Global biogeography of chemosynthetic symbionts reveals both localized and globally distributed symbiont groups

Jay T. Osvatic, Laetitia G. E. Wilkins, Lukas Leibrecht, Matthieu Leray, Sarah Zauner, Julia Polzin, Yolanda Camacho, Olivier Gros, Jan A. van Gils, Jonathan A. Eisen, Jillian M. Petersen, and Benedict Yuen

*Centre for Microbiology and Environmental Systems Science, University of Vienna, 1090 Vienna, Austria; †Genome and Biomedical Sciences Facility, Genome Center, University of California, Davis, CA 95616; ‡Department of Symbiosis, Max Planck Institute for Marine Microbiology, 28209 Bremen, Germany; §Smithsonian Tropical Research Institute, Apartado 0943-03092, Balboa, Republic of Panama; ¶Centro de Investigación en Ciencias del Mar y Limnología, Escuela de Biología, Universidad de Costa Rica, San Pedro 11501-2060, Costa Rica; ‡UMR 7205, Institut de Systématique, Évolution, Biodiversité, Équipe Biologie de la Mangrove, Département de Biologie, Université des Antilles, 97159 Pointe-à-Pitre Cedex, Guadeloupe; †Royal Netherlands Institute for Sea Research, 1790 AB Den Burg, The Netherlands; ¤Department of Evolution and Ecology, University of California, Davis, CA 95616; and ‡Department of Medical Microbiology and Immunology, University of California, Davis, CA 95616

Edited by Edward F. DeLong, University of Hawaii at Manoa, Honolulu, HI, and approved May 25, 2021 (received for review March 5, 2021)

In the ocean, most hosts acquire their symbionts from the environment. Due to the immense spatial scales involved, our understanding of the biogeography of hosts and symbionts in marine systems is patchy, although this knowledge is essential for understanding fundamental aspects of symbiosis such as host–symbiont specificity and evolution. Lucinidae is the most species-rich and widely distributed family of marine bivalves hosting autotrophic bacterial endosymbionts. Previous molecular surveys identified location-specific symbiont types that “promiscuously” form associations with multiple divergent cooccurring host species. This flexibility of host–microbe pairings is thought to underpin their global success, as it allows hosts to form associations with locally adapted symbionts. We used metagenomics to investigate the biodiversity, functional variability, and genetic exchange among the endosymbionts of 12 lucinid host species from across the globe. We report a cosmopolitan symbiont species, Candidatus Thiodiazotrophia taylori, associated with multiple lucinid host species. Ca. T. taylori has achieved more success at dispersal and establishing symbioses with lucinids than any other symbiont described thus far. This discovery challenges our understanding of symbiont dispersal and location-specific colonization and suggests both symbiont and host flexibility underpin the ecological and evolutionary success of the lucinid symbiosis.

Significance

Knowledge of host–symbiont biogeography is critical to understanding fundamental aspects of symbiosis such as host–symbiont specificity. Marine animals typically acquire their symbionts from the environment, a strategy that enables the host to associate with symbionts that are well-suited to local conditions. In contrast, we discovered that in the chemosymbiotic bivalve family Lucinidae several host species distributed across the globe are all associated with a single cosmopolitan bacterial symbiont. The genetic cohesiveness of this global symbiont species is maintained through homologous recombination across its extensive geographical range. The remarkable flexibility in the lucinid association is advantageous to both host and symbiont as it increases the likelihood of locating a compatible partner across diverse habitats spanning the globe.

Author contributions: J.T.O., L.G.E.W., J.A.E., J.M.P., and B.Y. designed research; J.T.O., L.G.E.W., LL, M.L., S.Z., J.P., Y.C., O.G., J.A.V.G., and B.Y. performed research; L.G.E.W. and J.M.P. contributed new reagents/analytic tools; J.T.O., L.G.E.W., and B.Y. analyzed data; J.T.O., L.G.E.W., and B.Y. wrote the paper; M.L., S.Z., J.P., Y.C., O.G., and J.A.V.G. collected samples; and J.A.E. and J.M.P. contributed to funding and conceptualization. The authors declare no competing interest.

This article is a PNAS Direct Submission. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

1To whom correspondence may be addressed. Email: jillian.petersen@univie.ac.at or benedict.yuen@univie.ac.at.

2This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2104378118/-/DCSupplemental.

Published July 16, 2021.
thrive in a wide array of shallow and deep-water marine habitats, which suggests a great diversity of lucinid symbionts remains to be discovered (17). This diversity of hosts and habitats, distributed throughout the globe, makes lucinids a powerful system for unraveling the ecological, biogeographic, and systematic factors influencing diversity and flexibility in horizontally transmitted nutritional symbioses.

We used deep-coverage metagenomes to study diversity and flexibility in the association between lucinids and their chemosynthetic symbionts, focusing on populations and species in the Caribbean and the Mediterranean. We report a lucinid symbiont—*Candidatus Thiodiazotropha* taylori—associated with eight host species, representing three lucinid subfamilies, from distant locations across the globe. Our findings suggest a high degree of flexibility in partner choice could be an important factor in the ecological and evolutionary success of the lucinid symbiosis. We further examined how homologous recombination has shaped symbiont biology and carried out comparative genomes to investigate the functional variability among symbiont species across different host populations and species. Finally, we discuss how both symbiotic and free-living life stages influence the biology of the symbionts and their lucinid hosts.

**Results**

**Individual Lucid Hosts Can Harbor More Than One Symbiont Species Simultaneously.** We sequenced, assembled, and binned metagenomes from 47 individual clams. This resulted in 63 MAGs (metagenome-assembled genomes), 53 of which were considered high-quality with greater than 90% completeness and less than 10% contamination (18) (Dataset S1). All newly assembled MAGs from this study are available under GenBank accession numbers SAMN16825223 to SAMN16825285, including raw reads (SRA database project number PRJNA679177). All 63 MAGs were assigned to the genus *Ca.* Thiodiazotropha. Sixteen metagenomes yielded two distinct *Ca.* Thiodiazotropha MAGs in one gill. The majority of lucinids we investigated hosted a single symbiont species abundant enough for its genome to be assembled from gill metagenomes. Of these, 15 *Clathrolucina costata* and 1 *Loripes orbiculatus* contained two distinct MAGs (Dataset S1). Using DESMAN, no more than two strains were detected within any MAG, indicating that there is also limited symbiont strain diversity in each host individual (Dataset S1).

