# Extreme genomic makeover: evolutionary history and niche adaptation of maternally-transmitted clam symbionts

Short title:

Evolution and ecology of maternally-transmitted symbiont genomes

Maëva Perez1\*†, Corinna Breusing2†, Bernard Angers1, Roxanne A. Beinart2, Yong-Jin Won3, and C. Robert Young4

†Authors contributed equally

1Department of Biological Sciences, Université de Montréal, Montreal, Canada

2Graduate School of Oceanography, University of Rhode Island, Narragansett, USA

3Division of EcoScience, Ewha Womans University, Seoul, South Korea

4National Oceanography Centre, Southampton, UK

\* Corresponding author

  E-mail: [maeva.perez@umontreal.ca](mailto:maeva.perez@umontreal.ca)

# Abstract

Vertical transmission of bacterial endosymbionts is accompanied by virtually irreversible gene loss that results in a progressive reduction in genome size. While the evolutionary processes and patterns of genome reduction have been well described in some terrestrial symbioses, they are less understood in marine systems where vertical transmission is relatively rare. The association between deep-sea vesicomyid clams and chemosynthetic Gammaproteobacteria is one example of maternally inherited symbioses in the ocean. Here, we assessed the contributions of drift, recombination and natural selection to genome evolution in two extant vesicomyid symbiont clades by comparing 15 representative symbiont genomes (1.017–1.586 Mb) to those of closely related bacteria and the hosts’ mitochondria. Our analyses suggest that drift is the main driver of genome evolution in vesicomyid symbionts, but imply a role of natural selection and inter-specific recombination in maintaining symbiont functional integrity and creating divergent patterns of gene conservation. For instance, the two symbiont clades show putative functional differences in sulfide physiology, anaerobic respiration, and dependency on environmental vitamin B12, which likely reflect adaptations to different ecological niches available to each symbiont group. Overall, these results contribute to our understanding of the eco-evolutionary processes shaping reductive genome evolution in vertically transmitted symbioses.

# Introduction

Heritable symbioses with intracellular bacteria are observed across the eukaryotic domain of life [1]. These symbioses have profound consequences for both host and symbiont, ranging from alterations of sex-ratios in a population, to providing nutrients that are otherwise unavailable in the host’s habitat, to enhancing resistance to predators and pathogens [1–3]. Regardless of host or habitat, vertical transmission of bacterial lineages from parent to off-spring inevitably leads to genome reduction in the symbionts (i.e., reductive genome evolution, RGE) [3, 4]. This process is thought to be a consequence of successive bottleneck events during transovarial transmission, which decrease the effective population size and genetic diversity of endosymbiont populations [5, 6]. The genetic homogeneity of vertically transmitted symbionts is further amplified by reduced rates of horizontal gene transfer (i.e., homologous recombination between bacterial strains or species). The higher the degree of host restriction, the lower the opportunities for these genetic exchanges [7]. As a consequence, genetic drift increases relative to selection in these taxa, favoring the accumulation of slightly deleterious mutations (Muller’s ratchet) [6, 8, 9]. The pea aphid/*Buchnera* symbiosis and several other well studied insect/bacteria models largely support this neutral hypothesis [10]. At the same time, other metazoan/microbial symbioses highlight the importance of selection in shaping RGE. For instance, Red Queen dynamics are predicted to maintain specificity and the functioning of cyto-nuclear interactions between host and symbiont [3]. In addition, symbiont traits that are beneficial for the host are likely to experience increased selective pressures, while selection may be relaxed on genes that are functionally redundant [11]. Thus, differences in gene content among related symbionts can reveal how host-symbiont pairs diverged in their ecological niches over evolutionary time. For example, divergence in plant host use between insect species is evident in the biosynthetic pathways encoded in the genomes of their obligate endosymbionts [11]. Ultimately, niche differentiation mediated by differential gene loss has the potential to be a significant driver of host evolution through ecological speciation, and host community structure through habitat partitioning between host-symbiont pairs. Despite its importance for both ecological and evolutionary processes, there is still a significant gap in our understanding of the selective processes influencing patterns of genome reduction in divergent vertically transmitted bacterial endosymbionts. This is especially true for the heritable endosymbionts of marine organisms, since vertical transmission is less common among aquatic symbioses [1].

Relatively strict vertical transmission of bacterial endosymbionts has been observed in deep-sea clams of the family Vesicomyidae (subfamily Pliocardiinae) [12], providing an opportunity to examine neutral and selective processes shaping RGE in the marine environment. Vesicomyid clams represent the most diverse group of deep-sea bivalves, with 173 described species present in reducing habitats ranging from hydrocarbon seeps on continental margins to hydrothermal vents on mid-ocean ridges [13–15]. All symbiont-bearing taxa are nutritionally dependent on their chemosynthetic gammaproteobacterial partners, which derive chemical energy from the oxidation of reduced sulfur compounds to produce nutrition for their hosts [16–18]. It is assumed that symbiont capture was a single event that happened before their radiation about 45 Mya [13, 19], an acquisition that is much more recent than that of well-studied terrestrial symbioses (~ 100–200 Mya) [20, 21]. Based on ribosomal sequence data vesicomyid clam symbionts are classified into two divergent clades: Clade I (associated with hosts of the *gigas*-group, including the nominal genera *Akebiconcha*, *Archivesica*, *Laubiericoncha* and *Phreagena*), and Clade II (associated with all other lineages of vesicomyid hosts) [13, 22]. Topological congruences between host mitochondrial and symbiont phylogenetic trees suggest that symbionts co-evolve with their hosts [12], although disruptions of these relationships have been reported due to infrequent horizontal transmission events that allow for homologous recombination and lateral gene transfer between bacterial lineages [23–28]. Previous analyses of one representative symbiont lineage from each clade (*Candidatus* Ruthia magnifica for Clade I and *Candidatus* Vesicomyosocius okutanii for Clade II) suggest that the processes underlying RGE are still operating in vesicomyid symbionts and that the genomes of Clade I symbionts are in an advanced state of reduction compared to Clade II [29]. Indeed, *Ca.* Ruthia magnifica and *Ca.* Vesicomyosocius okutanii possess intermediate genome sizes (1.16 and 1.02 Mbp, respectively) and levels of AT enrichment (66% and 68%, respectively) compared to other host-restricted symbionts, while contrasting levels of gene decay and GC content for 10 housekeeping genes were observed across their respective clades [30].Variations in host affiliation and genome reduction between symbiont clades do not appear to be driven by adaptation to different broad-scale habitat types, as host species from both clades have been found at hydrothermal vents and hydrocarbon seeps and often co-occur at the same locality [13, 18, 31, 32]. However, limited genetic data suggest that representatives of the two symbiont clades differ in physiological characteristics related to nitrate reduction and sulfur metabolism, which may affect microhabitat exploitation [18, 31], and could, thus, influence patterns of gene conservation. In fact, niche partitioning has been linked to patterns of gene loss in a variety of marine and freshwater bacteria [33, 34].

In this study, we aimed to assess the contributions of neutral and selective processes to RGE in vesicomyid clam symbionts by comparing their genome characteristics to those of their hosts’ mitochondria and free-living bacterial relatives. We test the hypothesis that genetic drift is the main driver of RGE in these symbionts and determine to what extent selection has shaped the genetic makeup of the symbionts throughout their evolutionary history.

# Materials and Methods

Detailed methods are available in the Supplementary Material.

