Highly variable fidelity drives symbiont community composition in an obligate symbiosis

Anna Mankowski[*1], Manuel Kleiner[2], Christer Erséus[3], Nikolaus Leisch[1], Yui Sato[1], Jean-Marie Volland[4], Bruno Hüttel[5], Cecilia Wentrup[6], Tanja Woyke[4], Juliane Wippler[1], Nicole Dubilier[*1], Harald Gruber-Vodicka[*1]

[1]Max Planck Institute for Marine Microbiology, Bremen, Germany

[2]North Carolina State University, Department of Plant and Microbial Biology, Raleigh, North Carolina, USA

[3]University of Gothenburg, Department of Biological and Environmental Sciences, Gothenborg, Sweden

[4]Deparment of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA

[5]Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding Research, Cologne, Germany

[6]University of Vienna, Department of Microbiology and Ecosystem Science, Vienna, Austria

[*]Corresponding authors:

Anna Mankowski: amankows@mpi-bremen.de, phone +49 (421) 2028 9050

Nicole Dubilier: ndubilie@mpi-bremen.de, phone +49 (421) 2028 9320

Harald R. Gruber-Vodicka: hgruber@mpi-bremen.de, phone +49 (421) 2028 7600

Keywords: chemosynthetic symbiosis, symbiont diversity, phylosymbiosis, metagenomics, phylogenetic distance

Abstract

Many animals are obligately associated with microbial symbionts that provide essential services such as nutrition or protection against predators. It is assumed that in such obligate associations fidelity between the host and its symbionts must be high to ensure the evolutionary success of the symbiosis. We show here that this is not the case in marine oligochaete worms, despite the fact that they are so dependent on their bacterial symbionts for their nutrition and waste recycling that they have lost their digestive and excretory systems. Our metagenomic analyses of 64 gutless oligochaete species from around the world revealed highly variable levels of fidelity not only across symbiont lineages, but also within symbiont clades. We hypothesize that in gutless oligochaetes, selection within host species for locally adapted and temporally stable symbiont communities leads to varying levels of symbiont fidelity and shuffles the composition of symbiont assemblages across geographic and evolutionary scales.

Introduction

Obligate symbioses where animals depend on microbial services are ubiquitous and have been independently established in multiple animal phyla[1–6]. Some of these, such as attine ants and giant tube worms substantially shape their ecosystems. Given their obligate nature, a difficult question to resolve is how they evolve, particularly in light of the observation that these associations maintain high symbiont fidelity where a given host lineage is stably associated with a symbiont lineage over evolutionary scales[7]. High symbiont fidelity can be achieved via two strategies - transmission from parent to offspring (vertical transmission) or host genotype dependent and specific acquisition of bacteria that is independent of parental transmission (horizontal transmission). Both modes of transmission have been documented across high fidelity nutritional symbioses[8,9] but their roles and dynamics are unexplored. From the point of host adaptation, the most extreme obligate associations are mouth and gutless animals that live in nutritional symbioses with chemosynthetic symbionts. Such chemosynthetic symbioses are predominantly characterized by a low number of partners, high symbiont fidelity or both[10–13]. A striking deviation from the common low diversity / high fidelity model are the gutless oligochaetes that depend on symbiont communities of four to six partners for nutrition as well as waste product recycling[14–17]. The communities are not sampled from the same six symbiont clades, but instead the seven host species analyzed up to this study associate with 14 symbiont groups[14,17–22]. In addition, low symbiont fidelity between these hosts and the primary symbiont – *Cand*. Thiosymbion – is strikingly common[23,24]. In one out of the seven species of gutless oligochaetes studied so far, even the main symbiont has been replaced with an unrelated lineage of gammaproteobacterial symbionts (Gamma4)[17]. In another species, this low fidelity also has been detected at a microevolutionary scale, as the 6 symbiont clades associated with specimens sampled around a small Mediterranean island had highly variable fidelity when compared to host mitochondrial lineages[24]. Most of these fidelity patterns were however assessed by comparing host mitochondrial and symbiont phylogenetic patterns. The maternal transmission of mitochondrial lineages could have led to such conflicting patterns if horizontal transmission and stringent selection of symbionts had been the basis of high fidelity. As nuclear evolution often deviates from mitochondrial genetic lineages, the effects of a stringent horizontal transmission established on nuclear encoded factors likely could not be tracked with mitochondrial markers[25–28]. Based on the observed variable fidelity between symbiont clades and mitochondrial genomes, we postulated that host species specific communities of gutless oligochaetes would be mediated by nuclear encoded factors, as it has been observed in other obligate associations[8]. In this study, we set out to test, if and how much host nuclear traits explain and drive the evolution of variable and multipartite symbiont communities. We therefore used a primer-free metagenomic approach to generate a host nuclear gene set (28S rRNA gene), a host mitochondrial gene set (mtCOI gene) and 16S rRNA gene based data on community composition. As such a community census vs. host genotypes analysis needs to cover large diversity of hosts from diverse habitats to be able to unlink host species and habitat effects, we collected 233 specimens from 17 globally distributed sites that represent 64 host species that all groups of gutless oligochaetes.

We compared host nuclear and host mitochondrial relationships with community structure and evolution of individual symbiont clades using multivariate statistics, symbiont clade-wise comparisons between host and symbiont phylogenies as well as divergence times and ancestral state reconstructions. We would have predicted that patterns of host nuclear and symbiont phylogenetic linkage also hold up at a microevolutionary. In contrast to our prediction that symbiont fidelity and community composition are linked to the host nuclear genome, we detected a pattern of forever changing communities, both unlinked to host nuclear and host mitochondrial lineages. The gutless oligochaete symbioses are characterized by variability and versatility, with the different symbiotic partners flexibly switching levels of fidelity that appear to only become temporally stable in a given environment.

Results and discussion

Nuclear and mitochondrial genomes of gutless oligochaetes evolved differently

Across animal diversity, several examples of conflict between the phylogenetic signal in nuclear and mitochondrial genomes have

been observed[25–28]. We therefore reconstructed both a nuclear marker gene, the 28S rRNA gene, and a mitochondrial marker gene, the mtCOI gene, for the 64 species-level host taxa from the metagenomic datasets. Our comparison of the 28S rRNA and mtCOI gene phylogenies showed that they were indeed not completely congruent (Figure S1). This incongruence indicates that the nuclear and mitochondrial genomes in gutless oligochaetes evolved differently. Thus, in the following paragraphs, we analyzed symbiont community composition as well as individual symbiont clade phylogenies separately in relation to both the 28S rRNA and the mtCOI gene phylogeny. Links between symbiont data to one or the other of the two host phylogenies would have different evolutionary explanations. Links to the nuclear phylogeny indicates that nuclear genome encoded traits mediate the overall community composition, the association with a given symbiont clade or both. In contrast, mitochondria do not encode for symbiosis mediating traits but are assumed to be maternally transmitted. We therefore assume that a link to the mitochondrial phylogeny indicates vertical, maternal transmission of symbiont clades.

Primer-free assessment of symbiont communities

To describe the symbiont communities across a wide range of host species and geographic locations in a primer-free approach, we analyzed 233 shot-gun metagenomes of gutless oligochaete individuals from 17 globally distributed sites in the Atlantic, Mediterranean, Indian, Pacific and Red Sea. We reconstructed full-length 16S rRNA genes of the symbionts and defined genus-level symbiont clades to characterize the composition of the symbiont community in each individual. For the detection of genus-level clades, we clustered the sequences at 95% identity, grouped the resulting clusters into potential symbiont clades according to their phylogenetic relationships and removed likely contaminants (see Note S1). After contamination removal, we obtained 33 genus-level symbiont clades of which 14 were not known to be associated with gutless oligochaetes before[14,17–22]. The detected clades belonged to the classes of Actinobacteria, Alpha-, Delta-, Gammaproteobacteria, Marinimicrobia and Spirochaetia.

