Article

Episodic Diversifying Selection, Recombination, and Symbiont Capture Characterize the Evolutionary History of Maternally-transmitted Clam Symbionts

Short title:
Evolution of maternally-transmitted symbiont genomes

C. R. Young

Corresponding Author:
C. Robert Young
National Oceanography Centre
University of Southampton Waterfront Campus
European Way
Southampton SO14 3ZH
United Kingdom

Tel: +44 (0)23 8059 6356
email: rob.young@noc.ac.uk

Classification:
Evolution, Genetics, Microbiology, Selection, Recombination, Symbiosis, Genomics, Vesicomyidae

Abstract

The shift from a free-living bacterial lifestyle to a vertically transmitted intracellular lifestyle is hypothesized to result in major genome-scale changes including a shift to AT-rich genomes, reduced recombination, and a reduction in genome size. In this study, host-symbiont coevolution, genome structure, recombination and diversifying selection were examined by applying comparative methods to symbiont and mitochondrial genomes from 11 vesicomyid deep-sea clam taxa. A pervasive pattern of natural selection affects many loci with functions such as DNA metabolism, transcription and translation as well as loci that are central to the chemosynthetic role of these symbionts reflecting the diverse selective pressures of an intracellular symbiont lifestyle. These signatures of selection are not limited to one or a few lineages, but are distributed throughout the history of the group. Recombination occurs frequently, but it is not randomly distributed among taxa. The non-random distribution of recombination among symbiont lineages might reflect constraints imposed by host distribution and physiology. One recent instance of a symbiont genome replacement was observed between two different host genera. These results suggest that both recombination and natural selection have been important features of adaptation to a vertically transmitted, intracellular lifestyle in this group of symbionts.

Introduction

Symbiosis is hypothesized to have been the mechanism that led to the development of cellular organelles (see reviews of Rand et al 2004, Kutschera and Niklas 2005 and references therein). The discovery of unique intracellular symbioses in deep-sea organisms provides an opportunity to study early organelle evolution and could help to elucidate ancient evolutionary processes involved in the acquisition of mitochondria and plastids (Kutschera and Niklas 2005), although the delineation between symbiont and organelle in this and other systems is debateable (McCutcheon and Moran 2011). The symbiosis between gammaproteobacteria and deep-sea clams is only about 50 MY old (Johnson et al. 2017). Vesicomyid clams harbour intracellular symbionts that provide sustenance to the host by utilizing reduced sulphide to fix carbon. The symbionts are located within bacteriocytes in the host gills, but are co-

localised in the host ovaries and transmitted by way of eggs each generation (Endow and Ohta 1990; Cary and Giovannoni 1993; Cary et al. 1993). Vertical transmission through eggs coupled with an intracellular lifestyle might drastically reduce opportunities for recombination (Silvia et al. 2003; Papke and Ward 2004; Wernegreen 2005) and characteristically leads to symbiont genomes that are reduced in size and gene content relative to environmentally acquired endosymbionts (Kuwahara et al. 2007; Newton et al. 2007). In contrast to species with horizontally transmitted symbionts (e.g., Won et al. 2006), concordant host and symbiont phylogenies (Distel, Felbeck and Cavanaugh 1994; Peek et al. 1998) and within-population genetic disequilibrium (Hurtado 2003) are observed among vesicomyid taxa. This pattern of co-speciation and genetic linkage has been hypothesized to be the result of mitochondrial and symbiont co-transmission through eggs, i.e., vertical transmission.

A vertical transmission mode through host eggs produces a symbiont population bottleneck each generation. Symbiont population sizes associated with *Phraegena okutanii* eggs have been estimated to be 400 symbiont cells containing 10 genomic copies per symbiont cell (Ikuta et al. 2016). Indeed, evidence of increased evolutionary rates in vesicomyid clam symbionts has previously been observed and is hypothesized to result from increased fixation of nearly neutral mutations due to reduced effective population sizes (N_e) arising from transmission bottlenecks and an intracellular lifestyle (Peek et al. 1998). The speed of adaptation of intracellular symbionts could also be affected by a reduced recombination rate through clonal interference between symbiont genomic mutations (Hill and Robertson 1966, Fogle et al. 2008) or genetic interactions between symbiont genes and host cytoplasmic or nuclear genes. Initial genomic studies reveal that vesicomyid symbionts lack DNA recombination and repair enzymes (Newton et al. 2007). This observation coupled with reduced access to divergent bacterial strains due to a strictly intracellular lifestyle might further affect evolutionary rate. Muller's ratchet (Felsenstein 1974, Muller 1964) has been evoked to explain the increased rate of fixation of deleterious mutations in symbiont genes (e.g., Moran 1996, Rand et al 2004). Genetic deletions are hypothesized to lead to genome size reduction due to reduced influx of genetic material due to recombination

(Moran and Mira 2001; Perez-Brocal et al. 2006), although it might be equally reasonable to hypothesize that deletions of unrequired genomic regions are selectively advantageous in symbionts. Nevertheless, strict vertical transmission coupled with an intracellular lifestyle is expected to reduce the efficacy of natural selection through reduced effective population sizes, reduced rates of recombination, reduced access to divergent strains of bacteria, and potential host-symbiont genetic interactions (McCutcheon and Moran 2011). Indeed, genomic stasis has been observed in *Buchnera aphidicola* over a time scale of 50MY, suggesting that this symbiont no longer contributes to ecological innovations of its hosts (Tamas et al. 2002).

