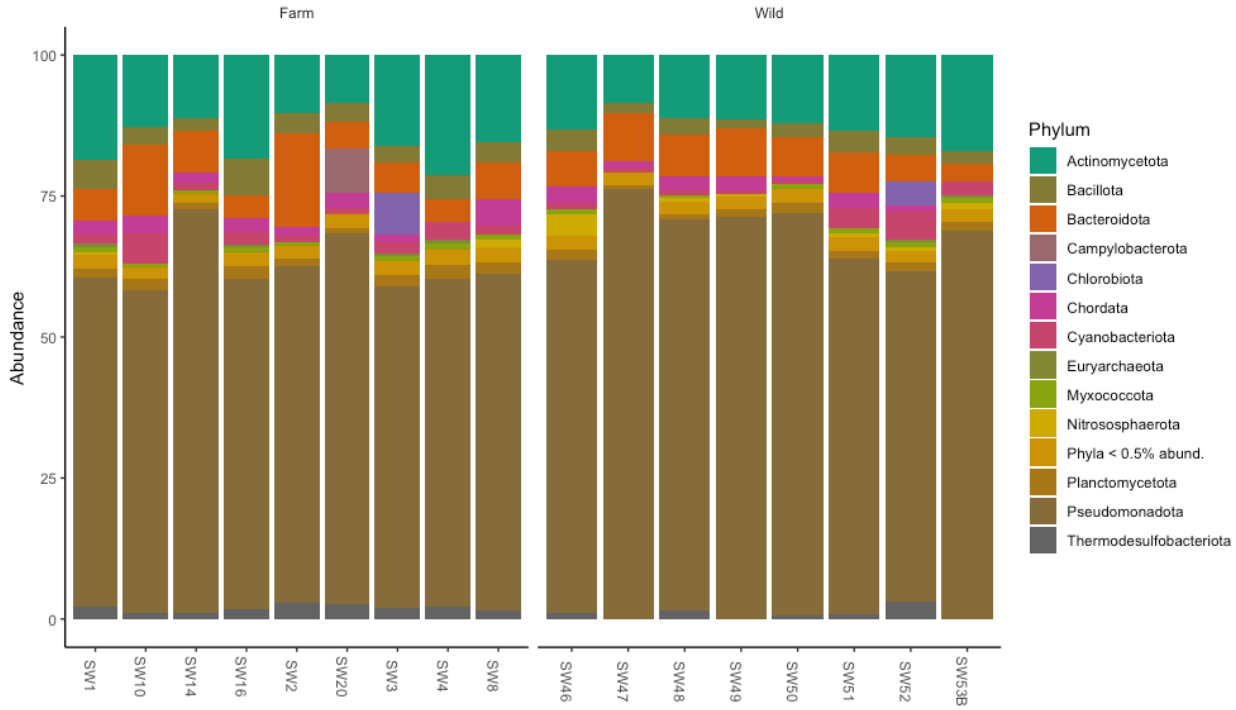

Figure 4.6. Relative abundance at the phylum level

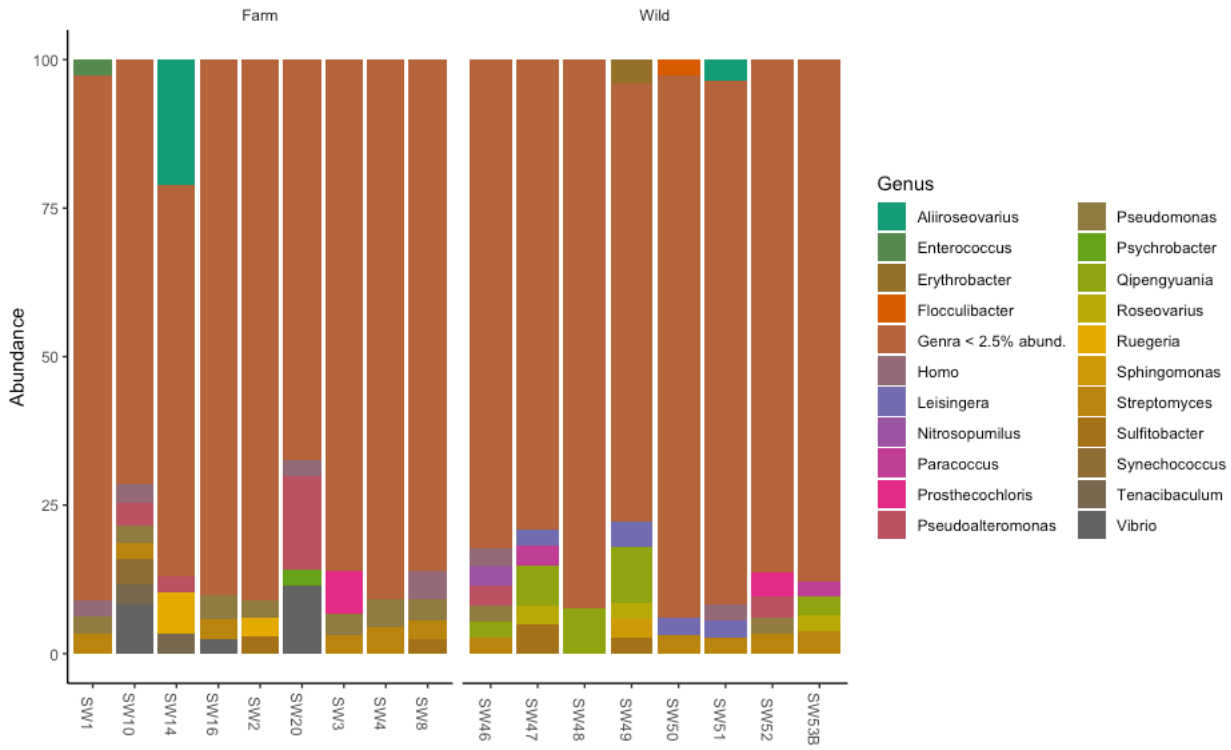

Figure 4.7. Relative abundance at the genus level

To further explore the observed differences, differential abundance analysis was performed using DESeq2 at the species level. A total of 776 taxa, of which bacteria constitute approximately 98%, exhibited significant differences between the two environments. Consistent with the genus-level relative abundance analysis, *Qipengyania* spp. and *Leisingera* spp. were significantly more abundant in the wild oyster samples while *Vibrio* spp. and *Tenacibaculum* spp. were significantly more abundant in the farm-raised oyster samples. Additionally, *Staphylococcus* spp., *Kushneria* spp., *Hyphococcus* spp., and *Spinghomicrobium* spp. showed significant abundance in wild oyster samples whereas *Burkholderia* spp. and *Mycolicibacterium* spp. showed a higher abundance in farm-raised oysters. Interestingly, all significantly abundant viral taxa were present only in the wild oyster samples, including *Dosirivirus* 49B3 and *Galateavirus* PVA5. (Fig. 4.8).

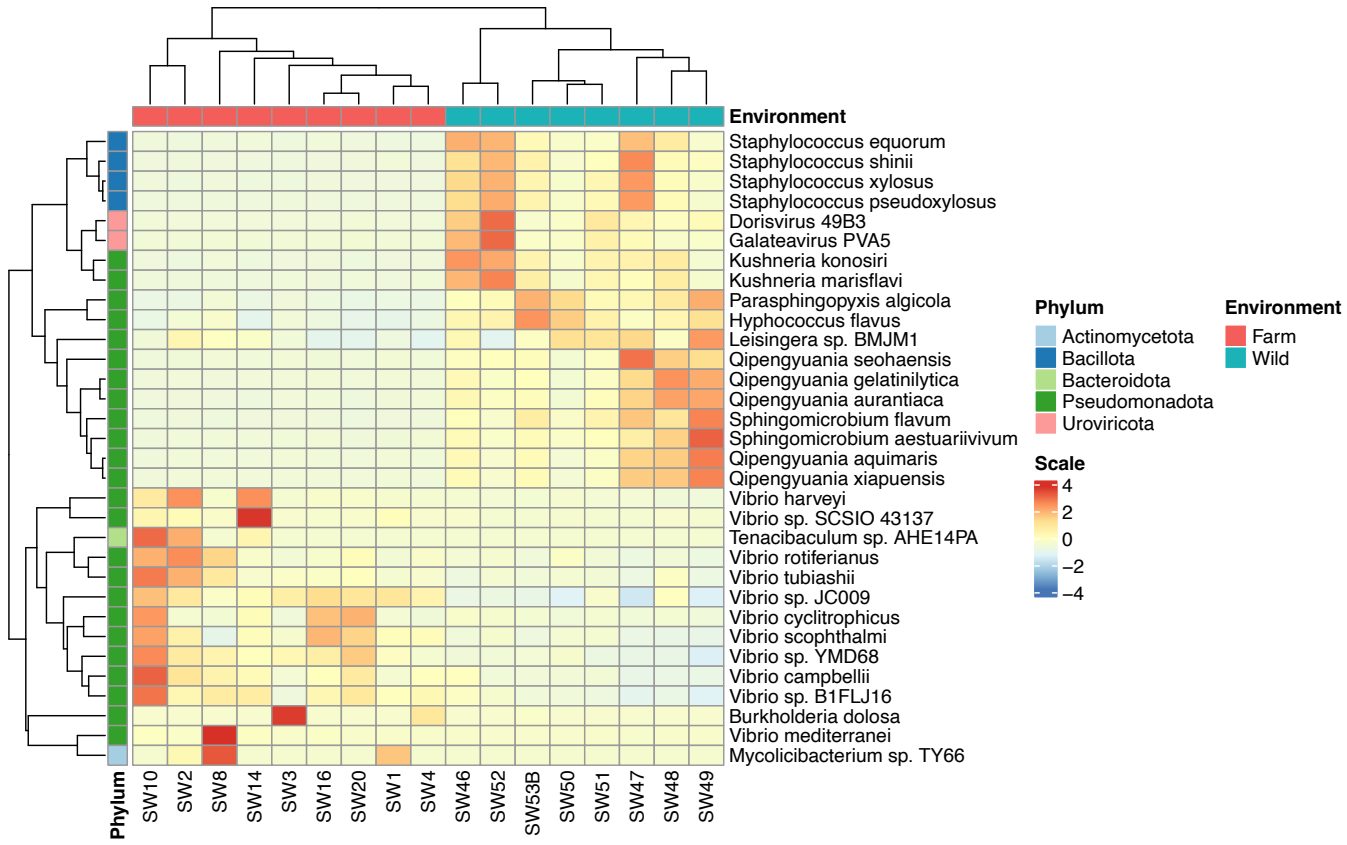

Figure 4.8. Top 20 differentially abundant taxa, including other taxa visibly different in the relative abundance plot.