To identify other hosts or environments that members of our

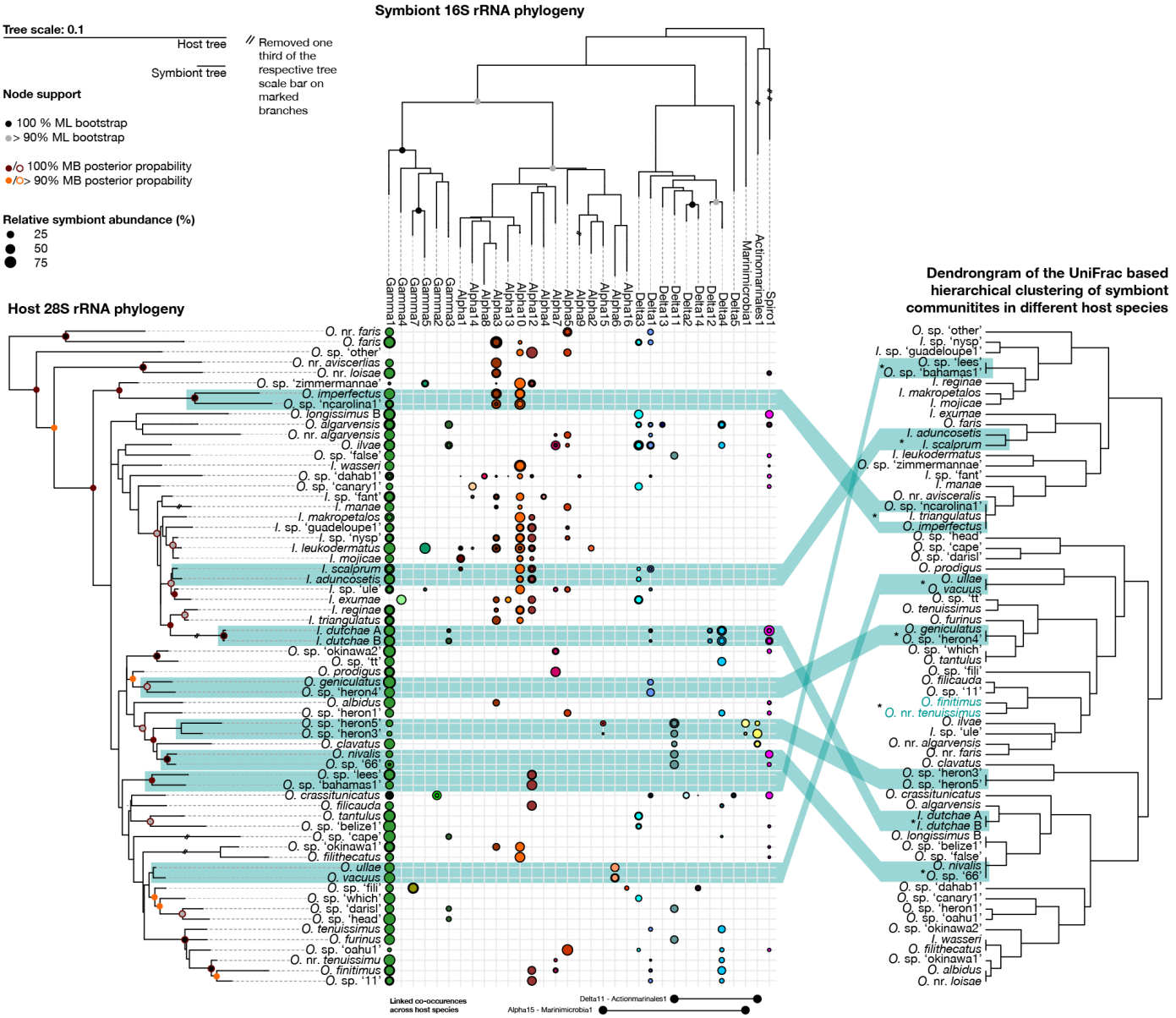

Figure 1: The composition of symbiont communities across 64 host species is not linked to host phylogeny. Top and left tree: Maximum-likelihood trees of the host 28S rRNA gene phylogeny (left) and the symbiont 16S rRNA gene phylogeny including one individual per host species/symbiont clade. The scale bar indicates 10% estimated sequence divergence. Nodes with none-parametric bootstrap support > 90% are highlighted in both trees. In addition, nodes with posterior probabilities > 90% estimated with MrBayes are highlighted in the host tree. Middle panel: Averaged relative abundance of symbiont clades per host species as estimated with EMIRGE. Right tree: Dendrogram generated by the hierarchical clustering of similarity of symbiont community composition of different host species calculated using UniFrac metric of host species wise averaged abundances of symbiont clades. Green boxes highlight host species that showed congruence between the 28S rRNA gene phylogeny and the community composition based dendrogram. *-marked host species showed congruence between the mtCOI gene phylogeny and the community composition based dendrogram.

symbiont clades are associated with, we screened publicly available 16S rRNA gene sequence data, and analyzed the phylogenetic relations between the symbionts and their closest relatives. Members of four clades were previously detected in environmental samples (Delta1, Delta3, Delta4 and Gamma3). Three of the gutless oligochaete symbiont clades were also detected in other unrelated marine invertebrates: the Gamma1 symbionts are also associated with Stilbonematinae and Astomonema nematodes, the Gamma5 and Gamma7 symbionts with Stilbonematinae nematodes and the Gamma4 symbionts with Kentrophoros ciliates (Figure S4-36). To understand whether gutless oligochaete symbiont clades were associated specifically with gutless hosts or also occurred in related marine oligochaetes in general we screened ten specimens that were morphologically identified as members of the closely related gut-bearing Phallodrilinae. Two symbiont clades, the Alpha3 and Alpha8, could also be detected in association with these gut-bearing relatives, indicating that most of the symbiont clades are linked to the gutless lifestyle of their hosts (Figure S37). Based on these results, we grouped the symbiont clades into three categories: i) symbionts only associated with gutless oligochaetes, ii) symbionts also associated with other marine invertebrates and iii) symbionts that are phylogenetically intermixed with free-living populations.

Limited numbers of community members form host species specific communities

The relative abundance of symbiont clades across host species revealed distinct communities (Figure 1). The communities of individuals of a single host species were relatively similar. Across host species, communities were made up of two to ten different symbiont clades. For the symbiont clades, the host ranges were highly variable ranging from being present in almost all host species to being present in a single species (for exact host ranges see Figure 1 and Table S2).

Our sampling effort of 64 host species resulted in a surprisingly low diversity of 33 symbiont clades. Although we systematically covered the host diversity and sampled approximately nine times more host species than all previous studies combined, we only identified roughly three times more symbiont clades[14,17–22]. This discrepancy and our rarefaction analysis showed a saturation of the detection of new symbiont clades, indicating that symbionts are only acquired from a limited pool of bacterial taxa (Fig. S2). This limited symbiont pool that was initially acquired from highly diverse sediment communities suggests that host traits, such as the immune system play an important role in controlling the symbiont communities of gutless oligochaetes as shown for other symbioses[29–33].

Besides host traits, symbiont-symbiont interactions could alter community composition as shown in other highly specific symbiont consortia in e.g. plant hosts[34]. Therefore, we tested for linked co-occurrences as well as linked exclusion patterns of symbiont clades using an unbiased network analysis. We found no examples for symbiont exclusion suggesting that community composition is not based on symbiont-symbiont competition. In addition, we found six linked co-occurrences of which four were detected between symbiont clades that co-existed in only a single host species. The other two were detected between symbiont clades that co-existed in closely related sister species of hosts. The low number of stable positive interactions that were limited to symbiont clades present in a single clade of hosts indicate that overall symbiont-symbiont interactions only play a very minor role in mediating community composition across gutless oligochaetes.