The strictly vertical co-transmission hypothesis for vesicomyid clams was subsequently rejected in favour of a leaky vertical transmission mode characterized by occasional instances of horizontal transmission. This model is supported by observations of some discordance between mitochondrial and symbiont phylogenies (Stewart, Young and Cavanaugh 2008), occasional instances of divergent symbiont lineages co-occurring in a single host (Stewart and Cavanaugh 2009), and phylogenetic evidence of recombination among some symbiont lineages (Stewart, Young and Cavanaugh 2009). These observations suggest that positive selection might be freer to operate on protein coding gene sequences than previously expected. Occasional instances of recombination, either between divergent lineages or between closely related lineages within the same host population, could reduce linkage disequilibrium among symbiont loci or cytoplasmic genomes and mitigate clonal interference (but see, Hurtado et al. 2003). Likewise, genetic interactions between symbiont and host genes could be affected. In this comparative genomics study, we examined sequenced data from eleven vesicomyid symbiont and host mitochondrial genomes. We assess symbiont gene content among lineages, rates of recombination as well as host-symbiont switching, and estimate the strength of episodic diversifying selection shaping the diversity of protein coding genes.

Results

Host mitochondrial genomes

Host taxa examined in this study were chosen from the deepest diverging lineages in the pliocardiin Vesicomyidae and are distributed globally in the northern hemisphere (Fig. 1) and across the known host diversity (Fig. 2A). Host mitochondrial genomes from the lineages examined in this study possess identical gene orders and contents as previously published mitochondrial genomes (Liu et al., 2015; Ozawa et al., 2017). The phylogeny constructed with mitochondrial genome data (Fig. 2A) is congruent with the known host phylogenetic relationships based on multilocus sequence data and the COI phylogeny (Johnson et al. 2017). Structural variation is, however, present. The COX2 gene varies in length among taxa (range: 999-1443bp). We also observe the previously described noncoding structural variation, hypothesized to be the control region, between the $tRNA^{Trp}$ or $tRNA^{His_2}$ and ND6 loci (Liu et al., 2015; Ozawa et al., 2017). However, we were unable to resolve this region with the current sequence data. All protein-coding genes in the mitochondrial genomes were screened for selection using the adaptive branch-site random effects likelihood method. The COX2 gene exhibited evidence for selection, however, the hypervariable region contained many ambiguous codon alignments. Removal of the ambiguously aligned regions rendered the tests not significant for this gene. Therefore, we conclude that no unambiguous evidence of selection was detected acting on the mitochondrial genomes examined in this study.

Symbiont capture

Examination of the mitochondrial genome phylogeny (Fig. 2A), the symbiont genome phylogeny (Fig. 2b and 2c), and genome-wide symbiont protein divergences (>99%) revealed a striking pattern of a recent symbiont replacement in the *P. extenta* host lineage. The host mitochondrial lineages are divergent, yet the symbiont lineages are nearly identical. The donor lineage appears to be *A. diagonalis*. It is noteworthy that these clams were both collected from sites in Monterey Canyon.

Gene content analysis

In the following sections, symbiont lineages are referred to by the previously erected genera for this group, *Candidatus Vesicomyosocius* and *Candidatus Ruthia* followed by host species names. The genus *Candidatus Vesicomyosocius* refers to the symbionts found in hosts referred to as the *gigas* clade, whereas *Candidatus Ruthia* refers the symbionts found in all other hosts (see Table 1). The symbionts found in both *A. diagonalis* and *P. extent*a are referred to as *C.V. diagonalis-extenta*. These designations, as we argue in the following sections, are supported by both functional genomic and evolutionary characteristics of the groups.

Gene content varied among symbiont lineages. Sequence-based comparisons to the *C.R. magnifica* genome confirm the observation of Newton et al. (2008) that many genes present only in *C.R. magnifica* and not in the other lineages are hypothetical genes with unknown function. A dissimilatory nitrate reductase operon flanked by MogA and MoaD is present in all lineages. This operon is functional in the *C.V. okutanii* genome (Kuwahara et al. 2007; Newton et al. 2008) but has been reduced to approximately 1.8kb in the *C.R. magnifica* genome. Here, we confirm the presence of a functional operon in all *C. Vesicomyosocius* symbionts. The *C.R. Pliocardia* sp. symbiont lineage (2.3kb), *C.R. pacifica* (3.2kb), *C.R. rectimargo* (3.5kb), *C.R. southwardae* (11kb) and *C.R. phaseoliformis* (10kb) lineages appear to contain a degenerate remnant of this operon, an inference based on premature stop codons, frame shifts, and loss of genes.

Within the *C. Vesicomyosocius* symbionts, gene content did vary among genomes. Many of the genes present in the previously sequenced *C.V. okutanii* genome but lost in one or more other *C. Vesicomyosocius taxa* are hypothetical genes, so predicting the potential functional consequences of their loss is not possible. Nevertheless, a few genes of known function were lost. Ribosomal RNA small subunit methyltransferase D, involved in transcriptional regulation, was lost from the *C.V. gigas* and *C.V. soyoae (kilmeri)* lineages, and the *C.V. gigas* lineage also lost ribosomal RNA small subunit methyltransferase E. The *C.V. diagonalis-extenta* lineage has lost a peptidoglycan-associated outer membrane lipoprotein, a gene whose homologues are involved in immune responses to pathogens in many taxa. Indeed, a few membrane-associated genes have been lost within the clade,

including the membrane-bound lytic murein transglycosylase B and D precursors.

The *C.V. southwardae* symbiont has the most fragmented genome. Contig coverage for this genome is characteristically high, approximately 150x. The genome contains 21 mobile element proteins, two of which share 85% and 90% protein identity to *Bathymodiolus septemdierum* symbiont transposases (IS5/IS1182 and IS256 families, respectively), and others that share up to 80% protein identity with transposases from the SUP05 group of bacteria. Thus, the fragmentation of this assembly is due to structural repeat variation that cannot be resolved with the current short reads.