**Symbionts Form Distinct Clades within the Gammaproteobacterial Family Sedimenticaceae.** We reconstructed the phylogenetic relationships of 63 *Ca.* Thiodiazotropha MAGs we generated to previously described lucinid symbionts and their closest free-living relatives, *Sedimenticola thiotaurini* and *Sedimenticola selenatireducens*, using a concatenated amino acid alignment of 43 universally conserved marker genes, with *Allochronium vinosum* as an outgroup [alignment on Figshare (19)]. The maximum likelihood phylogenetic reconstruction yielded 10 unique clades, each with an average nucleotide identity (ANI) above 95%, indicating that each clade likely corresponds to a distinct species, two of which have not been previously identified (Dataset S2). Apart from the symbionts of *Phacocidium pectinatum*, *Ca.* Sedimenticola endophagoecides, all other lucinid symbionts belong to a clade that most likely represents a single genus, named *Ca.* Thiodiazotropha. The first of the two species we identified was represented by 36 MAGs assembled from lucinid hosts from across the Pacific, Atlantic, and Indian Oceans, making it the most widely distributed lucinid symbiont known to date. We propose the name *Candidatus Thiodiazotropha* taylori for this species after Dr. John Taylor (*SI Appendix, Supplementary Discussion*), to honor his enormous contribution to understanding the biodiversity and evolution of the Lucinidae. *Ca.* T. taylori was associated with eight host species from three different subfamilies of Lucinidae (*Lucininae*, *Leucosphaerinae*, and *Codakinae*), making this the most diverse group of hosts that any lucinid symbiont species is known to associate with (Fig. 1). The second species, *Ca.* T. sp. “RUGA,” was found in a single *Rugulucina munda* metagenome. The *Ca.* T. sp. “RUGA” MAG shared an ANI of 94.543% with all *Ca.* T. taylori MAGs, making it the closest identified relative to the *Ca.* T. taylori group.

Reanalysis of *L. orbiculatus* metagenomes containing a previously described symbiont species, *Ca.* T. endoloripes, revealed that *Ca.* T. endoloripes is actually two distinct lineages with an ANI of 91.938%, which indicates they are separate species (Dataset S2). We propose the names *Ca.* T. weberae and *Ca.* T. lotti for these two species after Dr. Miriam Weber and Christian Lott, who discovered the population of *L. orbiculatus* in Elba, Italy. Its symbionts *Ca.* T. weberae and *Ca.* T. lotti both share an ANI of roughly 89% with *Ca.* T. taylori and *Ca.* T. sp. “RUGA,” their next-closest relatives (Fig. 2). The lucinid symbionts from the genus *Ca.* Thiodiazotropha formed two major clades, one comprising the four above-mentioned species and another containing *Ca.* T. endoloricida and the *Thiodiazotropha*-like species associated with *Ctena orbiculata* from Florida, previously reported by Lim et al. (20) (Fig. 2). Like *Ca.* T. taylori, *Ca.* T. endoloricida was also found in multiple host species (five species from subfamilies *Lucininae*, *Codakinae*, and *Leucosphaerinae*) but its geographic distribution appears to be restricted to the Caribbean based on samples that are so far available (February 2021).

**Comparative Genomics of Lucinidae Symbionts.** We compared the predicted functional capabilities of *Ca.* T. taylori, the most widespread symbiont in our dataset, to 1) its closest relatives (*Ca.* T. weberae, *Ca.* T. lotti, and *Ca.* T. sp. “RUGA”) and 2) the sympatric *Ca.* T. endoloricida, which can cooccur with *Ca.* T. taylori in the same host gill, to investigate whether functional potential reflects symbiont geography or phylogeny. All five symbiont species shared most core metabolic functions (Table 1). A maximum of 7.2%, or 170 of the 2,361 protein families (Pfams) found in these genomes, were enriched in *Ca.* T. endoloricida, i.e., they were significantly more frequent within this species group (*P*adj < 0.05), while the closely related species *Ca.* T. taylori, weberae, and lotti had far fewer enriched Pfam functions, suggesting that phylogenetic relationships are to some extent reflected in genomic functional potential (details in *SI Appendix, SI Results, Table S3*, and Fig. S2A and B and Dataset S5). In other words, the more closely related the symbionts, the more likely they are to share genome content.

