Insights into Symbiont Population Structure among Three Vestimentiferan Tubeworm Host Species at Eastern Pacific Spreading Centers

Maëva Perez, S. Kim Juniper
School of Earth and Ocean Sciences, University of Victoria, Victoria, BC, Canada; Department of Biology, University of Victoria, Victoria, BC, Canada

ABSTRACT
The symbiotic relationship between vestimentiferan tubeworms and their intracellular chemosynthetic bacteria is one of the more noteworthy examples of adaptation to deep-sea hydrothermal vent environments. The tubeworm symbionts have never been cultured in the laboratory. Nucleotide sequences from the small subunit rRNA gene suggest that the intracellular symbionts of the eastern Pacific vent tubeworms Oasisia alvinae, Riftia pachyptila, Tevnia jerichonana, and Ridgeia piscesae belong to the same phytype of gammaproteobacteria, “Candidatus Endoriftia persephone.” Comparisons of symbiont genomes between the East Pacific Rise tubeworms R. pachyptila and T. jerichonana confirmed that these two hosts share the same symbionts. Two Ridgeia symbiont genomes were assembled from trophosome metagenomes from worms collected from the Juan de Fuca Ridge (one and five individuals, respectively). We compared these assemblies to those of the sequenced Riftia and Tevnia symbionts. Pangenome composition, genome-wide comparisons of the nucleotide sequences, and pairwise comparisons of 2,313 orthologous genes indicated that “Ca. Endoriftia persephone” symbionts are structured on large geographical scales but also on smaller scales and possibly through host specificity.

IMPORTANCE
Remarkably, the intracellular symbionts of four to six species of eastern Pacific vent tubeworms all belong to the same phytype of gammaproteobacteria, “Candidatus Endoriftia persephone.” Understanding the structure, dynamism, and interconnectivity of “Ca. Endoriftia persephone” populations is important to advancing our knowledge of the ecology and evolution of their host worms, which are often key species in vent communities. In this paper, we present the first genomes for symbionts associated with the species R. piscesae, from the Juan de Fuca Ridge. We then combine these genomes with published symbiont genomes from the East Pacific Rise tubeworms R. pachyptila and T. jerichonana to develop a portrait of the “Ca. Endoriftia persephone” pangenome and an initial outline of symbiont population structure in the different host species. Our study is the first to apply genome-wide comparisons of “Ca. Endoriftia persephone” assemblies in the context of population genetics and molecular evolution.

A defining characteristic of hydrothermal vent ecosystems is the diversity and ubiquity of mutualistic partnerships between metazoa (multicellular organisms) and chemolithoautotrophic bacteria. Among these associations, one of the most remarkable is the well-studied model symbiosis between the giant tubeworm Riftia pachyptila and its unique sulfide-oxidizing gammaproteobacterial partner, “Candidatus Endoriftia persephone” (1). These intracellular symbionts are hosted within the specialized cells (bacteriocytes) of an organ known as the trophosome, which occupies most of the space in the cephalic cavity of the animal’s trunk. In this mutualistic association, the worm supplies the bacterial partner with the inorganic compounds necessary for coupling sulfide oxidation to CO2 fixation: dioxygen, carbon dioxide, and hydrogen sulfide (mostly as its sulfhydryc anion HS−). These substances diffuse across the gills into the blood of the animal and are then transported to the trophosome. In return, the endosymbionts provide the tubeworm with the organic molecules necessary for growth and metabolism, either by excreting those molecules or by being digested directly (2, 3). The symbiotic bacteria are transmitted horizontally, that is to say, acquired de novo from the environment at each generation (4). The symbionts penetrate the worm tissues through the epidermis and migrate to a region between the dorsal blood vessel and the foregut to form the prototrophosome. As the metatrochophore larva develops into an adult, its digestive tract atrophies in favor of the trophosome (5). The vestimentiferan adult thus becomes completely dependent on its bacteria for nutrition. For the symbionts, however, this association seems to be facultative. Free-living “Ca. Endoriftia persephone” symbionts have been detected in biofilms and seawater surrounding R. pachyptila aggregations (4), and it has been demonstrated recently that the Riftia symbionts can return to their free-living stage upon the death of the worm (6), thereby maintaining/sustaining environmental populations.