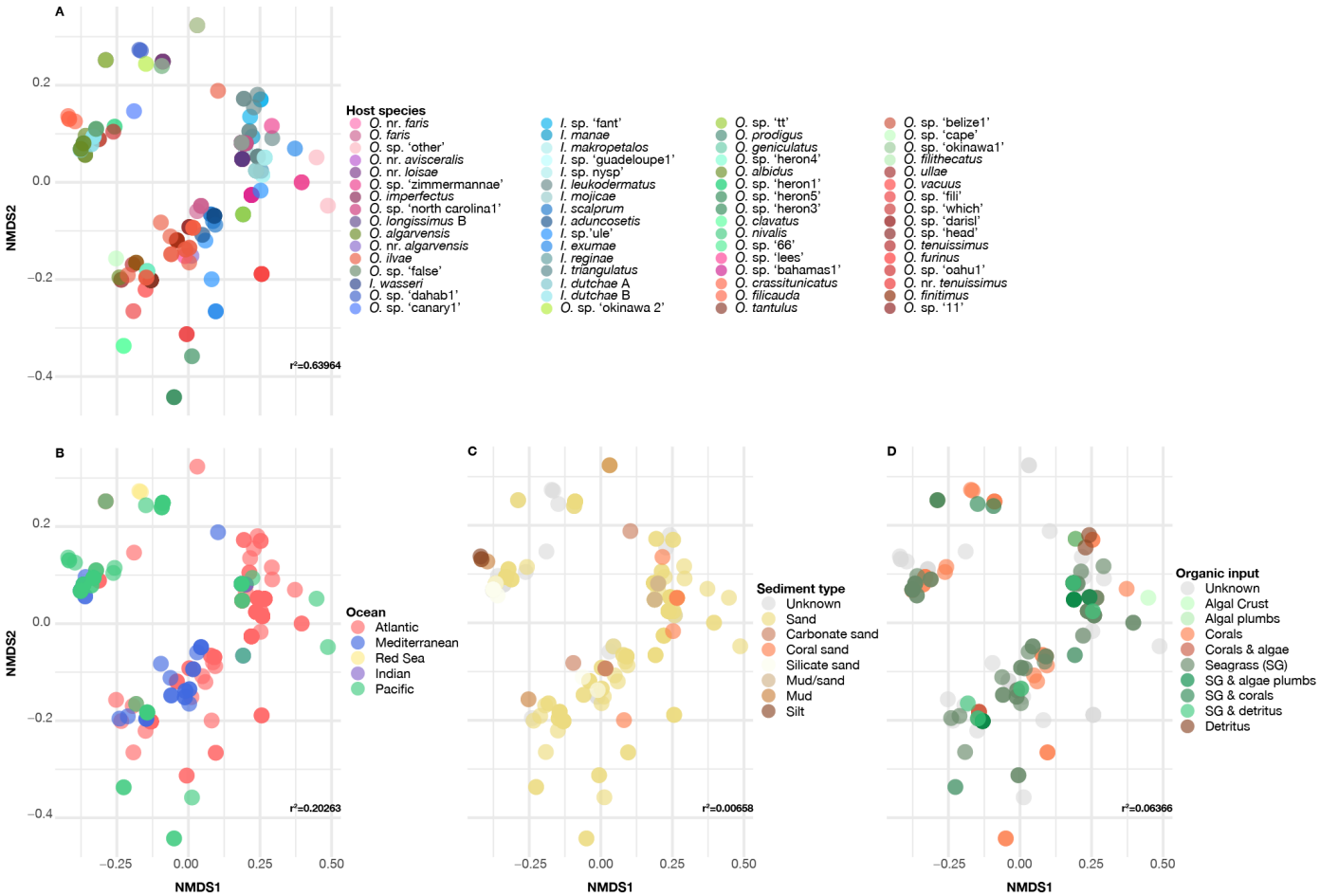

Figure 2: Symbiont community composition is linked to host species and largely unlinked to environmental parameters. All panels show the same NMDS plot of UniFrac dissimilarity values calculated from estimated relative abundances of symbiont clades in different host individuals. UniFrac stress value: 0.136. Different panels highlight samples differently according to certain metadata categories. For each metadata category, PERMANOVA r2 values are indicated within the respective plot. r2 values that are printed in bold were statistically significant.

To understand potential evolutionary and ecological drivers of symbiont community composition in gutless oligochaetes, we analyzed UniFrac distances of symbiont communities between host individuals in respect to host taxonomy and geographical, chemical and physical parameters of the environments. Host species dominated over environmental parameters as the major discriminatory factor to explain the composition of symbiont communities (PERMANOVA: host species: 63.96%, ocean: 20.26%, organic input: 6.37, sediment type: 0.66%, Mantel test for geographic distance: 13.16%, Figure 2). Despite the strong statistical link between symbiont community composition and host species, we also detected minor variations between individuals of the same host species, especially when they were sampled at different field sites (Table S2). Between individuals of the same species and from the same location, we detected the same set of symbionts but not all symbiont clades were always associated with all host individuals. In contrast, the set of symbiont clades of a given host species from different locations varied, suggesting that certain clades could provide special adaptations to their host in one environment that might not be needed at another location.

Symbiont community compositions evolved independently from host diversity over macroevolutionary scales

Although host species appeared to have a major influence on symbiont community composition, we found no link between host phylogeny and changes in symbiont community composition, a phenomenon described as phylosymbiosis. This applies to both the nuclear and the mitochondrial phylogeny when testing their

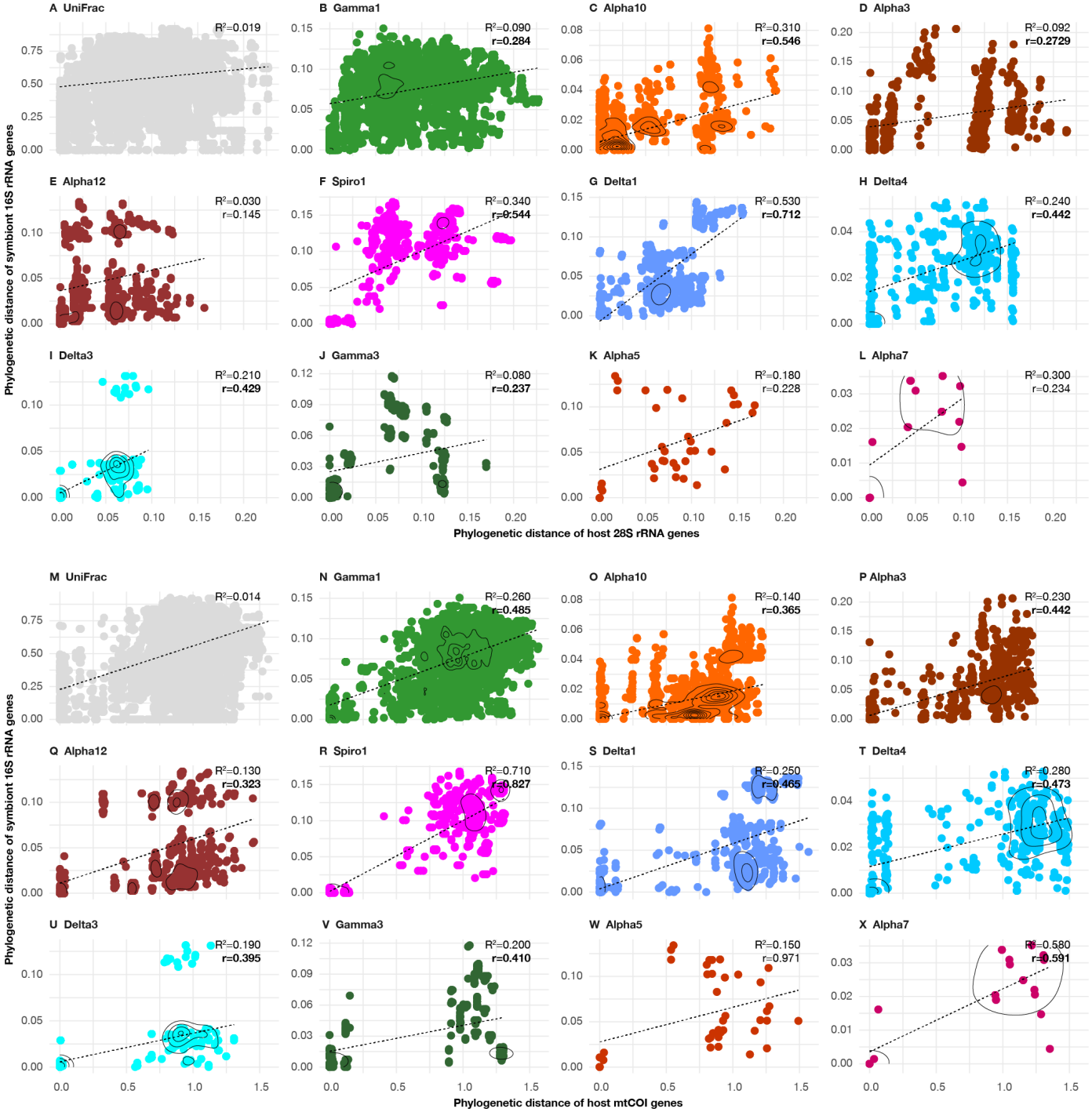

Figure 3: Both symbiont community composition and phylogenetic relations within symbiont clades are variably linked to phylogenetic relations between their respective host individuals. Each panel shows the pairwise UniFrac dissimilarity of the overall symbiont community composition or pairwise 16S rRNA nucleotide dissimilarity of symbionts from one of the elven most abundant symbiont clades versus the nucleotide dissimilarity of host marker genes of pairs of host individuals. **(A-L)** Host marker: 28S rRNA gene. **(M-X)** Host marker: mtCOI gene. Dots represent the UniFrac/16S rRNA nucleotide dissimilarity over the host marker gene nucleotide dissimilarity of each pair of host individuals. Solid lines highlight areas with high density of data points. Dotted lines represent the regression curve estimated by applying a linear model. R^2 values of each linear regression are given in the respective plot panel. Additionally, r values resulting from testing correlation between UniFrac/16S rRNA nucleotide distance vs. host marker gene nucleotide distance using the Mantel test are also given in each plot panel. Bold r values were statistically significant ($p \leq 0.05$).

