



RESEARCH

Open Access



Comparative genomics analyses of Actinobacteriota identify Golgi phosphoprotein 3 (GPP34) as a widespread ancient protein family associated with sponge symbiosis

Cláudia Ferreira¹, Ilia Burgsdorf¹, Tzipora Perez¹, Gustavo Ramírez^{1,2}, Maya Lalarz³, Dorothée Huchon^{4,5}  and Laura Steindler^{1*} 

Abstract

Background Sponges harbor microbial communities that play crucial roles in host health and ecology. However, the genetic adaptations that enable these symbiotic microorganisms to thrive within the sponge environment are still being elucidated. To understand these genetic adaptations, we conducted a comparative genomics analysis on 350 genomes of Actinobacteriota, a phylum commonly associated with sponges.

Results Our analysis uncovered several differences between symbiotic and free-living bacteria, including an increased abundance of genes encoding prokaryotic defense systems (PDSs) and eukaryotic-like proteins (ELPs) in symbionts. Furthermore, we identified GPP34 as a novel symbiosis-related gene family, found in two symbiotic Actinobacteriota clades, but not in their closely related free-living relatives. Analyses of a broader set of microbes showed that members of the GPP34 family are also found in sponge symbionts across 16 additional bacterial phyla. While GPP34 proteins were thought to be restricted to eukaryotes, our phylogenetic analysis shows that the GPP34 domain is found in all three domains of life, suggesting its ancient origin. We also show that the GPP34 family includes genes with two main structures: a short form that includes only the GPP34 domain and a long form that encompasses a GPP34 domain coupled with a cytochrome P450 domain, which is exclusive to sponge symbiotic bacteria.

Conclusions Given previous studies showing that GPP34 is a phosphatidylinositol-4-phosphate (PI4P)-binding protein in eukaryotes and that other PI4P-binding proteins from bacterial pathogens can interfere with phagolysosome maturation, we propose that symbionts employ GPP34 to modulate phagocytosis to colonize and persist within sponge hosts.

Keywords Symbiosis, Holobiont, Comparative genomics, GPP34, Host-symbiont interactions, Eukaryotic like, Actinobacteriota, Phosphatidylinositol 4-phosphate, Phagocytosis, Sponge microbiome

*Correspondence:

Laura Steindler

lsteindler@univ.haifa.ac.il

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Introduction

Microorganisms thrive in the most diverse environments, encompassing both free-living and host-associated lifestyles. To adapt to new ecological niches, microorganisms rely on several molecular mechanisms, including horizontal gene transfer (HGT) and recombination [1]. Adaptations, such as those related to the association with an animal host, leave in microbial symbionts genomic signatures that are absent in the genomes of taxonomically closely related free-living species [2]. Therefore, by analyzing the genomes of symbiotic microorganisms and comparing them with their close free-living counterparts, one can unravel the genetic changes associated with the shift from free-living to host-associated lifestyles.

Sponges (phylum Porifera) harbor exceptionally diverse, yet specific, microbial communities and provide an opportunity to investigate some of the earliest animal-microbe symbioses that have persisted to this day [3, 4]. The sponge-associated microorganisms support host health by contributing to its nutrition [5–7], processing toxic metabolic waste [8], and producing secondary metabolites that act as chemical defenses [9]. While previous comparative genomic studies of free-living versus sponge-symbiotic microorganisms have elucidated some of the genetic basis underlying the distinct fates of these microbes within sponges [10–18], the mechanisms that enable symbiotic microorganisms to survive within sponges are still largely unknown.

Metagenomic studies focusing on sponge microbial communities have revealed a significant enrichment of genes encoding eukaryotic-like proteins (ELPs) in symbiotic microorganisms across diverse taxonomic groups [11, 19–22], thought to have been acquired by prokaryotes through HGT [20, 23, 24]. Among the prominent types of ELPs discovered in sponge symbionts are ankyrin-rich repeat (ANK), leucine-rich repeat (LRR), and tetratricopeptide repeat (TPR) proteins, all of which are involved in protein–protein interactions [11, 13–15]. When ANK proteins encoded by genes from bacterial sponge symbionts are heterologously expressed in *Escherichia coli*, they interfere with the phagocytic pathway of the amoeba *Acanthamoeba castellanii*, suggesting that in sponge symbionts these ANK proteins may also be involved in phagocytosis evasion [25]. Importantly, ELPs are not exclusive to sponge symbionts and have been identified in bacteria interacting with other hosts. For example, ANK proteins from the pathogen *Legionella pneumophila* can interfere with phagosome maturation, thereby protecting it from the host's immune system [23, 26, 27]. These findings suggest that ELPs may facilitate a host-associated lifestyle in both pathogenic and symbiotic bacteria. Nevertheless, the enrichment of ANK proteins and other ELPs in sponge symbionts is not

uniform. A comprehensive genomic study, encompassing 780 genomes of sponge symbionts and 85 genomes of free-living bacteria, showed that the abundance and distribution of ELPs are not homogenous across different taxonomic phyla, with Actinobacteriota displaying a lower enrichment of ELPs compared to Poribacteria, Latescibacterota, and Acidobacteriota [28]. Actinobacteriota are among the most abundant symbionts found in sponges [3], and their lower enrichment of ELPs suggests that these symbionts may employ distinct, still undiscovered, molecular strategies to maintain stable associations with sponges [28].

In recent years, a substantial number of metagenomically assembled genomes (MAGs) of sponge-associated Actinobacteriota has become available through public databases. We employed these MAGs along with additional genomes of free-living Actinobacteriota to conduct a wide phylogenomic analysis and determine the relationships among symbiotic and free-living lineages. Focusing on two distinct monophyletic clades which both included a sponge-symbiotic lineage sister to a free-living lineage (the TK06 and UBA11606 clades), we reveal that members of GPP34 protein family (also known as Golgi phosphoprotein 3 and GOLPH3) are present in most members of the sponge symbiont lineages while lacking in their free-living counterparts. Phylogenetic analyses of GPP34 proteins suggest that this is an ancient protein family shared by all domains of life. Interestingly, members of this family that contained additional functional domains (e.g., cytochrome P450) appear to be restricted to sponge symbiotic taxa. We propose that sponge bacterial symbionts use proteins of the GPP34 family to modulate phagocytosis and escape degradation by host cells.

Materials and methods

Phylogenomic analysis

Actinobacteriota genomes were obtained from public databases until September 2021. Specifically, 160 Actinobacteriota MAGs associated with 8 sponge species from the class Demospongiae (*Aplysina aerophoba*, *Ircinia ramosa*, *Ircinia variabilis*, *Petrosia ficiformis*, *Theonella swinhoei*, *Carteriospongia foliascens*, *Coscinoderma matthewsi*, and *Rhopaloeides odorabile*) from different locations (Australia: Great Barrier Reef and Davies Reef; Israel: Achziv Nature Marine Reserve, Mediterranean Sea and Gulf of Aqaba, Red Sea; Croatia: Gulf of Piran, Adriatic Sea; Slovenia: Marine Biology Station Piran, Mediterranean Sea) were obtained from previous studies [5, 22, 28, 29]. Thirteen additional Actinobacteriota genomes, which were derived from bacteria isolated in culture from sponges, were also added to the analysis [30–41]. Reference genomes from non-sponge-associated Actinobacteriota were selected based on taxonomic

proximity. First, the taxonomy of the sponge-associated Actinobacteriota was determined using GTDB-Tk v1.3.0 (classify_wf) with release r95 [42]. Next, the taxonomically closest reference genomes were obtained and added to the analysis. In total, 350 Actinobacteriota genomes were analyzed (accession numbers listed in Data Set S1), including 81 which are cultured isolates, as reported on GTDB and NCBI [43, 44].

The phylogenomic tree was constructed using the concatenated alignment of 3021 amino acid sites. The alignment was generated with GTDB-Tk v2.4.0 (identify and align) and release r220 [43] and was trimmed using trimAl v1.5.rev0 (-gt 0.9 -cons 60) (<https://doi.org/https://doi.org/10.1093/bioinformatics/btp348>). Maximum-likelihood trees were inferred using RAxML v8.2.12 (<https://doi.org/https://doi.org/10.1093/bioinformatics/btu033>) with the PROTGAMMALG model of evolution and 1000 bootstrap replications. The tree was visualized and edited using iTOL [45] and Adobe Illustrator v27.5.

Functional and statistical analysis of Actinobacteriota genomes

Based on the phylogenomic Actinobacteriota tree, we selected two monophyletic groups that included both symbiotic and free-living bacteria for comparative genomics: family TK06 and genus UBA11606 (Fig. 1, Fig. S1, Data Set S1). For functional annotation, the protein sequences were searched against the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO) database using standalone KofamKOALA 1.3.0 [46]. To compare functional profiles of the here analyzed genomes, we used R version v4.1.1, in RStudio as a platform for statistical analysis. Specifically, we performed a principal coordinates analysis (PCoA) where Bray–Curtis dissimilarities were calculated based on the KOs using the vegan package [47] implemented in phyloseq [48]. To explore the statistically significant differences in the abundance of KOs in the subgroups (free living vs. symbionts) of the genomes studied, a differential abundance analysis was performed using the R package DESeq2 v1.36.0 [49] implemented in phyloseq v1.40.0 [48]. Specifically, for each KO, count data were fitted to a negative binomial distribution model, which allows coefficients (log₂-fold counts) and standard error estimates for each sample (genome). A Wald test, using maximum likelihood estimates of our KO model coefficients, was used to identify statistically significant (*P*-value < 0.05) differences between genome groups (free living vs. symbionts).

From our two groups comparisons (family TK06 and genus UBA11606), K15620 (Pfam PF05719), a gene annotated as Golgi phosphoprotein 3 (GPP34, GOLPH3), was selected for a more detailed analysis.

Functional and statistical analysis on prokaryotic defense systems (PDSs) and ELPs for family TK06 and genus UBA11606

Genes related to the PDS, selected based on previous studies (see the “Discussion” section), under the categories CRISPR-Cas, DND, R-M, and TA systems, were identified using KEGG annotation with KofamKOALA [46] (Data Set S1). The raw count for each category (specific gene family) was calculated, for each genome, by summing all the family-related genes. Relative abundances were calculated by dividing the raw counts of total PDS-coding genes by the total number of KEGG-annotated proteins in the genome (Data Set S1).

Eukaryotic-like domains (ELDs) were identified using Pfam v35 [50] implemented in InterProScan v5.16–55.0 [51], followed by selection of 11 ELD class types that were previously found enriched in sponge symbionts (see the “Discussion”): ANK: PF13606, PF13637, PF13857, PF12796, and PF00023; cadherin domain proteins (CAD: PF00028, PF12733, PF17803, PF17892); LRR: PF00560, PF12799, and PF13855; TPR: PF13371, PF07721, PF13432, PF13414, and PF09976; WD40 domain proteins (PF00400, PF07676); ncl-1, HT2A, and lin-41 (NHL: PF01436); pyrrolo-quinoline quinone domain proteins (PQQ: PF01011); eukaryotic-type carbonic anhydrase (PF00194); fibronectin type III domain proteins (fn3: PF00041); Calx-beta motif (PF03160); and GPP34 (PF05719). Annotations were considered for *e*-values < 1e-5. Pfam annotations represent domains; thus, each protein can have multiple Pfam annotations. Raw count frequencies of ELPs for each category (specific gene family class type) were calculated for each genome according to the following: first, we started with the number of Pfam references per protein, for each genome (note that each ELD class is represented by 1–5 Pfam reference domains). Next, we summed the Pfam references under the ELD class per protein and binary transformed it (multiple domains belonging to the same ELD were counted as 1). Last, we summed all the proteins per genome under the same ELD class type, to obtain raw counts of ELPs per genome for each ELD class type. It should be noted that some proteins are counted twice since they contain two different ELD classes. Relative abundances were calculated by dividing the raw counts of total ELP-coding genes by the total number of Pfam-annotated proteins in the genome (Data Set S1).