In addition to having a viable free-living stage, the symbionts
showed no visible tissue damage were rinsed with 70% ethyl alcohol until further processing. In our laboratory, the contents of the worms’ trunks (which include the trophosome) were removed by dissection and rinsed with 70% EtOH. Finally, the lab contents of the worms’ trunks were processed for DNA extraction. DNA extracts from six individual worms were sequenced and assembled at Genome Quebec, Montreal, Canada. Samples were prepared using standard protocols and were sequenced on the Illumina HiSeq 2000 platform. A subset of these samples was also sequenced on the Illumina MiSeq platform. The resulting metagenomes were assembled into two high-quality symbiont genome assemblies essentially free of host and bacterial contaminant sequences (see the supplemental material for a description of data quality assessment, curation to remove contaminant sequences, and in silico filtering).

The symbiont genome assemblies were performed using the IMG-ER platform of the Joint Genome Institute.

MATERIALS AND METHODS

Ridgeia symbiont genome assembly. (i) Sample collection and symbiont genome sequencing. Specimens of R. pachyptila were collected from the Axial Volcano and the Main Endeavor Field, on the Juan de Fuca Ridge, during a remotely operated vehicle (ROV) cruise on the R/V Thomas G. Thompson in July 2010. As described by Forget et al. (16), the worms were recovered and brought to the ship in sealed bioboxes. On board, individual worms were carefully removed from their tubes, and those that showed no visible tissue damage were rinsed with 70% ethyl alcohol (EtOH) and were flash frozen at −80°C until further processing. In our laboratory, the contents of the worms’ trunks (which include the trophosome) were removed by dissection and rinsed with 70% EtOH. Finally, the DNA from each dissected trunk was extracted using the QIAGEN DNeasy blood and tissue kit. DNA extracts from six individual worms were sequenced and assembled at Genome Quebec, Montreal, Canada. Samples were prepared using standard protocols and were sequenced on the Illumina HiSeq 2000 platform. A subset of these samples was also sequenced on the Illumina MiSeq platform. The resulting metagenomes were assembled de novo into two high-quality symbiont genome assemblies essentially free of host and bacterial contaminant sequences (see the supplemental material for a description of data quality assessment, curation to remove contaminant sequences, and de novo assembly).

(ii) Gene annotations. Gene calling for the Ridgeia 1 and Ridgeia 2 symbiont assemblies was performed using the IMG-ER platform of the Joint Genome Institute.

Genome-wide comparisons of Ridgeia symbionts with all other published vestimentiferan genomes. (i) Genome alignments. The genomes of the Ridgeia 1 and Ridgeia 2 symbiont assemblies were aligned with the
closely related Riftia 1, Riftia 2, and Tenvia symbionts (15) using progressive Mauve (17). This anchored alignment algorithm finds so-called locally colinear blocks (LCBs)—genomes that appear free of chromosomal rearrangement—and outputs the aligned sequences of each LCB in XMFA multiple alignment format as well as in a file containing the positions of the LCBs in each of the genomes (backbone file).

(ii) Pangenome composition. The pangenome composition was determined by the presence or absence and sizes of LCB sequences in each genome. Since colinear blocks have been found previously to be informative for phylogenetic analysis (18), we further used the presence or absence of individual LCBs (>100 bp) to compute faccard distances with the vegan package in R (19) and built a neighbor-joining tree using Populations software, version 1.2.32 (http://bioinfomatics.org/~tryphon/populations/). Bootstrap values were obtained from 100 bootstrap subsamples by use of the boot.phylo function from the ape package in R. Finally, a custom Python script was used to extract the annotations of all genes within the LCBs of interest from GenBank files. These genes were further annotated through visual inspection against the Mauve-generated multiple-genome alignment in order to record additional information, such as the representation of neighboring LCBs across the assemblies and nucleotide conservation.

(iii) Core genome nucleotide heterogeneity. Our analysis of core genome nucleotide heterogeneity for symbionts from the three tubeworm species also included the assembly for the Riftia symbionts (1). This first published assembly was not used in subsequent, more in-depth analyses because of its lower quality, as explained below.