## *Sample collection and genome reconstruction*

## New mitochondrial and symbiont genomes for nine species of vesicomyid clams were sequenced in this study, while genomes for another four species were retrieved from previous publications [7, 29, 35–42]. Bacterial relatives of the SUP05 clade (*Bathymodiolus thermophilus* symbiont [Won et al. unpubl.], *Ca.* Thioglobus autotrophicus [38]) were selected as outgroups (Figure 1, Table S1). Quality-trimmed symbiont Illumina reads were assembled *de novo* with Velvet [43], Spades [44] or Geneious [45]. Symbiont contigs were scaffolded and circularized through read mapping and reassembly of contig ends. Mitochondrial genomes were assembled with MitoBim [46] using as seed a set of initial contigs constructed with the read mapping and assembly functions in Geneious. Mitochondrial and symbiont genomes were annotated with GeSeq [47] and Rast [48], respectively (Table S2). Pseudogenes in the symbiont genomes were identified with PseudoFinder (<https://github.com/filip-husnik/pseudofinder/>). Genome similarities and taxonomic affiliations were assessed with FastANI [49] and GTDB-Tk [50].

## *Comparative genomics*

Sequence homology between symbiont genomes was inferred through assessment of positional orthology as well as orthogroup identification with OrthoFinder [51] (Table S3, S4). ProgressiveMauve [52] and Grimm [53] were used to identify large-scale structural differences among mitochondrial and symbiont genomes based on 13 and 716 conserved protein-coding genes, respectively. Phylogenetic trees were produced from these gene sets in MrBayes [54] using a GTR nucleotide substitution rate with a Gamma + I rate variation across sites. Concordance among tree topologies was assessed with Bucky [55]. Pairwise synonymous substitution rates for the concatenated mitochondrial and symbiont core genomes were computed using the Maximum-Likelihood method [56] implemented in the Biopython [57] codonalign package.

## *Selection analyses*

Episodic diversifying selection on individual lineages was identified based on non-recombining core protein-coding genes using aBSRel [58], with corrections for multiple testing based on the Holm-Bonferroni procedure (alpha = 0.05). Changes in the strength of selection on core protein-coding genes were inferred through quantifications of codon usage bias [59] and phylogenetic hypothesis testing with Relax [60]. In addition, we used Fubar [61] and Meme [62] to assess signatures of pervasive and episodic site-specific positive selection in 17 candidate genes that showed marked differences in presence/absence or duplication patterns between the two symbiont clades.

## *Data availability*

Symbiont genomes (CP060680–CP060688, JACRUR000000000, JACRUS000000000) and associated raw reads are available at the National Center for Biotechnology Information under BioProject PRJNA641445. Genome annotations and metabolic reconstructions can be found on the Rast webserver through the guest login (login: guest, password: guest). Host mitochondrial *COI* sequences and genomes have been deposited in GenBank under accession numbers MT894120–MT894130 and MT947381–MT947391, respectively. Genome alignment files and all scripts used in this study are available at <https://github.com/maepz/VesicSymb_Evolution>.

# Results

## *Host mitochondrial and symbiont genomes and phylogenies*

Host mitochondrial genomes examined in this study possess identical gene orders and contents within the vesicomyid taxa. The genome-wide mitochondrial phylogeny is congruent with the known host phylogenetic relationships based on multilocus and *COI* sequence data [13] (Figure 2). Structural variation is, however, present. Although we were not able to circularize the new mitochondrial genomes, our data confirm the previously described noncoding structural variation between the *tRNATrp*/*tRNAHis\_2*and *ND6* loci and indicate length variation in the *COX2* gene among taxa (1005–1452 bp, Figure S1).

Intra-host symbiont populations were genetically homogeneous with frequency distributions of genetic variants typical of monoclonal populations (Table S5, Figure S2). Genome size, GC content and number of intact protein-coding regions for the 15 vesicomyid symbiont assemblies investigated in our study ranged from 1.02–1.59 Mb, 31–37% and 896–1455, respectively (Table S5), with Clade I symbionts having consistently lower values for these genomic characteristics than Clade II symbionts. Following initial nomenclature, the symbiont lineages are referred to by the previously erected genera for the two groups, *Candidatus* Vesicomyosocius for Clade I, and *Candidatus* Ruthia for Clade II, followed by host species names [36, 37, 41]. This classification at the genus level is consistent with the phylogenetic definition based on *16S* rRNA identity (inter-genus identity < 95%) [63], clustering based on average nucleotide identity and alignment fraction [64] (Figure S3), taxonomic classification based on the Genome Taxonomy Database (Table S6) and genetic isolation between the two symbiont clades (see below).

Examination of the mitochondrial and symbiont phylogenies (Figure 2, S4) shows good concordance for all lineages except one. The symbiont lineages of *Ca*. V. diagonalis and *Ca*. V. extenta are nearly identical, whereas their respective host mitochondrial lineages are divergent. Genome size, GC content and gene composition also support the close phylogenetic relationship between *Ca*. V. diagonalis and *Ca*. V. extenta (Figure S3–S5, Table S5). The donor lineage in this recent symbiont replacement appears to be *Archivesica diagonalis*, which co-occurs with *Phreagena extenta* at hydrocarbon seep sites in Monterey Canyon. Within the symbionts, signatures of elevated substitution rates are evident on the branch leading to Clade I. Relative to the branch leading to Clade II, the Clade I ancestor branch is longer in the phylogeny of the symbionts compared to the mitochondria (Figure 2). Furthermore, the symbiont pairs across the Clade I-Clade II bipartition have significantly higher divergence than the others even when controlled for host divergence (1 < dSmito < 2; Figure S6).

## *Symbiont genome structure and recombination*

The *B. thermophilus* symbiont and *Ca.* T. autotrophicus shared about 1 Mbp of their genomes with the clam symbionts, with at least 22 and 3 inversion events being present relative to the *Ca.* R. magnifica reference, respectively (Figure S7).

Genome structure among the clam symbionts was also variable (Figure 2, S7). Three distinct inversions compared to the *Ca.* R. magnifica genome were found in the genomes of *Ca*. V. okutanii, *Ca*. V. gigas, and the monophyletic group composed of *Ca.* R. fausta, *Ca.* R. pacifica and *Ca.* R. rectimargo. In addition, inversions between the *tufA* and *tufB* paralogs, hotspots for chromosomal inversions [65], seem to have happened multiple times throughout the symbiont phylogeny (Figure 2, S7). Evidence for intra-host structural variation was found within *Ca.* R. phaseoliformis and *Ca.* R. southwardae.

Bayesian concordance analysis detected a large amount of recombination among symbiont lineages, though recombination is not randomly distributed. We observe no recombination between members of Clade I and II, but recombination is occurring within these two genera(Figure 2). Relatively little topological concordance was found in Clade II, with 37 different topologies being necessary to fully represent the diversity of conflicting phylogenetic signals compared to only 11 in Clade I (Figure S8). Within Clade I, conflict arises from the uncertainty of the position of *Ca.* V. gigas and *Ca.* V. marissinica(Figure 2). Within Clade II, only the grouping of *Ca.* R. fausta, *Ca.* R. rectimargo and *Ca.* R. pacifica is supported by all gene trees, while the positions of all other species have low support.

## *Patterns of gene conservation*

The outgroup genomes contained many large (> 5kb) contiguous sections that were not found in the symbionts. These genomic islands were mostly composed of unannotated genes and mobile elements (transposases, integrases, prophage genes), but also genes related to heavy metal tolerance and anti-viral defense (in the *B. thermophilus* symbiont) as well as motility and nitrogen metabolism (in *Ca*. T. autotrophicus) (Table S4).