**Core Metabolic Pathways Shared by All Symbiont Species.** The *Ca.* Thiodiazotropha symbionts rely on multiple pathways to oxidize reduced sulfur compounds to sulfate through a polysulfur or elemental sulfur intermediate and there were no major variations in sulfur oxidation capabilities across all the MAGs (Table 1 and *SI Appendix, Table S2* and Dataset S3). The energy generated from sulfur oxidation is used to power inorganic carbon fixation through the CBB cycle (Table 1). The MAGs of *Ca.* T. taylori, *T. weberae*, *T. lotti*, T. sp. “RUGA,” and *T. endoloricida* each contained a gene encoding Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) form I (Table 1). *Ca.* T. endoloricida was the only species that possessed two distinct types of RuBisCO, forms I and II. All five species also encoded the capacity for heterotrophic growth (Table 1). The genus name *Thiodiazotropha* was proposed following the recent discovery that lucinid symbionts are capable of fixing inorganic nitrogen from the atmosphere (14, 15). Indeed, all five symbiont species encoded large gene clusters involved in nitrogen fixation, dinitrogenase reductase subunit (*nifH*), the structural genes (*nifD* and *nifK*), ferredoxins, and regulatory factors (Dataset S3). All species had a complete denitrification pathway for reducing nitrate and nitrite to nitrogen gas, and thus all had the potential to use nitrate and nitrite as alternative electron acceptors (Table 1).
Fig. 1. Geographic distribution and host specificity of lucinid symbiont species. (A) Global biogeography of symbiont species in the genus Ca. Thiodiazotropha reveals both localized (Mediterranean and Caribbean) and globally distributed symbiont groups. Shapes represent host species and colors represent symbiont species. Ca. T. taylori (pink) was found in association with eight lucinid species across the globe. Ca. T. sp. “RUGA” (teal) is the endosymbiont of a R. munda specimen from Tin Can Bay, Queensland, Australia. Ca. T. endoloripes, previously described as a single species by Petersen et al. (15), is in fact two closely related species (Ca. T. weberae in orange and Ca. T. lotti in purple), so far found exclusively within L. orbiculatus in the Mediterranean. (B) Phylogenetic relationships of lucinid endosymbionts. Shown is a maximum likelihood phylogenetic tree reconstructed from 43 conserved marker genes. Circles indicate bootstrap support values above 95%. Colors indicate geographic origin of the sample.
Fig. 2. Phylogenetic relationships of Ca. Thiodiazotropha taylori MAGs and distribution across host species. (A) Phylogenetic relationship among Ca. T. taylori MAGs (high quality) and the individual host species where these MAGs were found in (b–h; a = outgroup symbiont found in the host *Rugalucina munda* colored in gray). The core gene alignment of Ca. T. taylori was constructed by aligning all shared genes in progressiveMauve and correcting for recombination in ClonalFrameML (SI Appendix, Fig. S1). This corrected alignment [alignment available on Figshare (90)] was uploaded to the IQtree web interface to construct a maximum likelihood phylogenetic tree with 1,000 bootstraps. The best-fit substitution model was TVM+F+ASC according to BIC scoring. Circles indicate support values above 95%. Note that Ca. T. taylori was found in eight host species but one host MAG was found to have low completion and high contamination and was therefore not included in this analysis (*C. orbicularis*, Florida). Shell images reprinted with permission from ref. 91. Lucinid shells are not to scale. (B) Spatial distribution of Ca. T. taylori in the gills of *C. costata* and (C) *L. orbiculatus*. Magenta, Ca. T. taylori; cyan, T. endolucinida; yellow, DAPI-labeled nuclei; green dashed lines, zone of ciliated epithelial cells.
Accessory Metabolic Functions in Carbon and Nitrogen Metabolism.

Despite sharing most metabolic functions that sustain the nutritional lucinid symbiosis, we observed notable variations in symbiotic carbon and nitrogen metabolic pathways. Ca. T. taylori, Ca. T. endolucinida, and Ca. T. sp. “RUGA” possessed genes required for growth on reduced one-carbon compounds (20), including lanthanide-dependent methanol dehydrogenase (soxF-like), genes for the synthesis of its cofactor pyrroloquinoline quinone (pqABCDE), and the full tetrahydromethanopterin-dependent pathway for formaldehyde oxidation, which was found in close proximity to methanol dehydrogenase (21, 22). The presence of this cluster of genes suggests these symbionts are able to convert waste product of lucinid clams. Only MAGs of Ca. T. taylori and Ca. T. weberae encoded genes for the uptake and conversion of urea to ammonia (ureACDEFGJ). MAGs of Ca. T. taylori, Ca. T. lotti, Ca. T. weberae, and Ca. T. “RUGA” also encoded genes for an alternative pathway for urea metabolism through urea amidolyase (SI Appendix, Table S2). Finally, Ca. T. endolucinida was the only species with MAGs encoding genes for the respiratory nitrite reductase complex (narGHJI), indicating that it can reduce nitrate for generating a proton motive force. These observations were based on the presence/absence of entire operons/gene clusters and were consistent across all the high-quality MAGs examined, of which we had multiple samples from each symbiont species (minimum of nine).

Recombination Rates. Nucleotide differences in closely related bacterial sequences can be caused by two primary mechanisms, either homologous recombination (HR), where a fragment of the recipient’s genome is replaced by that of a donor in a single generation (genetic exchange), or by mutations from one generation to the next (23). In the absence of HR, all differences in genomic sequences reflect a clonal genealogy. We calculated average HR to mutation rates (R/m) for each species group and for groups containing multiple species. The alignment lengths of shared genes used to calculate R/m ranged from 1,751,700 to 3,149,280 base pairs (bp), which for some groups covered more than half of the total length of the MAGs (Table 2). The alignment of shared genes was shortest in Ca. T. taylori due to the strict alignment parameters chosen (i.e., aligned genes must be present in all samples) and a greater genetic diversity within this species group. Just as Ca. T. taylori is unique among lucinid symbiont species for its extensive geographic range, its average R/m ratio was at least 10 times higher than the average R/m ratios in all the other species groups (0.814; 0.809 to 0.819; 95% confidence intervals; Table 2). The average R/m ratios for Ca. T. taylori populations from different locations were fairly consistent and ranged from 0.859 in Florida to 1.521 in Mauritania (SI Appendix, Table S4). The lowest R/m values were between Ca. T. endolucinida and Ca. T. taylori (0.053; 0.052 to 0.054), even though these can cooccur within host individuals. The lengths of all alignments, average number of recombination events, average length of events, and all R/m ratios and their 95% confidence intervals are shown in SI Appendix, Table S4. The average nucleotide diversity outside the recombining gene sequences was 3.6, 2.4, 1.9, and 0.6 single-nucleotide variants per bp for Ca. T. taylori, Ca. T. lotti, Ca. T. weberae, and Ca. T. endolucinida, respectively (Dataset S6). Although MAGs consist of consensus sequences of the most abundant representatives of an entire bacterial population, the low strain estimates from DESMAN (two or fewer strains) across all MAGs in this study provide confidence that these results are not confounded by strain differences within the gill metagenomes (SI Appendix, Table S1).