We used the stripSubsetLCBs command to extract the large (>100-bp) LCBs represented in all of the assemblies from the XMFA file. For each LCB, a FASTA file was generated, and the sequences were aligned with MAFFT (20). Subsequently, all the resulting alignments were concatenated to form a single genome-wide alignment of 2,580,528 bp with 75,472 variable sites. Finally, we used SeaView (21) to calculate the pairwise genetic distances using the Hasegawa-Kishino-Yano (HKY) model (22) and to generate a 100-bootstrap neighbor-joining tree.

(iv) Pairwise comparison of homologous genes. A file containing a table of all the homologous protein-coding genes was obtained using the Export Positional Homologs command from Mauve’s menu and the following parameters: minimum identity, 80; minimum coverage, 50. This table was then curated to keep only the entries of genes present in all of the genomes. Subsequently, we generated local protein and nucleic acid Blast databases of all of the coding sequences of the pangenome, from which we extracted, in FASTA format, the nucleotide and amino acid sequences of these genes by use of the blastdbcmd tool of BLAST++ (23). Then we aligned the amino acid sequences and generated protein sequence identity matrices with Clustal Omega (24). Subsequently, protein alignments were converted into codon-based nucleotide alignments with PAL2NAL (25). Finally, the nucleotide sequence identity matrices and the ratios of non-synonymous to synonymous substitution rates (dN/dS ratios) were calculated using Clustal Omega (24) and the YN00 program of the PAML package (26, 27), respectively. Genome-wide dN/dS ratios were generated from the concatenated codon-based alignments.

Mauve’s transitive algorithm identifies positionally homologous sequences. In closely related genomes, these positional homologs are also orthologs, but the algorithm could still mistakenly catch recently duplicated genes (paralogs). To prevent comparisons between paralogs, the protein sequences of all homologs with nucleotide identities lower than 50% were subjected to reciprocal BLAST searches against the five reference genomes. The homologous associations were then adjusted to include the true orthologs or were removed from the data set if orthologous sequences were missing in at least one genome. Fewer than a dozen homologous genes among the 2,324 identified were thereby curated (excluded from further analyses). We believe that the remaining cases of paralogous associations are limited to just a few extra genes and do not significantly affect our results.

Accession number(s). The Ridgeia 1 and Ridgeia 2 symbiont assemblies are published on the Joint Genome Institute IMG system with genome IDs 2651869500 and 2643221413, respectively, as well as in GenBank under accession numbers LDXT00000000 and LMXI00000000, respectively. The versions described in this paper are LDXT01000000 and LMXI01000000.

RESULTS
Metagenome assemblies of Ridgeia symbionts. (i) Assembly quality. The quality of a genome assembly depends on its completeness and coverage. These and other features of all available “Ca. Endoriftia persephone” genome assemblies are compared in Table 2. Completeness can be estimated from the number and size distribution (N50) of contigs, while coverage, calculated as the average per-base sequencing density, is a measure of the sampling effort. Higher sequencing depth results in higher sequence accuracy but also improves the completeness of isolate genomes.

The Ridgeia 1 symbiont assembly was of high quality (Table 2). It contained fewer and longer contigs than the Riftia 1 symbiont assembly, and its coverage was 7 to 16 times greater than those of all previously published assemblies of “Ca. Endoriftia persephone” (1, 15). Yet even with such high sequencing depth, we were not able to close the genome. We suspect that chromosome rearrangement within the symbiont population might be the cause of this fragmentation, because it can create ambiguous links during the scaffolding step of the assembly.

The “pooled” Ridgeia 2 symbiont genome was generally lower in quality than the assemblies of Gardebrecht et al. (15) but still notably superior in completeness and coverage to the assembly of Robidart et al. (1) (Table 2). Because of the lower overall quality and considerable differences in gene annotations, the assembly of Robidart et al. (1) was not used in our analyses.