The symbionts of Clade I and Clade II possessed essentially a subset of the genes found in the outgroup lineages. Many genes unique to the vesicomyid symbiont lineages appeared to be pseudogenes resulting from the degeneration of ancestral homologs. Patterns of pseudogenization were notably more prevalent and variable in the genomes of Clade II than Clade I symbionts (Figure S5). In many instances, homologous regions within the Clade I symbiont genomes were instead characterized by large deletions. Genes found in the outgroup bacterial genomes were most conserved within the genomes of *Ca.* R. southwardae, *Ca.* R. phaseoliformis and *Ca.* R. pliocardia. Among theClade IIsymbionts, gene degeneration was most pronounced in *Ca.* R. magnifica, which possessed a conservation pattern closer to that of the Clade I lineages (Figure S4, S5).

Both symbiont clades shared a core genome related to chemoautotrophic metabolism, but showed notable differences in presence/absence, duplication and degeneration patterns for genes related to a diversity of other metabolic processes (Supplementary Results, Figure S5, Table S4). For instance, the genomes of Clade I and Clade II symbionts encoded different types of methionine synthase. While Clade I symbionts contained genes for the cobalamin-dependent homocysteine methyltransferase *metH* and associated genes for cobalamin (precursor) transport and conversion (*btuM*, *btuR/cobA*), Clade II symbionts contained the cobalamin-independent version of this enzyme (*metE*) along with its transcriptional activator (*metR*). However, all symbiont lineages lacked pathways for *de novo* cobalamin biosynthesis. Genomes of both symbiont clades also differed in the presence of operons for dissimilatory (*narGHIJ*: Clade I) and assimilatory (*nasA*: Clade II) nitrate reductases, genes for putative nickel transporters (*hupE*) and nickel-dependent enzymes (*gloA*), as well as genes involved in glyoxylate regeneration (*icl*) and transcriptional repression of certain ribonucleotide reductases (*nrdR*) (only in Clade II). Surprisingly, *nasA* was annotated as pseudogene in almost all Clade II lineages and *Ca.* T. autotrophicus. This is possibly a misclassification as functional expression of *nasA* is observed in deep-sea SUP05 populations [66]. Alternatively, this gene might be in an early stage of pseudogenization as all variants were >74% of the average hit length. An operon encoding cysteine dioxygenase type I (*cdo*) and an aspartate aminotransferase superfamily protein was exclusively found in Clade I. The aminotransferase has similarity to cysteine sulfinic acid decarboxylase (*csad*) and possesses the same substrate recognition motif as the protein described from the *B. azoricus* symbiont (W1aa19S2aaC3; GenBank: SEH86284). Unlike their Clade II congeners, the genomes of almost all Clade I symbionts were characterized by a duplication event in the sulfide:quinone oxidoreductase type I gene (*sqrI*).

## *Genome-wide patterns of relaxed and intensified selection*

Both symbiont clades showed a reduction in codon usage bias (Figure 3A, B) and dN/dS rate-class extremes (Figure 3C) compared to the outgroup (Figure 3C), indicative of a genome-wide decline in the efficacy of natural selection, i.e., a reduction in selective constraint (Table S7, S8). Relaxation was in the range of that observed in insect endosymbionts and appeared exacerbated in Clade I (Figure 3B, C) [59]. Genes exhibiting intensified and relaxed selection in the clam symbionts represented a multitude of metabolic categories, although some functions were predominantly affected by shifts in selection regimes (Figure S9). Genes under relaxed selection were mostly involved in protein, amino acid and nucleoside/nucleotide metabolism, cell division and cell cycle, whereas genes under intensified selection were largely associated with respiration and sulfur metabolism.

## *Patterns of positive selection in core and clade-specific genes*

114 protein-coding core genes exhibited evidence for episodic diversifying selection along branches in the phylogeny (Table S9). Selection appears to be distributed throughout the evolutionary history of the group (Figure 4), acting mostly on the branches discriminating the outgroup, Clade I, and Clade II, as well as on the branches of the *B. thermophilus* symbiont and *Ca.* T. autotrophicus. Eighty-five percent of loci that showed signs of selection were classified into SEED categories (Figure 4). These loci were equally represented among cellular functions of the core genome except for a few categories along the branches separating the two symbiont clades. These categories included genes involved in nucleotide synthesis and defense (Figure 4).

With the exception of *gloA, narI*, and *narJ*, all investigated metabolic genes that were differentially preserved between clades showed evidence of pervasive or episodic site-specific diversifying selection that affected structural or functional regions in the encoded proteins (Table S10). Pervasive positive selection based on Fubar analyses was observed at 1–3 sites across the entire phylogeny in ten of the 17 genes tested: *btuM*, *btuR*, *csad*, *hupE*, *icl*, *metR*, *narG*, *narH, nasA*, *sqrI*. In addition, episodic positive selection based on Meme analyses was detected at 1–7 sites along a proportion of branches in all tested genes except for *btuR*, *gloA, narI*, and *narJ*. In the case of *cdo*, *csad* and *nasA*, these episodes of site-specific selection seemed to have mostly occurred in the ancestral lineages as no evidence for selection was found along the extant symbiont branches (Table S10).

# Discussion

## *RGE is ongoing and driven by neutral processes*

Current insights into the evolutionary processes shaping RGE in maternally inherited symbionts stem mostly from well-studied terrestrial insect-bacteria associations, where genetic drift has been shown to be the dominant force driving patterns of endosymbiont gene loss [10, 67]. Our comparisons of 15 vesicomyid symbiont genomes with those of relatives that contain a free-living phase suggest that neutral processes might play an equally important role in marine vertically transmitted symbioses. As in other models of recently acquired bacteria [68, 69], gene content differed greatly between vesicomyid symbiont genomes, indicating that the different lineages are independently losing genes. The presence of structural variation and varying degrees of gene degeneration within the symbiont genomes imply that these lineages have not yet reached a stable streamlined state compared to many insect endosymbionts [70], as suggested previously [29]. All clam symbionts exhibited a reduced GC%, decrease in codon usage bias, and a genome-wide trend of relaxation in selective pressures relative to the selected bacterial outgroup lineages. Taken together, these observations support the nearly neutral theory of RGE, driven by a reduction of effective population size in these taxa [71].

In agreement with previous findings [23–25, 28], we detected no recombination between Clade I and Clade II symbionts, even though some of the host taxa co-occur [25, 31]. These findings imply that there is enough molecular and ecological divergence between the two clades for clonal interference and/or strong host-symbiont epistatic interactions to constrain symbiont exchange [3]. The two groups are also discriminated based on measures of genomic relatedness and functional genomic traits. Thus, our results support the nomenclature initially put forward by Newton et al. [37] and Kuwahara et al. [41] classifying the symbionts from Clade I and Clade II into two distinct bacterial genera, *Ca.* Vesicomyosocius and *Ca.* Ruthia. For clarity, we will keep referring to these two genera as Clade I and Clade II in the rest of the discussion.

## *RGE is exacerbated in non-recombining symbionts*

Symbionts belonging to Clade I appear to be in a more advanced state of RGE than those of Clade II, as their genomes are smaller and lower in GC%, possess fewer genes and pseudogenes, exhibit less codon usage bias and are in general more homogeneous. Patterns of gene conservation suggest that much of the loss in this group happened after its speciation but before its radiation, a relatively short period of roughly 20 Myrs [12, 13]. Together with increased substitution rates on its diverging branch these results imply that the ancestral Clade I lineage experienced an acute episodic acceleration of RGE.

Based on genome-wide levels of topological disagreement, horizontal gene transfer through inter-specific homologous recombination is widespread among symbionts of Clade II but it is almost absent in Clade I. A reduction of the rate of infection by environmental symbionts and/or drift-driven loss of the recombination machinery [22] may have strongly reduced the rate of genetic exchange across symbiont lineages within Clade I, thereby setting this genus on a divergent evolutionary path.

