**SUPPLEMENTARY MATERIAL**

**Supplementary Methods**

## *Sample collection, DNA extraction, and metagenomic sequencing*

Host (and symbiont) taxa examined in this study were chosen from the deepest diverging lineages within the Vesicomyidae that are distributed globally in the northern hemisphere (Figure 1, Table S1) and are representative of the known host diversity [1]. Mitochondrial and symbiont genome assemblies of *Abyssogena mariana* [2], *A. phaseoliformis* (Japan) [2], *Archivesica marissinica* [3, 4], *Calyptogena fausta* [5], *C. magnifica* [6, 7] and *Phreagena okutanii* [2, 8] were obtained from previous studies. New clam specimens were collected from eleven hydrocarbon seep and hydrothermal vent sites in the Pacific and Atlantic Ocean during research expeditions between 1994 and 2004. Upon recovery of the submersibles, samples were dissected and frozen at –80°C. DNA was extracted onshore from symbiont-bearing gill tissues with the DNeasy Blood & Tissue kit (Qiagen, Hilden, Germany) following manufacturer’s instructions. Barcoded 2x300 bp metagenomic libraries for mixed host and symbiont DNA samples were prepared with the KAPA Hyperplus Library Preparation kit (KAPA Biosystems, Wilmington, MA, US) and sequenced on an Illumina MiSeq system at the National Oceanography Centre (Southampton, UK). Clam host species were identified via mitochondrialcytochrome-c-oxidase I(*COI*) sequencing using vesicomyid-specific primers [9]. Bacterial relatives with free-living phase of the SUP05 clade (*Bathymodiolus thermophilus* symbiont [Won et al. unpubl.], *Ca.* Thioglobus autotrophicus [10]) were selected as outgroups (Table S1).

## *Mitochondrial and symbiont genome reconstruction and annotation*

After initial quality checks with FastQC v0.11.5 [11], reads were adapter-clipped with Trimmomatic v0.36 [12] and assembled *de novo* with Velvet v1.2.10 [13], Spades v3.13.1 [14] or Geneious v10.1.3 [15] using manual optimizations of k-mer size distribution and read depth. Mitochondrial genomes were assembled *de novo* with MitoBim v1.9 [16] using as seed a set of initial contigs constructed with the read mapping and assembly functions in Geneious. Scaffolding of the symbiont genomes was done with Sspace v2.0 [17] and final circularization was performed by re-mapping, extracting and reassembling reads that mapped to the extremities of contigs using Bowtie2 v2.4.2 [18], Samtools v1.12 [19] and Spades, respectively. Mitochondrial genome annotations were produced by GeSeq [20] using Arwen v1.2.3 [21] for tRNA prediction, and manually curated with the aid of previously annotated mitochondrial genomes [2, 6] in Geneious. Symbiont genome assemblies were annotated with Rast v2.0 [22] (Table S2). Genes were classified as pseudogenes with PseudoFinder v1.0 (<https://github.com/filip-husnik/pseudofinder/>) if protein length was <80% of the average length of matches in the RefSeq database. We used this relatively conservative threshold to ensure exclusion of all potential pseudogenes from downstream selection and phylogenetic analyses. Pseudogenes annotated as fragmented were manually curated to distinguish actual fragmentation from functional gene copy number variations. Assembly quality and statistics were assessed with Quast v5.0.0 [23], while taxonomic classification was performed with GTDB-Tk v1.4.0 [24]. Average nucleotide identities and alignment fractions between genomes were calculated with FastANI v1.32 [25].