(ii) Confirmation that the Ridgeia symbionts are “Ca. Endoriftia persephone.” The Ridgeia 1 and Ridgeia 2 16S rRNA, 23S rRNA, and ITS sequences were 100% identical to each other and

<table>
<thead>
<tr>
<th>Symbiont*</th>
<th>Genome size (Mb)</th>
<th>No. of contigs</th>
<th>N50</th>
<th>Coverage (fold)</th>
<th>No. of reads</th>
<th>GC%</th>
<th>Total</th>
<th>Protein coding</th>
<th>rRNA</th>
<th>tRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridgeia 1 symbiont</td>
<td>3.44</td>
<td>97</td>
<td>83.9</td>
<td>180</td>
<td>7,436,749</td>
<td>58.9</td>
<td>3,188</td>
<td>3,132</td>
<td>3</td>
<td>47</td>
</tr>
<tr>
<td>Ridgeia 2 symbiont</td>
<td>3.42</td>
<td>693</td>
<td>7.6</td>
<td>17</td>
<td>993,690</td>
<td>58.9</td>
<td>3,698</td>
<td>3,641</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>Tenvia symbiont</td>
<td>3.64</td>
<td>184</td>
<td>92.7</td>
<td>15</td>
<td>212,833</td>
<td>58.2</td>
<td>3,277</td>
<td>3,230</td>
<td>3</td>
<td>44</td>
</tr>
<tr>
<td>Riftia 1 symbiont</td>
<td>3.48</td>
<td>197</td>
<td>28.4</td>
<td>25</td>
<td>467,070</td>
<td>58.2</td>
<td>3,254</td>
<td>3,209</td>
<td>3</td>
<td>42</td>
</tr>
<tr>
<td>Riftia 2 symbiont</td>
<td>3.71</td>
<td>414</td>
<td>24.6</td>
<td>13</td>
<td>205,880</td>
<td>58.2</td>
<td>3,566</td>
<td>3,515</td>
<td>4</td>
<td>47</td>
</tr>
<tr>
<td>“Candidatus Endoriftia persephone”</td>
<td>3.20</td>
<td>2170</td>
<td>1.9</td>
<td>11</td>
<td>130,000</td>
<td>57.9</td>
<td>6,450</td>
<td>6,414</td>
<td>4</td>
<td>32</td>
</tr>
</tbody>
</table>

* Data for the Ridgeia 1 and Ridgeia 2 symbionts are from this paper; data for the Tenvia, Riftia 1, and Riftia 2 symbionts are from the work of Gardebrecht et al. (15); and data for “Candidatus Endoriftia persephone” are from the work of Robidart et al. (1).
The relatively large size of the unconserved genome (3% of the “Ca. Endoriftia persephone” pangenome) was likely the result of gaps in the genome assemblies and the small sample size. We expect that increasing the quality of the data and the number of samples would reduce the relative importance of the unconserved genome in favor of the conserved pangenome or the regional core genome.

Genes encoded in the accessory genome. (i) Region-specific genome. The LCBs that were exclusive to the JdFR or EPR symbiont genomes both carried unique genes coding for transposases, integrases, and other phage-associated proteins, as well as a few genes involved in cell wall/membrane/envelope biogenesis (see Tables S1 and S2 in the supplemental material).

Interestingly, two clustered regularly interspaced short palindromic repeat (CRISPR–CRISPR-associated protein (Cas) systems (28) were found in all of the genome assemblies. The first was well conserved, but the spacers were notably different in the two genomes for which the CRISPR locus was successfully assembled (the Ridgeia 1 and Tevnia symbionts). In the second CRISPR–Cas system, the cas operon was not conserved across symbionts from the JdFR and the EPR; half of the cas genes were not homologous (see Tables S1 and S2 in the supplemental material).

Finally, a 17-kbp scaffold with genes encoding a type VI secretion system was found uniquely in the Ridgeia symbionts (see Table S1 in the supplemental material).

(ii) Vent site–versus species–specific genomes at the EPR. The LCBs unique to Riftia symbiont assemblies were limited to three contiguous scaffolds (see Table S3 in the supplemental material). The first was about 60 kbp long and contained genes typically found in fertility factors, i.e., the gene coding for the OmpA/MotB domain-containing protein, tra genes, IS200-like genes coding for transposases, and two genes coding for nucleotide-binding proteins. The other two scaffolds were smaller (about 10 kbp) and contained, respectively, six genes of the CRISPR-Cas3 system and four genes: two unannotated genes, one gene encoding a transcriptional regulator, and one encoding a putative relaxase. Because of the incompleteness of the Tevnia symbiont assembly, we could not rule out the possibility that these differences resulted from a biased sampling of the Tevnia symbiont’s metagenome. However, given the large size of the missing scaffolds, we suggest that this was unlikely.