Recombination can alter rates of evolution due to Hill-Robertson interference [72] by randomizing the associations between mutations that otherwise would be in linkage disequilibrium. These effects can vary, depending on the population size, mutation rate and recombination rate. In small populations, deleterious alleles fix through drift, reducing the mean fitness of the population (i.e., Muller’s ratchet; [73]). Additionally, selection against deleterious alleles can remove linked beneficial alleles from the population (i.e., background selection), reducing the rate of adaptation [74, 75]. Selection for beneficial alleles can also cause linked deleterious alleles to fix in a population (i.e., hitchhiking; [76, 77]). Low rates of recombination increase Hill-Robertson interference, reducing the rate of adaptation ([72, 78]). For example, in *Drosophila*, the rate of adaptive amino acid substitution is positively correlated to both recombination rate and mutation rate ([79]). Furthermore, when recombination is completely absent, clonal interference can reduce the rate of adaptation ([80]).

Strong linkage disequilibrium forces whole genomes to sweep in populations that lack capabilities for genetic exchange. Hence, loss of potential for homologous recombination should favor symbiont replacement in cases where the divergence between native and foreign symbionts is low enough to avoid host-symbiont incompatibilities. In fact, we find multiple examples of symbiont replacement among lineages of Clade I. For instance, individual clams of the species *P. extenta* have acquired the symbionts of the sympatric species *A. diagonalis.* Likewise, Breusing et al. [26] found a population of *A. gigas* carrying the symbionts of the host species *P. soyoae.* Symbiont replacement occurs in several vertically transmitted symbioses [81–84] and is speculated to constitute a mechanism for escaping the evolutionary rabbit hole caused by Muller’s ratchet [3].

Despite the lack of recombination machinery in Clade I, two lineages in this genus, *Ca.* V. gigas and *Ca.* V. marissinica, showed signs of genetic exchange. Perhaps recombination in these species is mediated via symbiont-derived host-encoded proteins. Evidence of symbiont gene transfer to the host nuclear genome was recently found by Ip et al. [36] who identified ancestral symbiont gene homologs in the *A. marissinica* host genome. Taken together, these observations support a crucial role of genetic exchange, and its associated machinery, in maintaining symbiont genome integrity [7] and moderating the ecological consequences of increased clonality. Examination of the symbionts at the population-level, both within and across individual hosts, will help to decipher the contributions of host physiology, symbiont fitness, cytonuclear incompatibilities, and rates of lateral symbiont transfer to their evolution.

## *Selective processes might be tied to genetic and environmental contexts*

Although our data indicate that genetic drift is the predominant force controlling RGE in vesicomyid symbionts, fractions of the symbiont genomes are affected by natural selection. Selective constrains are predicted to primarily act on genes involved in host-symbiont interactions as these molecular pathways must experience reciprocal adaptations to persist through speciation and niche exploitation. Diversifying selection affecting genes that play a role in host-symbiont communication, such as lipopolysaccharides and peptidoglycans, was observed in divergent clades of several terrestrial obligate and facultative endosymbionts [85–87]. Surprisingly, our data do not confirm these predictions and instead suggest a pervasive pattern of positive selection affecting all cellular functions. These results could indicate that the accumulation of slightly deleterious mutations in the symbiont genomes enhances selective pressures for compensatory mutations. Cases of such mutations have been observed in cellular organelles and bacterial endosymbionts of insects and fungi [88–90]. Alternatively, these patterns might be reflections of the early state of RGE or rapid adaptations of the symbionts to the host intracellular habitat [67].

Strong functional contrasts in gene loss between bacteria with free-living phase and symbionts as well as between both symbiont clades further suggest a role of niche differentiation in shaping symbiont genome composition through RGE, which has consequences for ecological processes, like habitat use, and evolutionary processes, like host speciation. Genes enabling bacteria to face the challenges of a free-living environment, such as metal detoxification, anti-viral defense and inter-species competition, were not conserved in the clam symbionts, while both clades differed in retention of genes affecting diverse metabolic traits, including the dependency on enzyme cofactors, the potential for anaerobic respiration, and sulfur physiology.

*Symbiont clades show putative differences in physiology and ecological niche*

### Clade I and II symbionts encoded different, convergently evolved types of methionine synthase [91], which vary in their requirement for vitamin B12. Measurements of cobalamin in deep-sea environments are very challenging and we, therefore, do not currently have informative data on vitamin B12 concentrations experienced by the clams. However, as both clades appeared unable to synthesize cobalamin *de novo* (like their eukaryotic hosts), these findings indicate that the environmental availability of vitamin B12 has the potential to be an important factor influencing the distribution of these taxa. Cobalamin independence in Clade II may offer an obvious selective advantage, by allowing these symbioses to exploit niches that would otherwise be inaccessible. By contrast, the requirement for exogeneous vitamin B12 (or its derivatives) might limit the range of (micro)habitats Clade I-based associations can colonize, unless cobalamin is acquired from a (currently unknown) secondary symbiont. Despite this potential cost, the retention of a cobalamin-dependent methionine synthase in Clade I likely provides an evolutionary benefit, given that MetH has a fifty-fold higher catalytic rate constant than MetE and thus enables faster growth [91]. Indeed, comparative measurements of vesicomyid growth rates suggest that species hosting symbionts of Clade I typically grow faster than species with symbionts of Clade II, despite a less efficient sulfur physiology [92]. Since growth is influenced by a variety of factors [92], it is possible that the enzymatic differences in methionine biosynthesis among symbiont clades contribute to an accelerated anabolism in Clade I-based associations, although this remains to be experimentally tested. The preservation of cobalamin-dependent enzymes as a result of conferred physiological advantages appears to be common across the eubacterial domain [93]. A recent genomic analysis showed that 86% of bacterial lineages have at least one cobalamin-dependent enzyme despite the existence of a cobalamin-independent alternative, and that many of these lineages rely on vitamin B12 production from other microbes in their environment [93]. The importance of vitamin B12 for the biology of the two symbiont groups is also evident in the fact that onlyClade II symbiontsencode a transcriptional repressor (NrdR) for the ribonucleotide reductase NrdAB, a key enzyme that controls the synthesis of DNA [94]. In Clade I symbionts, expression of NrdAB is probably regulated by cobalamin, which has been shown to repress NrdAB transcription through riboswitches [95]. There is evidence thatthe two symbiont cladesdiffer in their requirements for other enzyme cofactors, such as nickel. The genomes of Clade II symbionts encoded a specific transporter for nickel uptake, and most of these lineages contained at least one confirmed nickel-dependent enzyme (glyoxalase I [96]), all of which were absent in the symbionts of Clade I.