*Identification of orthologous groups and gene duplication events*

Sequence homology between symbiont genomes was inferred via two independent, complementary methods. First, single-copy core orthologs were identified based on homology and position [26] with the function “getOrthologList” from Mauve v2.4.0 [27], using a minimum identity of 35% and a minimum coverage of 51% (Table S3). Second, broader orthologous gene groups and gene duplications were determined with OrthoFinder v2.5.2 [28] using Mafft v7.310 [29] for multiple sequence alignment, FastTree v2.1.11 [30] for gene tree inference and Blast v2.9.0+ [31] for sequence search. A rooted species tree based on the core positional orthologs was used as prior information. Paralogous groups that originated before the divergence of the first extant clades were split into separate orthogroups (Table S4). An overview figure comparing genomic characteristics and relatedness of symbionts and bacterial relatives was produced with the ComplexHeatmap package in R v4.0.3 [32] using Manhattan distances for clustering.

## *Genetic variation and phylogenomic analyses*

Symbiont and mitochondrial intra-host genetic heterogeneities were estimated from the abundance of single nucleotide polymorphisms (SNPs). For each species, raw reads were mapped to the symbiont and mitochondrial genomes with Bowtie2 (--very-sensitive-local) and SNPs were called with VarScan v2.4.2 [33] using the following filters: --min-coverage 2 --min-reads2 1 --min-avg-qual 28 --min-var-freq 0.01. To remove false positives amongst symbiont SNPs, the “vcf” files were filtered with the accessory *fpfilter* script [33] using default parameters except for two: --max-var-mmqs 150 --max-mmqs-diff 100. To avoid the detection of putative false positive mitochondrial variants, we excluded the control region between ND6 and tRNA-Ala, which could not be fully resolved and contained repeats resulting in dubious mappings of the metagenomic reads. Because of low coverage for the mitochondrial genomes, the *fpfilter* script was not used. ProgressiveMauve v2.4.0 [34] and Grimm v2.01 [35] were used to identify large-scale structural differences among mitochondrial and symbiont genomes. For the host mitochondria, we concatenated alignments of 13 conserved protein-coding genes. For the symbionts and bacterial relatives, we extracted, realigned and concatenated 716 locally collinear blocks (LCBs) longer than 100 bp that were found in all bacterial genomes. Phylogenetic trees were produced from these core genes in MrBayes v3.2.7a [36] using a General Time Reversible (GTR) nucleotide substitution rate with a Gamma + I distribution across sites. The prior for the branch lengths was set to Unconstrained:Exp(50.0). Ten independent MCMC chains were each run for 2,000,000 generations after an initial 100,000 generations burn-in period. Trees were sampled every 10,000 generations to avoid autocorrelation. Parameter optimization for the MCMCs was performed by assessing convergence and mixing of both the continuous parameters of the model and the tree topologies using the R package RWTY v1.0.2 [37]. For the symbionts, additional neighbor-joining trees based on Jaccard distance were built from gene presence/absence patterns.

## *Bayesian concordance analyses*

We used Bucky v1.2 [38] to estimate the proportion of syntenic blocks – defined here as ProgressiveMauve’s Locally Colinear Blocs (LCBs) – supporting each symbiont topology. Putative recombination breakpoints within the core LCBs ≥ 100 bp were identified with Gard v0.2 [39] based on AICc ratio tests and a 5% false positive discovery rate threshold. Input posterior distributions of LCB topologies were each obtained from 2,000 trees generated in MrBayes with the same parameters as for the genome-wide tree construction. Two independent MCMC runs were carried out under the prior assumption that almost all genes shared the same topology (alpha = 0.001). MCMC runs were updated 1,000,000 times after an initial 10% burn-in period. One cold and three heated chains (swapping frequency = 10) were used to improve mixing and convergence of all of the MCMC runs.

## *Host and symbiont evolutionary rates*

We compared host and symbiont evolutionary rates by estimating the genome-wide divergence at synonymous sites between each host-symbiont pair. Nucleic and amino acid sequences of the 13 conserved mitochondrial and 555 non-recombining bacterial core protein-coding genes were extracted with Biopython v1.76 [40]. Amino acid sequences were then aligned with Muscle [41] and reverse translated into codon alignments using the “build” function from the Biopython “codonalign” package. The mitochondrial and bacterial codon-based alignments were each concatenated into two genome-wide alignments with lengths of 12,558 bp and 484,320 bp, respectively. We assessed substitution saturation by plotting transitions and transversions against adjusted F84 genetic distance. Pairwise synonymous substitution rates were computed using the Maximum-Likelihood method [42] implemented in the Biopython codonalign package. The source code was slightly modified to accommodate for ambiguous bases in the mitochondrial genomes.