Discussion

A Single Cosmopolitan Symbiont Species Associates with Multiple Diverse Lucinid Host Species at Locations around the World. We investigated the diversity and the predicted functional variability of the symbionts associated with lucinids on a global scale. Fresh samples were complemented with specimens from museum collections to
expands the scope of the study beyond the most intensively studied lucinid habitats in the Caribbean and Mediterranean Seas (15, 20, 24). We identified two lucinid symbiont species, Ca. T. taylori and Ca. T. “RUGA.” Ca. T. taylori is found in a remarkable eight different lucinid species from three different subfamilies (Figs. 1 and 2), which makes Ca. T. taylori the most promiscuous lucinid symbiont described thus far and the first chemosynthetic endosymbiont species with a globally distributed population (25, 26). We similarly found Ca. T. endolucinida in five divergent host species from three distinct lucinid subfamilies (Fig. 1). These findings corroborate previous reports that the same symbiont 16S ribosomal RNA (rRNA) gene sequence, identical to that of Ca. T. endolucinida (SI Appendix, Fig. S3), was present in four distinct host species at one location in Guadeloupe (25). We predict that future surveys of lucinid symbiont diversity are likely to reveal similar instances of promiscuity. The remarkable flexibility in the association between lucinids and Ca. Thiodiazotropha species could thus be an important feature underlying the evolutionary success of this ancient and widespread symbiosis.

**HR Maintains the Cohesiveness of Ca. T. taylori as a Single “Species.”** Genetic exchange plays a critical role in bacterial genome evolution and can either drive the divergence or homogenization of a population (27). HR drives homogenization by maintaining genomic cohesion of bacterial clades even across distant locations (20, 29). We reconstructed recombination events in four symbiont genomic cohorts of bacterial clades even across distant locations (20, 29). We reconstructed recombination events in four symbiont species groups—Ca. T. taylori, T. endolucinida, T. weberae, and T. lotti—to investigate the role of HR in maintaining genomic cohesion of Ca. T. taylori as a single species-level group. The rate of recombination to mutation (R/m) in Ca. T. taylori (0.814) is above the theoretical threshold for preventing population divergence (0.25) (29–31). In contrast, the R/m rates of Ca. T. endolucinida, T. weberae, and T. lotti are well below this 0.25 threshold (0.04 to 0.08), a pattern that correlates with the limited geographic range of these three species compared to Ca. T. taylori (Fig. 1). With such a large distribution range spanning many diverse habitats, selection pressures or neutral drift could cause local Ca. T. taylori populations to diverge, which might result in gene content and functional differences (Table 1). This is supported by our phylogenetic reconstruction of the Ca. T. taylori clonal frame (Fig. 2), which reveals that clams from each distinct geographic location are colonized by site-specific phylogenetic lineages that were poorly resolved in the phylogenomic tree (Fig. 1). Our recombination analyses indicate that HR has a cohesive effect that maintains the integrity of Ca. T. taylori as a single cosmopolitan species-level group. With an ANI of about 94.5%, the Ca. T. taylori and Ca. T. “RUGA” MAGs represent distinct, albeit closely related, genetic units (Fig. 1 and Dataset S1). Given that HR rates drop exponentially with decreasing sequence similarity, declining steeply between 90 to 95% ANI (32), we speculate that this seemingly small amount of nucleotide divergence likely prevents recombination between Ca. T. “RUGA” and Ca. T. taylori. Indeed, an ANI of 95% is widely used as a threshold for delineating bacterial species, as this is thought to correspond to a level of divergence that prevents genetic exchange by HR (33–35). One could speculate that Ca. T. “RUGA” and Ca. T. taylori previously belonged to a single species-level group but were relatively recently separated and that if the local geographic Ca. T. taylori populations encountered a barrier to HR they too might become new and distinct species-level groups. It is currently unclear which barriers could “remove” local populations from the cohesive forces of HR and allow them to diverge independently. Greater efforts to sample and sequence more Ca. T. “RUGA” and Ca. T. taylori symbionts will allow us to test these and other theories about the emergence of symbiont diversity in future.

The cosmopolitan distribution of lucinid clams hosting Ca. T. taylori suggests this symbiont species is able to disperse over great geographic distances. How Ca. T. taylori achieves this feat of dispersal remains a mystery but hitchhiking on their bivalve hosts during the planktonic larval phase appears unlikely because the lucinid larvae studied thus far have been found to be apysymbiotic (11). Furthermore, the range of this symbiont species far exceeds the limited distribution range of any lucinid host species. Ca. T. taylori therefore likely migrates during its free-living phase rather than its host-associated phase. The dormant endospores of thermophilic Firmicutes achieve extensive distributions through long-distance passive dispersal in oceanic currents, but even these tough endospores with enhanced survival capacities face more substantial dispersal limitations than Ca. T. taylori (36). It seems unlikely that free-living Ca. T. taylori cells would be able to survive harsh open-ocean conditions long enough to traverse the globe. To our knowledge, there are no molecular data to date indicating Ca. Thiodiazotropha spp. are present at a high abundance in coastal sediments or the water column, which suggests they are members of the rare biosphere present only at low relative abundance (37). Troussellier et al. put forward the intriguing hypothesis that macroorganisms could sustain the rare biosphere by serving as dissemination vectors for marine microbes (38). A recent meta-analysis of publicly available ampiclon sequencing data revealed the presence of Ca. Thiodiazotropha 16S rRNA gene sequences associated with the roots of various seagrass species around the globe, a discovery the authors subsequently verified with microscopic imaging (37). Vegetative seagrass fragments (shoots and rhizomes) can reestablish in distant locations after detaching from their parents and have great potential for long-distance dispersal (39, 40). It is interesting to speculate that lucinid clams and seagrasses from across the globe could form a network of source habitats facilitating the dissemination and dispersal of Ca. T. taylori vectored by oceanic circulation, thereby overcoming potential barriers to dispersal between geographically distant bodies of water. Despite its global distribution range, we may have identified one potential barrier to dispersal of Ca. T. taylori. Rigorous sampling of L. orbiculatus along the Atlantic and Mediterranean coasts of Europe did not reveal a single instance of Ca. T. taylori associated with L. orbiculatus in any of these locations (Fig. 1). Instead, L. orbiculatus in these locations all hosted the closely related sister species Ca. T. weberae and T. lotti (Fig. 1). The colonization of L. orbiculatus by Ca. T. lotti from the United Kingdom all the way south to Kotor, Montenegro suggests the strait of Gibraltar is