Functional profiling

A total of 4,752 KEGG ortholog (KO) genes profile and 408 pathway abundance profile (excluding unmapped or ungrouped features) were obtained from HUMAnN3.

Differential abundance analysis using MaAsLin2 revealed significant differential abundance only in two predicted MetaCyc pathways – PWY-6126: superpathway of adenosine nucleotides de novo biosynthesis II and SER-GLYSYN-PWY: superpathway of L-serine and glycine biosynthesis I. These pathways were significantly more abundant in the wild oysters than the farm-raised oysters. The PWY-6126 pathway facilitates the biosynthesis of adenosine nucleotides which are crucial building blocks of DNA and RNA. Additionally, adenosine nucleotides are components of ATP hence PWY-6126 pathway also plays a role in energy production.

The SER-GLYSYN-PWY pathway facilitates biosynthesis of L-serine and glycine. L-serine serves as a precursor for biomolecules such as proteins, nucleotides, and other amino acids. The reaction that converts L-serine to glycine also generates tetrahydrofolate (THF) which is essential for biosynthesis of purines and thymidylate involved in DNA synthesis. The SER-GLYSYN-PWY

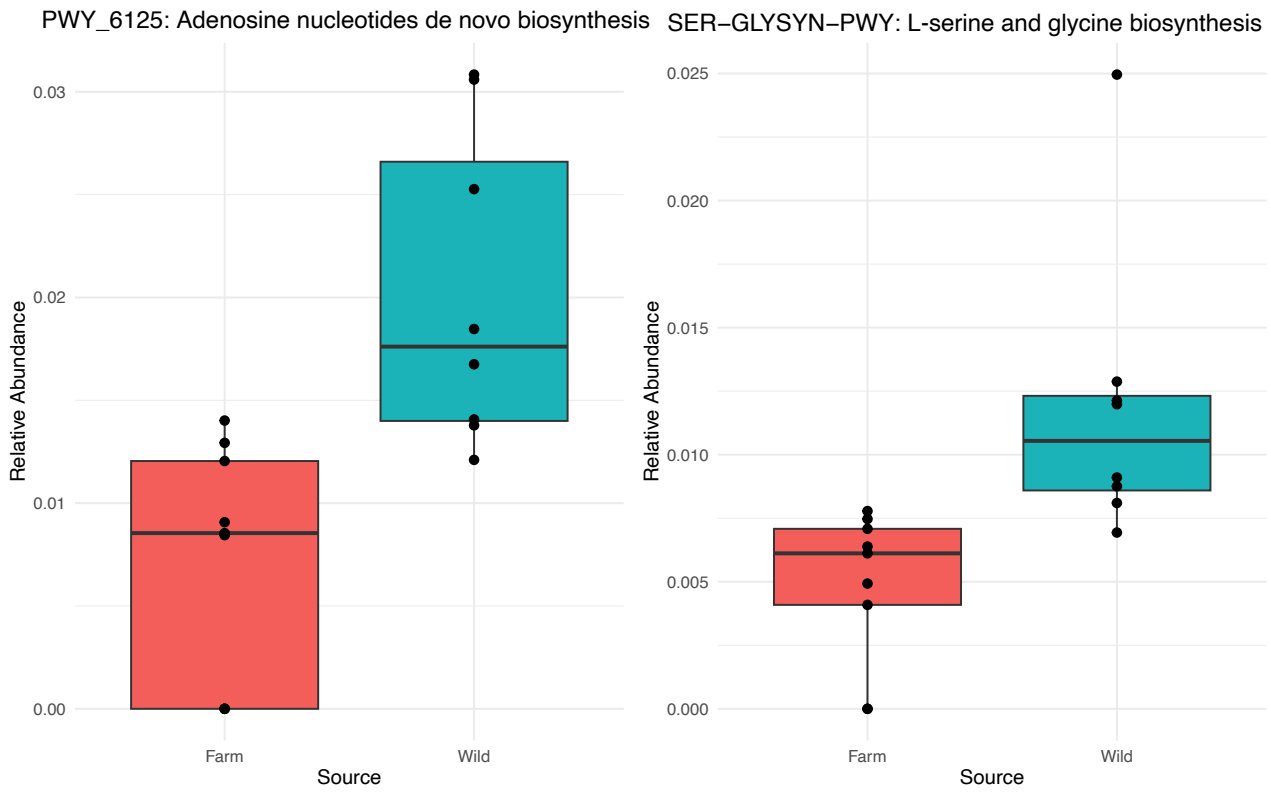

Figure 4.9. Boxplot of differentially abundant MetaCyc pathway.

Further analysis using the stratified HUMAnN3 pathway abundance table aimed to identify major contributors to these pathways in both wild oyster samples and farm-raised oyster samples revealed that most of the species contributing to these pathways could not be classified using the default HUMAnN3 taxonomic database (ChocoPhlAn pangenome database). For the SER-GLYSYN-PWY, the majority of the contributing specie could not be classified, and only one sample containing *Vibrio* spp. showed that *Vibrio* contributed to this pathway albeit in a very low abundance (0.002%) in the wild oyster samples. In contrast, the analysis showed that the major contributing genus in the farm-raised oyster samples included unclassified microbes, *Aliiroseovarius*, *Prosthecochloris*, and *Winogradskyella*. A similar trend was also seen in the PWY-6125 pathway, as most of the contributing species could not be classified in both wild and farm-raised oyster samples. However, *Prosthecochloris* genus was revealed as a contributing species in both environments. Farm-raised oysters also had *Aliiroseovarius* and *Winogradskyella* in addition to the *Prosthecochloris* (Fig. 4.10 and 4.11).

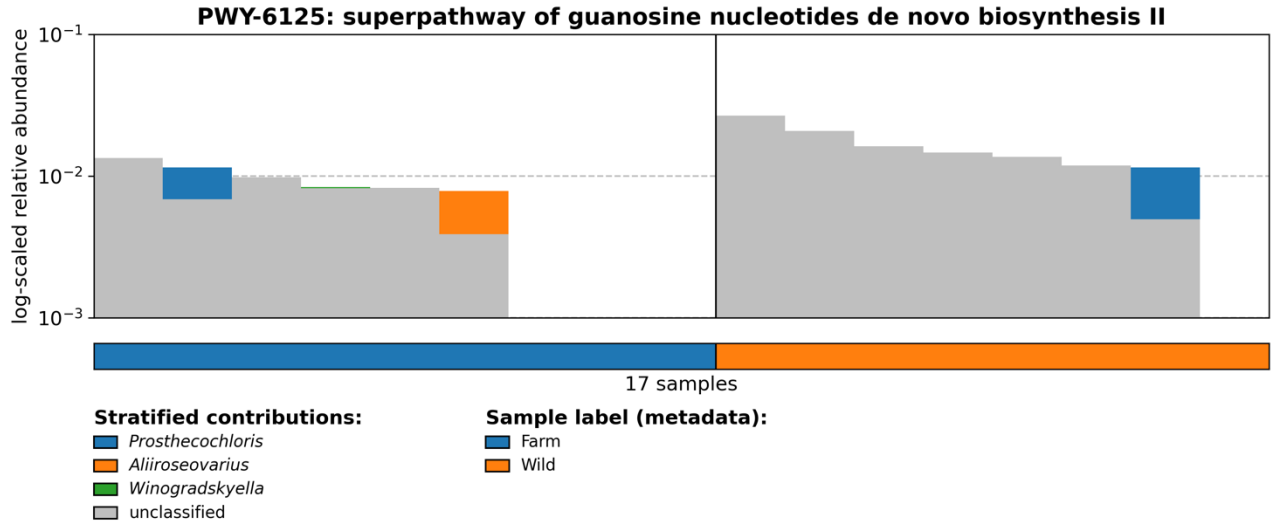

Figure 4.10. PWY-6125 pathway contributing genus plot.

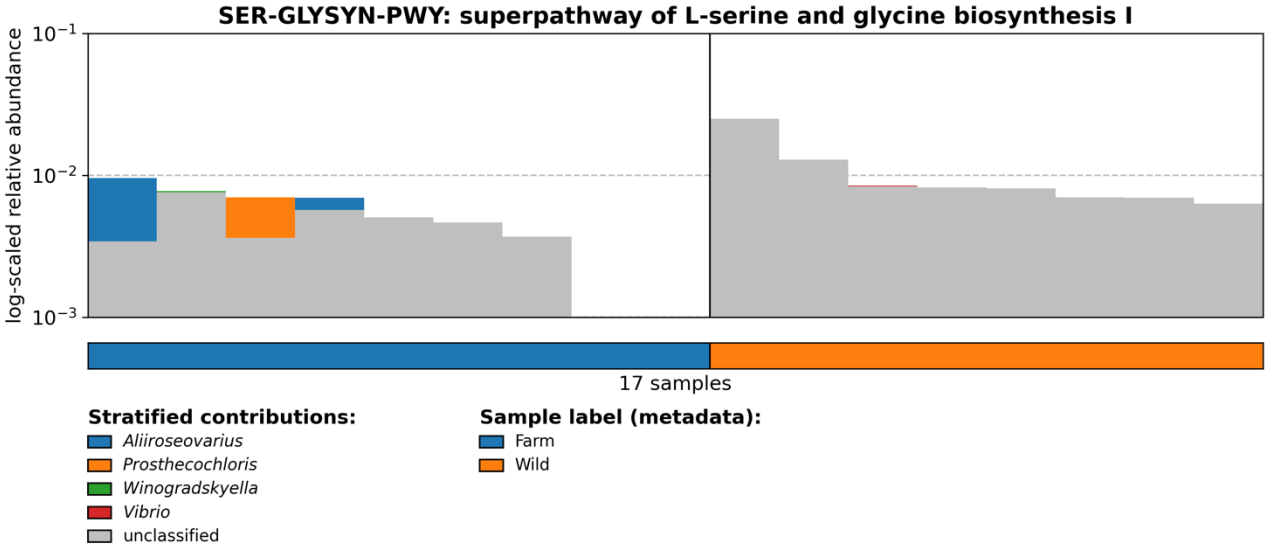

Figure 4.11. SER-GLYSN-PWY pathway contributing genus plot.

