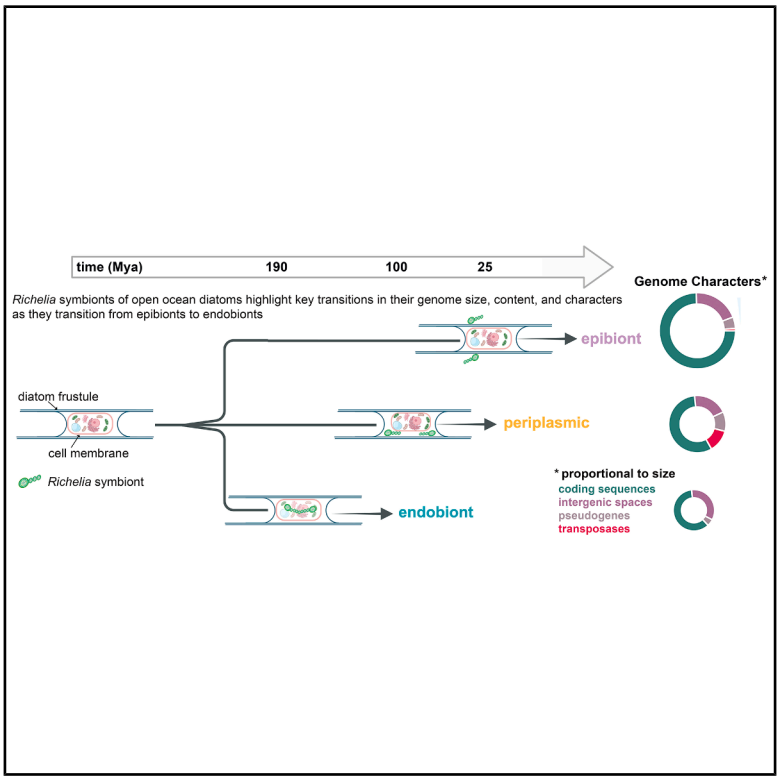
Article

Current Biology

Stepwise genome evolution from a facultative symbiont to an endosymbiont in the N_2-fixing diatom-*Richelia* symbioses

Graphical abstract



Highlights

- Diatom-*Richelia* symbioses are a valuable model for studying symbiont genome evolution
- Transposases and pseudogenes drive transitions in genomes of *Richelia* symbionts
- Pangenome analyses reveal core genome streamlining in *Richelia* endobionts
- Comparative genomics reveals host-driven evolution in symbiotic diazotrophs

Authors

Vesna Grujcic, Maliheh Mehrshad, Theo Vigil-Stenman, Daniel Lundin, Rachel A. Foster

Correspondence

vesna.grujcic@slu.se (V.G.), rachel.foster@su.se (R.A.F.)

In brief

Grujcic et al. compare genomes of N_2-fixing cyanobacterial symbionts of diatoms. They show that transitional stages toward host dependency involve genome reduction, pseudogenization, and loss of specific metabolic pathways. Their findings reveal how different symbiotic lifestyles shape genome reduction and increase host dependency in marine diazotrophic cyanobacteria.

Grujcic et al., 2025, Current Biology *35*, 4479–4493
September 22, 2025
https://doi.org/10.1016/j.cub.2025.08.003

CellPress

Current Biology

CellPress
OPEN ACCESS

Article

Stepwise genome evolution from a facultative symbiont to an endosymbiont in the N_2-fixing diatom-*Richelia* symbioses

Vesna Grujcic,[1,2,*] Maliheh Mehrshad,[2] Theo Vigil-Stenman,[1] Daniel Lundin,[3] and Rachel A. Foster[1,4,*]
[1]Department of Ecology, Environment, and Plant Sciences, Stockholm University, 106 91 Stockholm, Sweden
[2]Department of Aquatic Sciences and Assessment, Science for Life Laboratory, Swedish University of Agricultural Sciences, 750 07 Uppsala, Sweden
[3]Department of Biology and Environmental Science, Linnaeus University, 391 82 Kalmar, Sweden
[4]Lead contact
*Correspondence: vesna.grujcic@slu.se (V.G.), rachel.foster@su.se (R.A.F.)
https://doi.org/10.1016/j.cub.2025.08.003

SUMMARY

A few genera of diatoms that form stable partnerships with N_2-fixing filamentous cyanobacteria *Richelia* spp. are widespread in the open ocean. A unique feature of the diatom-*Richelia* symbioses is the symbiont cellular location spans a continuum of integration (epibiont, periplasmic, and endobiont) that is reflected in the symbiont genome size and content. In this study, we analyzed genomes derived from cultures and environmental metagenome-assembled genomes of *Richelia* symbionts, focusing on characters indicative of genome evolution. Our results show an enrichment of short-length transposases and pseudogenes in the periplasmic symbiont genomes, suggesting an active and transitionary period in genome evolution. By contrast, genomes of endobionts exhibited fewer transposases and pseudogenes, reflecting advanced stages of genome reduction. Pangenome analyses identified that endobionts streamline their genomes and retain most genes in the core genome, whereas periplasmic symbionts and epibionts maintain larger flexible genomes, indicating higher genomic plasticity compared with the genomes of endobionts. Functional gene comparisons with other N_2-fixing cyanobacteria revealed that *Richelia* endobionts have similar patterns of metabolic loss but are distinguished by the absence of specific pathways (e.g., cytochrome *bd* ubiquinol oxidase and lipid A) that increase both dependency and direct interactions with their respective hosts. In conclusion, our findings underscore the dynamic nature of genome reduction in N_2-fixing cyanobacterial symbionts and demonstrate the diatom-*Richelia* symbioses as a valuable and rare model to study genome evolution in the transitional stages from a free-living facultative symbiont to a host-dependent endobiont.

INTRODUCTION

Some of the most striking symbioses are widespread in the surface ocean and involve diverse single-celled eukaryotes (protists), which host a broad group of symbiotic partners, including bacteria, archaea, and other eukaryotes.[1–3] Despite their ecological importance, the specificity, functional roles, and evolutionary trajectories for many of these planktonic symbioses remain poorly understood. Few systems can be maintained in stable culture for extended periods (>2 years), and this limitation has hindered the establishment of robust model systems for studying planktonic symbioses. While more than a century of observations has highlighted the prevalence of planktonic symbioses and their importance in nutrient cycling,[4,5] detailed genomic and evolutionary studies on protist symbioses remain sparse. Yet, in the several examples of organellogenesis (mitochondria, plastid, and chromatophore), including the recently discovered nitroplast (N_2-fixing organelle), the newly formed organelles were derived from endosymbiotic events involving protists and prokaryotes.[6–9] Thus, modern planktonic symbioses provide unique opportunities for studying genome and organelle evolution.

One fascinating microbial symbiotic system involves a few genera of diatoms as hosts and three species of the terminal heterocyst-forming cyanobacteria *Richelia* as symbionts.[10] Heterocysts are specialized cells for N_2 fixation; thus, the role of *Richelia* as a nitrogen source is obvious and has been shown on the cellular level.[11] The number and length of *Richelia* filaments vary in each of the symbioses. In general, two short filaments (3–4 cells/filament) of *Richelia* are observed associated with the *Hemiaulus* spp., while two and up to 35 longer filaments (5–6 cells/filament) have been reported with the *Rhizosolenia* spp. host diatoms.[12] In *Chaetoceros* sp. diatoms, the number of *Richelia* symbionts can vary dramatically as well (e.g., 1–9), and filaments tend to be short.[13,14]

Diatoms are widespread and highly diverse photosynthetic eukaryotic microalgae that are important primary producers in the modern oceans and contribute to carbon (C) burial.[15]

Genomic and ultrastructural evidence suggests diatom plastids are derived from a secondary endosymbiosis involving a red algal endosymbiont.[16,17] Furthermore, horizontal gene transfer (HGT) has contributed to diatom evolution by introducing numerous bacteria-derived genes, while diatom-specific transposases are also recognized as important mechanisms for gene acquisitions and losses.[18]

The sequencing of the first four genomes of the symbiotic *Richelia* spp. highlighted how the symbiont's cellular location has influenced its genome size and genetic content.[19–21] Several environmental metagenome-assembled genomes (MAGs) have been reported that are closely related to *Richelia*.[10] The two endosymbiotic *R. euintracellularis* strains that associate with *Hemiaulus* spp. diatoms possess the smallest genomes (ReuHH01; 3.2 Mbp, ReuHM01; 2.2 Mbp) with the lowest guanine plus cytosine (GC) content (33.7% and 33.8%, respectively) and lack several nitrogen assimilatory pathways common to free-living heterocyst-forming cyanobacteria.[19,20] The partially integrated symbiont, *R. intracellularis*, that resides in the periplasmic space of *Rhizosolenia* spp. diatoms has a slightly larger genome and higher GC content (RintRC01; 5.48 Mbp; 39.2%). The periplasm of a diatom refers to the space between the outer silicified cell wall (frustule) and the inner cell membrane of the diatom. *Richelia rhizosoleniae* are epibionts that attach to the outside of *Chaetoceros* sp. diatoms and possess the largest genome with the highest GC content (RrhiSC01; 5.97 Mbp and 39.5%) of all *Richelia* spp. They are also the only symbionts that can be maintained freely in culture.[22]

Small genomes are common among obligate symbionts, and typically their genomes lack DNA-repair genes and have decreased GC content.[23,24] Symbiont genomes also tend to lose genes that encode functional pathways redundant with their respective hosts while retaining essential genes.[25,26] Earlier comparative genomic studies on symbiotic bacteria of insects and some ciliates[27–29] identified that the initial stages of genome degradation involve the proliferation of transposases, followed by gene inactivation, pseudogene accumulation, and genomic deletion.[30,31] Transposases can drive genome evolution as well by enabling the movement and/or interruption, duplication, and rearrangement of genetic information.[32] As symbionts transition to a more obligate and permanent state (e.g., toward an organelle state), the prevalence of transposases and intergenic regions tends to decrease.[26] Over time, ongoing deletions remove pseudogene fragments, and gene loss continues, resulting in small and more compact genomes.[33,34]

Collectively, the diatom-*Richelia* symbioses provide an ideal system for investigating the evolutionary trajectory of symbiont genomes due to the continuum of symbiont cellular integration. However, the impact of this integration on the symbiont genome content remains poorly characterized. In this study, we analyzed four *Richelia* spp. genomes derived from cultures and ten MAGs using comparative genomics analyses. We focused on parameters indicative of ongoing genome reduction processes, including changes in intergenic spacer (IGS) size, the presence of transposases, and pseudogenes. Finally, we compared the functional gene content of *Richelia* with other obligate endosymbiotic cyanobacteria to identify common patterns of functional gene retention and loss in N_2-fixing symbionts.

RESULTS AND DISCUSSION

Diversity and specificity of diatom-*Richelia* symbiosis

The cellular location and associated host for the four *Richelia* draft assemblies were previously reported.[13,19,20,35,36] To establish which species each MAG belongs to and, by extension, to predict their expected cellular location (endobiont, periplasmic, or epibiont) and associated host, we performed a phylogenomic analysis using the Genome Taxonomy Database toolkit (GTDB-Tk), which uses 120 conserved marker proteins to establish a *de novo* phylogeny that places each unknown genome in the phylogenetic context of the Genome Taxonomy Database (GTDB) taxonomy.[37] The full phylogenetic tree of *Nostocaceae* (the family to which *Richelia* belongs to) is illustrated in Data S1. The new phylogeny agrees with and expands upon earlier studies based on single marker genes, which separates the various *Richelia* spp. according to host and corresponding cellular location.[38,39] Three additional MAGs together with the epibiont *R. rhizosoleniae* (hereafter RrhiSC01) form a sister clade to the periplasmic symbionts and endobionts (Figure 1). The genome of the epibiont RrhiSC01 falls outside of GTDB's "relative evolutionary divergence" criterion for belonging in the *Richelia* genus. Moreover, the RrhiSC01 does not fall in the *Calothrix* part of the tree, and the genus name under which it was formerly known.[22] In cases like this, when a new taxon name is needed but in the absence of a formal description of the taxon, GTDB assigns placeholder names characterized by an underscore and a single capital letter. The clade in which RrhiSC01 belongs was hence given the name Calothrix_A. Recently, we described the *Richelia* genus in the Bergey's Manual of Systematics of Archaea and Bacteria and decided to include RrhiSC01 in the genus because of its close phylogenetic relationship to *Richelia* spp. and its association with diatoms.[10] To be consistent with this choice, we here name the other species in this clade *Richelia* spp., awaiting a potential future formal description of the genus. Based on their phylogenetic position, we cannot draw any conclusions about the potential association with diatoms for the other genomes in the RrhiSC01 clade, which are derived from marine sediment samples in a fringing reef,[40] but note that the common ancestor of all species we here call *Richelia* spp. potentially had an association with diatoms, presumably based on N_2 fixation. The five MAGs that form a clade with the two known *Richelia* endobiont genomes (ReuHH01 and ReuHM01) possess >95% average nucleotide identity (ANI) and are considered to be the same species and therefore endobionts. The DT-104 MAG that clustered with RintRC01 at 98% ANI is considered a periplasmic symbiont. One additional MAG (Candidatus *Richelia exalis*, hereafter TARA_PON), reported from the Tara Oceans project,[41] formed a sister clade to the endobionts. However, ANI was <95%, and therefore the symbiont cellular location and host association remain unknown (Figure 1; Table S1). We do, however, include TARA_PON and the three other MAGs (MO_192.B10, MO_167.B12, and MO_167.B42) in the analyses for comparison with their closest *Richelia* spp. relatives (RintRC01 and RrhiSC01, respectively). Shortened genome identifiers used throughout the text are listed together with their full names and accession numbers in Table S1.