Heatmaps representing the abundance of PDS-coding and ELP-coding genes were created based on the normalization of the relative abundance into ratios. Specifically, the relative abundance of coding genes per genome devoted to each gene category (each family of PDS and ELPs) was obtained for each genome by dividing the relative abundance of genes by the highest relative abundance

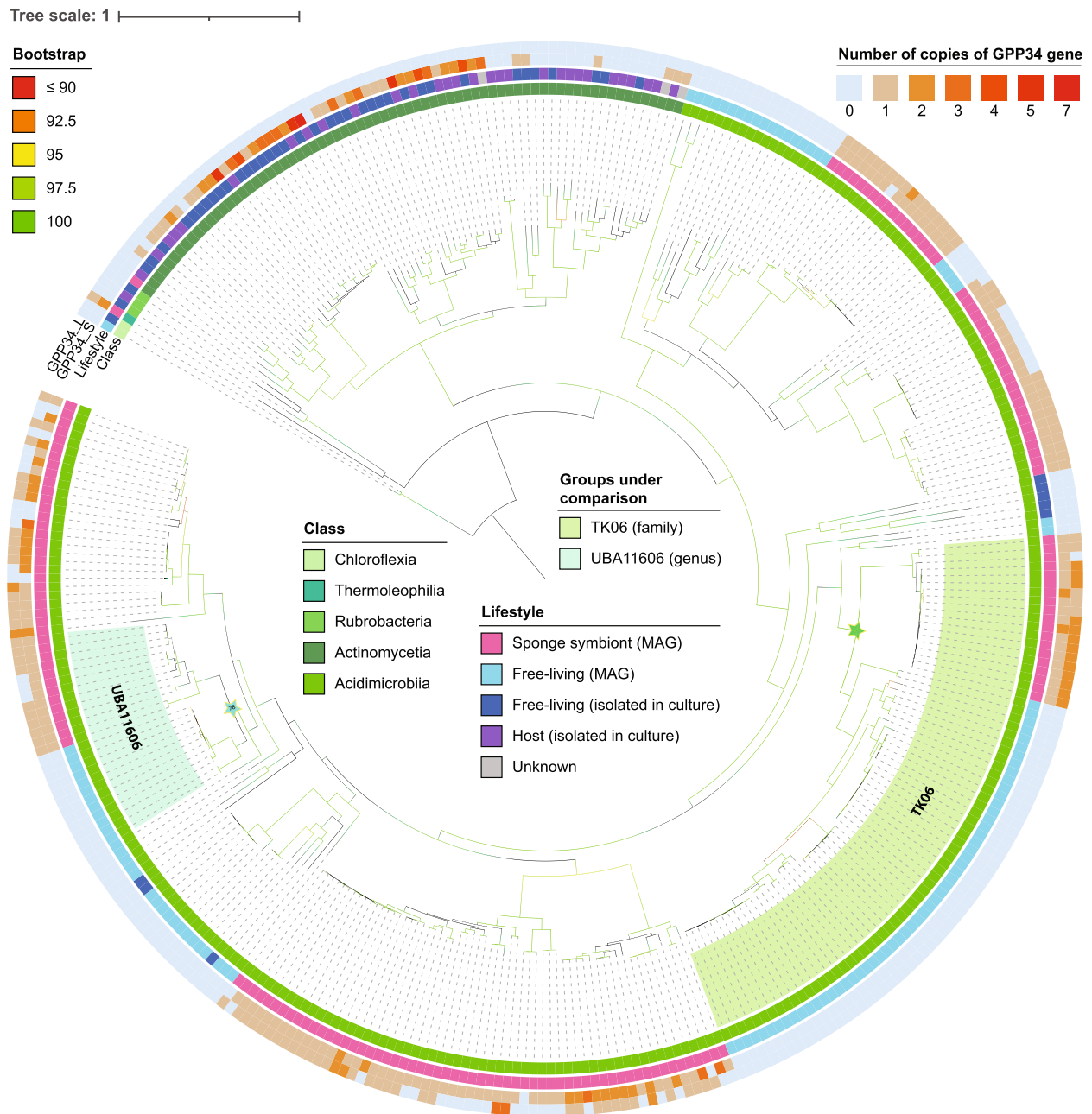


Fig. 1 Phylogenomic tree of Actinobacteriota phylum. Phylogenomic tree and distribution of Golgi-related protein (GPP34) among 350 Actinobacteriota genomes and two Chloroflexota genomes, which were used as an outgroup. The phylogenomic tree includes 160 sponge-symbiotic genomes (lifestyle: sponge symbiont, MAG) and 190 additional genomes of non-sponge symbionts (lifestyle: free living (MAG), free living (isolated in culture), host (isolated in culture), and unknown). Stars and background color indicate the groups which were used for the comparative genomic analysis, within family TK06 and genus UBA11606. Actinobacteriota classes are represented using colored strips within the inner circle followed by lifestyle indicators that are denoted by colored strips in the second circle. Additionally, the number of gene copies for GPP34_S and GPP34_L is indicated by colored strips in the outer circles. Bootstrap values above 90 are shown by colored branches. The bootstrap support for genus UBA11606 is shown as number (78). The same phylogenomic tree, including the names of all genomes, is presented in Fig. S1. GPP34, Golgi phosphoprotein 3; GPP34_S, short protein with single GPP34 domain; GPP34_L, long protein with two domains — GPP34 and cytP450. Raw phylogenetic tree and alignment can be found at <https://doi.org/10.6084/m9.figshare.27324744.v1>

under each category (for example, if GPP34 was present in one to three copies per genome, we divided the GPP34 counts per genome per the maximal copy number, which is three in this example).

The following statistical analyses were done for each gene family category under PDS and ELP groups: frequency distributions of differences in the proportion of coding genes derived from the free-living and symbiotic Actinobacteriota genomes were observed in random permutation tests (10,000 permutations) in R as described [16, 52]. *P*-values were estimated as the proportion of times that the permutation test produced a difference smaller, equal to or greater than the observed difference [16, 52].

GPP34 distribution across Bacteria and Archaea

We searched for the GPP34 domain in the genomes of the additional Actinobacteriota taxa (outside of the genus UBA11606 and family TK06) (Fig. 1, $n=251$) using functional annotation of GPP34 (K15620) against KO database using standalone KofamKOALA 1.3.0 [46]. In addition, we annotated all the genomes under study using Pfam v35 [50] implemented in InterProScan v5.16–55.0 [51] and filtered our results for proteins annotated as GPP34 (PF05719). This analysis served the purpose of identifying the domains of the proteins annotated as GPP34, dividing them into GPP34_S (short protein sequence, single domain: GPP34) and GPP34_L (long protein sequence, combination of two domains: GPP34 and cytP450). The same annotation methodologies were used to search for the presence of GPP34 proteins among 865 bacterial and 19 archaeal MAGs derived from a recently published dataset [28]. It should be noted that symbiont genomes used in the analysis were all derived from sponges of the class Demospongiae, which represents the majority of known sponges to date [53].

We added GPP34 annotation data (GPP34_S and GPP34_L) to ELPs that had been annotated in a previously published dataset [28]. Copy numbers of each annotated gene were corrected for the completeness of each genome by dividing each ELP count by the completeness of their genome. Nonmetric multidimensional scaling (NMDS) with Bray–Curtis dissimilarity was used to display the distribution of all the genomes ($n=865$, [28]) based on their ELP profile. This was done using the metaMDS function (package *vegan* v2.6–2 [47] and *ggplot2* v3.3.6 in R version v4.2.0. For graphically plotting all ELP frequency numbers in a circular heatmap, we further normalized the data by transforming it into ratios. Specifically, the ratio of coding genes per MAG devoted to each ELD class was obtained for each genome by dividing the number of gene copies per the highest number of gene copies under each ELD class. This was

done since the absolute counts varied greatly between each ELD class, and we wanted to plot all ELPs using the same heatmap range. For creating the phylogenomic tree with the circular heatmap showing the ELPs distribution across all genomes, we used *ggtree* package v3.4.4 [54] in R version v4.2.0.

The analyzed dataset from Robbins et al. [28] included a relatively small number of free-living bacteria. To investigate the distribution of the GPP34 gene between symbiotic and free-living bacteria, with broader representation of free-living representatives, we performed a focused analysis on three phyla commonly found as sponge symbionts: Chloroflexota, Proteobacteriota, and Acidobacteriota. First, to identify evolutionary relationships of the sponge symbionts with non-symbionts, we ran a phylum-specific phylogenomic analysis using GToTree software [55]. Specifically, GTDB representatives were fetched using the `gtt-get-accessions-from-GTDB` command. To ensure broad diversity coverage and avoid redundancy, we used the `-GTDB-representatives-only` argument. To track homologues of GPP34 gene across sampled genomes, we provided the associated Pfam accession (PF05719) to the software which implements a profile hidden Markov model search against the annotated features from all genomes. The analysis included the following: 4767 genomes of Chloroflexota, of which 473 were symbionts, 3175 genomes of Acidobacteriota, of which 163 symbionts, and 7635 genomes of Proteobacteriota, of which 154 symbionts. All phylogenetic trees were generated using GToTree command with GTDB accessions, symbiont genome paths, Pfam targets, and phylum-specific single gene copy markers as inputs. The analyses were performed in a Conda-enabled HPC environment with access to 128 processors and 1 TB of RAM. Resulting trees along with Pfam metadata were transferred to iTol for visualization. In iTol, we manually retained only clades containing sponge symbionts together with their closest free-living counterparts. Other clades were collapsed, and each genome was marked with color coding to indicate lifestyle (e.g., symbiont versus non-symbiont) and the presence/absence of GPP34. The final visualization displayed 916 Chloroflexota genomes, of which 473 were derived from sponge symbionts, 274 Acidobacteriota genomes, of which 163 derived from sponge symbionts, and 296 Proteobacteriota genomes, of which 154 derived from sponge symbionts.

GPP34 protein architecture and flanking genes

A representative for each protein architecture was drawn with IBS v1.0.3 [56], based on the conserved domains annotated using online DELTA-BLAST [57] against the NCBI database using default parameters (*e*-value cutoff of 0.05). The genes present in the flanking region (defined

as five genes upstream and five genes downstream) of GPP34 were annotated and aligned using DiGAlign server v1.0 (www.genome.jp/digalign/).

MAG-specific expression analysis

In the present study, we conducted a metatranscriptomic analysis using a previously published metatranscriptomics dataset [58] obtained from the sponge species *A. aerophoba*. A total of 107 genome assemblies, all deriving from *A. aerophoba*, underwent metatranscriptomic short-read mapping using Bowtie2 v2.4.5 (with default parameters) [59] and open reading frame annotation using *Prokka* [60]. The list of MAGs used in this analysis, with accession numbers and references, is available in the supplemental file Data Set 1. Count reads per feature were generated using *htseq-count* [61]. Both *recA* and GPP34 gene homolog identification were performed using BLASTp against all predicted proteins from each genome. GPP34 feature counts of interest are reported as *recA* normalized transcripts per million (TPM) values.

Of these analyzed 107 MAGs, 88 have *recA*, of which 60 are transcribed (>1 TPM), and 49 of those have at least 1 GPP34 annotated gene. For each genome with a transcribed *recA*, GPP34 TPM counts were retrieved, counting a total of 45 genomes (36 above 85% completeness and 9 under 85% completeness) that have *TPMs* > 1 for at least 1 GPP34. The gene *recA* was absent in genomes containing GPP34-PotA combined domain proteins; thus, it was not possible to analyze the expression of these proteins.

Phylogeny of GPP34 protein

The GPP34 protein sequences ($n=317$) used for the phylogenetic analysis were selected to include the following: (i) representatives of both short and long GPP34 protein sequences derived from sponge symbionts described in [28] (annotated by Pfam v35 [50] implemented in InterProScan v5.16–55.0 [51], across 16 bacterial phyla ($n=103$), (ii) representatives from Actinobacteriota that were isolated from sponge samples ($n=13$), (iii) representatives of GPP34 protein sequences derived from non-sponge symbionts of 10 different bacterial phyla ($n=55$), (iv) 97 protein sequences representative of animal diversity and 18 representative of fungal diversity (genome assemblies with highest assembly level were selected), (v) 18 protein sequences representative of non-fungi/non-animal eukaryotes, and (vi) 13 archaeal sequences.

To identify protist sequences, BLASTp searches were conducted using all protist GPP34 sequences annotated in UniProt as queries against the RefSeq database. The search was limited to Eukarya (taxid:2759), excluding Fungi (taxid:4751) and Metazoa (taxid:33208). From the BLAST hits obtained, we selected proteins that (i)

originated from genome assemblies sequenced at, or above, the scaffold level, (ii) contained a GPP34 domain (pfam05719) based on an InterProScan domain search, and (iii) were longer than 130 amino acids. Notably, plant sequences that clustered within fungi were considered contaminants and excluded.

