Distribution and physiology of Alphaproteobacteria living in symbiosis with marine gutless oligochaetes

Dissertation
zur Erlangung des Grades eines
Doktor der Naturwissenschaften
– Dr. rer. nat. –

dem Fachbereich Biologie/Chemie
der Universität Bremen vorgelegt von

Tina Enders
Bremen, Mai 2022

Die vorliegende Arbeit wurde in der Zeit von Juli 2017 bis Mai 2022 in der Abteilung Symbiose am Max-Planck-Institut für Marine Mikrobiologie in Bremen unter der Leitung von Prof. Dr. Nicole Dubilier und direkter Betreuung durch Dr. Harald R. Gruber-Vodicka angefertigt.

The research presented in this thesis was conducted from July 2017 to May 2022 in the Department of Symbiosis at the Max Planck Institute for Marine Microbiology in Bremen under the leadership of Prof. Dr. Nicole Dubilier and direct supervision by Dr. Harald R. Gruber-Vodicka.

Gutachtende | Reviewers

Prof. Dr. Nicole Dubilier

Dr. Pierre Offre

Tag des Promotionskolloquiums | Date of doctoral defense

20.07.2022

Diese zur Veröffentlichung erstellte Version der Dissertation enthält Korrekturen.

Alles zu seiner Zeit

Index

Summary

Symbioses are ubiquitous on earth and can be found across all kingdoms of life. The term symbiosis describes two organisms of different kind living together in a long-term and intimate association. One specific type are nutritional animal-microbe symbioses in which microorganisms provide metabolic functions and enable their animal hosts to thrive in otherwise inaccessible environments. Gutless oligochaetes from shallow water marine habitats are one example of such symbioses. The annelid worms entirely lost their digestive and excretory organs. Instead, they rely on nutrients and energy provided by a diverse but specific set of subcuticular but extracellular symbionts. Research has mostly focused on their chemosynthetic gammaproteobacterial and the sulfur-reducing deltaproteobacterial symbionts although Alphaproteobacteria have been shown early on to be partners in these symbioses. Alphaproteobacteria are an abundant, diverse and multifunctional class of bacteria and are promising to provide a plethora of functions to the gutless oligochaete symbiosis. However, little research has focused on their role in the gutless oligochaete symbiosis, which remains to be elucidated.

This thesis is a first step to characterize the taxonomic and functional diversity of Alphaproteobacteria in marine gutless oligochaetes. In the first part **(Chapter I)**, I give an overview of the Alphaproteobacteria that associate with gutless oligochaetes and provide directions for promising clades for further research. In the second part, I describe two specific, yet distinct symbiont species. One is a globally distributed representative of the common subcuticular symbiont community and is most abundant in *Olavius ilvae* in the Mediterranean Sea. This Alphaproteobacterium is a heterotroph and thrives on a variety of sugar compounds **(Chapter II)**. The other one is a low abundant symbiont in the intensively studied *Olavius algarvensis* community. It stands out by its highly reduced genome and represents one of the smallest bacterial genomes described from a marine animal-microbe symbiosis to date **(Chapter III)**. I used metagenomics, -transcriptomics and -proteomics approaches along with fluorescence *in situ* hybridization to gain insight into the symbiont taxonomy, their metabolic function and integration into the symbiont community, and their location inside the hosts' body.

Overall, my research sheds light on the most diverse class of gutless oligochaete symbionts – the Alphaproteobacteria – and provides the basis for future research on individual exciting symbiont clades.

Zusammenfassung

Symbiosen sind ein globales Phänomen, das Einzug in alle Domänen des Lebens gefunden hat. „Symbiose“ beschreibt eine Form des Zusammenlebens von Organismen verschiedener Art, die über einen längeren Zeitraum in engem Kontakt stattfindet. Eine besondere Form ist die Symbiose zwischen Tieren und Mikroorganismen, bei denen die Mikroorganismen den Stoffwechsel der Tiere erweitern und so die Besiedelung neuer Lebensräume ermöglichen. Ein Beispiel stellen Darmlose Wenigborster dar, die im küstennahen Meeresboden leben. Die Würmer haben völlig reduzierte Verdauungs- und Exkretionsorgane und sind stattdessen auf die Versorgung durch eine Gemeinschaft vielfältiger Bakterien angewiesen, die sich außerhalb der Wirtszellen unter der Cuticula angesiedelt haben. Bisher hat sich die Forschung hauptsächlich auf die chemosynthetischen Gammaproteobakterien und die Schwefel reduzierenden Deltaproteobakterien fokussiert. Dabei war schon früh bekannt, dass Alphaproteobakterien ebenso Teil dieser Symbiosen sein können. Alphaproteobakterien sind eine weit verbreitete und diverse Klasse der Bakterien, die mit ihren vielseitigen Stoffwechselwegen eine vielversprechende Menge an Funktionen zur Symbiose der Darmlosen Wenigborster beitragen könnte. Trotzdem wurde ihre Rolle in der Symbiose bisher nicht im Detail erforscht und bleibt dadurch eine spannende, offene Frage.

Diese Doktorarbeit ist ein erster Schritt, die taxonomische und funktionelle Diversität der Alphaproteobakterien in Symbiose mit marinen Darmlosen Wenigborstern zu beschreiben. Zu Beginn **(Kapitel I)** gebe ich einen Überblick über die Alphaproteobakterien, die symbiotisch mit Darmlosen Wenigborstern leben, und verweise auf Gruppen, die für zukünftige Forschung bedeutsam sein können. Im weiteren Verlauf gebe ich Einblick in zwei detaillierte Beispiele sehr unterschiedlicher Symbionten. Das erste Beispiel stellt ein Vertreter der typischen extrazellulären Symbionten-Schicht dar, der weltweit zu finden ist. Dieses Bakterium ist ein zahlreicher Symbiont der Wirts-Art *Olavius ilvae* aus dem Mittelmeer und lebt von organischen Kohlenstoffen wie Zuckern **(Kapitel II).** Das zweite Beispiel ist ein Symbiont aus der viel erforschten *Olavius algarvensis* Symbiose. Dieses Bakterium sticht durch seine geringe Häufigkeit und sein stark reduziertes Genom heraus und ist eines der kleinsten bekannten Bakterien aus dem marinen Lebensraum **(Kapitel III).** Zur Beschreibung der Taxonomie, der metabolischen Fähigkeiten und Funktionen in den symbiotischen Lebensgemeinschaften sowie zur Visualisierung der Bakterien im Wurm, habe ich „Omics“ (Metagenomik, -transcriptomik und –proteomik) sowie Fluoreszenz-*in-situ*-Hybridisierung angewendet.

Zusammenfassend trägt diese Arbeit dazu bei, die vielseitigste der bakteriellen Klassen in Symbiose mit Darmlosen Wenigborstern – die Alphaproteobakterien – besser zu verstehen und bietet damit Grundlagen und Aussichten für zukünftige Forschung.

Introduction

Symbiosis shapes all life on earth

We live in a world of sheer endless connectivity. There is almost no living being on this planet that is not in close interaction with its surrounding organisms. Symbiosis, stemming from the Greek words for together (sym-) and living (-biosis), is the clue to this phenomenon. Today's most widely used definition for symbiosis, especially in science, is based on concepts that the mycologist Heinrich Anton de Bary presented in a talk in 1879[1]. Throughout this thesis, symbiosis is considered as the living together of two organisms of different kind in an intimate and long-term association. Usually, the larger member of the association is referred to as the host and the smaller member or members in multipartite associations are referred to as the symbionts. De Bary's definition includes a variety of possible outcomes to the participants of a given association. The association can be to the benefit of both – mutualism, the detriment of one for the benefit of the other – parasitism, a neutral outcome for one or both of the partners – commensalism. Outcomes of symbioses can furthermore change from mutualism to parasitism and vice versa in short time frames within the same community, depending on the environmental factors. One example would be the most widespread insect symbiosis of *Wolbachia* spp. (Figure 1A). On one hand, *Wolbachia* can cause male killing in the host as a strategy to enforce its own distribution but at the same time it can also be necessary for host fitness and survival[2]. Symbiosis is prevalent across all kingdoms of life[3-6]. Well-studied examples that are by now common knowledge are the interactions of eukaryote macro-organisms with bacteria, for example root nodule forming rhizobia in legumes or the rumen microbiome in cattle (Figure 1B-C). Even more rare interactions have been characterized in great detail, such as archaea of the anaerobe methanotroph (ANME) group that engage with sulfur-reducing Deltaproteobacteria (Figure 1D). Not only can symbionts of the most distinct types form intricate symbioses, these associations can also serve remarkably diverse purposes. Prominent examples are nutritional symbioses in which one partner provides essential substrates such as microbiota that provide essential amino acids to sap-feeding insects, or defensive symbioses like the Squid-*Vibrio* symbiosis where bacteria aid the host to disguise as moonlight shimmer (Figure 1E-F)[7, 8]. Overall, symbiosis is ubiquitous on this planet, and it would fill books to elaborate on its diverse and fascinating outcomes.

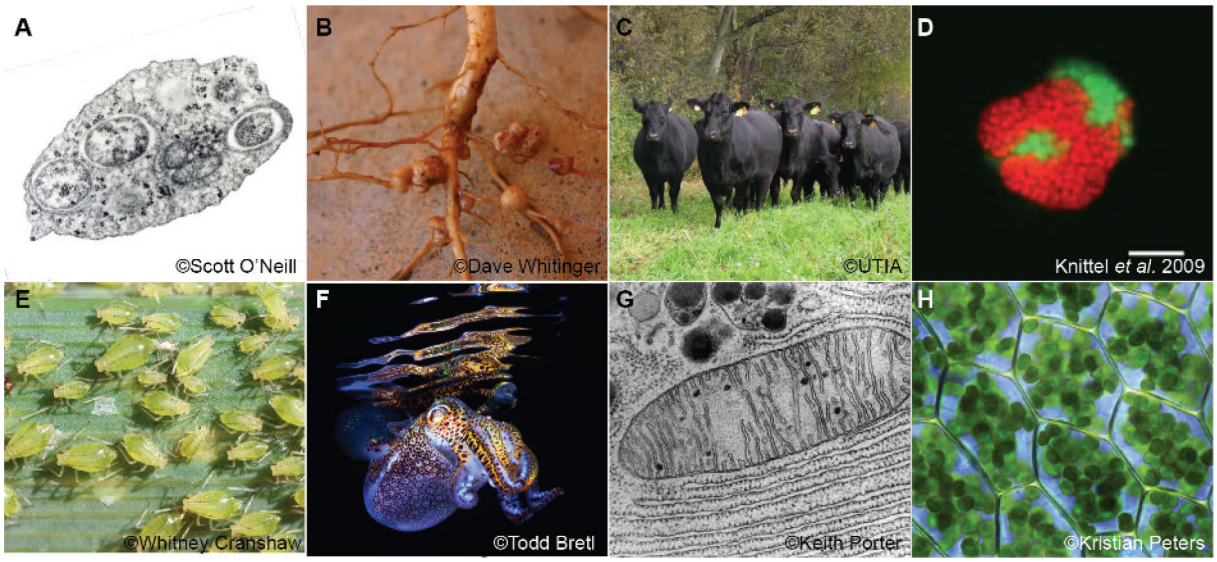


Figure 1: Symbioses are ubiquitous across all kingdoms of life.
A: *Wolbachia* are widespread endosymbionts in insects. B: Root-nodule forming Rhizobia. C: Anaerobe methane oxidizing Archaea with their sulfate-reducing bacterial partners. D: Cattle have diverse rumen microbiomes. E: Sap-feeding insects host vitamin provisioning *Buchnera*. F: *Euprymna* squids in symbiosis with bioluminescent *Vibrio*. G: Mitochondrium in an animal cells. H: Chloroplasts in plant cells.

Two major events of symbiosis had great impact on the evolution of life on earth[9]. The first event was the endosymbiosis of an aerobic prokaryote by an anaerobe to survive in the early oxic conditions on earth. The exact origin of the partner is still under debate, but scientists agree that relatives to today's Alphaproteobacteria were engulfed and are what we now call mitochondria (Figure 1G)[10, 11]. These organelles enabled cells to conserve more energy and grow into more complex life forms – the eukaryotes. The second event was the endosymbiosis of photosynthetic organisms to form plastids (Figure 1H). These organelles enable eukaryotes to build up carbohydrates autotrophically by fixing carbon from the atmosphere using light energy. With endosymbiosis as a fundament for complex life, it is no wonder that to date no multicellular organism is known without symbiotic interaction of some sort. Symbiosis enabled and still enables both pro- and eukaryotic organisms to thrive in niches that seemed uninhabitable at first glance. This shows the great potential that arises when combining intricate sets of functions.

Despite of all the benefits of the association for both partners, such intimate symbioses can also come with challenges. One striking example are insect symbioses where the bacterial symbionts have such tremendously reduced genomes that the symbionts' only metabolic function is to supply their hosts with amino acids unavailable through the animals' restricted diet[12, 13]. This example already shows that intimate symbiosis, despite the general benefit for the association

can have long-term adverse consequences for one or both partners. Genome reduction of the symbiont resulting from isolation from a broader genetic gene pool is one of these consequences. Starting with some deleterious mutations, this can lead to loss of functions and finally extinction of the bacterial symbiont. Consequently, the host might go extinct as well if there is no alternative to compensate for the function of the symbiont. Fatal consequences of genome reduction can be overcome by a variety of factors. One can be the transmission mode of the symbionts. The transmission can be vertical – from the host to its direct offspring, or horizontal – from other hosts or the environment to the new host, or a mixed mode spanning all degrees between the two extremes. Connectivity of the symbiont to the environment during transmission is one way to allow access to external gene pools. Another factor might be whether the symbiont pool is separated within the host or whether it is open to other symbiont species[14]. The latter allows for secondary access to new genes via horizontal gene transfer, or other types of recombination become possible. Flexibility in the gene and symbiont pool minimizes the impact of isolation and might make the partners of an association less dependent on the survival of the initial symbiont community over longer evolutionary periods.

Symbiosis has led to multiple complex life forms on earth and has shaped the life as we know it today tremendously, not only by the event of endosymbiosis to form complex organisms with chloroplasts and mitochondria. Symbiosis has proven as a concept of success and, despite the ever-growing research in this field, we have just started to understand these associations and find new and astonishing forms of interaction. Hence, it is of utmost importance to further explore symbioses around us if we want to understand how organisms can engage and communicate with each other.

Animal-microbe symbioses in shallow water marine ecosystems

Animal-microbe symbioses are one prominent example for symbioses. Animals, which are eukaryotes, and as such arose from the symbiosis of a bacterium and an archaeon to form the eukaryote cell, have a very long history of engaging with bacteria and continued successful interaction[9]. Well-studied examples are insect symbioses where bacteria provide essential amino acids or vitamins and co-factors to many sap-feeding hosts or provide enzymes to access carbon sources not accessible by the host itself such as lignocellulose degrading enzymes in wood-feeding termites[15, 16].

Chemosynthetic nutritional symbioses

Chemosynthesis is the ability to fix carbon with energy derived from the oxidation of inorganic compounds, as opposed to light energy in photosynthesis. Chemosynthetic symbiosis was discovered in the 70s in the deep sea where no light reaches the sea floor and even light-derived nutrients from the surface are scarce[17, 18]. Back then, scientists were surprised to discover large beds of mussels, clams and limpets and other species in this harsh environment[19]. Only when understanding the concept of chemosynthetic symbiosis in these far-out ecosystems, researchers realized that chemosymbiotic organisms must also be present in shallow water sediments[20]. Shallow water marine sediments with their own forms of extreme conditions also necessitate the collaboration of chemosynthetic bacteria with animals. The sediments are often characterized by low organic nutrient availability and long chemical gradients[21]. Chemosynthetic habitats are regularly associated to seagrass meadows, coral reefs or mangrove forests, which provide at least some substrate influx. A group of organisms that mastered to overcome the challenges in these sediments are animals living in a chemosynthetic nutritional symbiosis. Animals that are studied from these habitats are interstitial and small forms like ciliates, paracatenulid flat worms or gutless oligochaetes, but also macrofauna such as the lucinid clams[5, 22-24].

From a scientific perspective, the advantage of shallow water marine habitats, especially compared to deep-sea hydrothermal vents systems, is their easy accessibility. Located in coastal regions, scientists can reach them directly from the shore or by boat. With sampling sites ranging from knee-deep water down to a few tens of meters, scuba divers can access most sampling sites. Sometimes even walking in the sediment is sufficient to find chemosynthetic organisms. Particularily in the tropics and subtropics, a single bucket of sediment can contain a considerable diversity of hosts. The opportunity to repeatedly retrieve almost unlimited amounts of samples at very little cost enables in-depth research of these organisms despite the fact that we are not able to cultivate any of these exciting organisms[5, 25]. Densely sampled sites are the Mediterranean Sea and the Caribbean and suitable sites range from southern to northern warm temperate ecosystems, as the diversity of chemosynthetic symbioses drops considerably in cold temperate waters. Diverse chemosymbioses have been described to date[5]. However, with vast areas in the oceans yet unstudied, there will be more to uncover.

Gutless oligochaetes as diverse chemosynthetic host systems

Gutless oligochaetes (Clitellata, Annelida) are marine interstitial meiofauna from shallow water sediments and are a chemosynthetic animal-microbe symbiosis that has been studied for more than 40 years[26-28]. Scientists have described over 100 species from globally distributed sampling sites (Figure 2)[29]. These belong to the paraphyletic genera of *Inanidrilus* and *Olavius*. As the name implies, gutless oligochaetes entirely lack digestive and excretory organs and instead rely on a consortium of nutritional endosymbionts[30]. The bacteria reside below the worm's cuticle outside of the host's epidermal cell layer. The composition of the consortia is diverse and has been shown to be characteristic to the host species and the sampling location[30]. Most host species share a chemosynthetic gammaproteobacterial symbiont that fixes carbon by oxidizing sulfur compounds. Deltaproteobacteria were shown to reduce those compounds and replenish the sulfur stocks for the gammaproteobacterial symbiont[31]. The worms have a bright white appearance that is due to the storage of sulfur inclusions in the symbionts. The host provides the symbionts with their distinct favorable conditions by shuttling them through chemical gradients of the sediment[32].

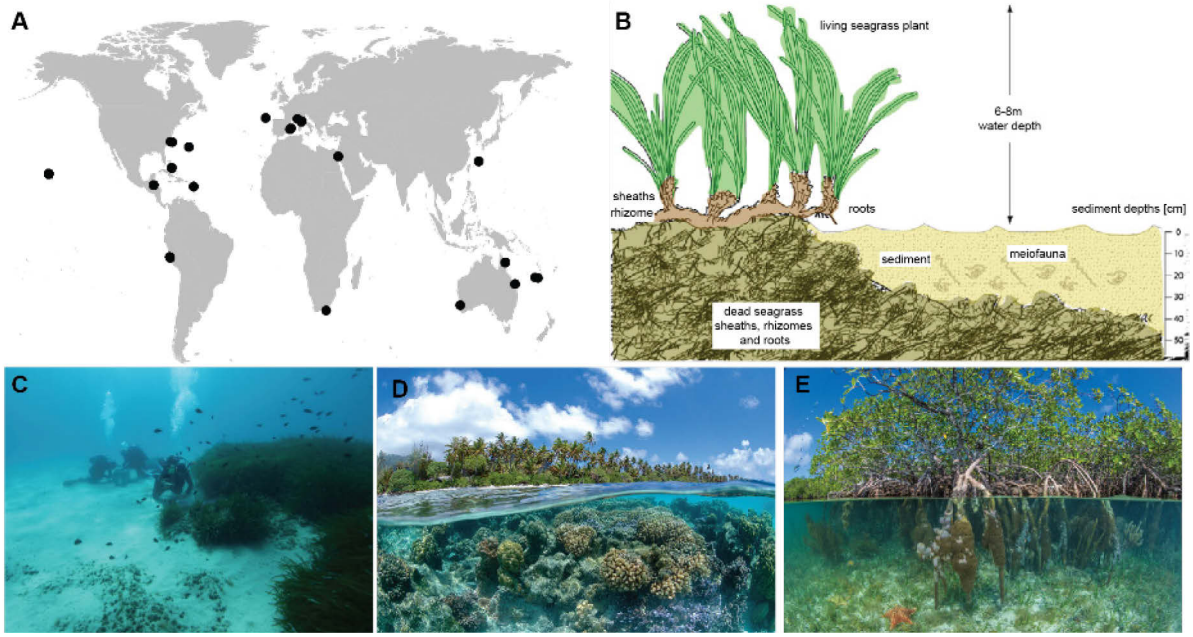


Figure 2: Globally distributed sampling sites of gutless oligochaetes are often associated with seagrass meadows, coral reefs or mangrove forests.
A: Globally distributed sampling sites of gutless oligochaetes. B: Detailed schematic of seagrass meadow associated sampling sites of *Olavius algarvensis* and *Olavius ilvae* from around Elba, Italy (figure modified from Kleiner *et al.* 2015 [33]). C: Divers close to seagrass meadow. D:Coral reef. E: Mangrove forest.

The best-studied host species is *Olavius algarvensis* from bays around the island of Elba in Italy that hosts a main and a secondary Gammaproteobacterium, up to four distinct Deltaproteobacteria and a Spirocheate (Figure 3)[24, 30, 31, 34]. The metabolism and the interaction of the Gammaproteobacteria and the Deltaproteobacteria have been intensely studied[21, 24, 35]. Other well studied host species were collected on Elba in the Mediterranean Sea and the Caribbean islands of the Bahamas, Bermuda and Belize. Host species that also have been investigated for their symbiont community in individual studies are *I. leukodermatus*, *I. makropetalos*, *Olavius crassitunicatus*, *O. loisae* and *O. ilvae*[34, 36-38]. These studies showed that the symbiont composition can greatly differ between host species and sampling sites. Although the gutless oligochaetes have been methodically studied throughout the years, there is still a lot of potential for new discoveries and mechanistic insights. One such area of great research potential is the presence of Alphaproteobacteria in these associations that have been observed in multiple hosts[36, 38]. However, only the latest metagenome study revealed that Alphaproteobacteria are even more diverse and widely distributed as secondary symbionts than the Deltaproteobacteria and were even overlooked in those species that have been investigated more intensively[30].

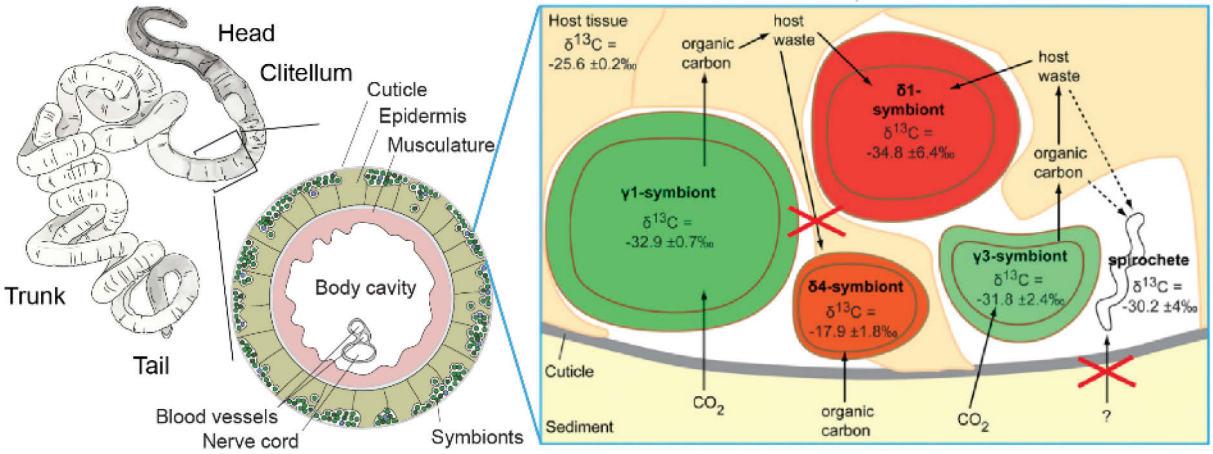


Figure 3: Gutless oligochaetes lack a digestive system. Instead they rely on a set of metabolically diverse subcuticular symbionts.
Left: Schematic representation of *Olavius algarvensis* (©Alina Esken) and a cross section through its trunk region (©Alexander Gruhl). Right: Stable-isotope fingerprints of *Olavius algarvensis* symbionts indicate that the symbionts use internal and external carbon sources (figure modified from Kleiner *et al.* 2018[39]).

The successful and ubiquitous class of Alphaproteobacteria

The class of Alphaproteobacteria in the Phylum Proteobacteria is one of the most successful clades of bacteria on earth. Wherever scientists take a sample and characterize the bacterial diversity with modern sequencing approaches, Alphaproteobacteria are present and very often even among the more abundant clades. This includes moderate environments such as soil or rivers where Alphaproteobacteria can make up to 50% of the community[40, 41]. But also in more extreme environments like deserts or hypersaline lakes Alphaproteobacteria are successful inhabitants[42, 43]. Many free-living Alphaproteobacteria are holding records and play crucial roles for human life. SAR11 is a group of marine bacteria of which *Candidatus* Pelagibacter ubique has the potential to be the most numerous bacterium on this planet[44]. Other Alphaproteobacteria have been tamed and are used to produce industrial scale amounts of acetate, alcohol and other substrates to feed and aid humans. Alphaproteobacteria are functionally versatile. The downside of this is that there are no pathways that are characteristic for them, which means that very often, their function cannot be inferred from their higher level taxonomy. The Genome Taxonomy Database (GTDB) lists 16016 entries in nearly 100 alphaproteobacterial orders of which 27 have a taxonomic name. These include the eight most well described orders, namely Acetobacterales, Caulobacterales, Pelagibacterales, Rhizobiales, Rhodobacterales, Rhodospirillales, Rickettsiales and Sphingomonadales[45]. However, the taxonomy of the Alphaproteobacteria has been under constant debate due to a lack of representatives in several clades or because historical taxonomy that deviates from recent sequencing insights, both complicating the matter[46]. As introduced earlier, Alphaproteobacteria play a crucial role in the origin of the eukaryotes. It is still unclear in which relation these early Alphaproteobacteria that have led to the mitochondria can be placed into what is the recent diversity of the class of Alphaproteobacteria but it seems they have been ascendants or sister to one of the larger recent clades[10, 11, 47]. Based on that we can assume that members of the Alphaproteobacteria might be among the most well-adapted groups of bacteria to engage with eukaryotic organisms as many of them likely have the tools and mechanisms to manipulate the eukaryotes' cellular biology.

Host-associated Alphaproteobacteria evolved to exert a variety of roles that range from successful free-living examples to pathogens and mutualistic symbionts. Orders prone for symbiotic interactions are the Rickettsiales and Hyphomicrobiales (former Rhizobiales) but also Rhodospirillales. Rickettsiales have evolved mostly pathogenic lifestyles and have been

mainly described to intracellularly infect animals but also many other eukaryote hosts. They are characterized by small genomes (<2Mb genome size) that have been reduced to only the necessary functions to infect and spread within their host. A well-studied example is *Wolbachia*, a successful insect symbiont, which has both parasitic and mutualistic functions depending on the setting[2]. Rickettsiales are potentially the closest relatives to ancestors of the Alphaproteobacteria that have been involved in the mitochondrial endosymbiosis event and the rise of eukaryotes[10, 11]. For example, the Rickettsiales, namely Midichloriaceae, are microbiome members of the simplest animal phyla, such as Placozoa or Cnidaria, and might even engage in mutualistic symbiosis[48-50]. Hyphomicrobiales are often found in soils and interact with plants in both commensalistic and mutualistic ways such as root nodule formation. They also engage in insect symbiosis and include the smallest bacterium that has been described to date – *Candidatus* Hodgkinia cicadicola with only 0.2 Mbp genome size[51].

Methods to study the uncultivable

One issue with intricate animal-microbe symbioses is that we are still not able to cultivate the bacterial symbionts without their hosts. This complicates the ability to examine the symbionts' physiology and manipulate their genomic information. Fixation of fresh animals provide a solid basis of preserved tissue for a multitude of possible analyses ranging from sequencing technologies, proteomics, metabolite investigations and other omics. In the best case, good access to shallow water symbiosis and the ability to rear or cultivate the holobiont enable us to study the association as a whole. Somewhat old-fashioned but powerful methods like fixation in methanol allow a range of different extractions later on with no need to decide on the wanted outcome ad hoc. Specific fixatives like RNAlater provide improved sample quality for genomics and proteomic approaches at storage temperatures more suited during field sampling[52]. Fixation in paraformaldehyde is still the go-to method for *in situ* hybridization. Fixation methods are still under development and allow for a broad range of applications. The development of the so-called second and third generation high throughput sequencing technologies have made sequencing faster and cheaper than ever before and together with library preparation improvements currently enables sequencing of large amounts of low-input samples (500 pg – 20 ng) to an in-depth degree[53]. These technological breakthroughs have enabled projects that investigate community structures and that can process marker genes in high throughput across hundreds or even thousands of samples and with that identify indistinguishable or low abundant community members[54, 55]. With such approaches, we can

for example study single individuals of small animals like the gutless oligochaetes and draw conclusion on e.g. the role of single members in their symbiotic community. The same is also true for the development of metatranscriptomics and metaproteomics approaches, with which we can gain a snapshot of the expression and the metabolic response of the individual partners at the time of fixation in the field. Sequencing tools cannot yet provide detailed spatial information, so we rely on hybridization techniques to localize information[56, 57]. With the current state-of-the-art sample preparations such as high pressure freezing and 3D reconstruction using e.g. micro-CT, we can integrate techniques and pinpoint expression in single community members and describe their morphological connectedness and potential modes of interaction to the host[58]. Overall, the shallow water symbiosis has come a long way, from pools of 10 000s of individuals necessary for breakthrough research only a decade ago, we are now able to specifically collect or manipulate single specimens, extract genomic, expression or metabolomic information of single individuals and link morphologically informed data via imaging[21, 24]. We can now retrieve multiple layers of data without cultivation and hence are able to somewhat overcome the barrier of the uncultivable. This progress enables us to work on an abundance of questions without being able to directly manipulate the symbiotic system. Consequently, we are currently not limited by the technical possibilities of investigation, but by the manpower to conduct the experiments and analyze the wealth of data.

Aims of this thesis

This thesis aims to provide an overview of the diverse Alphaproteobacteria that associate with gutless oligochaetes, especially with respect to their distribution, their taxonomy and their potential roles in their symbiont communities. Before I started to investigate Alphaproteobacteria in gutless oligochaetes, only few representative associations were published and little has been described on their role and relevance. Findings by Mankowski *et al.* concerning the whole gutless oligochaete community composition sparked the idea for this thesis and tremendously shaped its directions throughout my research[30]. Only the impressive collaborative effort of sample collection and the ongoing development of cost-efficient high throughput metagenomic studies allowed us to shed light on such a large variety of host species and symbiont communities. The data at hand enables us to have a broader overview on this animal clade and its symbionts, and draw conclusions for a range of species. Here, I addressed two main questions in more detail. I focused on one of the major but understudied symbiont clades – the Alphaproteobacteria. What is the diversity of Alphaproteobacteria that associate

with gutless oligochaetes? And which roles have Alphaproteobacteria in symbiosis with gutless oligochaetes? I used metagenomics, metatranscriptomics and metaproteomics approaches on the computational side, and fluorescence *in situ* hybridization as microscopy based method on the other side to gain answers and insights to these questions. In summary, I characterize the diversity, I put the spotlight on two symbionts that could not be more distinct and I point towards future directions on which and how to study specific gutless oligochaete host-symbiont communities.

Question 1: What is the diversity of Alphaproteobacteria that associate with gutless oligochaetes?

Several Alphaproteobacteria that associate with gutless oligochaetes have been published in independent manuscripts with first observations from 1999[36, 38]. Alphaproteobacteria have often been underrepresented in the investigations on gutless oligochaetes, resulting in no comparative studies that have been published. Mankowski *et al.* conducted the first overarching study including the full diversity of Alphaproteobacteria associating with gutless oligochaetes we know to date[30]. However, the frame of this study did not allow an in-depth characterization of the single clades of Alphaproteobacteria. **Chapter I** of this thesis provides this missing general overview on the single clades and gives more details on global abundance and distribution patterns as well as the scope of functions they bring into the symbiosis. Chapter I furthermore highlights clades of alphaproteobacterial symbionts where future research seems especially worthwhile.

Question 2: Which roles have Alphaproteobacteria in symbiosis with gutless oligochaetes?

Olavius algarvensis is the most well studied gutless oligochaete. It's gamma- and deltaproteobacterial symbionts have been studied intensively for their metabolism and interactions both with the hosts and among themselves. *O. algarvensis* does not host subcuticular Alphaproteobacteria. *O. ilvae* co-localizes with *O. algarvensis* around Elba and has a similar symbiont community. This might be one of the reasons why Alphaproteobacteria have not been the focus of investigation with these two easy to access gutless oligochaetes. In **Chapter II** I close this gap and describe an Alphaproteobacterium in the host species *O. ilvae* that, despite previous research on its community composition, has been overlooked until the advent of metagenomics[34]. Based on complementing meta-omics approaches I describe its metabolism and develop a hypothesis on its role in its host system. I also show with

fluorescence *in situ* hybridization that this Alphaproteobacterium is co-localized with the other subcuticular symbionts. In **Chapter III**, I describe a rather unusual member of the gutless oligochaete community that challenges the concept of mutualistic symbiosis in gutless oligochaetes. The investigation of large metagenomic datasets for population structure investigations led to the discovery of a novel alphaproteobacterial phylotype in *O. algarvensis*. According to my results, this Alphaproteobacterium is distinct to all other symbionts described to date as it has a drastically reduced genome, appears in low abundance and is likely not localized in the subcuticular symbiont layer. In addition to reconstructing the metabolism, I used metagenomics based analyses to describe its abundance patterns and draw conclusions on its potential role in the host. Taken together, Chapter II and III represent two specific extreme cases of the diverse roles Alphaproteobacteria can play in the gutless oligochaete symbiosis.