congruences to the community composition dendrogram based on UniFrac dissimilarities (Figure 1, 28S rRNA topology comparison: nRF=0.9937, p-value=1.0, nMC= 0.7847, p-value=0.8754; mtCOI topology comparison: nRF=0.8852, p-value=0.0, nMC=0.6126, p-value=0.00252). Although there was no overall congruence between the host phylogenies and the variation in symbiont community composition, we detected nine non-random examples of host sister species that were associated with very similar symbiont communities (t-test p-values for 28S rRNA and mtCOI: $< 2.2^{-16}$, Figure 1). These host sister species tended to be more closely related than other sister species with divergent symbiont communities (Mann-Whitney-U test p-values: $< 2.2^{-16}$ for 28S rRNA phylogeny and 1.482^{-13} for mtCOI phylogeny). In concordance with the topology-based analysis of phylosymbiosis, the analysis of the relation between phylogenetic distances and the symbiont community composition distances of all host individuals showed a weak linear correlation (Figure 3A and 4M). This weak correlation was based on the fact that many host pairings had more different symbiont communities than we would have expected in a strict phylosymbiotic relation. Also, the opposite extreme was true, namely that some hosts shared more similar symbiont communities than we would have expected from their phylogenetic distance, but these cases were rarer. These examples illustrated that the host, either via vertical transmission or inheritable traits that convey specificity, can influence symbiont community composition. These mechanisms at work at the host species level are apparently overpowered over macroevolutionary scales.

Symbiont clade-level variability of symbiont fidelity allows the evolution of variable and adaptive symbiont communities

We identified varying symbiont fidelity as one of the factors that might influence symbiont community composition and disrupt phylosymbiosis. Our clade wise analyses of symbiont community composition revealed several examples of low symbiont fidelity, including *de novo* acquisition, host switching and loss of symbiont clades in various host lineages. One main source of variability of symbiont communities is the acquisition of new symbiont clades. We used ancestral state reconstruction of symbiont occurrences and divergence time estimates of the symbiont clades to reconstruct the time frame of the primary acquisition of a given symbiont clade (Figure S4-36 and S38, Note S2). Our analysis suggests that only few clades were acquired early in the gutless oligochaete evolution and *de novo* acquisition of new symbiont clades continuously increased over time. This points towards a high importance of the recent time window for symbiont clade establishment as well as for accelerating evolutionary flexibility and potential specializations.

In addition, repeated uptake in divergent host lineages or low fidelity inherence of a given symbiont clade after its primary acquisition caused variability in community composition. Many of the rather young symbiont clades were likely acquired only once by a rather recent last common ancestor of their extant host species and only rarely switched between distantly related host lineages. Thus, they were mainly found in small, monophyletic host groups. In contrast, many of the older symbiont clades were not confined to monophyletic groups of hosts but showed patchy distributions across the host diversity suggesting frequent uptakes or losses (Figure 1 and Figure 3). Ancestral state reconstructions of symbiont occurrence patterns suggested that the majority of the symbiont clades independently established their symbioses with distinct host lineages several times (Figure S4-36). As most of these symbiont clades were not phylogenetically intermixed with free-living relatives, we assume that the majority of these repeated acquisitions happened via host switching rather than environmental acquisition (Figure S39-71). In addition to the previously published broad host range of the Gamma1symbiont, the Alpha3, Alpha8, Gamma4, Gamma5 and Gamma7 symbionts were also associated with other marine invertebrates that share the same environment (Figure S39-71)[23]. Thus, symbiont acquisition and host switching likely do not only happen between gutless oligochaete individuals but also between different invertebrate phyla.

On microevolutionary scales, fidelity varies across symbiont clade – host species pairings

To understand microevolutionary dynamics within symbiont clades, we analyzed phylotype exchange across the host diversity for each symbiont clade. We assessed fidelity of symbiont host associations by testing for a possible correlation between the host individuals' genetic distances and each of the symbiont clade genetic distances (Figure 3). We assumed that high fidelity between symbionts from a certain clade and their hosts would lead to a linear correlation between the genetic distance of host pairs and the genetic distance of the respective symbiont phylotypes. We detected significant correlations for a majority of the tested symbiont clades (Mantel test p-value < 0.05: 8/11 symbiont clades vs. 28S rRNA genetic distance and 10/11 vs. mtCOI, Figure 4). Despite this statistical significance, low correlation coefficients however point to a low predictive power of host relations for symbiont selection and rather illustrate rampant and ongoing symbiont phylotype exchange between host individuals from different species (Figure 3).

This striking result and insights from a single gutless oligochaete species, *O. algarvensis,* pointed us to analyze the symbiont fidelity on the smallest evolutionary scale our data provided – the phylotype association patterns within host species (Figure 4)[24]. Only few symbiont clade - host species pairings showed perfect correlation between host and symbiont genetic distances and overall, statistical significance for the correlation between host and symbiont genetic distances was low (28S rRNA: 1 of 77, mtCOI: 3 of 81). All other symbiont clade - host species pairings exhibited a variable range of correlations across the tested host species, independent of the symbiont phylogeny and host marker gene (Figure 4A and B). This suggests that symbiont fidelity is controlled by factors independent of symbiont phylogeny.

In a second step, we compared the fidelity patterns of symbiont communities from the three most sampled host species from a single location. Within a given host species the community members had variable levels of fidelity as also reported by Sato *et al.* (2021) for 80 specimens of *O. algarvensis* sampled from a single Mediterranean island[24]. Each host species had specific patterns of fidelity for its symbiont community members that were different for both host marker genes and diverged from the other host species. The shared community members that occurred in two or three of these host species also showed diverging fidelity patterns in different host species and for both marker genes (Gamma1, Gamma3, Delta1, Delta3, Figure 4C and D). This indicated that certain levels of fidelity are not general traits of either the symbiont clade or the host species but specific to a given symbiont clade – host species pairing and specific to the host marker gene used (Figure 4C and D). Given the observation that high symbiont fidelity is often correlated with a high degree of dependence of an association, we speculate that the dependence in a given symbiont clade – host species pairing could be encoded in the observed degree of fidelity[35].

Obligate symbiont communities gain flexibility through varying levels of symbiont fidelity

Taken together, our results of symbiont to host specificity patterns across evolutionary scales indicate that versatility dominates over stringent specificity. Clade-level community analyses would underestimate the versatility in the gutless oligochaete symbiosis as both across and within host species switches of symbiont phylotypes of a given clade are frequent but do not alter clade-level community composition. Microevolutionary patterns of phylotype mismatches