Among the LCBs found exclusively in the assemblies of symbionts from the vent site at 9°N, however, many seemed to have resulted from a poor sampling of the fragmented genome of the Riftia 1 symbiont. They represented stretches of DNA of a few thousand base pairs, often located at the extremities of conserved contigs. In contrast, other unique sequences seemed to represent real chromosomal differences, tended to be larger (up to 16.3 kbp), were flanked by regions of low nucleotide conservation, and contained unique mobile elements, toxin/antitoxin genes, and transcriptional regulator genes typically found in phage genomes (see Table S4 in the supplemental material).

Population structure of “Ca. Endoriftia persephone.” (i) Cluster analyses. The first sequenced metagenome of “Ca. Endoriftia persephone” (1) clustered apart from the more recent assemblies. This is probably a result of sequencing errors due to the overall lower quality of reads associated with the sequencing methods used at the time.

The Ridgeia symbionts cluster apart from the EPR symbionts.

FIG 1 Pangenome of “Candidatus Endoriftia persephone” based on the relative sizes of the locally colinear blocks shared by five “Ca. Endoriftia persephone” assemblies from two distinct geographical regions. The Ridgeia 1 and Ridgeia 2 symbionts are from the Juan de Fuca Ridge, and the Tevnia, Riftia 1, and Riftia 2 symbionts are from the East Pacific Rise. The five genome assemblies were aligned with progressiveMauve (17).

differed from the Tevnia and Riftia 1 and 2 symbiont sequences by 1, 0, and 3 nucleotides, respectively. This is consistent with the hypothesis that the same species of symbionts, “Ca. Endoriftia persephone,” is associated with Riftia, Tevnia, and Ridgeia tube-worms.

This hypothesis is further supported by the fact that the majority of “Ca. Endoriftia persephone” genes had homologs in the Ridgeia symbiont assemblies (see Fig. 3; also Data Set S1 in the supplemental material).

Like the symbionts associated with Riftia and Tevnia tube-worms, the Ridgeia symbionts have a diverse metabolism and possess genes for sulfide oxidation, carbon fixation through the Calvin-Benson-Bassham and reverse tricarboxylic acid (rTCA) cycles, denitrification, motility, and chemotaxis (see Table S5 in the supplemental material).

Pangenome composition of “Ca. Endoriftia persephone.” Figure 1 shows the composition of the pangenome of “Ca. Endoriftia persephone” based on the nucleotide sequences of the five most recent “Ca. Endoriftia persephone” assemblies. A core genome, representing 89% of the pangenome of “Ca. Endoriftia persephone,” was shared across all of the assemblies of Riftia, Tevnia, and Ridgeia symbionts. In addition, 4% of the pangenome was region specific, i.e., found in and shared among symbionts from the same geographical region only (the JdFR symbionts associated with R. piscesae or the EPR symbionts associated with R. pachyptila and T. jerichonana) (Fig. 1). Symbionts from the same geographical region shared as much as 98% of their genomes.

Finally, we found that 0.7 to 2.9% of the pangenome was unique to the specific assemblies and was in part composed of contaminant sequences and/or exogenous genetic material acquired recently through horizontal transfer. This is supported by the fact that the GC content of the unique genome for some assemblies was notably different from that of the core genome. In the Tevnia symbiont, for example, the GC content of the unique genome (96 kbp) was 42%, while that of the core genome was 60%.
both in terms of nucleotide distances in the core genome and in terms of the compositions of their accessory genomes (Fig. 2).

Within the EPR, the symbionts cluster by host species when classified by the nucleotide sequences of the core genome, while they cluster by vent site when classified by the composition of the accessory genome. Thus, at the area of the EPR (near 9°N) for which there are genome assemblies for both Riftia and Tevnia symbionts, the symbionts from these two hosts shared more exclusive LCBs than did Riftia symbionts collected from different EPR areas (9°N versus 13°N). Interestingly, the accessory genome exclusive to the vent site at 9°N was composed of shorter LCBs, and was slightly smaller overall, than the symbiont genome exclusive to the Riftia host species (70 kbp and 80 kbp, respectively) (see Tables S3 and S4 in the supplemental material).