## *Genome-wide screen for positive selection and changes in selective pressures*

Episodic diversifying selection on individual lineages was identified on the core non-recombining protein-coding genes using the adaptive Branch-site Random Effects Likelihood method (aBSRel v2.2) [43] with corrections for multiple testing based on the Holm-Bonferroni procedure (alpha = 0.05). Changes in the strength of selection were inferred via two independent methods. First, we used the Codon Deviation Coefficient [44] to quantify codon usage bias on all core protein-coding genes because this index does not require *a priori* knowledge of gene expression and is not biased by GC content. Second, we used Relax [45] to detect changes in the strength of selection between group pairs. We compared Clade I, Clade II, and both clades together to the outgroup. To test whether genes under episodic positive, relaxed or intensified (diversifying and purifying) selection represented a random subsample according to SEED categories [46], we estimated the probability of each distribution using the dmvhyper function from the extraDistr v1.8.11 R package (<https://cran.r-project.org/web/packages/extraDistr/index.html>) and compared it to that of 100 distributions obtained from randomly sampling the non-recombining core gene dataset. Fisher’s exact tests [47] were applied to find SEED categories that were over-represented in the genes under relaxed or intensified selection.

## *Tests for site-specific adaptive evolution in metabolic candidate genes*

We assessed signatures of site-specific positive selection in 17 candidate genes that showed marked differences in presence/absence or duplication patterns between the two symbiont clades (see below): vitamin B12 transporter component (*btuM*), cob(I)alamin adenosyltransferase (*btuR*), cysteine dioxygenase type I (*cdo*), putative cysteine sulfinic acid decarboxylase (*csad*), lactoylglutathione lyase (*gloA*), hydrogenase/urease accessory protein (*hupE*), isocitrate lyase (*icl*), methionine synthase and transcriptional activator (*metE*, *metH*, *metR*), dissimilatory/assimilatory nitrate reductase (*narGHIJ*, *nasA*), ribonucleotide reductase regulator (*nrdR*), and sulfide:quinone oxidoreductase type I (*sqrI*). Tests for pervasive and episodic diversifying selection were performed using Bayesian approximation and mixed-effects maximum likelihood approaches implemented in Fubar v2.2 [48] and Meme v2.1.2 [48], respectively. Sequence alignments were partitioned according to recombination breakpoints identified with Gard. Fubar analyses included 5 MCMC chains, with chain lengths of 2,000,000, a burn-in of 1,000,000 and a sample size of 100, while Meme analyses were run with default settings testing 1) the complete symbiont phylogeny and 2) only symbiont branches. Because site-level tests for positive selection are relatively conservative, we chose recommended p-value thresholds of 0.1 for Meme and posterior probability thresholds of 0.9 for Fubar to assess statistical significance [49]. To assess the implications of site-specific selection in the investigated genes we predicted structural and functional features of the encoded proteins with PredictProtein [51].

**Supplementary Results**

*Energy metabolism*

The genomes of all symbiont lineages contained genes for the oxidation of reduced sulfur compounds that serve as energy sources for chemoautotrophic growth [52]. All genomes encoded the sulfur oxidation (SOX) multienzyme pathway (without *soxCD*), the reversible dissimilatory sulfite reductase (rDSR) pathway as well as the adenosine 5’-phosphosulfate (APS) reductase pathway, indicating that these symbiont lineages are able to oxidize sulfide, thiosulfate and/or sulfite for energy production [53, 54]. In addition, all genomes comprised genes for sulfide:quinone oxido-reductase type I and VI (SQR), which can convert sulfide to sulfane sulfur [54]. With the exception of *Ca.* V. soyoae 2 and *Ca*. V. okutanii, the Clade I lineages contained two copies of the gene encoding SQR type I, whereas this gene was present as a single copy in Clade II.