Our data suggest that Clade I and Clade II symbionts further show notable differences in encoded gene clusters for nitrate reduction. These results extend previous findings by 11 genomes [18], confirming that these patterns are truly clade-specific characteristics. The use of NarGHIJ for nitrate reduction in Clade I likely enables these symbioses to inhabit hypoxic environments [18], since the use of nitrate as an electron acceptor would reduce the symbiont’s requirement for oxygen and, consequently, allow the host to utilize the limited oxygen present without competition with the symbiont. These assumptions agree with field observations that indicate that clam species hosting Clade I symbionts typically occupy micro-niches with higher levels of hydrogen sulfide (and, thus, presumably lower oxygen) than those hosting Clade II symbionts in the same habitat [31]. Niche partitioning based on environmental sulfide levels has also been suggested by physiological comparisons of *P. soyoae* and *C. pacifica*, which imply that *P. soyoae* symbionts have lower sulfide oxidation capacities and therefore require higher H2S concentrations for chemosynthesis [31]. This could be a consequence of a less efficient sulfide metabolism in *Ca.* V. soyoae (Clade I) resulting from an increased load of deleterious mutations in accordance with its more advance state of genome reduction compared to *Ca.* R. pacifica (Clade II). The presence of two tandem copies of *sqrI* displaying evidence of concerted evolution suggests increasing gene dosage as compensating mechanism in Clade I. Our data indicate that there might be additional adaptations (or ancestral restrictions) to contrasting sulfide environments between symbiont clades. Only Clade I symbionts encode genes for CDO and putatively CSAD, which are key enzymes in the biosynthesis of taurine/hypotaurine. These non-proteinogenic amino acids are important for sulfide detoxification in a variety of symbiont-bearing invertebrates that inhabit sulfidic environments [97–99]. Considering that Clade I-based associations appear to be prevalent in high-sulfide habitats, it is possible that these symbionts directly or indirectly contribute to H2S tolerance of their hosts. An alternative (though mutually non-exclusive) function of CDO and CSAD could involve replenishment of metabolic intermediates. Vesicomyid symbionts do not possess a complete TCA cycle and must recycle succinate through other means [18]. In Clade II symbionts succinate regeneration occurs via the glyoxylate shunt, while the mechanism in Clade I symbiontsis unclear [18]. Perhaps taurine is further metabolized via taurine dioxygenase, which would generate succinate as end product. If this pathway can be confirmed through physiological experiments, autonomous recycling of succinate by the symbiont could make important contributions to the holobiont’s carbon budget.

**Conclusions**

Our comparative genomic analyses show that patterns of genome reduction in vesicomyid symbionts are primarily shaped by genetic drift and that factors affecting symbiont clonality strongly influence the rate of RGE, with potential consequences for speciation and niche partitioning. However, the pervasive nature of episodic diversifying selection across functional traits in the vesicomyid symbiont genomes suggests that neutral evolutionary processes (genetic drift and recombination) are not the sole drivers of molecular evolution in these taxa. Similarly, differential patterns of gene loss between Clade I and Clade II symbionts reiterate that RGE does not follow a universal trajectory but is a reflection of the ecological and evolutionary context of the respective host-symbiont association. Convergent gene loss and pseudogenization imply common evolutionary pressures for some genes, whereas selection on codons and lineage-specific gene retention imply niche-specific adaptation in others. Future studies linking environmental data with symbiont genomic information will be helpful to obtain further insights into the ecological basis of RGE in the symbionts of deep-sea vesicomyid clams.

**Acknowledgements**

We thank the able captains and crews of the R/Vs *Atlantis*, *Western Flyer* and *Point Lobos* as well as the pilots of the submersibles *Alvin*, *Tiburon* and *Ventana* for supporting the sample collections. The Monterey Bay Aquarium Research Institute kindly provided samples from the Vrijenhoek collection for this study. We further thank N. Pratt and A. Baylay for their contributions to library preparation and high-throughput sequencing at the National Oceanography Centre Genomics Facility. This work was supported by grants of the David and Lucile Packard Foundation (to MBARI), the UK Natural Environment Research Council (grant number NE/N006496/1 to C.R.Y.), the United States National Science Foundation (grant number OCE-1736932 to R.A.B.), the German Research Foundation (BR 5488/1-1 to C.B.), the Natural Science and Engineering Research Council of Canada (grant number 238600 to B.A., Alexander Graham Bell graduate scholarship and Michael Smith Foreign Study Supplements to M.P.) and National Capability funding to the National Oceanography Centre (grant number NE/R015953/1). The symbiont genome of *Bathymodiolus thermophilus* was sequenced as part of a project titled ‘Understanding the deep-sea biosphere on seafloor hydrothermal vents in the Indian Ridge’ funded by the Korean Ministry of Oceans and Fisheries (grant number 20170411 to Y.J.W.). Bioinformatic analyses were in part performed on ComputeCanada HPC clusters.

# Competing Interests

The authors declare no competing interests.

# References

1. Russell SL. 2019 Transmission mode is associated with environment type and taxa across bacteria-eukaryote symbioses: a systematic review and meta-analysis. *FEMS Microbiol Lett* **366**, fnz013.

2. Fisher RM, Henry LM, Cornwallis CK, Kiers ET, West SA. 2017 The evolution of host-symbiont dependence. *Nat Commun* **8**, 15973. (doi:10.1038/ncomms15973)

3. Bennett GM, Moran NA. 2015 Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. *PNAS* **112**, 10169–10176. (doi:10.1073/pnas.1421388112)

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

5. Nei M, Maruyama T, Chakraborty R. 1975 The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10. (doi:10.1111/j.1558-5646.1975.tb00807.x)

6. Moran NA. 1996 Accelerated evolution and Muller’s rachet in endosymbiotic bacteria. *PNAS* **93**, 2873–2878.

7. Russell SL, Pepper-Tunick E, Svedberg J, Byrne A, Castillo JR, Vollmers C, Beinart RA, Corbett-Detig R. 2020 Horizontal transmission and recombination maintain forever young bacterial symbiont genomes. *PLoS Genet* **16**, e1008935. (doi:10.1371/journal.pgen.1008935)

8. Wernegreen JJ, Moran NA. 1999 Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol Biol Evol* **16**, 83–97.

9. Kuo C-H, Moran NA, Ochman H. 2009 The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19**, 1450–1454. (doi:10.1101/gr.091785.109)

10. Wernegreen JJ. 2011 Reduced selective constraint in endosymbionts: elevation in radical amino acid replacements occurs genome-wide. *PLoS ONE* **6**, e28905. (doi:10.1371/journal.pone.0028905)

11. Hansen AK, Moran NA. 2014 The impact of microbial symbionts on host plant utilization by herbivorous insects. *Mol Ecol* **23**, 1473–1496. (doi:10.1111/mec.12421)

12. Peek AS, Feldman RA, Lutz RA, Vrijenhoek RC. 1998 Cospeciation of chemoautotrophic bacteria and deep sea clams. *PNAS* **95**, 9962–9966. (doi:10.1073/pnas.95.17.9962)

13. Johnson SB, Krylova EM, Audzijonyte A, Sahling H, Vrijenhoek RC. 2017 Phylogeny and origins of chemosynthetic vesicomyid clams. *Syst Biodivers* **15**, 346–360. (doi:10.1080/14772000.2016.1252438)

14. Krylova EM, Sahling H. 2010 Vesicomyidae (Bivalvia): current taxonomy and distribution. *PLoS ONE* **5**, e9957. (doi:10.1371/journal.pone.0009957)

15. Audzijonyte A, Krylova EM, Sahling H, Vrijenhoek RC. 2012 Molecular taxonomy reveals broad trans-oceanic distributions and high species diversity of deep-sea clams (Bivalvia: Vesicomyidae: Pliocardiinae) in chemosynthetic environments. *Syst Biodivers* **10**, 403–415. (doi:10.1080/14772000.2012.744112)

16. Childress JJ, Fisher CR, Favuzzi JA, Sanders NK. 1991 Sulfide and carbon dioxide uptake by the hydrothermal vent clam, *Calyptogena magnifica*, and its chemoautotrophic symbionts. *Physiol Zool* **64**, 1444–1470. (doi:10.1086/physzool.64.6.30158224)