Finally, Tevnia symbionts seemed to be closer to Ridgeia than to Riftia symbionts in terms of nucleotide identity.

(ii) Comparisons of orthologous genes. Nucleotide heterogeneity was <1% within the EPR tubeworm symbionts and within the Ridgeia symbionts from the JdFR but ca. 2% between symbionts from the two different regions (Fig. 3A). Yet many homologous proteins were highly conserved across the assemblies from the JdFR and the EPR, indicating strong purifying selection acting on “Ca. Endoriftia persephone” (Fig. 3B).

The Ridgeia symbionts appeared more homogeneous than those from the EPR. Eighty-nine percent of the genes in the two Ridgeia symbiont assemblies had identical nucleotide sequences, compared to 54% on the EPR. These results were corroborated with the overall synonymous substitution rates (dS) and the ratio of nonsynonymous to synonymous substitution rates (dN/dS ratio) between pairs of symbiont assemblies (Fig. 3C).

The synonymous substitution rate is the ratio of the number of synonymous substitutions to the number of synonymous sites. Because synonymous substitutions tend to be selectively neutral, they accumulate over time and thus can be used as a proxy for divergence between genomes (29, 30). Assuming allopatric diver-
gence, we can then derive a molecular clock for the substitution rate \( r \) for the \( \text{“Ca. Endoriftia persephone”} \) sybionts by use of the equation \( r = dS/2T \), where \( dS \) is the divergence observed between the vicariant populations at the synonymous sites and \( T \) is the time of last contact between the East Pacific Rise and northeast Pacific ridge systems. Following the work of Vrijenhoek (31), we used a \( T \) of 28.5 million years ago and obtained a substitution rate of 0.14% (± 0.01%) per million years.

This substitution rate is lower than the rates observed for \( \text{Escherichia coli} \) in culture (0.45%) (32) and for the host vestimentiferan tubeworms themselves (−0.2%) (33). However, both of the latter substitution rates were based on comparisons of isolate genomes, while our rate was determined from comparisons of genome assemblies that resulted from the concatenation of multiple sybionts with potentially multiple genotypes. Thus, we may have underestimated the genetic diversity within and across the sybiont populations and therefore the rate of divergence between the two populations. Furthermore, the divergence between two populations depends on their respective reproduction rates and on the parameters that affect their respective genetic diversity over generations (i.e., their underlying biological mutation rate, effective size, and clonality [34, 35]). Current knowledge of doubling times and genetic diversity for these sybionts does not permit confident estimation of most of these parameters.

Our data can be used for an initial consideration of effective sybiont population sizes for the three host tubeworm species considered here. The genome-wide \( dN/dS \) ratio of \( \text{“Ca. Endoriftia persephone”} \) sybionts falls into the upper range of what has been observed in closely related obligate sybionts (36). For closely related genomes, the \( dN/dS \) ratio is also intrinsically dependent on the time since divergence and the effective population size. More closely related lineages or lineages with smaller population sizes tend to have higher \( dN/dS \) ratios due to a time lag or delay in the curation of slightly deleterious mutations (37,38). \( \text{“Ca. Endoriftia persephone”} \) sybiont populations showed this pattern in that the \( dN/dS \) ratios were negatively correlated with the divergence between sybiont pairs. The highest divergence between the lowest \( dN/dS \) ratio was seen in the comparison between \( \text{Ridgea} \) and EPR sybionts (\( dS, \sim 0.08 \)), and the lowest divergence and the highest \( dN/dS \) ratio were between the two \( \text{Ridgea} \) sybiont assemblies (Fig. 3C).

Interestingly, while the divergences between EPR sybionts were quite similar (0.0101 < \( dS < 0.0136 \)), the \( dN/dS \) ratio for the two \( \text{Riftia} \) assemblies was notably higher than those for the other pairs. He et al. (39) and Luo et al. (40) made similar observations for the pathogen \( \text{Clostridium difficile} \) and lineages of the marine alphaproteobacterium \( \text{Roseobacter} \), respectively. This suggests that the sybionts in association with \( \text{Riftia} \) tubeworms have a smaller effective population size than the overall EPR \( \text{“Ca. Endoriftia persephone”} \) sybiont populations, and thus, that the latter might be further structured either spatiotemporally, according to environmental conditions, or through host specificity.