Based on their gene content, it is likely that all symbiont lineages can use a variety of different enzymes to conserve energy via cross-membrane electron transport, including NADH-ubiquinone oxidoreductase (Complex I), SQR, bacterial Rnf complex, cytochrome *bc1* complex (Complex III), terminal cbb3-type cytochrome-c-oxidase (Complex IV) and F0F1-type ATP synthase (Complex V).

*Inorganic carbon fixation and biosynthetic processes*

Members of both clades encoded a form II ribulose bisphosphate carboxylase (cbbM) and other key enzymes for carbon assimilation via the Calvin-Benson-Bassham cycle as well as a complete gene set for the non-oxidative branch of the pentose phosphate pathway. Both symbiont clades lacked the gene for sedoheptulose-bisphosphatase and might instead rely on a reversible pyrophosphate-dependent phosphofructokinase (PPi-PFK) to interconvert between sedoheptulose 1,7-bisphosphate and sedoheptulose 7-phosphate. PPi-PFK is likely also used to catalyze the phosphorylation of fructose-6-phosphate to fructose 1,6-bisphosphate during glycolysis, as the gene for its ATP-dependent homolog was absent in all vesicomyid symbiont genomes [7].

All symbiont lineages have the potential to further metabolize glycolytic intermediates and end products via a partial tricarboxylic acid (TCA) cycle and pentose phosphate pathway to produce precursors for the generation of several macronutrients, coenzymes and nucleotides.

Functional gene copies of oxoglutarate dehydrogenase and fumarate reductase appeared to be missing from all genomes. Both clades contained complete gene sets for the independent biosynthesis of 19 amino acids and a variety of enzyme cofactors, including most vitamins and their derivatives (e.g., coenzyme A, FAD, NAD+), hemes and sirohemes, porphyrins, molybdopterin, ubiquinone and glutathione. The gene encoding homoserine kinase (*thrB*), an essential enzyme in threonine biosynthesis, was missing from all symbiont genomes [7], although it is possible that its function might be performed by a serine/threonine kinase that was present in genomes from both the Clade I and Clade II symbionts. Similarly, a separate gene for histidinol phosphatase involved in histidine biosynthesis was lacking from all symbiont genomes. However, the genomes of symbionts from both clades contained homologs of the *hisB* gene, which encodes a bifunctional imidazoleglycerol-phosphate dehydratase/histidinol-phosphatase. Pathways for the generation of retinol, cobalamin, ascorbic acid, cholecalciferol, menaquinone and tocopherol were incomplete, while protoheme biosynthesis appeared to occur through a novel form of protoporphyrinogen IX oxidase (HemJ), which has so far mostly been described in cyanobacteria [55]. As previously noted, the *ubiD*/*ubiX* gene complex for ubiquinone biosynthesis was absent in all symbiont lineages [7]. The lack of UbiD/UbiX might be compensated through acquisition of metabolic intermediates from the host or through an alternative, currently undescribed pathway.

*Methionine synthase*

Clade I and Clade II symbionts appear to use different enzymes for the synthesis of methionine. The gene for the cobalamin-dependent homocysteine methyltransferase (*metH*) as well as genes for cobalamin (precursor) transport and conversion (*btuM*, *btuR/cobA*) were conserved in genomes of Clade I but were missing or degenerated in all of the Clade II lineages, except for *Ca.* R. phaseoliformis and *Ca.* R. southwardae. Conversely, the gene for the cobalamin-independent version of this enzyme (*metE*) along with its transcriptional activator (*metR*) were exclusively found in the Clade II symbiont genomes. Notably, almost all genes for *de novo* cobalamin biosynthesis were absent from the investigated symbiont genomes, with the exception of cobyrinic acid A,C-diamide synthase (*cbiA*) (all genomes), adenosylcobalamin/alpha-ribazole phosphatase (*cobC*) (not in *Ca.* R. magnifica) and the high affinity cobalamin transporter BtuB (*Ca*. V. gigas, *Ca.* V. marissinica, *Ca.* R. southwardae, *Ca.* R. phaseoliformis).