17. Robinson JJ, Cavanaugh CM. 1995 Expression of form I and form II Rubisco in chemoautotrophic symbioses: implications for the interpretation of stable carbon isotope values. *Limnol Oceanogr* **40**, 1496–1502. (doi:10.4319/lo.1995.40.8.1496)

18. Newton IL, Girguis PR, Cavanaugh CM. 2008 Comparative genomics of vesicomyid clam (Bivalvia: Mollusca) chemosynthetic symbionts. *BMC Genomics* **9**, 585. (doi:10.1186/1471-2164-9-585)

19. Peek AS, Gustafson RG, Lutz RA, Vrijenhoek RC. 1997 Evolutionary relationships of deep-sea hydrothermal vent and cold-water seep clams (Bivalvia: Vesicomyidae): results from the mitochondrial cytochrome oxidase subunit I. *Mar Biol* **130**, 151–161. (doi:10.1007/s002270050234)

20. Moran NA, Munson MA, Baumann P, Ishikawa H. 1993 A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc Biol Sci* **253**, 167–171. (doi:10.1098/rspb.1993.0098)

21. Ferri E, Bain O, Barbuto M, Martin C, Lo N, Uni S, Landmann F, Baccei SG, Guerrero R, de Souza Lima S *et al.* 2011 New insights into the evolution of *Wolbachia* infections in filarial nematodes inferred from a large range of screened species. *PLoS ONE* **6**, e20843. (doi:10.1371/journal.pone.0020843)

22. Kuwahara H, Takaki Y, Shimamura S, Yoshida T, Maeda T, Kunieda T, Maruyama T. 2011 Loss of genes for DNA recombination and repair in the reductive genome evolution of thioautotrophic symbionts of *Calyptogena* clams. *BMC Evol Biol* **11**, 285. (doi:10.1186/1471-2148-11-285)

23. Stewart FJ, Young CR, Cavanaugh CM. 2008 Lateral symbiont acquisition in a maternally transmitted chemosynthetic clam endosymbiosis. *Mol Biol Evol* **25**, 673–687. (doi:10.1093/molbev/msn010)

24. Stewart FJ, Young CR, Cavanaugh CM. 2009 Evidence for homologous recombination in intracellular chemosynthetic clam symbionts. *Mol Biol Evol* **26**, 1391–1404. (doi:10.1093/molbev/msp049)

25. Decker C, Olu K, Arnaud-Haond S, Duperron S. 2013 Physical proximity may promote lateral acquisition of bacterial symbionts in vesicomyid clams. *PLoS ONE* **8**, e64830. (doi:10.1371/journal.pone.0064830)

26. Breusing C, Johnson SB, Vrijenhoek RC, Young CR. 2019 Host hybridization as a potential mechanism of lateral symbiont transfer in deep-sea vesicomyid clams. *Mol Ecol* **28**, 4697–4708. (doi:10.1111/mec.15224)

27. Ikuta T, Igawa K, Tame A, Kuroiwa T, Kuroiwa H, Aoki Y, Takaki Y, Nagai Y, Ozawa G, Yamamoto M *et al.* 2016 Surfing the vegetal pole in a small population: extracellular vertical transmission of an ‘intracellular’ deep-sea clam symbiont. *R Soc Open Sci* **3**, 160130. (doi:10.1098/rsos.160130)

28. Ozawa G, Shimamura S, Takaki Y, Takishita K, Ikuta T, Barry JP, Maruyama T, Fujikura K, Yoshida T. 2017 Ancient occasional host switching of maternally transmitted bacterial symbionts of chemosynthetic vesicomyid clams. *Genome Biol Evol* **9**, 2226–2236.

29. Kuwahara H, Takaki Y, Yoshida T, Shimamura S, Takishita K, Reimer JD, Kato C, Maruyama T. 2008 Reductive genome evolution in chemoautotrophic intracellular symbionts of deep-sea *Calyptogena* clams. *Extremophiles* **12**, 365–374. (doi:10.1007/s00792-008-0141-2)

30. Shimamura S, Kaneko T, Ozawa G, Matsumoto MN, Koshiishi T, Takaki Y, Kato C, Takai K, Yoshida T, Fujikura K *et al.* 2017 Loss of genes related to Nucleotide Excision Repair (NER) and implications for reductive genome evolution in symbionts of deep-sea vesicomyid clams. *PLoS ONE* **12**, e0171274. (doi:10.1371/journal.pone.0171274)

31. Goffredi SK, Barry JP. 2002 Species-specific variation in sulfide physiology between closely related vesicomyid clams. *Mar Ecol Prog Ser* **225**, 227–238. (doi:10.3354/meps225227)

32. Cruaud P, Decker C, Olu K, Arnaud-Haond S, Papot C, Le Baut J, Vigneron A, Khripounoff A, Gayet N, Cathalot C *et al.* 2019 Ecophysiological differences between vesicomyid species and metabolic capabilities of their symbionts influence distribution patterns of the deep-sea clams. *Mar Ecol* **40**, e12541. (doi:10.1111/maec.12541)

33. Luo H, Huang Y, Stepanauskas R, Tang J. 2017 Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. *Nat Microbiol* **2**, 1–9. (doi:10.1038/nmicrobiol.2017.91)

34. Baumgartner M, Roffler S, Wicker T, Pernthaler J. 2017 Letting go: bacterial genome reduction solves the dilemma of adapting to predation mortality in a substrate-restricted environment. *ISME J* **11**, 2258–2266. (doi:10.1038/ismej.2017.87)

35. Ozawa G, Shimamura S, Takaki Y, Yokobori S-I, Ohara Y, Takishita K, Maruyama T, Fujikura K, Yoshida T. 2017 Updated mitochondrial phylogeny of pteriomorph and heterodont bivalvia, including deep-sea chemosymbiotic *Bathymodiolus* mussels, vesicomyid clams and the thyasirid clam *Conchocele* cf. *bisecta*. *Mar Genomics* **31**, 43–52. (doi:10.1016/j.margen.2016.09.003)

36. Ip JC, Xu T, Sun J, Li R, Chen C, Lan Y, Han Z, Zhang H, Wei J, Wang H *et al.* 2020 Host-endosymbiont genome integration in a deep-sea chemosymbiotic clam. *Mol Biol Evol* **38(2)**, 502–518. (doi:10.1093/molbev/msaa241)

37. Newton IL, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, Fisher MC, Fontanez KM, Lau E, Stewart FJ, Richardson PM *et al.* 2007 The *Calyptogena magnifica* chemoautotrophic symbiont genome. *Science* **315**, 998–1000. (doi:10.1126/science.1138438)

38. Shah V, Morris RM. 2015 Genome sequence of “*Candidatus* Thioglobus autotrophica” strain EF1, a chemoautotroph from the SUP05 clade of marine Gammaproteobacteria. *Genome Announc* **3**, e01156–15. (doi:10.1128/genomeA.01156-15)

39. Yang M, Gong L, Sui J, Li X. 2019 The complete mitochondrial genome of *Calyptogena marissinica* (Heterodonta: Veneroida: Vesicomyidae): Insight into the deep-sea adaptive evolution of vesicomyids. *PLoS ONE* **14(9)**, e0217952. (doi:10.1371/journal.pone.0217952)

40. Liu H, Cai S, Zhang H, Vrijenhoek RC. 2016 Complete mitochondrial genome of hydrothermal vent clam *Calyptogena magnifica*. *Mitochondrial DNA Part A* **27**, 4333–4335. (doi:10.3109/19401736.2015.1089488)

41. Kuwahara H, Yoshida T, Takaki Y, Shimamura S, Nishi S, Harada M, Matsuyama K, Takishita K, Kawato M, Uematsu K *et al.* 2007 Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. *Curr Biol* **17**, 881–886. (doi:10.1016/j.cub.2007.04.039)