References

1. De Bary A. Die Erscheinung der Symbiose: Vortrag gehalten auf der Versammlung deutscher Naturforscher und Aerzte zu Kassel: Trübner; 1879.
2. Kaur R, Shropshire JD, Cross KL, Leigh B, Mansueto AJ, Stewart V, et al. Living in the endosymbiotic world of *Wolbachia*: A centennial review. Cell Host & Microbe. 2021.
3. Wrede C, Dreier A, Kokoschka S, Hoppert M. Archaea in symbioses. Archaea. 2012;2012.
4. Husnik F, Tashyreva D, Boscaro V, George EE, Lukeš J, Keeling PJ. Bacterial and archaeal symbioses with protists. Current Biology. 2021;31(13):R862-R77.
5. Sogin EM, Kleiner M, Borowski C, Gruber-Vodicka HR, Dubilier N. Life in the dark: Phylogenetic and physiological diversity of chemosynthetic symbioses. Annual review of microbiology. 2021;75:695-718.
6. Trench R. The cell biology of plant-animal symbiosis. Annual Review of Plant Physiology. 1979;30(1):485-531.
7. McFall-Ngai MJ. The importance of microbes in animal development: Lessons from the squid-*Vibrio* symbiosis. Annual review of microbiology. 2014;68:177-94.
8. Moran NA, Tran P, Gerardo NM. Symbiosis and insect diversification: An ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. Applied and Environmental Microbiology. 2005;71(12):8802-10.
9. Sagan L. On the origin of mitosing cells. Journal of theoretical biology. 1967;14(3):225-IN6.
10. Fan L, Wu D, Goremykin V, Xiao J, Xu Y, Garg S, et al. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. Nature ecology & evolution. 2020;4(9):1213-9.
11. Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJ. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature. 2018;557(7703):101-5.
12. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology. 2012;10(1):13-26.
13. Moran NA, Bennett GM. The tiniest tiny genomes. Annual review of microbiology. 2014;68:195-215.
14. Perreau J, Moran NA. Genetic innovations in animal–microbe symbioses. Nature Reviews Genetics. 2022;23(1):23-39.
15. Bennett GM, Moran NA. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome biology and evolution. 2013;5(9):1675-88.
16. Brune A. Symbiotic digestion of lignocellulose in termite guts. Nature Reviews Microbiology. 2014;12(3):168-80.
17. Jannasch HW, Wirsen CO. Chemosynthetic primary production at East Pacific sea floor spreading centers. Bioscience. 1979;29(10):592-8.
18. Karl D, Wirsen C, Jannasch H. Deep-sea primary production at the Galapagos hydrothermal vents. Science. 1980;207(4437):1345-7.

19. Corliss JB, Dymond J, Gordon LI, Edmond JM, von Herzen RP, Ballard RD, et al. Submarine thermal springs on the Galapagos Rift. Science. 1979;203(4385):1073-83.
20. Dubilier N, Bergin C, Lott C. Symbiotic diversity in marine animals: The art of harnessing chemosynthesis. Nature Reviews Microbiology. 2008;6(10):725-40.
21. Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. Proceedings of the National Academy of Sciences. 2012;109(19):E1173-E82.
22. Seah BK, Volland J-M, Leisch N, Schwaha T, Dubilier N, Gruber-Vodicka HR. *Kentrophoros magnus* sp. nov.(Ciliophora, Karyorelictea), a new flagship species of marine interstitial ciliates. bioRxiv. 2020.
23. Gruber-Vodicka HR, Dirks U, Leisch N, Baranyi C, Stoecker K, Bulgheresi S, et al. *Paracatenula*, an ancient symbiosis between thiotrophic Alphaproteobacteria and catenulid flatworms. Proceedings of the National Academy of Sciences. 2011;108(29):12078-83.
24. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature. 2006;443(7114):950-5.
25. Sogin EM, Leisch N, Dubilier N. Chemosynthetic symbioses. Current Biology. 2020;30(19):R1137-R42.
26. Erséus C. *Inanidrilus bulbosus* gen. et sp. n., a marine tubificid (Oligochaeta) from Florida, USA. Zoologica Scripta. 1979;8(1-4):209-10.
27. Giere O. Studies on marine Oligochaeta from Bermuda, with emphasis on new *Phallodrilus* species (Tubificidae). Cahiers de Biologie Marine. 1979;20:301-14.
28. Felbeck H, Liebezeit G, Dawson R, Giere O. CO_2 fixation in tissues of marine oligochaetes (*Phallodrilus leukodermatus* and *P. planus*) containing symbiotic, chemoautotrophic bacteria. Marine Biology. 1983;75(2):187-91.
29. Dubilier N, Blazejak A, Rühland C. Symbioses between bacteria and gutless marine oligochaetes. Molecular basis of symbiosis. 2005:251-75.
30. Mankowski A, Kleiner M, Erséus C, Leisch N, Sato Y, Volland J-M, et al. Highly variable fidelity drives symbiont community composition in an obligate symbiosis. bioRxiv. 2021.
31. Dubilier N, Mülders C, Ferdelman T, de Beer D, Pernthaler A, Klein M, et al. Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. Nature. 2001;411(6835):298-302.
32. Giere O, Conway N, Gastrock G, Schmidt C. 'Regulation' of gutless annelid ecology by endosymbiotic bacteria. Marine Ecology Progress Series. 1991:287-99.
33. Kleiner M, Wentrup C, Holler T, Lavik G, Harder J, Lott C, et al. Use of carbon monoxide and hydrogen by a bacteria–animal symbiosis from seagrass sediments. Environmental microbiology. 2015;17(12):5023-35.
34. Ruehland C, Blazejak A, Lott C, Loy A, Erséus C, Dubilier N. Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. Environmental Microbiology. 2008;10(12):3404-16.
35. Wippler J, Kleiner M, Lott C, Gruhl A, Abraham PE, Giannone RJ, et al. Transcriptomic and proteomic insights into innate immunity and adaptations to a symbiotic lifestyle in the gutless marine worm *Olavius algarvensis*. BMC genomics. 2016;17(1):1-19.
36. Dubilier N, Amann R, Erséus C, Muyzer G, Park S, Giere O, et al. Phylogenetic diversity of bacterial endosymbionts in the gutless marine oligochete *Olavius loisae* (Annelida). Marine Ecology Progress Series. 1999;178:271-80.
37. Blazejak A, Erséus C, Amann R, Dubilier N. Coexistence of bacterial sulfide oxidizers, sulfate reducers, and spirochetes in a gutless worm (Oligochaeta) from the Peru margin. Applied and Environmental Microbiology. 2005;71(3):1553-61.
38. 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. Applied and Environmental Microbiology. 2006;72(8):5527-36.
39. Kleiner M, Dong X, Hinzke T, Wippler J, Thorson E, Mayer B, et al. Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. Proceedings of the National Academy of Sciences. 2018;115(24):E5576-E84.
40. Spain AM, Krumholz LR, Elshahed MS. Abundance, composition, diversity and novelty of soil Proteobacteria. The ISME Journal. 2009;3(8):992-1000.
41. Jackson CR, Millar JJ, Payne JT, Ochs CA. Free-living and particle-associated bacterioplankton in large rivers of the Mississippi River Basin demonstrate biogeographic patterns. Applied and environmental microbiology. 2014;80(23):7186-95.
42. Van Goethem MW, Makhalanyane TP, Cowan DA, Valverde A. Cyanobacteria and Alphaproteobacteria May Facilitate Cooperative Interactions in Niche Communities. Frontiers in Microbiology. 2017;8.

43. Donachie SP, Bowman JP, Alam M. *Nesiotobacter exalbescens* gen. nov., sp. nov., a moderately thermophilic alphaproteobacterium from an Hawaiian hypersaline lake. International Journal of Systematic and Evolutionary Microbiology. 2006;56(3):563-7.
44. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11 clade dominates ocean surface bacterioplankton communities. Nature. 2002;420(6917):806-10.
45. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature biotechnology. 2018;36(10):996-1004.
46. Hördt A, López MG, Meier-Kolthoff JP, Schleuning M, Weinhold L-M, Tindall BJ, et al. Analysis of 1,000+ type-strain genomes substantially improves taxonomic classification of Alphaproteobacteria. Frontiers in microbiology. 2020;11:468.
47. Muñoz-Gómez SA, Hess S, Burger G, Lang BF, Susko E, Slamovits CH, et al. An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. Elife. 2019;8:e42535.
48. Driscoll T, Gillespie JJ, Nordberg EK, Azad AF, Sobral BW. Bacterial DNA sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont. Genome biology and evolution. 2013;5(4):621-45.
49. Gruber-Vodicka HR, Leisch N, Kleiner M, Hinzke T, Liebeke M, McFall-Ngai M, et al. Two intracellular and cell type-specific bacterial symbionts in the placozoan *Trichoplax* H2. Nature Microbiology. 2019;4(9):1465-74.
50. Baker LJ, Reich HG, Kitchen SA, Grace Klinges J, Koch HR, Baums IB, et al. The coral symbiont *Candidatus* Aquarickettsia is variably abundant in threatened Caribbean acroporids and transmitted horizontally. The ISME Journal. 2022;16(2):400-11.
51. McCutcheon JP, McDonald BR, Moran NA. Origin of an alternative genetic code in the extremely small and GC–rich genome of a bacterial symbiont. PLoS genetics. 2009;5(7):e1000565.
52. Jensen M, Wippler J, Kleiner M. Evaluation of RNA later as a field-compatible preservation method for metaproteomic analyses of bacterium-animal symbioses. Microbiology spectrum. 2021;9(2):e01429-21.
53. Raley C, Munroe D, Jones K, Tsai Y-C, Guo Y, Tran B, et al. Preparation of next-generation DNA sequencing libraries from ultra-low amounts of input DNA: Application to single-molecule, real-time (SMRT) sequencing on the Pacific Biosciences RS II. bioRxiv. 2014:003566.
54. Gruber-Vodicka HR, Seah BK, Pruesse E. phyloFlash: Rapid small-subunit rRNA profiling and targeted assembly from metagenomes. Msystems. 2020;5(5).
55. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology. 2014;15(3):1-12.
56. Moraru C, Lam P, Fuchs BM, Kuypers MM, Amann R. GeneFISH – an *in situ* technique for linking gene presence and cell identity in environmental microorganisms. Environmental microbiology. 2010;12(11):3057-73.
57. Bi S, Yue S, Zhang S. Hybridization chain reaction: a versatile molecular tool for biosensing, bioimaging, and biomedicine. Chemical Society Reviews. 2017;46(14):4281-98.
58. Geier B, Oetjen J, Ruthensteiner B, Polikarpov M, Gruber-Vodicka HR, Liebeke M. Connecting structure and function from organisms to molecules in small-animal symbioses through chemo-histo-tomography. Proceedings of the National Academy of Sciences. 2021;118(27).

:all

15 distinct clades of Alphaproteobacteria make up the most abundant and diverse symbiont class in globally distributed gutless oligochaete hosts. These symbionts have various taxonomic backgrounds and a great potential to extend the nutritional access of their hosts in their environments.

Chapter 1 | Alphaproteobacteria associated with gutless oligochaetes

Alphaproteobacteria are the most abundant and diverse class of symbionts across the marine gutless oligochaete (Oligochaeta, Annelida) diversity

Tina Enders[1], Anna Mankowski[1,2], Marlene Jensen[3], Manuel Kleiner[3,] Nicole Dubilier[1], Harald R. Gruber-Vodicka[1]

[1] Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

[2] Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

[3] Department of Plant & Microbial Biology, North Carolina State University, Raleigh 27695, North Carolina, USA

Corresponding authors:
Nicole Dubilier, ndubilie@mpi-bremen.de,
Harald R. Gruber-Vodicka, hgruber@mpi-bremen.de

Competing interest

The authors declare no competing financial interests.

Abstract

Gutless oligochaetes are marine worms that lost their digestive and excretory organs and instead rely on the nutrition and waste product recycling from a set of subcuticular bacterial symbionts. A recent metagenomic study on 64 host species from 17 global sampling sites showed that Alphaproteobacteria are the most diverse and abundant class of symbionts of gutless oligochaetes. However, little is known on the diversity and function of these Alphaproteobacteria. Here, we analyzed the phylogenetic diversity and biogeographic distribution of alphaproteobacterial symbionts in gutless oligochaetes and provide first insights in their metabolic potential. Alphaproteobacteria reside in more than half of the studied gutless oligochaete host species at nearly all locations sampled. The detected clades form mostly novel families or genera in the orders Rhizobiales, Rhodobacterales and Rhodospirillales and are distantly related to published bacteria. Alphaproteobacteria are a functionally diverse class of bacteria and could provide promising functions to extend the nutrient and thus habitat spectrum of gutless oligochaetes. This study provides an overview on current knowledge of the global distribution and taxonomy of Alphaproteobacteria in symbiosis with gutless oligochaetes and points to clades on which future research efforts could be focused.

Keywords

Alphaproteobacteria, marine bacterial-animal symbiosis, gutless oligochaete, diversity

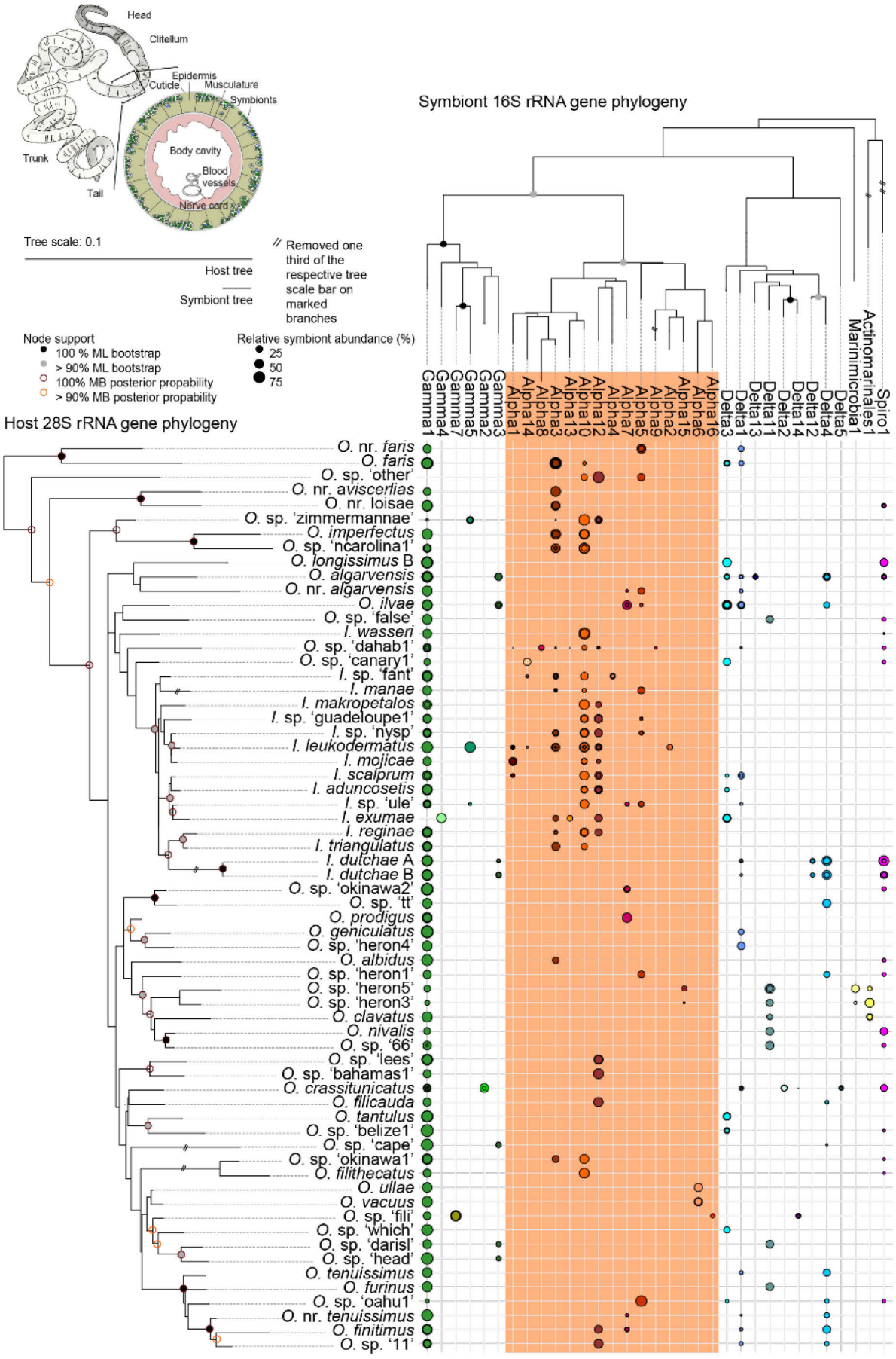
Introduction

Gutless oligochaetes are small marine annelids from shallow water sediments that have no digestive and excretory organs. Instead, they rely on a chemosynthetic community of diverse and yet host species specific set of bacteria to provide nutrition and recycle waste products[1, 2]. More than 100 gutless oligochaete species have been described in the past, and they can dominate the interstitial animal fauna at their collection sites in tropical to warm temperate habitats[3]. 16S ribosomal RNA (rRNA) based studies indicated that these highly successful hosts largely share their main gammaproteobacterial symbiont *Ca.* Thiosymbion that often makes up the bulk of the symbiont biomass, but can host a variety of secondary symbionts[4-10]. However, research into symbiont composition and the roles of the individual partners in the multipartite symbiosis has focused mainly on one specific host species, *Olavius algarvensis*, from seagrass meadow associated habitats around the island of Elba, Italy[1, 11-14]. In *Olavius algarvensis*, the chemoautotrophic *Ca.* Thiosymbion interacts with the second major symbiont

group, sulfur-reducing Deltaproteobacteria via a syntrophic sulfur cycling[1]. These Deltaproteobacteria live off organic substrates, including host waste products[11]. Spirochaetia were also regularly detected in *Olavius algarvensis* but their role and function in the *Olavius algarvensis* symbiosis remains elusive to date[8, 14]. A recent metagenomics analysis on a large dataset of 64 gutless oligochaetes species from 17 globally distributed sampling sites challenges the overall importance of Deltaproteobacteria in gutless oligochaetes (Figure 1)[2]. This untargeted approach based on 16S rRNA gene sequences revealed that the main Gammaproteobacterium is indeed member of the symbiosis in all but one of the studied host species[10, 15]. Furthermore, Deltaproteobacteria make up a diverse secondary symbiont group with nine distinct clades from currently known host species, which all have been previously detected in individual studies[10]. But overall, Alphaproteobacteria make up the most diverse and the most abundant secondary symbiont group around the globe[2]. Only five of the 15 alphaproteobacterial genera had been previously detected, largely in gutless oligochaete species from the Caribbean (Figure 2)[2, 10]. Investigations on Alphaproteobacteria in association with gutless oligochaetes have mainly been based on 16S rRNA genes and fluorescence *in situ* hybridization (FISH)[5, 7]. So far, no genomic data was available to describe the metabolism of the Alphaproteobacteria and their resulting function in the gutless oligochaete symbiosis.

Here we use a dataset of 233 gutless oligochaete metagenomes to characterize the different clades of Alphaproteobacteria that are associated with this globally sampled host diversity[2, 15]. We describe the geographic ranges and the host ranges of all clades by integrating our datasets with publicly available data, reconstruct their taxonomic affiliations within the class of Alphaproteobacteria and provide a comparative analysis on the possible metabolic roles.

Figure 1 (next page): Alphaproteobacteria are the most abundant and diverse symbiont clade associating with 64 globally sampled gutless oligochaetes host species.
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 non-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. The alphaproteobacterial clade is highlighted in orange. The figure was modified from Mankowski *et al.* 2021[2].


Head
Clitellum
Epidermis
Cuticle
Musculature
Symbionts
Body cavity
Blood vessels
Nerve cord
Trunk
Tail
Tree scale: 0.1
Host tree
Symbiont tree
// Removed one third of the respective tree scale bar on marked branches
Node support
100 % ML bootstrap
> 90% ML bootstrap
100% MB posterior propability
> 90% MB posterior propability
Relative symbiont abundance (%)
25
50
75
Symbiont 16S rRNA gene phylogeny
Host 28S rRNA gene phylogeny
Gamma1
Gamma4
Gamma7
Gamma5
Gamma2
Gamma3
Alpha1
Alpha14
Alpha8
Alpha3
Alpha13
Alpha10
Alpha12
Alpha4
Alpha7
Alpha5
Alpha9
Alpha2
Alpha15
Alpha6
Alpha16
Delta3
Delta1
Delta13
Delta11
Delta2
Delta14
Delta12
Delta4
Delta5
Marinimicrobia1
Actinomarinales1
Spiro1
O. nr. faris
O. faris
O. sp. 'other'
O. nr. aviscerlias
O. nr. loisae
O. sp. 'zimmermannae'
O. imperfectus
O. sp. 'ncarolina1'
O. longissimus B
O. algarvensis
O. nr. algarvensis
O. ilvae
O. sp. 'false'
I. wasseri
O. sp. 'dahab1'
O. sp. 'canary1'
I. sp. 'fant'
I. manae
I. makropetalos
I. sp. 'guadeloupe1'
I. sp. 'nysp'
I. leukodermatus
I. mojicae
I. scalprum
I. aduncosetis
I. sp. 'ule'
I. exumae
I. reginae
I. triangulatus
I. dutchae A
I. dutchae B
O. sp. 'okinawa2'
O. sp. 'tt'
O. prodigus
O. geniculatus
O. sp. 'heron4'
O. albidus
O. sp. 'heron1'
O. sp. 'heron5'
O. sp. 'heron3'
O. clavatus
O. nivalis
O. sp. '66'
O. sp. 'lees'
O. sp. 'bahamas1'
O. crassitunicatus
O. filicauda
O. tantulus
O. sp. 'belize1'
O. sp. 'cape'
O. sp. 'okinawa1'
O. filithecatus
O. ullae
O. vacuus
O. sp. 'fili'
O. sp. 'which'
O. sp. 'darisl'
O. sp. 'head'
O. tenuissimus
O. furinus
O. sp. 'oahu1'
O. nr. tenuissimus
O. finitimus
O. sp. '11'

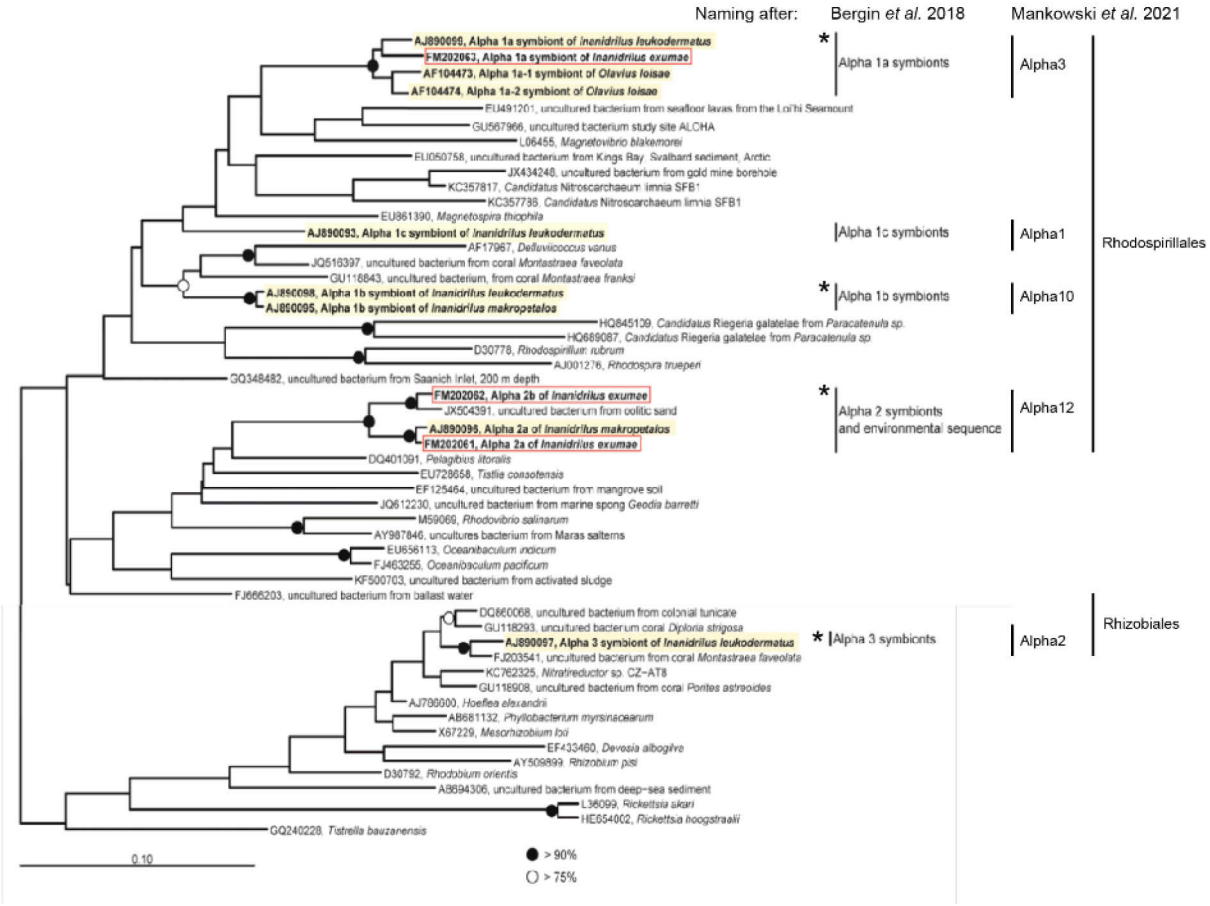


Figure 2: Five clades of Alphaproteobacteria associating with four gutless oligochaete host species from the Caribbean were published before 2021.
Five clades of Alphaproteobacteria hosted by gutless oligochaetes have been detected on 16S rRNA level prior to a global metagenome study by Mankowski *et al.*[2]. These Alphaproteobacteria were hosted by four host species from the Caribbean and belong to the orders of Rhodospirillales and Rhizobiales. Their affiliation and naming in the different studies from 2018 and 2021 are stated beside the 16S rRNA gene phylogenetic tree[2, 10]. Clades labelled with an asterisk have been visualized with fluorescence *in situ* hybridization in the subcuticular symbiont layer[7, 10]. Sequences from gutless oligochaete symbionts are highlighted in yellow and red boxes. The consensus tree shown is based on maximum likelihood analysis. Scale bars represent 10% estimated phylogenetic divergence. The tree was modified from Bergin *et al.* 2018.

Methods and materials

Sample collection, processing and metagenomic analysis

All gutless oligochaete samples were treated as described in detail by Mankowski *et al.*[2, 15]. In summary, 233 specimens of gutless oligochaetes were collected at 17 globally distributed shallow water sampling sites. DNA of single worm individuals was extracted, sequencing and quality control were done at the DOE JGI (Walnut Creek, California, USA) and the Max Planck Genome Centre (Cologne, Germany). Reads were quality trimmed and filtered and assembled to obtain symbiont metagenomes-assembled genomes (MAGs). Host species and symbiont

species were determined based on ribosomal and mitochondrial marker genes. Symbionts were clustered in functional genera with 95% similarity of the 16S rRNA gene sequence.

Analysis of abundance and distribution

We aimed to determine the global distribution and abundance of Alphaproteobacteria in association with gutless oligochaetes. Therefore, we determined occurrences of Alphaproteobacteria in distinct host species and sampling locations based on presence of detected 16S rRNA gene sequences[2]. Sampling sites and occurrences of Alphaproteobacteria at sampling sites and in host species were plotted in RStudio (RStudio v1.4.1106, R v4.0.4) using a variety of packages. Final figures were edited with Adobe illustrator v25.03.

Phylogenetic reconstruction of symbiont taxonomy

We used the metagenomic pipeline as described in detail by Mankowski *et al.* (github.com/amankowski/MG-processing_from-reads-to-bins)[15] for the 233 gutless oligochaete metagenomes and detected 216 Alpha bins in 140 specimens. For phylogenetic placement, we used MAGs with the best statistics from a multi-tool binning approach per host species and location. Available MAGs of all Alphaproteobacteria associated with gutless oligochaetes were placed in the GTDB tree at version 1.5.0 with the R202 reference data using the GTDB-Tk software[16]. Trees were visualized and annotated with iTOL and final figures edited with Adobe illustrator v25.03 (www.adobe.com/products/illustrator)[17].

We calculated the average nucleotide identity (ANI) of the distinct clades of Alphaproteobacteria associating with gutless oligochaetes to gain insight into the amount of species per clade and their clustering (enve-omics.ce.gatech.edu/g-matrix/)[18].

Profile analysis of clusters of orthologous groups (COGs)

We analyzed the metabolic potential of MAGs from gutless oligochaete associated Alphaproteobacteria based on clusters of orthologous groups (COG) categories that represent larger functional classifications. We obtained COG profiles for the different alphaproteobacterial clades and representatives of 149 alphaproteobacterial families with eggNOG-mapper v2.1.6 and diamond as protein aligner[19-22]. To obtain alphaproteobacterial family representatives, the bacterial tree bac120_r86.2 was downloaded from the GTDB database (gtdb.ecogenomic.org/downloads) and trimmed to the Alphaproteobacteria and Magnetococcia outgroup in iTOL[17]. Branch lengths for each leaf were extracted with Newick utils (nw_distance), the average branch length was calculated for each family, and for each

family the genome that was closest to the average branch length was chosen for further phylogenomic analyses (genomes listed in zenodo.org/record/6514058)[23]. For the 16S rRNA gene tree, the corresponding 16S rRNA gene sequences were extracted from the bac120_ssu_r86.2.fna sequence collection provided by GTDB using faSomeRecords (github.com/santiagosnchez/faSomeRecords). Relative abundances of genes of the MAGs in larger metabolic categories was quantified and plotted with non-metric multidimensional scaling (NMDS) and principal component analysis (PCA) calculated in RStudio (RStudio v1.4.1106, R v4.0.4) using a variety of packages[24]. Final figures were edited with Adobe illustrator v25.03.

Protein extraction and peptide preparation

We extracted proteins from five specimens of five marine gutless oligochaete host species (*O. ilvae, I. aduncosetis, I. leukodermatus, I. reginae,* and *I. scalprum)*. We conducted a tryptic protein digestion following the filter-aided sample preparation (FASP) protocol, adapted from Wisniewski *et al*., 2009, for all samples[25]. We added 60 µL SDT-lysis buffer (4% , w/v, SDS, 100 mM Tris-HCl pH 7.6, 0.1 M DTT) and heated samples to 95 °C for 10 min. To minimize sample loss, we did not do the 5 minutes centrifugation step at 21,000g as described in the original FASP protocol[25]. Instead, only a short spin down was conducted. Subsequently, we mixed the lysate with 400 µL UA solution (8 M urea in 0.1 M Tris/HCl pH 8.5) in a 10 kDa MWCO 500 µL centrifugal filter unit (VWR International) and centrifuged the mixture at 14,000g for 20 min. Next, we added 200 µL of UA solution and centrifugal filter spun again at 14,000g for 20 min. Subsequently, we added 100 µL of IAA solution (0.05 M iodoacetamide in UA solution) and incubated samples at 22 °C for 20 min in the dark. We removed the IAA solution by centrifugation following three washing steps with 100 µL of UA solution. Subsequently, filters were washed three times with 100 µL of ABC buffer (50 mM ammonium bicarbonate). We added 0.5 µg of Pierce MS grade trypsin (Thermo Fisher Scientific) in 40 µL of ABC buffer to each filter. We incubated filters overnight in a wet chamber at 37 °C. The next day, we eluted the peptides by centrifugation at 14,000g for 20 min followed by addition of 50 µL of 0.5 M NaCl and another centrifugation step. Peptides were quantified using the Pierce MicroBCA Kit (Thermo Fisher Scientific) following the instructions of the manufacturer. For the 2D-LC-MS/MS analysis, we desalted the peptides with Sep-Pak C18 Plus Light Cartridges (Waters). Acetonitrile from the peptide elution step was exchanged for 0.1% formic acid (v/v) using a centrifugal vacuum concentrator. The desalting step was

necessary to enable binding of peptides to the SCX column during sample loading for the 2D-LC method.

1D-LC-MS/MS

All samples were analyzed by 1D-LC-MS/MS as described in Hinzke *et al*. 2019[26]. 1500 ng of peptides were loaded in loading solvent A (2% acetonitrile, 0.05% trifluoroacetic acid) onto a 300 µm i.d. x 5 mm trap cartridge column packed with Acclaim PepMap100 C18, 5 µm, 100 Å (Thermo Fisher, 160454) using an UltiMate 3000 RSLCnano Liquid Chromatograph (Thermo Fisher Scientific). The trap was connected to a 75 µm x 75 cm analytical EASY-Spray column packed with PepMap RSLC C18, 2 µm material (Thermo Fisher Scientific), which was heated to 60 °C via the integrated heating module. The analytical column was connected via an Easy-Spray source to a Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific). Peptides were separated on the analytical column at a flow rate of 225 nl min^{-1} using a 460 min gradient. The gradient went from 98% buffer A (0.1% formic acid) to 31% buffer B (0.1% formic acid, 80% acetonitrile) in 364 min, then from 31% to 50% buffer B in 76 min and ending with 20 min at 99% buffer B. Eluting peptides were ionized via electrospray ionization (ESI) and analyzed in Q Exactive HF-X. Full scans were acquired in the Orbitrap at 60,000 resolution. The 15 most abundant precursor ions were selected in a data dependent manner, isolated with the quadrupole with a 1.2 m/z isolation window size, fragmented in the HCD cell with a NCE of 25, and measured in the Orbitrap at 7,500 resolution. The mass (m/z) 445.12003 was used as lock mass as described in Olsen *et al*. 2005[27]. Lock mass use was set to 'best'. Singly charged ions were excluded from MS/MS analysis. Dynamic exclusion was set to 30 s. On average 209,160 MS/MS spectra were acquired per sample with the 460 min gradient.

2D-LC-MS/MS

For the 2D-LC-MS/MS runs, we used the same instrumentation as for the 1D run. We followed the LC methods described in Hinzke *et al*. 2019 for the pH plug runs[26]. We loaded 9000 ng of peptide mixture with loading solvent B (10 mM phosphate buffer pH 3, 20% ACN and 600 mM NaCl) onto a 10 cm, 300 µm Micro SCX LC column (Thermo Fisher Scientific) at a flow rate of 300 µl min^{-1}. The specific plumbing scheme used in the RSLCnano corresponded to the standard set up recommended by the manufacturer for on-line 2D pH plug separations. During loading, the C18 trap (see above) was in-line downstream of the SCX column to capture peptides that did not bind to the SCX column (breakthrough). After loading, the C18

pre-column was switched in-line with the 75 µm x 75 cm analytical column (same as for 1D) and the breakthrough was separated using an 120 min[26]. Subsequently, elution of peptides from the SCX to the C18 trap (same as for 1D-LC) took place by injection of 20 µl of 8 different pH plugs with increasing pH (CTIBiphase buffers, Column Technology, Inc.) from the autosampler. The C18 trap was then again switched in-line with the analytical column and peptides separated with gradients of eluent A and B. Data acquisition in the mass spectrometer was done as described by Mordant and Kleiner, 2021[28].

Protein identification and quantification

We used a custom database which contained 1,362,363 protein sequences, including protein sequences predicted from host transcriptomes, non-redundant host sequences from the closely related species *Olavius algarvensis* and symbiont protein sequences predicted from metagenome assembled genomes, as well as a cRAP protein sequence database (http://www.thegpm.org/crap) of common laboratory contaminants. The database is available from the PRIDE repository (see data availability). Searches of the MS/MS spectra against this database were performed with the Sequest HT node in Proteome Discoverer version 2.2.0.388 (Thermo Fisher Scientific) as described in Jensen *et al*., 2021[29]. The following parameters were used: trypsin (full), maximum two missed cleavages, 10 ppm precursor mass tolerance, 0.1 Da fragment mass tolerance and maximum of 3 equal dynamic modifications per peptide, namely: oxidation on N (+ 15.995 Da), carbamidomethyl on C (+ 57.021 Da) and acetylation on the protein N terminus (+ 42.011 Da). False discovery rates (FDRs) for peptide spectral matches (PSMs) were calculated and filtered using the Percolator Node in Proteome Discoverer. Percolator was run with a maximum delta Cn 0.05, a strict target FDR of 0.01, a relaxed target FDR of 0.05 and validation based on q-value. The Protein FDR Validator Node in Proteome Discoverer was used to calculate q-values for inferred proteins based on the results from a search against a target-decoy database. Proteins with a q-value of <0.01 were categorized as high-confidence identifications and proteins with a q-value of 0.01–0.05 were categorized as medium-confidence identifications. Search results for all samples were combined into a multiconsensus report in Proteome Discoverer and only proteins identified with medium or high confidence were retained, resulting in an overall protein-level FDR of 5%. For protein quantification, normalized spectral abundance factors (NSAFs) were calculated per species and multiplied by 100, to give the relative protein abundance in %[30]. The mass spectrometry metaproteomics data and protein sequence database have been deposited to the

ProteomeXchange Consortium via the PRIDE (Vizcaino *et al*. 2016) partner repository with the dataset identifier XXX (access for reviewers: Username: Password:).