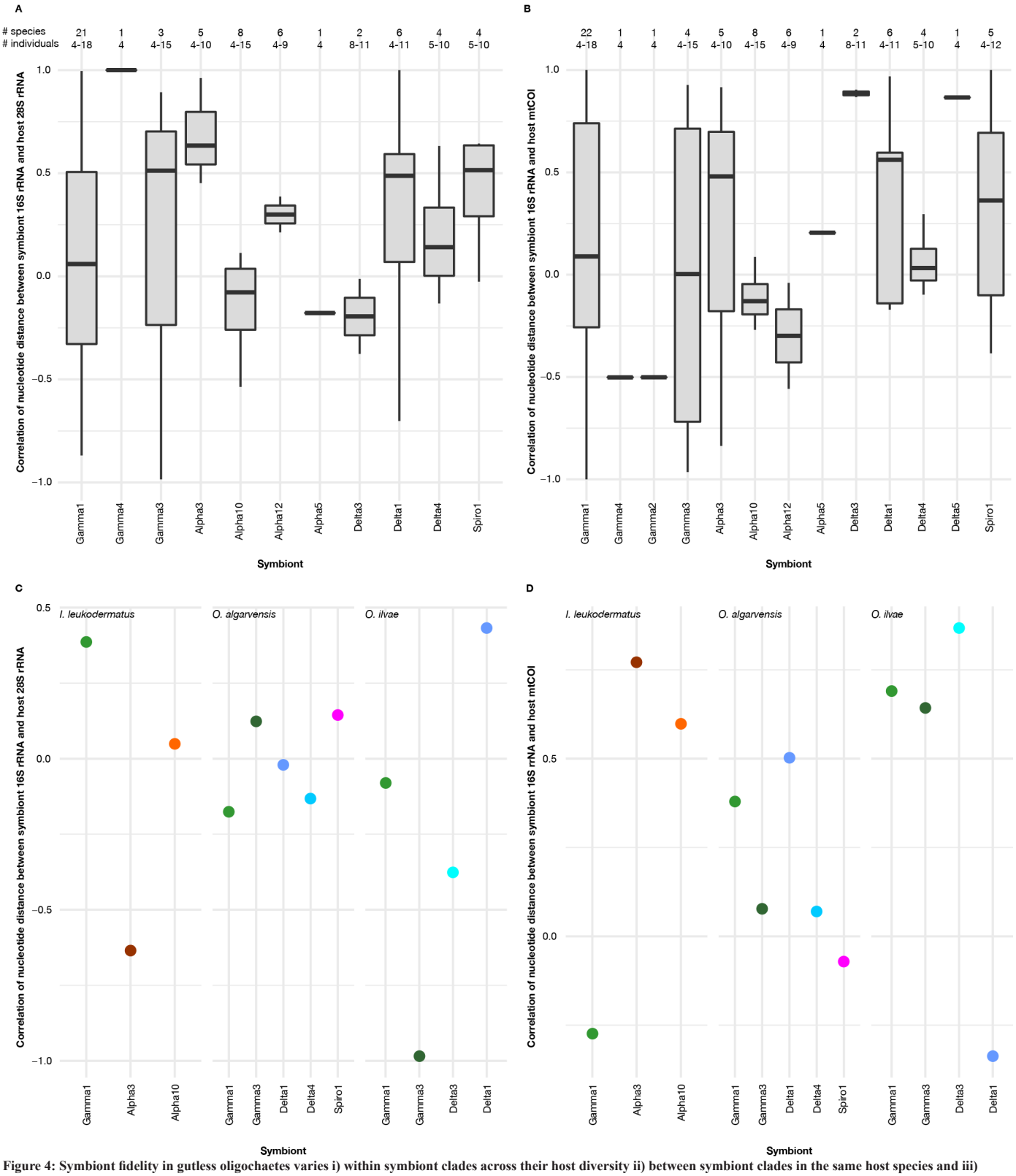

Figure 4: Symbiont fidelity in gutless oligochaetes varies i) within symbiont clades across their host diversity ii) between symbiont clades in the same host species and iii) between host nuclear or mitochondrial markers. All panels show the degree of symbiont fidelity between a certain symbiont clade in a given host species as correlation coefficient estimated by the Mantel test. **(A & B)** Correlation coefficients of 16S rRNA gene distance of different symbiont clades from different host species and the genetic distance of host marker genes of host individuals the symbionts were associated with. **(C & D)** Correlation coefficients of 16S rRNA gene distance of different symbiont clades and gene distance of host marker genes from the three host species where most individuals were sequenced of from one sampling site. **(A & C)** Host marker gene: 28S rRNA. **(B & D)** Host marker gene: mtCOI.

suggest that macroevolutionary patterns are based on population-level flexibility that also varies between host species.

The observed linkage patterns to the mitochondrial genotypes for some symbiont clade – host species pairings suggested that for some settings, high symbiont fidelity appears to be ensured via vertical transmission from mother to offspring. This apparent vertical transmission does not imply long term stability as we did not observe fidelity between symbiont phylotypes of a given clade and mitochondrial host genotypes across host species. In other settings, high symbiont fidelity appears to be selected for by nuclear host traits as shown by examples of strong linkage to the nuclear genotypes for other symbiont clade – host species pairings. Similar to the linkage patterns between symbiont phylotype and host mitochondrial genotype, the linkage between symbiont phylotypes and host nuclear genotype are also not stable across host species.

Our results indicate that adaptive symbiont communities can flexibly

evolve even when they are obligate to their hosts. In the strict obligate symbiosis of gutless oligochaetes, variability appears to be achieved by varying degrees of fidelity between host individuals and symbiont phylotypes. The local environment then appears to select and stabilize associations on a temporal scale. Any observed fidelity is likely linked to the dependence of a given host on a given symbiont in a given setting. Thus, the presence of the essential symbionts is selected for while the variability of other, non-essential clades could provide evolutionary and metabolic flexibility of the whole community. In an abstract way, symbiont communities of gutless oligochaetes could be compared to an evolutionary kaleidoscope that appears to have gained complexity and currently forms very distinct patterns of a limited set of symbiont clades that appear to be unlinked to host phylogeny over evolutionary time.

Variation of symbiont fidelity leads to 'forever young' symbiont communities

Several factors can lead to variation in the symbiont communities across the individuals of a given population: parental inheritance, host switching, loss and *de novo* environmental acquisition. Together these factors can balance between the benefits and the trade-offs of high fidelity symbioses. In such high fidelity symbioses, the biggest trade-off is symbiont genome streamlining that often leads to deleterious genome reductions and the decay of the symbiotic association[36–44]. At the genome level, a departure from this one-way from the 'cradle to the grave' scenario was suggested by Russel *et al.* (2020) who showed that symbionts with low fidelity due to frequent phylotype exchanges between host individuals have higher homologous recombination rates[45]. These recombinations prevent massive genome erosion and keep symbiont genomes 'forever young' when compared to symbiont genomes of high fidelity symbioses of a similar age [45].

We extend this concept to the community level. Based on our data we argue that in gutless oligochaetes, symbiont fidelity varies on an even broader level as we observed not only phylotype exchange but also symbiont acquisition, loss and host switching leading to genus-level variation in symbiont communities across a broad host diversity. Besides homologous recombination within a single symbiont clade described by Russell *et al.*, (2020) the observed community versatility could also allow for constant variation in the pool of metabolic functions encoded in a given symbiont community and at the same time likely prevents the loss of key functions[45]. We would therefore argue that this level of varying symbiont fidelity might not only keep the genomes of the symbionts 'forever young' but also could enable 'forever young' symbiont communities.

Conclusion

Symbiont community composition is the result of the evolutionary dynamics of its single members. Understanding the dynamics between host species and symbiont clades is only possible when a broad host diversity is analyzed. Given such a broad taxon sampling and a sufficient phylogenetic resolution, we can start to link microevolutionary patterns of symbiont fidelity of individual symbiont clades in a given host species and macroevolutionary patterns of symbiont community composition across the host diversity.

In gutless oligochaetes, symbiont fidelity is much more variable than anticipated for most examples of obligate, chemosynthetic symbioses and is building the foundation for variable yet stable symbiont community composition. Fidelity is not linked to symbiont clades or host species but to a given symbiont clade – host species pairing in given environment. Thus, symbiont community compositions appear to be highly specific to closely related hosts from the same environment but become unlinked from host evolutionary or geographic patterns over time.

Overall, varying symbiont fidelity seems to be a useful evolutionary strategy to balance the benefits of stable and flexible associations in obligate symbiont communities. So far, obligate symbiont communities have been mainly found to display low symbiont diversity and high symbiont fidelity. However, unbiased metagenomic community assessment of such associations has been rarely performed. Considering the evolutionary and ecological benefits that are connected to varying symbiont fidelity and community composition, it might be worthwhile to extend the taxon sampling of obligately dependent hosts and analyze the prevalence of symbiont variability across evolutionary scales.

Material and methods

Sample collection, processing and metagenomic sequencing

233 individuals of gutless and 10 individuals of gutbearing oligochaetes were sampled at various field sites between 1991 and 2018 (for overview see Table S1). Individual specimens were either flash-frozen in liquid nitrogen and stored at -80°C or fixed in RNAlater (Thermo Fisher Scientific, Waltham, MA, USA) and stored at 4°C or -20°C. DNA was extracted from single specimens with the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Library construction, quality control and sequencing were performed at the DOE JGI (Walnut Creek, California, USA) and the Max Planck Genome Centre (Cologne, Germany). Information on library preparation and sequencing details are listed in Table S2. Some samples were sequenced twice to generate a higher number of reads. In these cases, resulting reads from both sequencing runs were combined for further analyses. Also, two samples were extracted twice using different library preparation methods and individually sequenced. The resulting sequences were also pooled for further analyses.

Assembly of host marker genes

28S rRNA and mtCOI genes of all specimens were assembled by mapping the metagenomic reads to respective databases using bbmap v38.34 (https://sourceforge.net/projects/bbmap/). For the 28S rRNA, we used the SILVA database v138 [46,47]. Mapped reads were assembled using SPAdes v3.11.0 setting k-mer sizes to 99, 111 and 127 bp [48]. Final 28S rRNA gene sequences were predicted from the assembled sequences using barrnap v0.9 (https://github.com/tseemann/barrnap). MtCOI genes were assembled by adapting the PhyloFlash pipeline to operate on a custom mtCOI reference database and predict mtCOI genes from assembled sequences[49]. In case that several mtCOI genes were assembled, we only considered the most abundant one.