Genome structure and recombination

Genome structures among symbionts were variable. The *C.V. soyoae (kilmeri)* lineage shares the same genome structure previously reported for *C.V. okutanii* (Kuwahara et al. 2007). The *C.V. diagonalis-extenta* symbiont possesses one inferred inversion, and the *C.V. gigas* symbiont contains three inversions relative to the *C.V. okutanii* genome. The *C.R. rectiomargo* symbiont genome appears to have a single inversion relative to *C.R. magnifica*, whereas the *C.R. pacifica* symbiont genome possesses two inversions. The *C.R. Pliocardia* sp. symbiont possesses a single inversion relative to *C.R. magnifica*. We did not attempt to infer the number of inversions of the *C.R. phaseoliformis* and *C.R. southwardae* symbiont genomes due to the fragmented nature of assemblies. Applying phylogenetic likelihood methods to individual gene alignments, we detect a large amount of recombination among symbiont lineages, though recombination is not randomly distributed among lineages. We observe no recombination between members of the *C. Vesicomyosocius* and *C. Ruthia* genera, but recombination is occurring within these genera (Fig. 2b and 2c). Fifty genes (7%) exhibited within-gene recombination in the full dataset. The average symbiont phylogeny exhibits a low percentage of loci supporting the placement of the *C.R. phaseoliformis* and *C.R. southwardae* and *C.R. Pliocardia* sp. symbiont lineages (Fig. 2B and 2C), which is due to the conflicting phylogenetic signal among genes in these lineages.

Tests for selection

The number of symbiont genes that passed the inclusion criteria to be screened for selection (see methods) included 606 loci. The application of the adaptive branch-site random effects likelihood method yielded 107 genes with significant evidence of selection along branches in the phylogeny. Of these 107 genes, 11% exhibited intra-gene topological incongruence (reciprocal topology rejection using the KH test on breakpoints identified with GARD). Within locus recombination has the potential to create false positives in phylogenetic selection tests (Pond et al. 2015). Excluding loci with significant evidence of intra-locus recombination (see Table S2), a total of 16% of loci exhibited unambiguous evidence for episodic diversifying selection. These loci represent a variety of cellular functions (Table 2) and selection is distributed throughout the evolutionary history of the group (Table 3).

Seventy-seven of the loci that exhibited unequivocal evidence of selection were assigned to SEED categories (Table 2, Table S2). The most abundant categories include genes involved in DNA metabolism and translation. Nine genes involved in amino acid biosynthesis of 5 amino acids (Arg, Gly, Ser, Thr, Trp) exhibited evidence of selection. Several large subunit and small subunit ribosomal proteins were under selection (*n*=6; L1p, L3p, L4p, L7/L12, S2p, S3p), as well as several Class I and Class II tRNA aminoacetylation genes. Two genes involved in tRNA methylation were under selection (tRNA (Guanine 37-N1) methyltransferase, and tRNA (guanine 46-N7-)-methyltransferase). Several DNA replication and repair genes were under selection (ATP-dependent DNA helicase *pcrA*, DNA topoisomerases I and IV (subunits A and B), DNA gyrase subunit A, and DNA polymerase III subunits gamma and tau) as well as an RNA helicase (*RhlE*). Together, these results suggest widespread selection on genes involved in DNA metabolism, transcription and translation. In addition, a few metabolic genes that are central to the chemosynthetic role of these symbionts showed evidence for selection. Several genes involved in sulphur assimilation and metabolism were under selection (*DsrP*, *SoxB*, cobyrinic acid A,C-diamide synthase, *Bcp*-type thiol peroxidase, and a sulfur deprivation response regulator). One gene involved in carbon fixation (Rubisco activation protein: *cbbO*), and three genes involved in electron donating/accepting reactions (*CcoG*, *RnfB*, *RnfC*)

were under selection. A single gene involved in central carbohydrate metabolism (*GlcE*) and a single gene involved in ammonia assimilation also exhibited evidence of selection ([Protein-PII] uridylyltransferase).

Discussion

Intracellular symbioses, and in particular, maternal co-transmission of symbionts and mitochondria is expected to alter evolutionary pressures on the symbiont lineages. These effects are hypothesized to include: 1) a reduction in N_e due to transmission bottlenecks; 2) reduced recombination and horizontal gene transfer; 3) accelerated evolutionary rate; 4) shifts in selection pressures due to a change from an extracellular to an intracellular environment. Previous studies of vesicomyid hosts and symbionts have shown a general pattern of co-speciation indicative of maternal transmission (Distel, Felbeck and Cavanaugh 1994; Peek et al. 1998). In one clam species, approximately 400 symbionts are estimated to be transmitted through each egg (Ikuta et al. 2016). Horizontal transmission occasionally occurs in the group (Stewart, Young and Cavanaugh 2008; Decker et al. 2013), which leads to occasional instances of recombination among symbiont lineages (Stewart, Young and Cavanaugh 2009). Though one study has shown evidence for an elevated substitution rate in these symbionts compared to environmentally acquired symbionts (Peek et al. 2000), this study is the first to provide evidence for positive selection in this group. We find that episodic diversifying selection on protein coding genes affects a large number of loci across all lineages examined, and has been a strong force throughout the evolutionary history of this group. Furthermore, many of the loci exhibiting selection are involved in DNA metabolism, transcription and translation as well as chemosynthetic metabolism, reflecting the diversity of selective pressures experienced by intracellular symbionts.