### Table 2. Overall statistics from individual species and cooccurring pairs in ClonalFrameML

| Species                        | Length of genome alignment, bp | Recombination to mutation ratio (95% CI) |
|-------------------------------|-------------------------------|----------------------------------------|
| Ca. T. taylori                | 1,751,700                     | 0.814 (0.809–0.819)                     |
| Ca. T. weberae                | 3,149,280                     | 0.043 (0.041–0.046)                     |
| Ca. T. endolucinida           | 3,428,600                     | 0.084 (0.083–0.087)                     |
| Ca. T. endolucinida           | 3,428,600                     | 0.084 (0.083–0.087)                     |
| Cooccurring pairs             |                               |                                        |
| Ca. T. taylori–Ca. T. endolucinida | 1,664,616                   | 0.0066 (0.0063–0.0069)                  |
| Ca. T. weberae–T. lotti       | 3,149,280*                    | 0.053 (0.052–0.054)                     |

*The same core gene alignment was used for Ca. T. weberae and T. lotti together as well as each one individually.*
unlikely to pose a barrier to the dispersal of Ca. T. taylori. Nor are these distribution patterns driven by host-symbiont specificity, because L. orbiculatus in Mauritania hosts Ca. T. taylori (Fig. 1). A possible explanation is that environmental factors associated with a temperate climate prevent Ca. T. taylori from establishing in clams along the coasts of Europe. This is consistent with the tropical distribution of all lucinid species so far found to host L. orbiculatus. Lim et al. (20) recently reported multiple distribution patterns driven by host-symbiont specificity, 

able to outcompete T. taylori (Fig. 1). An alternate nonmutually exclusive explanation could be that Ca. T. weberae and T. lottii are better adapted and able to outcompete Ca. T. taylori for establishment in lucinid hosts throughout temperate Europe. Further sampling of other lucinid species and seagrasses along European coasts will be necessary to address these questions.

**Coexisting Ca. Thiodiazotropha Species Have Distinct Metabolic Capabilities.** Lim et al. (20) recently reported multiple Ca. Thiodiazotropha 16S rRNA gene amplicon sequence variants and MAGs of diverse Ca. Thiodiazotropha spp. associated with C. orbiculata individuals from Florida, providing the first indications that multiple symbiont species from the same genus may coexist in the same host gill. We were able to assemble and bin MAGs of two distinct symbiont species, Ca. T. taylori and Ca. T. endolucinida, from the metagenomes of 15 Ca. C. orbiculata samples from across the Caribbean. Using the same methods, a reanalysis of new and previously published gill metagenome data (LVIZ00000000) from L. orbiculatus (Elba, Italy) similarly revealed that MAGs from two distinct symbiont species, Ca. T. lottii and Ca. T. weberae, can be assembled and binned from metagenomes of single host individuals. Our fluorescence in situ hybridization (FISH) results showed that in C. costata, Ca. T. taylori and Ca. T. endolucinida both inhabit gill epithelial cells. These findings add to a growing number of studies reporting the coexistence of closely related chemosynthetic symbiont species within the same invertebrate host and indicate that this is more common in lucinids than previously assumed (41–43). Ecological models predict that cooccurring symbionts with the same resource requirements will compete for limited resources and that this competition is detrimental to the symbiosis (e.g., ref. 44). The cooccurring symbiont pairs we identified, Ca. T. taylori/endolucinida and Ca. T. weberae/lottii, shared core metabolic functions fundamental to their symbiosis with lucinid clams (15), including identical pathways for oxidizing sulfur, fixing inorganic carbon, and fixing nitrogen. Ca. T. taylori and Ca. T. endolucinida both had the additional potential to utilize methanol as a source of energy and carbon with an xox-type methanol dehydrogenase and the serine pathway for C1-carbon incorporation into biomass. Compartmentalizing coexisting symbiont species into separate bacteriocysts could prevent direct competition by allowing the host to partition resources and discriminate cooperative symbionts from potential cheaters that might destabilize the symbiosis (45). Consistent with this, our FISH analyses showed that although symbiont species cooccurred in host individuals they never cooccurred in single host bacteriocysts (Fig. 2). Despite their prolific productivity, habitats abundant with lucinids, such as seagrass meadows and coral reef lagoons, tend to occur in oligotrophic waters that are nitrogen-limited (14). It is noteworthy that the ability to hydrolyze urea to ammonia, which can subsequently be assimilated, is a conserved function of Ca. T. taylori and Ca. T. weberae that is absent in Ca. T. endolucinida and Ca. T. lottii (Table 1). The precise composition of lucinid nitrogenous waste remains unknown, but some bivalves do excrete urea as a waste product and the urea transporter DUR3 is highly expressed in the gills of L. orbiculatus (46, 47). Future studies are required to investigate whether the ability to utilize urea as a nitrogen source confers any advantage to Ca. T. taylori and Ca. T. weberae in the host or the external environment.