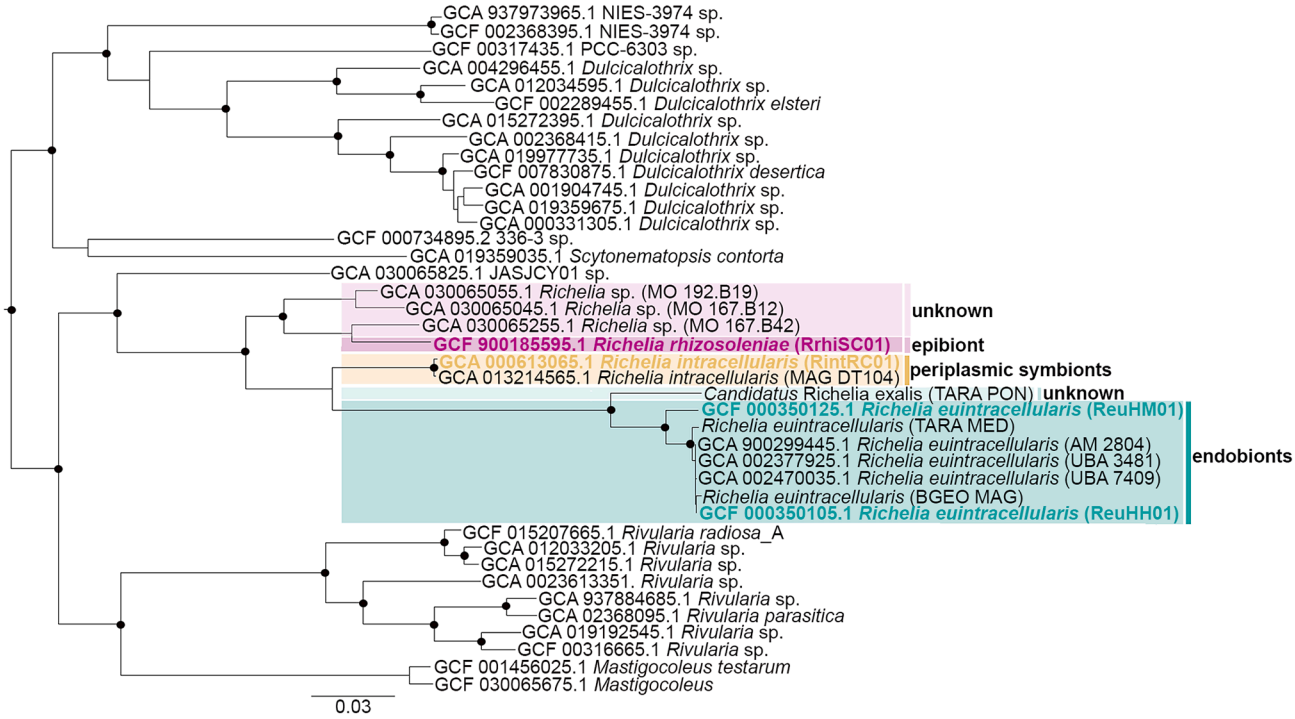


Figure 1. Phylogenetic reconstruction of *Richelia* spp. and its closest relatives within the *Nostocaceae* family

The reference *Richelia* spp. (draft) assemblies that were derived from enrichment cultures are highlighted in bold text, and MAGs are named according to the GTDB database and phylogeny. Shortened identifiers used throughout this study are shown in parentheses after the genome names. MAGs are colored according to their presumed cellular location with their respective host diatoms: epibiont (purple), periplasmic symbiont (yellow), endobionts (green), and unknown (light purple, light green). The tree was reconstructed using the GTDB-Tk tool in "*de novo*" mode, using 120 concatenated conserved marker proteins.[37] Scale bar refers to a phylogenetic distance of 0.03 nucleotide substitutions per site. The root of the tree follows the assumption that the *Nostocaceae* family in the full tree is monophyletic. Support values of 95 and greater are represented by a black circle on the branching point. The displayed tree was extracted from the complete bacterial tree. An extraction of the complete *Nostocaceae* family is illustrated in Data S1. See STAR Methods for details on the tree construction.

Coding and noncoding fractions reflect genome degradation stages in *Richelia* symbionts

Genome statistics were calculated for the four *Richelia* genomes derived from the cultures and ten MAGs (Figure 2; Table S1; Figure S1). There is a direct correlation between genome size and GC content in the *Richelia* spp.: endobionts possess smaller genomes and lower GC content (3.39 Mb ± 0.34 and 34.03% ± 0.88%) compared with periplasmic symbionts (5.17 Mb ± 0.78 and 39% ± 0.07%) and the epibiont (5.98 Mb and 40%) (Figure 2A; Table S1). The number and percentage of coding sequences (CDSs) follows a similar trend where genomes of endobionts have fewer CDSs (2,038 ± 175; 56% ± 4.81%) compared with periplasmic symbionts (6,029 ± 1,548; 67% ± 1.50%) and the epibiont (4,954; 76.09%) (Figures 2B, 2C, and S1A; Table S1).

To identify other features indicative of genome degradation, we examined the presence, abundance, and size of genomic regions associated with transposases. Insertion sequences (ISs), which typically consist of a transposase gene flanked by terminal inverted repeats, are widespread in bacterial and archaeal genomes,[42–44] but their proportion in prokaryotic genomes is usually below 3%.[45,46] Furthermore, ISs tend to proliferate in endosymbiotic microbial genomes, particularly those that have recently transitioned to a host-restricted lifestyle.[46]

We used Transposeek2, a BLASTx (basic local alignment search tool for translated nucleotide sequences using six-frame translation)-based pipeline, to identify transposase sequences by searching genomes and MAGs against a curated database (ISfinder).[47] Importantly, our analyses focused on the protein-coding region of the transposase genes and did not include the flanking terminal inverted repeats that define full IS elements. The number (1.56 ± 1.13), median length (231 ± 107 bp), and proportion (0.02% ± 0.02%) of transposase encoding regions in the genomes of the endobionts (Figures 2D–2F; Table S1) were very low and similar to that reported for other obligate symbionts (e.g., symbionts of insects, clams, and amoebae[48–50]). The same parameters for transposases in the epibiont genome were similarly low (Figures 2D–2F; Table S1). By contrast, the periplasmic symbiont genomes (RintRC01, MAG DT-104), however, contained the highest number of transposases among all genomes examined (1,537 ± 684), with a longer median length (251 ± 28 bp), resulting in an unusually high fraction of their genomes occupied by transposases (14.58% ± 9%) (Figures 2D–2F and S1; Tables S1 and S2).

Given the latter results, we further compared the transposases of the periplasmic symbionts to those of other microbial endosymbionts reported with high abundances (e.g., Figure 3; Table S2). Notably, the one genome (RintRC01) contained the highest detected percentage (21%) of transposases of any

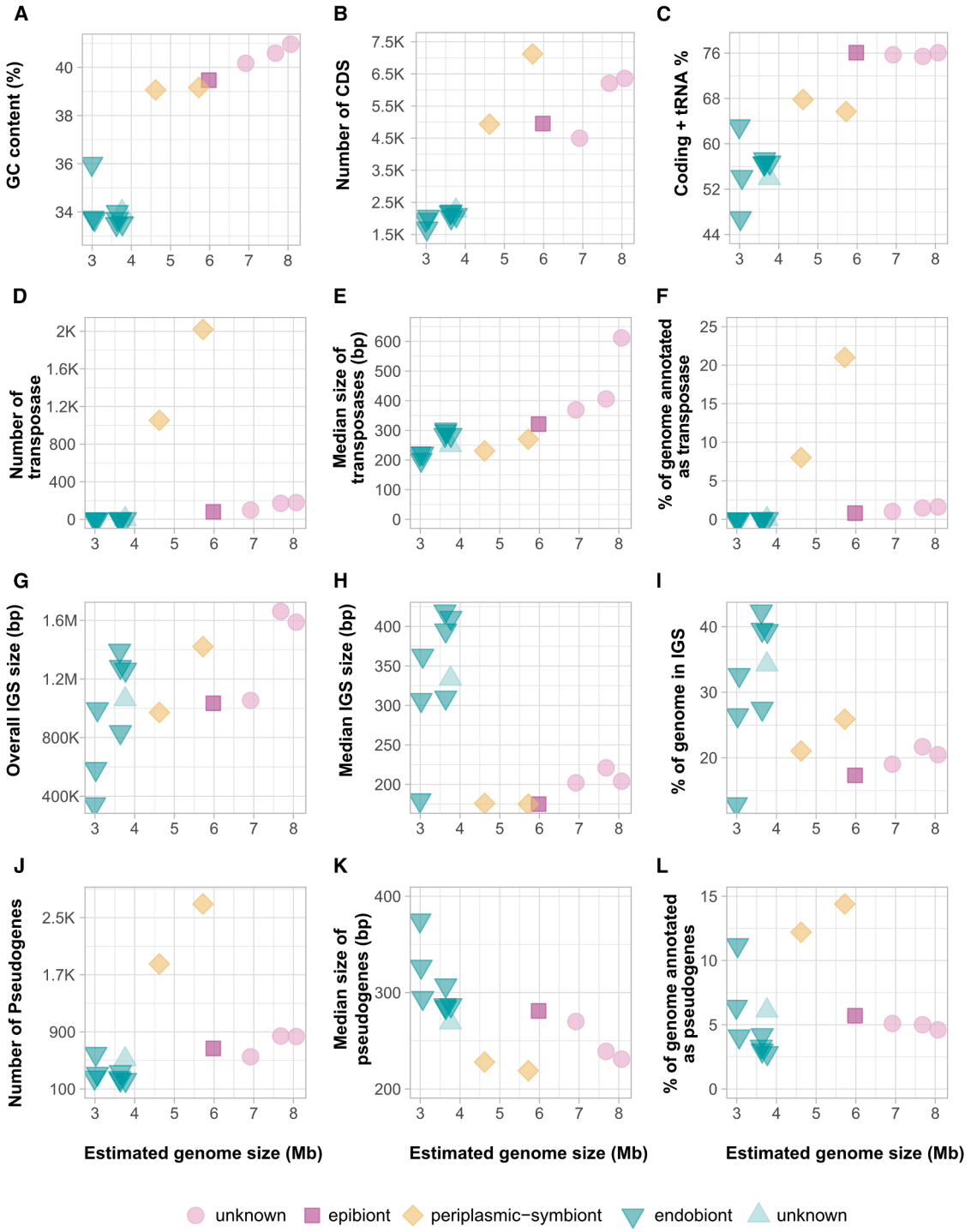


Figure 2. Key genomic features of *Richelia* spp. genomes and MAGs in relation to genome size (Mb)

(A–L) (A) GC content (%), (B) number of coding sequences (CDSs), (C) overall CDS + tRNA percentage in the genome (%), (D) number of transposase, (E) median length of transposase (bp), (F) overall length of transposases as a percentage (%) of the genome, (G) overall length of intergenic spacers (IGS) (bp), (H) median length of IGS (bp), (I) overall length of IGS as a percentage of the genome, (J) number of pseudogenes, (K) median length of pseudogenes (bp), and (L) overall length of pseudogenes as a percentage (%) of the genome. Genomes are grouped according to their cellular location: unknown (circles, triangle facing up), epibiont (square), periplasmic symbionts (diamonds), and endobionts (triangle facing down). Detailed statistics for each genome are provided in Table S1. Additional genomics features in relation to genome size are presented in Figure S1.

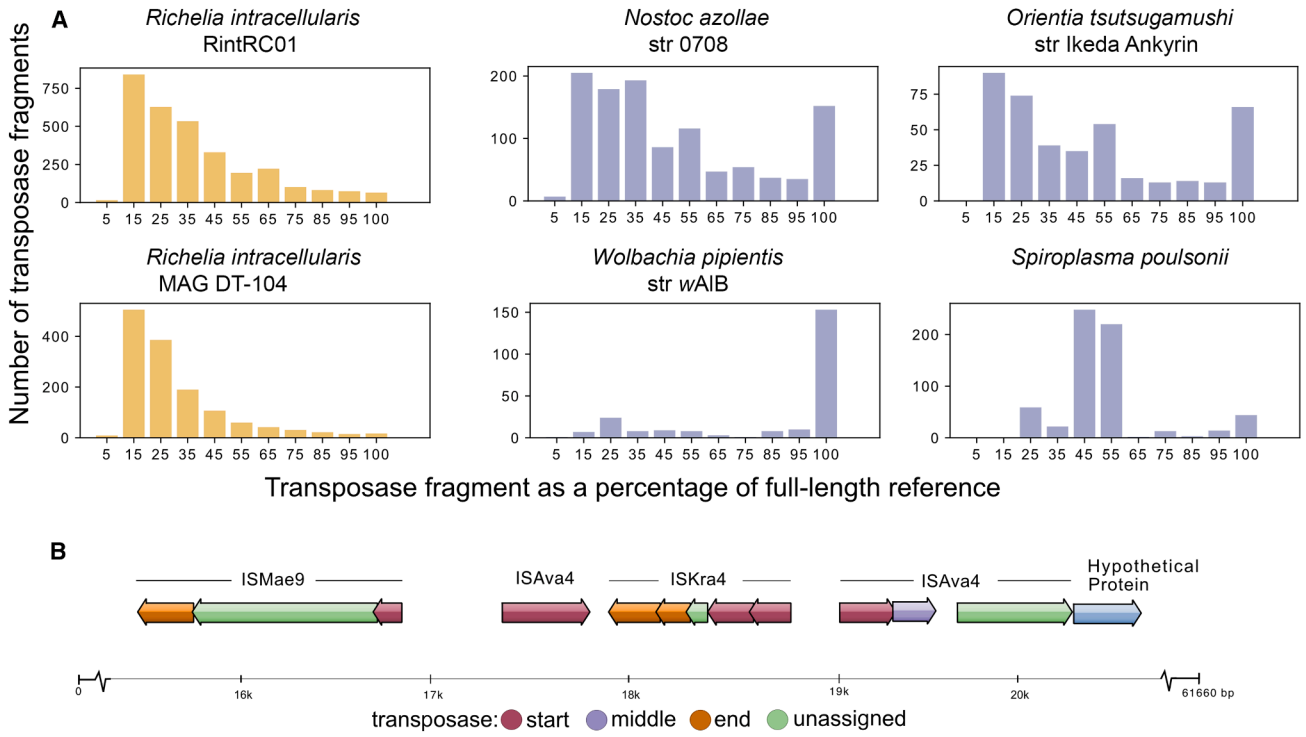


Figure 3. Length distribution and fragmentation of transposase sequences

(A) Distribution of transposase fragment lengths in the genomes of periplasmic *Richelia* symbionts (RintRC01 and MAG DT-104) compared with other IS-rich obligate symbionts. Fragment lengths are shown as a percentage of the full-length reference transposase sequences from ISFinder.