Significant differential abundance was observed in 13 KO genes (Fig. 4.12). One of these, peptide deformylase, was more abundant in the farm-raised oyster samples while the rest were more abundant in the wild oyster samples. Genes more abundant in the wild oyster samples include nitrite reductase (NO-forming), site-specific DNA-methyltransferase (adenine-specific), 5-phospho-L-glutamate reductase, 3-oxoacid CoA-transferase subunit B, methane/ammonia monooxygenase

subunit C, methyltransferase, transcription initiation factor TFIIB, small nuclear ribonucleoprotein, putative transcriptional regulator, phosphomethylpyrimidine synthase, ALAS (5-aminolevulinate synthase), and hydroxyacylglutathione hydrolase. These genes are involved in various biosynthesis and metabolic pathways including secondary metabolite biosynthesis, nitrogen metabolism, biosynthesis of co-factors, amino acid metabolism, pyruvate metabolism, thiamine metabolism, arginine biosynthesis, Glycine, serine and threonine metabolism, methane metabolism, butanoate metabolism, and porphyrin metabolism.

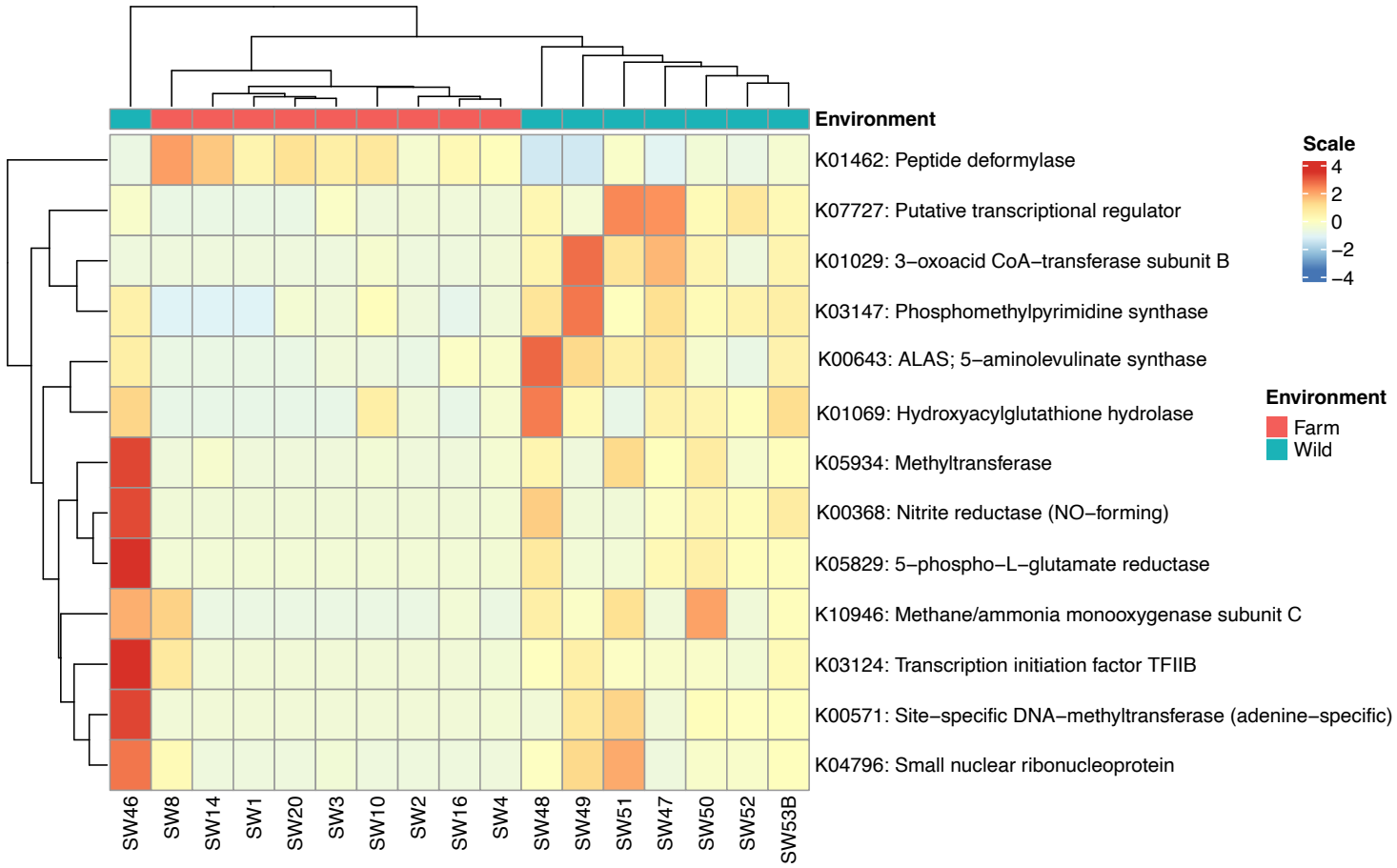

Figure 4.12. Heatmap of differentially abundant KO genes

MAG assembly and binning

A total of 187,763,242 reads that were unmapped to the eastern oyster or human genome were co-assembled using MetaSPAdes resulting in 10,698,461 unique scaffolds. After removing scaffolds less than 500bp, only 1,146,805 unique scaffolds remained. Filtered scaffolds were binned using MaxBin and MetaBat. 117 bins were retrieved from MaxBin while 79 bins were retrieved from

MetaBat2. These bins were further curated using DAStool, resulting in 47 final bins. The bins were accessed using CheckM. A total of 25 bins that were above the defined threshold (completeness higher than 70% and contamination lower than 10%) were considered good quality MAGs and were used for further downstream analysis. GTDB-Tk was used to assign taxonomy to the MAGs. Most MAGs belonged to *Desulfobacterota*, *Bacteriodota*, *Proteobacteria,* and *Spirochaetota* phylum (Table 4.1). However, 16S rRNA was missing in most of the MAGs (22/24), hence the assigned taxonomy couldn't be further verified based on their 16S rRNA gene.

Table 4.2. GTDB-Tk taxonomy classification of Isolated MAGs

ID	GTDB-Tk - Phylum	rRNA present	Closest relative (GenBank)	Percent Identity
MAG 1	Bacteroidota	N		
MAG 2	Desulfobacterota	5S		
MAG 3	Bacteroidota	N		
MAG 4	Desulfobacterota	23S		
MAG 5	Desulfobacterota	N		
MAG 6	Acidobacteriota	5S		
MAG 7	Bacteroidota	N		
MAG 8	Desulfobacterota	N		
MAG 9	Spirochaetota	N		
MAG 10	Proteobacteria - Alphaproteobacteria	N		
MAG 11	Proteobacteria - Gammaproteobacteria	N		
MAG 12	Bacteroidota	5S		
MAG 13	Cyanobacteria	N		
MAG 14	Spirochaetota	16S	*Spirochaetaceae* bacterium	93.7%
MAG 15	Bacteroidota	5S		
MAG 16	Proteobacteria	N		
MAG 17	Thermoproteota	5S		
MAG 18	Desulfobacterota	23S		
MAG 19	Verrucomicrobiota	N		
MAG 20	Bacteroidota	N		
MAG 21	Bacteroidota	N		
MAG 22	Spirochaetota	16S	Uncultured spirochete	98.2%
MAG 23	Bacteroidota	N		
MAG 24	Cyanobacteria	N		
MAG 25	Proteobacteria	5S		

Discussion

Our analysis reveal that the microbial community of wild and farm-raised oysters showed difference in the abundances of microbial community members (> 772 taxa) despite overall similarity in composition. Additionally, functional profiling analysis suggests that microbes present in the wild oysters sampled possess a higher number of genes needed to synthesize essential biomolecules, such as adenosine nucleotides, L-serine, and glycine, which are crucial for their growth and survival.

Taxonomic differential abundance analysis focused on the top 20 differentially abundant taxa and *Vibrio*, a known human pathogen associated with oysters. In the farm-raised oyster samples, we observed a higher abundance of *Vibrio*, *Tenacibaculum*, *Burkholderia*, and *Mycolicibacterium*. In contrast, wild oyster samples displayed higher abundance in genera such as *Qipengyuania*, *Staphylococcus*, *Kushneria*, *Spinghomicrobium,* and phage viruses *Dorisvirus* and *Galateavirus,* isolated from *Vibrio* spp. Although some of these microbes are typically associated with marine environments, they are known opportunistic pathogens of humans.

Sediments have been implicated as reservoirs of oyster-associated *Vibrio* infections [188]. Hence, we expected a higher abundance of *Vibrio* spp. in the wild oysters since they were collected at low tides and were closer to the sediments. However, our results showed a higher abundance of *Vibrio* spp. in the farm-raised oysters. This may be explained by the high abundance of *Vibrio* spp. infecting phage viruses (bacteriophages) in wild-raised oysters, as viral infections of bacterial hosts may lead to bacterial lysis and consequently reduce the bacterial population. Additionally, the three most common *Vibrio* pathogens, *V. parahaemolyticus*, *V. vulnificus*, or *V. cholera,* were not among the differentially abundant *Vibrio* spp. present in farm-raised oyster samples. This suggests that the oyster farming practices did not enhance human pathogens in these oysters at the time of sampling.

Furthermore*, Burkholderia* and *Tenacibaculum* have previously been identified in juvenile and adult oysters [189–191]. These genera are also known to include opportunistic pathogens of humans and fish. *Burkholderia cepacia* is a known pathogen in individuals with compromised immune systems particularly patients with cystic fibrosis [192] while *Tenacibaculum* spp. primarily affect aquatic animals, causing lesions or shell deformation in shellfish and skin ulcers or fin rots in fishes thus impacting the health and market value of the affected animals.