42. Lee Y, Kwak H, Shin J, Kim S-C, Kim T, Park J-K. 2019 A mitochondrial genome phylogeny of Mytilidae (Bivalvia: Mytilida). *Mol Phylogenet Evol* **139**, 106533. (doi:10.1016/j.ympev.2019.106533)

43. Zerbino DR, Birney E. 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–829. (doi:10.1101/gr.074492.107)

44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al.* 2012 SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455–477. (doi:10.1089/cmb.2012.0021)

45. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C *et al.* 2012 Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649. (doi:10.1093/bioinformatics/bts199)

46. Hahn C, Bachmann L, Chevreux B. 2013 Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res* **41**, e129–e129. (doi:10.1093/nar/gkt371)

47. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017 GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res* **45**, W6–W11. (doi:10.1093/nar/gkx391)

48. 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* **9**, 75. (doi:10.1186/1471-2164-9-75)

49. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018 High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114. (doi:10.1038/s41467-018-07641-9)

50. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019 GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* , btz848. (doi:10.1093/bioinformatics/btz848)

51. Emms DM, Kelly S. 2019 OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238. (doi:10.1186/s13059-019-1832-y)

52. Darling AE, Mau B, Perna NT. 2010 progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147. (doi:10.1371/journal.pone.0011147)

53. Tesler G. 2002 GRIMM: genome rearrangements web server. *Bioinformatics* **18**, 492–493. (doi:10.1093/bioinformatics/18.3.492)

54. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012 MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**, 539–542. (doi:10.1093/sysbio/sys029)

55. Larget BR, Kotha SK, Dewey CN, Ané C. 2010 BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* **26**, 2910–2911. (doi:10.1093/bioinformatics/btq539)

56. Goldman N, Yang Z. 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725–736. (doi:10.1093/oxfordjournals.molbev.a040153)

57. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al.* 2009 Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423. (doi:10.1093/bioinformatics/btp163)

58. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015 Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**, 1342–1353. (doi:10.1093/molbev/msv022)

59. Zhang Z, Li J, Cui P, Ding F, Li A, Townsend JP, Yu J. 2012 Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics* **13**, 43. (doi:10.1186/1471-2105-13-43)

60. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015 RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol* **32**, 820–832. (doi:10.1093/molbev/msu400)

61. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013 FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol* **30**, 1196–1205. (doi:10.1093/molbev/mst030)

62. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012 Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8**, e1002764. (doi:10.1371/journal.pgen.1002764)

63. Stackebrandt E, Goebel BM. 1994 Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Evol Microbiol,* **44**, 846–849. (doi:10.1099/00207713-44-4-846)

64. Barco RA, Garrity GM, Scott JJ, Amend JP, Nealson KH, Emerson D. 2020. A genus definition for bacteria and archaea based on a standard genome relatedness index. *mBio* **11**, e02475–19. (doi:10.1128/mBio.02475-19)

65. Hughes D. 2000 Co-evolution of the tuf genes links gene conversion with the generation of chromosomal inversions. *J Mol Biol* **297**, 355–364. (doi:10.1006/jmbi.2000.3587)

66. Anantharaman K, Breier JA, Sheik CS, Dick GJ. 2013 Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *PNAS* **110**, 330–335. (doi:10.1073/pnas.1215340110)

67. Martínez-Cano DJ, Reyes-Prieto M, Martínez-Romero E, Partida-Martínez LP, Latorre A, Moya A, Delaye L. 2015 Evolution of small prokaryotic genomes. *Front Microbiol* **5**, 742. (doi:10.3389/fmicb.2014.00742)

68. Andersson JO, Andersson SG. 1999 Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol* **16**, 1178–1191. (doi:10.1093/oxfordjournals.molbev.a026208)

69. Burke GR, Moran NA. 2011 Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biol Evol* **3**, 195–208. (doi:10.1093/gbe/evr002)

70. Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson A-S, Wernegreen JJ, Sandström JP, Moran NA, Andersson SGE. 2002 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376–2379. (doi:10.1126/science.1071278)

71. 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. (doi:10.1093/oxfordjournals.molbev.a025879)

72. Hill WG, Robertson A. 1966 The effect of linkage on limits to artificial selection. *Genet Res* **8**, 269–294.

73. Muller HJ. 1964 The relation of recombination to mutational advance. *Mutat Res Fundam Mol Mech Mutagen* **1**, 2–9. (doi:10.1016/0027-5107(64)90047-8)

74. Charlesworth B, Morgan MT, Charlesworth D. 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303. (doi:10.1093/genetics/134.4.1289)

75. Charlesworth B. 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res* **63**, 213–227. (doi:10.1017/S0016672300032365)

76. Smith JM, Haigh J. 1974 The hitch-hiking effect of a favourable gene. *Genet Res* **23**, 23–35.

77. Gillespie JH. 2000 Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* **155**, 909–919. (doi:10.1093/genetics/155.2.909)

78. Felsenstein J. 1974 The evolutionary advantage of recombination. *Genetics* **78**, 737–756.

79. Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. 2016 Adaptive evolution is substantially impeded by Hill-Robertson interference in *Drosophila*. *Mol Biol Evol* **33(2)**, 442–455. (doi:10.1093/molbev/msv236)

80. Gerrish PJ, Lenski RE. 1998 The fate of competing beneficial mutations in an asexual population. *Genetica* **102**, 127. (doi:10.1023/A:1017067816551)

81. Pérez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A, Latorre A. 2006 A small microbial genome: the end of a long symbiotic relationship? *Science* **314**, 312–313.

82. Koga R, Moran NA. 2014 Swapping symbionts in spittlebugs: evolutionary replacement of a reduced genome symbiont. *ISME J* **8**, 1237–1246. (doi:10.1038/ismej.2013.235)

83. Sudakaran S, Kost C, Kaltenpoth M. 2017 Symbiont acquisition and replacement as a source of ecological innovation. *Trends Microbiol* **25**, 375–390. (doi:10.1016/j.tim.2017.02.014)

84. Chong RA, Moran NA. 2018 Evolutionary loss and replacement of *Buchner*a, the obligate endosymbiont of aphids. *ISME J* **12**, 898. (doi:10.1038/s41396-017-0024-6)

85. Brownlie JC, Adamski M, Slatko B, McGraw EA. 2007 Diversifying selection and host adaptation in two endosymbiont genomes. *BMC Evol Biol* **7**, 68. (doi:10.1186/1471-2148-7-68)

86. Chong RA, Park H, Moran NA. 2019 Genome evolution of the obligate endosymbiont *Buchnera aphidicola*. *Mol Biol Evol* **36(7)**, 1481–1489 (doi:10.1093/molbev/msz082)

87. Dale C, Moran NA. 2006 Molecular interactions between bacterial symbionts and their hosts. *Cell* **126**, 453–465. (doi:10.1016/j.cell.2006.07.014)

88. Howe DK, Denver DR. 2008 Muller’s Ratchet and compensatory mutation in *Caenorhabditis briggsae* mitochondrial genome evolution. *BMC Evol Biol* **8**, 62. (doi:10.1186/1471-2148-8-62)

89. Castillo DM, Pawlowska TE. 2010 Molecular evolution in bacterial endosymbionts of fungi. *Mol Biol Evol* **27**, 622–636. (doi:10.1093/molbev/msp280)

90. Lambert JD, Moran NA. 1998 Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. *PNAS* **95**, 4458–4462.

91. González JC, Banerjee RV, Huang S, Sumner JS, Matthews RG. 1992 Comparison of cobalamin-independent and cobalamin-dependent methionine synthases from *Escherichia coli*: two solutions to the same chemical problem. *Biochemistry* **31**, 6045–6056. (doi:10.1021/bi00141a013)

92. Barry JP, Kochevar RE. 1998 A tale of two clams: differing chemosynthetic life styles among vesicomyids in Monterey Bay cold seeps. *Cah Biol Mar* **39**, 329–331.