Genome size and gene content are similar but variable across the symbionts examined in this study, though the *C. Vesicomyosocius symbionts*, characterized by the smallest genomes, seems to be more stable in this regard than other lineages. The *C. Vesicomyosocius* symbionts possess a slightly higher AT content than the other lineages examined (Stewart et al 2009). The largest two genomes (and the most fragmented) belonged to the

C.R. southwardae and *C.R. phaseoliformis* symbionts. Stewart et al. (2008) hypothesized that the phylogenetically inferred recombination between *C.R. rectimargo* (*V.* sp. mt-II in their study) and *C.R. southwardae* symbionts (MAR in their study) might actually be occurring with the *C.R. phaseoliformis* symbiont. *C.R. phaseoliformis* is a close relative of *C.R. southwardae*, but its presence in the North Pacific supports this more likely scenario. Therefore, Stewart et al. (2008) suspected that their observations were an artefact of incomplete taxonomic sampling. Indeed, this study confirms their prediction. Within the gene set experiencing episodic diversifying selection, *C.R. phaseoliformis* is often basal to the symbionts of the *Calyptogena* host taxa and is occasionally more closely related to *C.R. rectimargo*. The fragmented nature of the *C.R. phaseoliformis* and *C.R. southwardae* genome assemblies likely reflect the proclivity of theses symbionts to recombine.

These results confirm the presence of an apparently functional dissimilatory nitrate reductase operon in all of the members of *C. Vesicomyosocius* as observed in the *C.V. okutanii* genome. The *C. Ruthia* symbiont genomes contain a degenerate remnant of this operon. Thus, we hypothesize that members of *C. Vesicomyosocius* are potentially capable of using nitrate as a terminal electron acceptor, which might be a significant adaptation to ecological niches in which the hosts are found. Indeed, comparative physiological studies of *C. pacifica* and *P. soyoae (kilmeri*) reveal that *P. soyae* possesses a lower H_2S consumption rate and sulfide-binding ability (Goffredi and Barry 2002), and is generally exposed to lower environmental O_2 levels (Barry et al. 1997). Additionally, *P. soyae* exhibits a lower O_2 consumption relative to sulfide consumption in whole clam respiration experiments (Goffredi and Barry 2002). The utilization of nitrate by the symbiont could explain these observations (Newton et al., 2008), and might be a general characteristic of the *gigas* host clade. We observe no instances of recombination between members of the *C. Vesicomyosocius* and *C. Ruthia* symbiont genera, despite the fact that some of the host taxa co-occur. This observation might be indicative of strong host-symbiont epistatic interactions constraining genetic transfer between these groups. Indeed, FISH hybridization revealed a low level of symbiont transfer between sympatric hosts of *Archivesica chuni* and *Chrisitineconcha regab*, and data from several

symbiont genes collected from these samples support the hypothesis of no genetic exchange between the two symbiont lineages (Decker et al., 2013). The authors conclude that even though horizontal association is occurring at a low rate, it appears to be inconsequential for these hosts. Decker et al (2013) hypothesize that niche specialization due to the *A. chuni* symbiont using nitrate as an electron acceptor could explain their observations. Our results provide further evidence for the general hypothesis that host niche diversification produces evolutionary constraints on the horizontal acquisition of symbionts and the potential for symbiont genetic exchange. Our observation of a symbiont replacement in the *P. extenta* host and widespread historical recombination among *C. Ruthia* symbiont genera, however, suggests that horizontal acquisition might be an on-going and significant source of ecological and evolutionary innovation for the hosts, both through horizontal gene acquisition and symbiont competition.

These results provide the first estimates of the relative contributions of selection and recombination/gene transfer shaping this vertically transmitted symbiosis. Surprisingly, we find a significant number of loci exhibiting evidence of episodic diversifying selection affecting protein sequences. We observe a surprisingly large amount of recombination between lineages, but recombination is not occurring at equal rates between all taxa. Indeed, the *C. Vesicomyosocius* symbionts, associated with unique host physiology and the presence of an apparently functional dissimilatory nitrate reductase operon, do not appear to recombine with other taxa outside of this clade. Additionally, we find evidence for a recent whole genome replacement between two different host genera within the *gigas* clade, suggesting that a selective sweep has recently occurred in the *P. extenta* host. Despite the maternal co-transmission of symbionts and mitochondria, the mitochondrial lineage of *P. extenta* was not apparently affected by this recent symbiont genome replacement. This observation might support the environmental acquisition hypothesis of Stewart et al. (2008). Perhaps the symbiont genetic effective size is large enough to allow symbiont sweeps through individuals and populations without similarly affecting mitochondrial genomes through hitchhiking (but see Hurst and Jiggins 2005). Symbiont transmission effective size or its variability among individuals is unknown for most clam taxa,

including *P. extenta*. Our mitochondrial data, however, confirm variability in the COX2 mitochondrial gene for these taxa. Hurtado et al. (2003) provide evidence for linkage disequilibrium at the population level between mitochondrial and symbiont genes in '*C.' magnifica*. Selective sweeps driven by the symbiont lineage could have contributed to the fixation of slightly deleterious mitochondrial mutations in this species. Further population genomic sampling in '*C.' magnifica* could help to resolve whether on-going selective sweeps within the symbiont genome of this vent species might be affecting mitochondrial diversity through genetic hitchhiking.

Finally, our results highlight the influence of metacommunity ecological processes in shaping the evolutionary history of this and other symbioses. Symbiont loci that have experienced episodic diversifying selection among members of these taxa not only involve genetic exchange among co-occurring genera within the pliocardiin vesicomyids, but also, in one instance, appear to have been drawn from the larger pool of sulfur-oxidizing bacteria inhabiting these environments. Therefore, the geographical distribution and connectivity of diverse pliocardiin host taxa as well as the distribution and connectivity of microbial and viral populations acting as reservoirs of microbial genetic diversity likely determine the ultimate trajectory of adaptive evolution in this and other symbiotic systems.