Results and discussion

Alphaproteobacteria - A diverse and global class of symbionts

We used single animal metagenomics and 16S rRNA gene sequences based profiling to track the 15 alphaproteobacterial genera in 64 gutless oligochaete host species sampled from 17 globally distributed sites (for other clades see Mankowski *et al.* 2021[2]). Looking at the Alphaproteobacteria in a host centered view, they are present in all major lineages of the gutless oligochaete phylogeny and their association is not linked to host phylogeny[2]. The host species have diverse and species-specific colonization patterns with alphaproteobacterial genera. Individual host species can host from none up to six symbionts from alphaproteobacterial genera.

Figure 3: Alphaproteobacteria associate with gutless oligochaetes at 14 of 17 global regions. Based on presence patterns of 16S rRNA genes detected in metagenomes of gutless oligochaete host specimen, Alphaproteobacteria are hosted by gutless oligochaetes at 14 out of 17 globally distributed regions, except Peru, South Africa and New Caledonia.

In a symbiont centered perspective, we could detect Alphaproteobacteria in hosts sampled at 14 of the 17 regions, except in New Caledonia, South Africa, and Peru (Figure 3, Figure 4a). The Alphaproteobacteria were consistently present across samples from the Southern Pacific and the Caribbean (Figure 3). On a local scale, the Alphaproteobacteria were particularly diverse and abundant in samples from, the Great Barrier Reef (Australia), Okinawa (Japan), the Red Sea (Egypt), as well as in most sites in the Belize Barrier Reef (Belize), in Bermuda, and in Guadeloupe. In contrast, Alphaproteobacteria occurred in only few individuals from Hawaii and the Mediterranean Sea, where Deltaproteobacteria were the dominant secondary symbiont class (Figure 4B-D). Three alphaproteobacterial genera were present in a broad variety of the 64 host species at 14 of the 17 sampling locations: Alpha10 was the most widely distributed and was found in 21 host species at 15 sites from 9 regions, Alpha12 followed similar distribution patterns in 17 host species at 13 sites from 6 regions, and Alpha3 was found in 14 hosts at 14 sites from 8 regions (Figure 4C-D and Table1). The genera Alpha5, Alpha7 and Alpha1 were also globally distributed, with 5 to 9 host species and 7 to 12 sites from 4-6 regions. The remaining nine clades of Alphaproteobacteria were rather rare and only detected in few locations and host species. However, they do not show specific patterns of co-occurrence or exclusion, neither among themselves, nor with other symbiont lineages[2]. Remarkably, the abundant clades Alpha10, Alpha12 and Alpha3 seem more abundant in early branching gutless oligochaetes and the *Inanidrilus* clade endemic to the Caribbean (for detailed host and symbiont sampling locations see Supplement Figure S1 and S2).

Alphaproteobacteria can serve as the sole secondary symbiont, e.g. Alpha6 in *O. vacuus* and *O. ullae*, Alpha7 in *O. prodigus* or Alpha12 in *O.* sp. lees and *O.* sp. bahamas1, or make up the only class of secondary symbiont in addition to Cand. Thiosymbion, e.g. in a combination of Alpha10 with Alpha12 and Alpha3 in *I. reginae*. Alphaproteobacteria can also co-occur with a great variety of secondary symbionts from the Gammaproteobacteria, Deltaproteobacteria, Spirochaetia, Actinomarinales and Marinimicrobiales (Figure 1)[2]. In summary, the most abundant genera constitute six of the 10 most abundant secondary symbiont clades of a total of 32 bacterial genera associated with gutless oligochaetes, which underpins the importance of Alphaproteobacteria in these successful and widely distributed chemosynthetic hosts.

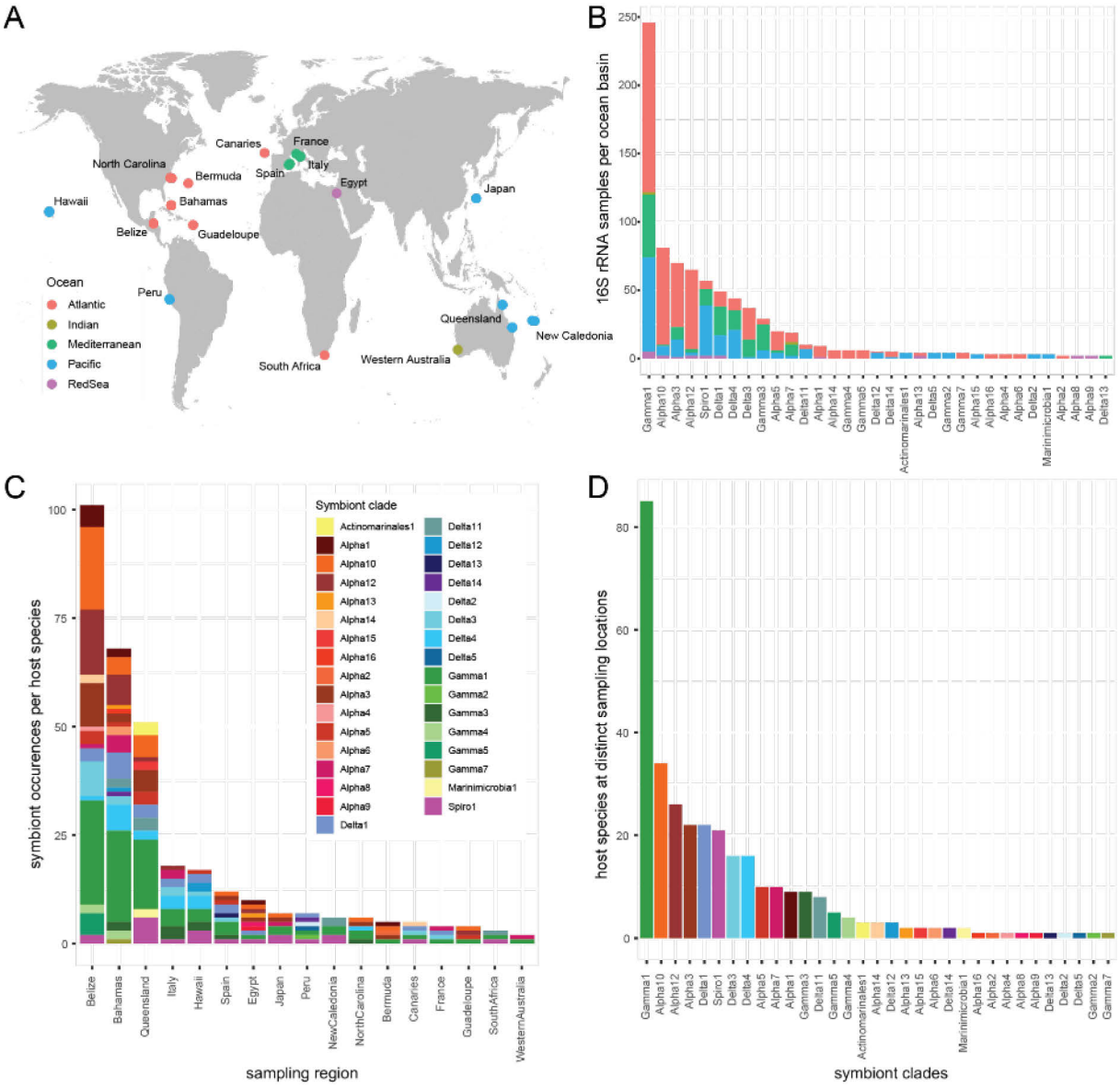


Figure 4: Alphaproteobacteria make up three of the five most abundant symbiont clades in gutless oligochaete hosts and are particularly abundant and diverse in host species from the Caribbean.

(A) Gutless oligochaete sampling regions colored according to ocean basins. (B) Occurrence of all sampled symbiont 16S rRNA gene sequences associating with gutless oligochaetes colored by ocean basin they were sampled at. Presence of 33 symbiont clades per 64 host species from 17 globally distributed sampling sites (C) plotted by region and (D) plotted by symbiont.

Table 1: Statistics on 16S rRNA genes and MAGs of 15 clades of Alphaproteobacteria associated with gutless oligochaetes.

The table lists: available 16S rRNA gene sequences and MAGs per clade; whether these clades are symbiont exclusive; the GTDB based taxonomy of the clades and the level of clade novelty; the amount of host species that host the clades; the amount of locations where the clades where detected; how many species the clades comprise based on average nucleotide identity (ANI) level.

Clade	**Samples** 16S/MAG	**Symbiont-only** GTDB	**Taxonomy** GTDB: Order;family	**Novelty level**	**Host species**	**Locations** site/region	**Species in clade** ANI
Alpha1	9/5	yes	Rhodospirillales	family	5	7/4	2
Alpha2	2/2	yes	Rhizobiales;Rhizobiaceae	genus	1	1/1	1
Alpha3	70/46	yes	Rhodospirillales;Casp-alpha2	genus	14	14/8	11
Alpha4	3/0	nd	nd	nd	1	1/1	nd
Alpha5	20/14	yes	GCA-2731375	family	9	8/6	4-6
Alpha6	3/4	yes	Rhodobacterales;Rhodobacteracea	genus	2	1/1	2
Alpha7	19/18	yes	UBA6615	family	7	12/6	6
Alpha8	2/2	yes	Rhodospirillales;Casp-alpha2	genus	1	1/1	1
Alpha9	2/2	yes	Rhizobiales;Ancalomicrobiaceae	genus	1	1/1	1
Alpha10	81/69	yes	Rhodospirillales	family	21	15/9	8
Alpha12	65/48	yes	UBA9366	family	17	13/6	14
Alpha13	4/2	yes	Rhodospirillales	family	2	2/2	1
Alpha14	6/0	nd	nd	nd	3	2/2	nd
Alpha15	3/0	nd	nd	nd	2	1/1	nd
Alpha16	3/3	yes	Rhodobacterales;Rhodobacteraceae	genus	1	1/1	1

Novel families and genera from five orders

To improve our phylogenetic and taxonomic assessments we generated metagenome assembled genomes (MAGs) using state of the art and automated assembly and binning approaches. We were able to retrieve high quality MAGs for 12 of the 15 genera detected via 16S rRNA based analysis (high quality: 80% complete and <5% contamination based on CheckM analysis. No MAGs could be obtained for Alpha4, Alpha14 and Alpha15. Using these MAGs, we then created taxonomic classifications and investigated phylogenetic relations to bacterial reference genomes based on the Genome Taxonomy Database (GTDB). According to GTDB based classification, the symbiont MAGs belong to six orders of Alphaproteobacteria, namely Rhizobiales, Rhodobacterales, and Rhodospirillales, as well as three novel orders, two related to Kiloniellales and one related to Sphingomonadales (Figure 5 and Table 1). Specifically, Alpha5 is placed at the root of Sphingomonadales in the order GCA-2731375 and Alpha7 and Alpha12 are related to the Kiloniellales *sensu strictu*, in the orders UBA6615 and UBA9366. The six most abundant alphaproteobacterial symbiont clades are all members of the Rhodospirillales or related to Kiloniellales.

All alphaproteobacterial MAGs recovered from gutless oligochaetes form symbiont genus specific clusters, and none are intermixed with other MAGs from the databases. Some of the symbiont genera are the first representatives for novel family-level clades and are only distantly related to reference MAGs (see Data availability for GTDB tree shared in iTOL). The symbiont genera are distributed throughout the alphaproteobacterial taxonomy, which suggests multiple uptakes of clades that have no prior history of symbiosis with invertebrates.

The alphaproteobacterial genera associated to gutless oligochaetes were based on 95% similarity of the 16S rRNA gene sequence that represent the genus level and were supported by our GTDB analysis (Table 1)[2]. To analyze the within genus heterogeneity, we used ANI values to cluster MAGs for each genus and detected between 1 and 14 species (ANI matrices are provided at zenodo.org/record/6514664). Most genera clustered into site specific species. Within the 14 regions many of the site specific species occurred in several host species (Table 1). Overall, we found that alphaproteobacterial symbiont clades are specific to gutless oligochaetes, represent diverse and novel families and genera and can encompass a broad diversity of species.

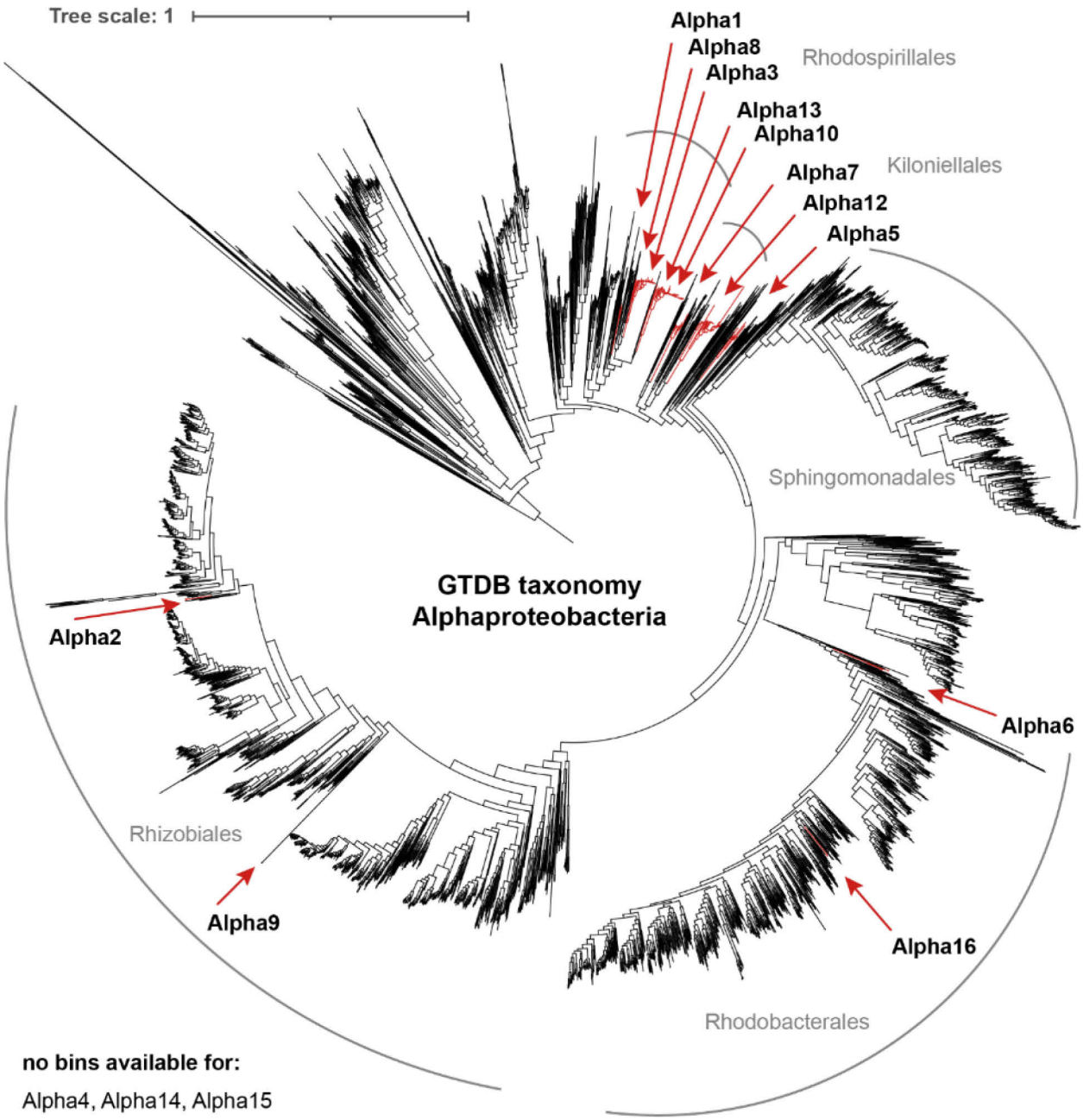


Figure 5: Alphaproteobacteria associating with gutless oligochaetes belong to five orders. Metagenome-assembled genomes of 15 clades of Alphaproteobacteria cluster within five orders based on marker genes of the GTDB taxonomy. The majority of gutless oligochaete associated Alphaproteobacteria belong to the orders of Rhodospirillales and Kiloniellales. All clades are symbiont-exclusive and often form novel genera or families. The tree scale represents the average number of substitutions per site.

Research on gutless oligochaetes focused on *O. algarvensis* from the Mediterranean Sea or Caribbean samples of the species *I. leukodermatus, I. loise, I. makropetalos and I. manae*[4-8, 10, 31, 32]. From these five host species, five of the 15 alphaproteobacterial symbiont genera we characterize here were known prior to the large metagenomic study by Mankowski *et al.*[2].

These genera had to be renamed due to inconsistencies and are, in order of decreasing host range, Alpha10, Alpha12, Alpha3, Alpha1 and Alpha2 (Figue1). It is not surprising that Alpha10, Alpha12 and Alpha3 had previously been detected, as these were also the most abundant alphaproteobacterial clades in our dataset. Alpha1 and Alpha2 in the contrary were more rare in our dataset, but both are hosted by *I. leukodermatus* that has been the focus of a dedicated study of its symbiont community[2, 7]. Other widely distributed clades such as Alpha7 have been overlooked, despite their presence in host species that have been the focus of 16S rRNA gene and FISH based analyses such as *O. ilvae* that co-occurs with the well-studied host *O. algarvensis*. Overall, the three most abundant alphaproteobacterial symbiont clades associating with gutless oligochaete have been detected by low throughput and PCR based studies. The great majority of alphaproteobacterial clades, even abundant ones like Alpha5 and Alpha1 have however escaped detection via PCR and clone library analyses and were only identified through the combination of broad taxon sampling and untargeted metagenomics[2].

Functional diversity underlies the diversity of alphaproteobacterial symbionts

We set out to compare the metabolic capabilities of the diverse alphaproteobacterial genera using the recovered MAGs by analyzing gene space attribution to larger metabolic functions. To this end, we calculated relative abundances of genes in clusters of orthologous groups (COG) categories, as they provide a unified framework to elucidate larger metabolic functions that each of the alphaproteobacterial genera might bring to the symbiosis (Figure 6, 7 and S3). In addition to within and between genera comparisons, we included representative genomes for 149 alphaproteobacterial families to assess the embedding in the larger alphaproteobacterial landscape of metabolic life-styles. Overall, the alphaproteobacterial symbionts clustered within the spectrum of metabolic profiles from the 149 representatives, indicating that no outstanding genomic streamlining had occurred in any of the symbiont genera. The most abundant alphaproteobacterial symbiont clades formed separate clusters which supports our hypothesis that they bring distinct functions to their hosts (Figure 6). Only clades of close taxonomic relation cluster together, indicating a similar functional background. The comparative analysis of relative gene abundances in COG categories of all investigated MAGs revealed that members of the Kiloniellales, i.e. Alpha7 and Alpha12, have higher relative gene abundances in carbohydrate metabolism (COG category G; Figure 7). Some widely distributed, but also some less abundant symbionts alike have relatively more genes in the category of co-enzyme metabolism (COG H; Alpha6, Alpha13 and Alpha16). Alpha1 and Alpha5 have highest relative

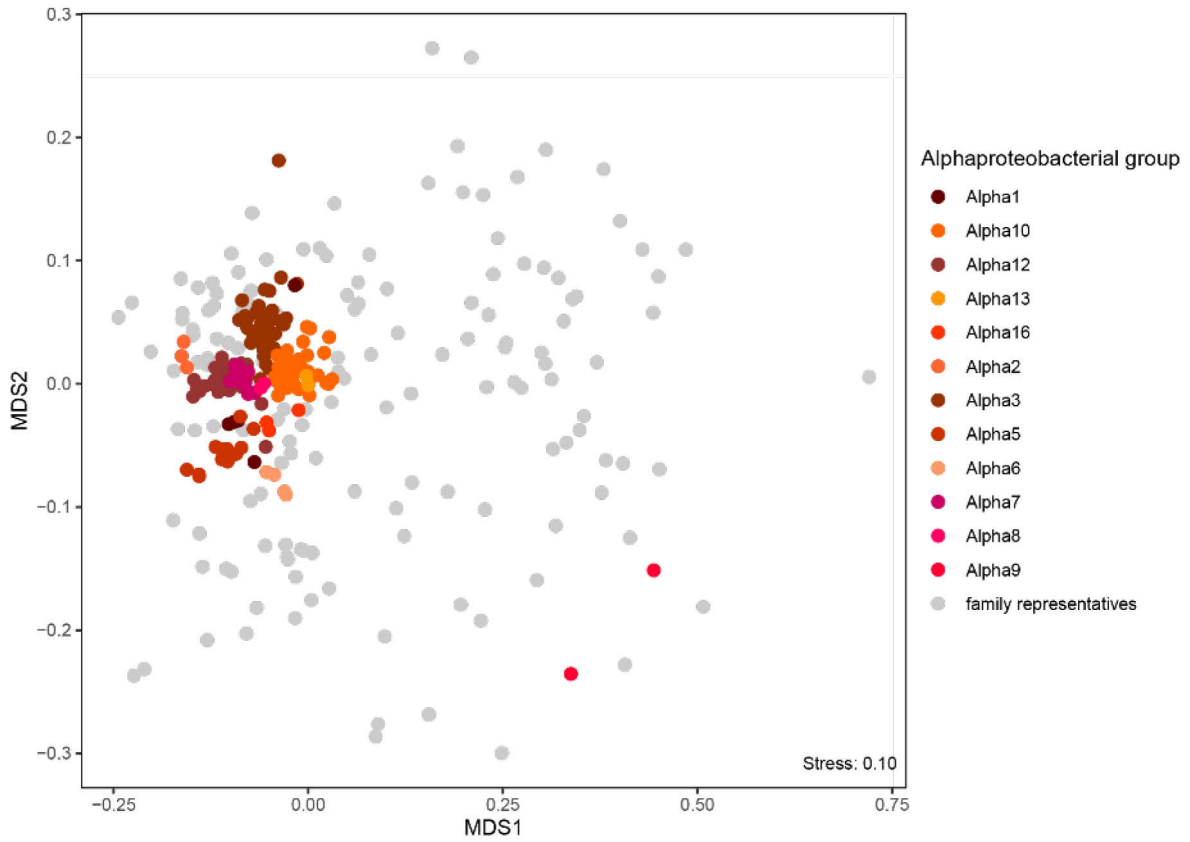


Figure 6: Alphaproteobacteria associating with gutless oligochaetes have distinct COG profiles and cluster within the diversity of alphaproteobacterial family representatives.
Non-metric dimensional scaling (NMDS) of the functional profiles based on relative abundances of COG categories from alphaproteobacterial MAGs of 233 gutless oligochaete host specimens (shades of red) and representatives of 149 alphaproteobacterial families (light grey). Driving factors are plotted by principal component analysis (PCA) in Supplement Figure S3.

abundance of genes involved in lipid metabolism (COG I) and Alpha5 in addition for secondary metabolite synthesis and degradation (COG Q) while Alpha3 has the highest relative gene abundance for signal transduction mechanisms (COG T). With the most divergent metabolic profile, Alpha9 appeared to be an outlier within the alphaproteobacterial symbionts, largely caused by high genome space allocation to internal trafficking and secretion (COG U), replication and repair (COG L) and translation (COG J). Notably, the most widely distributed symbiont clade Alpha 10 does not have a remarkable metabolic profile compared to the other symbionts but rather features a well-balanced genomic investment across all COG categories. This could indicate that Alpha10 has a generalist's metabolism and might be an explanation for its success in colonizing a variety of gutless oligochaete host species at many locations. Overall, the Alphaproteobacteria in association with gutless oligochaetes were very diverse, with largely heterotrophic roles in the symbiosis. Alphaproteobacteria likely also provide additional specific functions to the host, as suggested by the genomic investment into e.g. secondary metabolites

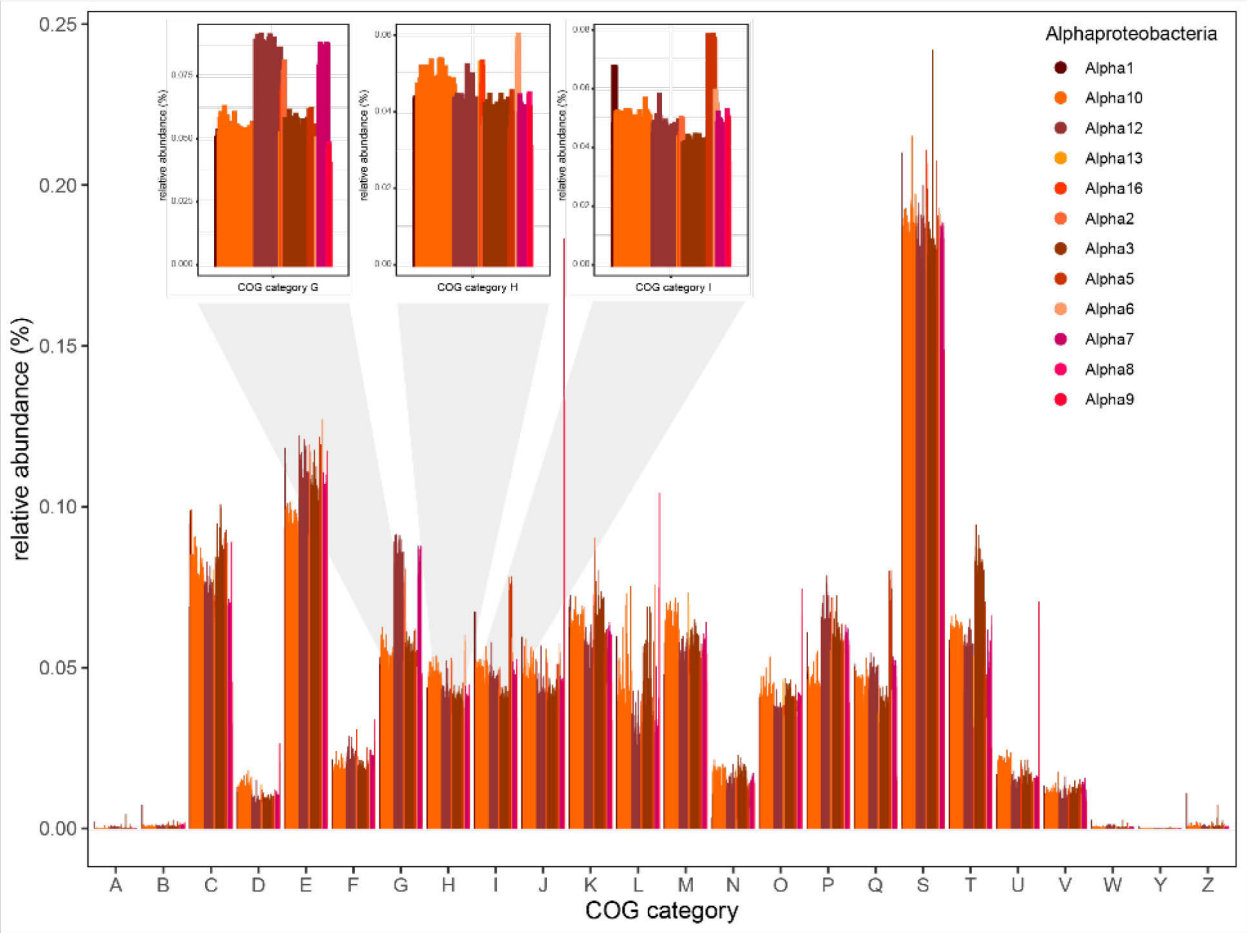


Figure 7: Distinct relative abundances of genes in COG categories indicate metabolic functions that Alphaproteobacteria might bring to the gutless oligochaete symbiosis.
Relative gene abundances in % per COG category were assigned with eggNOG and are depicted for all alphaproteobacterial MAGs from 233 gutless oligochaete specimens. The categories of carbohydrate metabolism (COG G), co-enzyme metabolism (COG I) and lipid metabolism (COG H) are examples for markedly different gene abundances in the symbiont MAGs.

in Alpha5. Host community specific and in-depth analyses of the genomic potential and expressed metabolism of the individual symbionts will be necessary to grasp each symbionts precise role in each of the gutless oligochaete symbiosis that it contributes to.

Divergent metabolic activities in the three main alphaproteobacterial symbiont genera

We demonstrated that Alphaproteobacteria are the most diverse and wide-spread secondary class of symbionts associated with gutless oligochaetes. Most clades have versatile metabolic capabilities, and to understand the exact contributions of a given symbiont in a given gutless oligochaete symbiosis, targeted research using expression data was necessary. Symbiont clades that appeared most relevant to study based on their broad distribution and abundance were Alpha3, Alpha10 and Alpha12. Of these three, Alpha3 stands out with its wide host range within gutless oligochaetes and its presence in the microbiomes of gutbearing relatives to the gutless

oligochaetes (Table 1)[2]. Proteomic expression data indicated that Alpha3 is a sulfur oxidizing chemoautotroph that employs the sox pathway for sulfur oxidation and a ribulose bisphosphate carboxylase for carbon fixation, suggesting key functional overlap to the main gammaproteobacterial symbiont.

Alpha10 is the most widely distributed alphaproteobacterial symbiont within gutless oligochaetes. Our genome and proteome analyses suggest that Alpha10 symbionts are heterotrophic symbionts, likely living off internal carbon sources such as waste products from the hosts' metabolism. Often co-occuring with Alpha10 and similarly widely distributed are symbionts of clade Alpha12[2]. Our expression data indicated a metabolism focused on saccharide degradation and an ability to respire carbon monoxide under aerobic conditions, hence Alpha12 might provide access to additional external carbon sources from surrounding seagrass meadows or coral reefs[12]. Overall, the expressed functions of the main genera of Alphaproteobacteria in symbiosis with gutless oligochaetes significantly expand their consortia towards mixotrophic symbiont communities, a role that cannot be underestimated given the gigatons of sugars that are bunkered under seagrass meadows worldwide[33].

Conclusion

Untargeted metagenomic approaches revealed that Alphaproteobacteria make up the most abundant and most widely distributed class of secondary symbionts in the gutless oligochaete symbiosis. The symbiont genera come from several orders, and can form several closely related species that were taken up multiple times by different hosts. We found that the symbionts have diverse metabolic potential and might be able to recycle sulfur compounds, fix carbon, and provide access to additional external carbon sources. Studying the main clades of Alphaproteobacteria in more detail will provide important insights in the function of the intricate and multipartite gutless oligochaete symbioses as well as into many more hosts that co-occur in these vast coastal hot-spots of remineralization and sulfur cycling.

Data and code availability

The 16S rRNA gene sequence data and assembled genomes used and generated in this study will be deposited in the European Nucleotide Archive (ENA) upon peer-review submission and are currently available upon request.

The scripts that were used for data visualization concerning mapping and COG profiling are available at github.com/TinaEnd/Alpha_all_GO.

Phylogenetic 16S rRNA gene and MAG based trees are available in project Alphas_all_GO at itol.embl.de/shared/tenders.

Acknowledgements

We thank all people involved in sampling and processing of this incredible amount of specimens, they are listed in Mankowski et al. 2021[2]. We are grateful to the Max Planck Society supporting this work and the Max Planck Genome Centre Cologne for conducting the sequencing. LC-MS/MS measurements were made in the Molecular Education, Technology, and Research Innovation Center (METRIC) at North Carolina State University. Funding was provided by the German Academic Exchange Service DAAD (M.J.). The U.S. National Science Foundation (grant IOS 2003107 to M.K.) supported this work. We thank Ben Francis and Maggie Sogin for essential input on R plotting. This work is contribution XXX from the Carrie Bow Cay Laboratory, Caribbean Coral Reef Ecosystem Program, National Museum of History, Washington DC.

References

1. Dubilier N, Mülders C, Ferdelman T, de Beer D, Pernthaler A, Klein M, et al. Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. Nature. 2001;411(6835):298-302.
2. Mankowski A, Kleiner M, Erséus C, Leisch N, Sato Y, Volland J-M, et al. Highly variable fidelity drives symbiont community composition in an obligate symbiosis. bioRxiv. 2021.
3. Dubilier N, Blazejak A, Rühland C. Symbioses between bacteria and gutless marine oligochaetes. Molecular basis of symbiosis. 2005:251-75.
4. Dubilier N, Giere O, Distel DL, Cavanaugh CM. Characterization of chemoautotrophic bacterial symbionts in a gutless marine worm (Oligochaeta, Annelida) by phylogenetic 16S rRNA sequence analysis and *in situ* hybridization. Applied and Environmental Microbiology. 1995;61(6):2346-50.
5. Dubilier N, Amann R, Erséus C, Muyzer G, Park S, Giere O, et al. Phylogenetic diversity of bacterial endosymbionts in the gutless marine oligochete *Olavius loisae* (Annelida). Marine Ecology Progress Series. 1999;178:271-80.
6. Blazejak A, Erséus C, Amann R, Dubilier N. Coexistence of bacterial sulfide oxidizers, sulfate reducers, and spirochetes in a gutless worm (Oligochaeta) from the Peru margin. Applied and Environmental Microbiology. 2005;71(3):1553-61.
7. 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. Applied and Environmental Microbiology. 2006;72(8):5527-36.
8. Ruehland C, Blazejak A, Lott C, Loy A, Erséus C, Dubilier N. Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. Environmental Microbiology. 2008;10(12):3404-16.
9. Zimmermann J, Wentrup C, Sadowski M, Blazejak A, Gruber-Vodicka HR, Kleiner M, et al. Closely coupled evolutionary history of ecto-and endosymbionts from two distantly related animal phyla. Molecular ecology. 2016;25(13):3203-23.