(B) Annotated IS-dense genomic region from contig650 of the RintRC01 genome. Each transposase-associated element is labeled with its best ISFinder match. Detailed statistics of different transposase families are illustrated in Figure S2 and Table S2.

known prokaryotic genome. In our analyses, we also identified an unusual profile of transposase length (Figure 3). Although IS-rich genomes typically contain a high number of full-length transposases, we found that most transposase-associated sequences in RintRC01 and MAG DT-104 were truncated, with only 65 and 17 full-length matches, respectively. Instead, the majority of transposases were detected as short fragments, and most of them encoded less than 15% of the full-length transposase. Upon closer inspection, some of the fragments were in clusters (Figure 3B), suggesting they are remnants of highly degraded IS elements or the result of multiple IS insertions. This pattern is similar to what has been observed in other endosymbionts, such as *Wolbachia*, where transposase-rich genomes accumulate numerous truncated and degraded elements, likely due to reduced transpositional activity after an initial phase of proliferation.[51]

To further explore the diversity of transposases in the RintRC01 genome, we classified each transposase sequence according to family. We discovered in total 14 families and 10 groups (Figures S2A and S2B). Two families with the highest representation were ISKra4 (n = 813) and IS5 (n = 663) (Figures S2A, S2B, and S2D). Additionally, we were interested to investigate which of the transposases are potentially functional by comparing the length of the transposase sequences to the full-length reference transposase proteins (Figures S2E and S2F). The low number of full-length (Figure S2C), potentially functional homologs, combined with the high number of short, fragmented sequences (Figures S2E and S2F), suggests that transposase activity in the RintRC01 genome is currently in decline after a period of proliferation.

Following examination of CDSs, we calculated statistics for IGS, which tend to be less constrained by selective pressure compared with CDS. The general trend was that IGS length was consistent with genome size and symbiont cellular location; therefore, the longest IGS was present in the epibiont (1,034,196 bp), followed by the periplasmic symbionts (1,196,254 ± 317,383 bp) and finally the endobionts (957,936 ± 310,050 bp) (Figure 2G; Table S1). Surprisingly, both the median length of IGS (341 ± 84 bp; Figure 2H; Table S1) and the percentage of IGS in the endobiont genomes (31.55% ± 10.26%; Figure 2I; Table S1) were the highest of all symbionts (Figure 2I; Table S1). Notably, the IGS percentage in the endobionts is approximately twice the number reported for free-living bacteria.[52] These large IGS regions in the endobiont genomes show no homology to known genes and are void of transposases. Combined, we interpret the latter results as evidence for an intermediate stage of genome degradation, where formerly functional genes have been inactivated but not yet purged from the genome.[31,53] However, it is important to note that the overall IGS proportion varied among endobiont genomes (12%–42%; Figure 2I) and further suggests that among the endobionts there are different stages of genome degradation, which should be expected as the MAGs are derived from environmental populations. Genomes with the low IGS could already be in advanced stages of genome degradation with an increased host dependency.

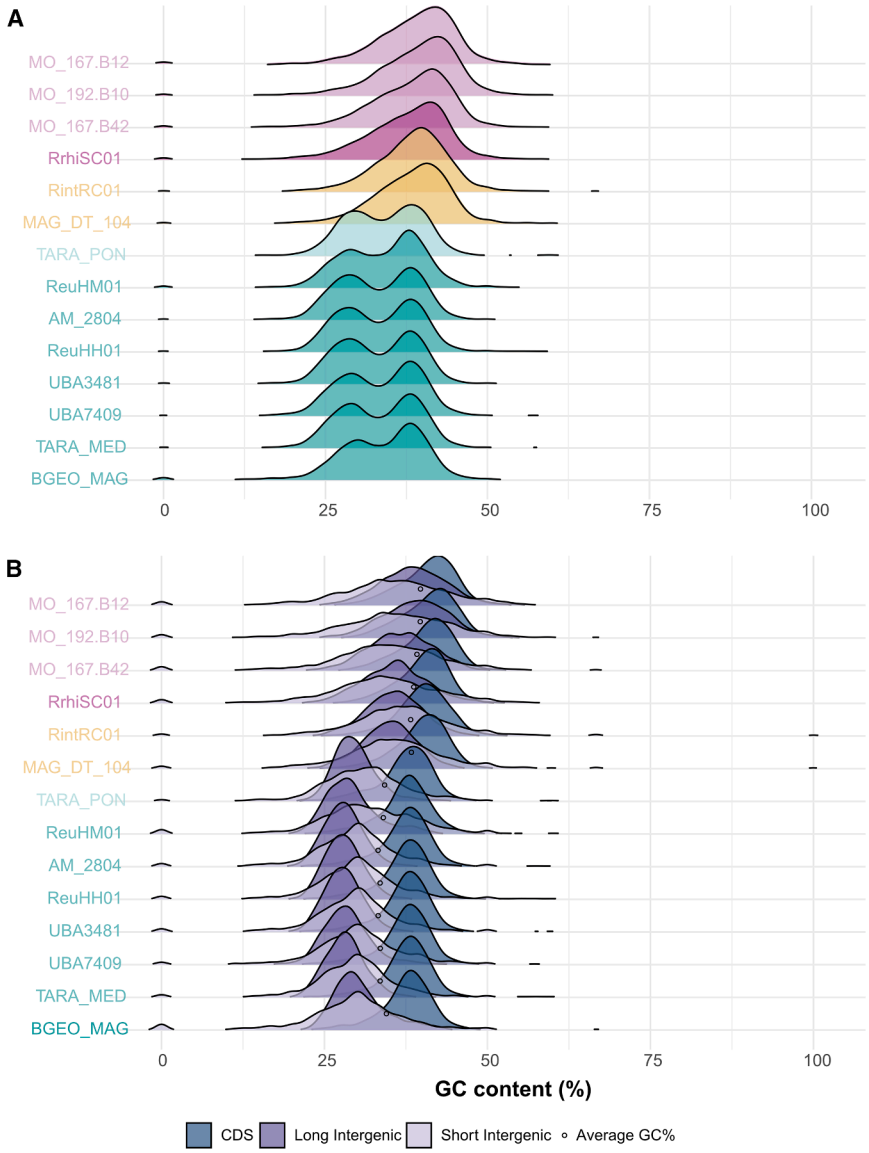


Figure 4. Distribution of GC content in CDSs and IGSs in *Richelia* spp. genomes and MAGs

(A) Ridge density plots showing the distribution of GC content (%) for all genomic regions. Genomes are color-coded according to their cellular location: unknown (light purple, light green), epibiont (purple), periplasmic symbionts (yellow), and endobionts (green).

(B) GC content distributions for three genomic categories: CDSs (blue), long IGSs (≥300 bp; purple), and short IGSs (<300 bp; light purple). Each category is plotted as a density ridge, and black dots indicate the average GC content (%) of each genome. See also Figure S3.

Variation of the IGS in the *Richelia* genomes raises important questions about the role of pseudogenes in their genome degradation. Pseudogenization is a key mechanism for gene loss and results from the accumulation of mutations in protein-CDSs and can often lead to the introduction of premature stop codons.[23,54] In prokaryotes, pseudogenes typically make up between 1% and 5% of the genome,[55] indicating there is purifying selection to keep genes functional.[54] However, intracellular pathogens and endosymbionts in transitional stages of genome reduction often exhibit high numbers of pseudogenes (10%–50% of their genome), which reduce their coding capacity significantly.[56,57]

The prevalence of pseudogenes varied in the *Richelia* genomes and MAGs (Table S1). As expected, the highest number of pseudogenes was found in the periplasmic symbionts (2,272 ± 594), followed by the epibiont (669) and endobionts (318 ± 128) (Figure 2J; Table S1). While endobionts contained fewer pseudogenes overall, they had the largest median pseudogene size (309 ± 33 bp) (Figure 2K; Table S1). This is likely due to a slower rate of DNA loss, reflecting reduced selective pressure for genome streamlining at later stages of symbiosis.[58] Alternatively, differences in pseudogene length may result from lineage-specific mechanisms of genome erosion, which tend to vary in rate and mode across symbiotic systems.

Despite their smaller pseudogene size, periplasmic symbionts had the highest proportion of pseudogenes relative to their genome assembly size (13.3% ± 1.5%), which was a significantly larger fraction compared with the epibiont (5.75%) and endobionts (5.02% ± 2.97%) (Figure 2L; Table S1). We interpret the high prevalence of pseudogenes in the periplasmic symbiont genomes, along with the abundance of IS elements and relatively large genome size, as indicative of early-stage genome reduction. The latter is also consistent with the stepwise model of genome reduction proposed by Lo et al.[31] By contrast, the low number of pseudogenes and fewer CDSs in the endobiont genomes suggests stronger selection for genome streamlining, possibly to minimize replication costs during cell division and decrease redundancy with functions of their host diatom. Additionally, over 40% of IGS in some of the endobionts were composed of detectable pseudogenes (Figures S1D–S1F; Table S3), potentially resulting from a slower rate of deletion or an accumulation of larger, non-functional genomic regions.[59] The lack of purifying selection for function in IGS regions often leads to an increased mutational bias toward adenine-thymine (AT) richness in, e.g., endosymbionts of insects and some obligate pathogenic bacteria (e.g., *Rickettsiales* and *Chlamydiales*).[25] This same pattern of AT richness was observed in the *Richelia* endobiont genomes (Figure 4).

Lower GC content in endobionts is driven by non-coding regions

The distribution frequency of GC content in a genome can reveal underlying processes such as mutational biases and

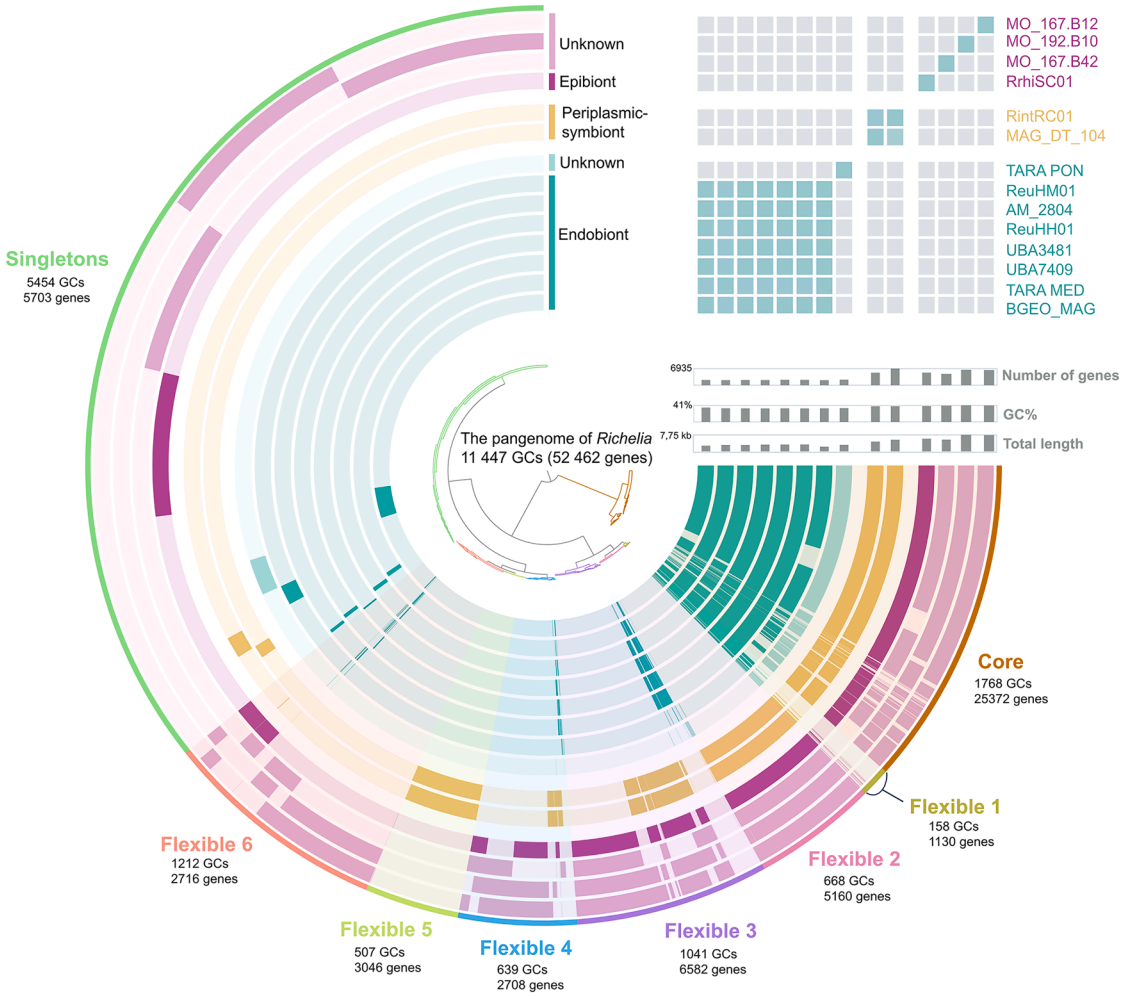


Figure 5. The pangenome of genus *Richelia* spp. and close relatives

Pangenome covers 52,462 genes and 11,447 gene clusters from four *Richelia* genomes derived from enrichment cultures and ten environmental MAGs. Genomes are organized based on their placement in the phylogenetic tree illustrated in Figure 1. Colors correspond to different cellular locations: epibiont, periplasmic symbionts, endobionts, and unknown. The top right corner shows genome clustering based on average nucleotide identity (ANI), with a 95% threshold delineating eight species-level groups. For detailed numbers and percentages of core, flexible, and singleton gene clusters in each genome, check Table S3. Distribution of gene clusters across KEGG functional categories in the pangenome is represented in Figures S4 and S5 and Table S4.