*Nitrate reductase*

An operon coding for the membrane-bound nitrate-reductase complex NarGHIJ was conserved in all Clade I symbiont genomes, but not in those of Clade II, which appeared to contain non-functional remnants of this operon. Conversely, the Clade II symbionts encode the cytoplasmic assimilatory nitrate reductase NasA, which was degenerated in Clade I.

*Cysteine dioxygenase, cysteine sulfinic acid decarboxylase and isocitrate lyase*

The gene coding for cysteine dioxygenase type I (*cdo*), which catalyzes the conversion of L-cysteine to cysteine sulfinic acid, was conserved in all Clade I lineages, but was absent or degenerated in most Clade II symbiont genomes (with the exception of *Ca.* R. phaseoliformis and *Ca.* R. pliocardia). CDO occurs in an operon with a pyridoxal phosphate dependent enzyme of the aspartate aminotransferase superfamily, which likely has cysteine sulfinic acid decarboxylase activity based on sequence homology with corresponding genes of *Bathymodiolus* mussel endosymbionts. By contrast, onlythe Clade II symbionts encode genes for isocitrate lyase (*icl*), a key enzyme of the glyoxylate cycle.

*Transcription, translation and post-translational modification*

All vesicomyid symbiont genomes contained an operon for a Class Ia ribonucleotide reductase (*nrdAB*), but only the Clade II lineages appeared to also encode the gene for its transcriptional repressor (*nrdR*). In addition, we found genes for several enzymes involved in protein modification and response to cellular stress in the Clade II genomes that were absent in Clade I. For instance, all Clade II lineages contained genes for the GTP-binding protein HflX (exception: *Ca*. R. pliocardia), and the peptide methionine sulfoxide reductase MsrB, which play a role in dissociation of translationally arrested ribosomes [56], and protein repair after oxidative damage, respectively. Likewise, most Clade II lineages encoded genes for GidB and other methyltransferases, which are involved in RNA modification.

*Cell wall and membrane biosynthesis*

The two symbiont clades differed in several genes that are involved in biogenesis of the cellular envelope. Although we found complete pathways for the production of the common membrane lipid phosphatidylethanolamine in the genomes of all vesicomyid symbiont lineages, genes for diacylglycerol kinase (*dgkA*), which is necessary for phospholipid recycling, was only present in the Clade II symbionts. Similarly, all Clade II symbionts encoded a 1,6-anhydro-N-acetylmuramate kinase (AnmK) and an outer membrane lipoprotein (SlyB), which are important for cell wall recycling and integrity, respectively. The Clade II lineages also contained a small-conductance mechanosensitive channel involved in osmoregulation (MscS), a lipopolysaccharide (LPS) export system protein (LptA) involved in LPS-translocation across the periplasm, and an N-acetyl-anhydromuramyl-L-alanine amidase (AmpD) involved in cell wall degradation. Homologs of these genes were either completely missing or pseudogenized in the Clade I symbionts. Both symbiont clades possessed genes for peptidoglycan biosynthesis, although MurD, MraY and MurG enzyme functionalities might be impaired or altered by the presence of internal stop codons in the case of *Ca.* V. diagonalis and *Ca.* V. extenta.