10. Bergin C, Wentrup C, Brewig N, Blazejak A, Erséus C, Giere O, et al. Acquisition of a novel sulfur-oxidizing symbiont in the gutless marine worm *Inanidrilus exumae*. Applied and environmental microbiology. 2018;84(7):e02267-17.
11. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature. 2006;443(7114):950-5.
12. Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. Proceedings of the National Academy of Sciences. 2012;109(19):E1173-E82.
13. Kleiner M, Wentrup C, Holler T, Lavik G, Harder J, Lott C, et al. Use of carbon monoxide and hydrogen by a bacteria–animal symbiosis from seagrass sediments. Environmental microbiology. 2015;17(12):5023-35.
14. Wippler J, Kleiner M, Lott C, Gruhl A, Abraham PE, Giannone RJ, et al. Transcriptomic and proteomic insights into innate immunity and adaptations to a symbiotic lifestyle in the gutless marine worm *Olavius algarvensis*. BMC genomics. 2016;17(1):1-19.
15. Mankowski A. From genomes to communities - Evolution of symbionts associated with globally distributed marine invertebrates: Universität Bremen; 2021.
16. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press; 2020.
17. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic acids research. 2019;47(W1):W256-W9.
18. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: A toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints; 2016. Report No.: 2167-9843.
19. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. 2019;47(D1):D309-D14.
20. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular biology and evolution. 2017;34(8):2115-22.
21. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics. 2010;11(1):1-11.
22. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015;12(1):59-60.
23. Junier T, Zdobnov EM. The Newick utilities: High-throughput phylogenetic tree processing in the Unix shell. Bioinformatics. 2010;26(13):1669-70.
24. R Core Team. R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
25. Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nature methods. 2009;6(5):359-62.
26. Hinzke T, Kouris A, Hughes R-A, Strous M, Kleiner M. More is not always better: Evaluation of 1D and 2D-LC-MS/MS methods for metaproteomics. Frontiers in microbiology. 2019;10:238.
27. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, et al. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Molecular & cellular proteomics. 2005;4(12):2010-21.
28. Mordant A, Kleiner M. Evaluation of sample preservation and storage methods for metaproteomics analysis of intestinal microbiomes. Microbiology spectrum. 2021;9(3):e01877-21.
29. Jensen M, Wippler J, Kleiner M. Evaluation of RNA later as a field-compatible preservation method for metaproteomic analyses of bacterium-animal symbioses. Microbiology spectrum. 2021;9(2):e01429-21.
30. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. Journal of proteome research. 2006;5(9):2339-47.
31. Giere O. Studies on marine Oligochaeta from Bermuda, with emphasis on new *Phallodrilus* species (Tubificidae). Cahiers de Biologie Marine. 1979;20:301-14.
32. Felbeck H, Liebezeit G, Dawson R, Giere O. CO_2 fixation in tissues of marine oligochaetes (*Phallodrilus leukodermatus* and *P. planus*) containing symbiotic, chemoautotrophic bacteria. Marine Biology. 1983;75(2):187-91.
33. Sogin EM, Michellod D, Gruber-Vodicka HR, Bourceau P, Geier B, Meier DV, et al. Sugars dominate the seagrass rhizosphere. Nature Ecology & Evolution. 2022.

Chapter 1 | Alphaproteobacteria associated with gutless oligochaetes

Alphaproteobacteria are the most abundant and diverse class of symbionts across the marine gutless oligochaete (Oligochaeta, Annelida) diversity

Supplementary information

Tina Enders[1], Anna Mankowski[1,2], Marlene Jensen[3], Manuel Kleiner[3,] Nicole Dubilier[1], Harald R. Gruber-Vodicka[1]

[1] Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

[2] Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

[3] Department of Plant & Microbial Biology, North Carolina State University, Raleigh 27695, North Carolina, USA

Corresponding authors:
Nicole Dubilier, ndubilie@mpi-bremen.de,
Harald R. Gruber-Vodicka, hgruber@mpi-bremen.de

1 Sampling sites of hosts and symbionts

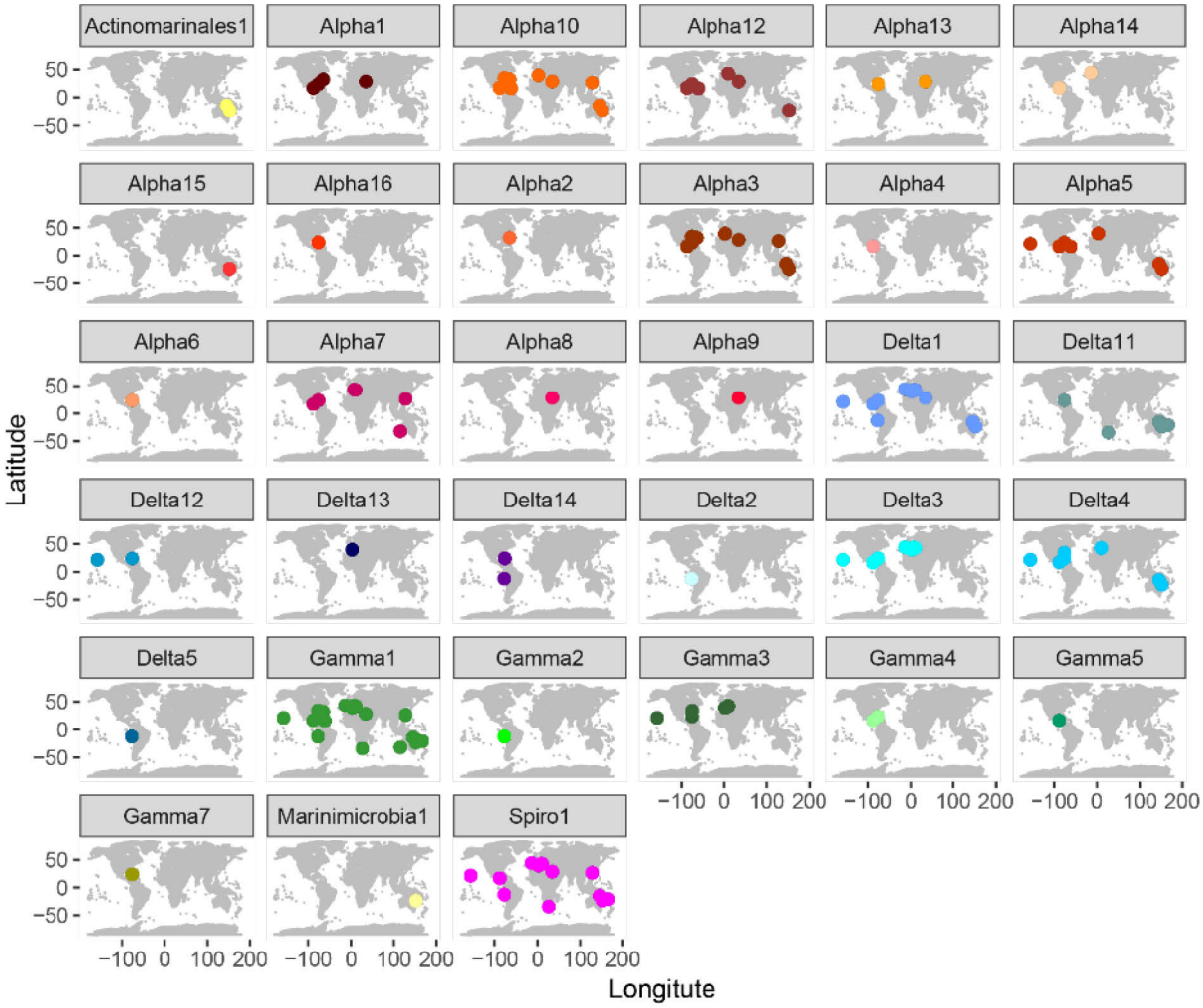


Figure S1: Based on presence of 16S rRNA genes detected in metagenomes of 233 gutless oligochaetes, 33 clades of bacteria associate with 64 gutless oligochaete species sampled at 17 globally distributed sites.

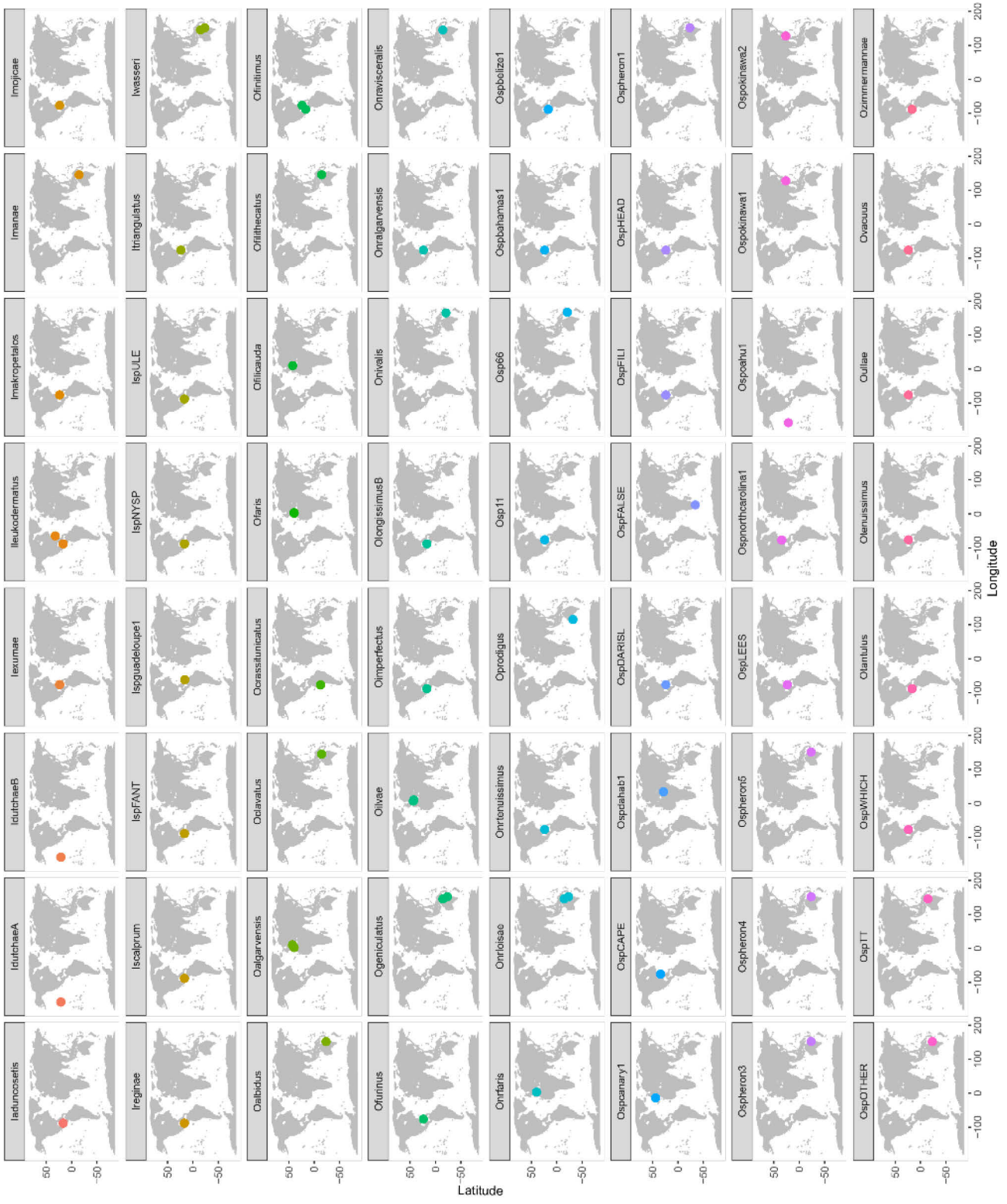


Figure S2: 233 gutless oligochaete specimen of 64 species were sampled at 17 globally distributed sites.

2 COG analysis

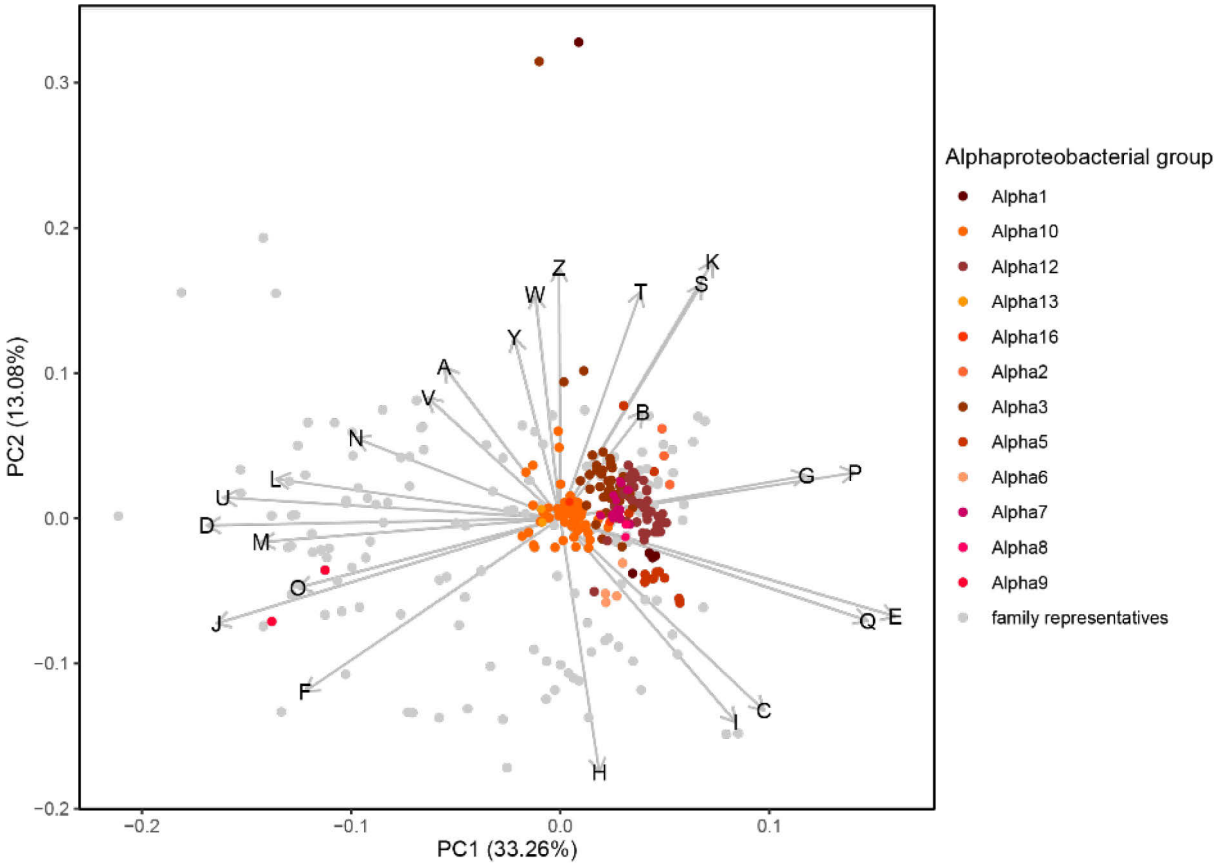


Figure S3: Alphaproteobacteria associating with gutless oligochaetes have distinct COG profiles and cluster within the diversity of alphaproteobacterial family representatives.
Principal component analysis with driving COG categories of the functional profiles based on relative abundances of COG categories from alphaproteobacterial MAGs of 233 gutless oligochaete host specimens (shades of red) and representatives of 149 alphaproteobacterial families (light grey). Non-metric dimensional scaling (NMDS) of the same dataset as is plotted in Figure 6.

Data and code availability

The 16S rRNA gene sequence data and assembled genomes used and generated in this study will be deposited in the European Nucleotide Archive (ENA) upon peer-review submission and are currently available upon request.

The scripts that were used for data visualization concerning mapping and COG profiling are available at github.com/TinaEnd/Alpha_all_GO.

Phylogenetic 16S rRNA gene and MAG based trees are available in project Alphas_all_GO at itol.embl.de/shared/tenders.

:abundant

The proposed *Candidatus* Saccharisymbium is one of the most widely distributed alphaproteobacterial genera in symbiosis with several globally distributed gutless oligochaete species. We can find it in the host species *Olavius ilvae*, which is abundant in seagrass meadow associated sediments around the island of Elba, Italy.

Chapter 2 | *Candidatus* Saccharisymbium – a globally distributed symbiont

A novel family of Rhodospirillales (Alphaproteobacteria) in symbiosis with globally distributed gutless marine worms (Oligochaeta, Annelida)

Tina Enders[1], Grace D'Angelo[1], Marlene Jensen[2], Manuel Kleiner[2], Nikolaus Leisch[1], Anna Mankowski[1,3], Dolma Michellod[1], Nicole Dubilier[1], Harald R. Gruber-Vodicka[1]

[1] Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

[2] Department of Plant & Microbial Biology, North Carolina State University, Raleigh 27695, North Carolina, USA

[3] Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Corresponding authors:
Nicole Dubilier, ndubilie@mpi-bremen.de,
Harald R. Gruber-Vodicka, hgruber@mpi-bremen.de

Competing interest

The authors declare no competing financial interests.

Abstract

Alphaproteobacteria are an abundant and diverse class of endosymbionts of gutless oligochaete worms from shallow water marine habitats associated with seagrass meadows and coral reefs. However, little is known about the function of Alphaproteobacteria in the multipartite gutless oligochaete symbiosis. Here, we describe six closely related species of Alphaproteobacteria that form a symbiont-exclusive genus that represents a novel family within the order of Rhodospirillales. These species are associated with seven host species that were collected from the Mediterranean Sea, the Atlantic, the Pacific and the Indian Ocean. Fluorescence *in situ* hybridization of the host species *Olavius ilvae* showed that the symbionts co-localized with previously described Gamma- and Deltaproteobacteria in the subcuticular symbiont layer. The high quality metagenome-assembled genomes for all six species had an average size of 4.4 Mbp and a GC content of 69.8%. The symbiont genomes encoded a variety of heterotrophic metabolic pathways and expression analysis using metatranscriptomes and metaproteomes indicated that they base their metabolism on the import and degradation of various oligo- and polysaccharides. Enzymes to use both oxygen and nitrous oxide as terminal electron acceptors in the respiration chain are expressed. We hypothesize that the newly discovered symbiont family of Rhodospirillales extends the metabolic access of the hosts to organic substrates from the environment such as sugars excreted from nearby seagrass meadows.

Keywords

Alphaproteobacteria, marine bacterial-animal symbiosis, gutless oligochaete, sugar degradation

Introduction

Nutritional animal-microbe symbioses have enabled animals to thrive in ecological niches that would not be accessible without their bacterial partners. One striking example are marine chemosynthetic symbioses in which bacteria fix carbon dioxide from the atmosphere driven by chemical energy from the oxidation of inorganic molecules such as sulfur or methane[1]. First detected in the deep sea, chemosynthetic symbioses are also abundant in oligotrophic shallow water sediments in marine coastal regions. One of the most diverse and most widely distributed chemosynthetic symbioses are annelids called ‘gutless oligochaetes’ from the genera *Olavius* and *Inanidrilus*[2, 3]. More than 100 species are described from tropical and subtropical regions and their prime habitats are connected to seagrass meadows and coral reef sediments[4]. Gutless oligochaetes have no digestive system and no excretory organs and instead rely on a consortium

of diverse, yet specific bacteria to maintain their metabolism[5, 6]. Their main chemosynthetic symbiont is *Candidatus* Thiosymbion, a Gammaproteobacterium from the purple sulfur bacteria, the Chromatiaceae[7]. These symbionts fix carbon dioxide based on energy derived from oxidizing sulfur compounds[8, 9]. Since reduced sulfur compounds are scarce and need to be replenished, the Gammaproteobacteria have been shown to engage in an internal sulfur cycle with sulfate-reducing Deltaproteobacteria[10]. These make up a large and diverse group of secondary symbionts in several host species of gutless oligochaetes and have been studied in detail for their metabolic function and role in the symbiosis[8-11]. Other secondary symbiont clades have been detected on the 16S ribosomal RNA (rRNA) gene level, belonging to the Gammaproteobacteria, Alphaproteobacteria and Spirochaetia[6, 12, 13]. However, only few studies have described the role and function of these secondary symbionts based on genomic data, and all of them have focused on the Gamma- and Deltaproteobacteria dominated consortium of the host species *Olavius algarvensis* from the Mediterranean Sea. Recent metagenomic studies on a large set of gutless oligochaete hosts showed that Alphaproteobacteria are an abundant and even more diverse group of secondary symbionts than the Deltaproteobacteria[14]. Several alphaproteobacterial symbionts can also serve as the only additional symbiotic partner next to the main chemosynthetic Gammaproteobacterium but their metabolism and function in the symbiosis remain elusive[6]. Generally, Alphaproteobacteria are a diverse and multifunctional clade of bacteria prone to engage in symbiosis with a wide range of hosts from different animal phyla. The most prominent examples come from terrestrial systems, e.g. *Wolbachia* that associate with almost all terrestrial arthropod clades or *Candidatus* Hodgkinia cicadicola that have the smallest described genome[15, 16]. Marine examples are manifold and have been described from several host phyla, including sponges, placozoans, corals and flatworms[17-22].

Given the stable association of several alphaproteobacterial clades in globally abundant gutless oligochaetes, we expect these clades to have beneficial traits and important roles in these symbioses[14]. One of the most prevalent alphaproteobacterial symbiont clades occurs in the highly abundant and easily accessible Mediterranean host species *Olavius ilvae*[14]. Focusing on this host species and a Caribbean host species, we here characterize this novel and globally distributed family of Alphaproteobacteria with phylogenetic, light-microscopic and physiological analyses. The symbiont belongs to the Rhodospirillales and resides in the subcuticular layer, co-localizing with the main and other secondary symbionts. Based on physiological reconstruction and expression analysis, the symbiont degrades various

saccharides as its main metabolic function. For this reason, we propose the novel genus *Candidatus* Saccharisymbium (from here on referred to as Saccharisymbium).

Methods and materials

Sample collection

Gutless oligochaete specimens were sampled in the years between 1991 and 2020 from globally distributed sampling sites of shallow water sediments associated with seagrass meadows or coral reefs in water depths ranging from 0.3 to 7 m. Detailed metadata is noted in Supplement Table S1a-c and also described in a previous study[14]. For genomic, transcriptomic or proteomic analysis, specimens were fixed in RNAlater (Thermo Fisher Scientific, Waltham, MA, USA) and stored at -20 °C or 4 °C or flash frozen in liquid nitrogen until used. Specimens used for fluorescence *in situ* hybridization were fixed in 4% paraformaldehyde (PFA) and stored in methanol at -20 °C.

Short-read DNA/RNA extraction, sequencing, assembly and binning

DNA and RNA from individual specimens was obtained by homogenizing the samples with bead beating or proteinase K digestion and extracted using the DNeasy Blood & Tissue Kit or the AllPrep DNA/RNA/Protein Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Metagenomic library preparation, quality control and sequencing was carried out at the Max Planck Genome Centre (Cologne, Germany). Paired-end reads with 2x150 bp read lengths were sequenced aiming for an average of 30 million reads per sample. Reads were quality filtered and trimmed with bbduk.sh (BBtools, jgi.doe.gov/data-and-tools/bbtools/) and error corrected with bayeshammer implemented in SPAdes[23, 24]. Read quality was assessed with FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) before and after trimming. Species identification based on 28S rRNA, mtCOI for the host and 16S rRNA genes for the symbionts from short-read libraries was done as previously described[14, 25]. Genome assembly was done with metaSPAdes and genomes binned with a multi-tool binning approach as described by Mankowski *et al.*[25]. The quality of metagenome-assembled genomes (MAGs) such as completeness and contamination was assessed with checkM[26].

Long-read DNA extraction, sequencing, assembly and binning

Genomic DNA was extracted from one fresh *Olavius ilvae* specimen. High molecular weight genomic DNA was isolated with the MagAttract HMW DNA Kit (Qiagen, Hilden, Germany). Quality was assessed by the Agilent FEMTOpulse and DNA quantified by the Quantus dsDNA

kit (Promega, Madison, USA). DNA was processed to obtain a PacBio Sequencing-compatible library following the recommendations outlined in "Procedure & Checklist – Preparing HiFi Libraries from Low DNA Input Using SMRTbell Express Template Prep Kit 2.0". Libraries were sequenced on a Sequel II instrument at the Max-Planck Genome Centre with sequencing chemistry 2.0, binding kit 2.0 on one 8M SMRT cell for 30 h applying continuous long read (CLR) sequencing mode. The CLR reads were mapped using the most complete short read MAG as reference (OilvSAN1). The mapped reads were then assembled using Flye (v 2.8)[27]. The completeness of the assembly was assessed with QUAST and checkM[26, 28]. MAGs were visually inspected with Bandage (Supplement Figure S1)[29].

Phylogenomic placement of symbionts

Phylogenetic placement of Saccharisymbium spp. was done based on full-length 16S rRNA gene sequences and marker gene sets of available MAGs. A 16S rRNA gene tree was calculated using a maximum likelihood approach and maximum likelihood ratio tests (mafft-xinsi, IQ-TREE –m MFP –alrt 1000)[30, 31]. Reference sequences included were related hits to Saccharisymbium spp. from NCBI's environmental and type strain nucleotide collections (nr/nt, env_nt) and relevant GTDB relatives[32, 33]. MAGs of Saccharisymbium spp. were placed in the GTDB tree at version 1.5.0 with the R202 reference data using the GTDB-Tk software[33]. All trees were visualized and annotated with iTOL and final figures edited with Adobe illustrator v25.03 (www.adobe.com/products/illustrator)[34]. Average nucleotide identity (ANI) values between the individual MAGs of the symbiont clade were calculated with the ANI/AAI-Matrix Genome-based distance matrix calculator (http://enve-omics.ce.gatech.edu/g-matrix/)[35].

Gene expression analysis

We analyzed gene expression in three individuals each of the two host species *Olavius ilvae* from Sant'Andrea, Elba, and *Inanidrilus sp.* ULE from Curlew Cay, Belize. Metagenomic and total RNA metatranscriptomics libraries aiming at 30 million paired-end reads (2x150 bp) were prepared for single specimens. MAGs were binned as described previously[25]. We were able to obtain a high quality MAG of Saccharisymbium spp. from two host specimens of each species. These MAGs were annotated with prokka (v1.14.5, --compliant) and the coding sequences (CDS) were used as reference for mapping the corresponding transcriptomic reads with bbmap.sh (v38.90, minid=0.99 t=32 trimq=20 qtrim=lr untrim=t pairedonly=t pairlen=600 mintrimlength=100 32bit=t covstats=covstats.txt statsfile=statsfile.txt bamscript=bs.sh;

sh bs.sh)[36, 37]. Expression patterns were evaluated focusing on genes that had an average or median fold expression greater than 1.

Symbiont fractionation for proteome analysis

O. ilvae worms were collected from Sant'Andrea bay, Elba, Italy, homogenized in batches of 25-30 individuals, and fractionated via density centrifugation following a protocol adapted from Hinzke *et al.*[38]. Specimen batches were placed in glass Duall homogenizers (tissue grind pestle and tube SZ22, Kontes Glass company, Vineland, New Jersey) with 0.5 ml sterile artificial seawater and homogenized by grinding. The homogenate was transferred to 2 ml screw cap cryovial tubes and filled to a final volume of 1.5 ml with sterile artificial seawater. The samples were centrifuged for 2 min at 4000x g at 4 °C and the supernatant was transferred to a new cryovial (designated S1). The remaining pellet (P1) was resuspended in 1.5 ml sterile artificial seawater. The centrifugation conditions were repeated on these two fractions (S1 & P1). The supernatant from P1 was transferred into a new cryovial (S2) and the remaining pellet was resuspended in sterile artificial seawater. The supernatant from S1 was transferred to a new cryovial (S3) and the pellet was frozen at -80 °C. The centrifugation conditions were repeated on fractions P1 and S2. The supernatant from S2 was transferred to a new cryovial (S4) and the pellet was frozen at -80 °C. The supernatant from P1 was discarded and the pellet was frozen at -80 °C. Fractions S3 and S4 were centrifuged for 7 min at 21,000x g at 4 °C. The supernatant of S4 was discarded and the supernatant of S3 was transferred to a new cryovial (S5). All cryovials were finally frozen at -80 °C for long term storage. From each batch, fractions P1 and S3, representing the larger symbiont fraction and the smaller symbiont fraction, were selected for protein extraction.

Protein extraction and peptide preparation

We extracted proteins from five specimens of the marine gutless oligochaete *Olavius ilvae*, distinct from the samples used for genomic analysis. We conducted a tryptic protein digestion following the filter-aided sample preparation (FASP) protocol, adapted from Wisniewski *et al*., 2009, for all samples[39]. We added 60 µL SDT-lysis buffer (4%, w/v, SDS, 100 mM Tris-HCl pH 7.6, 0.1 M DTT) and heated samples to 95 °C for 10 min. To minimize sample loss, we did not do the 5 minutes centrifugation step at 21,000g as described in the original FASP protocol[39]. Instead, only a short spin down was conducted. Subsequently, we mixed the lysate with 400 µL UA solution (8 M urea in 0.1 M Tris/HCl pH 8.5) in a 10 kDa MWCO 500 µL centrifugal filter unit (VWR International) and centrifuged the mixture at 14,000g for 20 min.

Next, we added 200 µL of UA solution and centrifugal filter spun again at 14,000g for 20 min. Subsequently, we added 100 µL of IAA solution (0.05 M iodoacetamide in UA solution) and incubated samples at 22 °C for 20 min in the dark. We removed the IAA solution by centrifugation following three washing steps with 100 µL of UA solution. Subsequently, filters were washed three times with 100 µL of ABC buffer (50 mM ammonium bicarbonate). We added 0.5 µg of Pierce MS grade trypsin (Thermo Fisher Scientific) in 40 µL of ABC buffer to each filter. We incubated filters overnight in a wet chamber at 37 °C. The next day, we eluted the peptides by centrifugation at 14,000g for 20 min followed by addition of 50 µL of 0.5 M NaCl and another centrifugation step. Peptides were quantified using the Pierce MicroBCA Kit (Thermo Fisher Scientific) following the instructions of the manufacturer. For the 2D-LC-MS/MS analysis, we desalted the peptides with Sep-Pak C18 Plus Light Cartridges (Waters). Acetonitrile from the peptide elution step was exchanged for 0.1% formic acid (v/v) using a centrifugal vacuum concentrator. The desalting step was necessary to enable binding of peptides to the SCX column during sample loading for the 2D-LC method.

1D-LC-MS/MS

All samples were analyzed by 1D-LC-MS/MS as described in Hinzke *et al*. 2019[40]. 1500 ng of peptides were loaded in loading solvent A (2% acetonitrile, 0.05% trifluoroacetic acid) onto a 300 µm i.d. x 5 mm trap cartridge column packed with Acclaim PepMap100 C18, 5 µm, 100 Å (Thermo Fisher, 160454) using an UltiMate 3000 RSLCnano Liquid Chromatograph (Thermo Fisher Scientific). The trap was connected to a 75 µm x 75 cm analytical EASY-Spray column packed with PepMap RSLC C18, 2 µm material (Thermo Fisher Scientific), which was heated to 60 °C via the integrated heating module. The analytical column was connected via an Easy-Spray source to a Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific). Peptides were separated on the analytical column at a flow rate of 225 nl min^{-1} using a 460 min gradient. The gradient went from 98% buffer A (0.1% formic acid) to 31% buffer B (0.1% formic acid, 80% acetonitrile) in 364 min, then from 31% to 50% buffer B in 76 min and ending with 20 min at 99% buffer B. Eluting peptides were ionized via electrospray ionization (ESI) and analyzed in Q Exactive HF-X. Full scans were acquired in the Orbitrap at 60,000 resolution. The 15 most abundant precursor ions were selected in a data dependent manner, isolated with the quadrupole with a 1.2 m/z isolation window size, fragmented in the HCD cell with a NCE of 25, and measured in the Orbitrap at 7,500 resolution. The mass (m/z) 445.12003 was used as lock mass as described in Olsen *et al*. 2005[41]. Lock mass use was set to 'best'. Singly charged ions were excluded from MS/MS analysis. Dynamic

exclusion was set to 30 s. On average 209,160 MS/MS spectra were acquired per sample with the 460 min gradient.

2D-LC-MS/MS

For the 2D-LC-MS/MS runs, we used the same instrumentation as for the 1D run. We followed the LC methods described in Hinzke *et al*. 2019 for the pH plug runs[40]. We loaded 9000 ng of peptide mixture with loading solvent B (10 mM phosphate buffer pH 3, 20% ACN and 600 mM NaCl) onto a 10 cm, 300 µm Micro SCX LC column (Thermo Fisher Scientific) at a flow rate of 300 µl min^{-1}. The specific plumbing scheme used in the RSLCnano corresponded to the standard set up recommended by the manufacturer for on-line 2D pH plug separations. During loading, the C18 trap (see above) was in-line downstream of the SCX column to capture peptides that did not bind to the SCX column (breakthrough). After loading, the C18 pre-column was switched in-line with the 75 µm x 75 cm analytical column (same as for 1D) and the breakthrough was separated using an 120 min[40]. Subsequently, elution of peptides from the SCX to the C18 trap (same as for 1D-LC) took place by injection of 20 µl of 8 different pH plugs with increasing pH (CTIBiphase buffers, Column Technology, Inc.) from the autosampler. The C18 trap was then again switched in-line with the analytical column and peptides separated with gradients of eluent A and B. Data acquisition in the mass spectrometer was done as described by Mordant and Kleiner, 2021[42].

Protein identification and quantification

We used a custom database which contained 1,362,363 protein sequences, including protein sequences predicted from host transcriptomes, non-redundant host sequences from the closely related species *Olavius algarvensis* and symbiont protein sequences predicted from metagenome assembled genomes, as well as a cRAP protein sequence database (http://www.thegpm.org/crap) of common laboratory contaminants. The database is available from the PRIDE repository (see data availability). Searches of the MS/MS spectra against this database were performed with the Sequest HT node in Proteome Discoverer version 2.2.0.388 (Thermo Fisher Scientific) as described in Jensen *et al*., 2021[43]. The following parameters were used: trypsin (full), maximum two missed cleavages, 10 ppm precursor mass tolerance, 0.1 Da fragment mass tolerance and maximum of 3 equal dynamic modifications per peptide, namely: oxidation on N (+ 15.995 Da), carbamidomethyl on C (+ 57.021 Da) and acetylation on the protein N terminus (+ 42.011 Da). False discovery rates (FDRs) for peptide spectral matches (PSMs) were calculated and filtered using the Percolator Node in Proteome

Discoverer. Percolator was run with a maximum delta Cn 0.05, a strict target FDR of 0.01, a relaxed target FDR of 0.05 and validation based on q-value. The Protein FDR Validator Node in Proteome Discoverer was used to calculate q-values for inferred proteins based on the results from a search against a target-decoy database. Proteins with a q-value of <0.01 were categorized as high-confidence identifications and proteins with a q-value of 0.01–0.05 were categorized as medium-confidence identifications. Search results for all samples were combined into a multiconsensus report in Proteome Discoverer and only proteins identified with medium or high confidence were retained, resulting in an overall protein-level FDR of 5%. For protein quantification, normalized spectral abundance factors (NSAFs) were calculated per species and multiplied by 100, to give the relative protein abundance in %[44]. The mass spectrometry metaproteomics data and protein sequence database have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino *et al*. 2016) partner repository with the dataset identifier XXX (access for reviewers: Username: Password:).