**DISCUSSION**

**Divergence of JdFR and EPR sybionts.** We used five high-quality genome assemblies of \( \text{“Ca. Endoriftia persephone”} \) to analyze the structure of the \( \text{“Ca. Endoriftia persephone”} \) population though pairwise comparisons of (i) the composition of the pan-genome, (ii) the nucleotide identity within the core genome, and (iii) the synonymous and nonsynonymous substitution rates for a large subsample of the core genome genes. Our results were consistent with those obtained from phylogenetic analyses based on 16S rRNA gene (41) and ITS sequences, as well as repetitive element palindromic PCR (rep-PCR) fingerprints (7), and indicated that the population of \( \text{“Ca. Endoriftia persephone”} \) sybionts in association with \( \text{R. piscesae} \) on the Juan de Fuca Ridge (JdFR) was distinct from the \( \text{“Ca. Endoriftia persephone”} \) population on the East Pacific Rise (EPR), which is associated with \( \text{R. pachyptila} \) and \( \text{T. jerichonana} \).

(i) **Allopatry.** Comparisons of the composition of vent-associated macrofauna communities (42) and the genetic structure of vestimentiferan worms (33) provide evidence that the northeast Pacific and EPR vent communities have been isolated by the development of discontinuities along the Pacific midocean ridge caused by the tectonic fracturing of the Farallon Plate about 30 million years ago (42, 43). Similar dichotomies attributed to later plate fragmentation events were observed in populations of various invertebrate species spanning multiple ridge systems in the northeast Pacific (44–46). It is therefore reasonable to assume that \( \text{“Ca. Endoriftia persephone”} \) populations were similarly affected by the emergence of these geographical barriers. Our results indicate that the divergence of the JdFR and EPR sybionts was dominated by passive processes/genetic drift. On the one hand, the core genome was characterized by overall low \( dN/dS \) ratios and a conserved codon bias (data not shown), suggesting that the same selective constraints acted on both populations. Additionally, when we compared the functional distribution of core genome genes with median \( dN/dS \) values to that of genes with extreme \( dN/dS \) values (5% highest \( dN/dS \) ratios), no Clusters of Orthologous Groups (COGs) or KEGG Orthology (KO) categories appeared to be overrepresented in the outliers (\( P > 0.05 \) by the chi-square test of independence). On the other hand, the accessory genome of each population of sybionts was composed of many mobile elements and selfish sequences, as well as unique CRISPR spacers, all of which suggest two distinct histories of interactions that have independently modified the EPR and JdFR sybiont genomes.

(ii) **Adaptations to viral predation.** The presence of phage DNA as well as two to three (for \( \text{Ridgea} \) 1 sybionts) CRISPR spacers can be seen as evidence that viruses are an important “enemy” of free-living and/or intracellular \( \text{“Ca. Endoriftia persephone”} \) and that the sybiont genomes carry these markers of phage infections.

Although little considered until recently, there is accumulating evidence for a viable and presumably metabolically active free-living stage of \( \text{“Ca. Endoriftia persephone”} \) (4, 6). Viruses are known to be abundant at deep-sea hydrothermal vents and are likely an important cause of mortality for free-living bacteria (47). Alternatively, the trophosome might also be a favorable environment for the proliferation of phages among the dense and fast-growing intracellular sybiont population.

The presence of CRISPR spacers that differ between the JdFR and EPR sybiont populations could suggest the existence of different \( \text{“Ca. Endoriftia persephone”} \)-specific viruses on these two midocean ridges, although we have also found CRISPR spacer variability within the sybiont population of a single worm (M. Perez and S. K. Juniper, submitted for publication).