Similar criteria were used for archaeal sequences. BLASTp searches were performed using all archaea sequences from UniProt with a PF05719 domain as queries against the NCBI nr protein database. The search was limited to Archaea (taxid:2157). We selected sequences that (i) were annotated as archaea in the NCBI genome database (GPP34 sequences originating from marine sediment metagenomes were excluded); (ii) originated from genome assemblies sequenced at, or above, the scaffold level; (iii) were longer than 130 amino acids; and (iv) contained a GPP34 domain (PF05719). Only sequences from Asgardarchaeota met these criteria, although our search was not restricted to this lineage. To select GPP34 sequences from non-sponge symbiotic bacteria, we conducted a comprehensive search using InterProScan. We first identified GPP34 sequences within each of the phyla represented by sponge symbionts in our dataset. Out of 16 phyla screened, GPP34 sequences were identified in 10. We then downloaded the sequences for all hits from these 10 phyla and performed BLAST searches against the NCBI database. If the top BLAST hit indicated that the sequences came from a nonsymbiotic bacterium, it was included in our dataset for further analysis.

Phylogenetic analysis proceeded as follows: the sequence alignment of the GPP34 domain of each sequence was obtained using the function *hmmalign* of HMMER 3.4 [62] with the PF05719.hmm profile file downloaded from InterPro (on September 24, 2024) and with the *-trim* option. The *-trim* option was used to remove, from the alignment, the terminal tails of residues that did not belong to the Pfam domain. All other positions were retained in the alignment. The obtained alignment included 472 amino-acid sites across 317 sequences. The phylogenetic tree of the GPP34 domain was reconstructed using IQ-TREE 1.6.12 [63]. First, we used ModelFinder Plus (MFP) to identify the best evolutionary model. Several complex models were added to the model search (i.e., option *-madd* C10, C20, C30, C40, C50, C60, LG4M, LG4X, CF4, EX2, EX3, EHO, UL2, UL3, EX_EHO, C10+G4, C20+G4, C30+G4, C40+G4, C50+G4, C60+G4, LG4M+G4, LG4X+G4, CF4+G4, EX2+G4, EX3+G4, EHO+G4, UL2+G4, UL3+G4, EX_EHO+G4, LG+C10, LG+C20, LG+C30, LG+C40, LG+C50, LG+C60, LG+LG4M, LG+LG4X, LG+CF4, LG+C10+G4, LG+C20+G4, LG+C30+G4, LG+C40+G4, LG+C50+G4, LG+C60+G4, LG+LG4M+G4, LG+LG4X+G4, LG+CF4+G4). The best model selected under the BIC criterion was the EX_EHO+G4 model. Tree reconstruction was performed using the best model with the *-mwopt* option, and branch

support was determined with 1000 ultrafast bootstrap replicates (i.e., option `-bb 1000`). Furthermore, branch supports were also computed under the EX_EHO+G4 model using 100 nonparametric bootstrap replicates (i.e., option `-b 100`).

Results

Phylogenomics of Actinobacteriota

Our phylogenomic analysis of Actinobacteriota comprised 160 MAGs of uncultured symbionts derived from sponges (class Demospongiae), 13 genomes from cultured bacteria isolated from sponges (class Demospongiae), 18 genomes from bacteria isolated from non-sponge hosts, and 159 genomes from either free-living Actinobacteriota or from Actinobacteriota of unknown lifestyle. The majority of symbionts are taxonomically affiliated to the class Acidimicrobiia, and they were identified in at least eight different sponge species from seven geographic locations (Fig. 1, Fig. S1).

This phylogenomic analysis revealed two clades, which both included a symbiotic lineage sister to a free-living lineage. The first clade belongs to the family Poriferisodallaceae fam. nov. (replacing the placeholder name TK06) with maximal bootstrap support, and the second clade is classified under the genus UBA11606 from the family Aldehydirespiratoraceae fam. nov. (replacing the placeholder name UBA11606) [64], with bootstrap support of 78. Hereafter, we use the nomenclature family TK06 and genus UBA11606, according to the taxonomy taking place at the time of the analysis [43]. We assume that the free-living lineages represent the ancestral lifestyle of the symbiotic clades.

Symbionts and free-living Actinobacteriota are functionally divergent

Symbiotic and free-living representatives of the family TK06 and the genus UBA11606 cluster into four separate clades based on the PCoA of the KO terms present in the genes encoded in their genomes. Symbionts from the two different taxonomic groups had more similar profiles than the two free-living groups (Fig. 2). To gain insights into the factors contributing to similar functional profiles in symbionts, we conducted a differential abundance analysis of the KOs in either symbionts or free-living bacteria. The analysis revealed 49 and 242 differentially distributed KOs between sponge symbionts and free-living members of the genus UBA11606 and the family TK06, respectively (Fig. S2 and Data Set S1) (DESeq2, Wald test followed by Benjamini–Hochberg multiple-inference correction, p -value < 0.05).

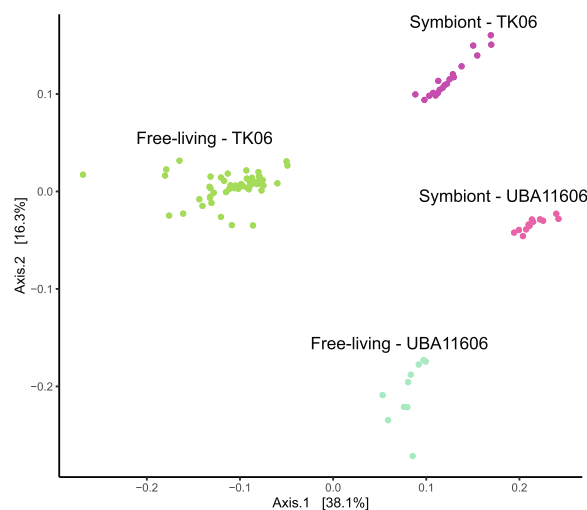


Fig. 2 Functional comparison of genomes of symbiotic and free-living members of family TK06 and genus UBA11606. In this PCoA with Bray–Curtis dissimilarity, the dots represent the genomes, which cluster based on their KO composition. The percentage on the axis represents dissimilarity. Groups are represented by colors: free living in blue (UBA11606) and green (TK06) and symbionts in pink (UBA11606) and purple (TK06)

Symbiotic Actinobacteriota are enriched in PDSs and ELPs

One KEGG category that is particularly enriched in genomes of symbiotic Actinobacteriota compared to free-living relatives (in both family TK06 and genus UBA11606) is PDS, which includes genes involved in clustered regularly interspaced short palindromic repeats (CRISPR)-Cas systems, DNA phosphothiolation (DND) systems, restriction-modification (R-M) systems, and toxin-antitoxin (TA) system (Fig. S3 and Fig. S4A).

Sponge symbionts are also enriched in several ELPs (Fig. 3), with symbiotic genomes harboring a greater number of ELPs per genome (Table S1 and Data Set S1). Compared to free-living relatives, symbionts possess a higher number of ELDs per protein (Table S1 and Data Set S1), in either a single- or mixed-domain architecture (Fig. S4B).

A random permutation analysis confirmed the significant enrichment of genes with ELDs from the CAD, LRR, fn3, Calx-beta classes, and in GPP34 in symbiotic Actinobacteriota compared to their free-living counterparts, in both the genus UBA11606 and the family TK06 ($p < 0.0001$).

However, not all ELP-encoding genes were enriched in symbionts across both clades. For example, TPR genes were enriched in symbiotic genomes from the genus UBA11606 ($p = 0.007$), whereas in the family TK06 they were enriched in the free-living genomes ($p < 0.0001$). WD40 genes showed significant enrichment exclusively in symbiont genomes of the family TK06 ($p < 0.0001$).

ANK and NHL genes did not display enrichment in either the symbiotic or free-living lifestyles, both for the genus UBA11606 (ANK, $p=0.473$; NHL, $p=0.223$) and the family TK06 (ANK, $p=0.401$; NHL, $p=0.286$) (Fig. S4C).

GPP34 distribution across genomes of the Actinobacteriota phylum

One notable finding was the differential distribution of a gene coding for a protein annotated as GPP34, belonging to a eukaryotic protein family. Within the selected Actinobacteriota taxonomic groups (the family TK06 and the genus UBA11606), the gene coding for GPP34 is consistently present in all symbiotic genomes ($n=19$ and 13 , respectively) while being absent in all analyzed free-living counterparts ($n=55$ and 12 , respectively).

To assess whether this difference between lifestyles is a general characteristic of Actinobacteriota, we conducted a comprehensive search for GPP34-coding genes across all 350 analyzed Actinobacteriota genomes. This resulted in the identification of a total of 422 GPP34 proteins. Among the analyzed genomes, 153 out of 160 sponge symbiotic genomes displayed 1 to 4 copies of the GPP34 protein. Within the other 31 host-derived genomes (isolated from various host organisms), 15 genomes exhibited 1 to 7 copies of the GPP34 gene. These hosts encompassed a range of organisms, including sponges (*A. aerophoba*, *Acanthostrongylophora* sp., *Sphaciospongia confederata*, *Sphaciospongia vagabunda*, *Lissodendoryx nobilis*, *Phakellia ventilabrum*, and marine sponge SP-1), a coral (Gorgonacea), a mangrove root (*Avicennia marina*), plants (*Solanum tuberosum* and *Citrus reticulata*), and rhizospheric soil (*Colocasia esculenta*, wild tea plants, and vegetable garden soil) (Data Set S1). Out of the 47 genomes derived from Actinobacteriota that were isolated in culture, 33 exhibited 1 to 5 copies of the GPP34_S gene. The 3 Actinobacteriota genomes of unknown lifestyle had 1-3 copies of GPP34_S gene. Lastly, the GPP34 gene was absent in all 109 free-living Actinobacteriota MAGs (Fig. 1). These results support the differential distribution of GPP34 genes among symbiotic and free-living bacteria.

GPP34 distribution beyond the phylum Actinobacteriota

Previous research has shown that ELPs are not uniformly distributed across different phyla of sponge symbiotic bacteria [28]. Therefore, we determined whether the newly discovered GPP34 symbiotic feature was unique to Actinobacteriota or present across other phyla. For this, we reanalyzed a selection of genomes investigated in a previous study [28], including genomes from 19 additional bacterial phyla and an archaeal phylum (Crenarchaeota).

Remarkably, GPP34 was found across 16 of the 20 analyzed bacteria phyla (Table 1). No GPP34 was annotated in the archaeal symbionts from the Crenarchaeota present in this dataset. Among bacteria, GPP34 was annotated exclusively in genomes of symbiotic bacteria, while it was absent in all the analyzed taxonomically related free-living counterparts ($n=85$). It should be noted that certain phyla, such as Myxococcota, Verrucomicrobiota, and Marinisomatota, had either no or only one symbiotic representative in the dataset. Furthermore, out of 780 bacterial sponge symbiont genomes examined, 604 contained 1 to 6 copies of GPP34 proteins, resulting in a total of 1381 GPP34 proteins (Fig. 4, Fig. S5, and Fig. S6). Interestingly, Poribacteria ($n=24$), which were previously reported to be highly enriched in ELPs [28], had only one genome with a single GPP34 gene. By contrast, Latescibacterota ($n=27$) and Acidobacteriota ($n=76$) exhibited a high abundance of both ELPs and GPP34. In some phyla, such as Chloroflexota and Bacteroidota, the distribution of GPP34 was not equal across different classes (Table 1). It should be noted that symbiont genomes of the analyzed dataset derived all from sponges from the class Demospongiae.

Since the dataset used in Robbins et al. [28] included a relatively small number of genomes from free-living bacteria ($n=85$) compared to symbiont genomes ($n=780$), we expanded our analysis to increase the representation of free-living bacterial genomes within three phyla commonly associated with sponges — Chloroflexota, Acidobacteriota, and Proteobacteriota. We focused specifically on genomes of free-living bacteria that were closely related to sponge-symbiotic

(See figure on next page.)