Differential abundance analysis of the pathways present in the microbial communities of the sampled oyster revealed that pathways involved in biosynthesis were more abundant in wild oyster

samples. The PWY-6126 pathway plays a crucial role in the biosynthesis of adenosine nucleotides, which are essential building blocks for DNA and RNA and also important components of ATP while the SER-GLYSYN-PWY pathway is involved in the biosynthesis of L-serine which serves as a precursor for various biomolecules, including proteins, nucleotides, and other amino acids. The abundance of these pathways in wild oysters suggests a potentially higher capacity for energy production and increased capacity for growth and survival due to enhanced protein synthesis and biomolecule production in the organisms present. Further analysis to identify the major contributors to these pathways revealed that most of the species contributing to the higher abundance of SER-GLY-PWY and PWY_6125 were unclassified with the default Chocophlan database. Additionally, the species identified as contributing to these pathways were only found in a few samples in the various environments and were not consistent throughout the samples in the same environment, hence we concluded that the results of this analysis were inconclusive. Efforts to change the database used for the analysis were unsuccessful due to computational limitations.

Gene differential abundance analysis revealed that peptide deformylase gene, which is essential for proper protein maturation and functionality was more abundant in farm-raised oysters while genes involved in various biosynthesis and metabolic pathways including biosynthesis of co-factors, amino acid metabolism, pyruvate metabolism, thiamine metabolism, arginine biosynthesis, glycine, serine and threonine metabolism, were more present in the wild oyster samples. Of noteworthy is the higher abundance of nitrite reductase gene in wild oyster samples. The presence of this denitrification-associated gene suggests the higher occurrence of denitrifying bacteria within the wild oyster samples ecosystem. These denitrifying bacteria are able to convert bioavailable nitrogen to gaseous form, facilitating the removal of nitrogen from their habitat.

Most of the Isolated MAGs lacked 16S rRNA genes, hence the taxonomic assignments by GTDB-Tk could not be further verified. This indicates that these MAGs were of low quality, likely due to a significant loss of reads during the host removal step, making it challenging to assemble whole metagenomes from the remaining reads.

In conclusion, this work explores the microbial diversity and functional capabilities of the microbial communities of farm-raised and wild oysters and contributes to existing knowledge for ensuring the safety of their consumption and understanding the mechanisms underlying their overall fitness. Our analysis amongst other things highlights the presence of opportunistic pathogens in these communities. Therefore, it is imperative that the concentration of these pathogens be monitored prior

to human consumption. Additionally, determining the factors that may impact the concentration of these pathogens prior to human consumption should be a point of future studies.

Chapter V. Conclusion

Marine invertebrates inhabit various marine ecosystems and constitute a vast array of largely unexplored organisms. Within these ecosystems, they assume crucial roles as ecosystem engineers, creating habitats for other organisms, acting as filter feeders to enhance water quality, and serving as indicator species for identifying potential issues like pollution [193] or climate change impacts [194]. Symbiotic associations between marine invertebrates and microbes are very essential for their survival, adaptation, and evolution [1,4]. Several studies have explored these associations resulting in many interesting findings that shed light on the critical role they play in shaping marine biodiversity and ecosystem dynamics. In the face of climate change, the study of these associations has become even more crucial as they can potentially shed light on whether symbiosis will aid marine organisms cope with threats to the biosphere [13]. However, due to the large ecosystem that marine invertebrates inhabit, they are still largely unexplored, hence more studies are needed to gain further insight into these relationships, illuminating their role in evolution, diversity, and survival. In line with this objective, my dissertation aims to make meaningful contributions by presenting a unique perspective on marine invertebrate-microbial symbiosis.

LTR retrotransposons which are transposable elements characterized by long terminal repeats that integrate into a host genome and influence the host evolution, function, and gene regulation, typically serve as a model for the study of retroviruses [16], this is because they are structurally similar and phylogenetically. Endogenous retrovirus provides a perfect example of a symbiotic relationship between virus and its host, a research area that is highly limited in marine invertebrates. Additionally, existing studies on retroelements typically focused on model organisms such as Drosophila *melanogaster* [68], *Caenorhabditis elegans* [70], *Bombyx mori* [71], etc, hence, studies on non-model organisms, particularly marine invertebrates are limited. Furthermore, some of these studies typically focus on these elements in terms of their role in genome composition rather than a detailed assessment of the elements and their evolution.

In **Chapter 2** of my dissertation, I aimed to expand on the limited knowledge of LTR retrotransposons in marine invertebrates by conducting a detailed assessment of LTR-retrotransposons in *Lamellibrachia luymesi*, a non-model deep-sea tubeworm that inhabits hydrocarbon seeps in the Gulf of Mexico. The study provides a robust bioinformatic pipeline, based on our knowledge of LTR retrotransposon structure, to characterize intact LTR retrotransposons in

non-model organisms. I focused on intact LTR-retrotransposon in order to conduct a detailed assessment of these retroelements beyond their role in genome composition. This analysis revealed the presence of a reservoir of novel LTR-retrotransposon families that were different from those in terrestrial species. I also found that some of the retroelements discovered had identical long terminal repeats and further analysis showed that they were recently inserted into the genome, thus raising the possibility of recent or ongoing retrotransposon activity. Furthermore, through phylogenetic analysis, I confirmed the family assignments of the identified retroelements and inferred their evolutionary placement within existing families. This study provides a framework that can be built upon to further explore the function and diversity of LTR-retrotransposons in non-model organisms and to further investigate retrotransposition activities in these organisms.

Furthermore, beyond studying elements that serve as models for understanding symbiotic associations in non-model organisms. I sought to investigate specific symbiotic relationships in marine invertebrates. To this end, in **Chapter 3**, I closely assessed the genome of *Gorgonocephalus chilensis,* a basket star, for the presence of symbiotic microbes. Ultimately, this study revealed the presence of a novel *Mycoplasma* symbiont in the genome of *G. chilenesis*, this symbiont was present in *Gorgonocephalus* samples from both North Pacific and Argentinean waters. Functional annotation of the symbiont revealed a reduced metabolic pathway and broad substrate transport systems, indicating a host-dependent lifestyle. Phylogenetic analysis provided insights into the evolutionary placement of this symbiont, revealing that they belonged to a recently characterized non-free-living lineage of mycoplasmas specifically associated with marine invertebrates. Further analysis of the organisms making up this monophyletic clade showed that they all lacked any known virulence factor, pointing to a commensal symbiotic relationship between this symbiont and its host. This study represents the first step to understanding the biology of this symbiont and lays the foundation for further research on the metabolic interplay between this species and its basket star host, elucidating how they jointly thrive and adapt to their niche.

Moreover, while some symbiotic associations are sometimes unharmful to the host and the microbes, some of these microbes are detrimental to human health when consumed. Hence, the microbiome of marine invertebrates that serve as food to humans is of significant importance. Particularly, oysters are typically associated with various microbes that occur naturally in their environment due to their filter-feeding behavior. However, some of these microbes are known pathogens to humans, hence when eaten raw or undercooked, oysters can serve as vectors of these

pathogens to humans. Several studies have explored oyster-associated microbes [155,157,158]**,** and particular attention has been paid to pathogenic strains of *Vibrio* spp. that are known to cause severe illness or even death when consumed [195]**.** While oysters can be either wild-caught or farm-raised, most existing studies on their microbial composition focus on the farm-raised ones, which are more commonly consumed. Hence, there is a gap in knowledge about the microbial composition of wild oysters.

To bridge this gap**, Chapter 4** of my dissertation sought to explore the microbial communities of wild oysters and how they vary from their farm-raised counterparts. Our analysis employed a metagenomic approach to have a more comprehensive overview of the microbes present (beyond bacterial species) and to overcome the shortcomings of 16S rRNA approach such as low taxonomic resolution and limited functional information. Our analysis revealed that *Vibrio* spp and other opportunistic pathogens such as *Burkholderia* spp. and *Mycolicobacterium* spp. were more abundant in farm-raised oysters. Additionally, we observed a higher abundance of *Vibrio* spp. infecting phage viruses, *Galeavirus* and *Dorisvirus*, in the wild oyster samples, potentially explaining the lower abundance of *Vibrio* spp. in these samples. Functional analysis indicated a potential higher abundance of denitrifying bacteria in the wild oyster samples than in the farm-raised oysters as indicated by the higher abundance of these genes in this environment, illuminating the roles of these microbes in their environment. Our analysis expands upon limited studies and highlights the microbial diversity of wild oysters, laying a framework for ensuring the safety of their consumption. Additionally, it emphasizes the need to monitor the concentration of known pathogens in farm-raised oysters prior to human consumption. Moreover, understanding the dynamics of oyster-associated microbial communities contributes to the broader field of food safety.

In consolidation, these studies significantly expand our understanding of the symbiotic interaction between microbes and the largely unexplored marine invertebrates and lay the foundation for future research in exploring microbial symbiosis in a wider range of marine invertebrates.