(iii) **Host adaptation.** Some genes possibly involved in the symbiosis had relatively high \( dN/dS \) ratios (e.g., the chemotaxis protein CheY, the cell division protein DamX, an outer membrane...
protein), but the divergence between the two populations was too small for detection of the signature of positive selection (48). Nevertheless, the large scaffolds containing genes associated with a type VI secretion system, found only in Riftia symbionts, could be part of a mechanism of host adaptation. Indeed, the type VI secretion system can act as a virulence factor against eukaryotic cells or competing bacteria (49, 50). It has also been found to be key in determining host specificity in Rhizobium leguminosarum (51). Other genes involved in cell wall/membrane biogenesis could be involved in the expression of microbe-associated molecular patterns (MAMPs), hypothesized to be critical in mediating host-symbiont interactions (52). Genomic and proteomic comparisons with a sympatric population of “Ca. Endoriftia persephone” symbionts associated with a different host species (e.g., a Lamellibrachia sp. [7]) might tell us more about host specificity.

EPR symbionts are further structured into populations that might be relatively isolated spatially or temporally. Our results show little evidence for geographic differentiation of symbionts from the two sites on the EPR for which genome sequence data are available. Symbionts from a single vent site, at 9°N, were no more similar to each other than symbionts from two different sites, at 9°N and 13°N. In contrast, when symbionts from the two EPR host tubeworms were compared, the nucleotide sequences of symbionts hosted by the same tubeworm species were more homogeneous and had a higher dN/dS ratio, suggesting that Riftia symbionts formed a subpopulation within the EPR. Additionally, Riftia symbionts carried scaffolds with genes typically found in F-type conjugative plasmids. These genes have been speculated to play a role in horizontal gene transfer (15) and might allow for a high degree of genetic exchange between Riftia symbionts, thus keeping this population homogeneous.

While free-living symbionts can probably disperse on large scales and colonize new surfaces/vents independently of their hosts, small-scale spatial or temporal differences in environmental conditions could favor particular strains of symbionts, resulting in population partitioning. This local increase in homogeneity might be amplified or maintained in the presence of the tubeworm hosts through pseudoverdual transfer of symbionts (6).

Molecular mechanisms controlling host specificity might also exist, but a higher resolution of genetic diversity would be needed to clearly characterize differences in the symbionts’ accessory genomes.

Toward a better characterization of “Ca. Endoriftia persephone” populations. Whereas previous studies presenting “Ca. Endoriftia persephone” genomes focused on the metabolism of the symbiont (1, 15), this study was the first to apply genome-wide comparisons of “Ca. Endoriftia persephone” assemblies in the context of population genetics and molecular evolution. These comparisons underline the importance of viruses and genetic drift in shaping the genetic makeup of the symbionts and defining populations. Our findings suggest that, as with vent animal species, midocean ridge discontinuities in the eastern Pacific Ocean have resulted in allopatric divergence of symbiont populations on the Juan de Fuca Ridge and the East Pacific Rise. Furthermore, within a single ridge system, the symbiont populations are not panmictic and are possibly structured according to environmental conditions or host specificity, or both. Finally, genome-wide comparisons revealed that the population-specific functional genes are likely encoded in the accessory genome and potentially in plasmids.

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

We suggest that future studies focus on assessing the number and diversity of “Ca. Endoriftia persephone” genotypes. To this end, we propose that CRISPR spacers and extrachromosomal genetic material may have the potential to be used for high-resolution differentiation of populations of symbionts. For example, “CRISPR typing” has been used for genotyping human bacterial pathogens (53–57) and aquatic bacteria (58–60). In the meantime, sequencing of the complete genomes of individual Endoriftia cells would allow us to detect chromosomal differences.

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

ACKNOWLEDGMENTS

We thank the crews of the R/V Atlantis and R/V Thomas G. Thompson, as well as the pilots of the submersibles Alvin and ROPOS. This research was enabled in part by computing assistance provided by Belaid Moa, West-Grid (westgrid.ca), and Compute Canada/Calcul Canada (compecanada.ca). We also thank Nathalie Forget, Steve Perlman, Sebastien Duperon, the members of Verena Tunicliffe’s lab, Lee Katz, Diana Varela, Real Roy, Francis Nano, and all of the contributors to seqanswers.com and biostars.org.

We declare no conflict of interest.

FUNDING INFORMATION

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

REFERENCES

9. Nees HA, Lutz RA, Shank TM, Luther GW, III. 2009. Pre- and post-