Identification of host taxa

Species-level host taxa were defined based on mtCOI gene phylogenies that also included the gene sequences of previously identified specimens. Specimens that could not be assigned to published species based on morphological or molecular data were treated as new taxa and were assigned provisional names with consecutive numbers and the sampling location. The gutbearing oligochaete specimens could be identified based on morphological traits.

Host marker gene phylogenies

28S rRNA and mtCOI gene sequences were aligned using mafft-linsi v7.407[50–52] . The mtCOI alignment was manually trimmed in Geneious v11.1.5 and bases 40-695 were kept (https://www.geneious.com). The best suited model for the Bayesian inference based phylogeny was estimated using the MODELTEST function of iqtree v1.6.10[53]. Bayesian inference based phylogenies were calculated using MrBayes 3.2.7a, using 4 chains, running for 4,000,000 generations and applying the GTR+G+I model[54,55]. The

sample frequency was set to 1000 and the print frequency was set to 500. 1,000 trees were discarded as initial burn-in. All estimated parameters were controlled to show an effective sampling size (ESS) > 200 in Tracer v1.7.1[56]. Maximum-likelihood based phylogenies were calculated using iqtree, including automatic selection of the best suited model and generation of 100 none-parametric bootstrap replicates. The sequences of one gutbearing oligochaete specimen (*Phallodrilinae* gen. sp. 'strang') were used to root the phylogenies. The original tree was calculated on the full alignment, subtrees that were used in subsequent analyses were obtained by manually pruning the tree in iTol[57].

Symbiont clade definition and quantification

16S rRNA genes were assembled from the metagenomic libraries of gutless oligochaetes using phyloFlash. using the –all option and in addition specifying the read length. For subsequent analyses, we only considered sequences that were i) assembled with SPAdes, ii) longer than 1000 bp and iii) did not contain more than 20 ambiguous bases. The resulting sequences were clustered at 95% sequence similarity using usearch v10.0.240[58]. We used the SINA search and classify algorithm to add the 16S rRNA gene sequences of close relatives from the SILVA database v132 that shared at least 90% sequence similarity for each of our assembled symbiont sequences[59]. All assembled sequences and the SILVA database hits were aligned using mafft-linsi and a phylogenetic tree was calculated from the resulting alignment using FastTree v2.1.1[60]. We mapped the 95% clusters to this tree and manually merged monophyletic clades that consisted of several of the 95% clusters into single symbiont clades. We analyzed the prevalence of all phylogenetically defined symbiont clades across the gutless oligochaete metagenomic libraries. We excluded clades that had distribution patterns that suggested they were contaminations or spurious associations (Note S1). The abundances of the remaining clades (symbiont clades from here on) were quantified across all metagenomic libraries using EMIRGE v.0.61.1 following the standard workflow for custom emirge databases[61].

Phylogeny of all symbionts and their relatives

All sequences included in the symbiont clades defined above were used to obtain sequences from closely related bacteria from the SILVA and the RefSeq public databases[62]. For SILVA, we used the SINA search and classify algorithm to obtain up to 10 relatives for each sequence that shared at least 99% and 95% sequence similarity for each of our input sequences. In addition, we also screened the RefSeq database using BLAST implemented in Geneious v11.1.5 to obtain the ten most similar 16S rRNA genes[63]. Duplicated sequences were removed from the collection of sequences of the symbionts' relatives. In addition, we included the 16S rRNA gene sequence of *Crenarchaeotal* sp. clone JP41 (NCBI accession: L25301.1) as outgroup. The resulting sequence collection was aligned using mafft-linsi and a phylogenetic tree was calculated using iqtree including automatic selection of the best suited model and generation of 100 none-parametric bootstrap replicates. Subtrees that were used in further analyses were pruned from the resulting tree using iTol.

Individual symbiont clade phylogenies

For the calculation of trees of individual symbiont clades, the symbiont 16S rRNA gene sequences of each clade were treated individually. We used the SINA search and classify algorithm to obtain up to 10 relatives that shared at least 90% sequence similarity for each of the input sequences from the SILVA database v138.1. In case of the Gamma7 and the Alpha9 symbiont clade, we did not obtain any relative sequence at this threshold. For these clades, we obtained up to 10 relatives that shared at least 85% sequence similarity for each input sequence instead. In addition, we clustered the symbiont sequences at 98% sequence similarity using the cluster_fast algorithm of usearch. We used the resulting centroids of every symbiont clade to obtain the 5 most similar 16S rRNA gene sequences from the RefSeq database using BLAST implemented in Geneious v11.1.15. Duplicated sequences were removed from the collection of sequences of the symbionts' relatives. In addition, we included the 16S rRNA gene sequence of Crenarchaeotal sp. clone JP41 (NCBI accession: L25301.1) as outgroup. The resulting sets of sequences of each symbiont clade were aligned using mafft-linsi. A maximum-likelihood phylogeny was calculated using iqtree including automatic selection of the best suited model and generation of 100 none-parametric bootstrap replicates.

Estimates of divergence times for host and symbiont clades and reconstruction of ancestral states of symbiont association patterns

For the estimation of host divergence times, we used a Bayesian phylogenetic framework and a relaxed molecular clock model. We constructed a matrix of eight 28S rRNA gene sequences of gutless host and eight publicly available 28S rRNA gene sequences of other *Oligochaeta* and one representative of the *Polychaeta*. The oligochaete representatives were selected to i) cover a broad diversity of the phylum and ii) to include the following calibration points for our molecular clock model: the last common ancestor of the *Goniadidae* (323 Mya), the last common ancestor of the *Hormogastridae* (82 ± 15 Mya), the divergence between *Hirudienea* and *Lumbriculidae* (201 Mya) and the last common ancestor of the *Phyllodocida* (4.85 ± 1.9)[64,65]. The polychaete sequence was included to root the tree and to include the last common ancestor of all *Annelida* (510 ± 10 Mya) as additional calibration point. All calibration points were considered as uniform priors. All sequences were aligned using mafft-linsi and the time calibrated tree was calculated in BEAST v2.6.3[66] using the GTR+G+I model and the relaxed log normal clock model. Besides the mentioned priors for time calibration, we also set Alpha and Beta values of birtRate.Y.t prior to 0.001 and 1,000, respectively. We also defined priors that considered the oligochaetes and the gutless oligochaetes as monophyletic groups. The chain length was set to 100,000,000 generations and a 10% burn-in was defined. All estimated parameters were controlled to show an ESS > 200 in Tracer.

We used the same approach as for the host to estimate the symbiont clade divergence times based on a subset of our symbiont sequence matrix that combined 2-3 symbiont 16S rRNA gene sequences per symbiont clade with 50 typestrain sequences from RefSeq databases (Suppl. Data). Typestrains of the *Chromatiaceae* were used to include their previously published divergence estimate as calibration point for the symbiont analysis[67]. In addition, we included the 16S rRNA gene sequence of *Crenarchaeotal* sp. clone JP41 (NCBI accession: L25301.1) as outgroup. The time calibrated tree was calculated in BEAST v2.6.3, using the GTR+G+I model and the relaxed log normal clock model. The divergence time of the *Chromatiaceae* was considered as an exponential prior with a mean value of 0.1 and an offset of 1.64. We additionally constrained the analyses by setting a uniform prior from 3.5-4.5 billion years for the whole dataset to account for the maximum age of life on earth. Additional priors were used to define monophyletic clades for all bacteria, the Delta1, Delta4 and Delta12 clades as well the combined Delta4-Delta12 clade that were observed in the previous phylogenetic analyses of the symbiont clades. We also set Alpha and Beta values of birtRate.Y.t prior to 0.001 and 1,000, respectively. We ran 4 parallel chains, setting the chain length to 500,000,000 generation and a 10% burn in was defined. All estimated parameters were controlled to show an ESS > 200 in Tracer.

Ancestral states of symbiont presence/absence patterns were calculated using the phytools package in R v3.6.3 and mapped onto the 28S rRNA gene phylogeny of the hosts (R Core Team, 2020) [68].