Metabolic reconstruction and COG profiling

We analyzed the metabolic potential of Saccharisymbium spp. with a focus on *O. ilvae* specimens from Sant'Andrea, Elba, and I. sp. ULE specimen from Curlew Cay, Belize, that were used for transcriptomic and proteomic analysis as described above (libraries 4515_G-L, 4514_G-L, 4410). The genomes were functionally annotated with the automated tools prokka and RAST and both annotations from the five samples were considered[37, 45, 46]. Effector proteins and secretion systems were predicted using effectiveDB[47]. Hypothetical proteins with high expression values were manually annotated with a combination of psi-blast, nt-blast and hmmscan against the Pfam database (www.ebi.ac.uk/Tools/hmmer/search/hmmscan)[32, 48]. Transcriptomic and proteomic expression was considered based on presence-absence for the metabolic cell overview (Figure 6). Pathway Tools and MetaCyc were used to reconstruct metabolic pathways and compare the five MAGs used for metabolic reconstruction [49, 50]. Figure 6 was constructed in Adobe illustrator v25.3 (www.adobe.com/products/illustrator).

All 19 available Saccharisymbium spp. MAGs for annotated for their clusters of orthologous genes (COG) categories that represent larger functional classifications[51]. We used eggNOG-mapper v2.1.6 with diamond as protein aligner to obtain COG profiles of Saccharisymbium spp., its relatives from GTDB taxonomy and representatives of 149 alphaproteobacterial families[52-55]. To obtain alphaproteobacterial family representatives, the bacterial tree bac120_r86.2 was downloaded from the GTDB database (gtdb.ecogenomic.org/downloads)

and trimmed to the Alphaproteobacteria and Magnetococcia outgroup in iTOL[34]. Branch lengths for each leaf were extracted with Newick utils (nw_distance), the average branch length was calculated for each family, and for each family the genome that was closest to the average branch length was chosen for further phylogenomic analyses (genomes listed in zenodo.org/record/6514058)[56]. Genomic investment in different metabolic categories was quantified and plotted with bar-charts, non-metric multidimensional scaling (NMDS) and principal component analysis (PCA) calculated in RStudio (RStudio v1.4.1106, R v4.0.4) using a variety of packages (Supplement Table S2).

Estimation of global distribution and abundance

We used the integrated microbial NGS platform to assess the global distribution of related sequences from public sequence read archives (SRAs) to the 16S rRNA gene sequence of Saccharisymbium spp. (IMNGS, similarity threshold at 99% and 97%, Min size 200, www.imngs.org/)[57]. We reduced the dataset of all seven available full-length 16S rRNA genes longer than 1000 bp with less than 20 Ns of Saccharisymbium spp. by using three centroids (OilvsSAN2, 4148_4829_B and 4148_4829_P) clustered with VSEARCH v2.6.2 (--cluster_fast, --id 0.99, --centroids)[58]. Metadata to SRAs with hits were downloaded from NCBI with Batch Entrez and location data plotted in R (RStudio v1.4.1106, R v4.0.4) using a variety of packages (Supplement Table S2).

Fluorescence *in situ* hybridization

We used fluorescence *in situ* hybridization (FISH) to localize Saccharisymbium spp. inside the body of the host species *O. ilvae* from Sant'Andrea on Elba, Italy. The ARB software Design Probes function was used to design specific double-labelled FISH probes that match the symbionts' 16S rRNA gene[59]. The database used was SSU Ref NR 99. The probes were tested *in silico* for self-dimer, hetero-dimer and hairpin binding with the OligoAnalyzer Tool (eu.idtdna.com/pages/tools/oligoanalyzer). The two best probes 5'- GCC TGA CTG ACC GTT CCG -3' and 5'- TAG ATC GGC AGT ATC AAG CG -3' with Atto 550 5'-end and 3'-end modification were purchased at biomers.net (Ulm, Germany, www.biomers.net/en/index.html?lang=en). All probes used in this study are listed in Supplement Table S3. Whole worms of *O. ilvae* were fixed in 4% paraformaldehyde for 4 h and stored in methanol at -20 °C. Prior to FISH, samples were stepwise rehydrated in phosphate buffered saline with Tween20 (20%, 1:25 v:v, PTw). To increase probe accessibility, a digestion step with proteinase K (0.05 mg ml^{-1} at 37 °C for 5 min) stopped with glycine

(2 mg ml^{-1}) and re-fixation with 4% PFA for 1 h at room temperature followed. FISH was conducted slightly modified after the standard protocol described by Manz *et al.*[60]. We used a formamide (FA) concentration of 30%, hybridized in an over-night step with 8.4 pmol μl^{-1} of oligonucleotide probes and used extended steps for washing (see Supplement Section 5 for the whole protocol). Nucleic acid in the samples was counter-stained with DAPI. Epifluorescence images were acquired on a Zeiss LSM 780 (Carl Zeiss, Jena, Germany) using the ZEN software (v2.3 blue edition, www.zeiss.de/mikroskopie/produkte/mikroskopsoftware/zen.html), and final figures edited with Fiji (ImageJ v1.53f51) with the bioformats plugin (github.com/ome/bioformats) and Adobe illustrator v25.3 (www.adobe.com/products/illustrator)[61].

Results and discussion

Candidatus Saccharisymbium – a novel genus of Rhodospirillales associated with gutless marine worms

A novel cluster of alphaproteobacterial phylotypes identified during the investigation of 233 gutless oligochaete individuals from globally distributed sampling sites represent a novel family in the order Rhodospirillales and lack close relatives with known function[14]. The seven high-quality sequences of the total of 21 sequences (>1000 bp, <20 N) had a similarity of 97.7-99.5%. The phylotypes of this cluster occurred in seven distinct host species sampled at eight spots from the Mediterranean Sea, the Atlantic, Pacific and Indian Ocean (Figure 1, Table 1, metadata in Supplement Table S1). Phylogenetic placement based on full-length 16S rRNA gene sequences and maximum likelihood approaches place the 21 phylotypes from previous studies and five phylotypes obtained in this study as a novel symbiont-only genus and the first representatives of a novel family within the order of Rhodospirillales (Figure 2, see Data availability for trees shared in iTOL)[14]. Phylotypes of the newly proposed genus of Saccharisymbium spp. form two separate clades of which one contains the Caribbean and Japanese samples from the Pacific and Atlantic Ocean (for genus description see Supplement Section 2). The other clade contains samples from the Australian coast in the Indian Ocean and from the Mediterranean Sea. Closest relatives to Saccharisymbium spp. available from NCBI are an uncultured bacterial fosmid library clone from the Mediterranean Sea (EU686599) with a sequence similarity of 91.4-92.5% to Saccharisymbium spp. phylotypes and an uncultured bacterial clone from the Mediterranean Medea hypersaline anoxic basin (JF809761) with 92.6-93.3% sequence similarity (Figure 1 and 2). These two samples share a 16S rRNA gene

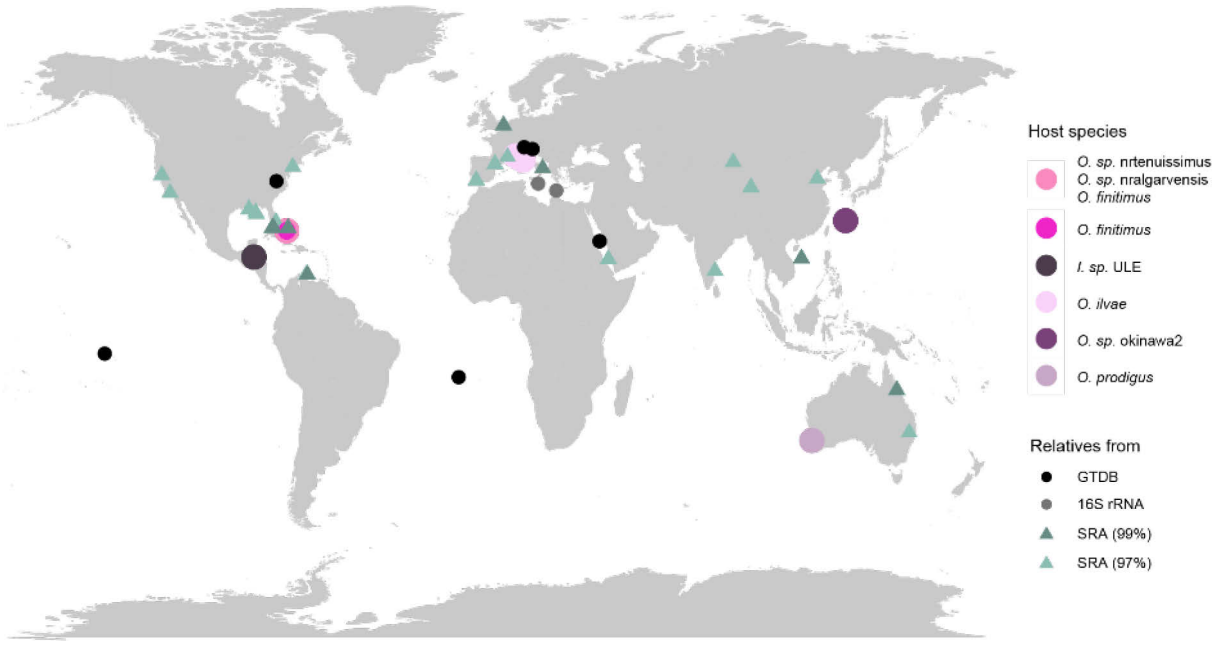


Figure 1: Six *Candidatus* Saccharisymbium spp. are globally distributed based on sampling sites and related sequences from public databases.
The map shows sampling sites and host species of the Saccharisymbium species sampled in this study (colored dots) along with the sampling sites of close relatives from 16S rRNA (black dots) and GTDB phylogeny (grey dots) presented in Figure 2 and 3. Furthermore, sequence read archive (SRA) hits with a similarity of ≥99% (dark green) and ≥97% (light green) to representative full-length 16S rRNA sequences of Saccharisymbium spp. (centroid of 99% similarity, n=3) are plotted (triangles).

sequence similarity of 94.5% and likely represent members of a distinct genus within the same, yet undescribed, family of Rhodospirillales as Saccharisymbium spp. The taxonomy for the order Rhodospirillales is still under debate and current polyphyletic structuring needs to be resolved[62]. The sister-family to Saccharisymbium spp. contains a large variety of sequences from uncultured bacteria. Together with these, Saccharisymbium spp. form the sister-clade to the core Rhodospirillales as described by Hördt *et al.*[62]. Saccharisymbium spp. and core Rhodospirillales form the sister clade to Rhodovibrionaceae and Kiloniellaceae. Based on 16S rRNA gene phylogeny, all symbionts in this cluster form a novel genus level clade. Its remote phylogenetic position compared to all members of described families suggest a novel family level clade at the base of core Rhodospirillales, despite low divergence rates on the 16S rRNA gene level.

Phylogenetic placement based on genomic marker sets from the Genome Taxonomy Database (GTDB) analyzed with GTDB-Tk supports clustering of Saccharisymbium spp. as a novel family branching as sister taxon to the Rhodovibrionaceae and Kiloniellaceae (Figure 3, see Data availability for trees shared in iTOL). Based on GTDB taxonomy, these form the order of

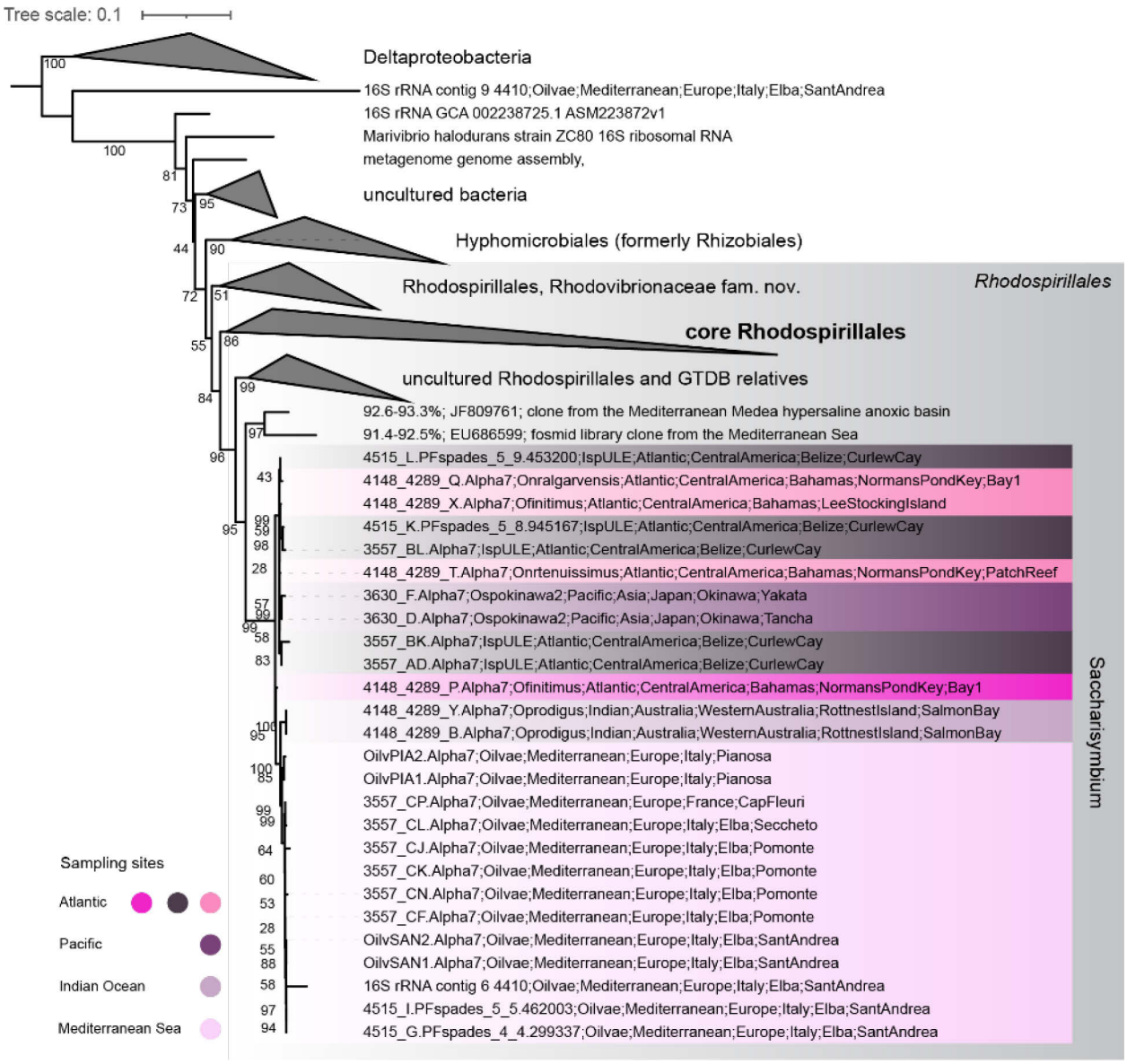


Figure 2: Based on the 16S rRNA gene, 21 Saccharisymbium phylotypes form a novel genus and family level clade and belongs to the sister clade to core Rhodospirillales[62].
Colored backgrounds correspond to sampling sites and host species in Figure 1 and symbiont species in Figure 3. Numbers at tree nodes represent bootstrap support values calculated using -aLRT in IQ-TREE. Selected Deltaproteobacteria were used as an outgroup (Supplement Section 3). The tree scale represents the average number of substitutions per site.

Kiloniellales instead of being part of the Rhodospirillales. Nevertheless, the overall position of Saccharisymbium spp. is congruent based on 16S rRNA gene and GTDB taxonomy. The GTDB reference includes uncharacterized MAGs from Tara Ocean and other sampling campaigns that cluster with Saccharisymbium spp. These samples were collected in the South Pacific and Atlantic, the Caribbean, the Mediterranean and the Red Sea (Figure 1). The related MAGs have an average nucleotide identity (ANI) of 75% or less to Saccharisymbium spp. and are thus distinct bacterial species, also compared to each other (Supplement Figure S2)[63]. Based on

ANI values, the MAGs from gutless oligochaetes represent five distinct species of Saccharisymbium that cluster by location and are host species specific. However, the MAGs of two of individual hosts recovered from neighboring islands of the Bahamas that are all hosted by individuals of *O. finitimus* had an ANI value of 88%. As this is below the common species cutoff of 95%, we propose six novel species within the genus of Saccharisymbium as listed in Table 1 (for species description see Supplement Section 2). Overall, our recovered MAGS share a minimum ANI of 84%. Earlier studies have shown that ANI values tend to be clearly separated and bacteria share either >95% ANI if they belong to the same species or <83% when they belong to distinct species[63]. Our observed range of ANI values of 84-92% between different symbiont species might point to a short separation time in the genus Saccharisymbium.

Table 1: Six *Candidatus* Saccharisymbium spp. occur in seven gutless oligochaete host species from eight globally distributed sampling sites.
The table lists gutless oligochaete host species, which Saccharisymbium species they host, and their sampling locations as represented in Figure 1. Quality statistics (completeness and contamination based on checkM) for the best MAG per proposed Saccharisymbium species are stated. Species descriptions are provided in Supplement Section 2.

Host species	**Proposed species** *Candidatus* Saccharisymbium	**Sampling site** Spot, region	**Best MAG** Compl. Cont.
Olavius prodigus	australicum	Rottnest Island, Australia	98.71% 1.12%
Olavius sp. nrtenuissimus	bahamense	Normans Pond Key, Bahamas	99.57% 0.65%
Olavius sp. nralgarvensis		Normans Pond Key, Bahamas	98.70% 1.30%
Olavius finitimus		Lee Stocking Island, Bahamas	99.57% 0.65%
Olavius finitimus	aequebahamense	Normans Pond Key, Bahamas	99.57% 0.43%
Inanidrilus sp. ULE	belizense	Curlew Cay, Belize	99.57% 0.43%
Olavius ilvae	mediterraneum	Elba, Italy	98.51% 0.00%
		Pianosa, Italy	99.13% 0.43%
		Cap Fleuri, France	99.0% 0.00%
Olavius sp. okinawa2	okinawense	Okinawa, Japan	85,57% 0.76%

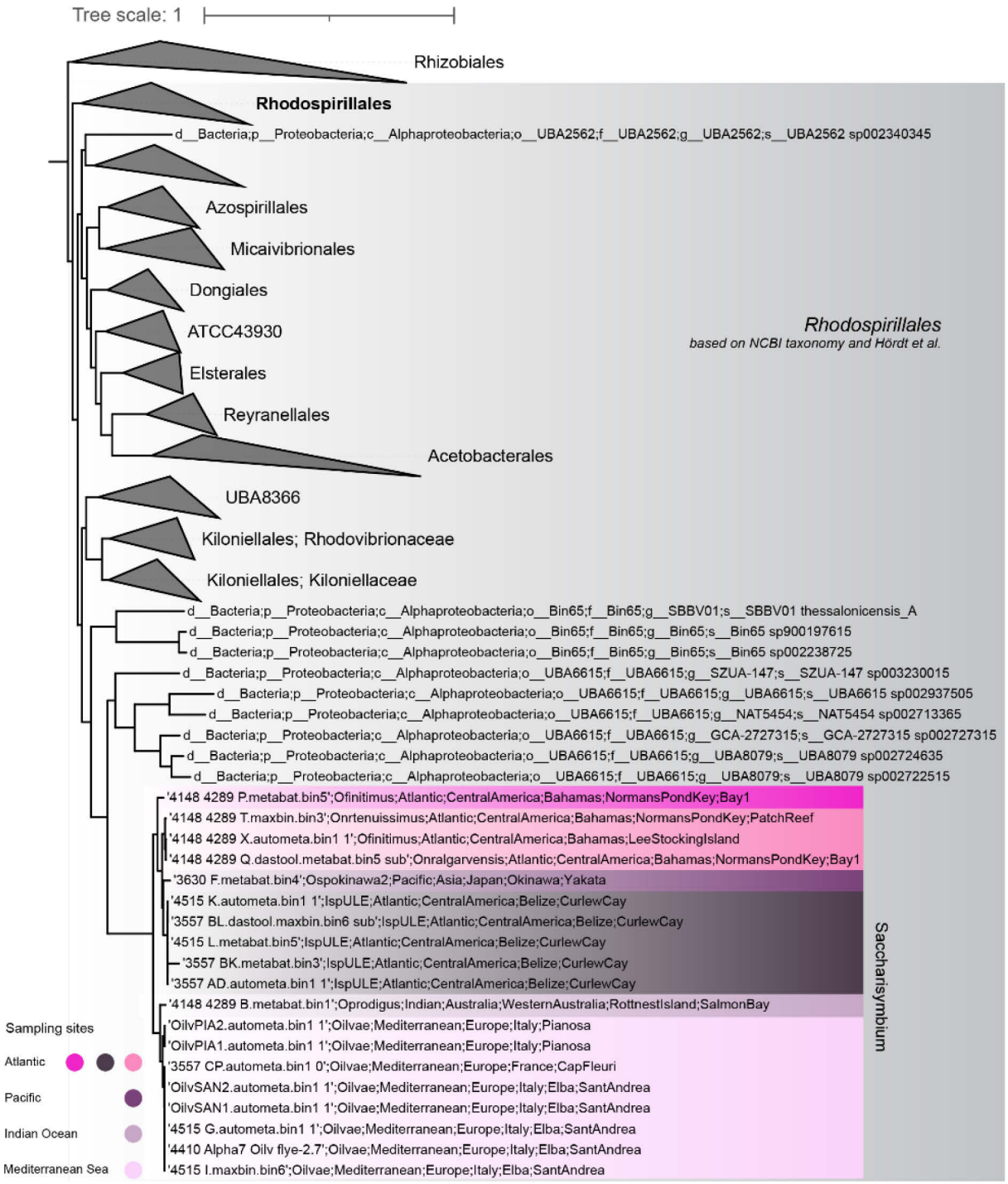


Figure 3: Based on phylogenomic analysis using GTDB-tk, 19 Saccharisymbium phylotypes form a novel genus and family level clade and belong to the sister clade to core Rhodospirillales[62].

Colored backgrounds correspond to sampling sites and host species in Figure 1 and samples in Figure 2. The tree scale represents the average number of substitutions per site. The default outgroup of the GTDB-tk *de novo* workflow was used to root the tree.

Saccharisymbium spp. thrive on sugars under oxic and anoxic conditions

We were able to generate high-quality metagenome-assembled genomes (MAGs) for Saccharisymbium spp. from seven of the eight host species that contained symbiont 16S rRNA genes, and a medium-quality draft for the symbionts of *O. sp.* from Okinawa. Of the high-quality MAGs of Saccharisymbium spp. 13 had a genome size of 4.19-4.63 Mbp and a GC content of 68.25-70.08%. Exceptions were one MAG from *O. sp.* nrtenuissimus and one MAG from *I. sp.* ULE that had a larger size of 5.2 Mbp and a GC content of 65%. The high-quality MAGs had a completeness above 98.26% and a contamination less than 1.2% except one case at 4.49% (extended information in Supplement Table S4a). Coding densities assigned with checkM were at 88.8-92.1%. We used COG categories to identify to which major metabolic functions Saccharisymbium spp. invest their genome space. We compared the 19 MAGs of Saccharisymbium spp. to its nine relatives from the GTDB taxonomy sister-clade and to 149 representatives of most alphaproteobacterial families (Figure 4, Supplement Figure S3). NMDS analysis showed that Saccharisymbium spp. have an ordinary metabolic profile compared to other Rhodospirillales and Alphaproteobacteria in general as they cluster within all family representatives (Figure 4). However, Rhodospirillales are a versatile order that encompasses

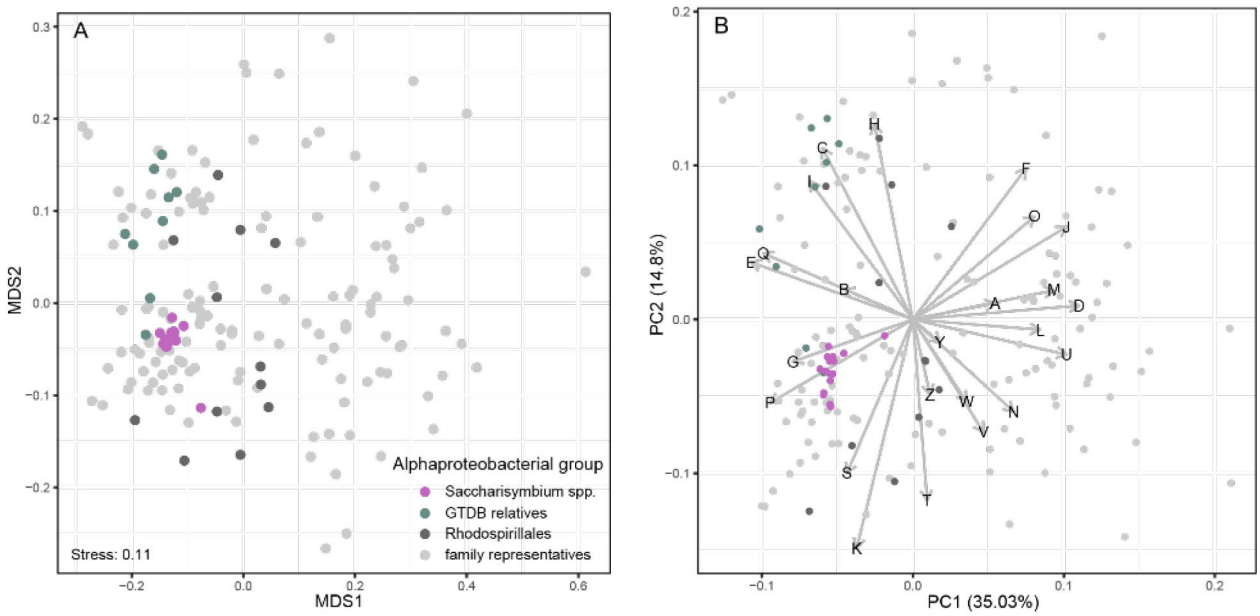


Figure 4: Based on relative abundances of COG categories, Saccharisymbium spp. functional profiles cluster within diverse Rhodospirillales and are distinct from GTDB next relatives.
A: Non-metric dimensional scaling (NMDS) of the functional profiles based on relative abundance of COG categories in the genomes of Saccharisymbium species (purple), relatives from GTDB (green), Rhodospirillales (dark grey) and representatives of alphaproteobacterial families (light grey).
B: Principal component analysis with driving COG categories of the same dataset as A.

bacteria with diverse metabolisms. We observed that NMDS analyses put the long-read MAG of sample 4410 in a position slightly apart from its related short-read MAGs. A reason for this might be the inclusion of the short contigs that potentially belong to other symbionts as also indicated in the GC average and skew in the circular genome representation (Figure 5). Caution must be taken when analyzing the metabolism of Saccharisymbium spp. based on the long-read MAG, especially the short contigs. We identified COG categories G and P to drive the position and clustering of Saccharisymbium spp. compared to other Alphaproteobacteria. These categories comprise genes for carbohydrate metabolism and transport (COG G) and inorganic ion transport and metabolism (COG P). Besides COGs with unknown function (COG S), amino

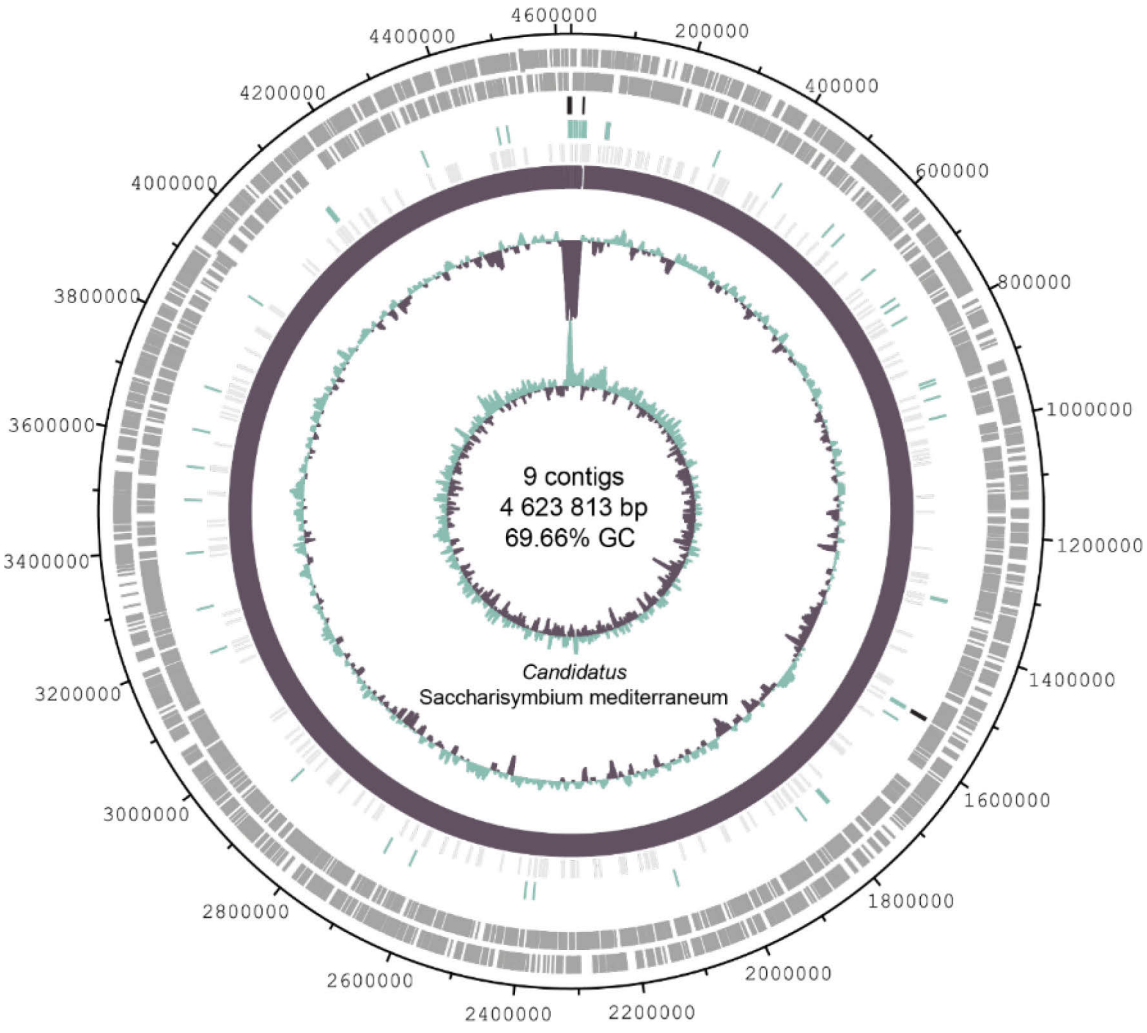


Figure 5: Circular genome representation of the S. mediterraneum long-read metagenome-assembled genome.
Circles from outside to inside represent base pairs, coding sequences (CDS) on the forward and on the reverse strand in grey, ribosomal 5S, 16S and 23S rRNA gene sequences in black, tRNAs in blue, contigs in purple, deviation from average GC content and GC skew in blue (positive) and purple (negative). Position 1 starts at contig 1.

acid metabolism and transport (COG E), carbohydrate metabolism and transport (COG G) and energy production, and conversion (COG C) are the COG categories where Saccharisymbium spp. invest most genome space. Compared to GTDB relatives, Saccharisymbium spp. invest especially in signal transduction (COG T), cell motility (COG N) and carbohydrate metabolism and transport (COG G; Supplement Figure S3). We went on to study the metabolism of Saccharisymbium spp. in more detail.

We investigated the genomic potential and metabolic gene expression of S. mediterraneum and S. belizense MAGs to study their role and function in the symbiosis (libraries 4515 and 4410). We identified metabolic pathways of relevance based on transcriptomic and proteomic expression data (detailed lists provided at zenodo.org/record/6514798). Main pathways of the potential cell metabolism and expressed pathways are depicted in Figure 6. Overall, Saccharisymbium spp. MAGs investigated in detail were 4.3-5.2 Mbp in size, encoded for 4127-4447 protein coding genes and 357-362 pathways according to Pathway tools. Ribosomal RNA was present in all samples and also tRNAs and reactions charging all essential amino acids. Our transcriptomic data covered 9-25% of the encoded genes, proteome data up to 2.3%.

Saccharisymbium spp. are likely heterotrophs living off saccharides and derivatives under oxic and anoxic conditions (Figure 6). As COG category investigation indicated, the symbionts focus their metabolism around sugar degradation. Both metagenomic potential and expression data points towards an import of sugars and sugar acids, degradation via glycolysis and complete oxidation of acetyl-CoA unit into the citric acid cycle (tricarboxylic acid cycle, TCA), the base for anabolism and energy conservation. Sugar importers and degradation pathways annotated in the genome are manifold and range from sugars such as sucrose, trehalose, xylose, galactose and lactose over sugar alcohols such as xylitol and myo-inositol to sugar acids. The genetic potential indicates that carbohydrates can be degraded via glycolysis, the pentose phosphate pathway for 5C-substrates and pathways specific to the substrates and products are channeled to the TCA cycle. Furthermore, carboxylates can be degraded and products fueled into the TCA cycle. We detected genes for both oxic and anoxic conditions expressed in the transcriptomes and proteomes. The ability to respire with both oxygen and nitrous oxide as terminal electron acceptors would render the symbionts independent of the environmental conditions generation of ATP. We were surprised to detect not only structural components for the typical F-type ATP synthase but also a V-type ATP synthases. The latter might play a role in controlling acidification of compartments in or around the cells.