REFERENCES

1. Dubilier N, Bergin C, Lott C. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. Nat Rev Microbiol 2008;6(10):725–740; doi: 10.1038/nrmicro1992.
2. Gould AL, Zhang V, Lamberti L, et al. Microbiome interactions shape host fitness. Proceedings of the National Academy of Sciences 2018;115(51); doi: 10.1073/pnas.1809349115.
3. Wilkins LGE, Leray M, O'Dea A, et al. Host-associated microbiomes drive structure and function of marine ecosystems. PLoS Biol 2019;17(11); doi: 10.1371/journal.pbio.3000533.
4. Li J, Zhang Y, Sun J, et al. Editorial: Interaction between Marine Invertebrates and Symbiotic Microbes in a Changing Environment: Community Structure and Ecological Functions. Front Mar Sci 2023;9; doi: 10.3389/fmars.2022.1128906.
5. Krediet CJ, Ritchie KB, Paul VJ, et al. Coral-Associated Micro-Organisms and Their Roles in Promoting Coral Health and Thwarting Diseases. Proceedings of the Royal Society B: Biological Sciences 2013;280(1755); doi: 10.1098/rspb.2012.2328.
6. Webster NS, Thomas T. The sponge hologenome. mBio 2016;7(2); doi: 10.1128/mBio.00135-16.
7. Distel DL, Delong EF, Waterbury2 JB. Phylogenetic Characterization and In Situ Localization of the Bacterial Symbiont of Shipworms (Teredinidae: Bivalvia) by Using 16S RRNA Sequence Analysis and Oligodeoxynucleotide Probe Hybridizationt. 1991.
8. Ruehland C, Blazejak A, Lott C, et al. Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. Environ Microbiol 2008;10(12):3404–3416; doi: 10.1111/J.1462-2920.2008.01728.X.
9. Nyholm S V., McFall-Ngai MJ. The winnowing: establishing the squid–vibrio symbiosis. Nature Reviews Microbiology 2004 2:8 2004;2(8):632–642; doi: 10.1038/nrmicro957.
10. O'brien PA, Webster NS, Miller DJ, et al. Host-Microbe Coevolution: Applying Evidence from Model Systems to Complex Marine Invertebrate Holobionts. 2019; doi: 10.1128/mBio.
11. Li Y, Tassia MG, Waits DS, et al. Genomic adaptations to chemosymbiosis in the deep-sea seep-dwelling tubeworm Lamellibrachia luymesi. BMC Biol 2019;17(1):1–14; doi: 10.1186/s12915-019-0713-x.
12. Sun J, Zhang Y, Xu T, et al. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nat Ecol Evol 2017;1(5):121; doi: 10.1038/s41559-017-0121.
13. Li J, Yang Q, Dong J, et al. Microbiome Engineering: A Promising Approach to Improve Coral Health. Engineering 2022; doi: 10.1016/j.eng.2022.07.010.
14. Kumar A, Bennetzen JL. Plant Retrotransposons. Annu Rev Genet 1999;33(1):479–532; doi: 10.1146/annurev.genet.33.1.479.
15. Casacuberta E, González J. The Impact of Transposable Elements in Environmental Adaptation. Mol Ecol 2013;22(6):1503–1517; doi: 10.1111/mec.12170.
16. Bowen NJ, McDonald JF. Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res 2001;11(9):1527–1540; doi: 10.1101/gr.164201.
17. Xiong Y, Eickbush TH. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol 1988;5(6); doi: 10.1093/oxfordjournals.molbev.a040521.
18. Aroh O, Halanych KM. Genome-wide characterization of LTR retrotransposons in the non-model deep-sea annelid Lamellibrachia luymesi. BMC Genomics 2021;22(1):1–11; doi: 10.1186/S12864-021-07749-1/FIGURES/6.

19. Anonymous. Environmental Benefits of Shellfish Aquaculture. n.d. Available from: https://portal.ct.gov/DOAG/Aquaculture1/Aquaculture/Environmental-Benefits-of-Shellfish-Aquaculture [Last accessed: 12/13/2021].
20. Boeke J, Stoye J. Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. Cold Spring Harbor Laboratory Press; 1997.
21. Smit AF. Interspersed Repeats and Other Mementos of Transposable Elements in Mammalian Genomes. Curr Opin Genet Dev 1999;9(6):657–663; doi: 10.1016/S0959-437X(99)00031-3.
22. SanMiguel P, Gaut BS, Tikhonov A, et al. The paleontology of intergene retrotransposons of maize. Nat Genet 1998;20(1):43–45; doi: 10.1038/1695.
23. Alzohairy AM, Sabir JSM, Gyulai G, et al. Environmental stress activation of plant long-terminal repeat retrotransposons. Functional Plant Biology 2014;41(6):557; doi: 10.1071/FP13339.
24. Muszewska A, Hoffman-Sommer M, Grynberg M. LTR Retrotransposons in Fungi. PLoS One 2011;6(12):29425; doi: 10.1371/journal.pone.0029425.
25. Thomas-Bulle C, Piednoël M, Donnart T, et al. Mollusc genomes reveal variability in patterns of LTR-retrotransposons dynamics. BMC Genomics 2018;19(1):1–18; doi: 10.1186/s12864-018-5200-1.
26. Wicker T, Sabot F, Hua-Van A, et al. A Unified Classification System for Eukaryotic Transposable Elements. 2007.
27. Zhang L, Yan L, Jiang J, et al. The Structure and Retrotransposition Mechanism of LTR-Retrotransposons in the Asexual Yeast Candida Albicans. Virulence 2014;5(6):655–664; doi: 10.4161/viru.32180.
28. Steinbauerová V, Neumann P, Novák P, et al. A widespread occurrence of extra open reading frames in plant Ty3/gypsy retrotransposons. Genetica 2011;139(11–12):1543–1555; doi: 10.1007/s10709-012-9654-9.
29. Neumann P, Novák P, Hoštáková N, et al. Systematic Survey of Plant LTR-Retrotransposons Elucidates Phylogenetic Relationships of Their Polyprotein Domains and Provides a Reference for Element Classification. Mob DNA 2019;10(1):1; doi: 10.1186/s13100-018-0144-1.
30. Carvalho M, Ribeiro T, Viegas W, et al. Presence of env-like sequences in Quercus suber retrotransposons. J Appl Genet 2010;51(4):461–467; doi: 10.1007/BF03208875.
31. Gómez-Orte E, Vicient CM, Martínez-Izquierdo JA. Grande retrotransposons contain an accessory gene in the unusually long 3′-internal region that encodes a nuclear protein transcribed from its own promoter. Plant Mol Biol 2013;81(6):541–551; doi: 10.1007/s11103-013-0019-2.
32. Steckbeck JD, Kuhlmann AS, Montelaro RC. Structural and Functional Comparisons of Retroviral Envelope Protein C-Terminal Domains: Still Much to Learn. Viruses 2014;6(1):284–300; doi: 10.3390/v6010284.
33. Mclane LM, Pulliam KF, Devine SE, et al. The Ty1 integrase protein can exploit the classical nuclear protein import machinery for entry into the nucleus. Nucleic Acids Res 2008;36(13):4317–4326; doi: 10.1093/nar/gkn383.
34. Vicient CM, Casacuberta JM. Additional ORFs in Plant LTR-Retrotransposons. Front Plant Sci 2020;11:555; doi: 10.3389/fpls.2020.00555.
35. Gonzalez P, Lessios HA. Evolution of sea urchin retroviral-like (SURL) elements: Evidence from 40 echinoid species. Mol Biol Evol 1999;16(7):938–952; doi: 10.1093/oxfordjournals.molbev.a026183.

36. Chen JE, Cui G, Wang X, et al. Recent expansion of heat-activated retrotransposons in the coral symbiont Symbiodinium microadriaticum. ISME Journal 2018;12(2):639–643; doi: 10.1038/ismej.2017.179.
37. Piednoël M, Donnart T, Esnault C, et al. LTR-Retrotransposons in R. exoculata and Other Crustaceans: The Outstanding Success of GalEa-Like Copia Elements. Kashkush K. ed. PLoS One 2013;8(3):e57675; doi: 10.1371/journal.pone.0057675.
38. Wang K, Shen Y, Yang Y, et al. Morphology and genome of a snailfish from the Mariana Trench provide insights into deep-sea adaptation. Nat Ecol Evol 2019;3(5):823–833; doi: 10.1038/s41559-019-0864-8.
39. Schulze A. Phylogeny of Vestimentifera (Siboglinidae, Annelida) inferred from morphology. Zool Scr 2003;32(4):321–342; doi: 10.1046/j.1463-6409.2003.00119.x.
40. Schulze A, Halanych KM. Siboglinid Evolution Shaped by Habitat Preference and Sulfide Tolerance. In: Hydrobiologia Springer; 2003; pp. 199–205; doi: 10.1023/A:1026192715095.
41. Halanych KM. Molecular Phylogeny of Siboglinid Annelids (a.k.a. Pogonophorans): A Review. In: Morphology, Molecules, Evolution and Phylogeny in Polychaeta and Related Taxa Springer-Verlag; 2005; pp. 297–307; doi: 10.1007/1-4020-3240-4_16.
42. Anonymous. RepeatModeler Download Page. n.d. Available from: http://www.repeatmasker.org/RepeatModeler/ [Last accessed: 2/19/2021].
43. Anonymous. RepeatMasker Home Page. n.d. Available from: http://www.repeatmasker.org/ [Last accessed: 2/19/2021].
44. Anonymous. LLUY_1.0 - Genome - Assembly - NCBI. n.d. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCA_009193005.1 [Last accessed: 4/13/2020].
45. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 2008;9(1):18; doi: 10.1186/1471-2105-9-18.
46. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007;35(Web Server issue):W265-8; doi: 10.1093/nar/gkm286.
47. Lerat E. Identifying Repeats and Transposable Elements in Sequenced Genomes: How to Find Your Way through the Dense Forest of Programs. Heredity (Edinb) 2010;104(6):520–533; doi: 10.1038/hdy.2009.165.
48. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol 2018;176(2):1410–1422; doi: 10.1104/pp.17.01310.
49. Finn RD, Bateman A, Clements J, et al. Pfam: The protein families database. Nucleic Acids Res 2014;42(D1):222–230; doi: 10.1093/nar/gkt1223.
50. Zhang R-G, Wang Z-X, Ou S, et al. TEsorter: lineage-level classification of transposable elements using conserved protein domains. bioRxiv 2019;800177; doi: 10.1101/800177.
51. Llorens C, Futami R, Covelli L, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. n.d.; doi: 10.1093/nar/gkq1061.
52. Vershinin A V., Ellis THN. Heterogeneity of the internal structure of PDR1, a family of Ty1/copia-like retrotransposons in pea. Molecular and General Genetics 1999;262(4–5):703–713; doi: 10.1007/s004380051132.
53. Neogi U, Engelbrecht S, Claassen M, et al. Mutational Heterogeneity in p6 Gag Late Assembly (L) Domains in HIV-1 Subtype C Viruses from South Africa. AIDS Res Hum Retroviruses 2016;32(1):80–84; doi: 10.1089/aid.2015.0266.