Analyses and plotting of symbiont community composition

The analyses of symbiont community composition were performed in R v3.6.3 unless differently stated. During the analyses, the following packages were used: phyloseq[69], ape[70], vegan (https://github.com/vegandevs/vegan), plyr[71], MASS[72], gdata (https://cran.r-project.org/web/packages/gdata/index.html), reshape2 (https://github.com/hadley/reshape), forcats (https://github.com/robjhyndman/forecast), igraph (https://github.com/igraph/rigraph), Hmisc (https://github.com/harrelfe/Hmisc/), optparse (https://github.com/trevorld/r-optparse), data.table (https://github.com/Rdatatable/data.table), ade4 (https://github.com/sdray/ade4), tidyverse[73] and spa (https://github.com/markvanderloo/rspa). Plots were generated using ggplot2 from the tidyverse package, gridExtra (https://cran.r-project.org/web/packages/gridExtra/index.html), ggpubr (https://cran.r-project.org/web/packages/ggpubr/index.html), maps (https://www.rdocumentation.org/packages/maps), mapdata (https://www.rdocumentation.org/packages/mapdata), and patchwork (https://github.com/thomasp85/patchwork).

Community composition analyses

The similarity between symbiont communities of host individuals were calculated based on the abundance patterns of the symbiont clades and the symbiont 16S rRNA gene phylogeny using the UniFrac metric as implemented in the phyloseq package in R. We tested for parameters that could explain differences in symbiont community composition between individuals using PERMANOVA and the Mantel test[74,75]. We only considered parameters that were collected for at least 50% of the samples. These included: host species, ocean, continent, field site, GPS coordinates, organic input, sediment type, water depth, sampling month and sampling year (Suppl. Table 3). All factors except for geographical distances were treated as categorical data and analyzed using PERMANOVA. Geographical distance was treated as the correlation between the UniFrac distances and actual geographic distances and analyzed using the Mantel test.

Co-occurrence patterns of symbiont clades were analyzed using Spearman's correlations and were corrected using the Benjamini-Hochberg standard false discovery rate correction.

Phylosymbiosis

UniFrac distances on the average symbiont abundances per host species were transformed into a dendrogram using hierarchical clustering. The congruences between the 28S rRNA and the mtCOI based host tree or and the symbiont community UniFrac dendrogram was assessed separately using the normalized Robinson-Fould metric and the normalized Matching Cluster metric, implemented in TreeCmp v1.0-b291[76–78]. Statistical significance was estimated by comparing the congruence between the host phylogeny vs. 1000 random trees as described by Brooks *et al.* 2016, https://github.com/awbrooks19/phylosymbiosis)[79]. The relation between host phylogenetic distances and the symbiont community composition distances was analyzed using linear regression and the Mantel test.

Correlation between host and symbiont phylogenetic distances

For all hosts with member sequences of a given symbiont clade we calculated pairwise phylogenetic distances of the hosts' 28S rRNA or the mtCOI genes as well as pairwise distances of the 16S rRNA gene sequences using from the respective phylogenies using the R's cophenetic function. We analyzed the correlation between the host and symbiont genetic distances using linear regression and the Mantel test.

Data and script availability

Raw metagenomic sequences as well as host and symbiont marker genes generated in this study will be deposited in the European Nucleotied Archive (ENA) upon peer-review submission and are currently available upon request. Reference sequences, nucleotide alignments, phylogenetic trees as well as abundance tables that were used and/or generated during this study are also available upon request. The scripts and data for analyzing symbiont community composition and phylogenetic correlations are available under: https://github.com/amankowski/GO_symbiont-diversity

Acknowledgments

We are thankful for sample collections and field assistant by Alexander Gruhl, Anna Ansebo, Anna Blazejak, Anne-Christin Kreutzmann, Christian Lott, Claudia Bergin, Dolma Michellod, Emilia M. Sogin, Erica Mejlon, Erich Mueller, Falk Warnecke, Fred Wells, Jörg Ott, Judith Zimmermann, Katrine Worsaae, Ken Halanych, K. B. Brandon Seah, Lisa Matamoros, Lena Gustavsson, Mario P. Schimak, Michael Hadfield, Miriam Sadowski, Miriam Weber, Olav Giere, Oliver Jäckle, Nicholas Bekkouche, Olivier Gros, Pamela Reid, Philippe Bouchet, Pierre De Wit, Ramon Rosello-Mora, Silke Wetzel, Silvia Bulgheresi, Stefan Sommer, and Tina Enders. In addition, we would like to thank the crew of the Meteor cruise M92, as well as the Carrie Bow Cay Laboratory, the Heron Island Research Station, the HYDRA Institute Elba, the Mediterranean Institute for Advanced Studies, the Lee Stocking Island Research Station, the Little Darby Island Research Station, the Lizard Island Research Station, and the Okinawa Institute of Science and Technology and their staff for supporting our sampling campaigns. This work was supported by the Max Planck Society, a Moore Foundation Marine Microbial Initiative Investigator Award to ND (Grant GBMF3811), a U.S. National Science Foundation award to MK (grant IOS 2003107), the USDA National Institute of Food and Agriculture Hatch project 1014212 (MK), and a Marie-Curie Intra-European Fellowship PIEF-GA-2011-301027 CARISYM (HRGV). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. This work is contribution XXX from the Carrie Bow Cay Laboratory, Caribbean Coral Reef Ecosystem Program, National Museum of History, Washington DC.

Author contributions

AM, MK, HRGV, JW and ND conceived the study and AM designed the workflows with help from HRGV and JW. AM, MK, CÉ, NL, YS, JMV, BH, CW, TW, JW and HRGV acquired specimens and generated metagenomic data. AM analyzed the data. AM, HRGV, MK and ND interpreted the results. AM drafted the manuscript and AM and HGV edited the manuscript. All authors provided revisions.

Literature

1. Felbeck, H. Chemoautotrophic Potential of the Hydrothermal Vent Tube Worm, *Riftia pachyptila* Jones (Vestimentifera). *Science* **213**, 336–338 (1981).

2. Cavanaugh, C. M., Gardiner, S. L., Jones, M. L. & Jannasch H. W. & Waterbury, J. B. Prokaryotic Cells in the Hydrothermal Vent Tube Worm *Riftia pachyptila* Jones: Possible Chemoautotrophic Symbionts. *Science (80-.).* **213**, 340–342 (1981).

3. Cavanaugh, C. M. Symbiotic chemoautotrophic bacteria in marine invertebrates from sulphide-rich habitats. *Nature* **302**, 58–61 (1983).

4. Ott, J., Rieger, G., Rieger, R. & Enderes, F. New Mouth less Interstitial Worms from the Sulfide System: Symbiosis with Prokaryotes. *Mar. Ecol.* **3**, 313–333 (1982).

5. Ott, J., Bright Monika & Bulgheresi, S. Symbioses between marine nematodes and sulfur-oxidizing chemoautotrophic bacteria. *Symbiosis* **36**, 103–126 (2004).

6. Schultz, T. R. & Brady, S. G. Major evolutionary transitions in ant agriculture. *Proc. Natl. Acad. Sci.* **105**, 5435 LP – 5440 (2008).

7. Douglas, A. E. & Werren, J. H. Holes in the Hologenome: Why Host-Microbe Symbioses Are Not Holobionts. *MBio* **7**, e02099-15 (2016).

8. Nussbaumer, A. D., Fisher, C. R. & Bright, M. Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* **441**, 345–348 (2006).

9. Gruber-Vodicka, H. R. *et al. Paracatenula*, an ancient symbiosis between thiotrophic *Alphaproteobacteria* and catenulid flatworms. *Proc. Natl. Acad. Sci.* **108**, 12078 LP – 12083 (2011).

10. Di Meo, C. A. *et al.* Genetic Variation among Endosymbionts of Widely Distributed Vestimentiferan Tubeworms. *Appl. Environ. Microbiol.* **66**, 651 LP – 658 (2000).

11. Polzin, J., Arevalo, P., Nussbaumer, T., Polz, M. F. & Bright, M. Polyclonal symbiont populations in hydrothermal vent tubeworms and the environment. *Proc. R. Soc. B Biol. Sci.* **286**, 20181281 (2019).

12. Distel, D., Felbeck, H. & Cavanaugh, C. Evidence for phylogenetic congruence among sulfur-oxidizing chemoautotrophic bacterial endosymbionts and their bivalve hosts. *J. Mol. Evol.* **38**, 533–542 (1994).

13. Currie, C. R. *et al.* Ancient Tripartite Coevolution in the Attine Ant-Microbe Symbiosis. *Science (80-.).* **299**, 386 LP – 388 (2003).

14. Dubilier, N. *et al.* Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. *Nature* **411**, 298–302 (2001).

15. Woyke, T. *et al.* Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**, 950–955 (2006).

16. Kleiner, M. *et al.* Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1173--82 (2012).

17. Bergin, C. *et al.* Acquisition of a Novel Sulfur-Oxidizing Symbiont in the Gutless Marine Worm *Inanidrilus exumae*. *Appl. Environ. Microbiol.* **84**, e02267-17 (2018).