We observed some additional predicted metabolic differences between Ca. T. taylori and Ca. T. endolucinida. This symbiont pair has one major difference: Ca. T. endolucinida MAGs encode both RubisCo forms I and II, while Ca. T. taylori only encoded form I (Table 1). Ca. T. endolucinida also has the genes encoding a respiratory nitrate reductase protein complex, which indicates the ability to use nitrate instead of oxygen as an electron acceptor for respiration (Table 1). This combination of functional traits suggests Ca. T. endolucinida may be better adapted to survival in a lower-oxygen environment than Ca. T. taylori. This could be highly beneficial for the symbiosis as the endosymbionts of both the lucinid clam Lucinoma aspizonata and the vent tube worm Riftia pachyptila are able to respire nitrate as an adaptation to living in deep-sea environments with fluctuating oxygen availability (48–50). Whether shallow-water lucinids in the Caribbean are able to exploit this unique metabolic capability of Ca. T. endolucinida remains unknown but there are several observations suggesting it does not. First, the vanishingly small nitrate reductase activity in the gills of Codakia orbicularis, an established host of Ca. T. endolucinida, and the absence of nitrates in the tissues of this host species together suggest nitrate is not used as an electron acceptor in this symbiosis (51). Second, nitrate is only sporadically present at low concentrations in the pore water of Thalassia testudinum sediments and is undetectable in the overlying waters (14, 51). Third, oxygen is abundantly available in Caribbean lucinid habitats, especially during periods of photosynthesis, and lucinid clams construct burrows leading to the surface to obtain oxygenated water from above the sediment layer (52). These observations are inconsistent with Ca. T. endolucinida’s requiring these metabolic adaptations while they are housed in the buffered environment of the lucinid bacteriocytes. Moreover, our analysis of the C. orbiculats metatranscriptome indicates that the genes encoding RubisCO form II and respiratory nitrate reductase are expressed at a much lower level compared to RubisCO form I and the assimilatory nitrate reductases, respectively (Dataset S7). These data are further supported by the much lower abundance of RubisCO form II proteins compared to form I in the C. orbiculats proteome (14). Based on these observations, we hypothesize that Ca. T. endolucinida relies on these metabolic functions primarily during its free-living phase and we speculate that Ca. T. endolucinida could occupy an external environmental niche—one characterized by low oxygen availability—distinct from that of Ca. T. taylori.

**Materials and Methods**

**Sample Collection.** Live clams were collected from sites throughout the Caribbean and Mediterranean (SI Appendix, SI Methods and Dataset S1). Gills were dissected in the field and preserved in RNAlater (Ambion; Life Technologies) or DNA/RNA Shield (Zymo Research) at −80 °C. Samples from locations in the Pacific and Indian Oceans came from the collections of the Natural History Museum (NHM) in London and the Florida Natural History Museum (FLMNH), Gainesville, FL (SI Appendix, SI Methods and Dataset S1). Access was permitted and organized by Dr. John Taylor (NHM) and Dr. Gustav Paulay and Dr. Amanda Bemis (FLMNH).

**DNA Extraction, Preparation, and Sequencing.** DNA was extracted from gill tissues using the Qiagen DNeasy Blood and Tissue kit (69506; Qiagen) or the animal tissue protocol from Analytikjena InnuPrep DNA Mini Kit (845-KS-1041250) (SI Appendix, SI Methods). Samples were either treated with RNase or directly quantified before DNA libraries were prepared using Illumina-compatible library prep kits (Dataset S1 and SI Appendix, SI Methods and Table S5). All libraries were sequenced with Illumina technology to generate paired-end reads of 150 bp or 250 bp length (SI Appendix, SI Methods and Dataset S1).

**Quality Filtering, Assembly, and Bacterial Genome Binning.** Read libraries were trimmed, PhiX contamination-filtered, and quality-checked using BBMAP s instructions and stored at −20 °C. Samples from locations in the Pacific and Indian Oceans came from the collections of the Natural History Museum (NHM) in London and the Florida Natural History Museum (FLMNH), Gainesville, FL (SI Appendix, SI Methods and Dataset S1). Access was permitted and organized by Dr. John Taylor (NHM) and Dr. Gustav Paulay and Dr. Amanda Bemis (FLMNH).
The lineage workflow was used to identify, align, and concatenate a default set of using FastANI v1.3 (34). The accuracy of these species boundaries was tested publicly available MAGs were quality-checked and filtered using the methods of taylori in gill sections of of Ca. T. endolucidina (GCA_001715975.1), an endosymbiont of C. orbicularis from Guadeloupe (53). FeatureCounts was used to quantify gene-level transcript abundances, which were subsequently converted to transcripts per million (78). Metatranscriptomic results can be found in Dataset 57.

Recombination Rates and Nucleotide Diversity of Bacterial Symbionts. To infer recombination events in bacterial genomes, we used the maximum likelihood implementation of ClonalFrameML (23). High-quality MAGs were aligned using progressiveMauve, in Mauve v2.0 (79). Nucleotide sequences (i.e., core genes) shared within the groups of MAGs analyzed were extracted with stripSubsetLCBs, a script previously described in ref. 80. These core genes were realigned with MUSCLE (81) and cleaned with trimAl (82) (parameters in SI Appendix, SI Methods). RAxML v8.2.10 (83) was used to build a phylogenetic tree from this new alignment as described in ref. 84. All alignments are available on Figshare (85–87). The resulting tree and alignment were fed into ClonalFrameML to calculate the ratio of recombination vs. mutation events with default parameters (See Dataset 51 for MAGs used in this analysis). Recombination events were computed and visualized in R v6.3.2 with the script cfml_results.R (SI Appendix, SI Methods). GUBBINS was used to estimate average nucleotide diversities inside and outside the clonal frame (88). We also used the clonal frame alignments to reconstruct and visualize the phylogenetic relationships of the Ca. T. taylori MAGs as previously described (SI Appendix, SI Methods). All scripts used can be found in the associated Jupyter Notebook (89).