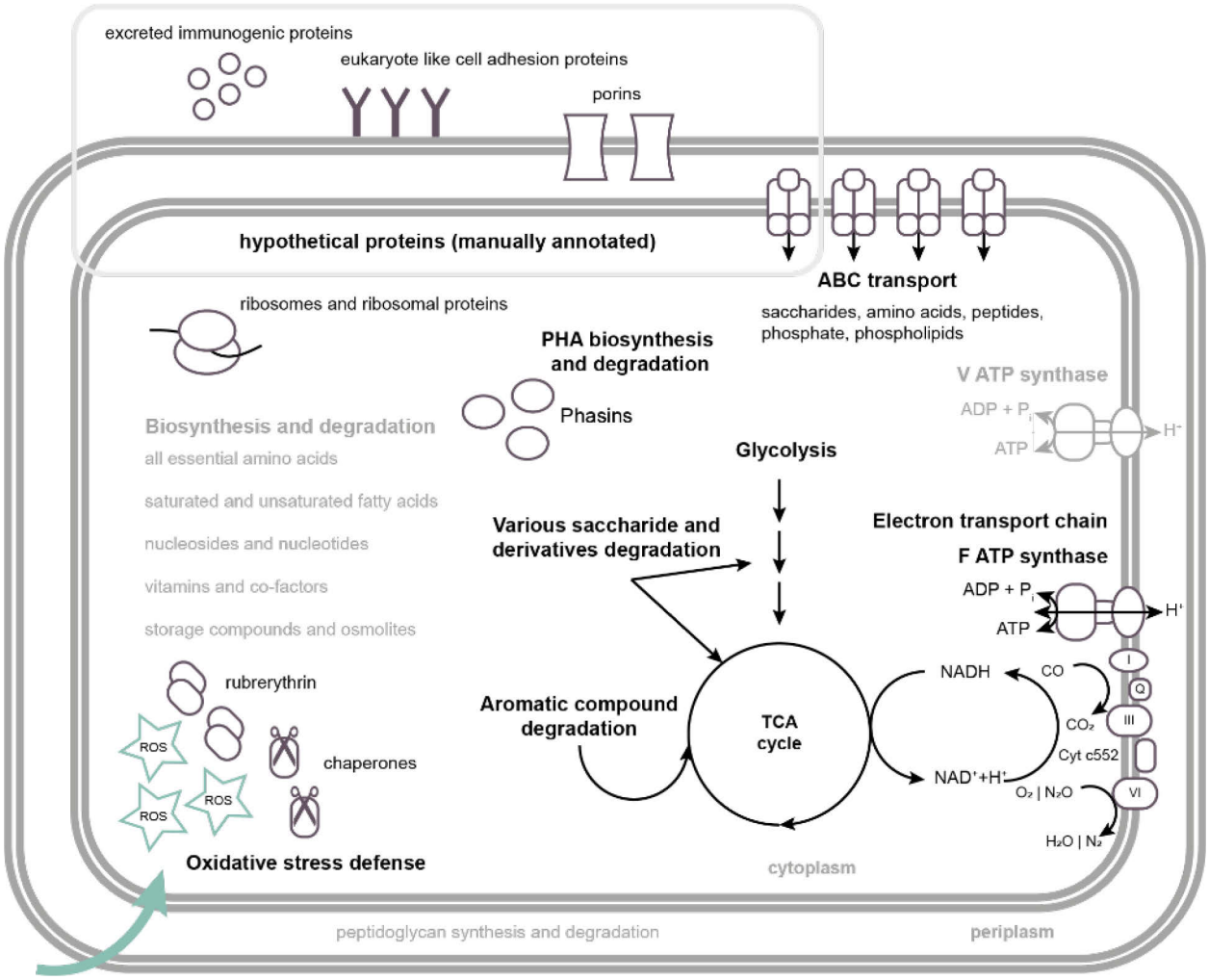


Figure 6: Saccharisymbium spp. expressed heterotrophic saccharide degradation in both oxic and anoxic conditions.
The reconstructed metabolism and cellular overview is based on metagenome-assembled genomes and their annotation with Prokka and RAST from the two host species *Olavius ilvae* from Sant'Andrea on Elba, Italy and *Inanidrilus* sp. ULE from Curlew Cay, Belize. Metabolic functions with expression evidence from transcriptomic and proteomic data are highlighted in black, and additional genomic potential is stated in grey. ROS: reactive oxygen species; TCA: tricarboxylic acid cycle; NAD: nicotinamide adenine dinucleotide; ATP: adenosine triphosphate; Q: quinone.

We found that Saccharisymbium spp. are genetically capable of decomposing plant derived substrates and express relevant enzymes. One-carbon compound metabolism in Saccharisymbium spp. includes carbon monoxide oxidation to CO_2 under aerobic conditions, which could come from degrading seagrass meadows in the vicinity of the worms[64]. We found degradation pathways for aromatic compounds such as catechol and protocatechuate encoded in the genome of Saccharisymbium spp. We could identify expression of subunits of key enzymes for aerobic and anaerobic aromatic compound degradation pathways in the transcriptome data. Among the expressed genes for aromatic compound degradation were

enzymes for toluene degradation. Together with the degradation of saccharides that could be available from surrounding seagrass meadows, we hypothesize that the symbionts might be able to scavenge plant derived substrates from the environment and use these as external resources to broaden the substrate spectrum of the worm[65]. To proof that external carbohydrates can be degraded and are used as carbon source, isotope fingerprints of the bacterial proteins in comparison to host and environment fingerprints would be needed.

Saccharisymbium spp. code for denitrification and are genetically able to reduce nitrate to nitrite and nitric oxide to nitrous oxide. Nitrous oxide may be further used as terminal electron acceptor for respiration to dinitrogen. Saccharisymbium spp. can take up ammonium via glutamate, and glutamine synthesis from glutamate. A urea cycle points towards the necessity of nitrogen waste excretion indicating that nitrogen compounds are not limited for the symbionts. Saccharisymbium spp. encode a variety of genes for sulfur compound metabolism. These include genes for sulfite oxidation to sulfate (soeABC), elemental sulfur reduction and most likely assimilatory sulfate reduction and hydrogen sulfide formation. We could not detect genes for sulfur metabolism highly expressed in both transcriptome and proteome and their role for Saccharisymbium spp. metabolism needs further investigation. We detected reactive oxygen species degradation proteins highly expressed from Saccharisymbium spp. genomes and conclude that the symbionts are under constant oxygen stress. Rubrerythrin was among the most highly expressed proteins in the transcriptome. It is a cytoplasmic protein typical for anaerobic sulfate-reducing bacteria providing oxidative stress protection via catalytic reduction of intracellular hydrogen peroxide.

Saccharisymbium spp. seem to use cell surface structures based on saccharides as the medium of choice to communicate with their hosts. The genomes of the investigated MAGs did not contain any evidence for secretion systems. Instead, genes for extracellular lipopolysaccharide and polysaccharide formation such as M-antigens are encoded in the genome and also expressed. Highly expressed hypothetical proteins contain domains typical for eukaryotic protocadherin like proteins. These are usually involved in cell adhesion in neurological contexts. Hence, Saccharisymbium spp. seem to be equipped with several cell surface tools to interact with or evade detection by the host.

Overall, Saccharisymbium spp. seem to encode a rather self-sufficient metabolism. The bacteria can synthesize and degrade peptidoglycan as a structural component for their cell walls. Saturated and unsaturated fatty acids and oleates can be synthesized and degraded. All essential

amino acids can be converted or synthesized *de novo* and degraded. Saccharisymbium can produce and degrade a variety of storage products. Among these are glycogen and glycan, trehalose, glycine betaine, polyhydroxyalkanoic acid and polyphosphates. Multiple vitamins such as thiamine and folic acid can be salvaged or synthesized. Notably, multiple copies of B12 import binding proteins are encoded. Another indication for a recent uptake as symbiont or a potential host-independent lifestyle are the multiple encoded genes for flagellum formation. Cultivation attempts informed by the metabolic potential could show whether Saccharisymbium spp. can thrive independent of and outside a hosts.

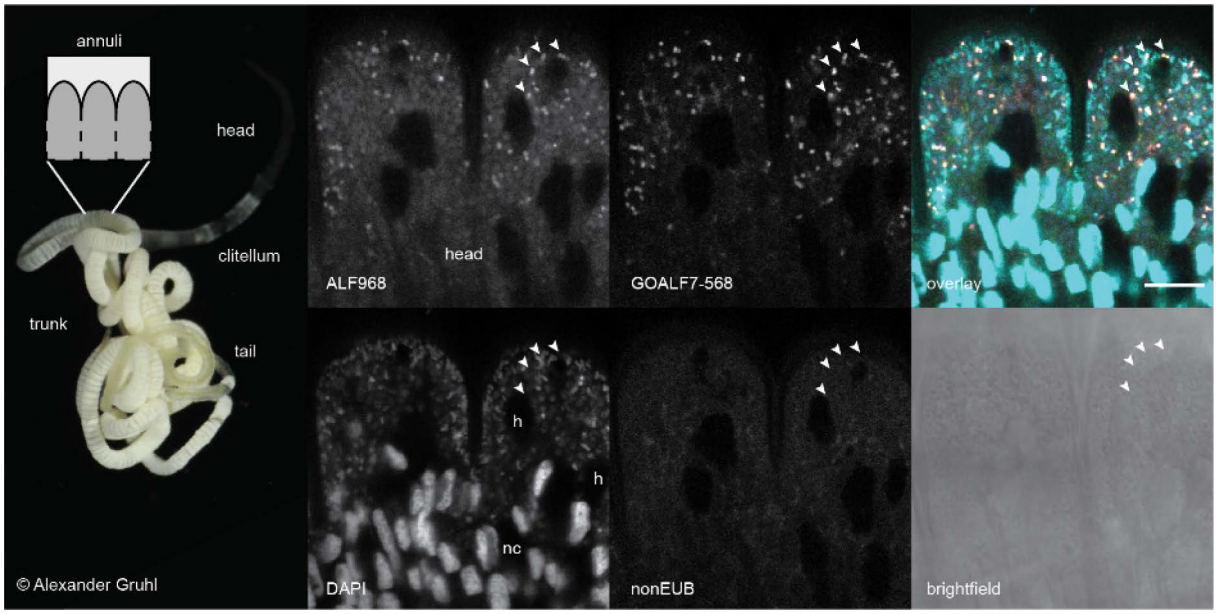


Figure 7: S. mediterraneum resides in the subcuticular symbiont layer based on 16S rRNA specific fluorescence *in situ* hybridization.
Specimens of *O. ilvae* from Elba, Italy were hybridized with a Saccharisymbium spp. specific probe (GOALF7-568, yellow), an alphaproteobacterial general probe (ALF968, magenta) and a non-binding probe (nonEUB, green). DNA was counterstained with DAPI (cyan). The colored image shows the overlay with probes colored as stated in brackets. White arrows point to S. mediterraneum cells. h – host cell, nc – host nuclei. Scale bar: 10 µm. For additional FISH images see Supplement Section 5.

S. mediterraneum inhabits the subcuticular symbiont layer throughout the host's body

To determine the distribution of Saccharisymbium spp. inside whole host individuals, we analyzed specimens of *O. ilvae* from Elba, Italy, using fluorescence *in situ* hybridization (FISH) with double-labelled oligonucleotide probes specific to the bacterial 16S rRNA gene (Figure 7, Supplement Figure S4-5). S. mediterraneum signals were located in the subcuticular symbiont layer throughout the worms' body. They co-localized with the main gammaproteobacterial symbionts in the trunk region of the worms but were also present and co-localizing with secondary gamma- and deltaproteobacterial symbionts in the head of the worms[66].

S. mediterraneum signals were of coccoid shape with a signal diameter of 757 ± 82 nm (n=31). S. mediterraneum seems to be more abundant towards the cuticle but was also detected in regions more closely to the base of the epidermal host cells. Proteome expression data indicates that S. mediterraneum has a much lower biomass compared to the main Gammaproteobacterium. This is likely caused by both its smaller size and lower abundance. Similar to the prevalence of Saccharisymbium spp. in metagenomic datasets, we could detect S. mediterraneum in roughly half of the *Olavius ilvae* specimens that we hybridized. We expect that our findings for S. mediterraneum also apply to Saccharisymbium spp. from other sampling locations and host species but investigation on respective samples would be necessary for proof.

A globally distributed genus with local and host specific species

We collected gutless oligochaete species hosting Saccharisymbium spp. at globally distributed, tropical and sub-tropical sampling sites in coastal regions of all major oceans. Location data of related bacteria from 16S rRNA and GTDB phylogeny analyses points to a similar global distribution of these samples (Figure 1). In contrast to the gutless oligochaete samples from sediments, potentially free-living relatives from the database were collected from open ocean water bodies, hydrothermal vents, associated to sponges and even from alpine spring ground water (Supplement Table S5). The species and genus of Saccharisymbium spp. appear to be specific to gutless oligochaete. The family and next relatives were mainly collected in the ocean but might not be restricted to marine or saline environments in general. To further investigate the potential distribution of Saccharisymbium spp. relatives, we analyzed the global occurrence of 16S rRNA sequences related to Saccharisymbium spp. in publicly available sequence read archives (SRAs). Using the IMNGS platform, we screened 500048 SRAs against the three 99% similarity centroid of the high quality Saccharisymbium 16S rRNA sequences (SRA hit lists provided at zenodo.org/record/6514783). We detected SRA hits with at least 99% similarity in 2, 12 and 10 of the SRAs and hits with at least 97% similarity in 32, 55 and 54 of the SRAs. Samples for which geographic data was available are mapped in Figure 1. All SRAs with hits were sampled at marine and freshwater aquatic locations, in sediments or soil or connected to corals. Hits to Saccharisymbium spp. made up at maximum 0.09% of the total reads per SRA at 99% similarity and 0.98% of the reads at 97% similarity. SRAs with highest read abundances were sampled at the Bahamas (SRR3185594) in an investigation of the formation of ooids, sedimentary grains, and in ballast sediment (SRR6081896) of unknown origin, which was studied for its microbial community. We conclude that bacterial species related to Saccharisymbium spp. might not exclusively be gutless oligochaete symbionts as they are

globally distributed in tropical and subtropical marine samples, but have also been detected in freshwater aquatic systems and soil samples. However, Saccharisymbium spp. related hits had low abundances throughout all positive SRAs and do not seem to be a dominant community member in any of the sampled environments.

Despite the symbionts being location- and hence host-specific, co-evolution of the symbionts with their hosts from one uptake event is unlikely, since host and symbiont phylogenetic patterns are not congruent. We observed that not all host specimen of the individual species have Saccharisymbium spp. as symbionts and also not all host species sampled at the specific sampling sites (Supplement Table S1)[14]. More samples would be necessary to estimate the prevalence of Saccharisymbium spp. per host species and sampling sites. We conclude that Saccharisymbium spp. might play a beneficial role for the host species but is unlikely an obligate member of the associations.

Concluding discussion

This study describes a novel family of Alphaproteobacteria in the order of Rhodospirillales. This novel family is a symbiont-exclusive clade of six distinct and gutless oligochaete specific species. Only the use of untargeted metagenomics approaches enabled the detection of this symbiont clade[14]. Phylotypes of this clade have been overlooked even in the host species *O. ilvae* that has been focus of 16S rRNA gene-based symbiont community studies in the past decades[66]. We found species of the proposed genus of *Candidatus* Saccharisymbium at globally distributed sampling sites. Most symbiont species appear host and location specific because host species are often only recovered from one sampling location. In one case, one host species has been sampled from two locations and this species hosts two distinct symbionts that appear to be location specific. However, a 100% identical MAG to one of these symbiont species has been identified from another host species at a close but distinct sampling site. Overall, it is currently not possible to disentangle whether specificity to their host species or sampling locations drive distribution patterns.

Saccharisymbium spp. focus their metabolism on the import and degradation of carbohydrates and are able to respire under oxic and anoxic conditions. The genomic potential indicates that Saccharisymbium spp. are able to degrade aromatic compounds. Taken together, we hypothesize that the symbionts are able to make use of compounds potentially available from plant material of seagrass meadows in the vicinity of the worms' habitats. Seagrass meadows

in the Mediterranean Sea and the Caribbean have been shown to be a source of high amounts of sucrose and other carbohydrates such as trehalose and glucose in lower amounts[65]. One hypothesis why these carbohydrates are available in high abundances is that bacteria are hindered to degrade these under anaerobic conditions by the presence of phenolic compounds. Saccharisymbium spp. might make use of the carbohydrates since they seem able to degrade aromatic compounds including phenolics also under anaerobic conditions. The heterotrophically acquired energy and carbon compounds could serve as an additional source for the whole symbiotic community and might enable the worms to inhabit a broader array of microniches. The alphaproteobacterial symbionts seem able to incorporate worm waste products such as carboxylates into their heterotrophic metabolism and thus might provide redundancy to the secondary Deltaproteobacteria in terms of waste product management. However, further investigation on the relevance of carbon sources from the environment or the host are necessary.

We were surprised that most Saccharisymbium species do not seem to be an obligate symbiont and present in all individuals of their respective gutless oligochaete host species. Notably, *O. prodigus* was the only species were all samples hosted Saccharisymbium spp. *O. prodigus* has a simple symbiont community with S. australicum as the only secondary symbiont in addition to the main Gammaproteobacterium, which might make it essential for its host[14]. Given that all other Saccharisymbium species manage to colonize host species around the globe, they likely have little negative impact but rather a commensal or even beneficial contribution to the symbiosis.

The metabolisms of S. mediterraneum and S. belizense as representatives for the genus appear rather self-sufficient and seem to encode a *de novo* synthesis of relevant cell compounds such as amino acids, vitamins and cell structural components. S. mediterraneum and S. belizense code for a variety of heterotrophic metabolic pathways to sustain a living and are able to run their own ATP synthase coupled with a functional respiration chain in oxic and anoxic conditions. Given the broad metabolic capabilities and the presence of genes for flagella formation, we expect free-living close relatives of Saccharisymbium spp. to exist. Results from sequence read archive searches indicate that close relatives of Saccharisymbium spp. are present in marine sediments and are associated with corals (SRR3145910) and *Euprymna* bobtail squids (ERR1777341). We can, however, not exclude that the related hits from marine sediment samples stem from invertebrate hosts.

Distribution patterns suggest that *O. ilvae* from Elba co-occur with *O. algarvensis* in similar but not identical niches, as *O. algarvensis* dominates sites on Elba but does not occur in the neighboring island Pianosa where *O. ilvae* is abundant[8, 14]. Both hosts have a similar set of symbionts except for the S. mediterraneum that is only present in *O. ilvae* and a Spirochaetia species that is only present in *O. algarvensis*. The exclusive pattern of Saccharisymbium spp. and the Spirochaetia in the two host species suggest a decisive role for these two symbiont groups in host niche realization. Such a niche differentiation might also be supported by the fact that of the six host species of Saccharisymbium only one also hosts the more widely distributed Spirochaetia symbiont.

Overall, our study provides insights in a widely distributed symbiont clade that has been overlooked in the past and that extends a chemoautotrophic symbiosis with heterotrophic capabilities. Saccharisymbium illustrates the important role secondary symbionts can have in the ecophysiology of similar co-occurring host species from the same animal clade.

Data and code availability

Raw metagenomic sequences, symbiont marker genes and symbiont MAGs generated in this study will be deposited in the European Nucleotide Archive (ENA) upon peer-review submission and are currently available upon request.

The scripts that were used for data visualization concerning mapping and COG profiling are available at github.com/TinaEnd/Alpha7_Saccharisymbium.

Phylogenetic 16S rRNA gene and MAG based trees are available in project Alpha7_Saccharisymbium at itol.embl.de/shared/tenders.

Acknowledgements

We thank Christian Lott, Miriam Weber and the HYDRA Marine Sciences team for sampling and field assistance. We also thank the Elba field teams from spring 2019 and 2020 and fall 2019, and especially Yui Sato for sampling. We are grateful to the Max Planck Society supporting this work and the Max Planck Genome Centre Cologne for conducting the sequencing. LC-MS/MS measurements were made in the Molecular Education, Technology, and Research Innovation Center (METRIC) at North Carolina State University. Funding was provided by the German Academic Exchange Service DAAD (M.J.). The U.S. National Science Foundation (grant IOS 2003107 to M.K.) supported this work. We acknowledge the thorough

attempts of Marlene Jensen to obtain protein stable isotope data for S. mediterraneum. Silke Wetzel, Miriam Sadowski and Wiebke Ruschmeier provided excellent technical support in the laboratory. This work is contribution XXX from the Carrie Bow Cay Laboratory, Caribbean Coral Reef Ecosystem Program, National Museum of History, Washington DC.

References

1. Dubilier N, Bergin C, Lott C. Symbiotic diversity in marine animals: The art of harnessing chemosynthesis. Nature Reviews Microbiology. 2008;6(10):725-40.
2. Giere O. Studies on marine Oligochaeta from Bermuda, with emphasis on new *Phallodrilus* species (Tubificidae). Cahiers de Biologie Marine. 1979;20:301-14.
3. Erséus C. *Inanidrilus bulbosus* gen. et sp. n., a marine tubificid (Oligochaeta) from Florida, USA. Zoologica Scripta. 1979;8(1-4):209-10.
4. Dubilier N, Blazejak A, Rühland C. Symbioses between bacteria and gutless marine oligochaetes. Molecular basis of symbiosis. 2005:251-75.
5. Dubilier N, Giere O, Distel DL, Cavanaugh CM. Characterization of chemoautotrophic bacterial symbionts in a gutless marine worm (Oligochaeta, Annelida) by phylogenetic 16S rRNA sequence analysis and *in situ* hybridization. Applied and Environmental Microbiology. 1995;61(6):2346-50.
6. Dubilier N, Amann R, Erséus C, Muyzer G, Park S, Giere O, et al. Phylogenetic diversity of bacterial endosymbionts in the gutless marine oligochete *Olavius loisae* (Annelida). Marine Ecology Progress Series. 1999;178:271-80.
7. Zimmermann J, Wentrup C, Sadowski M, Blazejak A, Gruber-Vodicka HR, Kleiner M, et al. Closely coupled evolutionary history of ecto-and endosymbionts from two distantly related animal phyla. Molecular ecology. 2016;25(13):3203-23.
8. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature. 2006;443(7114):950-5.
9. Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. Proceedings of the National Academy of Sciences. 2012;109(19):E1173-E82.
10. Dubilier N, Mülders C, Ferdelman T, de Beer D, Pernthaler A, Klein M, et al. Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. Nature. 2001;411(6835):298-302.
11. Kleiner M, Dong X, Hinzke T, Wippler J, Thorson E, Mayer B, et al. Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. Proceedings of the National Academy of Sciences. 2018;115(24):E5576-E84.
12. Bergin C, Wentrup C, Brewig N, Blazejak A, Erséus C, Giere O, et al. Acquisition of a novel sulfur-oxidizing symbiont in the gutless marine worm *Inanidrilus exumae*. Applied and environmental microbiology. 2018;84(7):e02267-17.
13. 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. Applied and Environmental Microbiology. 2006;72(8):5527-36.
14. Mankowski A, Kleiner M, Erséus C, Leisch N, Sato Y, Volland J-M, et al. Highly variable fidelity drives symbiont community composition in an obligate symbiosis. bioRxiv. 2021.
15. McCutcheon JP, McDonald BR, Moran NA. Origin of an alternative genetic code in the extremely small and GC–rich genome of a bacterial symbiont. PLoS genetics. 2009;5(7):e1000565.
16. Kaur R, Shropshire JD, Cross KL, Leigh B, Mansueto AJ, Stewart V, et al. Living in the endosymbiotic world of *Wolbachia*: A centennial review. Cell Host & Microbe. 2021.
17. Karimi E, Keller-Costa T, Slaby BM, Cox CJ, da Rocha UN, Hentschel U, et al. Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable Alphaproteobacteria. Scientific reports. 2019;9(1):1-15.
18. Karimi E, Slaby BM, Soares AR, Blom J, Hentschel U, Costa R. Metagenomic binning reveals versatile nutrient cycling and distinct adaptive features in alphaproteobacterial symbionts of marine sponges. FEMS Microbiology Ecology. 2018;94(6):fiy074.
19. Gruber-Vodicka HR, Leisch N, Kleiner M, Hinzke T, Liebeke M, McFall-Ngai M, et al. Two intracellular and cell type-specific bacterial symbionts in the placozoan *Trichoplax* H2. Nature Microbiology. 2019;4(9):1465-74.

20. Baker LJ, Reich HG, Kitchen SA, Grace Klinges J, Koch HR, Baums IB, et al. The coral symbiont *Candidatus* Aquarickettsia is variably abundant in threatened Caribbean acroporids and transmitted horizontally. The ISME Journal. 2022;16(2):400-11.
21. Gruber-Vodicka HR, Dirks U, Leisch N, Baranyi C, Stoecker K, Bulgheresi S, et al. *Paracatenula*, an ancient symbiosis between thiotrophic Alphaproteobacteria and catenulid flatworms. Proceedings of the National Academy of Sciences. 2011;108(29):12078-83.
22. Jäckle O, Seah BK, Tietjen M, Leisch N, Liebeke M, Kleiner M, et al. Chemosynthetic symbiont with a drastically reduced genome serves as primary energy storage in the marine flatworm Paracatenula. Proceedings of the National Academy of Sciences. 2019;116(17):8505-14.
23. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012;19(5):455-77.
24. Nikolenko SI, Korobeynikov AI, Alekseyev MA, editors. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC genomics; 2013: Springer.
25. Mankowski A. From genomes to communities - Evolution of symbionts associated with globally distributed marine invertebrates: Universität Bremen; 2021.
26. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research. 2015;25(7):1043-55.
27. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nature biotechnology. 2019;37(5):540-6.
28. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.
29. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of *de novo* genome assemblies. Bioinformatics. 2015;31(20):3350-2.
30. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.
31. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015;32(1):268-74.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10.
33. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press; 2020.
34. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic acids research. 2019;47(W1):W256-W9.
35. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: A toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints; 2016. Report No.: 2167-9843.
36. Bushnell B. BBmap. v38.34 ed2010.
37. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
38. Hinzke T, Kleiner M, Markert S. Centrifugation-based enrichment of bacterial cell populations for metaproteomic studies on bacteria–invertebrate symbioses. Microbial Proteomics: Springer; 2018. p. 319-34.
39. Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nature methods. 2009;6(5):359-62.
40. Hinzke T, Kouris A, Hughes R-A, Strous M, Kleiner M. More is not always better: Evaluation of 1D and 2D-LC-MS/MS methods for metaproteomics. Frontiers in microbiology. 2019;10:238.
41. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, et al. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Molecular & cellular proteomics. 2005;4(12):2010-21.
42. Mordant A, Kleiner M. Evaluation of sample preservation and storage methods for metaproteomics analysis of intestinal microbiomes. Microbiology spectrum. 2021;9(3):e01877-21.
43. Jensen M, Wippler J, Kleiner M. Evaluation of RNA later as a field-compatible preservation method for metaproteomic analyses of bacterium-animal symbioses. Microbiology spectrum. 2021;9(2):e01429-21.
44. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. Journal of proteome research. 2006;5(9):2339-47.
45. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic acids research. 2014;42(D1):D206-D14.

46. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific reports. 2015;5(1):1-6.
47. Eichinger V, Nussbaumer T, Platzer A, Jehl M-A, Arnold R, Rattei T. EffectiveDB — Updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. Nucleic acids research. 2016;44(D1):D669-D74.
48. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic acids research. 1997;25(17):3389-402.
49. Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, et al. Pathway Tools version 23.0 update: Software for pathway/genome informatics and systems biology. Briefings in bioinformatics. 2021;22(1):109-26.
50. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic acids research. 2014;42(D1):D459-D71.
51. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic acids research. 2015;43(D1):D261-D9.
52. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. 2019;47(D1):D309-D14.
53. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular biology and evolution. 2017;34(8):2115-22.
54. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics. 2010;11(1):1-11.
55. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015;12(1):59-60.
56. Junier T, Zdobnov EM. The Newick utilities: High-throughput phylogenetic tree processing in the Unix shell. Bioinformatics. 2010;26(13):1669-70.
57. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Scientific reports. 2016;6(1):1-9.
58. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: A versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
59. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: A software environment for sequence data. Nucleic acids research. 2004;32(4):1363-71.
60. Manz W, Amann R, Ludwig W, Wagner M, Schleifer K-H. Phylogenetic oligodeoxynucleotide probes for the major subclasses of proteobacteria: Problems and solutions. Systematic and applied microbiology. 1992;15(4):593-600.
61. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological image analysis. Nature Methods. 2012;9(7):676-82.
62. Hördt A, López MG, Meier-Kolthoff JP, Schleuning M, Weinhold L-M, Tindall BJ, et al. Analysis of 1,000+ type-strain genomes substantially improves taxonomic classification of Alphaproteobacteria. Frontiers in microbiology. 2020;11:468.
63. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications. 2018;9(1):5114.
64. Kleiner M, Wentrup C, Holler T, Lavik G, Harder J, Lott C, et al. Use of carbon monoxide and hydrogen by a bacteria–animal symbiosis from seagrass sediments. Environmental microbiology. 2015;17(12):5023-35.
65. Sogin EM, Michellod D, Gruber-Vodicka HR, Bourceau P, Geier B, Meier DV, et al. Sugars dominate the seagrass rhizosphere. Nature Ecology & Evolution. 2022.
66. Ruehland C, Blazejak A, Lott C, Loy A, Erséus C, Dubilier N. Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. Environmental Microbiology. 2008;10(12):3404-16.

Chapter 2 | *Candidatus* Saccharisymbium – a globally distributed symbiont

A novel family of Rhodospirillales (Alphaproteobacteria) in symbiosis with globally distributed gutless marine worms (Oligochaeta, Annelida)

Supplementary information

Tina Enders[1], Grace D'Angelo[1], Marlene Jensen[2], Manuel Kleiner[2], Nikolaus Leisch[1], Anna Mankowski[1,3], Dolma Michellod[1], Nicole Dubilier[1], Harald R. Gruber-Vodicka[1]

[1] Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

[2] Department of Plant & Microbial Biology, North Carolina State University, Raleigh 27695, North Carolina, USA

[3] Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Corresponding authors:
Nicole Dubilier, ndubilie@mpi-bremen.de,
Harald R. Gruber-Vodicka, hgruber@mpi-bremen.de

Table 1a-e: Overview of samples and metadata used in this study (pages 78 and 79).

Library	Host species	Ocean	Country/Region	Island/City	Bay/Spot	Organic input
a) short-read metagenome samples						
3557_AD	IspULE	Atlantic	Belize	CurlewCay		corals
3557_BK	IspULE	Atlantic	Belize	CurlewCay		
3557_BL	IspULE	Atlantic	Belize	CurlewCay		
4148_4289_CG	Ofinitimus	Atlantic	Bahamas	LittleDarbyIsland		seagrass_detritus
4148_4289_CH	Ofinitimus	Atlantic	Bahamas	LittleDarbyIsland		seagrass_detritus
3557_AL	Ofinitimus	Atlantic	Belize	TwinCays		seagrass
4148_4289_P	Ofinitimus	Atlantic	Bahamas	NormansPondKey	Bay1	
4148_4289_X	Ofinitimus	Atlantic	Bahamas	LeeStockingIsland		
3557_CF	Oilvae	Mediterranean	Italy	Elba	Pomonte	seagrass
3557_CJ	Oilvae	Mediterranean	Italy	Elba	Pomonte	seagrass
3557_CK	Oilvae	Mediterranean	Italy	Elba	Pomonte	seagrass
3557_CL	Oilvae	Mediterranean	Italy	Elba	Seccheto	seagrass
3557_CN	Oilvae	Mediterranean	Italy	Elba	Pomonte	seagrass
3557_CP	Oilvae	Mediterranean	France	CapFleuri		seagrass_detritus
3557_CQ	Oilvae	Mediterranean	France	CapFleuri		seagrass_detritus
3557_CG	Oilvae	Mediterranean	Italy	Elba	Pomonte	seagrass
3557_CO	Oilvae	Mediterranean	Italy	Elba	Pomonte	seagrass
3557_CM	Oilvae	Mediterranean	Italy	Elba	Seccheto	seagrass
OilvPIA1	Oilvae	Mediterranean	Italy	Pianosa		
OilvPIA2	Oilvae	Mediterranean	Italy	Pianosa		
OilvSAN1	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
OilvSAN2	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
4148_4289_Q	Onralgarvensis	Atlantic	Bahamas	NormansPondKey	Bay1	
4148_4289_T	Onrtenuissimus	Atlantic	Bahamas	NormansPondKey	PatchReef	
4148_4289_B	Oprodigus	Indian	WesternAustralia	RottnestIsland	SalmonBay	
4148_4289_Y	Oprodigus	Indian	WesternAustralia	RottnestIsland	SalmonBay	
3630_D	Ospokinawa2	Pacific	Japan	Okinawa	Tancha	seagrass_corals
3630_F	Ospokinawa2	Pacific	Japan	Okinawa	Yakata	seagrass_corals
3630_A	Ospokinawa2	Pacific	Japan	Okinawa	KinBay	detritus
3630_B	Ospokinawa2	Pacific	Japan	Okinawa	KinBay	detritus
3630_E	Ospokinawa2	Pacific	Japan	Okinawa	Yakata	seagrass_corals
4515_G	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
4515_H	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
4515_I	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
4515_K	IspULE	Atlantic	Belize	CurlewCay	unknown	corals
4515_L	IspULE	Atlantic	Belize	CurlewCay	unknown	corals
b) long-read metagenome samples						
4410	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
c) metatranscriptome samples						
4514_G	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
4514_H	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
4514_I	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
4514_K	IspULE	Atlantic	Belize	CurlewCay	unknown	corals
4514_L	IspULE	Atlantic	Belize	CurlewCay	unknown	corals
d) metaproteome samples						
Oilv_1	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
Oilv_2	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
Oilv_3	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
Oilv_4	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
Oilv_0371_P1	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
Oilv_0374_S3	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
OilvA_P1	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
OilvA_S3	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
e) FISH samples						
ELSP_2019_0062	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
ELSP_2019_0063	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
ELSP_2019_0089	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
ELSP_2019_0213	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass
2020_Elba_0723	Oilvae	Mediterranean	Italy	Elba	SantAndrea	seagrass

Water depth	Latitude	Longitude	Sampling year	Sampling month	16S Ca. S.	Best bin
a) short-read metagenome samples						
-2,0	16,790105	-88,082010	2017	April	yes	3557_AD.autometa.bin1_1
-2,0	16,790135	-88,081800	2017	April	yes	3557_BK.metabat.bin3
-2,0	16,790135	-88,081800	2017	April	yes	3557_BL.dastool.maxbin.bin6_sub
-1,0	23,856100	-76,225400	2013	April	no	
-1,0	23,856100	-76,225400	2013	April	no	
-1,5	16,823600	-88,105925	2017	March	no	
-7,0	23,761700	-76,123600	2002	April	yes	4148_4289_P.metabat.bin5
-3,0	23,773000	-76,105000	2002	April	yes	4148_4289_X.autometa.bin1_1
-0,3	42,743400	10,119400	2016	November	yes	
	42,743400	10,119400	2016	November	yes	
-0,3	42,743400	10,119400	2016	November	yes	
	42,733800	10,192700	2016	October	yes	
-0,3	42,743400	10,119400	2016	November	yes	
-1,5	43,719900	7,4009000	2017	June	yes	3557_CP.autometa.bin1_0
-1,5	43,719900	7,4009000	2017	June	no	
-0,3	42,743400	10,119400	2016	November	no	
-0,3	42,743400	10,119400	2016	November	no	
-4,0	42,733800	10,192700	2016	October	no	
-6,0	42,574217	10,066183	2012	May	yes	OilvPIA1.autometa.bin1_1
-6,0	42,574217	10,066183	2012	May	yes	OilvPIA2.autometa.bin1_1
-6,0	42,808160	10,142104	2014	June	yes	OilvSAN1.autometa.bin1_1
-6,0	42,808160	10,142104	2014	June	yes	OilvSAN2.autometa.bin1_1
-6,0	23,762000	-76,122700	2002	April	yes	4148_4289_Q.dastool.metabat.bin5_sub
-6,0	23,789700	-76,138800	2002	April	yes	4148_4289_T.maxbin.bin3
-1,0	-32,015500	115,513000	1991	January	yes	4148_4289_B.metabat.bin1
-4,0	-32,015500	115,513000	1991	January	yes	
-0,7	26,468700	127,823800	2017	November	yes	
-0,5	26,492400	127,841600	2017	November	yes	3630_F.metabat.bin4
-2,0	26,449800	127,852300	2017	November	no	
-2,0	26,450900	127,853700	2017	November	no	
-2,0	26,492600	127,842100	2017	November	no	
-6	nd	nd	2017	May	yes	4515_G.autometa.bin1_1.fasta
-6	nd	nd	2017	May	no	
-6	nd	nd	2017	May	yes	4515_I.maxbin.bin6.fasta
-2	16,790105	-88,082010	2017	April	yes	4515_K.autometa.bin1_1.fasta
-2	16,790105	-88,082010	2017	April	yes	4515_L.metabat.bin5.fasta
e) long-read metagenome samples						
-6	nd	nd	2019	September	yes	4410_Alpha7_Oilv_flye-2.7.fasta
c) metatranscriptome samples						
-6	nd	nd	2017	May	yes	
-6	nd	nd	2017	May	nd	
-6	nd	nd	2017	May	yes	
-2	16,790105	-88,082010	2017	April	yes	
-2	16,790105	-88,082010	2017	April	yes	
d) metaproteome samples						
-6	nd	nd	2017	May		
-6	nd	nd	2017	May		
-6	nd	nd	2017	May		
-6	nd	nd	2017	May		
-6	nd	nd	2020	October		
-6	nd	nd	2020	October		
-6	nd	nd	2020	October		
-6	nd	nd	2020	October		
e) FISH samples						
-6	nd	nd	2019	March		
-6	nd	nd	2019	March		
-6	nd	nd	2019	March		
-6	nd	nd	2019	April		
-7	42,808647	10,142022	2020	October		

1 Software and tools used

Table S2: Software used in this study with references and online availability.