selection pressures.[60] Our analyses showed a unimodal distribution of GC content in the *Richelia* epibiont and periplasmic symbiont genomes. The *Richelia* endobionts, however, exhibit a bimodal distribution (Figure 4A). By further separating CDSs and IGS sequences and categorizing IGS sequences by length (with a 300 bp threshold), we identified that the higher GC peak in the endobionts corresponds to CDS and the lower GC peak is primarily found in the IGS (Figures 4B and S3). This bimodal pattern, associated with genome reduction in obligate symbionts, reflects increased genetic drift and relaxed selection pressures.[23,26] The median sizes of IGS in the endobionts are longer compared with the IGS of the periplasmic symbionts and the epibiont, which, alongside fewer CDSs, contributes to their lower overall GC content. By contrast, the median CDS length is 767 bp in the epibiont, 374 ± 66 bp in periplasmic symbionts, and 668 ± 94 bp in endobionts (Table S1). The low GC content and low similarity of the IGS to known genes are indicative of *Richelia* endobionts being in more advanced stages of genome reduction compared with the other *Richelia* symbionts. Furthermore, the latter pattern is commonly observed in symbionts that have been in prolonged relationships with their hosts.[59,61]

Symbiotic lifestyle shapes the pangenome of *Richelia*

Our pangenome analysis of the *Richelia* genomes and closely related MAGs identified 11,447 unique gene clusters from a total of 52,462 genes (Figure 5). The unique gene clusters were categorized into eight bins based on their occurrences across the different genomes. The core genome bin contained 1,768 gene clusters (15.4%) present in all genomes, and we identified 6 flexible (sometimes referred to as accessory) genome bins based on their distribution in the pangenome (37%): flexible 1 (n = 158), flexible 2 (n = 668), flexible 3 (n = 1,041), flexible 4 (n = 639), flexible 5 (n = 507), and flexible 6 (n = 1,212) (Figure 5). Finally, gene

clusters unique to individual genomes were categorized as singletons (n = 5,454; 47.6%) (Figure 5).

The core genome represents functions conserved across all *Richelia* strains, which are critical for their survival.[62] The flexible genome likely reflects the adaptability to various environmental conditions, including, for *Richelia*, their specific cellular location. A high proportion of singletons (nearly 50%) indicates significant genomic plasticity, suggesting that individual genomes may retain and/or acquire unique functions for specific purposes.[63] Interestingly, a recent study[64] highlighted that a bacterium's lifestyle, particularly the degree of host integration, plays a significant role in shaping pangenome fluidity (also called genome fluidity). Pangenome fluidity refers to the average proportion of genes unique to any two genomes of the same species.[64] *Richelia* genomes exhibit different degrees of fluidity. For example, endobionts have a higher proportion of their gene clusters in the core (71.98% ± 4.72%) and far fewer gene clusters in the flexible (10.85% ± 4.33%) and singleton (6.06% ± 7.54%) parts of the pangenome (Table S4). Thus, the endobionts possess a more conserved pangenome structure, which is consistent with their presumably stable, intracellular environment. By contrast, greater pangenome fluidity was observed in the epibiont and periplasmic symbionts, as evidenced by their larger flexible genome fractions (47.32% ± 3% and 49.46% ± 0.32%, respectively) (Table S4). Unlike the endobionts, both the epibiont and periplasmic symbionts exist in a more variable environment and potentially have different interactions with their hosts (e.g., competition and cooperation).

To better understand the functional capacity of these genomes, we annotated gene clusters using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. More than 60% of the genes could be assigned to functional categories (Figure S4). Most of the annotated gene clusters were part of the core genome, while many of the unannotated genes were associated with the flexible or singleton parts of the pangenome (Figure S5). Our comparison of the functional genomic content retained and lost in different parts of the pangenome further underscored the impact of the different cellular locations on the *Richelia* symbiont genomes.

We started by examining functions related to C and N metabolism, given that the symbionts function as a N source and N_2 fixation in heterocyst-forming cyanobacteria is primarily fueled by photosynthesis. We detected the 12 *nif* genes for the nitrogenase complex required for N_2 fixation in the core genome: *nifH* (K02588), *nifD* (K02586), *nifK* (K02591), *nifE* (K02587), *nifN* (K02592), *nifX* (K02596), *nifB* (K02585), iscS/NFS1 (K04487), *nifT* (K02593), *nifV* (K02594), *nifW* (K02595), and *nifZ* (K02597). Out of the 12 *nif* genes, seven genes (*nifUBENXSV*) encode for biosynthesis of the iron molybdenum cofactor, which is essential for the catalytic activity of nitrogenase in the conversion of atmospheric N_2 to ammonia. One gene, *nifJ* (K03737), which encodes pyruvate (flavodoxin) oxidoreductase (PFOR), was detected only in the flexible 3 part of the pangenome of both the periplasmic symbionts and the epibiont genome (and the three related MAGs) but absent in all endobiont genomes. Without *nifJ*, the endobionts may rely solely on the pyruvate dehydrogenase complex (PDC) for pyruvate oxidation under oxic conditions, rather than utilizing PFOR's flavodoxin-dependent pathway, which is typically important under anaerobic or iron-limited conditions.[65] In endobionts, *nifJ* gene loss likely reflects an adaptation to the stable environment provided inside the host cytoplasm, where oxygen sensitivity of PFOR and a more stable iron availability could make the ferredoxin pathway sufficient for nitrogenase (or hydrogenase) activity without the need for a flavodoxin-based reduction.

Additionally, genomes of both the periplasmic symbionts and endobionts have lost genes associated with assimilatory nitrate reduction (*narB*, K00367; *nirA*, K00366) that remain in the flexible 3 region of the pangenome in epibiont RrhiSC01. For comparison, both *narB* and *nirA* are also absent from the spheroid body of the freshwater diatom *Epithemia turgida* (EtSB) and *Candidatus Atelocyanobacterium thalassa* (hereafter UCYN-A) genomes, but *nirA* is retained in the genome of the heterocyst-forming cyanobacteria *Nostoc azollae* endobiont. These findings highlight the influence of both the host and the oligotrophic natural environment in which these symbioses reside on their N assimilation strategies.

Recent evidence has shown that the *Richelia* spp. symbionts, including the periplasmic symbionts, differ in their C metabolic activity and host dependency.[66,67] To explore this further, we examined the photosynthesis module in KEGG and identified 45 photosynthetic genes in the core, 15 genes in the flexible, and no genes in the singleton part of the pangenome. Among the 15 genes in the flexible section, 6 were not redundant with core genes and included 4 genes that encode functions for Photosystem II (PSII) (*PsbJ*, *PsbP*, *PsbT*, and *Psb28-2*; K02711, K02717, K02718, and K08904), 1 gene for Photosystem I (PSI) (*Psal* and K02696), and 1 gene required for electron transport (*PetJ* and K08906). With the exception of *PsbT* (annotated in flexible1 of some endobionts), all six genes were missing from the genomes of endobionts but present in the genomes of the periplasmic symbionts, the epibiont, and related MAGs. The notable deletion of *PetJ*, which encodes cytochrome *c6*, in the endobiont genomes likely reflects an adaptation to the controlled intracellular environment of the host. In free-living heterocyst-forming cyanobacteria, cytochrome *c6* is essential in heterocysts, where it serves as the primary soluble electron donor to Cox2 under copper-replete conditions—a role that plastocyanin (PetE and K02638, present in all analyzed genomes) cannot fully substitute.[68,69] In vegetative cells, *PetJ* is typically expressed under copper-limited conditions, when plastocyanin cannot function due to insufficient copper availability.[69] However, *Richelia* endobionts reside within the diatom host cytoplasm[36] but are not enclosed by a host-derived membrane, which could provide access to both consistent electron donors and sufficient copper that would make *PetJ* redundant.

Another important aspect of photosynthesis is light capture. The light-harvesting proteins of cyanobacteria are organized as antennae called phycobilisomes, which are arranged on the thylakoid membranes. Interestingly, within the phycobilisome module, 18 of the genes were detected within the core genome, and 3 genes were identified in flexible regions of the periplasmic symbiont and epibiont genomes. Three of the latter genes were missing in the endobionts and included *cpcD* (K02287), *cpcE* (K02288), and *cpcF* (K02289), which are involved in the synthesis and assembly of phycocyanin, a key component of the phycobilisomes that captures light energy and transfers it to the reaction centers of photosystems I and II.[70] The absence of genes

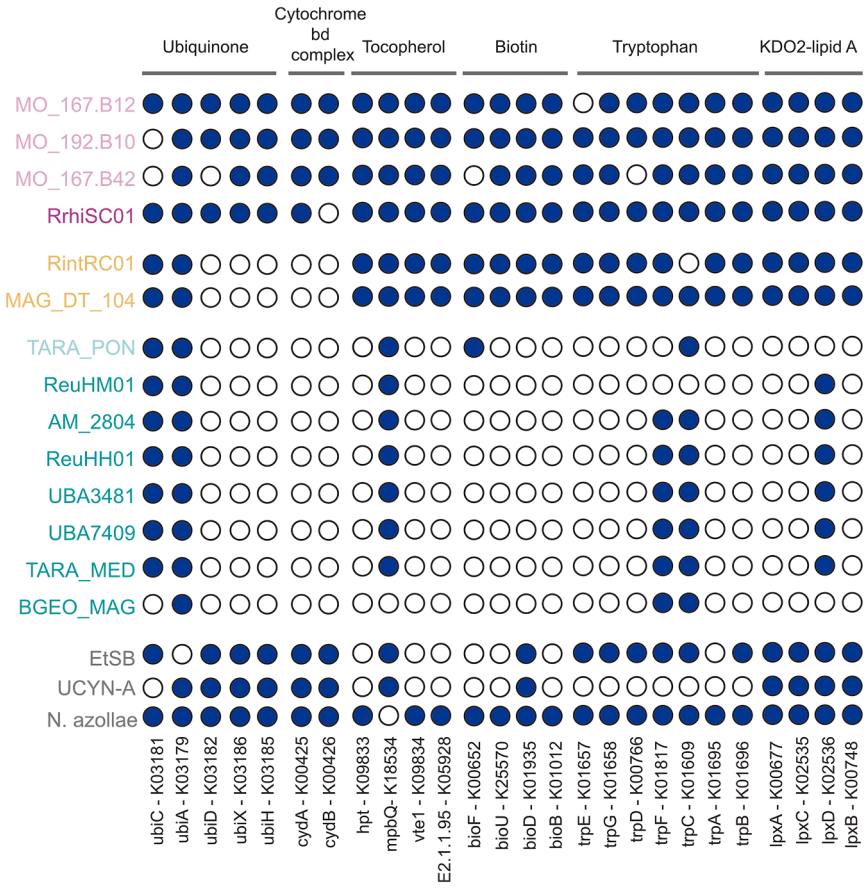


Figure 6. Presence and absence of KEGG-annotated genes involved in selected biosynthetic pathways

Columns represent individual genes identified by their KEGG Orthology (KO) numbers and corresponding gene names. *Richelia* genomes are color-coded according to their cellular locations, and the three other genomes belonging to obligate N_2-fixing endobionts and the nitroplast (UCYN-A) are colored in gray. Gene presence is marked as a blue circle and absence as a white circle. For additional functional analyses.

See also Figures S6 and S7.

important for phycocyanin, along with the loss of some genes in the photosynthesis module, suggests that endobionts have reduced light-capturing capacity and decreased genome content for their own photosynthetic apparatus. A higher dependency and capacity to obtain organic C (e.g., sugars) from the host diatom was recently confirmed experimentally and in wild populations of the *Hemiaulus-Richelia* (ReuHH01) symbioses.[66,67] The latter combined with our functional analyses here provides further evidence for a lower investment of the endobionts to perform their own photosynthesis.

Functional comparison with other obligate symbionts highlights that gene retention and loss differ among N_2-fixing endobionts

In order to identify how genome reduction and metabolic dependency have evolved in N_2-fixing cyanobacterial symbioses, we compared the genome content of the *Richelia* symbionts to other obligate N_2-fixing symbionts. These include the EtSB, the UCYN-A/nitroplast, and *N. azollae*, a heterocyst-forming cyanobacteria that is an obligate endosymbiont of a water fern (Figure 6).