Materials and Methods

Specimens were collected between 1996 and 2004 over eight research expeditions (Table 1). Depths of sampling locations ranged from 650–3550m. Samples were dissected aboard ship and then frozen at -80C or were frozen whole at -80C. DNA was extracted from symbiont bearing gill tissue using the DNAEasy extraction kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. Host species identification was initially confirmed by sequencing the host mitochondrial COI using vesicomyid-specific primers (Peek et al. 1997). Alignments and phylogenetic reconstruction of host COI for initial host verification were conducted in MEGA v.7 (Kumar et al. 2016).

All genomes were sequenced in-house on a MiSeq instrument. Genomic DNA libraries were prepared using the KAPA Hyperplus Library Preparation kit (KAPA Biosystems, Wilmington, MA, US) according to kit

instructions. Read quality of genomic data was assessed using FastQC (Andrews 2010). Symbiont and host mitochondrial assemblies were constructed using Velvet (Zirbeno and Berney 2008), manually optimizing for *k*-mer size distribution and read depth. Some assemblies were also constructed using the read mapping and assembly functions in Geneious version 10.1.3 (Kearse et al. 2012). Symbiont assemblies were generated de novo from mixed host and symbiont data yielding N50s ranging from 37 kbp to 1.0 Mbp (Table S1). Contigs were considered symbiont based on binning by GC content and read depth. Given the quite high and consistent symbiont coverage (>150x, Table S1), we are very confident that all contigs binned in this manner represent actual symbiont contigs and are not contaminates from the surface of the gill. Final assemblies for samples ranged from 1–84 contigs. The total genome sizes ranged from 1.0–1.6Mb (Table S1). The largest vesicomyid symbiont genome reported previously is the 1.2Mb *R. magnifica* genome (Newton et al. 2006), whereas the smallest is the 1.0Mb *V. okutanii* genome (Kuwarhara et al 2007). Mitochondrial genome annotations were produced manually with the aid of previously annotated mitochondrial genomes (Liu et al. 2016, Ozawa et al., 2017) in Geneious. The RAST pipeline was used to generate symbiont annotations (Aziz et al. 2008).

Symbiont and mitochondrial gene alignments and phylogenetic testing were performed using the hyphy phylogenetics package (Pond et al. 2005). Codon-based alignments for each gene were produced with MUSCLE (Edgar 2004). Progressive Mauve (Darling et al. 2010) and LASTZ (Harris 2007) were used to identify large-scale structural differences among genomes. GARD along with the KH test (Kishino and Hasegawa 1989) were used to identify breakpoints and contiguous non-recombining gene segments within genes (Pond et al. 2006). Episodic diversifying selection on individual lineages was identified using the adaptive Branch-site Random Effects Likelihood method (aBSRel, Smith et al. 2015). Gene length must be at least 300 bp, a gene must be present in all taxa examined, and it must be present in a single copy in all genomes to be included in the set of genes tested for selection. All genes included in the gene set are bidirectional best blast hits. Multiple test corrections were applied using the *qvalue* package (Bass et al. 2015) in R (R Core Team 2017) where appropriate.

Acknowledgements

The Monterey Bay Aquarium Research Institute kindly provided samples from the Vrijenhoek collection for this study. I thank the crews of the ships and pilots involved in the collections for this study without whose efforts this work would not have been possible. The manuscript benefitted from comments from C. Breusing, S. Johnson, H. Ruhl, R.C Vrijenhoek and S Thatje. This work was supported by NERC National Capability funding. N. Pratt and A. Baylay contributed to sequencing at the National Oceanography Centre Genomics Facility.

References

Andrews S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.

Bass AJ, Dabney A and Robinson D (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.8.0, http://github.com/jdstorey/qvalue

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

Cary SC, Giovannoni SJ. (1993) Transovarial inheritance of endosymbiotic bacteria in clams inhabiting deep-sea hydrothermal vents and cold seeps. Proc Natl Acad Sci USA. 90: 5695–5699.

Darling A, Mau B, Perna N. (2010) Progressive Mauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. Plos ONE. June 2010;5(6):1-17.

Distel, D. L., Felbeck, H., & Cavanaugh, C. M. (1994) Evidence for phylogenetic congruence among sulfur-oxidizing chemoautotrophic bacterial endosymbionts and their bivalve hosts. Journal of Molecular Evolution, 38, 533–542.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-1797

Endow K, Ohta S. (1990) Occurrence of bacteria in the primary oocytes of vesicomyid clam *Calyptogena soyoae*. Mar Ecol Prog Ser. 64:309–311.

Felsenstein J (1974) The evolutionary advantage to recombination Genetics 78: 737-756

Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Hill WG, Robertson A, (1966) The effect of linkage on limits to artificial selection. Genetical Research. 8: 269–294.

Hurst GDD, Jiggins FM (2005) Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc R Soc Lond B 272:1525–1534.

Hurtado LA, Mateos M, Lutz RA, Vrijenhoek RC. (2003) Coupling of Bacterial Endosymbiont and Host Mitochondrial Genomes in the Hydrothermal Vent Clam *Calyptogena magnifica*. Applied and Environmental Microbiology. 69(4):2058-2064.

Johnson SB, Krylova EM, Audzijonyte A, Sahling H, Vrijenhoek RC (2017) Phylogeny and origins of chemosynthetic vesicomyid clams, Systematics and Biodiversity, 15:4, 346-360

Ikuta T, Igawa K, Tame A, et al. (2016) Surfing the vegetal pole in a small population: extracellular vertical transmission of an “intracellular” deep-sea clam symbiont. Royal Society Open Science. 3(5):160130.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., & Drummond, A. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28(12), 1647-1649.