Fig. 3 Distribution of ELPs from nine different classes of ELDs and GPP34 across Actinobacteriota genomes within genus UBA11606 and family TK06. Phylogenomic trees (genus UBA11606 on the left and family TK06 on the right) with the heatmaps (based on Pfam annotation) representing the ratio of ELP and GPP34 abundances, based on relative abundance correction where the counts of the ELP and GPP34 were divided by the total number of proteins annotated on that genome. These values were further normalized between categories (ELD classes and total GPP34), by dividing them per the highest number of annotated proteins in each category (ELD class or GPP34). The trees were generated by extraction of the genus UBA11606 and family TK06 from the larger phylogenomic tree (Fig. 1). ANK, ankyrin repeat; CAD, cadherin domain; LRR, leucine-rich repeats; TPR, tetratricopeptide repeat; WD40, beta-transducin repeat; NHL, (ncl-1, HT2A, and lin-41); Carb_anhydrase, carbonic anhydrase; fn3, fibronectin type III domain; Calx-beta, Calx-beta motif; GPP34, Golgi phosphoprotein 3, GPP34_T stands for GPP34_total and represents the sum of GPP34_S and GPP34_L

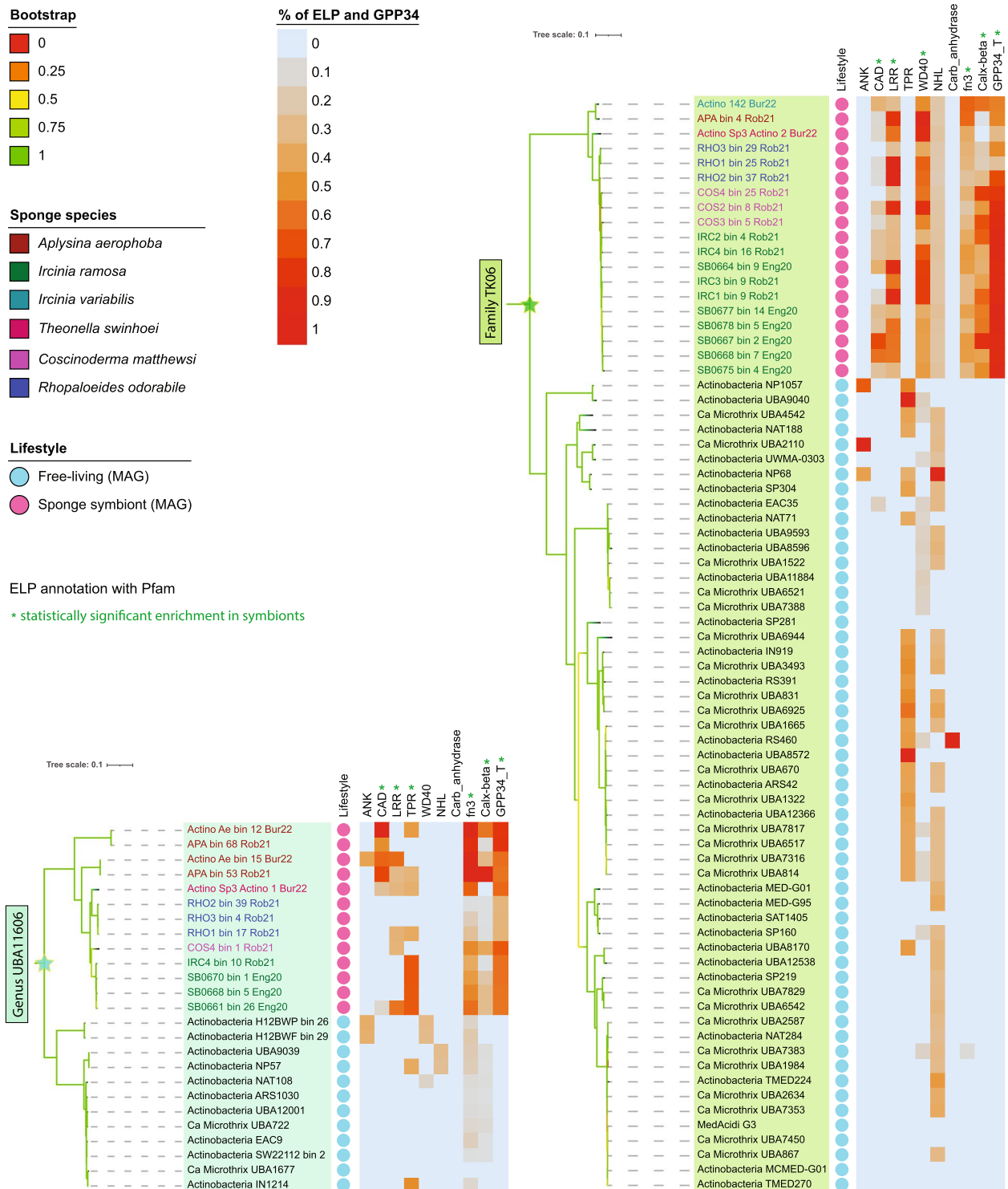


Fig. 3 (See legend on previous page.)

groups. In all three phyla, we observed that GPP34 was predominantly present in the genomes of symbionts (sponge and coral) and largely absent in their closely related counterparts (Figs. S7, S8, S9, Data Set 1).

GPP34 protein architecture

To gain further insights into GPP34, we analyzed the size and architecture of the various proteins identified in our searches. This revealed that GPP34 proteins

Table 1 Percentages of bacterial genomes with GPP34 across sponge symbiotic and free-living bacteria

Phylum *Class	Symbiotic genomes with GPP34 (%)			Free-living genomes with GPP34 (%)
	Total	GPP34_S	GPP34_L	
<i>UBP10</i> (n=10)	100	100	90	-
<i>Spirochaetota</i> (n=8)	100	100	62.5	-
<i>Nitrospinota</i> (n=4)	100	100	100	-
<i>Deinococcota</i> (n=3)	100	100	66.67	-
<i>Latescibacterota</i> (n=27)	100	96.30	92.59	-
<i>Gemmatimonadota</i> (n=54)	100	75.93	92.60	0 (n=3)
<i>Actinobacteriota</i> (n=117)	99.16	98.29	79.49	0 (n=3)
<i>Acidobacteriota</i> (n=76)	92.11	44.74	85.53	-
<i>Nitrospirota</i> (n=21)	85.71	0	85.71	-
<i>Bacteroidota</i> (n=34)	82.35	5.88	76.47	0 (n=13)
* <i>Rhodothermia</i> (n=27)	100	3.70	96.29	0 (n=1)
* <i>Bacteroidia</i> (n=7)	14.29	14.29	0	0 (n=12)
<i>Proteobacteria</i> (n=164)	82.32	74.39	58.54	0 (n=48)
* <i>Gammaproteobacteria</i> (n=81)	77.78	61.73	65.43	0 (n=13)
* <i>Alphaproteobacteria</i> (n=83)	86.75	86.75	51.81	0 (n=35)
<i>Dadabacteria</i> (n=16)	62.5	6.25	56.25	-
<i>Chloroflexota</i> (n=188)	62.23	54.79	53.13	-
* <i>UBA2235</i> (n=22)	100	100	81.82	-
* <i>Dehalococcoidia</i> (n=75)	97.33	93.33	80	-
* <i>Anaerolineae</i> (n=91)	24.18	12.09	21.98	-
<i>Bdellovibrionota</i> (n=3)	33.33	0	33.33	0 (n=1)
<i>Planctomycetota</i> (n=13)	15.38	15.38	0	0 (n=5)
<i>Poribacteria</i> (n=24)	4.17	4.17	0	-

Description of the number and percentage of symbiotic (n=780) and free-living (n=85) bacterial genomes with GPP34 gene among 20 phyla (and respective classes) using a previously published dataset for analysis [28]. Percentages reflect the most comprehensive annotation (Pfam) on genomes with >85% completeness (n=865). The percentages represent the proportion of genomes with GPP34 hits relative to the total number of symbiotic genomes (column 'Total'). The analysis was also performed separately for short and long GPP34 proteins, presented in columns 'GPP34_S' and 'GPP34_L', respectively. The same comparison was conducted for free-living genomes. Hyphen is used when no genome was analyzed. Phyla without GPP34 (not shown in the Table) are: *Cyanobacteriota* (symbionts: n=16; free-living: n=2), *Myxococcota* (symbionts: n=1; free-living: n=5), *Verrucomicrobiota* (symbionts: n=1; free-living: n=2), and *Marinisomatota* (free-living: n=3).

can be categorized into two major groups: short GPP34 (GPP34_S) and long GPP34 (GPP34_L) proteins (Fig. S10). Pfam annotation revealed that GPP34_S proteins contain a single GPP34 domain, while GPP34_L proteins consist of two domains, GPP34 combined with another domain (Fig. 5). Among GPP34_L proteins, the vast majority are characterized by the combination of the GPP34 and cytochrome P450 (cytP450, PF00067) domains (98.94%, n=656/663). Other domain combinations include GPP34 with the Rieske domain ([2Fe-2S] domain, PF00355), found in two genomes (*Proteobacteria* and *Nitrospinota*, 0.14%), and GPP34 combined with a PotA domain (ABC transporter, PF00005), identified in five *Chloroflexota* genomes (0.36%) (Fig. 5). Interestingly, all the proteins in which GPP34 is combined with other domains (GPP34_L) are derived from genomes of sponge symbionts.

In the *Actinobacteriota* genomes analyzed (Fig. 1), GPP34_S is the prevalent protein architecture, accounting for 67.77% of the total GPP34 proteins (n=286/422), whereas GPP34_L represents 32.23% (n=136/422). Among the 20 bacterial phyla represented in the wider dataset from Robbins et al. [28], the GPP34_S architecture is the most common, constituting 51.99% of the total GPP34 proteins (n=718/1381), while GPP34_L represents 47.50% (n=656/1381). Most phyla harboring GPP34 coding genes possess both GPP34_S and GPP34_L (n=12/16). However, some phyla exclusively have representatives of one of the two architectures: *Planctomycetota* and *Poribacteria* only contain GPP34_S, while *Nitrospirota* and *Bdellovibrionota* only have GPP34_L (Table 1).

When analyzing the GPP34_S flanking genes in the genomes of genus *UBA11606* and family *TK06*, we observed that in most of the analyzed genomes (8/13 in