54. Metzger MJ, Paynter AN, Siddall ME, et al. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. Proc Natl Acad Sci U S A 2018;115(18):E4227–E4235; doi: 10.1073/pnas.1717227115.
55. De La Chaux N, Wagner A. BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol 2011;11(1):154; doi: 10.1186/1471-2148-11-154.
56. Cao L, Yin G, Cao Z, et al. Identification and characterization of a LTR retrotransposon from the genome of Cyprinus carpio var. Jian. Genetica 2016;144(3):325–333; doi: 10.1007/s10709-016-9901-6.
57. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 2013;30(4):772; doi: 10.1093/MOLBEV/MST010.
58. Nguyen LT, Schmidt HA, Von Haeseler A, et al. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 2015;32(1):268; doi: 10.1093/MOLBEV/MSU300.
59. Rambaut A. Figtree v1.4.4. 2018.
60. Jukes TH, Cantor CR. Evolution of protein molecules. 1969.
61. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A 2004;101(34):12404–12410; doi: 10.1073/pnas.0403715101.
62. Marchler-Bauer A, Bryant SH. CD-Search: Protein domain annotations on the fly. Nucleic Acids Res 2004;32(WEB SERVER ISS.):327–331; doi: 10.1093/nar/gkh454.
63. Bae YA, Moon SY, Kong Y, et al. CsRn1, a novel active retrotransposon in a parasitic trematode, Clonorchis sinensis, discloses a new phylogenetic clade of Ty3/gypsy-like LTR retrotransposons. Mol Biol Evol 2001;18(8):1474–1483; doi: 10.1093/oxfordjournals.molbev.a003933.
64. Goodwin T, Poulter R. A group of deuterostome Ty3/gypsy-like retrotransposons with Ty1/copia-like pol-domain orders. Molecular Genetics and Genomics 2002;267(4):481–491; doi: 10.1007/s00438-002-0679-0.
65. Volff J-N, Körting C, Altschmied J, et al. Jule from the Fish Xiphophorus Is the First Complete Vertebrate Ty3/Gypsy Retrotransposon from the Mag Family. Mol Biol Evol 2001;18(2):101–111; doi: 10.1093/oxfordjournals.molbev.a003784.
66. Tubío JMC, Naveira H, Costas J. Structural and Evolutionary Analyses of the Ty3/gypsy Group of LTR Retrotransposons in the Genome of Anopheles gambiae. Mol Biol Evol 2005;22(1):29–39; doi: 10.1093/molbev/msh251.
67. McCarthy EM, Liu J, Lizhi G, et al. Long terminal repeat retrotransposons of Oryza sativa. Genome Biol 2002;3(10):research0053.1; doi: 10.1186/gb-2002-3-10-research0053.
68. Kaminker JS, Bergman CM, Kronmiller B, et al. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol 2002;3(12):research0084.1; doi: 10.1186/gb-2002-3-12-research0084.
69. Jin-Shan X, Qing-You X, Jun L, et al. Survey of long terminal repeat retrotransposons of domesticated silkworm (Bombyx mori). Insect Biochem Mol Biol 2005;35(8):921–929; doi: 10.1016/j.ibmb.2005.03.014.
70. Bowen NJ, McDonald JF. Genomic analysis of Caenorhabditis elegans reveals ancient families of retroviral-like elements. Genome Res 1999;9(10):924–935; doi: 10.1101/gr.9.10.924.

71. Michaille JJ, Mathavan S, Gaillard J, et al. The Complete Sequence of Mag, a New Retrotransposon in Bombyx Mori. Nucleic Acids Res 1990;18(3):674; doi: 10.1093/nar/18.3.674.
72. Springer MS, Davidson EH, Britten RJ. Retroviral-like Element in a Marine Invertebrate (Retrotransposon/Retrovirus/Long Terminal Repeat/Echinoderm/Mobile Elements). 1991.
73. Butler M, Goodwin T, Poulter R. An Unusual Vertebrate LTR Retrotransposon from the Cod Gadus morhua. Mol Biol Evol 2001;18(3):443–447; doi: 10.1093/oxfordjournals.molbev.a003822.
74. Simmen MW, Bird A. Sequence Analysis of Transposable Elements in the Sea Squirt, Ciona intestinalis. Mol Biol Evol 2000;17(11):1685–1694; doi: 10.1093/oxfordjournals.molbev.a026267.
75. Abe H, Ohbayashi F, Shimada T, et al. Molecular structure of a novel gypsy-ty3-like retrotransposon (Kabuki) and nested retrotransposable elements on the W chromosome of the silkworm Bombyx mori. Molecular and General Genetics 2000;263(6):916–924; doi: 10.1007/s004380000270.
76. Copeland CS, Brindley PJ, Heyers O, et al. Boudicca, a Retrovirus-Like Long Terminal Repeat Retrotransposon from the Genome of the Human Blood Fluke Schistosoma mansoni. J Virol 2003;77(11):6153–6166; doi: 10.1128/jvi.77.11.6153-6166.2003.
77. Terrat Y, Bonnivard E, Higuet D. GalEa retrotransposons from galatheid squat lobsters (Decapoda, Anomura) define a new clade of Ty1/copia-like elements restricted to aquatic species. Molecular Genetics and Genomics 2008;279(1):63–73; doi: 10.1007/s00438-007-0295-0.
78. Copeland CS, Mann VH, Morales ME, et al. The Sinbad retrotransposon from the genome of the human blood fluke, Schistosoma mansoni, and the distribution of related Pao-like elements. BMC Evol Biol 2005;5(1):20; doi: 10.1186/1471-2148-5-20.
79. Wicker T, Keller B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 2007;17(7):1072–1081; doi: 10.1101/gr.6214107.
80. Du C, Swigoňová Z, Messing J. Retrotranspositions in orthologous regions of closely related grass species. BMC Evol Biol 2006;6(1):62; doi: 10.1186/1471-2148-6-62.
81. Chénais B, Caruso A, Hiard S, et al. The Impact of Transposable Elements on Eukaryotic Genomes: From Genome Size Increase to Genetic Adaptation to Stressful Environments. Gene 2012;509(1):7–15; doi: 10.1016/j.gene.2012.07.042.
82. Halanych KM, Bacheller JD, Aguinaldo AMA, et al. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science (1979) 1995;267(5204):1641–1643; doi: 10.1126/science.7886451.
83. Razin S, Yogev D, Naot Y. Molecular Biology and Pathogenicity of Mycoplasmas. Microbiology and Molecular Biology Reviews 1998;62(4):1094; doi: 10.1128/mmbr.62.4.1094-1156.1998.
84. Rasmussen JA, Villumsen KR, Duchêne DA, et al. Genome-resolved metagenomics suggests a mutualistic relationship between Mycoplasma and salmonid hosts. Communications Biology 2021 4:1 2021;4(1):1–10; doi: 10.1038/s42003-021-02105-1.
85. Sellyei B, Varga Z, Cech G, et al. Mycoplasma infections in freshwater carnivorous fishes in Hungary. J Fish Dis 2021;44(3):297–304; doi: 10.1111/JFD.13283.

86. TATTAR TA. Mycoplasmas. Diseases of Shade Trees 1989;57–67; doi: 10.1016/B978-0-12-684351-4.50012-0.
87. Bano N, DeRae Smith A, Bennett W, et al. Dominance of Mycoplasma in the guts of the Long-Jawed Mudsucker, Gillichthys mirabilis, from five California salt marshes. Environ Microbiol 2007;9(10):2636–2641; doi: 10.1111/J.1462-2920.2007.01381.X.
88. Holben WE, Williams P, Saarinen M, et al. Phylogenetic Analysis of Intestinal Micro¯ora Indicates a Novel Mycoplasma Phylotype in Farmed and Wild Salmon. n.d.; doi: 10.1007/s00248-002-1011-6.
89. Neulinger SC, Gärtner A, Järnegren J, et al. Tissue-associated "Candidatus mycoplasma corallicola" and filamentous bacteria on the cold-water coral Lophelia pertusa (Scleractinia). Appl Environ Microbiol 2009;75(5):1437–1444; doi: 10.1128/AEM.01781-08/SUPPL_FILE/AEM___TISSUE_ASSOCIATED_BACTERIA_ON_LOPHELIA_PERTUSA___SUPPLEMENTAL_FILE.PDF.
90. Meziti A, Ramette A, Mente E, et al. Temporal shifts of the Norway lobster (Nephrops norvegicus) gut bacterial communities. FEMS Microbiol Ecol 2010;74(2):472–484; doi: 10.1111/J.1574-6941.2010.00964.X.
91. Iehata S, Valenzuela F, Riquelme C. Analysis of bacterial community and bacterial nutritional enzyme activity associated with the digestive tract of wild Chilean octopus (Octopus mimus Gould, 1852). Aquac Res 2015;46(4):861–873; doi: 10.1111/ARE.12240.
92. Ramírez AS, Vega-Orellana OM, Viver T, et al. First description of two moderately halophilic and psychrotolerant Mycoplasma species isolated from cephalopods and proposal of Mycoplasma marinum sp. nov. and Mycoplasma todarodis sp. nov. Syst Appl Microbiol 2019;42(4):457–467; doi: 10.1016/J.SYAPM.2019.04.003.
93. Huang Z Bin, Guo F, Zhao J, et al. Molecular analysis of the intestinal bacterial flora in cage-cultured adult small abalone, Haliotis diversicolor. Aquac Res 2010;41(11):e760–e769; doi: 10.1111/J.1365-2109.2010.02577.X.
94. Tanaka R, Ootsubo M, Sawabe T, et al. Biodiversity and in situ abundance of gut microflora of abalone (Haliotis discus hannai) determined by culture-independent techniques. Aquaculture 2004;241(1–4):453–463; doi: 10.1016/J.AQUACULTURE.2004.08.032.
95. Nakagawa S, Saito H, Tame A, et al. Microbiota in the coelomic fluid of two common coastal starfish species and characterization of an abundant Helicobacter-related taxon. Scientific Reports 2017 7:1 2017;7(1):1–10; doi: 10.1038/s41598-017-09355-2.
96. Cheng H, Concepcion GT, Feng X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 2021;18(2):170–175; doi: 10.1038/s41592-020-01056-5.
97. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2022;50(D1):D20; doi: 10.1093/NAR/GKAB1112.
98. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20(1); doi: 10.1186/S13059-019-1891-0.
99. Parks DH, Imelfort M, Skennerton CT, et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;25(7):1043; doi: 10.1101/GR.186072.114.
100. Chaumeil PA, Mussig AJ, Hugenholtz P, et al. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 2020;36(6):1925–1927; doi: 10.1093/BIOINFORMATICS/BTZ848.