Kishino, H., M.H. Hasegawa (1989) Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

Kumar S., Stecher G., Tamura K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol; 33 (7): 1870-1874.

Kuwahara H, Yoshida T, Takaki Y, et al. (10 co-authors). (2007) Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. Curr Biol. 17: 881–886.

Liu H.,Cai S., Zhang H., and R. C. Vrijenhoek (2016) Complete mitochondrial genome of hydrothermal vent clam *Calyptogena magnifica*. Mitochondrial DNA Part A Vol. 27, Iss. 6,2016

McCutcheon JP and Moran NA (2012) Extreme genome reduction in symbiotic bacteria. *Nature reviews. Microbiology*, *10*(1), 13.

Moran NA. (1996) Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. Proc Natl Acad Sci USA. 93:2873–2878.

Moran NA. (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell. 108:583–586.

Moran NA, Mira A. (2001) The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. 2:research0054.1–0054.12

Muller HJ (1964) The relation of recombination to mutational advance. Mutational Research 1:2-9

Newton ILG, Woyke T, Auchtung TA, et al. (16 co-authors). (2007) The Calyptogena magnifica chemoautotrophic symbiont genome. Science. 315:998–1000.

Newton, I. L. G., Girguis, P. R., & Cavanaugh, C. M. (2008) Comparative genomics of vesicomyid clam (Bivalvia: Mollusca) chemosynthetic symbionts. BioMed Central Genomics, 9, 585.

Ozawa G, Shimamura S, Takaki Y, Yokobori S, Ohara Y,Takishita K, Maruyama T, Fujikura K, Yoshida T. (2016) Updated mitochondrial phylogeny of Pteriomorph and Heterodont Bivalvia, including deep-sea chemosymbiotic Bathymodiolus mussels, vesicomyid clams and the thyasirid clam Conchocele cf. bisecta. Marine genomics. 10.1016/j.margen.2016.09.003.

Papke RT, Ward DM. (2004) The importance of physical isolation to microbial diversification. FEMS Microbiol Ecol. 48: 293–303.

Peek AS, Vrijenhoek RC, Gaut BS (1998) Accelerated evolutionary rate in sulfur-oxidizing endosymbiotic bacteria associated with the mode of symbiont transmission. Mol Biol Evol. 15 (11): 1514-1523.

Peek AS, Feldman RA, Lutz RA, Vrijenhoek RC. (1998) Co-speciation of

chemoautotrophic bacteria and deep-sea clams. Proc Natl Acad Sci USA. 95:9962–9966.

Peek AS, Gaut BS, Feldman RA, Barry JP, Kochevar RE, Lutz RA, Vrijenhoek RC. (2000) Neutral and nonneutral mitochondrial genetic variation in deep-sea clams from the family Vesicomyidae. J Mol Evol. 50:141–153.

Pond SLK, Frost SDW, Muse SV; HyPhy: hypothesis testing using phylogenies. Bioinformatics 2005; 21 (5): 676-679.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk and Simon DW Frost. 2006. Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. Molecular Biology and Evolution 23(10): 1891-1901

Silva FJ, Latorre A, Moya A. 2003. Why are the genomes of endosymbiotic bacteria so stable? Trends Genet. 19:176–180.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Pond SLK; Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Mol Biol Evol 2015; 32 (5): 1342-1353.

Stewart FJ, Young CR, Cavanaugh CM. 2008. Lateral symbiont acquisition in a maternally transmitted chemosynthetic clam endosymbiosis. Mol Biol Evol. 25:673–687.

Stewart FJ, Cavanaugh CM. 2009. Pyrosequencing analysis of endosymbiont population structure: co-occurrence of divergent symbiont lineages in a single vesicomyid host clam. Environmental Microbiology 11(8), 2136–2147

Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson A-S, Wernegreen JL, Sandström JP, Moran NA, Andersson SGE (2002) 50 Million Years of Genomic Stasis in Endosymbiotic Bacteria. Science. 28 Jun 2002: 2376-2379.

Wernegreen JJ, Funk DJ. 2004. Mutation exposed: a neutral explanation for extreme base composition of an endosymbiont genome. J Mol Evol. 59:849–858.

Wernegreen JJ. 2005. For better or for worse: genomic consequences of intracellular mutualism and parasitism. Curr Opin Genet Dev. 15:572–583.

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–829.

Table 1: Sampling information and accession numbers for taxa in this study

Species	Accession	Dive No.[b]	Locality[c]	Lat	Long	Depth (m)	Year
"gigas clade"[a]							
Archivesica gigas	XXXXXX	T 548	Guaymas Basin (v)	27.3400	-111.2700	1754	2003
Archivesica diagonalis	XXXXXX	**T 405**	Monterey Canyon (s)	36.2254	-122.8852	3455	2002
Phreagena soyoae (kilmeri)	XXXXXX	V 2059	Monterey Canyon (s)	36.7762	-122.0842	985	2001
Phreagena extenta	XXXXXX	T 406	Monterey Canyon (s)	36.6088	-122.4366	2889	2002
Vesicomyosocius okutanii (symbiont)	AP009247.1						
Phreagena okutanii (mtDNA)	AP014555						
Ruthia magnifica (symbiont)	CP000488.1		East Pacific Rise (v)				
"Calyptogena" magnifica (mtDNA)	KR862368		East Pacific Rise (v)				
Pliocardia sp. Blake Ridge	XXXXXX	A 3710	Blake Spur (s)	32.4948	-76.1847	2155	2001
Abyssogena southwardae	XXXXXX	A 3133	Logatchev, MAR (v)	14.7532	-44.9805	3038	1997
Abyssogena phaseoliformis	XXXXXX	TVG	Aleutian Trench (s)	54.3050	-157.2133	3550	1996
Abyssogena phaseoliformis (mtDNA)	AP014557						
Abyssogena mariana (mtDNA)	LC126311						
Calyptogena pacifica	XXXXXXX	V 2555	Monterey Canyon (s)	36.7739	-122.0488	650	2004
Calyptogena rectimargo	XXXXXXX	V 2338	Monterey Canyon (s)	36.6816	-122.1197	1540	2003

[a] group of taxa referred to as "gigas clade" in the text; [b] Abbreviations for vehicles: T=Tiburon, V=Ventana, A=Alvin, TVG=TV Grab; [c] v=vent, s=seep

Table 2: SEED categories and subcategories of loci exhibiting evidence for episodic diversifying selection according to the aBSRel analysis.