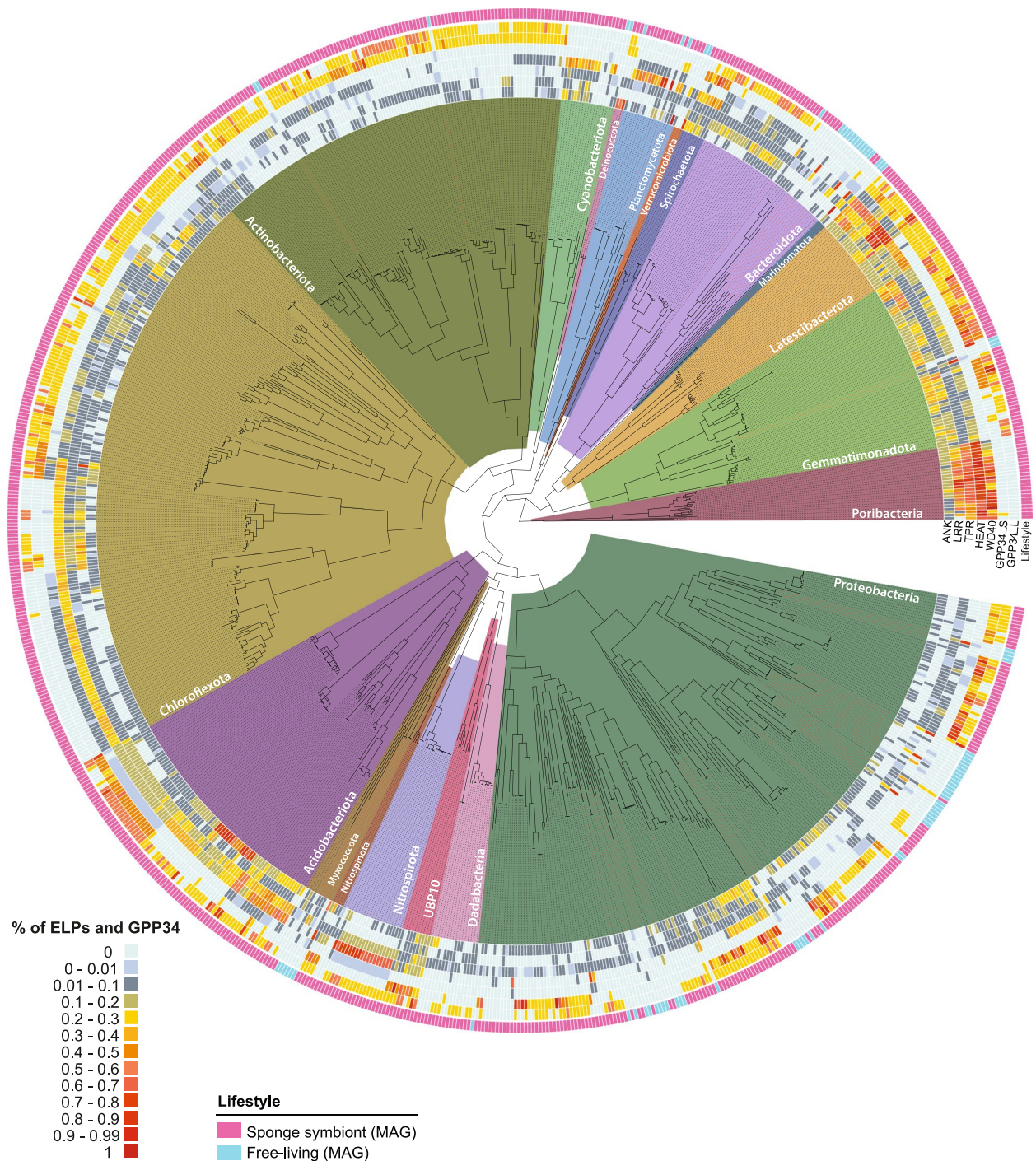


Fig. 4 Distribution of ELPs and GPP34 proteins across diverse Bacteria phyla. Phylogenomic tree based on the dataset of genomes used in Robbins et al. (2021) (Fig. 3 [28]), figuring 865 genomes (> 85% completeness), out of which 780 are symbiont MAGs (denoted by pink strips in external circle) and 85 are MAGs from seawater (denoted by light-blue strips in external circle). Values represent the percentage of coding genes per MAG devoted to each ELD class and to GPP34. The background to the tree branches is colored by phylum. The same phylogenomic tree, including the names of all genomes, is presented in Fig. S5. ANK, ankyrin repeat protein; LRR, leucine-rich repeats; TPR, tetratricopeptide repeat; HEAT, Huntington, elongation factor 3, PR65/A, and TOR; WD40, beta transducin repeat; GPP34, Golgi phosphoprotein 3 (GPP34_S, one domain GPP34; GPP34_L, GPP34 associated with another domain)

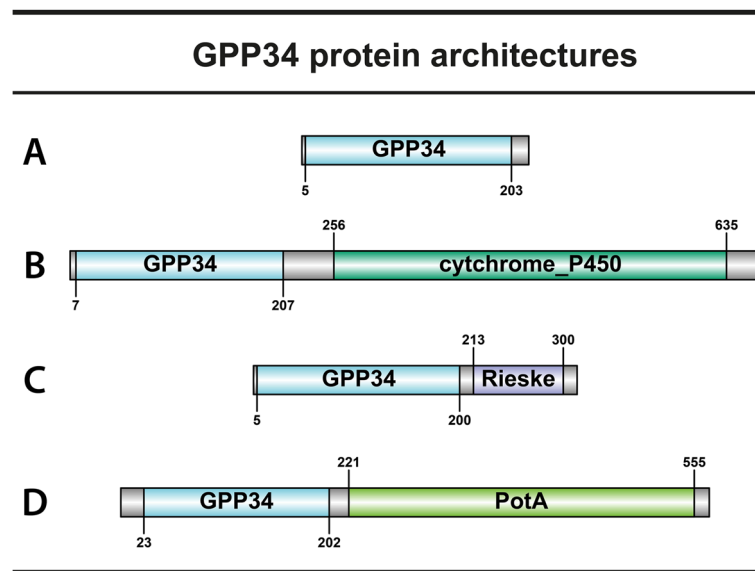


Fig. 5 Architecture of four different proteins containing the GPP34 domain, based in DELTA-BLAST. Proteins A and B belong to an acidimicrobial genome (Actino_Sp3_Actino1_Bur22) of a symbiont affiliated to the genus UBA11606, associated with the sponge species *T. swinhoei*. The architectures of protein A and B were categorized as GPP34_S (short protein sequence: one domain) and GPP34_L (long protein sequence: two domains), respectively. Further, protein C and D belong to two genomes (APA_bin_87_Rob21 and IRC_PAM_SB0661_bin_34_Rob21) of symbionts from the phyla Proteobacteriota and Chloroflexota, respectively. The first associated with the sponge species *A. aerophoba* and the second to *I. ramosa*. A, contig-70_45_length_50594_read-count_2396658_25; B, contig-70_71_length_41088_read-count_1976913_20; C, c_000000136236_5; D, NODE_379_length_81712_cov_64.429798_7

genus UBA11606 and 17/19 in family TK06), a gene coding for *cytP450* was detected within five genes up- or downstream of GPP34_S (Fig. S11).

GPP34 gene expression

To assess the expression of GPP34 coding genes by sponge symbionts in their natural environment, we analyzed a previously published metatranscriptomics dataset [58] obtained from the sponge *A. aerophoba*, along with 107 MAGs derived from the same sponge species. Our analysis revealed that GPP34 is actively expressed in 45 out of the 107 analyzed genomes. Representatives of both GPP34_S and GPP34_L architectures were found to be expressed. Furthermore, GPP34 was found to be expressed in 11 out of the 16 phyla that encode the gene (Fig. S12).

Evolutionary origin of the GPP34 domain

To gain deeper insights into the evolutionary history of GPP34, we performed a phylogenetic analysis that exclusively examined the GPP34 domain. This expanded beyond the GPP34 sequences identified in symbiotic bacterial genomes from the dataset used in Robbins et al. (2021) [28], incorporating GPP34 sequences from all domains of life — eukaryotes, archaea (with GPP34 only found in Asgardarchaeota), and bacteria. Among the bacterial sequences, we included those from both sponge symbionts and non-symbionts across diverse phyla (for a

detailed description of sequence selection, see the “[Materials and methods](#)”). Due to the large taxon sampling and the limited length of the alignment, the support values from nonparametric bootstraps were low.

Because GPP34 has been described as a eukaryotic domain, we expected domain sequences from bacterial origin to branch among eukaryotic sequences. However, the GPP34 domains of most eukaryotes formed a distinct monophyletic group (UF boot/NP boot support=99/34). The eukaryotic sequences that did not belong to this clade were all fast evolving and included Amoebozoa, Chlorophyta, Metamonada, and Ochrophyta sequences. The rapid evolution of these sequences, combined with the short length of the domain analyzed (~200 amino acids), likely explains why they did not cluster with other eukaryotes. Interestingly, most symbiotic GPP34 sequences, both short and long variants, also formed a monophyletic group (UF boot/NP boot support=99/26), suggesting a common ancestry for both symbiotic domains. These symbiotic GPP34 domains are distinct from domains present in the genomes of non-sponge symbionts yet more closely related to them than to eukaryotic GPP34 domains. This indicates that GPP34 proteins in sponge symbionts likely have a bacterial origin, rather than having been acquired recently through horizontal transfer from eukaryote hosts. Additionally, GPP34 sequences from archaea did not form a

monophyletic clade, suggesting possible horizontal gene transfers between archaea and bacteria (Fig. 6). Although the exact position of the root is unknown in this tree, the fact that both archaea and eukaryotes harbor GPP34_S sequences supports that this is the ancestral form of this protein. Additionally, proteins with a GPP34+cytP450 architecture form a monophyletic group (UF boot/NP boot support=100/45), indicating a single common origin for this protein form. Moreover, we confirmed that GPP34_L (i.e., proteins with a GPP34 domain and a cytP450, Rieske, or PotA domain) were found only in sponge symbionts. Within the genomes of sponge symbionts, GPP34 genes are found in varying copy numbers, ranging from 1 to 6, and these copies often belong to distinct and well-supported phylogenetic lineages suggesting that they did not originate from recent gene duplications (Fig. 6, Fig. S13, and Data Set S1).

Discussion

Symbiotic and free-living Acidimicrobia are functionally divergent

To investigate bacterial genetic features that are related to symbiosis, comparative genomics analysis needs to be done between taxonomically closely related symbionts and free-living counterparts [12, 14, 65]. Accordingly, for our comparative genomics analyses, we selected adequate comparisons within two monophyletic groups: genus UBA11606 and family TK06 (Fig. 1). Based on the functional genetic profile of the analyzed genomes, we show that sponge symbionts inhabiting different sponge species and deriving from different geographical locations are functionally more similar between them and distant from their free-living counterparts (Fig. 2). This observation is consistent with the previously reported functional convergence in sponge symbiotic communities [3, 14].

Genomic convergence across symbionts is driven by shared evolutionary pressures experienced within the sponge environment. The high rates of water pumping due to the sponge's filter-feeding activities lead sponge symbionts to be exposed to many viral particles [66, 67]. This has resulted in a higher abundance of PDS in sponge symbiont genomes [65, 68] to protect bacteria from exogenous DNA [12, 29, 69]. Actinobacteriota symbionts are no exception, and we found PDS to be enriched in sponge symbionts as compared to their free-living counterparts (Fig. S3 and Fig. S4A). Similarly, we found that symbiotic Acidimicrobia are enriched in ELPs compared to their free-living relatives. ELPs present in genomes of sponge symbionts were suggested to mediate host-microbial recognition and to enable evasion of symbionts from host digestion [12, 20, 21]. Examples are the enrichment of ANK, LRR, fn3, and CAD in Cyanobacteriota [13, 15], ANK and LRR in Thaumarchaeota [18],

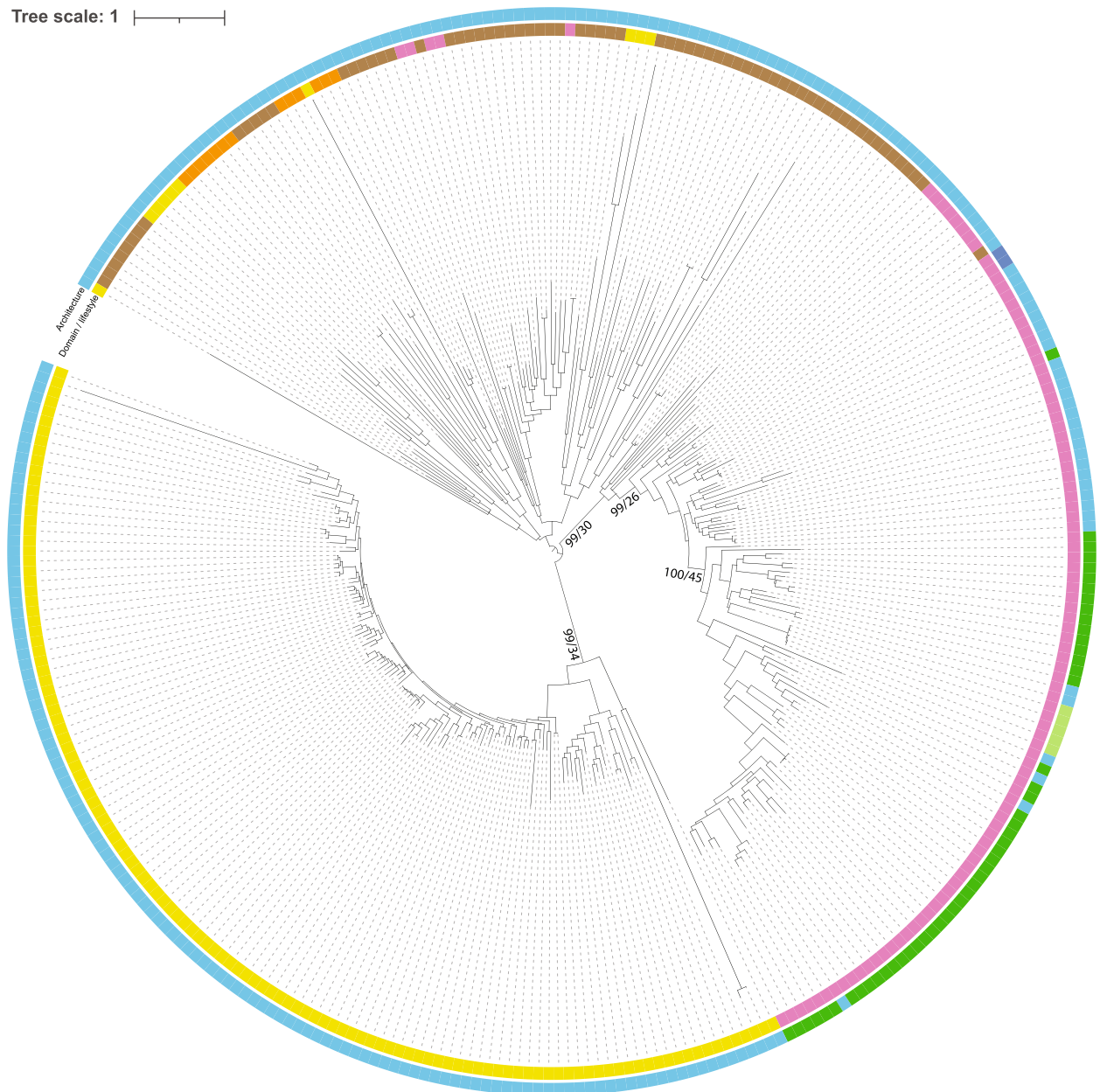
TPR in Poribacteria [11], and ANK and TPR in Chloroflexota [70]. Here, we found that sponge symbiotic Actinobacteriota genomes within the family TK06 and the genus UBA11606 exhibited significant enrichment in ELPs containing various domains, including LRR, which plays a role in protein-protein interactions; calx-beta, involved in signal transfer across the plasma membrane; and fn3 and CAD, implicated in functions like cell adhesion, morphology, and migration [71]. However, some ELP domains showed distinct enrichment patterns in TK06 as compared to UBA11606. For instance, genomes from genus UBA11606 were enriched in TPR genes, while those from family TK06 were enriched in WD40 genes, an ELP previously reported in the microbial community of the sponge *I. ramosa* [22]. By contrast, our analysis, in agreement with previous studies [22, 28], did not reveal a significant enrichment in ANK proteins in Actinobacteriota symbionts compared to their free-living counterparts. Interestingly, ANKs have been found to be enriched in other sponge symbiotic phyla, where their role in evading host phagocytosis was experimentally substantiated [20, 25, 66]. This suggests that Actinobacteriota likely employs an alternative, yet unknown, mechanism to evade host phagocytosis [22, 28].