Data Availability. The data (raw reads, metagenomic assemblies, and MAGs) have been deposited with links to BioProject accession number PRJNA679177 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject). The BioSample accession numbers for the MAGs are SAMN1682223–SAMN16825285 and SAMN16952162–SAMN16952207 are the corresponding raw read sets. Analysis scripts are available at https://doi.org/10.6084/m9.figshare.13295912.v5. Datasets are available on Figshare at https://figshare.com/projects/A_globally_distributed_symbiont_challenges_host specificity_in_ lucinid_clams/93998. All other study data are included in the article and/or supporting information.

Acknowledgments. We thank Gustav Paulay (FLMNH, Gainesville, FL), Amanda Bemis, and John Taylor (NHM, London) for providing preserved samples. John Taylor also provided high-resolution images of lucinid clams for Fig. 2. We thank Liz Hambleton (University of Vienna), Miriam Weber, Christian Lott, Boris Unger, Dorotheé Makarov, Natalie Elisabeth (University of California, Berkeley), Diana Chih, Tim Oortwijn, and Arthur Curry for providing fresh samples. We are grateful to Stephanie Markert for providing access to the proteomics data, Ipek Yamin Merci and Nora Grossschmidt for helping with DNA extractions, Marina DeLeón for help with fieldwork and travels in Costa Rica, Cassandra Ettinger for early discussions of the study, and Guillaume Jospin for support with the bioinformatic analyses. We thank Xavier Didelet for discussions on the recombination analysis. L.G.E.W. was supported by Grant GBMF5603 from the Gordon and Betty Moore Foundation (J.A.E. was Principal Investigator on the same grant). The work was also supported by the European Research Council Starting Grant EvoLucin and a Vienna Research Grant for Young Investigators from the Vienna Science and Technology Fund (WWTF, VRG14-021) and the Austrian Academy of Sciences. Sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the University of California, Davis Genome Center (supported by NIH Shared Instrumentation Grant 1S10OD010786-01), the Medical Sequencing Facility at CeMM (https://www.biomedical-Sequencing.org) in Vienna, Austria, and at the Joint Microbiome Facility (JMF) at the Medical University of Vienna and the University of Vienna (project ID 1911). Thanks to Petra Pjevac and Gudrun Kohl for sample processing at JMF. The Gill embedding was performed by the Histopathology Facility at Vienna Biocenter Core Facilities, a member of the Vienna Biocenter, Austria.

C. orbicularis Metatranscriptome Analysis. Three C. orbicularis specimens were collected in Guadeloupe (2016), preserved in RNA later, and stored at −80 °C. Total RNA was extracted using TRIzol (Thermo Fisher Scientific) according to the manufacturer’s instructions. RNA degradation and library construction were carried out at the Vienna Biocenter Core Facilities GmbH as described in Yuen et al. (47). The RNA-sequencing reads were trimmed and processed as previously described by Yuen et al. (47). We used BBMAP (slow = t, ambi- guous = best, minid = 0.99) to align the processed reads to the MAG of C. T. endolucidina (GCA_001715975.1), an endosymbiont of C. orbicularis from Guadeloupe (53). FeatureCounts was used to quantify gene-level transcript abundances, which were subsequently converted to transcripts per million (78). Metatranscriptomic results can be found in Dataset 57.

Phylogenetic Analyses. Only the MAGs taxonomically assigned to Sedimentisphaerae or Chromatiales, using GDTB v0.3.3 (63) were used in this study. We also downloaded publically available MAGs of other lucinid symbionts (Ca. Thiodiazotropha spp. and Sedimentisphaera spp.), alongside A. vinous as an out-group, for this analysis (complete list of accession numbers in Dataset 51). The publicly available MAGs were quality-checked and filtered using the methods described above (and in SI Appendix, SI Methods). The CheckM v1.1.3 (64) lineage workflow was used to identify, align, and concatenate a default set of 43 universal marker genes from all the available MAGs (Dataset 51); concatenated marker gene alignment can be found on Figshare (19). This concatenated amino acid alignment was then submitted to the W-IQ-TREE server (65) using default settings (SI Appendix, SI Methods) and the resulting maximum likelihood tree was visualized using the Interactive Tree Of Life (iTOL) v5 (66). All MAGs were placed into species groups based on ANI values above 95% using fastANI v1.3 (64). The accuracy of these species boundaries was tested with ANIb and ANIm through the species web server (67).

Localization of Symbionts in Clam Gills. We carried out FISH to visualize Ca. T. taylori in gill sections of L. orbicularus (Mauritania, 2018) and C. costata (Panama, 2019). Probes were designed to target the 16S rRNA gene sequence of Ca. T. taylori using DECIPHER’s design probes web tool (SI Appendix, SI Methods) (68). A formamide series from 0 to 60%, in 10% steps, was carried out to optimize the probe hybridization conditions (SI Appendix, SI Methods; probe sequences are used in SI Appendix, Table 56). Nonsense sequences of the target-specific probes were also tested as negative controls (69). Dissected gills were preserved in 4% paraformaldehyde, dehydrated into 70% ethanol, and stored at 4 °C (SI Appendix, SI Methods). The gills were embedded in paraffin wax by the Histopathology Facility at Vienna BioCenter Core Facilities, Austria (SI Appendix, SI Methods). We cut the embedded gills into 5-μm sections with a Leica microtome and mounted the sections on SuperfrostPlus adhesion slides (Thermo Scientific). The sections were dewaxed in Roti-Histol (Carl Roth) following the manufacturer’s instructions. Probes were hybridized to the gill sections in a 50% formamide buffer (details on hybridization conditions and washing protocols are in SI Appendix, SI Methods and Tables 56 and 57). Following the hybridization, the samples were DAPI-stained (1 μg/mL) and mounted in ProLong Glass antifade mounting media (Thermo Fisher Scientific) (SI Appendix, SI Methods). Images were captured on a Leica TCS SPS X confocal laser scanning microscope using a 63× objective (SI Appendix, SI Methods).