*Transport across membrane*

Multiple components of a type I secretion system (lapC, lapB, lapE, and the secreted agglutinin RTX) were found in all of the Clade II symbionts except for *Ca.* R. pliocardia. By contrast, this locus was missing in Clade I. The Clade II symbionts also encoded a putative hydrogenase/urease accessory protein (HupE), which is thought to be a nickel or cobalt transporter [57]. Although *hupE* is often associated with operons coding for [NiFe] hydrogenases, we did not find genes encoding hydrogenase subunits in any of the symbiont genomes. However, a gene encoding a nickel-dependent glyoxalase I (*gloA*) was present in the genomes of most Clade II symbionts.

*DNA repair and recombination*

In agreement with Kuwahara et al. [58] and Shimamura et al. [59], genes of the nucleotide excision repair pathway, *uvrA*, *uvrD*, *uvrD* paralog and *mfd*, were conserved in most symbiont genomes, while *uvrB* and *uvrC* were degenerated in all Clade I lineages. Within Clade II, *uvrA*, *uvrB*, *uvrD* paralog and *mfd* were present in all lineages, whereas *uvrC* was lost in *Ca.* R. pliocardia, and *uvrD* was lost in *Ca.* R. phaseoliformis. Many Clade II lineages contained genes for repair of alkylated DNA (*alkD*) and strand breaks (*radA*), while homologs of these genes were absent from all Clade I symbiont genomes. Furthermore, we found that essential genes involved in SOS response to DNA damage, *recA*, *recOR*, and *recX*, were lost in Clade I and *Ca.* R. magnifica. In the other Clade II lineages, these genes were conserved with the exception of *recO*, which was degenerated in *Ca.* R. phaseoliformis. Likewise, the gene coding for RuvC, an essential component of the last step of the recF and recBCD pathways for homologous recombination [60], as well as the genes coding for the XerCD recombinase system and the DNA recombination protein RmuC were lost in virtually all Clade I lineages, but conserved in most of the Clade II symbionts.

*Mobile elements and defense against pathogens*

The genomes of all vesicomyid symbionts are notably sparse in genes related to anti-viral defense and transposition. Phage-related genes except for a putative phage tape measure protein were completely missing in Clade I, while a few transposases, integrases and other phage-derived proteins were found in some of the Clade II lineages, in particular *Ca.* R. southwardae and *Ca.* R. phaseoliformis. In addition, remnants of type I restriction-modification systems (HsdMRS) and mRNA-degrading toxin-antitoxin systems (e.g., MazEF) were present in the genomes of *Ca.* R. fausta, *Ca.* R. pacifica, *Ca.* R. rectimargo, *Ca.* R. phaseoliformis and *Ca.* R. southwardae, but lost in all other symbiont genomes. *Candidatus* R. southwardae and *Ca.* R. phaseoliformis further contained degenerated operons for the 5-methylcytosine-specific restriction endonuclease McrBC. In addition, putatively defunct versions of Cascade complex genes that were previously part of a CRISPR/Cas system were found in *Ca.* R. pliocardia (*cas2*) and *Ca.* R. southwardae (*cas1*, *cas3*).

**References**

1. 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)

2. 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)

3. 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)

4. 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)

5. 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)

6. 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)

7. 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)

8. 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)

9. 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)

10. 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)

11. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. 2010 FastQC. See http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

12. Bolger AM, Lohse M, Usadel B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. (doi:10.1093/bioinformatics/btu170)

13. 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)

14. 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)

15. 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)

16. 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)

17. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011 Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579. (doi:10.1093/bioinformatics/btq683)

18. Langmead B, Salzberg SL. 2012 Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359. (doi:10.1038/nmeth.1923)

19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup*.* 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

20. 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)

21. Laslett D, Canbäck B. 2008 ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics* **24**, 172–175. (doi:10.1093/bioinformatics/btm573)

22. 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)

23. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013 QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075. (doi:10.1093/bioinformatics/btt086)

24. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019 GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36(6)**, 1925–1927. (doi:10.1093/bioinformatics/btz848)

25. 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)

26. Lemoine F, Lespinet O, Labedan B. 2007 Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC Evol Biol* **7**, 237. (doi:10.1186/1471-2148-7-237)