Sponge symbionts are enriched in GPP34

Our analysis revealed a notable genetic characteristic exclusive to symbiotic Actinobacteriota when compared to their free-living counterparts: the presence of genes encoding for GPP34 proteins. These proteins are annotated as eukaryotic proteins [71]. Early proteomic studies first described eukaryotic GPP34 (also called GMx33, GOLPH3, and Golgi phosphoprotein 3) as conserved cytosolic *trans*-Golgi-associated proteins in human, mouse, fruit fly *Drosophila melanogaster*, roundworm *Caenorhabditis elegans*, and yeasts *Saccharomyces cerevisiae* (in the later, reported under the name Vps74), *Schizosaccharomyces pombe*, and *Kluyveromyces lactis* [72–74]. Functionally, GPP34 is annotated as a phosphatidylinositol-4-phosphate (PI4P)-binding protein [75], a feature that enables GPP34 to localize to the Golgi by binding PI4P, which is abundant in the Golgi membrane. The eukaryotic GPP34 protein plays an important role in shaping the Golgi and promoting vesicle formation. Specifically, GPP34 links the PI4P to actomyosin, thus establishing a connection between the Golgi and F-actin which facilitates Golgi vesicle trafficking [76].

GPP34-like proteins have not been previously reported in bacteria. However, our expanded analysis including 19 additional bacteria phyla revealed the presence of GPP34 coding genes across 16 sponge symbiotic phyla and their expression in 11 of those (Fig. 4 and Fig. S12). When looking at the correlations between ELPs and GPP34, we

Tree scale: 1 | ——— |



Domain / Lifestyle

- Bacteria: non-symbiont
- Bacteria: symbiont
- Archaea
- Eukarya

Architecture

- Short (GPP34 only)
- Long (GPP34 combined with cyt450)
- Long (GPP34 combined with Rieske)
- Long (GPP34 combined with PotA)

Fig. 6 Evolution of the GPP34 domain. Phylogenetic tree reconstructed with IQ-TREE under the EX_EHO + G4 model of evolution. The tree was rooted at midpoint for the graphical representation; however, the exact position of the root is unknown. The phylogenetic tree includes 133 eukaryotic GPP34 domains (highlighted in yellow), 13 archaeal GPP34 proteins (highlighted in orange), and 171 sequences of bacterial origin, from which 103 are sponge symbiotic (highlighted in pink) and 68 are from non-sponge symbionts (highlighted in brown). The outer color ring represents protein architecture, i.e., short (GPP34_S) in light blue, or long (GPP34_L), with dark green for GPP34 combined with cyt450, dark blue for GPP34 combined with Rieske, light green for GPP34 combined with PotA; see Fig. 5 for examples of the protein architectures. The inner color ring represents the domain of life (Eukarya, Archaea, and Bacteria) and within the domain Bacteria also the lifestyle (symbiont versus non-symbiont). Ultrafast bootstraps/nonparametric bootstrap supports are indicated only for selected nodes. To view information on all UF boot and NP boot, see Fig. S13. Raw phylogenetic tree and alignment can be found at <https://doi.org/10.6084/m9.figshare.27325512.v1>

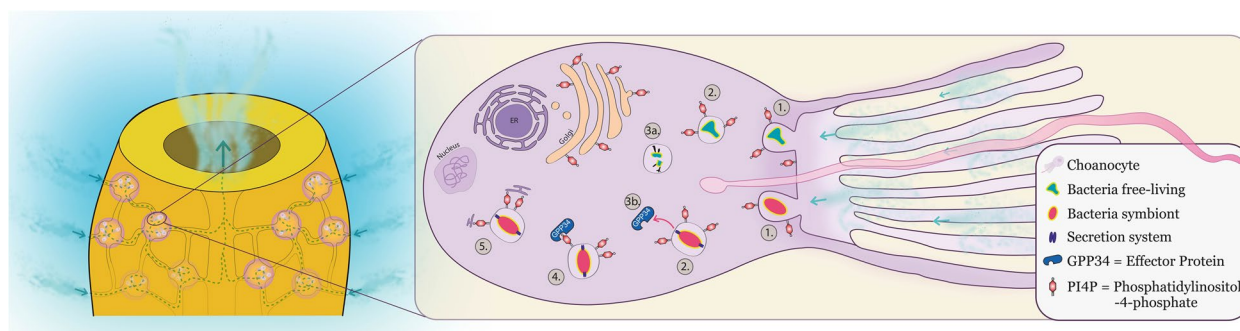


Fig. 7 Schematic hypothesis of the mechanism used by sponge symbionts to interfere in the host phagosome maturation. Scheme of the proposed new mechanism by which sponge symbionts evade host phagocytosis upon interaction with a sponge choanocyte, resulting in the establishment of a stable association with the host. The draft model represented is based on the strategies used by the pathogen *L. pneumophila* [86] to manipulate host phagocytosis. GPP34 protein may act as an effector protein that is potentially secreted by the symbiont and bind to phosphatidylinositol 4-phosphate (PI4P), thus hijacking phagolysosome maturation and degradation of the phagocytized bacterial cell. Legend of the figure steps: (1.) engulfment of the bacteria, (2.) host targets it towards phagosome maturation, (3a.) free-living bacteria is degraded as food, (3b.) symbiotic bacteria secrete GPP34 protein, (4.) GPP34 protein binds to PI4P, and (5.) recruitment of vesicles derived from the endoplasmic reticulum (ER), thus disguising the sponge symbiont as an ER vesicle

note different scenarios. GPP34 is scarce in Poribacteria that are highly enriched in ELPs, whereas Latescibacteriota and Acidobacteriota are enriched in both ELPs and GPP34. Additionally, Chloroflexota seems to differentiate its ELPs and GPP34 distribution among its classes. These findings suggest that each bacterial phylum/lineage, which includes sponge symbionts, has evolved unique sets of symbiosis-related genes, which play pivotal roles in mediating interactions between symbiotic bacteria and the sponge host, contributing to the establishment and maintenance of symbiosis.

GPP34 proteins may be involved in a mechanism for escaping symbiont degradation by the sponge host

As bacteria lack organelles such as the Golgi, the presence of proteins functionally annotated as PI4P-binding proteins suggests their potential involvement in modulating host-symbiont interaction. Notably, pathogenic bacteria and viruses have been reported to express proteins that interfere with the host's PI metabolism, allowing them to evade digestion by the host [77–79]. For instance, the pathogen *L. pneumophila* carries distinct effector proteins, namely SidC and SidM, which, like GPP34, also bind to PI4P [80, 81]. By employing these effector proteins, *L. pneumophila* can alter the composition of the phagosomal membrane [82, 83], remodeling phagosomes into a vacuole that resembles ER vesicles, called the *Legionella*-containing vacuole (LCV), where *L. pneumophila* can replicate [84]. Similarly, a fungal symbiont of plants can also interfere with host PI dynamics. The fungus *Laccaria bicolor* secretes a protein that is transported into the plant cell via phosphatidylinositol

3-phosphate-mediated endocytosis, promoting the colonization of the tree roots [85].

Based on these studies, we propose that phagocytized sponge symbionts may use GPP34 as a secreted effector protein that binds PI4P, thus hijacking the phagolysosome maturation that would normally lead to bacterial degradation and digestion, resulting in the stable establishment of the symbiont within the sponge host (Fig. 7).

Proteins containing the mixed-domain architecture of GPP34 with cytochrome P450 are unique to sponge symbionts

Across the 20 sponge symbiotic phyla analyzed, we found GPP34 proteins composed of four domain architecture groups. The most abundant architecture, harboring only a GPP34 domain (GPP34_S), is widespread across the three domains of life: Archaea, Bacteria, and eukaryotes. The second most abundant architecture, composed of GPP34 and cytochrome P450 domains (GPP34_L), is only found in sponge bacterial symbionts. When analyzing the flanking region of GPP34_S proteins in sponge-symbiotic bacteria, we often find a cytP450 protein within five genes up- or downstream of GPP34_S (Fig. S11). The presence of cytP450 in both the GPP34_S flanking region and as part of the GPP34_L architecture suggests that cytP450 is likely important for GPP34 function in the context of sponge symbiosis. While additional studies are needed to clarify its role and evolutionary origin within GPP34_L, previous work in eukaryotes revealed cytP450 as a component of the Golgi membrane proteome [73] and found that the upregulation of cytP450 1A1 (CYP1A1) can hinder macrophage phagocytosis [87], highlighting the

potential significance of cytP450 proteins in modulating sponge-microbe interactions.

Phylogeny of GPP34 proteins suggests a bacterial origin

Although GPP34 proteins were initially described as eukaryotic [72–74], our phylogenetic analysis reveals their presence across all three domains of life, suggesting an ancient evolutionary origin. The monophyly of GPP34 domains from sponge symbionts, along with their closer relationship to GPP34 domains from nonsymbiotic bacteria rather than to eukaryotic GPP34 sequences, supports the hypothesis that the symbiotic GPP34 protein family originated from bacteria and was not acquired through recent horizontal gene transfer from eukaryotes (Fig. 6). Furthermore, the divergence between GPP34 in sponge symbionts and those in non-sponge symbiont suggests that the symbiotic function of GPP34 in sponges evolved specifically within this group, while GPP34 in nonsymbiotic bacteria may serve a different role. GPP34 proteins are not present in all Eukaryota and Archaea lineages. Among eukaryotes, GPP34 proteins were mainly found to be present in amorpheans (i.e., amoebozoans, fungi, animals, and other single-celled opisthokonts). In few other eukaryote lineages, their presence is limited to one or two representatives. Similarly, in Archaea, the GPP34 domain could only be found in members of the Asgard clade, the sister clade of eukaryotes. This suggests that the GPP34 domain may have been lost in several eukaryotic and archaeal lineages. Alternatively, because of this patchy distribution, we cannot exclude a more complex evolutionary scenario (for example, involving several ancient horizontal transfers from bacteria).