Under symbiotic conditions, coding regions that provide little or no added value in a given environment may be lost.[46,71] Similarly, as observed in bacterial endosymbionts of insects,[23,46,71] repeated population bottlenecks may weaken selection pressures, even for essential genes,[25] and lead to the elimination of dispensable genes through genetic drift.[72] We identified several such examples, e.g., the *Richelia* endobionts and other cyanobacteria lack several genes for the complete biosynthesis of certain amino acids, vitamins, oxidase in the cytochrome *bd* complex, and the KDO2-lipid A biosynthesis pathway (Figure 6). However, despite all symbionts functioning as N sources for their respective hosts and, in the case of EtSB, being also associated with a diatom, the pattern of gene loss and retention was not identical. The variation in functional genome content likely reflects differences in their evolutionary trajectories, host interactions, and genomic constraints, which has led to distinct adaptations even within their similar symbiotic niches. For example, a recent comparative analysis of membrane transporter content in the three *Richelia* spp. highlighted a similar host dependency of the endobiont and periplasmic symbionts for metabolite exchanges, while the epibiont possesses the same transporters as the latter and additionally other transporters (e.g., ammonium, nitrite/nitrate, phosphonate, and ferric-siderophore complexes) necessary for life in the oligotrophic ocean.[21]

The *Richelia* endobionts, including the periplasmic symbionts, showed several instances of degraded biosynthetic pathways (Figure S6), including a disrupted ubiquinone synthesis pathway and the absence of cytochrome *bd* ubiquinol oxidase (cytochrome *bd* complex). Similarly, the genomes of EtSB and UCYN-A/nitroplast also show disrupted ubiquinone synthesis but retain cytochrome *bd* (Figure 6). By contrast, both pathways remain intact in *N. azollae*, the epibiont *Richelia* (RrhiSC01), and one of the three MAGs closely related to RrhiSC01. Ubiquinone/ubiquinol are integral components of both the respiratory and photosynthetic electron transport chains and are located in the thylakoids of cyanobacteria. Thus, the loss of one of these components limits the symbiont's capacity to generate ATP through oxidative phosphorylation and suggests that ReuHH01 and RintRC01 receive ubiquinone/ubiquinol from their respective hosts. In the obligate intracellular *Rickettsia* spp. symbionts, which lack a complete ubiquinone synthesis pathway, the endosymbionts rely on importing compounds from the host to complete several biosynthetic pathways.[73] A comparable scenario could occur in the various endobionts (*Richelia* and EtSB) and

UCYN-A/nitroplast, given that one to two of the five *ubi* genes have been conserved for ubiquinone synthesis. The absence of cytochrome *bd* in the *Richelia* endobionts was surprising; however, ReuHH01 retains other quinol oxidases such as Cox3, which is necessary for respiratory activity of heterocysts and contributes to oxygen protection.[74]

Similar metabolic degradation is seen in the biosynthesis of vitamins, such as α-tocopherol and biotin. α-Tocopherol is an antioxidant particularly effective at scavenging intracellular singlet oxygen.[75] The *Richelia* epibiont, periplasmic symbionts, and *N. azollae* possess a complete biosynthetic pathway for α-tocopherol (Figure 6). The *Richelia* endobionts, together with EtSB and the UCYN-A/nitroplast, however, have a degraded pathway with only one gene remaining for α-tocopherol. We interpret that the incomplete α-tocopherol pathway has resulted from a decreased need for antioxidant defense, and as such favors a scenario where endobionts and UCYN-A/nitroplast rely on their respective hosts for protection against oxidative stress. For example, electron micrographs and confocal imaging of the *Richelia* endobionts show filaments in close proximity to the mitochondria of their hosts.[13,36] In fact, electron micrographs have shown that *Richelia* endobionts possess outer-inner membrane-like vesicles near their host mitochondria,[36] which suggests a possible mechanism for the host to function in oxidative stress protection.

Biotin is a crucial cofactor for various metabolic enzymes involved in carboxylation reactions, such as fatty acid synthesis, amino acid metabolism, and gluconeogenesis. The biotin biosynthesis pathway in the epibiont RrhiSC01, two of the three related MAGs to RrhiSC01, and *N. azollae* is complete, containing four genes that encode the necessary enzymes (Figure 6). However, biotin biosynthesis is degraded in the *Richelia* periplasmic symbionts, EtSB, and UCYN-A/nitroplast and completely absent in all *Richelia* endobiont genomes. The inability to synthesize biotin and an intracellular location require endobionts (and the periplasmic symbionts) to obtain it directly from their host. Most marine algae, including diatoms, can synthesize biotin.[76] Additionally, several biotin transporters (K03523, K16785, K16786, and K16787) were detected in all the *Richelia* symbionts, UCYN-A/nitroplast, and the three MAGs related to RrhiSC01 (Figure S7).

We identified that many amino acid synthesis pathways were eroded in the periplasmic symbionts and the endobiont *Richelia* genomes. Furthermore, we noted that often just one gene was missing, except for tryptophan, where only two out of seven genes necessary for the full pathway remain in the *Richelia* endobiont genomes (Figure 6). Tryptophan synthesis is completely absent in the UCYN-A/nitroplast, and one gene is missing in EtSB, while the full synthesis pathway remains in the other *Richelia* symbionts, environmental MAGs, and *N. azollae* (Figure 6). Tryptophan is essential for cyanobacteria in electron transfer and therefore central to capturing sunlight and initiating photosynthesis for efficient energy conversion.[77] However, tryptophan is the most complex and energy-consuming among all amino acids,[78] which could explain why this pathway in particular has been extensively degraded in the genomes of the endobionts and the UCYN-A/nitroplast. It is also likely redundant with their diatom hosts; all diatoms can synthesize tryptophan.[79] Furthermore, amino acids can be imported, and *Richelia* symbionts contain homologs of solute-binding proteins for N-I and N-II amino acid transporters, two of which (e.g., NatB and NatF) have recently been functionally tested and affinities characterized for several substrates.[67]

Finally, the notable loss of genes involved in 3-deoxi-D-manno-octulosonic acid-lipid A (KDO2-lipid A) biosynthesis in *Richelia* endobionts and retention in the other *Richelia* genomes and their close relatives, EtSB, *N. azollae*, and UCYN-A/nitroplast, suggests a unique loss of structural components in their outer membranes. Cyanobacteria are gram-negative and possess an inner membrane and an outer membrane, which contains lipid A as a key component of lipopolysaccharides.[80] The pathway that produces lipid A is uniquely degraded in the *Richelia* endobionts, with only one gene out of four encoding enzymes remaining, which has been noted before in ReuHH01.[21] The arrangement of the heterocyst envelope of *Richelia* ReuHH01 has also been recently reported as modified in transmission electron micrograph observations[36] (Figure 6). By contrast, the full lipid A biosynthesis pathway remains in all other genomes, including in the heterocyst-forming obligate endobiont *N. azollae*. The loss of lipid A biosynthesis in the *Richelia* endobionts remains unclear, as some endosymbiotic bacteria have lost the capacity for synthesis, while others have not.[81] In some host-associated bacteria, including several *Gammaproteobacteria*, the loss of the structural barrier in outer cell membranes increases permeability to hydrophobic molecules.[82] Notably, earlier work on *Anabaena* sp. 7120, a heterocyst-forming cyanobacterium, demonstrated that inactivation of several genes for lipid A biosynthesis in *Anabaena* constructs, including an IpxC homolog, resulted in increased permeability and specifically heightened uptake of sucrose and glutamate but not other amino acids.[83] Lipopolysaccharides also stimulate host recognition and immune responses in some well-characterized pathogenic host-microbe interactions of animals and multicellular eukaryotes.[84] It is unclear if such interactions occur in these planktonic symbioses.

In summary, *Richelia* endobionts exhibit several examples of advanced host dependency, with genome reduction and loss of many important biosynthetic pathways compared with the periplasmic symbionts and the epibiont. This places the loss and retention of the functional genome content of *Richelia* endobionts closer to that of the obligate planktonic symbiont EtSB and the UCYN-A/nitroplast, while also still showing distinguishable losses, e.g., loss/incomplete cytochrome bd complex, tryptophan, biotin, and lipid A. The functional genome content of the *Richelia* epibiont RrhiSC01 reflects its facultative nature, and the periplasmic *Richelia* appear in a transitory state, sharing genome loss/retention with both endobionts and the epibiont.

Protist-prokaryote symbioses as models for studying the transitionary steps leading to endosymbiosis

In recent years, there has been growing recognition and remarkable discoveries in several protist-prokaryotic symbiotic systems that show evidence of endosymbionts transitioning to organelles. In several of these systems, the evolutionary path of the symbiont, including the cycles of establishment, degeneration, replacement, and replication, varies tremendously.[3] Likewise, the functions, including the necessity of the symbiont for the host and vice versa, are not always following the norms of interactions, i.e., mutualism, expected for endosymbiosis. Thus,

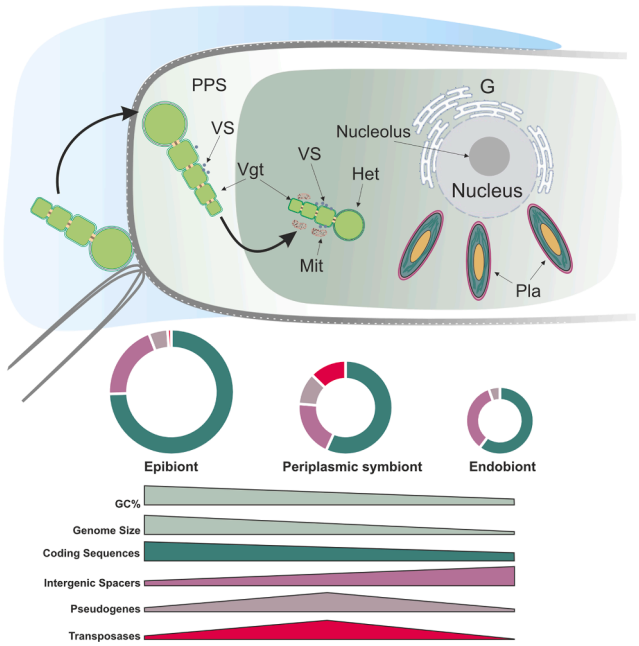


Figure 7. Conceptual model of the evolutionary trajectory of *Richelia* spp. symbionts across increasing levels of host cellular integration

The drawings are schematic and intended to illustrate genomic trends inferred from our data and a simple depiction of the cellular integration of *Richelia* with its respective host diatom based on earlier observations.[10,12,13,35,36] *Richelia* spp. filaments consist of varying numbers of vegetative cells (Vgt) and one terminal heterocyst (Het), and cells in a filament are connected by septal junctions that function in intercellular molecular exchange. From left to right: the facultative epibiont *R. rhizosoleniae* externally attach to the spines of their host diatoms with their Hets and have the largest genomes and highest GC%, the periplasmic *R. intracellularis* symbionts reside between the outer cell wall (frustule) and the cell membrane (plasmalemma) of their host diatoms and possess slightly smaller genomes and lower GC%, and the *R. euintracellularis* endobionts are fully integrated into the cytoplasm of their host diatoms and have highly reduced genomes and the lowest GC%. Inside of the diatom cell, there is a nucleus surrounded by Golgi bodies (G), several plastids (Pla), and numerous mitochondria (Mt), some of which are in tight associations with Vgt of the *Richelia* endobionts.[36] Small dots positioned on *Richelia* endobiont and periplasmic symbiont Vgt are cell envelope vesicles (VSs) present in the periplasm, which possibly participate in the transfer of metabolites from the cytoplasm of the cyanobacterium to the diatom.[35,36] The lower portion of the figure illustrates the genome trajectory of *Richelia* spp. symbionts as they transition from epibionts to endobionts. Epibionts are facultative symbionts, and their genomes resemble that of free-living bacteria with a high proportion of CDSs, a low proportion of IGSs, and few pseudogenes and transposases (as depicted in doughnut plots). Color coding in the donut plots corresponds to the genomic categories illustrated in the plots below. Periplasmic symbionts are transitioning from a facultative state to a more obligate one, and the transition is reflected in the several genome characters: a slight decrease in the percentage of CDSs and increases in IGSs, and a substantial increase of pseudogenes and transposases. Genomes of endobionts are characterized by further decreases in the CDSs and an increase in IGSs. In this stage, transposases are nearly purged from the endobiont genomes, and there are few pseudogenes accumulated in the IGSs.
See also Table S3.

establishing common "rules" in how diverse symbiotic models evolve is challenging.

One notable example is the thecate amoeba, *Paulinella chromatophora*, which recently (90–140 mya) has acquired its photosynthetic organelle, called a chromatophore, from an α cyanobacterium.[85] *P. chromatophore* represents the only known repeated instance of the primary endosymbiotic event, and hence studying the genome content of the chromatophore, especially in comparison to primary plastids, has identified several unique features.[86,87] Another interesting feature of the chromatophore genome is the loss of genes in the chromatophore is compensated by nuclear-encoded imported proteins that are of non-cyanobacterial origin. For example, the latter proteins are host-derived but acquired from HGT from diverse bacteria other than the α-cyanobacterium.[88] A second remarkable example is the calcareous haptophyte *Braarudosphaera bigelowii*, which has recently been recognized for its acquisition of a nitroplast, a N_2 fixing organelle, that originated some 100 mya from a cyanobacterial endosymbiont (UCYN-A).[89] Combining soft-X-ray tomography and proteomics, the integration and the coordination of the host and nitroplast cell cycles were revealed, along with that a significant fraction of the nitroplast proteins are encoded by and imported from the host genome.[89] The latter are the seminal characteristics of organelles. A third endosymbiotic system involves an omnivorous ciliate, *Euplotes*, and its beta-proteobacterium *Polynucleobacter* symbionts. A unique trait in the *Euplotes-Polynucleobacter* symbioses is the dynamic symbiont turnover, where obligate *Polynucleobacter* endosymbionts are not inherited from a single ancestral lineage but are repeatedly acquired from free-living environmental strains.[28] Despite being essential to the host, these symbionts rapidly undergo genome degradation, leading to their extinction and replacement by new strains, where each new strain enters a recurring cycle of genome erosion and host dependence.[90] With its short evolutionary timescales and repeated symbiont acquisition, *Euplotes* provides a powerful model for studying early stages of symbiosis, genome reduction, and symbiont integration.