18. Dubilier, N. *et al.* Phylogenetic diversity of bacterial endosymbionts in the gutless marine oligochete *Olavius loisae* (Annelida). *Mar. Ecol. Prog. Ser.* **178**, 271–280 (1999).

19. Dubilier, N., Giere, O., Distel, D. L. & Cavanaugh, C. M. Characterization of chemoautotrophic bacterial symbionts in a gutless marine worm (Oligochaeta, Annelida) by phylogenetic 16S rRNA sequence analysis and in situ hybridization. *Appl. Environ. Microbiol.* **61**, 2346–2350 (1995).

20. Blazejak, A., Erseus, C. & Amann Rudolf & Dubilier, N. Coexistence of bacterial sulfide oxidizers, sulfate reducers, and spirochetes in a gutless worm (Oligochaeta) from the Peru margin. *Appl. Environ. Microbiol.* **71**, 1553–1561 (2005).

21. Ruehland, C., Blazejak, A., Lott, C., Loy, A. & Erséus Christer & Dubilier, N. Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. *Environ. Microbiol.* **10**, 3404–3416 (2008).

22. Blazejak, A., Kuever, J., Erséus, C., Amann, R. & Dubilier, N. Phylogeny of 16S rRNA, Ribulose 1,5-Bisphosphate Carboxylase/Oxygenase, and Adenosine 5′-Phosphosulfate Reductase Genes from Gamma- and Alphaproteobacterial Symbionts in Gutless Marine Worms (Oligochaeta) from Bermuda and the Bahamas. *Appl. Environ. Microbiol.* **72**, 5527–5536 (2006).

23. Zimmermann, J. *et al.* Closely coupled evolutionary history of ecto- and endosymbionts from two distantly related animal phyla. *Mol. Ecol.* **25**, 3203–3223 (2016).

24. Sato, Y. *et al.* Fidelity varies in the symbiosis between a gutless marine worm and its microbial consortium. *bioRxiv* 2021.01.30.428904 (2021) doi:10.1101/2021.01.30.428904.

25. Cong, Q. *et al.* Complete genomes of Hairstreak butterflies, their speciation and nucleo-mitochondrial incongruence. *Sci. Rep.* **6**, 24863 (2016).

26. Sota, T. & Vogler, A. P. Incongruence of Mitochondrial and Nuclear Gene Trees in the Carabid Beetles Ohomopterus. *Syst. Biol.* **50**, 39–59 (2001).

27. Phillips, M. J., Haouchar, D., Pratt, R. C., Gibb, G. C. & Bunce, M. Inferring Kangaroo Phylogeny from Incongruent Nuclear and Mitochondrial Genes. *PLoS One* **8**, e57745 (2013).

28. Ting, N., Tosi, A. J., Li, Y., Zhang, Y.-P. & Disotell, T. R. Phylogenetic incongruence between nuclear and mitochondrial markers in the Asian colobines and the evolution of the langurs and leaf monkeys. *Mol. Phylogenet. Evol.* **46**, 466–474 (2008).

29. Nakabachi, A. *et al.* Transcriptome analysis of the aphid bacteriocyte, the symbiotic host cell that harbors an endocellular mutualistic bacterium, *Buchnera*. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5477 LP – 5482 (2005).

30. Login, F. H. *et al.* Antimicrobial Peptides Keep Insect Endosymbionts Under Control. *Science (80-.).* **334**, 362 LP – 365 (2011).

31. McFall-Ngai, M. Care for the community. *Nature* **445**, 153 (2007).

32. Anselme, C. *et al.* Identification of the Weevil immune genes and their expression in the bacteriome tissue. *BMC Biol.* **6**, 43 (2008).

33. Franzenburg, S. *et al.* Distinct antimicrobial peptide expression determines host species-specific bacterial associations. *Proc. Natl. Acad. Sci.* **110**, E3730 LP-E3738 (2013).

34. Uroz, S., Courty, P. E. & Oger, P. Plant Symbionts Are Engineers of the Plant-Associated Microbiome. *Trends Plant Sci.* **24**, 905–916 (2019).

35. Fisher, R. M., Henry, L. M., Cornwallis, C. K., Kiers, E. T. & West, S. A. The evolution of host-symbiont dependence. *Nat. Commun.* **8**, 15973 (2017).

36. Bennett, G. M. & Moran, N. A. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. *Proc. Natl. Acad. Sci.* **112**, 10169 LP – 10176 (2015).

37. Moran, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci.* **93**, 2873 LP – 2878 (1996).

38. Rispe, C. & Moran, N. A. Accumulation of Deleterious Mutations in Endosymbionts: Muller's Ratchet with Two Levels of Selection. *Am. Nat.* **156**, 425–441 (2000).

39. Hosokawa, T., Kikuchi, Y., Nikoh, N., Shimada, M. & Fukatsu, T. Strict Host-Symbiont Cospeciation and Reductive Genome Evolution in Insect Gut Bacteria. *PLOS Biol.* **4**, e337 (2006).

40. McCutcheon, J. P. & Moran, N. A. Functional Convergence in Reduced Genomes of Bacterial Symbionts Spanning 200 My of Evolution. *Genome Biol. Evol.* **2**, 708–718 (2010).

41. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2012).

42. Sloan, D. B. & Moran, N. A. Genome Reduction and Co-evolution between the Primary and Secondary Bacterial Symbionts of Psyllids. *Mol. Biol. Evol.* **29**, 3781–3792 (2012).

43. Jäckle, O. *et al.* Chemosynthetic symbiont with a drastically reduced genome serves as primary energy storage in the marine flatworm *Paracatenula*. *Proc. Natl. Acad. Sci.* **116**, 8505 LP – 8514 (2019).

44. McCutcheon, J. P., Boyd, B. M. & Dale, C. The Life of an Insect Endosymbiont from the Cradle to the Grave. *Curr. Biol.* **29**, R485–R495 (2019).

45. Russell, S. L. *et al.* Horizontal transmission and recombination maintain forever young bacterial symbiont genomes. *PLOS Genet.* **16**, e1008935 (2020).

46. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590--6 (2013).

47. Yilmaz, P. *et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).

48. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

49. Gruber-Vodicka, H. R., Seah, B. K. B. & Pruesse, E. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. *mSystems* **5**, e00920-20 (2020).

50. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

51. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

52. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

53. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

54. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407–415 (2004).

55. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* **61**, 539–542 (2012).

56. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

57. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).

58. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

59. Pruesse, E. & Peplies Jörg & Glöckner, F. O. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. 1823–1829 (2012).

60. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).

61. Miller, C. S., Baker, B. J., Thomas, B. C. & Singer Steven W. & Banfield, J. F. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol.* **12**, R44 (2011).

62. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).

63. Altschul, S. F., Gish, W., Miller, W. & Myers E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

64. Anderson, F. E. *et al.* Phylogenomic analyses of *Crassiclitellata* support major Northern and Southern Hemisphere clades and a Pangaean origin for earthworms. *BMC Evol. Biol.* **17**, 123 (2017).

65. Verdes, A. *et al.* Molecular phylogeny of *Odontosyllis*; (Annelida, Syllidae): A recent and rapid radiation of marine bioluminescent worms. *bioRxiv* 241570 (2018) doi:10.1101/241570.

66. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014).

67. Hugoson, E., Ammunét, T. & Guy, L. Host-adaptation in *Legionellales* is 2.4 Ga, coincident with eukaryogenesis. *bioRxiv* 852004 (2020) doi:10.1101/852004.

68. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

69. Bartram, A. K. *et al.* Exploring links between pH and bacterial community composition in soils from the Craibstone Experimental Farm. *FEMS Microbiol. Ecol.* **87**, 403–415 (2014).

70. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

71. Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. *J. Stat. Softw.* **40**, (2011).

72. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, 2002).

73. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

74. Mantel, N. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Res.* **27**, 209 LP – 220 (1967).

75. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26**, 32–46 (2001).

76. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).

77. Bogdanowicz, D., Giaro, K. & Wróbel, B. TreeCmp: Comparison of Trees in Polynomial Time. *Evol. Bioinforma.* **8**, EBO.S9657 (2012).

78. Bogdanowicz, D. & Giaro, K. On a matching distance between rooted phylogenetic trees. *Int. J. Appl. Math. Comput. Sci.* **23**, 669–684.

79. Brooks, A. W., Kohl, K. D., Brucker, R. M., van Opstal, E. J. & Bordenstein, S. R. Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History. *PLOS Biol.* **14**, e2000225 (2016).