93. Shelton AN, Seth EC, Mok KC, Han AW, Jackson SN, Haft DR, Taga ME. 2019 Uneven distribution of cobamide biosynthesis and dependence in bacteria predicted by comparative genomics. *ISME J* **13**, 789–804. (doi:10.1038/s41396-018-0304-9)

94. Torrents E. 2014 Ribonucleotide reductases: essential enzymes for bacterial life. *Front Cell Infect Microbiol* **4**, 52. (doi:10.3389/fcimb.2014.00052)

95. Borovok I, Gorovitz B, Schreiber R, Aharonowitz Y, Cohen G. 2006 Coenzyme B12 controls transcription of the *Streptomyces* class Ia ribonucleotide reductase nrdABS operon via a riboswitch mechanism. *J Bacteriol* **188**, 2512–2520. (doi:10.1128/JB.188.7.2512-2520.2006)

96. Boer JL, Mulrooney SB, Hausinger RP. 2014 Nickel-dependent metalloenzymes. *Arch Biochem Biophys.* **0**, 142–152. (doi:10.1016/j.abb.2013.09.002)

97. Pruski AM, Fiala-Médioni A. 2003 Stimulatory effect of sulphide on thiotaurine synthesis in three hydrothermal-vent species from the East Pacific Rise. *J Exp Biol* **206**, 2923–2930. (doi:10.1242/jeb.00513)

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

99. Joyner JL, Peyer SM, Lee RW. 2003 Possible roles of sulfur-containing amino acids in a chemoautotrophic bacterium-mollusc symbiosis. *Biol Bull* **205**, 331–338. (doi:10.2307/1543296)

# Figure Captions

**Figure 1** Global distribution of bacterial species compared in this study.

**Figure 2** Genome-wide host mitochondrial (left) and symbiont (right) trees. These phylogenies represent the Bayesian majority-rule consensus of 2, 000 independent trees (GTR + G + I model). Left: Consensus tree and posterior probabilities of the branches from the concatenated alignment of 13 core protein coding genes. Right: Consensus tree from the concatenated alignment of syntenic blocs shared between the symbiont (Clade I: blue; Clade II: green) and outgroup (light red) genomes. Chromosome schemes showing genome inversions and assembly fragmentation are displayed at the end of the branches (blue: inversions between TufA/B paralogs; green, orange and magenta: other inversions). Refer to text for a description of the genome structures. Numbers in red are the genome-wide mean covariance factors, which represent the percentage of non-recombining syntenic blocs supporting each split of the phylogeny.

**Figure 3** Codon usage bias in symbionts and bacteria with free-living phase. A) Codon Deviation Coefficient (CDC) spectra for each genome within the outgroup, yellow: *B. thermophilus* symbiont; red: *Ca.* T. autotrophicus. B) Correlation between the average CDC of outgroup, Clade I (blue) and Clade II (green) based on 555 core genes. Linear regressions are shown. CDC values vary from 0 (no bias) to 1 (maximum bias). C) Selection parameter (k) spectra of core genes for which a significant change in selection was detected by Relax. Note that k is on a log scale. CDC values were significantly lower in Clade I compared to Clade II, and CDC and k values were significantly lower in Clade I and Clade II with respect to the outgroup (paired Wilcoxon-Mann Whitney test p-value < 0.01).

**Figure 4** SEED category distribution of core genes under episodic diversifying selection within phylogenetic clades (A, B, C, D), and on partitioning branches (E, F, G, H). A) Distribution of all non-recombining core genes (dark grey, 555 loci) and loci under selection within the outgroup, Clade I and Clade II (light grey, 114 loci). The number of loci selected within each clade is presented in the inset. B) Genes under selection within the outgroup. C) Genes under selection within Clade II. D) Genes under selection within Clade I. E) Distribution of all non-recombining core genes (dark grey, 555 loci) and loci under selection on all partitioning branches (light grey, 71 loci). The number of loci selected on each branch is presented in the inset. F) Genes under selection on branch a. G) Genes under selection on branch b. H) Genes under selection on branch c. Note that genes may be present in multiple functional categories and multiple clades or branches. SEED categories significantly overrepresented (in red) and underrepresented (in blue) in the groups compared to the core genome are highlighted. Refer to text for further breakdown of these categories. NA: no functional annotation.

# Supplementary Figure Captions

**Figure S1** Multiple sequence alignments for the mitochondrial *cox2* gene.

**Figure S2** Variant frequency distributions for the intra-host symbiont populations sequenced in this study (20 bin histograms). Single nucleotide polymorphism frequencies were computed from the raw symbiont genome coverage for each species. Reads were mapped to the reference with Bowtie2 using the --very-sensitive-local parameter.

**Figure S3** Discrimination of symbiont genomes based on A) functional characteristics (SEED categories) and B) relatedness indices. The two clades segregate largely into two groups, with symbiont genomes of Clade I being more homogenous than those of Clade II.

**Figure S4** Jaccard distance-based neighbor-joining trees established from A) the presence/absence of syntenic blocs (LCBs > 100bp) and B) the presence/absence of positionally orthologous genes. Numbers above branches are bootstrap support values.

**Figure S5** A)Heatmap of gene presence/absence, duplication and pseudogenization patterns in symbiont and outgroup genomes based on Manhattan distances and complete clustering. Clade I, Clade II and their relatives from three separate groups based on these genomic characteristics, although *Ca*. R. magnifica assumes an intermediate position between symbiont clades. The presence of pseudogenes is more pronounced in Clade II compared to Clade I, in agreement with the less advanced state of genome reduction in this symbiont group. Gene duplications are almost completely absent in the symbiont genomes. B) Overview of gene presence/absence, duplication and pseudogenization patterns for genes that were differentially preserved between the two symbiont clades.

**Figure S6** Relationship between symbiont and mitochondrial divergence. For each holobiont species, host and symbiont divergences are expressed as genome-wide pairwise synonymous substitutions rates (dS) in their respective genomes. dS values were estimated from the concatenated alignments of 13 mitochondrial and 555 symbiont protein coding genes. Putative pseudogenes and non-core protein coding genes were excluded from the analyses. 〇 indicates mitochondrial and symbiont genomes isolated from a single individual.

**Figure S7** Genome inversions between *Ca*. R magnifica and other bacteria. Optimal inversion scenarios were computed by GRIMM from the permutation maps exported from the whole-genome alignments.

**Figure S8** Cumulative genome representation by tree topologies as estimated by Bucky.

**Figure S9** SEED category distribution of core genes under relaxed or intensified selection. Insets show the test (bright yellow) and reference (dark yellow) branches for each analysis. A) Distribution of all non-recombining core genes (dark grey, 555 loci) and loci under relaxed (grey, 346 loci) or intensified selection (light grey, 83 loci) within all symbiont clades. B) Genes with significant change in selection intensity in the symbionts compared to the outgroup. C) Genes with significant change in selection intensity in Clade II compared to the outgroup. D) Genes with significant change in selection intensity in Clade I compared to the outgroup. Note that genes may be present in multiple functional categories. SEED categories significantly overrepresented (in red) and underrepresented (in blue) in the groups compared to the core genome are highlighted. Refer to text for further breakdown of these categories. NA: no functional annotation.