As described above, GPP34 proteins are present under different architectures in symbionts: either in a short form with a single GPP34 domain or in a long form (which exists only in sponge symbionts), in which the GPP34 domain is combined with an additional domain (e.g., cytP450). Two evolutionary scenarios may explain this pattern. Under the first scenario, two different domains were present in the common ancestor clade, and these domains were combined, possibly by a recombination event, into a single long protein in some clades. Under the second scenario, the ancestral clade possessed a protein with the two domains, and in some lineages, the additional domain was lost. We speculate that the first scenario is more likely, as it assumes a simpler form for the ancestral protein and because the simple form is present in all domains of life, while the long form is only present in symbionts. Interestingly, few GPP34_S sequences are present within the GPP34_L clade suggesting that some GPP34_S sequences originated from a loss of the P450 domain. Furthermore, previous studies have reported

other cytP450-redox partner fusions, where the same protein harbors the cytP450 domain together with domains that aid in the electron transfer. For instance, P450 BM3, first characterized in the bacterium *Bacillus megaterium*, harbors cytP450 linked to NADPH-cytochrome P450 reductase (CPR, composed of FMN and FAD domains), which catalyzes electron transfer from NADPH to cytP450 [88].

Conclusions

This study presents a detailed comparative genomics analysis on 350 genomes of Actinobacteriota, a bacterial phylum commonly associated with sponges. Our analysis revealed several differences between symbiotic and free-living bacteria, advancing our knowledge of the mechanisms involved in sponge-microbe interactions. We identify a monophyletic group of GPP34 proteins found across sponge symbionts from diverse microbial phyla and absent from their closest free-living counterparts, underscoring its significance in the symbiotic context. While GPP34 was previously documented exclusively in eukaryotic organisms, we show that this protein family presents a patchy distribution across the three domains of life. In eukaryotes, GPP34 is predicted to bind PI4P, a function that is also attributed to secreted proteins in pathogenic bacteria that are involved in manipulating phagosome maturation, facilitating pathogen replication and establishment within the host. Accordingly, we propose that GPP34 represents a novel mechanism by which symbionts may inhibit phagosome maturation to enable bacterial establishment and maintenance within the sponge host, which requires experimental validation.

Abbreviations

ANK	Ankyrin repeat
CAD	Cadherin
CRISPR–Cas	Clustered regularly interspaced short palindromic repeats-associated
cytP450	Cytochrome P450
DELTA-BLAST	Domain enhanced lookup time accelerated basic local alignment search tool
DND system	DNA phosphothiolation
ELD	Eukaryotic-like domain
ELP	Eukaryotic-like protein
ER	Endoplasmic reticulum
fn3	Fibronectin type III
GTDB	Genome Taxonomy Database
HGT	Horizontal gene transfer
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
LCV	<i>Legionella</i> -containing vacuole
LRR	Leucine-rich repeat
MAG	Metagenome-assembled genome
NCBI	National Center for Biotechnology Information
NMDS	Nonmetric multidimensional scaling
NP boot	Nonparametric bootstrap
PCoA	Principal coordinates analysis
PDS	Prokaryotic defense system
PFAM	Protein families
PI	Phosphatidylinositol

PI3P	Phosphatidylinositol 3-phosphate
PI4P	Phosphatidylinositol 4-phosphate
R-M	Restriction-modification
T4SS	Type IV secretion systems
TA	Toxin-antitoxin
TPM	Transcripts per million
TPR	Tetratricopeptide repeat
UF boot	Ultrafast bootstrap

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-024-01963-1>.

Supplementary material: Data Set S1. Table S1: Genomic information on 350 actinobacteriota genomes. Table S2: Enriched functional KEGG annotation list, based on DESeq2 analysis ($P=0.05$), for the symbiotic and free-living genomes ($n=25$) from genus UBA11606. Table S3: Enriched functional KEGG annotation list, based on DESeq2 analysis ($P=0.05$), for the symbiotic and free-living genomes ($n=74$) from family TK06. Table S4: Raw counts and relative abundance for the four PDS categories for the symbiotic and free-living genomes under the genus UBA11606, for statistical analysis purpose (based on total counts). Table S5: Raw counts and relative abundance for the four PDS categories for the symbiotic and free-living genomes under the family TK06, for statistical analysis purpose (based on total counts). Table S6: The numbers represent the count of proteins in each genome, for each class type (9/11), with 1-2 annotated domains. Table S7: The numbers represent the count of proteins in each genome, for each class type (10/11), with 1-3 annotated domains. Table S8: Distribution of GPP34 (K15620) among the genomes present in the Actinobacteria tree. Table S9: ELPs and GPP34 distribution among genomes from Robbins et al., 2021. Table S10: List of GPP34 proteins found in genomes from Robbins et al., 2021. Table S11: Metadata for GPP34 proteins ($n=317$) in the GPP34 protein tree. Includes Bacteria, Eukaryota and Archaea. Table S12: List of the genomes used in the metatranscriptomics analysis. Table S13: List of genes used in the metatranscriptomics analysis and their respective expression. Table S14: Metadata and supporting information for the supplementary figures SF7, SF8, and SF9.

Supplementary Material S2. Table S1: Quantitative information on ELPs in genus UBA11606 and family TK06 (symbionts and free-living). Fig. S1: Phylogenomic tree of Actinobacteriota phylum, identical to Fig. 1, but including genome names. Fig. S2: Enrichment of functional categories in genomes of either free-living or symbiotic bacteria, belonging to the genus UBA11606 and family TK06. Fig. S3: Distribution of prokaryotic defense system (PDS) genes based on KEGG identifiers. This distribution is across genomes within the genus UBA11606 and family TK06, from two different lifestyles: symbiotic and freelifing. Fig. S4: Statistical analysis for the enrichment of PDS and ELP categories. Fig. S5: Distribution of ELPs and GPP34 proteins across genomes from the dataset of Robbins et al., 2021 (Fig. 3 [28]). The phylogenomic tree is the same as Fig. 4, but includes also the genome names. Fig. S6: Non-metric multidimensional scaling analysis of bacterial genomes based on their ELP and GPP34 content. Fig. S7: Distribution of the GPP34 gene within Chloroflexota among sponge symbionts and their closely related bacteria with different lifestyles. Fig. S8: Distribution of the GPP34 gene within Proteobacteriota among sponge symbionts and their closely related bacteria with different lifestyles. Fig. S9: Distribution of the GPP34 gene within Acidobacteriota among sponge symbionts and their closely related bacteria with different lifestyles. Fig. S10: Size distribution of GPP34 proteins. Two main protein size categories are observed: 200-260 aa, and 620-700 aa. Fig. S11: Synteny map of the genomic region flanking the GPP34 gene, for the sponge symbionts from genus UBA11606. Fig. S12: Expression of GPP34 genes across different MAGs from the sponge species *Aplysina aerophoba*. Fig. S13: Evolution of the GPP34 domain. Phylogenetic tree reconstructed with IQtree under the EX_EHO+G4 model of evolution. Non-parametric and ultrafast bootstrap supports are represented by branch colors in panels A and B, respectively.

Acknowledgements

The authors thank the Gordon and Betty Moore Foundation and the Israel Science Foundation for the financial support. We would also like to acknowledge Dr. Cláudio Nunes-Alves for editorial consulting.

Authors' contributions

C.F. wrote the main manuscript text and participated in all analyses. The study was based on preliminary phylogenomic analyses by I.B., who provided consultation throughout the project. T.P. contributed to the preparation of both phylogenomic trees presented. M.L. contributed to statistical analyses. G.R. conducted the analysis of gene expression data. D.H. guided the analysis of GPP34 phylogeny and contributed to the preparation and interpretation of the phylogenetic tree. L.S. supervised the project, co-wrote the manuscript with C.F., and secured the funding. All authors reviewed and approved the final manuscript.

Funding

This study was funded by the Gordon and Betty Moore Foundation, through Grant GBMF9352, and by the Israel Science Foundation (grant no. 933/23). GAR was supported by a Zuckerman Postdoctoral Research Fellowship.

Data availability

Sequences analyzed in this study are available on the NCBI GenBank database, where they were originally submitted as part of previously published articles. Genomes of the MAGs published in Robbins et al. (2021) are not available in NCBI but can be accessed via https://data.ace.uq.edu.au/public/sponge_mags/. Detailed information, including the articles and BioSample numbers for the analyzed genomes, is provided in the supplemental file: Data Set 1. Raw phylogenetic trees and alignments can be found at <https://doi.org/https://doi.org/10.6084/m9.figshare.27324744.v1> and <https://doi.org/https://doi.org/10.6084/m9.figshare.27325512.v1>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Marine Biology, Leon H. Charney School of Marine Sciences, University of Haifa, Haifa, Israel. ²Department of Biological Sciences, California State University, Los Angeles, CA, USA. ³Bioinformatic Services Unit, Faculty of Natural Sciences, University of Haifa, Haifa, Israel. ⁴George S. Wise Faculty of Life Sciences, School of Zoology, Tel Aviv University, Tel Aviv, Israel. ⁵The Steinhardt Museum of Natural History and National Research Center, Tel Aviv University, Tel Aviv, Israel.

Received: 18 June 2024 Accepted: 1 November 2024

Published online: 07 January 2025

References

- Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*. 2022;20:206–18.
- Chuckran PF, Hungate BA, Schwartz E, Dijkstra P. Variation in genomic traits of microbial communities among ecosystems. *FEMS Microbes*. 2021;2:xtab020.
- Thomas T, Moitinho-Silva L, Lurgi M, Björk JR, Easson C, Astudillo-García C, et al. Diversity, structure and convergent evolution of the global sponge microbiome. *Nat Commun*. 2016;7:11870.
- Webster NS, Thomas T. The sponge hologenome *mBio*. 2016;7:e00135-e216.