Software	Reference	Availability
Adobe Illustrator v25.03	Webpage	https://adobe.com/products/illustrator
ARB	[1]	http://www.arb-home.de/
Bandage	[2]	https://rrwick.github.io/Bandage/
bayeshammer	[3]	http://bioinf.spbau.ru/en/spades/bayeshammer
BBtools bbmap.sh v38.90	Webpage, [4]	https://jgi.doe.gov/data-and-tools/bbtools/
BLAST	[5]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
checkM	[6]	https://ecogenomics.github.io/CheckM/
effectiveDB	[7]	https://effectors.csb.univie.ac.at/
eggnog-mapper	[8, 9]	https://github.com/eggnogdb/eggnog-mapper
FastQC	Webpage	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Flye v2.8	[10]	https://github.com/fenderglass/Flye
Fiji (ImageJ v153f51)	Webpage	https://imagej.net/software/fiji/
GTDB-Tk	[11]	https://github.com/Ecogenomics/GTDBTk
Hmmscan (Pfam)	Webpage	https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan http://hmmer.org/
iTOL	[12]	https://itol.embl.de/
IQ-TREE	[13]	http://www.iqtree.org/
mafft	[14]	https://mafft.cbrc.jp/alignment/server/
OligoAnalyzer	Webpage	https://eu.idtdna.com/pages/tools/oligoanalyzer
PhyloFlash	[15]	http://hrgv.github.io/phyloFlash/
Prokka v1.14.5	[16]	https://github.com/tseemann/prokka
QUAST	[17]	http://quast.sourceforge.net/quast
RAST	[18-20]	http://rast.theseed.org/FIG/rast.cgi
R v4.0.4	[21]	https://www.r-project.org/
RStudio v1.4.1106	Webpage	https://www.rstudio.com/
devtools ggfortify ggplot2 maps mapdata plotly reshape2 seqinr tidyr vegan		https://www.r-project.org/nosvn/pandoc/devtools.html https://cran.r-project.org/web/packages/ggfortify/index.html https://ggplot2.tidyverse.org https://cran.r-project.org/web/packages/maps/index.html https://cran.r-project.org/web/packages/mapdata/index.html https://plotly.com/r/ https://cran.r-project.org/web/packages/reshape2/index.html https://cran.r-project.org/web/packages/seqinr/index.html https://tidyr.tidyverse.org/ https://cran.r-project.org/web/packages/vegan/index.html
SPAdes	[22]	http://cab.spbu.ru/software/spades/
VSEARCH	[23]	https://github.com/torognes/vsearch
ZenBlue	Webpage	https://www.zeiss.de/mikroskopie/produkte/mikroskopsoftware/zen.html

2 Proposal of the *Candidatus* Saccharisymbium fam. nov., gen. nov., spp. nov.

We propose *Candidatus* Saccharisymbium gen. nov. (Sa.ccha.ri.sym´bi.um; L. neut. n. *saccharum* sugar; Gr. pref. *sym*- together; Gr. masc. n. *bios* life; *Saccharsymbium* a symbiotic sugar utilizing organism) with the six species (1) S. mediterraneum (me.di.ter'ra.ne.um, L. neut. n. *mediterraneum* Latin name of the Mediterranean Sea), symbiont of the host species *Olavius ilvae* from Elba and Pianosa, Italy, and Cap Fleuri in France; (2) S. australicum (au.stra'li.cum, L. neut. n. *australicum* Neo-Latin for Australia), symbiont of the host species *Olavius prodigus* from Rottnest Island, Australia; (3) S. bahamense (ba.ha.men'se, L. neut. a. *bahamense* from the Bahamas), symbiont of the host species *Olavius finitimus* from Lee Stocking Island and Norman's Pond Cay, Bahamas, (4) S. aequebahamense (ae.que.ba.ha.men'se, L. neutr. a. *aeque* equally, *bahamense* from the Bahamas) symbiont of the host species *Olavius finitimus, Olavius sp.* nralgarvensis, and *Olavius sp.* nrtenuissimus from Norman's Pond Cay, Bahamas; (5) S. okinawense (o.ki.na.wen'se, L. neut. a. *okinawense* from Okinawa), symbiont of the host species *Olavius sp.* okinawa2 from Okinawa, Japan; and (6) S. belizense (be.li.zen'se, L. neut. a. *belizense* from Belize), symbiont of the host species *Inanidrilus sp.* ULE from Curlew Cay, Belize. Basis of assignment: SSU rRNA gene sequences (accession numbers: XXX) and positive match with the specific 16S rRNA oligonucleotide fluorescence *in situ* hybridization probes 5´- GCC TGA CTG ACC GTT CCG -3´ and 5´- TAG ATC GGC AGT ATC AAG CG -3´ for S. mediterraneum. The six species form a novel family and genus taxon within the order of Rhodospirillales in the class of Alphaproteobacteria in the phylum of Proteobacteria. *Candidatus* Saccharisymbium spp. were identified first by Anna Mankowski in a metagenomic dataset of globally distributed gutless oligochaete host species in 2018[24]. All *Candidatus* Saccharisymbium spp. remain uncultivable to date.

3 Phylogenomic placement

Reference sequences used as outgroup for the 16S rRNA gene based phylogenetic reconstruction (Figure 2) were:

- *Bdellovibrio bacteriovorus* HD100 (NC_005363)
- *Desulfovibrio vulgaris* str. Hildenborough (NC_002937)
- *Geobacter sulfurreducens* PCA (NC_002939)

4 Saccharisymbium spp. genomes and metabolism

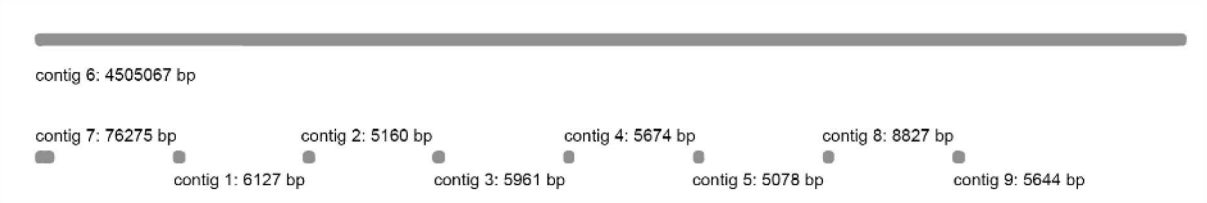


Figure S1: Contigs of the S. mediterraneum long-read metagenome-assembled genome.
The PacBio metagenome-assembled genome from S. mediterraneum sampled in Sant'Andrea Bay on Elba, Italy, has a total size of 4623813 bp assembled on nine contigs. Contig 6 holds 97.43% of the genomic information.

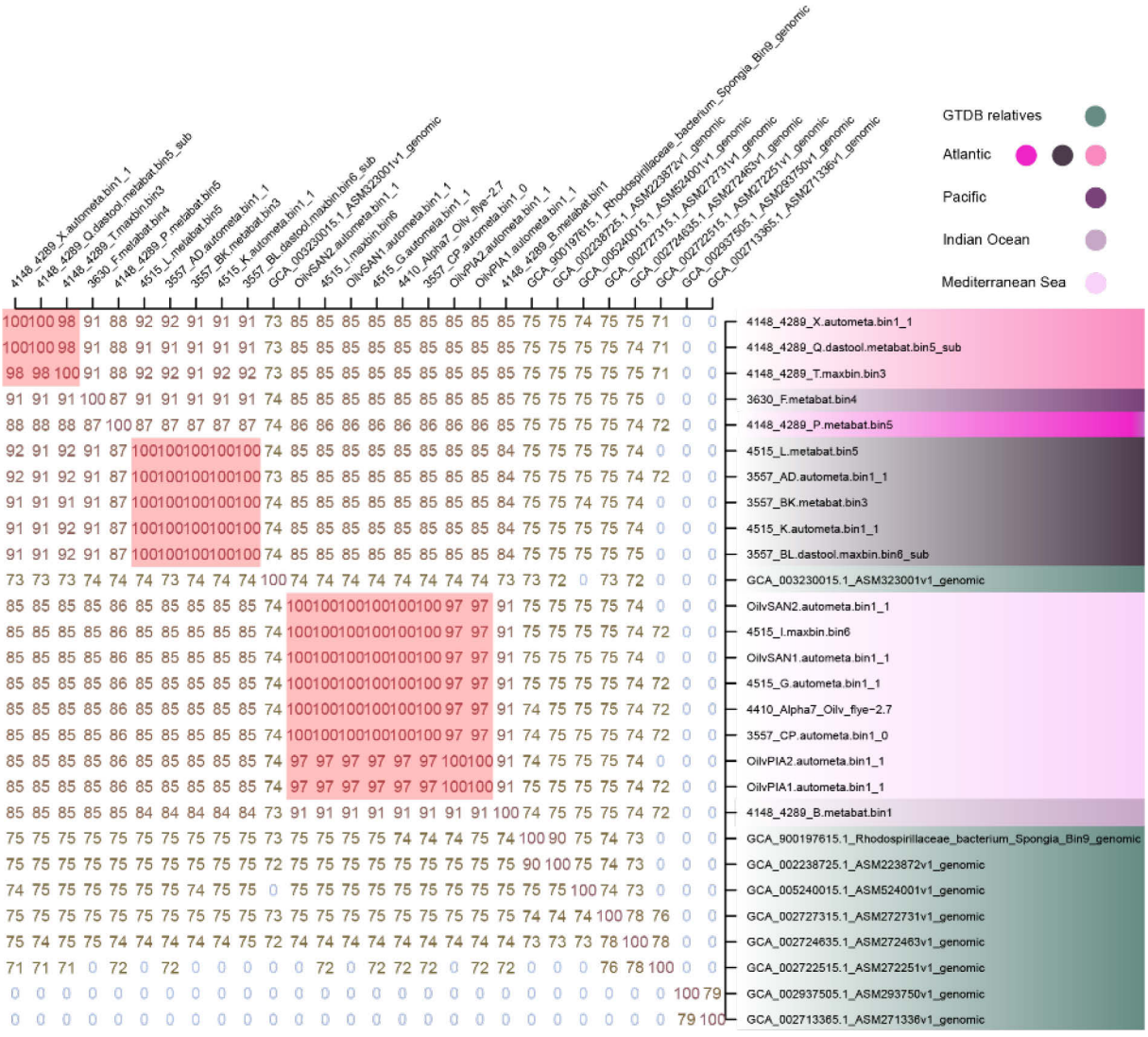


Figure S2: Saccharisymbium MAGs form six distinct species based on average nucleotide identity (ANI) values and are distinct to relatives from the GTDB taxonomy.
Saccharisymbium MAGs (shades of purple)form six species clusters with ANI values above 95%. Relatives from GTDB (green) are less than 75% similar to Saccharisymbium spp. and hence distinct species.

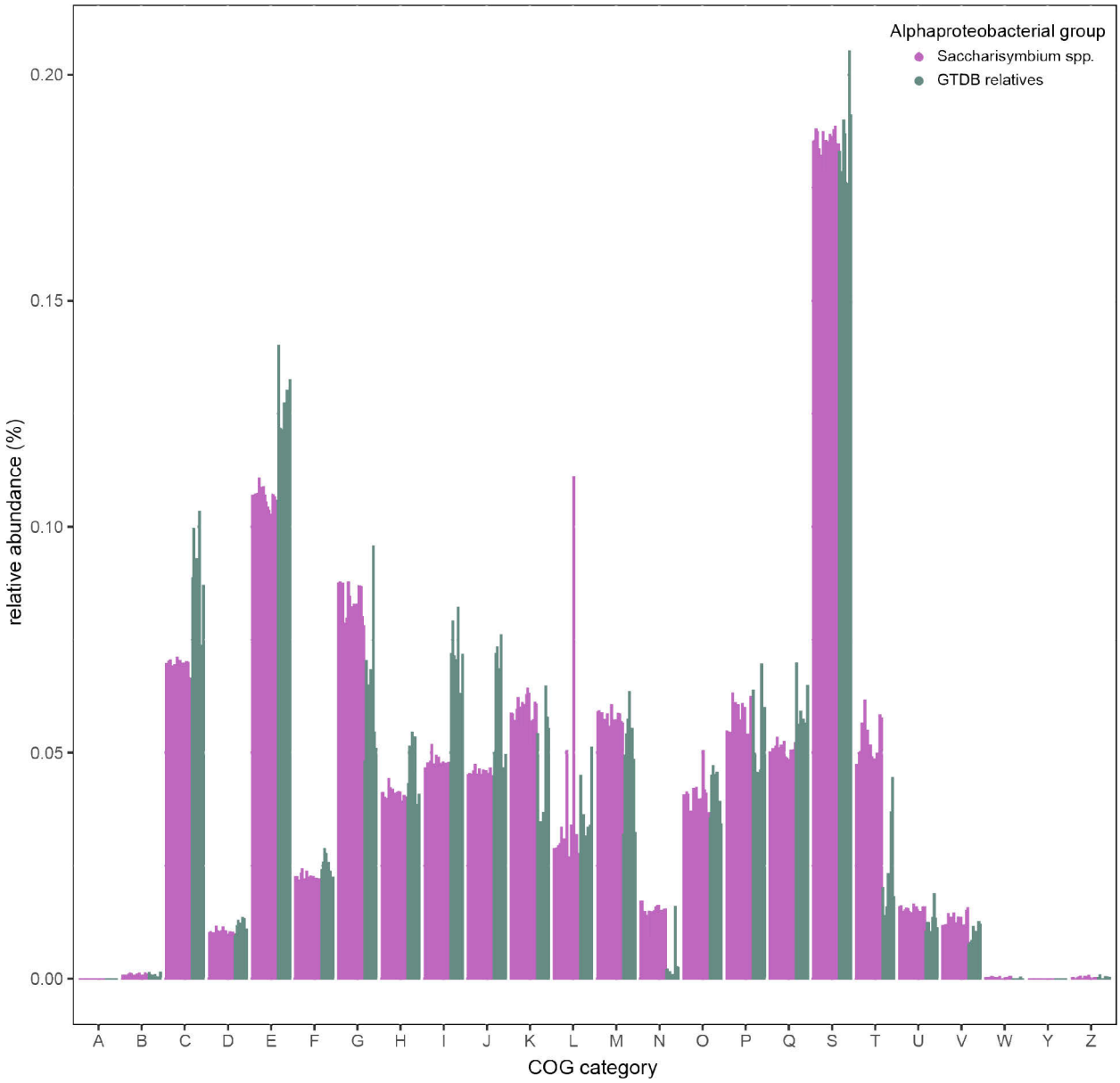
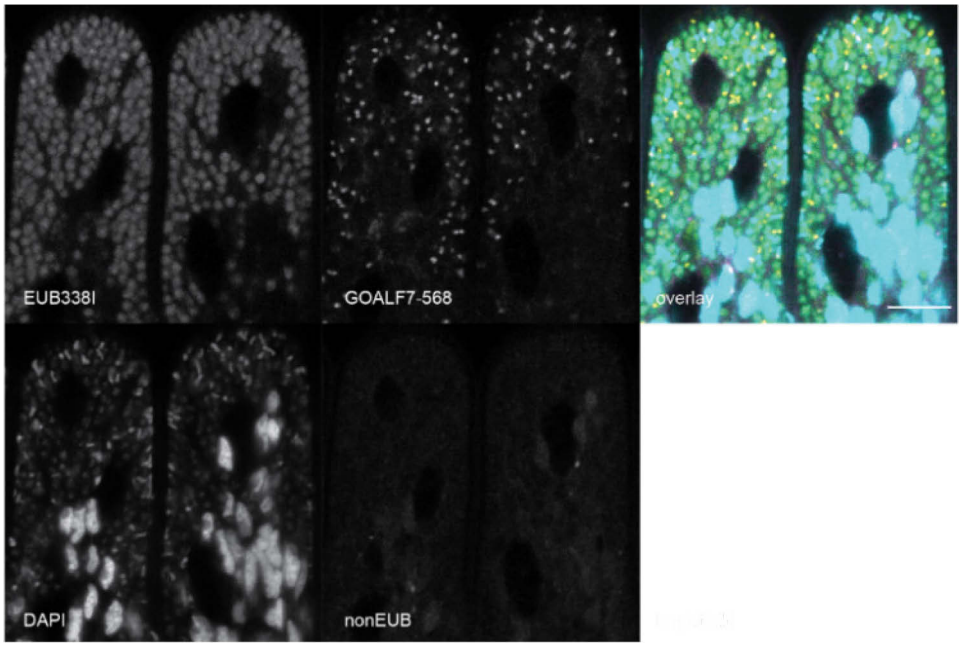


Figure S3: Saccharisymbium spp. have high relative gene abundances in COG categories for carbohydrate (COG G) and amino acid (COG E) metabolism and transport, and energy production and conservation (COG C).
Relative gene abundances in % per COG category were assigned with eggNOG and are depicted for MAGs of Saccharisymbium spp. (purple) and relative MAGs from GTDB. Saccharisymbium spp. has highest gene abundances in carbohydrate (COG G) and amino acid (COG E) metabolism and transport, and energy production and conservation (COG C). Compared to GTDB relatives Saccharisymbium spp. have more genomic space allocated to Signal Transduction (COG T), cell motility (COG N), and carbohydrate metabolism and transport (COG G).

5 Fluorescence *in situ* hybridization

Additional FISH images



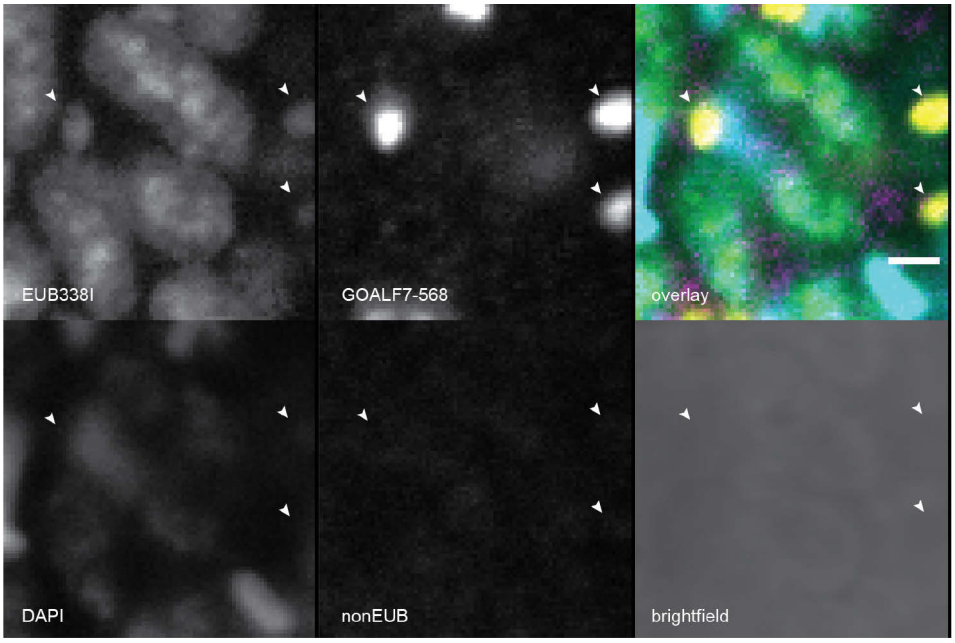


Figure S4 (top) and S5 (bottom): Saccharisymbium mediterraneum resides in the subcuticular symbiont layer.
Fluorescence *in situ* hybridization of Saccharisymbium mediterraneum in *O. ilvae* from Elba, Italy. Whole worms were hybridized with a Saccharisymbium spp. specific probe (GOALF7-568, yellow), a bacterial general probe (EUB338I, green) and the complementary non-binding probe for unspecific binding (nonEUB, magenta). DNA was counterstained with DAPI (cyan). The colored image shows the overlay with probes colored as stated in brackets. The lower right image shows a schematic of the annuli imaged. Scale bar: 10 µm (top), 1 µm (bottom).

Protocol – Whole-mount MiL-FISH on *Olavius algarvensis* worms[25]

A. Gruhl, MPI for Marine Microbiology, Bremen – Last update 05/04/22 by Tina Enders

Fixation (directly after sampling)

- Drop worms into freshly made (max. 1 day old) 4% PFA 0.05M PBS. Make sure to carry over as little seawater as possible (or change fixative once). Incubate 4h at 4°C.
- Wash 3x in 0.05M PBS
- Wash in 100% methanol (make sure to replace almost all water, repeat if necessary)
- Store @ -20°
- *Worms fixed and stored in Methanol could also be used*

Rehydration

- Wash 5 min @ RT with 60/40 methanol/PTw
- Wash 5 min @ RT with 30/70 methanol/PTw
- Wash 4x 5 min @ RT with PTw

Protein digestion, background removal, blocking of endogenous enzyme activity

- Cut worms in 2-3 pieces/worm
- Incubate samples with 0.05 mg/ml proteinase K (1:500 Prot K in PTw) in PTw for 5 min @ 37° **without shaking**
- Wash 2x 5 min with 2 mg/ml glycin in PTw
- Wash 2x 5 min PTw
- Re-fix (4% PFA in PTw) for 1h @ RT
- Wash 5x 5 min with PTw @ RT on shaker (700-800 rpm)

Mono/MiL FISH [or replace with CARD FISH etc. protocol]

- Prepare hybridization buffer (HyB)
- Incubate with HyB for 10 min at RT
- Incubate with pre-warmed (46 °C) HyB 1-24 h **(3 h)** in (shaker or hybridization oven)
- **Keep samples at hybridization temperature from now. All subsequent washes with pre-warmed buffers**
- Prepare hybridization mixture:
    - If using multiple probes simultaneously: make a combined working solution by mixing 4.2 µl of 100 pM stock from each probe and fill up with water to 50 µl. Dilute this 1:10 in hybridization buffer
    - Keep probe solutions dark and on ice
- warm up hybridization mixture to 46 °C
- Remove HyB from sample, add of probe hybridization mixture
- Incubate 2-24 h **[overnight]**
- prepare washing buffer (WB) and mixtures, pre-warm
- Wash in HyB buffer at 46 °C for 10 min
- Wash in HyB buffer 46 °C for 30 min
- Wash in 50% Hyb buffer / 50 % WB @ 46 °C for 30 min

- Prewarm WB to 48 °C
- Wash in WB at **48 °C** for 20 min
- Wash in WB at RT for 20 min
- Wash in 50% WB / 50% PTw @ RT for 10 min
- Wash in PTw @ RT for 10 min
- Wash in PTw for 30 min @RT @ slow shake

Counterstaining with DAPI

- Incubate with 1:500 DAPI in PTw for 30 min
- Wash 1x with PTw for 5 min shaking

Mounting

- Wash in 1:3 Glycerol/1x PBS for 30 min (from now on without shaking)
- Wash in 1:1 Glycerol/1x PBS for 30 min
- Wash in 3:1 Glycerol/1x PBS for 1h
- Mount worms in Vectashield on slide: use slides with 3 layers of tape (tesafilm) on each side to carry the coverslip.
- Seal coverslips with nail polish

Solutions

NOTE: Use of DEPC treated water/chemicals is not strictly necessary when using oligoDNA probes for 16S FISH!

4 x 0.05 M PB (0.2 M PB)

- Na_2HPO_4 x 2 H_2O – 26,12g
- NaH_2PO_4 x H_2O – 7,52g
- Dilute in 900 ml Milli-Q water
- Adjust ph to 7.4 (using >1M NaOH / HCl)
- Fill up to 1L

1.2 M NaCl

- Dissolve 70.13 g NaCl in 1 L Milli-Q

4% PFA in 0.05 M PBS

- 10 ml of 16 % PFA
- 10 ml of 4 x 0.05 M PB
- 10 ml of 1.2 M NaCl
- 10 ml Milli-Q

or:

- 10 ml of 20 % PFA
- 12.5 ml of 4 x 0.05 M PB
- 12.5 ml of 1.2 M NaCl
- 15 ml Milli –Q

H_2O [DEPC-treated]

- Under fume hood combine:
- 1 mL DEPC
- 1 L Milli-Q H2O
- in glass bottle, close lid, shake
- incubate at 37° for at least 2h
- autoclave

Proteinase K stock (25 mg/ml) in DEPC treated H_2O

- 0.125 g of Proteinase K in 5ml of DEPC H_2O
- Aliquot and freeze

10 x PBS [DEPC-treated]

- 70g NaCl
- 62,4g Na_2HPO_4 x 2 H_2O
- 3,4g KH_2PO_4
- Dilute in 900 ml Milli-Q
- Adjust to pH 7,4
- Fill to 1000 ml
- Add 1 mL DEPC
- Close bottle, shake
- incubate at 37° for at least 2h
- autoclave Add 1 mL DEPC
- Close bottle, shake
- incubate at 37° for at least 2h
- autoclave

PTw

- 100 ml 10 x PBS [DEPC treated]
- 896 ml DEPC-H_2O
- 4 ml 20% Tween-20
- store at 4°C

1% Triethanolamine in PTw :

- 1 mL of Triethanolamine
- 99 mL of PTw

Methanol/ PTw dilution series (50 mL)

- 60/40: 30 mL of MetOH + 20 mL of PTw
- 30/70: 15 mL of MetOH + 35 mL of PTw

20 mg/mL Glycine in PTw:

- 100 mg of Glycine

- 50 mL of PTw

Hybridisation buffer

Stock reagent	Volume	final concentration in hybridization buffer
5 M NaCl	360 µl	900 mM
1 M Tris / HCl	40 µl	20 mM
Formamide	% depending on probe	
distilled H_2O	add to 2 ml	
10% SDS	2 µl	0.01%
(add SDS last to avoid precipitation)		

Washing buffer

Table A: NaCl concentration in the washing buffer according to % formamide hybridization buffer (depending on probe).

	Washing at 48°C	
% formamide in hybridisation buffer	[NaCl] in M endconcentration	µl 5 M NaCl in 50 ml
0	0.900	9000
5	0.636	6300
10	0.450	4500
15	0.318	3180
20	0.225	2150
25	0.159	1490
30	0.112	1020
35	0.080	700
40	0.056	460
45	0.040	300
50	0.028	180
55	0.020	100
60	0.014	40
65	-	-
70	-	-

Stock reagent	Volume	final concentration in hybridization buffer
5 M NaCl	concentration depending on % formamide in hybridization buffer (see table A in appendix)	
1 M Tris / HCl	1 ml	20 mM
0.5 M EDTA (only if ≥ 20% formamide in hybridization!)	(500 µl)	5 mM
distilled H_2O	add to 50 ml	
10% SDS	50 µl	0.01%
(add SDS last to avoid precipitation)		

Table S3: Oligonucleotides used for fluorescence *in situ* hybridization.

Oligo name	Sequence	5'-modification	3'-modification	Specificity	Reference	probeBase
GOALF7-586	5'- GCC TGA CTG ACC GTT CCG -3'	Atto 550	Atto 550	Candidatus A7 mediterraneum	this study	
GOALF7-634	5'- TAG ATC GGC AGT ATC AAG CG -3'	Atto 550	Atto 550	Candidatus A7 mediterraneum	this study	
EUB338I	5'- GCT GCC TCC CGT AGG AGT -3'	Atto 488	Atto 488	most Bacteria	Aman et al., 1990	https://probebase.csb.univie.ac.at/pb_report/probe/159
nonEUB	5'- ACT CCT ACG GGA GGC AGC -3'	Atto 647	Atto 647	negative control for nonspecific binding	Wallner et al., 1993	https://probebase.csb.univie.ac.at/pb_report/probe/243
nonEUB	5'- ACT CCT ACG GGA GGC AGC -3'	Atto 488	Atto 488	negative control for nonspecific binding	Wallner et al., 1993	https://probebase.csb.univie.ac.at/pb_report/probe/243
ALF968	5'- GGT AAG GTT CTG CGC GTT -3'	Cy5	Cy5	Alphaproteobacteria, except of Rickettsiales	Alm et al, 1996	https://probebase.csb.univie.ac.at/pb_report/probe/21

Table S4a: CheckM statistics for Saccharisymbium spp. metagenome-assembled genomes.

Bin	Id	Marker lineage	genomes	markers	marker sets	Completeness	Contamination	Strain heterogeneity	Genome size (bp)	contigs	N50 (contigs)	Longest condit (bp)	GC	GC std	Coding density	predicted genes
3557_AD.autometa.bin1_1	c__Alphaproteobacteria	(UID3305)	564	349	230	96,96	7,26	9,52	4501395	596	12533	62916	69,84	2,30	91,40	4784
3557_BK.metabat.bin3	k__Bacteria	(UID203)	5449	104	58	60,27	8,78	10,00	2801683	757	4219	17047	69,51	2,09	91,97	3328
3557_BL.dastool.maxbin.bin6_sub	c__Alphaproteobacteria	(UID3305)	564	349	230	89,17	4,71	9,09	4279362	830	6947	26427	68,25	10,06	88,87	4587
3557_CP.autometa.bin1_0	o__Rhodospirillales	(UID3754)	63	336	201	99,00	0,00	0,00	4353733	212	31958	125234	69,81	2,06	90,25	4341
3630_F.metabat.bin4	o__Rhodospirillales	(UID3754)	63	336	201	85,57	0,76	33,33	3675859	748	6374	35450	69,55	2,10	90,11	4458
4148_4289_B.metabat.bin1	o__Rhodospirillales	(UID3754)	63	336	201	98,71	1,12	0,00	4198633	67	94297	303523	69,00	1,65	88,84	4276
4148_4289_P.metabat.bin5	c__Alphaproteobacteria	(UID3305)	564	349	230	99,57	0,43	0,00	4440562	86	115151	425876	70,18	1,99	92,13	4311
4148_4289_Q.dastool.metabat.bin5_sub	c__Alphaproteobacteria	(UID3305)	564	349	230	98,70	1,20	25,00	4405500	123	60054	164851	70,08	1,47	92,09	4291
4148_4289_T.maxbin.bin3	c__Alphaproteobacteria	(UID3305)	564	349	230	99,57	0,65	0,00	5295260	426	101429	283083	65,98	21,33	81,45	4463
4148_4289_X.autometa.bin1_1	c__Alphaproteobacteria	(UID3305)	564	349	230	99,57	0,91	25,00	4419300	64	127354	313524	70,03	1,44	92,02	4239
4410_Alpha7_Oilv_flye-2.7	o__Rhodospirillales	(UID3754)	63	336	201	98,51	0,50	0,00	4623813	9	4505067	4505067	69,66	22,63	89,12	4728
4515_G.autometa.bin1_1	o__Rhodospirillales	(UID3754)	63	336	201	98,51	0,00	0,00	4318694	209	35978	120919	69,83	1,80	90,17	4361
4515_I.maxbin.bin6	c__Alphaproteobacteria	(UID3305)	564	349	230	98,26	4,49	11,76	5183231	536	26450	125222	65,01	17,97	81,61	4697
4515_K.autometa.bin1_1	c__Alphaproteobacteria	(UID3305)	564	349	230	98,70	0,45	50,00	4431076	35	195834	390878	69,75	0,85	91,24	4327
4515_L.metabat.bin5	c__Alphaproteobacteria	(UID3305)	564	349	230	99,57	0,43	0,00	4421650	18	420376	818218	69,75	0,80	91,51	4279
OilvPIA1.autometa.bin1_1	c__Alphaproteobacteria	(UID3305)	564	349	230	98,70	0,43	0,00	4353633	194	31953	126594	69,82	1,76	90,61	4269
OilvPIA2.autometa.bin1_1	c__Alphaproteobacteria	(UID3305)	564	349	230	99,13	0,43	0,00	4352860	204	31718	148592	69,82	1,73	90,61	4283
OilvSAN1.autometa.bin1_1	o__Rhodospirillales	(UID3754)	63	336	201	98,51	0,00	0,00	4316411	201	35116	125260	69,83	1,73	90,18	4341
OilvSAN2.autometa.bin1_1	o__Rhodospirillales	(UID3754)	63	336	201	98,51	0,05	0,00	4309673	243	27063	115936	69,84	1,93	90,31	4367
average						95,55			4351701				69,24		89,71	
median						98,70			4353733				69,81		90,31	

Table S4b: Pathway Tools statistics for Saccharisymbium spp. Metagenome-assembled genomes.

Bin	Size (bp)	Total genes	Protein genes	RNA genes	Pathways	tRNAs	Transporters
4410_Alpha7_Oilv_flye-2.7	4623813	4885	4813	72	326	65	24
4515_G.autometa.bin1_1	4318694	4225	4172	53	362	49	68
4515_I.maxbin.bin6	5183231	4507	4447	60	361	56	69
4515_K.autometa.bin1_1	4431076	4365	4309	56	361	51	64
4515_L.metabat.bin5	4421650	4317	4261	56	357	51	62

Table S5: Related sequences to Saccharisymbium spp. from the Genome Taxonomy Database (GTDB).