Category	Count	Subcategory	Count
Amino Acids and Derivatives	11	Alanine, serine, and glycine	1
		Arginine; urea cycle, polyamines	2
		Aromatic amino acids and derivatives	4
		Branched-chain amino acids	1
		Lysine, threonine, methionine, and cysteine	3
Carbohydrates	3	Central carbohydrate metabolism	1
		CO2 fixation	1
		One-carbon Metabolism	1
Cell Division and Cell Cycle	1	Cell Division and Cell Cycle	1
Cofactors, Vitamins, Prosthetic Groups, Pigments	8	Biotin	1
		Cofactors, Vitamins, Prosthetic Groups, Pigments	1
		Folate and pterines	1
		NAD and NADP	1
		Pyridoxine	1
		Quinone cofactors	2
		Riboflavin, FMN, FAD	1
DNA Metabolism	6	DNA repair	1
		DNA replication	4
		DNA uptake, competence	1
Fatty Acids, Lipids, and Isoprenoids	3	Isoprenoids	1
		Phospholipids	2
Membrane Transport	1	Membrane Transport	1
Nitrogen Metabolism	1	Nitrogen Metabolism	1
Nucleosides and Nucleotides	8	Detoxification	2
		Purines	5
		Pyrimidines	1
Protein Metabolism	17	Protein biosynthesis	15
		Protein degradation	1
		Protein folding	1
Respiration	4	Electron accepting reactions	1
		Electron donating reactions	2
		Respiration	1
RNA Metabolism	9	RNA processing and modification	9
Stress Response	1	Oxidative stress / Heat Shock	1
Sulfur Metabolism	4	Inorganic sulfur assimilation	2
		Sulfur Metabolism	2

Table 3: Lineages under selection identified by aBSREL.

e	d	g	k	o	A	B	M	m	p	r	Branches
						x					21
x	x	x	x	x							17
					x						16
									x	x	14
							x				9
										x	9
								x			7
						x	x				7
					x				x	x	7
x	x										6
									x		5
					x	x	x				5
					x		x				4
					x	x	x		x	x	3
					x	x			x	x	3
		x									2
			x	x							2
x	x		x	x							1
			x								1
					x		x	x			1
					x			x			1
				x							1
							x		x	x	1
						x	x	x	x	x	1
						x	x	x			1
					x	x	x	x			1
							x	x	x	x	1
					x		x		x	x	1
x	x	x									1
						x			x	x	1
					x	x					1

"x" denotes members of the smaller set in the bipartition. Abbreviations: e= *extent* host; d= C.*V. diagonalis*; g=*C.V. gigas*; k= *C.V. soyoae (kilmeri)*; o= *C.V. okutanii*; A= *C. Ruthia phaeseoloformis* (Aleutian); B= *C.V. Plioicardia* sp. (Blake Ridge); M= *C.R. southwardae* (MAR); m= *C.R. magnifica*; p= *C.R. pacifica*; r= *C.R. rectimargo*.