101. Parks DH, Chuvochina M, Rinke C, et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res 2022;50(D1):D785–D794; doi: 10.1093/NAR/GKAB776.
102. Aziz RK, Bartels D, Best A, et al. The RAST Server: Rapid annotations using subsystems technology. BMC Genomics 2008;9(1):1–15; doi: 10.1186/1471-2164-9-75/TABLES/3.
103. Lagesen K, Hallin P, Rødland EA, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;35(9):3100; doi: 10.1093/NAR/GKM160.
104. Kanehisa M, Furumichi M, Sato Y, et al. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 2021;49(D1):D545; doi: 10.1093/NAR/GKAA970.
105. Gao F, Zhang CT. Ori-Finder: A web-based system for finding oriCs in unannotated bacterial genomes. BMC Bioinformatics 2008;9(1):1–6; doi: 10.1186/1471-2105-9-79/FIGURES/2.
106. Luo H, Quan CL, Peng C, et al. Recent development of Ori-Finder system and DoriC database for microbial replication origins. Brief Bioinform 2019;20(4):1114–1124; doi: 10.1093/BIB/BBX174.
107. Cantalapiedra CP, Hern Andez-Plaza A, Letunic I, et al. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 2021;38(12):5825–5829; doi: 10.1093/MOLBEV/MSAB293.
108. Chen L, Yang J, Yu J, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 2005;33(Database Issue):D325; doi: 10.1093/NAR/GKI008.
109. Couvin D, Bernheim A, Toffano-Nioche C, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res 2018;46(W1):W246–W251; doi: 10.1093/NAR/GKY425.
110. Yoon SH, Ha S min, Lim J, et al. A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology 2017;110(10):1281–1286; doi: 10.1007/S10482-017-0844-4/FIGURES/3.
111. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. 2016; doi: 10.7287/peerj.preprints.1900v1.
112. Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun 2019;10(1); doi: 10.1038/S41467-019-10210-3.
113. Meier-Kolthoff JP, Auch AF, Klenk HP, et al. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 2013;14(1):1–14; doi: 10.1186/1471-2105-14-60/TABLES/2.
114. Van Kuppeveld FJM, Van der Logt JTM, Angulo AF, et al. Genus- and species-specific identification of mycoplasmas by 16S rRNA amplification. Appl Environ Microbiol 1992;58(8):2606; doi: 10.1128/aem.58.8.2606-2615.1992.
115. Kearse M, Moir R, Wilson A, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 2012;28(12):1647–1649; doi: 10.1093/BIOINFORMATICS/BTS199.
116. Yoon SH, Ha SM, Kwon S, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 2017;67(5):1613; doi: 10.1099/IJSEM.0.001755.
117. Sayers EW, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Res 2020;48(D1):D84–D86; doi: 10.1093/NAR/GKZ956.

118. Weisburg WG, Tully JG, Rose DL, et al. A Phylogenetic Analysis of the Mycoplasmas: Basis for Their Classification. Public Health Resources 1989;9.
119. Wolf M, Müller T, Dandekar T, et al. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. Int J Syst Evol Microbiol 2004;54(3):871–875; doi: 10.1099/IJS.0.02868-0.
120. Mayor D, Jores J, Korczak BM, et al. Multilocus sequence typing (MLST) of Mycoplasma hyopneumoniae: A diverse pathogen with limited clonality. Vet Microbiol 2008;127(1–2):63–72; doi: 10.1016/j.vetmic.2007.08.010.
121. Manso-Silván L, Dupuy V, Lysnyansky I, et al. Phylogeny and molecular typing of Mycoplasma agalactiae and Mycoplasma bovis by multilocus sequencing. Vet Microbiol 2012;161(1–2):104–112; doi: 10.1016/j.vetmic.2012.07.015.
122. Tocqueville V, Ferré S, Nguyen NHP, et al. Multilocus sequence typing of Mycoplasma hyorhinis strains identified by a real-time TaqMan PCR assay. J Clin Microbiol 2014;52(5):1664–1671; doi: 10.1128/JCM.03437-13.
123. Dijkman R, Feberwee A, Landman WJM. Development and evaluation of a multi-locus sequence typing scheme for Mycoplasma synoviae. Avian Pathology 2016;45(4):426–442; doi: 10.1080/03079457.2016.1154135.
124. Bekő K, Kreizinger Z, Sulyok KM, et al. Genotyping Mycoplasma gallisepticum by multilocus sequence typing. Vet Microbiol 2019;231(January):191–196; doi: 10.1016/j.vetmic.2019.03.016.
125. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003;19(12):1572–1574; doi: 10.1093/BIOINFORMATICS/BTG180.
126. Darriba D, Taboada GL, Doallo R, et al. jModelTest 2: more models, new heuristics and high-performance computing. Nat Methods 2012;9(8):772; doi: 10.1038/NMETH.2109.
127. Fujikawa N, Kurumizaka H, Nureki O, et al. Structural basis of replication origin recognition by the DnaA protein. Nucleic Acids Res 2003;31(8):2077–2086; doi: 10.1093/NAR/GKG309.
128. Cordova CMM, Lartigue C, Sirand-Pugnet P, et al. Identification of the Origin of Replication of the Mycoplasma pulmonis Chromosome and Its Use in oriC Replicative Plasmids. J Bacteriol 2002;184(19):5426; doi: 10.1128/JB.184.19.5426-5435.2002.
129. Chopra-Dewasthaly R, Marenda M, Rosengarten R, et al. Construction of the first shuttle vectors for gene cloning and homologous recombination in Mycoplasma agalactiae. FEMS Microbiol Lett 2005;253(1):89–94; doi: 10.1016/J.FEMSLE.2005.09.021.
130. Papazisi L, Gorton TS, Kutish G, et al. The complete genome sequence of the avian pathogen Mycoplasma gallisepticum strain R low. n.d.; doi: 10.1099/mic.0.26427-0.
131. Messer W. The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication. FEMS Microbiol Rev 2002;26(4):355–374; doi: 10.1111/J.1574-6976.2002.TB00620.X.
132. Mackiewicz P, Zakrzewska-Czerwińska J, Zawilak A, et al. Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Res 2004;32(13):3781–3791; doi: 10.1093/NAR/GKH699.
133. Borukhov S, Lee J, Laptenko O. Bacterial transcription elongation factors: new insights into molecular mechanism of action. Mol Microbiol 2005;55(5):1315–1324; doi: 10.1111/J.1365-2958.2004.04481.X.
134. Ha S-M, Kim CK, Roh J, et al. Application of the Whole Genome-Based Bacterial Identification System, TrueBac ID, Using Clinical Isolates That Were Not Identified With Three Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry

(MALDI-TOF MS) Systems. Ann Lab Med 2019;39:530–536; doi: 10.3343/alm.2019.39.6.530.
135. Parker CT, Tindall BJ, Garrity GM. International code of nomenclature of Prokaryotes. Int J Syst Evol Microbiol 2019;69(1):S1; doi: 10.1099/IJSEM.0.000778/CITE/REFWORKS.
136. Jaffe JD, Stange-Thomann N, Smith C, et al. The Complete Genome and Proteome of Mycoplasma mobile. Genome Res 2004;14(8):1447–1461; doi: 10.1101/GR.2674004.
137. Himmelreich R, Hilbert H, Plagens H, et al. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res 1996;24(22):4420–4449.
138. Chambaud I, Heilig R, Ferris S, et al. The complete genome sequence of the murine respiratory pathogen Mycoplasma pulmonis. Nucleic Acids Res 2001;29(10):2145; doi: 10.1093/NAR/29.10.2145.
139. Chen S, Hao H, Zhao P, et al. Genome-wide analysis of the first sequenced Mycoplasma capricolum subsp. capripneumoniae strain M1601. G3: Genes, Genomes, Genetics 2017;7(9):2899–2906; doi: 10.1534/G3.117.300085/-/DC1.
140. Guimaraes AMS, Santos AP, SanMiguel P, et al. Complete Genome Sequence of Mycoplasma suis and Insights into Its Biology and Adaption to an Erythrocyte Niche. PLoS One 2011;6(5):e19574; doi: 10.1371/JOURNAL.PONE.0019574.
141. Inamine JM, Ho K-C, Loechel S, et al. Evidence that UGA Is Read as a Tryptophan Codon Rather Than as a Stop Codon by Mycoplasma pneumoniae, Mycoplasma genitalium, and Mycoplasma gallisepticum. J Bacteriol 1990;172(1):504–506.
142. Naderi Sima, Saier Jr MH. MicroCorrespondance. Mol Microbiol 1996;22(2):389–391; doi: 10.1046/J.1365-2958.1996.00033.X.
143. Yus E, Maier T, Michalodimitrakis K, et al. Impact of genome reduction on bacterial metabolism and its regulation. Science (1979) 2009;326(5957):1263–1268; doi: 10.1126/SCIENCE.1177263.
144. Boyd DA, Cvitkovitch DG, Hamilton IR. Sequence, Expression, and Function of the Gene for the Nonphosphorylating, NADP-Dependent Glyceraldehyde-3-Phosphate Dehydrogenase of Streptococcus mutans. J Bacteriol 1995;177(10):2622–2727.
145. Brocchi M, de Vasconcelos ATR, Zaha A. Restriction-modification systems in Mycoplasma spp. Genet Mol Biol 2007;30(SUPPL. 1):236–244; doi: 10.1590/S1415-47572007000200011.
146. Hille F, Richter H, Wong SP, et al. The Biology of CRISPR-Cas: Backward and Forward. Cell 2018;172(6):1239–1259; doi: 10.1016/J.CELL.2017.11.032.
147. Ipoutcha T, Tsarmpopoulos I, Talenton V, et al. Multiple Origins and Specific Evolution of CRISPR/Cas9 Systems in Minimal Bacteria (Mollicutes). Front Microbiol 2019;10:2701; doi: 10.3389/FMICB.2019.02701/BIBTEX.
148. Ben-Menachem G, Himmelreich R, Herrmann R, et al. The thioredoxin reductase system of mycoplasmas. Microbiology (Reading) 1997;143 (Pt 6)(6):1933–1940; doi: 10.1099/00221287-143-6-1933.
149. Li Y, Zheng H, Liu Y, et al. The Complete Genome Sequence of Mycoplasma bovis Strain Hubei-1. PLoS One 2011;6(6):e20999; doi: 10.1371/JOURNAL.PONE.0020999.
150. Bürki S, Frey J, Pilo P. Virulence, persistence and dissemination of Mycoplasma bovis. Vet Microbiol 2015;179(1–2):15–22; doi: 10.1016/J.VETMIC.2015.02.024.
151. Seymour LM, Jenkins C, Deutscher AT, et al. Mhp182 (P102) binds fibronectin and contributes to the recruitment of plasmin(ogen) to the Mycoplasma hyopneumoniae cell surface. Cell Microbiol 2012;14(1):81–94; doi: 10.1111/J.1462-5822.2011.01702.X.