27. Darling ACE, Mau B, Blattner FR, Perna NT. 2004 Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394–1403. (doi:10.1101/gr.2289704)

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

29. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**, 772–780. (doi:10.1093/molbev/mst010)

30. Price MN, Dehal PS, Arkin AP. 2010 FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490. (doi:10.1371/journal.pone.0009490)

31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009 BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421. (doi:10.1186/1471-2105-10-421)

32. Gu Z, Eils R, Schlesner M. 2016 Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849. (doi:10.1093/bioinformatics/btw313)

33. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK*.* 2012 VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568–576. (doi:10.1101/gr.129684.111)

34. 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)

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

36. 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)

37. Warren DL, Geneva AJ, Lanfear R. 2017 RWTY (R We There Yet): An R package for examining convergence of Bayesian phylogenetic analyses. *Mol Biol Evol* **34(4)**, 1016–1020. (doi:10.1093/molbev/msw279)

38. 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)

39. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006 GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098. (doi:10.1093/bioinformatics/btl474)

40. 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)

41. Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797. (doi:10.1093/nar/gkh340)

42. 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)

43. 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)

44. 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)

45. 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)

46. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M *et al.* 2014 The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* **42**, D206–D214. (doi:10.1093/nar/gkt1226)

47. Fisher SRA. 1962 Confidence limits for a cross-product ratio. *Aust J Stat* **4**, 41. (doi:10.1111/j.1467-842X.1962.tb00285.x)

48. 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)

49. 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)

50. Spielman SJ, Weaver S, Shank SD, Magalis BR, Li M, Kosakovsky Pond SL. 2019 Evolution of viral genomes: Interplay between selection, recombination, and other forces. *Methods Mol Biol* **1910**, 427–468. (doi:10.1007/978-1-4939-9074-0\_14)

51. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Hönigschmid P, Schafferhans A, Roos M, Bernhofer M *et al.* 2014 PredictProtein--an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res* **42**, W337-343. (doi:10.1093/nar/gku366)

52. Dubilier N, Bergin C, Lott C. 2008 Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nat Rev Micro* **6**, 725–740. (doi:10.1038/nrmicro1992)

53. Nakagawa S, Takai K. 2008 Deep-sea vent chemoautotrophs: diversity, biochemistry and ecological significance. *FEMS Microbiol Ecol* **65**, 1–14. (doi:10.1111/j.1574-6941.2008.00502.x)

54. Klatt JM, Polerecky L. 2015 Assessment of the stoichiometry and efficiency of CO2 fixation coupled to reduced sulfur oxidation. *Front Microbiol* **6**, 484. (doi:10.3389/fmicb.2015.00484)

55. Kato K, Tanaka R, Sano S, Tanaka A, Hosaka H. 2010 Identification of a gene essential for protoporphyrinogen IX oxidase activity in the cyanobacterium *Synechocystis* sp. PCC6803. *PNAS* **107**, 16649–16654. (doi:10.1073/pnas.1000771107)

56. Zhang Y, Mandava CS, Cao W, Li X, Zhang D, Zhang X, Qin Y, Mi K, Lei J *et al.* 2015 HflX is a ribosome-splitting factor rescuing stalled ribosomes under stress conditions. *Nat Struct Mol Biol* **22**, 906–913. (doi:10.1038/nsmb.3103)

57. Eitinger T, Suhr J, Moore L, Smith JAC. 2005 Secondary transporters for nickel and cobalt ions: theme and variations. *Biometals* **18**, 399–405. (doi:10.1007/s10534-005-3714-x)

58. 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)

59. 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)

60. Connolly B, Parsons CA, Benson FE, Dunderdale HJ, Sharples GJ, Lloyd RG, West SC. 1991 Resolution of Holliday junctions in vitro requires the *Escherichia coli* ruvC gene product. *PNAS* **88**, 6063–6067. (doi:10.1073/pnas.88.14.6063)