5. Burgsdorf I, Sizikov S, Squatrito V, Britstein M, Slaby BM, Cerrano C, et al. Lineage-specific energy and carbon metabolism of sponge symbionts and contributions to the host carbon pool. *ISME J.* 2022;16:1163–75.
6. Song H, Hewitt OH, Degnan SM. Arginine biosynthesis by a bacterial symbiont enables nitric oxide production and facilitates larval settlement in the marine-sponge host. *Curr Biol.* 2021;31:433–437.e3.
7. Wilkison CR, Fay P. Nitrogen fixation in coral reef sponges with symbiotic cyanobacteria. *Nature.* 1979;279:527–9.
8. Hoffmann F, Radax R, Woebken D, Holtappels M, Lavik G, Rapp HT, et al. Complex nitrogen cycling in the sponge *Geodia barretti*. *Environ Microbiol.* 2009;11:2228–43.
9. Wilson MC, Mori T, Rückert C, Uria AR, Helf MJ, Takada K, et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature.* 2014;506:58–62.
10. Tian RM, Wang Y, Bougouffa S, Gao ZM, Cai L, Bajic V, et al. Genomic analysis reveals versatile heterotrophic capacity of a potentially symbiotic sulfur-oxidizing bacterium in sponge. *Environ Microbiol.* 2014;16:3548–61.
11. Kamke J, Rinke C, Schwientek P, Mavromatis K, Ivanova N, Sczyrba A, et al. The candidate phylum *Poribacteria* by single-cell genomics: new insights into phylogeny, cell-compartmentation, eukaryote-like repeat proteins, and other genomic features. *PLoS ONE.* 2014;9:e87353.
12. Thomas T, Rusch D, DeMaere MZ, Yung PY, Lewis M, Halpern A, et al. Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J.* 2010;4:1557–67.
13. Gao ZM, Wang Y, Tian RM, Wong YH, Batang ZB, Al-Suwailem AM, et al. Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont "*Candidatus* *Synechococcus spongiarum*." *mBio.* 2014;5:e00079-14.
14. Fan L, Reynolds D, Liu M, Stark M, Kjelleberg S, Webster NS, et al. Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *Proc Natl Acad Sci USA.* 2012;109:E1878–1887.
15. Burgsdorf I, Handley KM, Bar-Shalom R, Erwin PM, Steindler L. Life at home and on the roam: genomic adaptations reflect the dual lifestyle of an intracellular, facultative symbiont. *mSystems.* 2019;4:e00057-19.
16. Sizikov S, Burgsdorf I, Handley KM, Lahyani M, Haber M, Steindler L. Characterization of sponge-associated *Verrucomicrobia*: microcompartment-based sugar utilization and enhanced toxin–antitoxin modules as features of host-associated *Opitutales*. *Environ Microbiol.* 2020;22:4669–88.
17. Burgsdorf I, Slaby BM, Handley KM, Haber M, Blom J, Marshall CW, et al. Lifestyle evolution in cyanobacterial symbionts of sponges. *mBio.* 2015;6:e00391-15.
18. Haber M, Burgsdorf I, Handley KM, Rubin-Blum M, Steindler L. Genomic insights into the lifestyles of Thaumarchaeota inside sponges. *Front Microbiol.* 2021;11:622824.
19. Liu MY, Kjelleberg S, Thomas T. Functional genomic analysis of an uncultured δ -proteobacterium in the sponge *Cymbastela concentrica*. *ISME J.* 2011;5:427–35.
20. Reynolds D, Thomas T. Evolution and function of eukaryotic-like proteins from sponge symbionts. *Mol Ecol.* 2016;25:5242–53.
21. Diez-Vives C, Moitinho-Silva L, Nielsen S, Reynolds D, Thomas T. Expression of eukaryotic-like protein in the microbiome of sponges. *Mol Ecol.* 2017;26:1432–51.
22. Engelberts JP, Robbins SJ, de Goeij JM, Aranda M, Bell SC, Webster NS. Characterization of a sponge microbiome using an integrative genome-centric approach. *ISME J.* 2020;14:1100–10.
23. Lurie-Weinberger MN, Gomez-Valero L, Merault N, Glöckner G, Buchrieser C, Gophna U. The origins of eukaryotic-like proteins in *Legionella pneumophila*. *Int J Med Microbiol.* 2010;300:470–81.
24. Bork P. Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins: Structure, Function, and Bioinformatics.* 1993;17:363–74.
25. Nguyen MTHD, Liu M, Thomas T. Ankyrin-repeat proteins from sponge symbionts modulate amoebal phagocytosis. *Mol Ecol.* 2014;23:1635–45.
26. Pan X, Lührmann A, Satoh A, Laskowski-Arce MA, Roy CR. Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. *Science.* 1979;208(320):1651–4.
27. Habyarimana F, Al-Khodori S, Kalia A, Graham JE, Price CT, Garcia MT, et al. Role for the ankyrin eukaryotic-like genes of *Legionella pneumophila* in parasitism of protozoan hosts and human macrophages. *Environ Microbiol.* 2008;10:1460–74.
28. Robbins SJ, Song W, Engelberts JP, Glasl B, Slaby BM, Boyd J, et al. A genomic view of the microbiome of coral reef demosponges. *ISME J.* 2021;15:1641–54.
29. Slaby BM, Hackl T, Horn H, Bayer K, Hentschel U. Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. *ISME J.* 2017;11:2465–78.
30. Fondi M, Orlandini V, Maida I, Perrin E, Papaleo MC, Emiliani G, et al. Draft genome sequence of the volatile organic compound-producing Antarctic bacterium *Arthrobacter* sp. strain TB23, able to inhibit cystic fibrosis pathogens belonging to the *Burkholderia cepacia* complex. *J Bacteriol.* 2012;194:6334–5.
31. Horn H, Hentschel U, Abdelmohsen UR. Mining genomes of three marine sponge-associated actinobacterial isolates for secondary metabolism. *Genome Announc.* 2015;3:e01106-e1115.
32. Karimi E, Gonçalves JMS, Reis M, Costa R. Draft genome sequence of *Microbacterium* sp. strain Alg239_V18, an actinobacterium retrieved from the marine sponge *Spongia* sp. *Genome Announc.* 2017;5:e01457-16.
33. Ian E, Malko DB, Sekurova ON, Bredholt H, Rückert C, Borisova ME, et al. Genomics of sponge-associated *Streptomyces* spp. closely related to *Streptomyces albus* J1074: insights into marine adaptation and secondary metabolite biosynthesis potential. *PLoS One.* 2014;9:e96719.
34. Huang X, Kong F, Zhou S, Huang D, Zheng J, Zhu W. *Streptomyces tirandamycinicus* sp. nov., a novel marine sponge-derived actinobacterium with antibacterial potential against *Streptococcus agalactiae*. *Front Microbiol.* 2019;10:482.
35. Orlandini V, Maida I, Fondi M, Perrin E, Papaleo MC, Bosi E, et al. Genomic analysis of three sponge-associated *Arthrobacter* Antarctic strains, inhibiting the growth of *Burkholderia cepacia* complex bacteria by synthesizing volatile organic compounds. *Microbiol Res.* 2014;169:593–601.
36. Schorn MA, Alanjary MM, Aguinaldo K, Korobeynikov A, Podell S, Patin N, et al. Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology (Reading).* 2016;162:2075–86.
37. Waters AL, Peraud O, Kasanah N, Sims JW, Kothalawala N, Anderson MA, et al. An analysis of the sponge *Acanthostrongylophora igensis* microbiome yields an actinomycete that produces the natural product manzamine A. *Front Mar Sci.* 2014;1:54.
38. Mangano S, Michaud L, Caruso C, Brilli M, Bruni V, Fani R, et al. Antagonistic interactions between psychrotrophic cultivable bacteria isolated from Antarctic sponges: a preliminary analysis. *Res Microbiol.* 2009;160:27–37.
39. Sun W, Zhang F, He L, Karthik L, Li Z. Actinomycetes from the South China sea sponges: isolation, diversity, and potential for aromatic polyketides discovery. *Front Microbiol.* 2015;6:1048.
40. Abdelmohsen UR, Pimentel-Elardo SM, Hanora A, Radwan M, Abou-El-Ela SH, Ahmed S, et al. Isolation, phylogenetic analysis and anti-infective activity screening of marine sponge-associated actinomycetes. *Mar Drugs.* 2010;8:399–412.
41. Harjes J, Ryu T, Abdelmohsen UR, Moitinho-Silva L, Horn H, Ravasi T, et al. Draft genome sequence of the antitrypanosomally active sponge-associated bacterium *Actinokineospora* sp. strain EG49. *Genome Announc.* 2014;2:e00160-14.
42. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol.* 2018;36:996.
43. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics.* 2020;36:1925–7.
44. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2022;50:D20–6.
45. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 2016;44:W242–5.
46. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics.* 2020;36:2251–2.
47. Oksanen J, Simpson G, Blanchet FG, Kindt R, Legendre P, Minchin P, et al. *Vegan*: community ecology package. R package version 2.6–2. 2022.
48. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE.* 2013;8:e61217.

49. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.
50. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* 2021;49:D412–9.
51. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30:1236–40.
52. Bruce P, Bruce A. Practical statistics for data scientists: 50 essential concepts. 2017. Available from: www.allitebooks.com
53. van Soest RWM, Boury-Esnault N, Vacelet J, Dohrmann M, Erpenbeck D, de Voogd NJ, et al. Global diversity of sponges (Porifera). *PLoS ONE.* 2012;7: e35105.
54. Guangchuang Yu. Data integration, manipulation and visualization of phylogenetic trees. Chapman & Hall/CRC. 2022.
55. Lee MD. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics.* 2019;35:4162–4.
56. Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, et al. IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics.* 2015;31:3359–61.
57. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. *Biol Direct.* 2012;7:12.
58. Ramírez GA, Bar-Shalom R, Furlan A, Romeo R, Gavagnin M, Calabrese G, et al. Bacterial aerobic methane cycling by the marine sponge-associated microbiome. *Microbiome.* 2023;11:49.
59. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9:357–9.
60. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30:2068–9.
61. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015;31:166–9.
62. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Res.* 2018;46:W200–4.
63. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;32:268–74.
64. Nguyen VH, Wemheuer B, Song W, Bennett H, Webster N, Thomas T. Identification, classification, and functional characterization of novel sponge-associated acidimicrobial species. *Syst Appl Microbiol.* 2023;46: 126426.
65. Liu M, Fan L, Zhong L, Kjelleberg S, Thomas T. Metaproteogenomic analysis of a community of sponge symbionts. *ISME J.* 2012;6:1515–25.
66. Jahn MT, Arkhipova K, Markert SM, Stigloher C, Lachnit T, Pita L, et al. A phage protein aids bacterial symbionts in eukaryote immune evasion. *Cell Host Microbe.* 2019;26:542–550.e5.
67. Vogel S. Current-induced flow through living sponges in nature. *Proc Natl Acad Sci USA.* 1977;74:2069–71.
68. Taylor JA, Díez-Vives C, Nielsen S, Wemheuer B, Thomas T. Communal-ity in microbial stress response and differential metabolic interactions revealed by time-series analysis of sponge symbionts. *Environ Microbiol.* 2022;24:2299–314.
69. Horn H, Slaby BM, Jahn MT, Bayer K, Moitinho-Silva L, Förster F, et al. An Enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes. *Front Microbiol.* 2016;7:1751.
70. Bayer K, Jahn MT, Slaby BM, Moitinho-Silva L, Hentschel U. Marine sponges as *Chloroflexi* hot spots: genomic insights and high-resolution visualization of an abundant and diverse symbiotic clade. *mSystems.* 2018;3:e00150–18.
71. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. *Nucleic Acids Res.* 2023;51:D418–27.
72. Wu CC, Taylor RS, Lane DR, Ladinsky MS, Weisz JA, Howell KE. GMx33: a novel family of trans-Golgi proteins identified by proteomics. *Traffic.* 2000;1:963–75.
73. Bell AW, Ward MA, Blackstock WP, Freeman HNM, Choudhary JS, Lewis AP, et al. Proteomics characterization of abundant Golgi membrane proteins. *J Biol Chem.* 2001;276:5152–65.
74. Bonangelino CJ, Chavez EM, Bonifacino JS. Genomic screen for vacuolar protein sorting genes in *Saccharomyces cerevisiae*. *Mol Biol Cell.* 2002;13:2486–501.
75. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics.* 2009;25:3045–6.
76. Dippold HC, Ng MM, Farber-Katz SE, Lee SK, Kerr ML, Peterman MC, et al. GOLPH3 bridges phosphatidylinositol-4-phosphate and actomyosin to stretch and shape the Golgi to promote budding. *Cell.* 2009;139:337–51.
77. Weber SS, Ragaz C, Hilbi H. Pathogen trafficking pathways and host phosphoinositide metabolism. *Mol Microbiol.* 2009;71:1341–52.
78. Poirier V, Av-Gay Y. Intracellular growth of bacterial pathogens: the role of secreted effector proteins in the control of phagocytosed microorganisms. *Microbiol Spectr.* 2015;3:VMBF-0003–2014.
79. Delang L, Paeshuyse J, Neyts J. The role of phosphatidylinositol 4-kinases and phosphatidylinositol 4-phosphate during viral replication. *Biochem Pharmacol.* 2012;84:1400–8.
80. Brombacher E, Urwyler S, Ragaz C, Weber S, Kami K, Overduin M, et al. Rab1 guanine nucleotide exchange factor SidM is a major phosphatidylinositol 4-phosphate-binding effector protein of *Legionella pneumophila*. *J Biol Chem.* 2009;284:4846–56.
81. Weber SS, Ragaz C, Reus K, Nyfeler Y, Hilbi H. *Legionella pneumophila* exploits PI(4)P to anchor secreted effector proteins to the replicative vacuole. *PLoS Pathog.* 2006;2:418–30.
82. Weber S, Steiner B, Welin A, Hilbi H. *Legionella*-containing vacuoles capture PtdIns(4)P-rich vesicles derived from the Golgi apparatus. *mBio.* 2018;9:e02420–18.
83. Ragaz C, Pietsch H, Urwyler S, Tüaden A, Weber S, Hilbi H. The *Legionella pneumophila* phosphatidylinositol-4 phosphate-binding type IV substrate SidC recruits endoplasmic reticulum vesicles to a replication-permissive vacuole. *Cell Microbiol.* 2008;10:2416–33.
84. Tilney LG, Harb OS, Connelly PS, Robinson CG, Roy CR. How the parasitic bacterium *Legionella pneumophila* modifies its phagosome and transforms it into rough ER: implications for conversion of plasma membrane to the ER membrane. *J Cell Sci.* 2001;114:4637–50.
85. Plett JM, Kemppainen M, Kale SD, Kohler A, Legué V, Brun A, et al. A secreted effector protein of *Laccaria bicolor* is required for symbiosis development. *Curr Biol.* 2011;21:1197–203.
86. Swart AL, Hilbi H. Phosphoinositides and the fate of *Legionella* in phagocytes. *Front Immunol.* 2020;11:25.
87. Tian LX, Tang X, Zhu JY, Luo L, Ma XY, Cheng SW, et al. Cytochrome P450 1A1 enhances inflammatory responses and impedes phagocytosis of bacteria in macrophages during sepsis. *Cell Communication and Signaling.* 2020;18:70.
88. Munro AW, Girvan HM, McLean KJ. Cytochrome P450-redox partner fusion enzymes. *Biochim Biophys Acta Gen Subj.* 2007;1770:345–59.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.