152. Pilo P, Vilei EM, Peterhans E, et al. A Metabolic Enzyme as a Primary Virulence Factor of Mycoplasma mycoides subsp. mycoides Small Colony. J Bacteriol 2005;187(19):6824; doi: 10.1128/JB.187.19.6824-6831.2005.
153. Anonymous. U.S. Aquaculture | NOAA Fisheries. n.d. Available from: https://www.fisheries.noaa.gov/national/aquaculture/us-aquaculture [Last accessed: 12/13/2021].
154. Pierce ML, Ward JE, Holohan BA, et al. The influence of site and season on the gut and pallial fluid microbial communities of the eastern oyster, Crassostrea virginica (Bivalvia, Ostreidae): community-level physiological profiling and genetic structure. Hydrobiologia 2016;765(1):97–113; doi: 10.1007/S10750-015-2405-Z/TABLES/6.
155. Pimentel ZT, Dufault-Thompson K, Russo KT, et al. Microbiome Analysis Reveals Diversity and Function of Mollicutes Associated with the Eastern Oyster, Crassostrea virginica. mSphere 2021;6(3); doi: 10.1128/MSPHERE.00227-21/ASSET/2AA43C91-6062-4AC5-84B5-E3D6AE115C4D/ASSETS/IMAGES/MEDIUM/MSPHERE.00227-21-F004.GIF.
156. Pierce ML, Ward JE. Microbial Ecology of the Bivalvia, with an Emphasis on the Family Ostreidae. https://doi.org/102983/0350370410 2018;37(4):793–806; doi: 10.2983/035.037.0410.
157. King GM, Judd C, Kuske CR, et al. Analysis of Stomach and Gut Microbiomes of the Eastern Oyster (Crassostrea virginica) from Coastal Louisiana, USA. PLoS One 2012;7(12):e51475; doi: 10.1371/JOURNAL.PONE.0051475.
158. Fernandez-Piquer J, Bowman JP, Ross T, et al. Molecular analysis of the bacterial communities in the live Pacific oyster (Crassostrea gigas) and the influence of postharvest temperature on its structure. J Appl Microbiol 2012;112(6):1134–1143; doi: 10.1111/J.1365-2672.2012.05287.X.
159. Froelich B, Oliver JD. The Interactions of Vibrio vulnificus and the Oyster Crassostrea virginica. Microb Ecol 2013;65(4):807–816; doi: 10.1007/S00248-012-0162-3/METRICS.
160. Ndraha N, Wong H chung, Hsiao HI. Managing the risk of Vibrio parahaemolyticus infections associated with oyster consumption: A review. Compr Rev Food Sci Food Saf 2020;19(3):1187–1217; doi: 10.1111/1541-4337.12557.
161. Woods JW, Calci KR, Marchant-Tambone JG, et al. Detection and molecular characterization of norovirus from oysters implicated in outbreaks in the US. Food Microbiol 2016;59:76–84; doi: 10.1016/j.fm.2016.05.009.
162. Williamson TR, Tilley DR, Campbell E. Emergy analysis to evaluate the sustainability of two oyster aquaculture systems in the Chesapeake Bay. Ecol Eng 2015;85:103–120; doi: 10.1016/j.ecoleng.2015.09.052.
163. Canty R, Blackwood D, Noble R, et al. A comparison between farmed oysters using floating cages and oysters grown on-bottom reveals more potentially human pathogenic Vibrio in the on-bottom oysters. Environ Microbiol 2020;22(10):4257–4263; doi: 10.1111/1462-2920.14948.
164. Unzueta-Martínez A, Downey-Wall AM, Cameron LP, et al. Ocean acidification alters the diversity and structure of oyster associated microbial communities. Limnol Oceanogr Lett 2021;6(6):348–359; doi: 10.1002/lol2.10214.
165. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform 2021;22(1):178–193; doi: 10.1093/bib/bbz155.
166. Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34(17):i884–i890; doi: 10.1093/bioinformatics/bty560.

167. Vasimuddin Md, Misra S, Li H, et al. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS) IEEE; 2019; pp. 314–324; doi: 10.1109/IPDPS.2019.00041.
168. Lu J, Breitwieser FP, Thielen P, et al. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput Sci 2017;2017(1):e104; doi: 10.7717/PEERJ-CS.104/SUPP-5.
169. Dabdoub S. Kraken-Biom: Enabling Interoperative Format Conversion for Kraken Results (Version 1.2). 2016.
170. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One 2013;8(4):e61217; doi: 10.1371/journal.pone.0061217.
171. Oksanen J, BFG, KR, LP, MPR, ORB, SGL, SP, SMHH and WH. Vegan: Community Ecology Package. R Package Version 2.2-0. n.d. Available from: http://CRAN.Rproject.org/package=vegan [Last accessed: 7/16/2023].
172. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. 2016. Available from: https://ggplot2.tidyverse.org. [Last accessed: 7/16/2023].
173. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12); doi: 10.1186/s13059-014-0550-8.
174. RStudio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. 2020. Available from: http://www.rstudio.com/ [Last accessed: 7/16/2023].
175. Suzek BE, Huang H, McGarvey P, et al. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23(10):1282–1288; doi: 10.1093/bioinformatics/btm098.
176. Caspi R, Altman T, Billington R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 2014;42(D1):D459–D471; doi: 10.1093/nar/gkt1103.
177. Beghini F, McIver LJ, Blanco-Míguez A, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife 2021;10; doi: 10.7554/eLife.65088.
178. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45(D1):D353–D361; doi: 10.1093/nar/gkw1092.
179. Mallick H, Rahnavard A, McIver LJ, et al. Multivariable association discovery in population-scale meta-omics studies. PLoS Comput Biol 2021;17(11):e1009442; doi: 10.1371/journal.pcbi.1009442.
180. Nurk S, Meleshko D, Korobeynikov A, et al. MetaSPAdes: A new versatile metagenomic assembler. Genome Res 2017;27(5):824–834; doi: 10.1101/GR.213959.116/-/DC1.
181. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 2016;32(7):1088–1090; doi: 10.1093/BIOINFORMATICS/BTV697.
182. Kang DD, Li F, Kirton E, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 2019;2019(7); doi: 10.7717/PEERJ.7359/SUPP-3.
183. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 2016;32(4):605–607; doi: 10.1093/BIOINFORMATICS/BTV638.
184. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–2079; doi: 10.1093/bioinformatics/btp352.

185. Sieber CMK, Probst AJ, Sharrar A, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology 2018 3:7 2018;3(7):836–843; doi: 10.1038/s41564-018-0171-1.
186. Yin Y, Mao X, Yang J, et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. n.d.; doi: 10.1093/nar/gks479.
187. Robert J Smith. Phytomosaic/Ecole: Ecole: School of Ecology Package. 2021. Available from: https://github.com/phytomosaic/ecole [Last accessed: 7/16/2023].
188. Chase E, Young S, Harwood VJ. Sediment and Vegetation as Reservoirs of Vibrio vulnificus in the Tampa Bay Estuary and Gulf of Mexico. Appl Environ Microbiol 2015;81(7):2489–2494; doi: 10.1128/AEM.03243-14.
189. Trabal Fernández N, Mazón-Suástegui JM, Vázquez-Juárez R, et al. Changes in the composition and diversity of the bacterial microbiota associated with oysters (Crassostrea corteziensis, Crassostrea gigas and Crassostrea sikamea) during commercial production. FEMS Microbiol Ecol 2014;88(1):69–83; doi: 10.1111/1574-6941.12270.
190. Arfken A, Song B, Allen SK, et al. Comparing larval microbiomes of the eastern oyster (Crassostrea virginica) raised in different hatcheries. Aquaculture 2021;531:735955; doi: 10.1016/j.aquaculture.2020.735955.
191. Asmani K, Petton B, Le Grand J, et al. Establishment of microbiota in larval culture of Pacific oyster, Crassostrea gigas. Aquaculture 2016;464:434–444; doi: 10.1016/j.aquaculture.2016.07.020.
192. Scoffone VC, Chiarelli LR, Trespidi G, et al. Burkholderia cenocepacia Infections in Cystic Fibrosis Patients: Drug Resistance and Therapeutic Approaches. Front Microbiol 2017;8; doi: 10.3389/fmicb.2017.01592.
193. Parra-Luna M, Martín-Pozo L, Hidalgo F, et al. Common sea urchin (Paracentrotus lividus) and sea cucumber of the genus Holothuria as bioindicators of pollution in the study of chemical contaminants in aquatic media. A revision. Ecol Indic 2020;113:106185; doi: 10.1016/j.ecolind.2020.106185.
194. Sully S, Burkepile DE, Donovan MK, et al. A global analysis of coral bleaching over the past two decades. Nat Commun 2019;10(1):1264; doi: 10.1038/s41467-019-09238-2.
195. Cole KM, Supan J, Ramirez A, et al. Suspension of oysters reduces the populations of Vibrio parahaemolyticus and Vibrio vulnificus. Lett Appl Microbiol 2015;61(3):209–213; doi: 10.1111/LAM.12449.