Functional Annotation and Pangenomic Analysis of Bacterial Genomes. Anvi’o’s pangenomic workflow was used for orthogroup cluster grouping and functional comparisons (SI Appendix, SI Methods). We annotated all features containing conserved orthologous (ORF) regions (v2) with eggNOG database v5.0 (71). All ORFs were also annotated with Pfam domains (72) using HMMER v3.3 (73) (parameters in SI Appendix, SI Methods). All high-quality MAGs were also used to create pangenome in Anvi’o with an mcl inflation value of 8; pangenomic genomes database and profiles are on Figshare (74, 75). The MAGs were also annotated on the Rapid Annotation using Subsystem Technology (RAST) web server (https://rast.nmpdr.org) using the RASTTK pipeline (76). Where necessary, genes of interest were manually screened using NCBI blast+ v2.1.1.77 (77) and BBMAP to search for genes potentially missing from the assemblies (SI Appendix, SI Methods). To determine if any annotated functions were present in a given species group or population at a higher frequency than expected under an uniform distribution, EggNOG and Pfam terms were statistically tested for enrichment across different species groups and populations through the “ani-get-enriched-functions-per-pan-group” function in Anvi’o with an “adjusted q value” cutoff of 0.05 (SI Appendix, SI Methods).

C. orbicularis Metatranscriptome Analysis. Three C. orbicularis specimens were collected in Guadeloupe (2016), preserved in RNA later, and stored at −80 °C. The bins were checked for completion using Checkm’s taxonomy specific workflow and manually refined using “ani-refine” (SI Appendix, SI Methods). MAGs that were determined to be 90% or more complete and less than 10% contaminated post refinement, referred to as high-quality MAGs, were used for further analyses. Potential strain numbers in individual clams/metagenomes were calculated using DESMAN v2.1.1 (61) and the snakemake workflow (62) on the program (SI Appendix, SI Methods).

Expression Analysis Cores at the University of California, Davis Genome Center.

8 of 10 | PNAS
https://doi.org/10.1073/pnas.2104378118

Global biogeography of chemosynthetic symbionts reveals both localized and globally distributed symbiont groups

1. M. McFall-Ngai et al., Animals in a bacterial world, a new imperative for the life sciences. Proc. Natl. Acad. Sci. U.S.A. 110, 3229–3236 (2013).
2. C. M. Cavanaugh, Microbial symbiosis: Patterns of diversity in the marine environment. Integr. Comp. Biol. 34, 79–89 (2015).
3. E. A. Hambleton et al., Sterol transfer by atypical cholesterol-binding NPC2 proteins in coral-algal symbiosis. eLife 8, e43923 (2019).
4. M. L. Jones, A. pachyphila Jones: Observations on the Vestimentiferan Worm from the Galapagos Rift. Science 213, 333–336 (1981).
71. J. Huerta-Cepas et al., eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47 (D1), D309–D314 (2019).

72. S. El-Gebali et al., The Pfam protein families database in 2019. *Nucleic Acids Res.* 47 (D1), D427–D432 (2019).

73. Howard Hughes Medical, hmmscan: Search sequence(s) against a profile database (2011). hmmer.org.

74. J. T. Osvatic, L. G. E. Wilkins, Pangenome genomes database for Anvi’o. Figshare. https://dx.doi.org/10.6084/m9.figshare.13325744.v1. Deposited 3 December 2020.

75. J. T. Osvatic, L. G. E. Wilkins, Pangenome profiles database for Anvi’o. Figshare. https://dx.doi.org/10.6084/m9.figshare.13325723.v2. Deposited 3 December 2020.

76. T. Brettin et al., RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* 5, 8365 (2015).

77. C. Camacho et al., BLAST+: Architecture and applications. *BMC Bioinformatics* 10, 421 (2009).

78. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930 (2014).

79. A. E. Darling, B. Mau, N. T. Perna, progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PloS One* 5, e11147 (2010).

80. A. C. Darling, B. Mau, F. R. Blattner, N. T. Perna, Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403 (2004).

81. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113 (2004).

82. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973 (2009).

83. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313 (2014).

84. L. G. E. Wilkins, C. L. Ettinger, G. Jospin, J. A. Eisen, Metagenome-assembled genomes provide new insight into the microbial diversity of two thermal pools in Kamchatka, Russia. *Sci. Rep.* 9, 3059 (2019).

85. L. G. E. Wilkins, Ca. T. endolucinida core gene nucleotide alignment to infer recombination to mutation ratio. Figshare. https://dx.doi.org/10.6084/m9.figshare.13296200.v1. Deposited 27 November 2020.

86. L. G. E. Wilkins, Ca. T. weberae and Ca. T. lotti core gene nucleotide alignment to infer recombination to mutation ratio. Figshare. https://dx.doi.org/10.6084/m9.figshare.13296401.v1. Deposited 27 November 2020.

87. L. G. E. Wilkins, Ca. T. taylori core gene nucleotide alignment to infer recombination to mutation ratio. Figshare. https://dx.doi.org/10.6084/m9.figshare.13296065.v1. Deposited 27 November 2020.

88. N. J. Croucher et al., Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 43, e15 (2015).

89. L. G. E. Wilkins, J. T. Osvatic, B. Yuen, Analysis scripts (Jupyter Notebook). Figshare. https://dx.doi.org/10.6084/m9.figshare.13295912.v5. Deposited 27 November 2020.

90. L. G. E. Wilkins, Ca. T. taylori core gene nucleotide alignment corrected for recombination events to construct phylogenetic tree of strains (ecotypes). Figshare. https://dx.doi.org/10.6084/m9.figshare.13295912.v5. Deposited 27 November 2020.

91. J. Taylor and E. Glover, *Biology, Evolution and Generic Review of the Chemosymbiotic Bivalve Family Lucinidae* (Ray Society, 2021).

92. P. Pyne et al., Homologs from sulfur oxidation (Sox) and methanol dehydrogenation (Xox) enzyme systems collaborate to give rise to a novel pathway of chemo-lithotrophic tetrathionate oxidation. *Mol. Microbiol.* 18, 29669166 (2018).