Our study here contributes new information to this symbiosis-driven theoretical framework by tracing the genomic evolution of cyanobacterial symbionts across three distinct stages of integration: from facultative epibionts to periplasmic symbionts and finally to fully integrated endobionts (Figure 7). The *diatom-Richelia* system offers a rare opportunity to observe stepwise genome degradation in action, including the progressive loss of essential metabolic functions and a marked expansion of pseudogenes and transposases in the periplasmic symbionts, features that appear to drive reductive evolution (Figure 7). As such, the *diatom-Richelia* symbioses represent an ideal model for studying how increasing levels of cellular integration shape symbiont genome architecture.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed toward, and will be fulfilled by, the lead contact, Rachel A. Foster (rachel.foster@su.se).

Materials availability

No new materials were generated in this study.

Data and code availability

- Genomes used in this study are all publicly available, and their accession numbers are mentioned in the STAR Methods and Table S1.

- All the material and code to reproduce the results of this study are deposited and are publicly available via GitHub: https://github.com/VesnaGr/Richelia_Comparative_Genomics/.
- Additional information required to analyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The computations and data handling were enabled by resources in projects NAISS 2023/23-493 and NAISS 2023/22-993 provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX, funded by the Swedish Research Council through grant agreement no. 2022-06725. T.V.–S. and V.G. were supported by a Knut och Alice Wallenbergs Stiftelse grant (grant no. 2019.0321) to R.A.F. M.M. was supported by a grant from the Swedish Research Council for Sustainable Development, FORMAS (grant no. 2021-00546). We want to thank two anonymous reviewers for their helpful insights and criticisms and Prof. Enrique Flores from CSIC (Seville, Spain) for generously providing feedback on our manuscript.

AUTHOR CONTRIBUTIONS

V.G., M.M., and R.A.F. conceptualized the study. V.G., M.M., and D.L. performed bioinformatic analyses and developed figures. T.V.-S. analyzed insertion sequences. V.G., M.M., and R.A.F. interpreted the data and wrote the manuscript. All authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Genomes and metagenomes collections
- METHOD DETAILS
  - Phylogeny reconstruction
  - Genome statistics
  - GC content analysis
  - Transposase analysis
  - Pseudogenes identification
  - Pangenome analyses
  - Annotations
  - KO annotations outside of the pangenomic pipeline
  - Computing average nucleotide identity
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2025.08.003.

Received: January 23, 2025
Revised: June 3, 2025
Accepted: August 4, 2025
Published: August 29, 2025

REFERENCES

1. Decelle, J., Colin, S., and Foster, R.A. (2015). Photosymbiosis in marine planktonic protists. In Marine Protists, S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not, eds. (Springer), pp. 465–500. https://doi.org/10.1007/978-4-431-55130-0_19.
2. Foster, R.A., and Zehr, J.P. (2019). Diversity, Genomics, and Distribution of Phytoplankton-Cyanobacterium Single-Cell Symbiotic Associations. Annu. Rev. Microbiol. *73*, 435–456. https://doi.org/10.1146/annurev-micro-090817-062650.
3. Husnik, F., Tashyreva, D., Boscaro, V., George, E.E., Lukeš, J., and Keeling, P.J. (2021). Bacterial and archaeal symbioses with protists. Curr. Biol. *31*, R862–R877. https://doi.org/10.1016/j.cub.2021.05.049.
4. de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., et al. (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science *348*, 1261605. https://doi.org/10.1126/science.1261605.
5. Pierella Karlusich, J.J., Pelletier, E., Lombard, F., Carsique, M., Dvorak, E., Colin, S., Picheral, M., Cornejo-Castillo, F.M., Acinas, S.G., Pepperkok, R., et al. (2021). Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. Nat. Commun. *12*, 4160. https://doi.org/10.1038/s41467-021-24299-y.
6. López-García, P., Eme, L., and Moreira, D. (2017). Symbiosis in eukaryotic evolution. J. Theor. Biol. *434*, 20–33. https://doi.org/10.1016/j.jtbi.2017.02.031.
7. Roger, A.J., Muñoz-Gómez, S.A., and Kamikawa, R. (2017). The origin and diversification of mitochondria. Curr. Biol. *27*, R1177–R1192. https://doi.org/10.1016/j.cub.2017.09.015.
8. Coale, T.H., Loconte, V., Turk-Kubo, K.A., Vanslembrouck, B., Mak, W.K.E., Cheung, S., Ekman, A., Chen, J.-H., Hagino, K., Takano, Y., et al. (2024). Nitrogen-fixing organelle in a marine alga. Science *384*, 217–222. https://doi.org/10.1126/science.adk1075.
9. Marin, B., Nowack, E.C.M., and Melkonian, M. (2005). A plastid in the making: evidence for a second primary endosymbiosis. Protist *156*, 425–432. https://doi.org/10.1016/j.protis.2005.09.001.
10. Foster, R.A., Villareal, T.A., Lundin, D., Waterbury, J.B., Webb, E.A., and Zehr, J.P. (2022). Richelia. In Bergey's Manual of Systematics of Archaea and Bacteria (Wiley), pp. 1–17.
11. Foster, R.A., Kuypers, M.M.M., Vagner, T., Paerl, R.W., Musat, N., and Zehr, J.P. (2011). Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. ISME J. *5*, 1484–1493. https://doi.org/10.1038/ismej.2011.26.
12. Villareal, T.A. (1992). Marine Nitrogen-Fixing Diatom-Cyanobacteria Symbioses. In Marine Pelagic Cyanobacteria: *Trichodesmium* and Other Diazotrophs, E.J. Carpenter, D.G. Capone, and J.G. Rueter, eds. (Springer), pp. 163–175. https://doi.org/10.1007/978-94-015-7977-3_10.
13. Caputo, A., Nylander, J.A.A., and Foster, R.A. (2019). The genetic diversity and evolution of diatom-diazotroph associations highlights traits favoring symbiont integration. FEMS Microbiol. Lett. *366*, fny297. https://doi.org/10.1093/femsle/fny297.
14. Tuo, S.-H., Mulholland, M.R., Taniuchi, Y., Chen, H.-Y., Jane, W.-N., Lin, Y.-H., and Chen, Y.L. (2021). Trichome Lengths of the Heterocystous N_2-Fixing Cyanobacteria in the Tropical Marginal Seas of the Western North Pacific. Front. Mar. Sci. *8*, fmars.2021.678607. https://doi.org/10.3389/fmars.2021.678607.
15. Peng, L., Xie, C., Wang, M., Gu, J., Zhang, Y., Jiang, T., Cui, Y., and Wang, Z. (2023). Metabarcoding of microeukaryotes in surface sediments from the Pacific Arctic and adjacent sea areas: The role of diatoms in the biological pump. Glob. Planet Change *230*, 104262. https://doi.org/10.1016/j.gloplacha.2023.104262.
16. Strassert, J.F.H., Irisarri, I., Williams, T.A., and Burki, F. (2021). A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat. Commun. *12*, 1879. https://doi.org/10.1038/s41467-021-22044-z.
17. Dorrell, R.G. (2024). Functional impacts of horizontal genome evolution across eukaryotic algae. PhD thesis (Université Paris-Saclay).
18. Deschamps, P., and Moreira, D. (2012). Reevaluating the green contribution to diatom genomes. Genome Biol. Evol. *4*, 683–688. https://doi.org/10.1093/gbe/evs053.
19. Hilton, J.A., Foster, R.A., James Tripp, H.J., Carter, B.J., Zehr, J.P., and Villareal, T.A. (2013). Genomic deletions disrupt nitrogen metabolism

pathways of a cyanobacterial diatom symbiont. Nat. Commun. *4*, 1767. https://doi.org/10.1038/ncomms2748.

20. Hilton, J.A. (2014). Ecology and Evolution of Diatom-Associated Cyanobacteria Through Genetic Analyses. PhD thesis (University of California).
21. Nieves-Morión, M., Flores, E., and Foster, R.A. (2020). Predicting substrate exchange in marine diatom-heterocystous cyanobacteria symbioses. Environ. Microbiol. *22*, 2027–2052. https://doi.org/10.1111/1462-2920.15013.
22. Foster, R.A., Goebel, N.L., and Zehr, J.P. (2010). Isolation of *Calothrix rhizosoleniae* (cyanobacteria) strain SC01 from *Chaetoceros* (Bacillariophyta) spp. diatoms of the subtropical north pacific ocean. J. Phycol. *46*, 1028–1037. https://doi.org/10.1111/j.1529-8817.2010.00885.x.
23. McCutcheon, J.P., and Moran, N.A. (2011). Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. *10*, 13–26. https://doi.org/10.1038/nrmicro2670.
24. McCutcheon, J.P., Boyd, B.M., and Dale, C. (2019). The life of an insect endosymbiont from the cradle to the grave. Curr. Biol. *29*, R485–R495. https://doi.org/10.1016/j.cub.2019.03.032.
25. Moran, N.A., McCutcheon, J.P., and Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. *42*, 165–190. https://doi.org/10.1146/annurev.genet.41.110306.130119.
26. Moran, N.A., and Bennett, G.M. (2014). The tiniest tiny genomes. Annu. Rev. Microbiol. *68*, 195–215. https://doi.org/10.1146/annurev-micro-091213-112901.
27. Clayton, A.L., Oakeson, K.F., Gutin, M., Pontes, A., Dunn, D.M., von Niederhausern, A.C., Weiss, R.B., Fisher, M., and Dale, C. (2012). A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect-bacterial symbioses. PLoS Genet. *8*, e1002990. https://doi.org/10.1371/journal.pgen.1002990.
28. Boscaro, V., Felletti, M., Vannini, C., Ackerman, M.S., Chain, P.S.G., Malfatti, S., Vergez, L.M., Shin, M., Doak, T.G., Lynch, M., et al. (2013). *Polynucleobacter necessarius*, a model for genome reduction in both free-living and symbiotic bacteria. Proc. Natl. Acad. Sci. USA *110*, 18590–18595. https://doi.org/10.1073/pnas.1316687110.
29. Manzano-Marín, A., and Latorre, A. (2014). Settling down: the genome of *Serratia symbiotica* from the aphid *Cinara tujafilina* zooms in on the process of accommodation to a cooperative intracellular life. Genome Biol. Evol. *6*, 1683–1698. https://doi.org/10.1093/gbe/evu133.
30. Martínez-Cano, D.J., Reyes-Prieto, M., Martínez-Romero, E., Partida-Martínez, L.P., Latorre, A., Moya, A., and Delaye, L. (2014). Evolution of small prokaryotic genomes. Front. Microbiol. *5*, 742. https://doi.org/10.3389/fmicb.2014.00742.
31. Lo, W.-S., Huang, Y.-Y., and Kuo, C.-H. (2016). Winding paths to simplicity: genome evolution in facultative insect symbionts. FEMS Microbiol. Rev. *40*, 855–874. https://doi.org/10.1093/femsre/fuw028.
32. Seidl, M.F., and Thomma, B.P.H.J. (2017). Transposable elements direct the coevolution between plants and microbes. Trends Genet. *33*, 842–851. https://doi.org/10.1016/j.tig.2017.07.003.
33. Moran, N.A., and Wernegreen, J.J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol. Evol. *15*, 321–326. https://doi.org/10.1016/s0169-5347(00)01902-9.
34. Bennett, G.M., and Moran, N.A. (2013). Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. Genome Biol. Evol. *5*, 1675–1688. https://doi.org/10.1093/gbe/evt118.
35. Jahson, S., Rai, A.N., and Bergman, B. (1995). Intracellular cyanobiont *Richelia intracellularis*: ultrastructure and immuno-localisation of phycoerythrin, nitrogenase, Rubisco and glutamine synthetase. Mar. Biol. *124*, 1–8. https://doi.org/10.1007/BF00349140.
36. Flores, E., Romanovicz, D.K., Nieves-Morión, M., Foster, R.A., and Villareal, T.A. (2022). Adaptation to an intracellular lifestyle by a nitrogen-fixing, heterocyst-forming cyanobacterial endosymbiont of a diatom. Front. Microbiol. *13*, 799362. https://doi.org/10.3389/fmicb.2022.799362.
37. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794. https://doi.org/10.1093/nar/gkab776.
38. Janson, S., Wouters, J., Bergman, B., and Carpenter, E.J. (1999). Host specificity in the *Richelia*-diatom symbiosis revealed by *hetR* gene sequence analysis. Environ. Microbiol. *1*, 431–438. https://doi.org/10.1046/j.1462-2920.1999.00053.x.
39. Foster, R.A., and Zehr, J.P. (2006). Characterization of diatom-cyanobacteria symbioses on the basis of *nifH*, *hetR* and 16S rRNA sequences. Environ. Microbiol. *8*, 1913–1925. https://doi.org/10.1111/j.1462-2920.2006.01068.x.
40. Chase, A.B., Bogdanov, A., Demko, A.M., and Jensen, P.R. (2023). Biogeographic patterns of biosynthetic potential and specialized metabolites in marine sediments. ISME J. *17*, 976–983. https://doi.org/10.1038/s41396-023-01410-3.
41. Delmont, T.O. (2021). Discovery of nondiazotrophic *Trichodesmium* species abundant and widespread in the open ocean. Proc. Natl. Acad. Sci. USA *118*, e2112355118. https://doi.org/10.1073/pnas.2112355118.
42. Touchon, M., and Rocha, E.P.C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. Mol. Biol. Evol. *24*, 969–981. https://doi.org/10.1093/molbev/msm014.
43. Vigil-Stenman, T., Larsson, J., Nylander, J.A.A., and Bergman, B. (2015). Local hopping mobile DNA implicated in pseudogene formation and reductive evolution in an obligate cyanobacteria-plant symbiosis. BMC Genomics *16*, 193. https://doi.org/10.1186/s12864-015-1386-7.
44. Aziz, R.K., Breitbart, M., and Edwards, R.A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res. *38*, 4207–4217. https://doi.org/10.1093/nar/gkq140.
45. Newton, I.L.G., and Bordenstein, S.R. (2011). Correlations between bacterial ecology and mobile DNA. Curr. Microbiol. *62*, 198–208. https://doi.org/10.1007/s00284-010-9693-3.
46. Moran, N.A., and Plague, G.R. (2004). Genomic changes following host restriction in bacteria. Curr. Opin. Genet. Dev. *14*, 627–633. https://doi.org/10.1016/j.gde.2004.09.003.
47. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. *34*, D32–D36. https://doi.org/10.1093/nar/gkj014.
48. Moya, A., Peretó, J., Gil, R., and Latorre, A. (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat. Rev. Genet. *9*, 218–229. https://doi.org/10.1038/nrg2319.
49. Bordenstein, S.R., and Reznikoff, W.S. (2005). Mobile DNA in obligate intracellular bacteria. Nat. Rev. Microbiol. *3*, 688–699. https://doi.org/10.1038/nrmicro1233.
50. Leclercq, S., Giraud, I., and Cordaux, R. (2011). Remarkable abundance and evolution of mobile group II introns in *Wolbachia* bacterial endosymbionts. Mol. Biol. Evol. *28*, 685–697. https://doi.org/10.1093/molbev/msq238.
51. Cerveau, N., Leclercq, S., Leroy, E., Bouchon, D., and Cordaux, R. (2011). Short- and long-term evolutionary dynamics of bacterial insertion sequences: insights from *Wolbachia* endosymbionts. Genome Biol. Evol. *3*, 1175–1186. https://doi.org/10.1093/gbe/evr096.
52. Thorpe, H.A., Bayliss, S.C., Hurst, L.D., and Feil, E.J. (2017). Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. Genetics *206*, 363–376. https://doi.org/10.1534/genetics.116.195784.
53. Dial, D.T., Weglarz, K.M., Aremu, A.O., Havill, N.P., Pearson, T.A., Burke, G.R., and von Dohlen, C.D. (2022). Transitional genomes and nutritional role reversals identified for dual symbionts of adelgids (Aphidoidea: Adelgidae). ISME J. *16*, 642–654. https://doi.org/10.1038/s41396-021-01102-w.