Figures

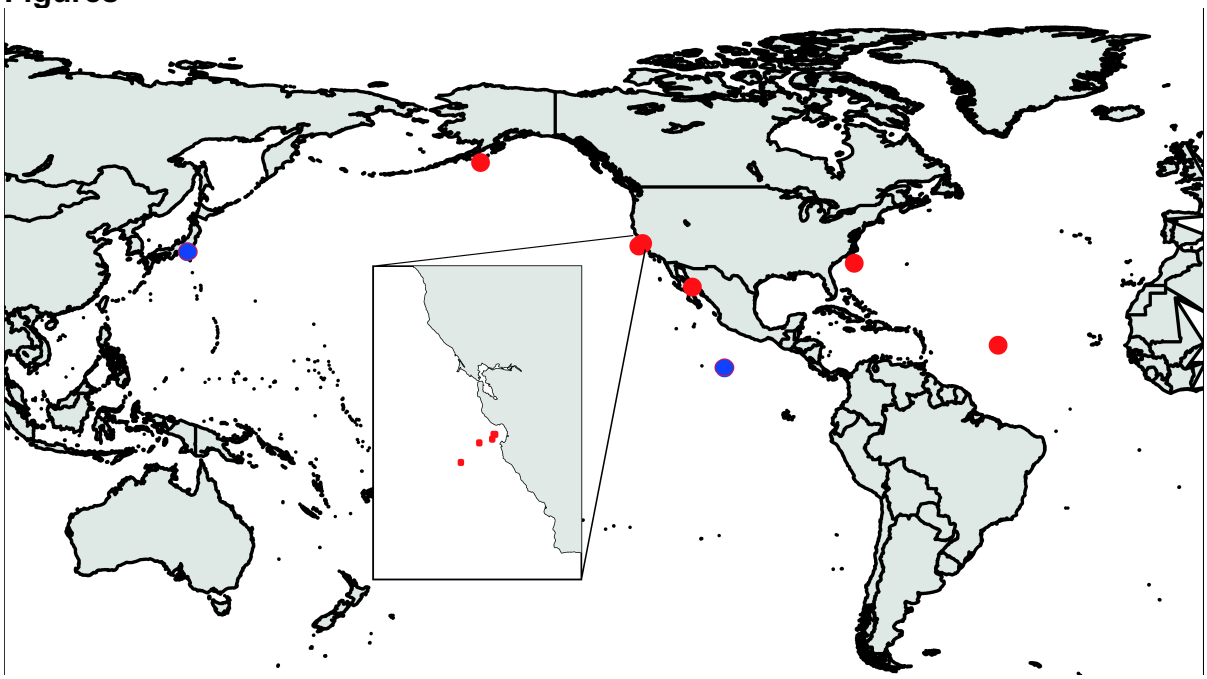

Figure 1: Sampling locations. Red circles: genomes sequenced in this study; blue circles: previously published genomes. Inset depicts samples collected from varying depths in Monterey Bay.

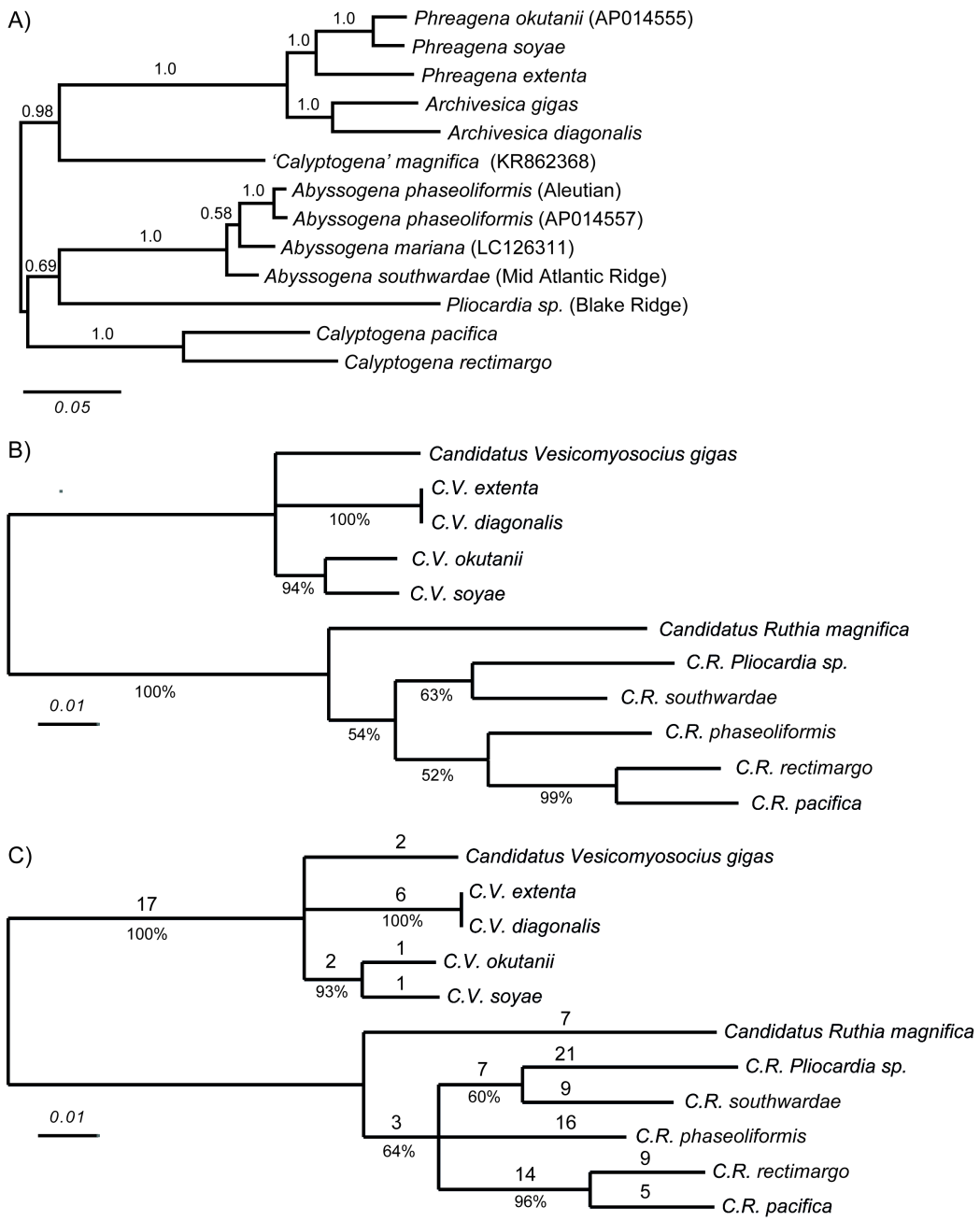

Figure 2: Host and symbiont ML phylogenomic estimates. A) ML phylogeny based on all mitochondrial protein-coding genes. Numbers above branches are bootstrap support values. B) Average ML phylogeny among all 546 symbiont loci exhibiting no evidence of episodic diversifying selection. Percentages below branches are the proportion of loci whose ML phylogeny estimate supports that taxon bipartition, and branch lengths are the averages of ML estimates across all genes. C) Average phylogeny among all 107 symbiont loci with at least one lineage exhibiting evidence of episodic diversifying selection. Branch lengths and percentages blow branches are the same as in (B). Numbers above branches are the number of loci exhibiting selection on that branch.

Supplemental Tables

Table S1: Annotation statistics for symbiont genomes in this study.
Table S2: Genes exhibiting significant evidence of selection.