54. Kuo, C.-H., and Ochman, H. (2010). The extinction dynamics of bacterial pseudogenes. PLoS Genet. *6*, e1001050. https://doi.org/10.1371/journal.pgen.1001050.

55. Liu, Y., Harrison, P.M., Kunin, V., and Gerstein, M. (2004). Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol. *5*, R64. https://doi.org/10.1186/gb-2004-5-9-r64.

56. Batut, B., Knibbe, C., Marais, G., and Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. Nat. Rev. Microbiol. *12*, 841–850. https://doi.org/10.1038/nrmicro3331.

57. Goodhead, I., Blow, F., Brownridge, P., Hughes, M., Kenny, J., Krishna, R., McLean, L., Pongchaikul, P., Beynon, R., and Darby, A.C. (2020). Large scale and significant expression from pseudogenes in *Sodalis glossinidius* – a facultative bacterial endosymbiont. Microb. Genom. *6*, e000285. https://doi.org/10.1099/mgen.0.000285.

58. Moran, N.A. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. Curr. Opin. Microbiol. *6*, 512–518. https://doi.org/10.1016/j.mib.2003.08.001.

59. Degnan, P.H., Ochman, H., and Moran, N.A. (2011). Sequence conservation and functional constraint on intergenic spacers in reduced genomes of the obligate symbiont *Buchnera*. PLoS Genet. *7*, e1002252. https://doi.org/10.1371/journal.pgen.1002252.

60. Van Leuven, J.T., and McCutcheon, J.P. (2012). An AT mutational bias in the tiny GC-rich endosymbiont genome of *Hodgkinia*. Genome Biol. Evol. *4*, 24–27. https://doi.org/10.1093/gbe/evr125.

61. Van Leuven, J.T., Meister, R.C., Simon, C., and McCutcheon, J.P. (2014). Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. Cell *158*, 1270–1280. https://doi.org/10.1016/j.cell.2014.07.047.

62. Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. *11*, 472–477. https://doi.org/10.1016/j.mib.2008.09.006.

63. Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. *15*, 589–594. https://doi.org/10.1016/j.gde.2005.09.006.

64. Dewar, A.E., Hao, C., Belcher, L.J., Ghoul, M., and West, S.A. (2024). Bacterial lifestyle shapes pangenomes. Proc. Natl. Acad. Sci. USA *121*, e2320170121. https://doi.org/10.1073/pnas.2320170121.

65. McNeely, K., Xu, Y., Ananyev, G., Bennette, N., Bryant, D.A., and Dismukes, G.C. (2011). *Synechococcus* sp. strain PCC 7002 *nifJ* mutant lacking pyruvate:ferredoxin oxidoreductase. Appl. Environ. Microbiol. *77*, 2435–2444. https://doi.org/10.1128/AEM.02792-10.

66. Foster, R.A., Tienken, D., Littmann, S., Whitehouse, M.J., Kuypers, M.M. M., and White, A.E. (2022). The rate and fate of N_2 and C fixation by marine diatom-diazotroph symbioses. ISME J. *16*, 477–487. https://doi.org/10.1038/s41396-021-01086-7.

67. Nieves-Morión, M., Camargo, S., Bardi, S., Ruiz, M.T., Flores, E., and Foster, R.A. (2023). Heterologous expression of genes from a cyanobacterial endosymbiont highlights substrate exchanges with its diatom host. PNAS Nexus *2*, gad194. https://doi.org/10.1093/pnasnexus/pgad194.

68. Torrado, A., Ramírez-Moncayo, C., Navarro, J.A., Mariscal, V., and Molina-Heredia, F.P. (2019). Cytochrome c6 is the main respiratory and photosynthetic soluble electron donor in heterocysts of the cyanobacterium *Anabaena* sp. PCC 7120. Biochim. Biophys. Acta Bioenerg. *1860*, 60–68. https://doi.org/10.1016/j.bbabio.2018.11.009.

69. Castell, C., Hervás, M., López-Maury, L., Roncel, M., and Navarro, J.A. (2022). Adaptation of cyanobacterial photosynthesis to metal constraints. In Expanding Horizon of Cyanobacterial Biology (Elsevier), pp. 109–128. https://doi.org/10.1016/B978-0-323-91202-0.00006-3.

70. Stadnichuk, I.N., and Kusnetsov, V.V. (2023). Phycobilisomes and phycobiliproteins in the pigment apparatus of oxygenic photosynthetics: from cyanobacteria to tertiary endosymbiosis. Int. J. Mol. Sci. *24*, 2290. https://doi.org/10.3390/ijms24032290.

71. Ochman, H., and Moran, N.A. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science *292*, 1096–1099. https://doi.org/10.1126/science.1058543.

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

73. Driscoll, T.P., Verhoeve, V.I., Guillotte, M.L., Lehman, S.S., Rennoll, S.A., Beier-Sexton, M., Rahman, M.S., Azad, A.F., and Gillespie, J.J. (2017). Wholly *Rickettsia*! Reconstructed metabolic profile of the quintessential bacterial parasite of eukaryotic cells. mBio *8*, e00859-17. https://doi.org/10.1128/mBio.00859-17.

74. Valladares, A., Maldener, I., Muro-Pastor, A.M., Flores, E., and Herrero, A. (2007). Heterocyst development and diazotrophic metabolism in terminal respiratory oxidase mutants of the cyanobacterium *Anabaena* sp. Strain PCC 7120. J. Bacteriol. *189*, 4425–4430. https://doi.org/10.1128/JB.00220-07.

75. Neely, W.C., Martin, J.M., and Barker, S.A. (1988). Products and relative reaction rates of the oxidation of tocopherols with singlet molecular oxygen. Photochem. Photobiol. *48*, 423–428. https://doi.org/10.1111/j.1751-1097.1988.tb02840.x.

76. Cohen, N.R., Ellis, A., Burns, W.G., Lampe, R.H., Schuback, N., Johnson, Z., Sañudo-Wilhelmy, S., and Marchetti, A. (2017). Iron and vitamin interactions in marine diatom isolates and natural assemblages of the Northeast Pacific Ocean. Limnol. Oceanogr. *62*, 2076–2096. https://doi.org/10.1002/lno.10552.

77. Alachkar, A. (2022). Aromatic patterns: Tryptophan aromaticity as a catalyst for the emergence of life and rise of consciousness. Phys. Life Rev. *42*, 93–114. https://doi.org/10.1016/j.plrev.2022.07.002.

78. Barik, S. (2020). The uniqueness of tryptophan in biology: Properties, metabolism, interactions and localization in proteins. Int. J. Mol. Sci. *21*, 8776. https://doi.org/10.3390/ijms21228776.

79. Jiroutová, K., Horák, A., Bowler, C., and Oborník, M. (2007). Tryptophan biosynthesis in stramenopiles: eukaryotic winners in the diatom complex chloroplast. J. Mol. Evol. *65*, 496–511. https://doi.org/10.1007/s00239-007-9022-z.

80. Burnat, M., Schleiff, E., and Flores, E. (2014). Cell envelope components influencing filament length in the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. J. Bacteriol. *196*, 4026–4035. https://doi.org/10.1128/JB.02128-14.

81. Zientz, E., Dandekar, T., and Gross, R. (2004). Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol. Mol. Biol. Rev. *68*, 745–770. https://doi.org/10.1128/MMBR.68.4.745-770.2004.

82. Zhang, G., Meredith, T.C., and Kahne, D. (2013). On the essentiality of lipopolysaccharide to gram-negative bacteria. Curr. Opin. Microbiol. *16*, 779–785. https://doi.org/10.1016/j.mib.2013.09.007.

83. Nicolaisen, K., Mariscal, V., Bredemeier, R., Pernil, R., Moslavac, S., López-Igual, R., Maldener, I., Herrero, A., Schleiff, E., and Flores, E. (2009). The outer membrane of a heterocyst-forming cyanobacterium is a permeability barrier for uptake of metabolites that are exchanged between cells. Mol. Microbiol. *74*, 58–70. https://doi.org/10.1111/j.1365-2958.2009.06850.x.

84. Perkins, S.L., Budinoff, R.B., and Siddall, M.E. (2005). New gammaproteobacteria associated with blood-feeding leeches and a broad phylogenetic analysis of leech endosymbionts. Appl. Environ. Microbiol. *71*, 5219–5224. https://doi.org/10.1128/AEM.71.9.5219-5224.2005.

85. Nowack, E.C.M., Melkonian, M., and Glöckner, G. (2008). Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. Curr. Biol. *18*, 410–418. https://doi.org/10.1016/j.cub.2008.02.051.

86. Nowack, E.C.M., and Grossman, A.R. (2012). Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. Proc. Natl. Acad. Sci. USA *109*, 5340–5345. https://doi.org/10.1073/pnas.1118800109.

87. Singer, A., Poschmann, G., Mühlich, C., Valadez-Cano, C., Hänsch, S., Hüren, V., Rensing, S.A., Stühler, K., and Nowack, E.C.M. (2017). Massive protein import into the early-evolutionary-stage photosynthetic organelle of the amoeba *Paulinella chromatophora*. Curr. Biol. *27*, 2763–2773.e5. https://doi.org/10.1016/j.cub.2017.08.010.

88. Morales, J., Kokkori, S., Weidauer, D., Chapman, J., Goltsman, E., Rokhsar, D., Grossman, A.R., and Nowack, E.C.M. (2016). Development of a toolbox to dissect host-endosymbiont interactions and protein trafficking in the trypanosomatid *Angomonas deanei*. BMC Evol. Biol. *16*, 247. https://doi.org/10.1186/s12862-016-0820-z.

89. Cornejo-Castillo, F.M., Cabello, A.M., Salazar, G., Sánchez-Baracaldo, P., Lima-Mendez, G., Hingamp, P., Alberti, A., Sunagawa, S., Bork, P., de Vargas, C., et al. (2016). Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. Nat. Commun. *7*, 11071. https://doi.org/10.1038/ncomms11071.

90. Boscaro, V., Kolisko, M., Felletti, M., Vannini, C., Lynn, D.H., and Keeling, P.J. (2017). Parallel genome reduction in symbionts descended from closely related free-living bacteria. Nat. Ecol. Evol. *1*, 1160–1167. https://doi.org/10.1038/s41559-017-0237-0.

91. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004. https://doi.org/10.1038/nbt.4229.

92. Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

93. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

94. Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat. Methods *20*, 1203–1212. https://doi.org/10.1038/s41592-023-01940-w.

95. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

96. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

97. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., 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. https://doi.org/10.1093/bioinformatics/btp163.

98. R Core Team (2025). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

99. Posit Team (2025). RStudio: Integrated Development Environment for R (Posit Software).

100. Kurtz, S. (2017). The Vmatch large scale sequence analysis software. http://www.vmatch.de/.

101. Syberg-Olsen, M.J., Garber, A.I., Keeling, P.J., McCutcheon, J.P., and Husnik, F. (2022). Pseudofinder: Detection of Pseudogenes in Prokaryotic Genomes. Mol. Biol. Evol. *39*, msac153. https://doi.org/10.1093/molbev/msac153.

102. Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*, e1319. https://doi.org/10.7717/peerj.1319.

103. Delmont, T.O., and Eren, A.M. (2018). Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ *6*, e4320. https://doi.org/10.7717/peerj.4320.

104. Buck, M., Mehrshad, M., and Bertilsson, S. (2022). mOTUpan: a robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. NAR Genom. Bioinform. *4*, lqac060. https://doi.org/10.1093/nargab/lqac060.

105. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

106. Eddy, S.R. (1996). Hidden Markov models. Curr. Opin. Struct. Biol. *6*, 361–365. https://doi.org/10.1016/s0959-440x(96)80056-x.

107. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.

108. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

109. Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J. Mol. Biol. *428*, 726–731. https://doi.org/10.1016/j.jmb.2015.11.006.

110. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

111. Galperin, M.Y., Wolf, Y.I., Makarova, K.S., Vera Alvarez, R., Landsman, D., and Koonin, E.V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. *49*, D274–D281. https://doi.org/10.1093/nar/gkaa1018.

112. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

113. Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics *36*, 2251–2252. https://doi.org/10.1093/bioinformatics/btz859.

114. Yang, M., Derbyshire, M.K., Yamashita, R.A., and Marchler-Bauer, A. (2020). NCBI's Conserved Domain Database and Tools for Protein Domain Analysis. Curr. Protoc. Bioinformatics *69*, e90. https://doi.org/10.1002/cpbi.90.

115. Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G., and Toth, I.K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal. Methods *8*, 12–24. https://doi.org/10.1039/C5AY02550H.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Prokka annotated genomes	This study	Github repository: https://github.com/VesnaGr/Richelia-Comparative-Genomics/tree/main/Prokka
Pseudofinder annotations	This study	https://github.com/VesnaGr/Richelia-Comparative-Genomics/tree/main/Pseudofinder
Custom Python scripts for calculating genome statistics related to coding and non-coding sequences, along with R code used to generate Figures 2 and S2 and Table S1	This study	https://github.com/VesnaGr/Richelia-Comparative-Genomics/tree/main/Genome%20stats
Custom Python scripts for calculating pseudogene statistics presented in Figures 2 and S5 and Table S1	This study	https://github.com/VesnaGr/Richelia-Comparative-Genomics/tree/main/Pseudogenes
Custom Python scripts to analyze GC content and related R code to generate Figures 3 and S5	This study	https://github.com/VesnaGr/Richelia-Comparative-Genomics/tree/main/GC_content
Anvi'o gene clusters summary, related data, and R code used to generate Figures 4 and S7–S11	This study	https://github.com/VesnaGr/Richelia-Comparative-Genomics/tree/main/Richelia_pangenomics
Blast Koala annotations and related R code necessary to reproduce Figures 5 and S12	This study	https://github.com/VesnaGr/Richelia-Comparative-Genomics/tree/main/KO_analysis
Tree files produced by GTDB-tk and Iqtree used to generate Figure 1	This study	https://github.com/VesnaGr/Richelia-Comparative-Genomics/tree/main/GTDB-Tk_Tree
Software and algorithms		
GTDB-Tk v.2.4.0	Parks et al.[37]	https://github.com/Ecogenomics/GTDBTk
HMMER v.3.1b2	Eddy[106]	https://github.com/EddyRivasLab/hmmer
FastTree v2.1.10	Price et al.[93]	https://github.com/morgannprice/fasttree
CheckM2 v.1.0.1	Chklovski et al.[94]	https://github.com/chklovski/CheckM2
Prokka v.1.14.6	Seeman[95]	https://github.com/tseemann/prokka
Prodigal v2.6.3	Hyatt et al.[96]	https://github.com/hyattpd/Prodigal
BioPython v1.83.	Cock et al.[97]	https://github.com/biopython/biopython
R v4.4.0	R Core Team[98]	https://cran.r-project.org/
RStudio	Posit Team[99]	https://posit.co/download/rstudio-desktop/
Transposeek2	This study	https://github.com/Omnistudent/transposeek2
ISfinder database	Siguier et al.[47]	https://isfinder.biotoul.fr/about.php
Vmatch	Kurtz[100]	http://www.vmatch.de/
Pseudofinder v1.1.0	Syberg-Olsen et al.[101]	https://github.com/filip-husnik/pseudofinder
Anvi'o v8	Eren et al.[102]	https://anvio.org/
mOTUpan v0.3.2	Buck et al.[104]	https://github.com/moritzbuck/mOTUlizer
BLASTKoala v3.0	Kanehisa et al.[109]	https://www.kegg.jp/blastkoala/
EGG-NOG mapper v2.1.9	Cantalapiedra et al.[108]	https://github.com/eggnogdb/eggnog-mapper
DIAMOND v2.1.8	Buchfink et al.[110]	https://github.com/bbuchfink/diamond
PyANI v0.2.13.1	Pritchard et al.[115]	https://github.com/widdowquinn/pyani

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Genomes and metagenomes collections

Ten *Richelia* and four related cyanobacterial genome assemblies were downloaded from NCBI and the Genoscope TARA Oceans database (https://www.genoscope.cns.fr/tara/) used in this study. Ten of the fourteen genomes are MAGs reconstructed from environmental metagenomes. No additional binning, filtering, or manual curation was performed. Accession numbers and associated metadata are listed in Table S1.

METHOD DETAILS

Phylogeny reconstruction

We used the "de novo workflow" of GTDB-Tk (v.2.4.0; GTDB v.R09-RS220)[91] to estimate a phylogeny for the genomes discussed here together with all bacterial species. Briefly, GTDB-Tk identifies 120 marker proteins, aligns them using tools from HMMER (v.3.1b2)[92] and, when using the "de novo workflow", estimates a maximum likelihood phylogeny using FastTree (v.2.1.10)[93] from the concatenated alignment with WAG and SH support values. From the resulting phylogeny, the *Nostocaceae* clade was extracted to create Figure 1.

Genome statistics

Completeness and contamination of the genome assemblies and MAGs were calculated using CheckM2 (v. 1.0.1).[94] Genomes were initially annotated with Prokka (v.1.14.6)[95] including gene recognition and translation initiation site identification with Prodigal (v2.6.3).[96] To analyze the genomic features and calculate various sizes such as the size of CDS, tRNA, and IGS, we employed a custom python script and used Prokka annotated FASTA and GFF files (available on GitHub repository). CDS and tRNA sizes were extracted from the GFF file by identifying feature types and calculating sizes based on start and end positions. IGS, defined as the segments of the genome between annotated CDS features, were calculated as the difference between the start position of the current CDS feature and the end position of the previous feature minus one.

GC content analysis

To analyze GC content across CDS and IGS regions, we used a custom python script (available at GitHub repository) that extracts CDS and IGS sequences from each genome. Prokka annotated FASTA files from each genome were first matched with their corresponding GFF annotation files. From these, all gene coordinates were used to extract CDSs, while IGS were defined as nucleotide regions located between annotated gene features on the same contig. The GFF features were sorted by start position to ensure correct parsing, and overlapping genes were accounted for by excluding ambiguous IGS boundaries. For each genome, CDS and IGS sequences were exported as separate FASTA files. IGS sequences were then classified into two categories: short IGS (<300 bp) and long IGS (≥300 bp). GC content for each sequence (CDS, short IGS, and long IGS) was calculated using BioPython's GC() function,[97] which computes the percentage of guanine and cytosine bases relative to total sequence length. Final GC distributions were visualized in R[98,99] using density ridge plots (Figures 4A and 4B).

Transposase analysis

To identify transposase, genomes were compared to a database of IS transposase genes with a two-step blastx process, using Transposeek2 [2024, https://github.com/Omnistudent/transposeek2]. The Transposeek2 python script compiles genomic footprints of regions with transposase blast hits, then divides these footprints into identified insertion sequences using the highest score. The amino acid database of transposases was originally downloaded from ISfinder [2016, https://isfinder.biotoul.fr/about.php].[47] To identify inactive transposase, the length of the identified sequence was expressed as a percentage of the full-length subject length. In addition to searching for remains of transposase proteins, the proliferation of mobile genetic elements was also investigated using Vmatch,[100] which detects repeated nucleotide sequences of any kind. Vmatch was applied to the genomes using parameters -showdesc 0 -d -l 200 -best 5000 -sort ld (show sequence description of match, find direct matches, minimum repeat length 200, show 5000 results, sort in descending order of length).

Pseudogenes identification

Pseudogenes were identified and annotated using Pseudofinder[101] with default settings and by using NCBI nr database. Detailed descriptions of files produced by pseudofiner (v1.1.0) can be found https://github.com/filip-husnik/pseudofinder/wiki/5.-Commands, and all corresponding files from this project are available in the GitHub repository. Using the GFF files produced by Pseudofinder, we calculated the numbers and lengths of pseudogenes in each genome. Furthermore, Pseudofinder divides pseudogenes into four categories based on the provided justification for pseudogenization: run-on, truncated, predicted fragmentation, and blast hits in intergenic spacers. Pseudogenes annotated as transposases in the BlastP and BlastX files produced by Pseudofinder were excluded from GFF files based on their identifiers and therefore from the overall analysis. Detailed calculations of the above are represented in the python script available on GitHub.

Pangenome analyses

Pangenome of the four draft *Richelia* genome assemblies and the 10 recovered MAGs was computed using anvi'o[102] standard pangenomics workflow (anvio V8) with some additions.[103] Briefly, we ran mOTUpan (v0.3.2) inside anvi'o with *anvi-script-compute-bayesian-pan-core* to computationally estimate whether gene clusters belong to the core or flexible genomes.[104] This classification was then used to visualize the core genome within the pangenome (Figure 5). Briefly, the pangenome workflow consists of 1) generating contigs database out of genome FASTA files with program *anvi-gen-contigs-database* 2) generating genome storage database with program *anvi-gen-genomes-storage* 3) generate anvi'o pangenome database with program *anvi-pan-genome* 4) visualize pangenome using program *anvi-display-pan*. The detailed description of what program *anvio-pan-genome* does can be found in Delmont and Eren.[103] Briefly, it begins by calculating amino acid sequence similarities using blastP,[105] filters out weak hits, and then employs

the MCL algorithm[106] to identify gene clusters. The gene clusters, as described previously,[103] represent sequences of one or more predicted open reading frames grouped together based on their homology at the translated DNA sequence level. Subsequently, it computes the distribution of these gene clusters across genomes, conducts hierarchical clustering analyses for both gene clusters and genomes, and finally generates an anvi'o pan database which was used to visualize *Richelia* pangenome.

Annotations

Genomes were initially annotated with Prokka and further annotated using EGG-NOG mapper (v2.1.9)[107,108] with the reference database v5 and BLASTkoal (v 3.0).[109] Prokka and emapper annotations were imported into pangenome as detailed in GitHub repository. Additionally, we performed functional annotations inside of anvio'o with COG annotations using the *anvi-run-ncbi-cogs* program with the –sensitive flag (runs sensitive version of DIAMOND (v2.1.8)[110] and the 2020 COG20 database.[111] KEGG/KOfam (v4),[112,113] annotations were also added to each genome database file, as well as hmm-hits (v3.3.1).[106] Summary file produced by anvio is available at repository. All our functional and metabolic analysis were based on annotations produced by emapper and KEGG database. All annotations discussed in this manuscript were further manually confirmed by inspecting their conserved domains using the NCBI conserved domain batch search.[114]

KO annotations outside of the pangenomic pipeline

In order to compare the genomes of *Richelia* with other symbiotic cyanobacteria associated with diatoms and to avoid gene duplication, we downloaded amino acid FASTA files for *EtSB* (accession number GCA_000829235.1), *nitroplast* (*UCYN-A*, accession number GCA_020885515.1), and *N. azollae* (GCA_000196515.1). We annotated these genomes, along with all *Richelia* genomes, using the online version of BlastKOALA (v 3.0).[109] Our functional analysis was based on the KEGG annotations generated from this process. Since many of the modules in the KEGG database are based on *E. coli* or other model bacteria, we took extra care when analyzing our pathways. Specifically, when we encountered missing genes, we did not immediately conclude that a pathway was incomplete. Instead, we only classified a pathway as incomplete after comparing them to genomes of well-studied cyanobacteria, such as *Synechocystis* sp. PCC 6803 (https://www.genome.jp/kegg-bin/show_organism?menu_type=genome_info&org=syn), *Nostoc* sp. PCC 7120 (*Anabaena* sp. PCC 7120, https://www.genome.jp/kegg-bin/show_organism?org=ana) and *Rivularia* sp. PCC 7116 (https://www.genome.jp/kegg-bin/show_organism?org=riv). While accounting for genome incompleteness, we considered a gene or function as truly absent in a group only if it is missing from all genomes within that group, and as present only if it appears in 90 % of genomes within that group. We also acknowledge the possibility that some functions may not be annotated, or that some genes for alternative pathways may have not yet been discovered or characterized.

Computing average nucleotide identity

Similarity of genomes in the pangenome was calculated in anvi'o with anvi-compute-genome-similarity using PyANI (v0.2.13.1).[115] This program calculated average nucleotide identity (ANI) which was used in for visualization together with the pangenome (Figure 5).

QUANTIFICATION AND STATISTICAL ANALYSIS

All quantifications and statistical analyses conducted are detailed in the method details, figure legends or as referred in the text, available at the Git hub repository. The data and statistical analysis were conducted with BioPython v1.83.[97] or with R v4.4.0[98] in Rstudio.[99]