GTDB-tk acession	GTDB-tk taxonomy	16S	Source
GCA_002713365.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__UBA6615; f__UBA6615; g__NAT5454; s__NAT5454 sp002713365	yes	marine pelagic biome, South Atlantic Ocean
GCA_002722515.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__UBA6615; f__UBA6615; g__UBA8079; s__UBA8079 sp002722515	no	marine pelagic biome, South Pacific Ocean
GCA_002724635.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__UBA6615; f__UBA6615; g__UBA8079; s__UBA8079 sp002724635	no	marine pelagic biome, Saudi Arabia: Red Sea
GCA_002727315.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__UBA6615; f__UBA6615; g__GCA-2727315; s__GCA-2727315 sp002727315	no	marine pelagic biome, Saudi Arabia: Red Sea
GCA_002937505.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__UBA6615; f__UBA6615; g__UBA6615; s__UBA6615 sp002937505	yes	saline water including plankton, Pacific Ocean: South Pacific Ocean
GCA_003230015.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__UBA6615; f__UBA6615; g__SZUA-147; s__SZUA-147 sp003230015	yes	Atlantic Ocean: Mid-Atlantic Ridge, Guaymas Basin, Black smoker hydrothermal vent wall
GCA_002238725.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Bin65; f__Bin65; g__Bin65; s__Bin65 sp002238725	yes	Aplysina aerophoba, Slovenia: Marine Biology Station Piran
GCA_005240015.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Bin65; f__Bin65; g__SBBV01; s__SBBV01 thessalonicensis_A	no	alpine spring ground water, Italy
GCF_900197615.1	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Bin65; f__Bin65; g__Bin65; s__Bin65 sp900197615	no	Bins of sponge-associated and environment from coastal water body of Portugal

References

1. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: A software environment for sequence data. Nucleic acids research. 2004;32(4):1363-71.
2. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of *de novo* genome assemblies. Bioinformatics. 2015;31(20):3350-2.
3. Nikolenko SI, Korobeynikov AI, Alekseyev MA, editors. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC genomics; 2013: Springer.
4. Bushnell B. BBmap. v38.34 ed2010.
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10.
6. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research. 2015;25(7):1043-55.
7. Eichinger V, Nussbaumer T, Platzer A, Jehl M-A, Arnold R, Rattei T. EffectiveDB — Updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. Nucleic acids research. 2016;44(D1):D669-D74.
8. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. 2019;47(D1):D309-D14.
9. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular biology and evolution. 2017;34(8):2115-22.
10. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nature biotechnology. 2019;37(5):540-6.
11. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press; 2020.
12. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic acids research. 2019;47(W1):W256-W9.
13. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015;32(1):268-74.
14. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.
15. Gruber-Vodicka HR, Seah BK, Pruesse E. phyloFlash: Rapid small-subunit rRNA profiling and targeted assembly from metagenomes. Msystems. 2020;5(5).
16. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
17. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.
18. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid annotations using subsystems technology. BMC genomics. 2008;9(1):1-15.
19. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic acids research. 2014;42(D1):D206-D14.
20. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific reports. 2015;5(1):1-6.
21. R Core Team. R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
22. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012;19(5):455-77.
23. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: A versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
24. Mankowski A, Kleiner M, Erséus C, Leisch N, Sato Y, Volland J-M, et al. Highly variable fidelity drives symbiont community composition in an obligate symbiosis. bioRxiv. 2021.
25. Schimak MP, Kleiner M, Wetzel S, Liebeke M, Dubilier N, Fuchs BM. MiL-FISH: Multilabeled oligonucleotides for fluorescence *in situ* hybridization improve visualization of bacterial cells. Applied and environmental microbiology. 2016;82(1):62-70.

:aberrant

The proposed *Candidatus* Pumilisymbium is different from the subcuticular and mutualistic gutless oligochaete symbionts. It has a strongly reduced genome of only 0.64 Mbp and is a symbiont of the host species *Olavius algarvensis* from seagrass meadow associated sediments around the island of Elba, Italy.

Chapter 3 | *Candidatus* Pumilisymbium – a strongly reduced symbiont

Candidatus Pumilisymbium abstrusum, a highly reduced and deeply branching Alphaproteobacterium in symbiosis with marine invertebrate gutless oligochaetes (Oligochaeta, Annelida)

Tina Enders[1], Alexander Gruhl[1], Manuel Kleiner[2], Nikolaus Leisch[1], Anna Mankowski[1,3], Yui Sato[1], Nicole Dubilier[1], Harald R. Gruber-Vodicka[1]

[1] Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

[2] Department of Plant & Microbial Biology, North Carolina State University, Raleigh 27695, North Carolina, USA

[3] Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Corresponding authors:
Nicole Dubilier, ndubilie@mpi-bremen.de,
Harald R. Gruber-Vodicka, hgruber@mpi-bremen.de

Competing interest

The authors declare no competing financial interests.

Abstract

Bacterial symbionts with highly reduced genomes are well known from terrestrial animals but few representatives in the marine realm are described. Here, we present *Candidatus* Pumilisymbium abstrusum, a deeply branching marine Alphaproteobacterium that is an endosymbiont of the gutless oligochaete *Olavius algarvensis* (Clitellata, Annelida). Phylogenetic analyses place *Ca.* P. abstrusum and its closest relative, a symbiont of the arctic ice worm, as a novel order-level clade distantly related to Rickettsiales *sensu stricto*. *Ca.* P. abstrusum has the smallest genome reported for a symbiont of a marine animal. The high quality metagenome-assembled genome has a size of only 641 817 bp, a GC content of 34.74% and comprises 3 rRNAs, 36 tRNAs and 658 protein coding genes. Based on metabolic reconstruction, the symbiont has a streamlined metabolism that potentially integrates with the host. *Ca.* P. abstrusum uses intermediates and waste products of the host's metabolism such as fructose and malate and can in return supply B-vitamins, co-factors and pyruvate. Core cell functions such as DNA and RNA metabolism, ribosomal protein biosynthesis and ATP synthesis via glycolysis and an ATP synthase are retained, whereas amino acid biosynthesis is almost absent. The reduced metabolic capabilities and a type III secretion system that can modulate host biology indicate an intracellular life-style. This combination of features, i.e. ATP synthesis and host modulation via a type III secretion system in a bacterium with such a strongly reduced genome, is unprecedented in the bacterial world and suggests a novel type of integration into the host biology.

Keywords

Type III secretion system (T3SS), genome reduction, marine animal-microbe symbiosis,

Introduction

Strongly reduced bacteria from marine animal-microbe symbioses are understudied

Genome size reduction is a common evolutionary trajectory for free-living and especially host associated bacteria[1, 2]. Free-living bacteria that can independently replicate need a minimum set of roughly 1000 genes, which translates to a genome size of ~1 Mbp[3, 4]. Bacteria in highly specialized and beneficial nutritional symbioses that have reduced and streamlined genomes often smaller than 0.75 Mbp (Figure 1) commonly undercut this mark. Nearly all beneficial symbionts with such highly reduced genomes have been characterized in terrestrial and especially insect hosts (Figure 1)[5]. In addition to beneficial bacteria, some pathogens also have

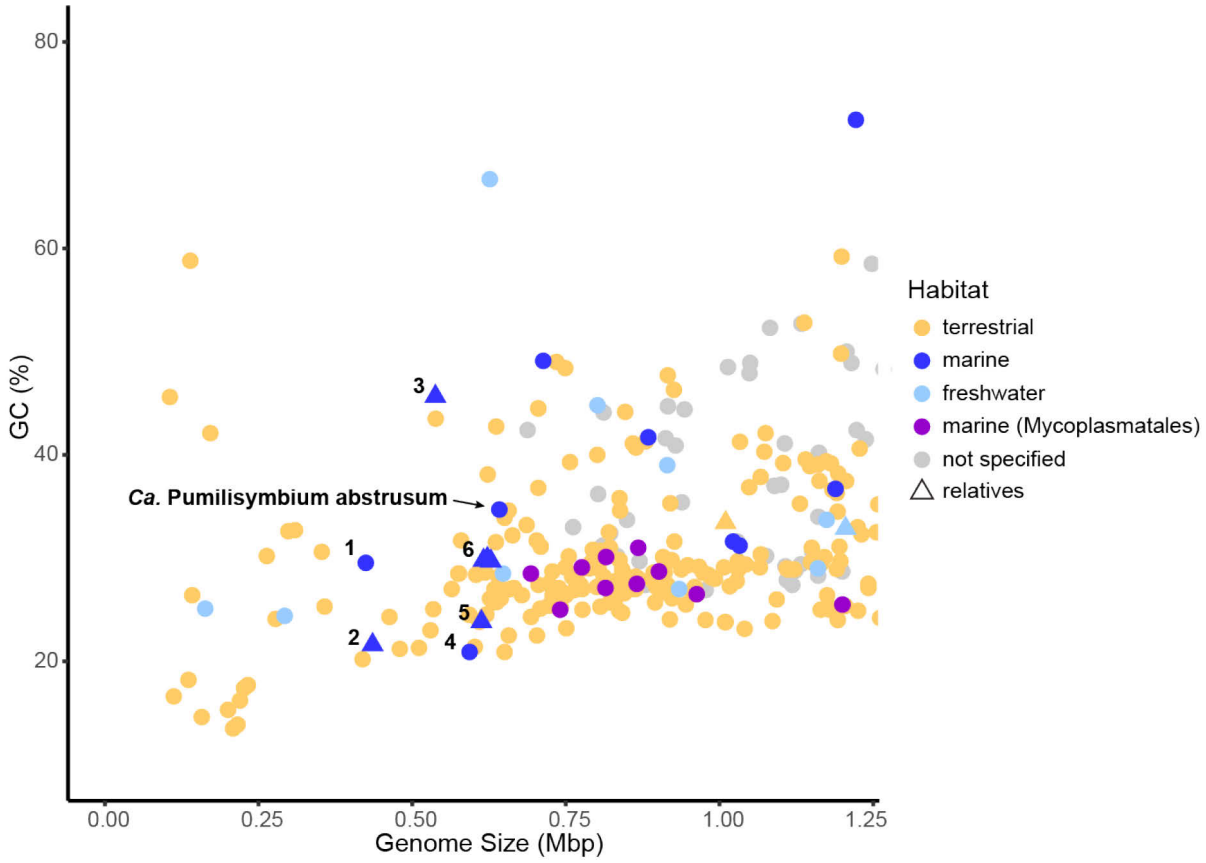


Figure 1: Complete bacterial genomes with a size below 1 Mbp were mainly collected from terrestrial symbioses.
Representatives of bacterial species deposited at NCBI (Size below 1.25 Mbp, accessed March 2021) and relatives from sister clades based on GTDB (see Figure 4) are plotted by genome size and GC content. Colors represent habitats of the bacteria: yellow: terrestrial, light blue: fresh-water, dark blue: marine, purple: Mycoplasma from large marine animals, grey: habitat not specified for this study. Relatives from the GTDB are shown as triangles. 1: *Ca.* Spiroplasma holothuricola; 2: AB1-6 bacterium associated with *Bugula neritina*; 3: bacterium from hydrothermal vent metagenome; 4: AB-1 bacterium associated with *Bugula neritina*; 5: bacterium from microbial mat; 6: *Ca.* Cytomitobacter primus, *Ca.* Cytomitobacter indipagum, *Ca.* Nesciobacter abundans.

limited gene sets such as Mycoplasmatales or Rickettsiales but have hardly been characterized in aquatic animal hosts (Figure 1)[6-8]. Given that the large part of animal diversity is aquatic and almost half of the animal phyla only occur in the ocean, this points towards a bias in symbiosis research.

Here, we address this gap and describe a highly reduced symbiont that associates with a species of gutless marine annelids. Gutless oligochaetes are a morphologically conspicuous group of millimeter-sized tubificid Annelida that do not possess a mouth, gut and nephridia. They occur in most tropical and temperate coastal sediments and rely on species-specific communities of chemosynthetic bacterial symbionts for nutrition and waste product removal[9-11]. In a

population genetic study conducted on *Olavius algarvensis* around the island of Elba, Italy, we detected a novel symbiont clade that did not match any of the typical community members for this host[12]. The recovered complete genome of this symbiont is one of the smallest genomes reported for marine bacteria. Based on our phylogenetic and comparative metabolic analysis, this bacterium belongs to an uncharacterized order-level clade of Alphaproteobacteria with a unique type of low-impact and intracellular life-style.

Methods and materials

Sample collection

Gutless oligochaete specimens of the species *Olavius algarvensis* were collected in the years 2013 to 2019 from shallow water sediments from different bays around Elba, Italy (Supplement section 3, metadata at zenodo.org/record/6515035). Samples were fixed in RNAlater and stored at 4 °C or -20 °C until further processing.

DNA extraction and sequencing

For DNA extractions, individual specimens were homogenized and extracted using either the DNeasy Blood&Tissue or the AllPrep DNA/RNA/Protein kit (Qiagen, Hilden, Germany). Metagenome library preparation, quality control and sequencing for all specimens was done at the Max Planck Genome Centre (Cologne, Germany). Paired-end reads of 2x100 or 2x150 bp were produced with an average of 14 million reads per sample. Reads were quality trimmed and filtered with bbduk.sh (BBtools, jgi.doe.gov/data-and-tools/bbtools/) and error corrected with bayeshammer as implemented in SPAdes[13, 14]. Quality before and after trimming was assessed with FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/). A detailed list of samples and treatments is available at zenodo.org/record/6515035.

Genome assembly, binning and annotation

Read based taxonomic profiling for community composition and contamination of each sample was done with PhyloFlash[15]. Reads were assembled with megahit (meta-sensitive, min_contig_len 500)[16, 17]. Reads were mapped to the assemblies with bbmap.sh (BBtools) and sorted with Samtools before metagenome-assembled genomes (MAGs) were created with metabat[18-21]. The target MAG was improved by iterative read mapping with bbmap.sh (minid=0.95) and reassembly with SPAdes[13]. The assembly was visually inspected in Bandage (Figure 2)[22]. The MAG and its genomic features were visualized with DNAplotter from the artemis software[23].

Phylogenetic placement

Phylogenetic trees were calculated with full-length 16S rRNA gene sequences using a maximum likelihood approach and approximate maximum likelihood ratio tests (mafft-xinsi, IQ-TREE -m MFP –alrt 1000)[24, 25]. Reference sequences were a selection of Rickettsiales bacteria, relevant GTDB-tk relatives and related database hits to *Ca.* P. abstrusum, from here on referred to as P. abstrusum, from NCBI's environmental and type strain nucleotide collection (nr/nt, env_nt)[26, 27]. Deltaproteobacterial 16S rRNA gene sequences from *Bdellovibrio bacteriovorus*, *Geobacter sulfurreducens* and *Desulfovibrio vulgaris* were used as outgroup. Phylogenomic analyses were calculated with P. abstrusum's MAG, which was placed in the GTDB tree at version 1.5.0 with the R202 reference data using the GTDB-Tk software[28]. Trees were visualized and annotated using iTOL and edited with Adobe illustrator v25.3 (adobe.com/products/illustrator)[29].

Metabolic reconstruction and COG profiling

Automated functional annotation of the MAG was performed with RAST and Prokka[30-33]. Annotations were combined using Genious v11.1.5 (www.geneious.com). Remaining hypothetical proteins were manually annotated with psi-blast, hmmscan against the Pfam database (www.ebi.ac.uk/Tools/hmmer/search/hmmscan) and SUPERFAMILY[26, 34-36]. Effector proteins were predicted with effectiveDB[37]. Proteins with transmembrane helices were predicted with TMHMM v2.0 (services.healthtech.dtu.dk/service.php?TMHMM-2.0). Reconstruction of the final metabolic cell overview was partially guided by the metabolic model from Pathway Tools, MetaCyc and KEGG maps[38-42]. Annotation information is provided at zenodo.org/record/6515015. The metabolic cell overview was designed in Adobe Illustrator v25.3 (Figure 5).

Additionally, protein sequences predicted with Prodigal implemented in Prokka were assigned to clusters of orthologous genes (COGs) with eggNOG-mapper v2.1.6 using diamond[43-46]. The presence of COG categories that represent larger functional classifications was quantified and compared to representatives of most alphaproteobacterial families and a selection of related bacteria (Supplement Table S4). To obtain alphaproteobacterial family representatives, the bacterial tree bac120_r86.2 was downloaded from the GTDB database (gtdb.ecogenomic.org/downloads) and trimmed to the Alphaproteobacteria and Magnetococcia outgroup in iTOL[29]. Branch lengths for each leaf were extracted with Newick utils (nw_distance), the average branch length was calculated for each family, and for each family

the genome that was closest to the average branch length was chosen for further phylogenomic analyses (genomes listed in zenodo.org/record/6514058)[47]. Relations and driving COG categories were represented with non-metric multidimensional scaling (NMDS) and principal component analysis (PCA) plots calculated in RStudio (RStudio v1.4.1106, R v4.0.4) employing a variety of packages (Supplement section 10).

Genome based abundance estimation

We estimated the abundance of P. abstrusum in the gutless oligochaete populations in several bays around Elba, Italy, by screening gutless oligochaete metagenomes versus the P. abstrusum coding sequences from its four main contigs. We used a conservative mapping approach to prevent false positives (bbmap.sh, minid=0.99, trimq=20, qtrim=lr, pairedonly=t, pairlen=600, mintrimlength=100). Only genes with a length greater than 400 were considered as positive hits. Mapping results were plotted in RStudio (RStudio v1.4.1106, R v4.0.4) employing a variety of packages (Supplement section 4).

We used the integrated microbial next generation sequencing (IMNGS) platform to assess the global distribution of P. abstrusum (accessed 03.03.2022)[48]. We used the full-length 16S rRNA gene of P. abstrusum as query to search against sequence read archives with a minimum hit similarity of 90% and a minimum hit lengths of 200 bp.

PCR based abundance estimation

PCR was performed using a primer pair specific to P. abstrusum, 379F (5'-GGAAACCTTGATCCGGTTATG-3') and 1026R (5'-AACACCTGTGATATGTATAGTG-3') that amplifies a fragment of 641 bp[49]. An amount of 1 µl template DNA extracted from individual specimens was used with Phusion DNA polymerase and the reaction mix according to manufacturer's recommendation. The amplification program was: 30 sec initial denaturation at 98 °C, 36 cycles of denaturation for 15 sec at 98 °C, 30 sec of annealing at 61 °C, extension for 30 sec at 72 °C and a final extension at 72 °C for 10 min. PCR amplifications were checked on agarose gel. Positive samples with the expected product length were purified and Sanger sequenced. The PCR product sequences were trimmed with FinchTV (digitalworldbiology.com/FinchTV) and aligned with the P. abstrusum 16S rRNA gene sequence using mafft[25].

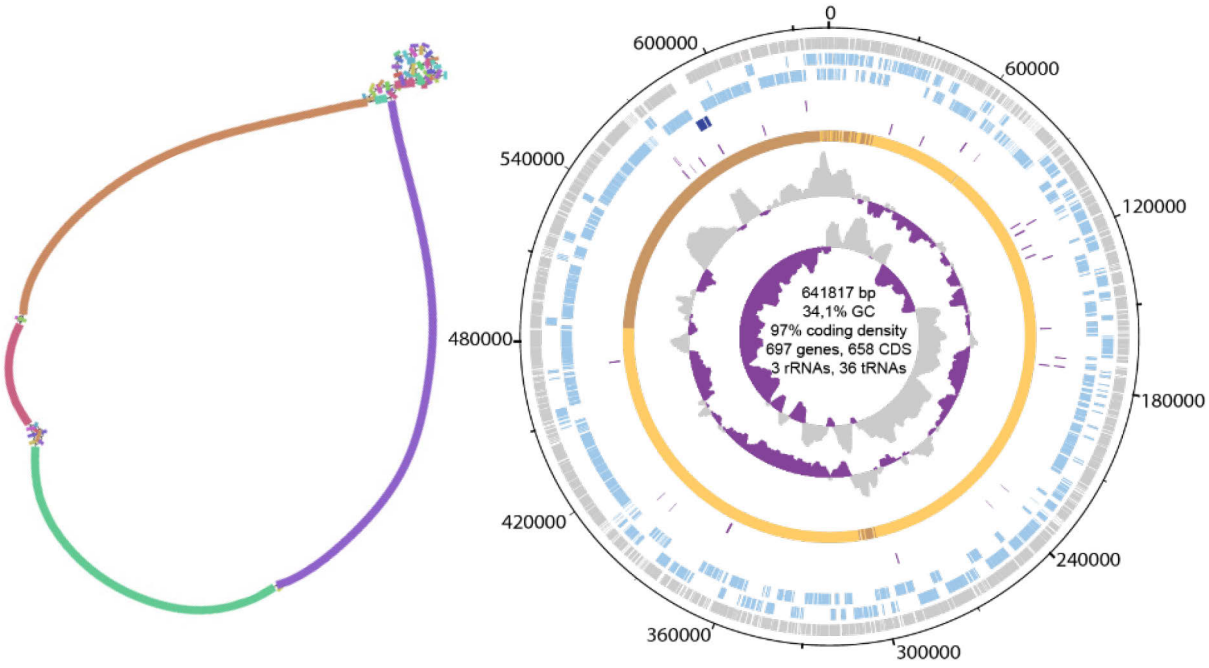


Figure 2: Genome representation of the P. abstrusum metagenome-assembled genome.
Left: Assembly graph in Bandage. Right: Circular genome representation. Circles from outside to inside represent base pairs, all coding sequences (CDS) in grey, CDS on the forward and on the reverse strand in light blue, ribosomal 5S, 16S and 23S rRNA gene sequences operon in dark blue, tRNAs on the forward and reverse strand in purple, contigs in yellow, deviation from average GC content and GC skew in grey (positive) and purple (negative). Position one was arbitrarily set.

Results

A novel alphaproteobacterial phylotype with a highly reduced genome associated with gutless marine annelids

We detected a previously unknown bacterial phylotype during investigation of 328 single individual metagenomic libraries of the Mediterranean gutless oligochaete *Olavius algarvensis*. PhyloFlash reconstructed two identical full-length 16S rRNA gene sequences for this bacterium from two worm individuals sampled in Cavoli Bay, Elba, Italy[15]. The 16S rRNA gene sequence is distinct from the described symbiont community of *O. algarvensis* and only 78.6% identical to its closest hit in the SILVA database (KC961300.1.1457, Rickettsiales;AB1;uncultured bacterium)[9, 15]. To provide a basis for phylogenomic analysis and metabolic reconstruction we generated a metagenome-assembled genome (MAG) using statistics based binning, read mapping and re-assembly[13, 18, 20]. Based on the final assembly graph we obtained a set of circularly connected contigs with a total size of 641 817 bp and an average GC content of 34.7% (Figure 2, Supplement section 5). The presence of all three rRNA coding genes, 36 tRNAs for all 20 common proteinogenic amino acids and the circular

assembly graph structure suggest high completeness of the MAG, despite a low reported completeness based on the checkM marker set for Bacteria (68.86% completeness, 0% contamination)[50]. Further evidence for a high completeness is the presence of all 21 SSU ribosomal proteins and 30 of 33 of the LSU ribosomal proteins, with only non-essential sub-units missing. Such a set of genome characteristics are typical for an endosymbiotic life-style.

Stable and specific association with *Olavius algarvensis* around Elba

We could already detect the novel phylotype in two specimens based on single marker gene analysis. To evaluate the nature of the association of the bacterium with the gutless oligochaete community, we investigated its distribution and a potential long-term association. Therefore, we used specific read mapping against the coding sequences of the MAG for the 328 *O. algarvensis* metagenomes from several sites around Elba, the neighboring island Pianosa and sites from Mallorca, Spain. We detected evidence for the bacterium in all sites around Elba but not in Pianosa or Mallorca. Around Elba, 5-28% of *O. algarvensis* specimens were infected, with highest infection rates in the northeastern bay Capo Vita (Figure 3, Supplement section 4, Figure S2 and S3). We corroborated these results with a PCR based approach at the two sites St'Andrea (n=15/29, 13%/24% infected) and Cavoli (n=18, 39% infected). This widespread occurrence around the island of Elba points to a long-term but facultative association. Hence, we propose the name *Candidatus* Pumilisymbium abstrusum (*pumili*-, lat.: dwarvish, *abstrusum*, lat.: hidden, Supplement section 7).

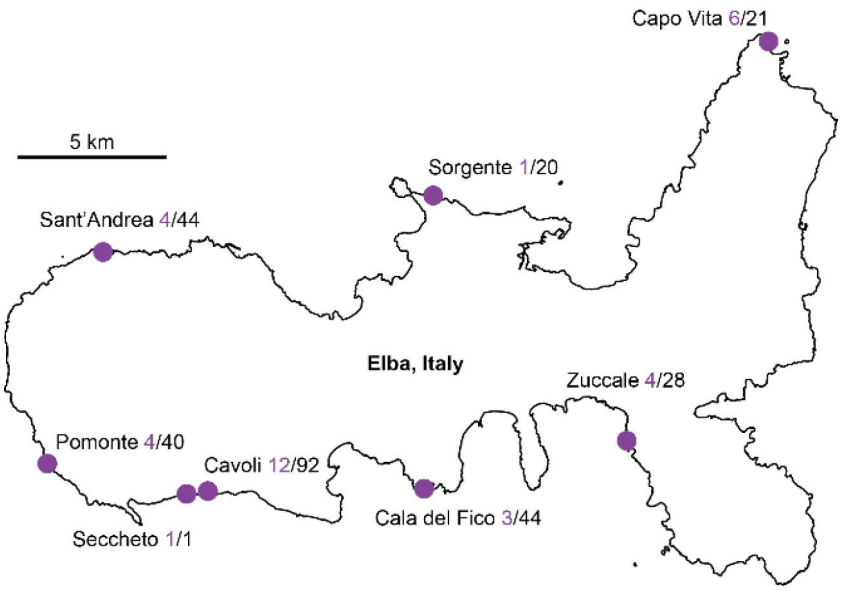


Figure 3: P. abstrusum is present in low number of individuals in all sampled bays around Elba. Specimens infected with P. abstrusum (purple) and total number of specimens investigated per sampled bays around the island of Elba, Italy.

Next, we asked whether P. abstrusum is only hosted by *O. algarvensis*. *Olavius ilvae* is a co-occurring gutless oligochaete at several sites that also shares several symbiont clades with *O. algarvensis*[11]. To probe if *O. ilvae* also hosts P. abstrusum, we screened 61 specimens from the same sites around Elba and Pianosa and in addition from a site at the French Riviera, Cap Fleuri. We could not detect P. abstrusum in any *O. ilvae* specimens indicating that P. abstrusum is not able to colonize *O. ilvae* (Supplement section 4, Figure S4). To assess if any other gutless oligochaetes can be infected, we used a global dataset of gutless oligochaete metagenomes but again found no hits[11]. Another possibility would be the association with other unrelated hosts. Using the IMNGS platform, we screened 500048 16S rRNA gene based amplicon data sets covering a wide range of environments. We could not identify a single sample with 16S rRNA gene amplicons more than 90% identical to P. abstrusum's 16S rRNA gene[48]. In summary, our results indicate that P. abstrusum is endemic to the Mediterranean Sea with *O. algarvensis* as its only host species.

Pumilisymbium abstrusum represents a deeply branching, novel order of Alphaproteobacteria

To place P. abstrusum in the tree of life, we generated phylogenetic reconstructions based on either the 16S rRNA gene or on the GTDB phylogenetic marker gene sets[28]. Based on the 16S rRNA gene analysis, P. abstrusum's closest relatives come from a microbiome clone library of the arctic ice worm *Mesenchytraeus solifugus* (Enchytraeidae, Annelida, sequence similarity of 84.06%)[26, 51]. All other database hits had a sequence similarity below 81.01%, suggesting that these two annelid associated phylotypes are the first representatives of a novel order-level clade[52]. Phylogenetic analyses using the 16S rRNA gene database hits place P. abstrusum and the ice worm associated sequences as a sister clade to the *incertae sedis* Holosporales within the order of Rickettsiales, albeit with no statistical support (Figure 4A, see Data availability for trees shared in iTOL). To probe for potential long-branch attraction effects, we recalculated the phylogenies without Holosporales sequences, but tree structure and position of P. abstrusum remained the same (Supplement Figure S5 and S6).

Phylogenomic reconstruction and taxonomic assignment based on GTDB reference taxonomy supported the rank of a novel order-level clade of Alphaproteobacteria[28, 53, 54]. P. abstrusum clustered with several taxa of *incertae sedis* status, i.e. Holosporales and paramecial and diplonemid symbionts[6, 55], that had previously been classified as Rickettsiales but that likely represent several separate order-level clades (Figure 4B, Supplement Section 8)[56].

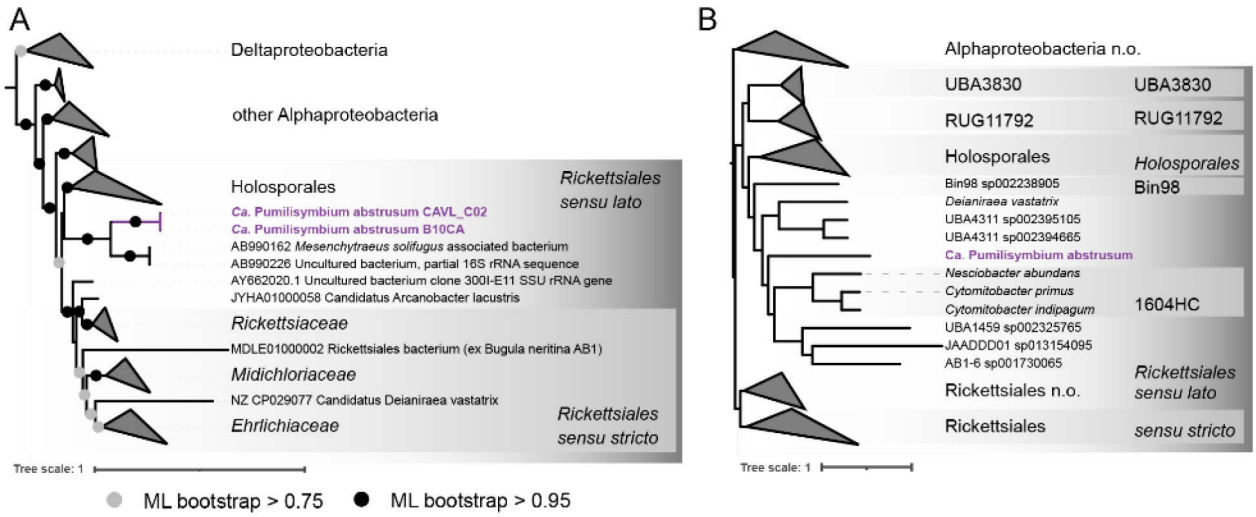


Figure 4: P. abstrusum and an ice worm symbiont form a novel bacterial order related to Rickettsiales (Alphaproteobacteria).
(A) Phylogenetic placement based on full-length 16S rRNA gene sequences from NCBI (nr/nt, nr/nt type strain material, env_nt) and relevant relatives from GTDB-tk phylogeny from (B). Taxonomic annotation based on Floriano et al. 2018[27]. Support values were calculated using -aLRT in IQ-TREE. Bootstrap values below 0.75 are not shown. (B) Phylogenomic placement with selected taxonomic marker genes (GTDB-tk). Order classification follows the Genome Taxonomy Database. The tree scale represents the average number of substitutions per site.

Pumilisymbium abstrusum's genome is streamlined for information processing

Using automated annotation tools and manual curation, we could functionally annotate 388 of the predicted 658 protein coding sequences. A high number of proteins of P. abstrusum are involved in basic cell functioning like DNA metabolism, transcription and translation (Figure 5, Supplement section 10 and at zenodo.org/record/6515015). This is similar to other reduced bacteria in host association that usually can perform essential information processing independent of their hosts[5]. In addition, P. abstrusum can regulate its transcription via regulators such as rpoH and helix-turn-helix family proteins. P. abstrusum might be a naturally competent bacterium since it encodes for competence factors like a ComF family protein, a ComEC/Rec2 family competence protein, Competence protein ComM and a Rossmann fold nucleotide-binding protein Smf possibly involved in DNA uptake. We observed an enrichment of short genes that are likely pseudogenes and 23 cases of potential overlapping open reading frames. These patterns suggests that the genome has undergone a process of pronounced simplification and streamlining, retaining only the necessary functions (Supplement section 9, Figure S7). We could not detect any insertion sequences or CRISPR elements in the MAG of P. abstrusum.

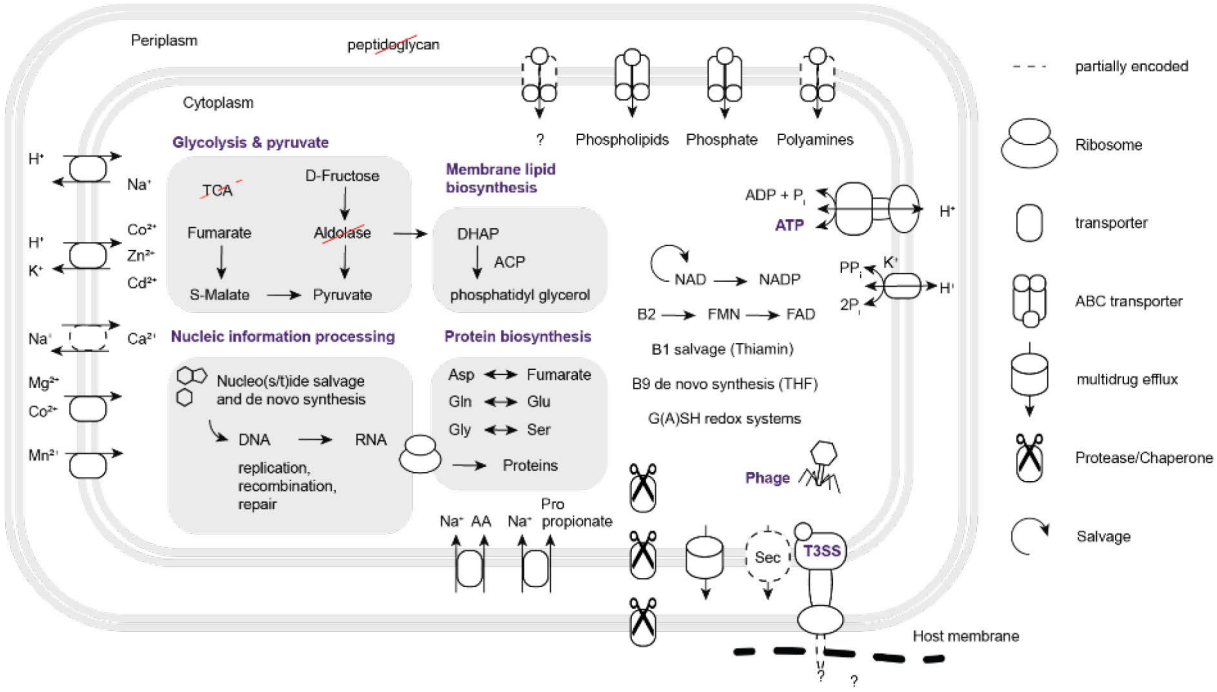


Figure 5: P. abstrusum retains a reduced basic cell metabolism with an ATP synthase and a type III secretion system as key features.
Proposed functions are based on metagenome assembled genome annotation with Prokka, RAST, psi-blast, hmm and superfamily. TCA: tricarboxylic acid cycle; DHAP: dihydroxyacetone phosphate; ACP: acyl carrier protein; NAD(P): nicotinamide adenine dinucleotide (phosphate); FMN: flavin mononucleotide; THF: tetrahydrofolate; Sec: Sec secretion system; G(A)SH: glutathione (amide); T3SS: type III secretion system.

Given the large proportion of functionally annotated genes that relate to information processing, we compared P. abstrusum's genome investment to a representative selection of the alphaproteobacterial diversity including close relatives. We therefore analyzed the cellular processes encoded in each genome based on the distribution of the genes to clusters of orthologous groups (COGs). COG ategory F (nucleotide metabolism and transport) and J (translation, ribosomal structure and biogenesis) are main drivers for a distinct pattern in P. abstrusum and cover over 6.5% and 30.1% of the genes, respectively. In summary, the genomic investments of P. abstrusum lead to a remarkable metabolic profile compared to representatives of other alphaproteobacterial families and other orders with *incertae sedis* status such as Holosporales (Figure 6, Supplement section 10, Figure S8 and S9).

Protein synthesis is based on external amino acid supply

P. abstrusum is dependent on amino acid import as it can only synthesize one amino acid *de novo* and encodes genes for partial synthesis or conversion of only three amino acids. The genome encodes for aspartate synthesis from fumarate and ammonia, asparagine formation from aspartate and glutamine, glutamine formation from glutamate and ammonia and glycine

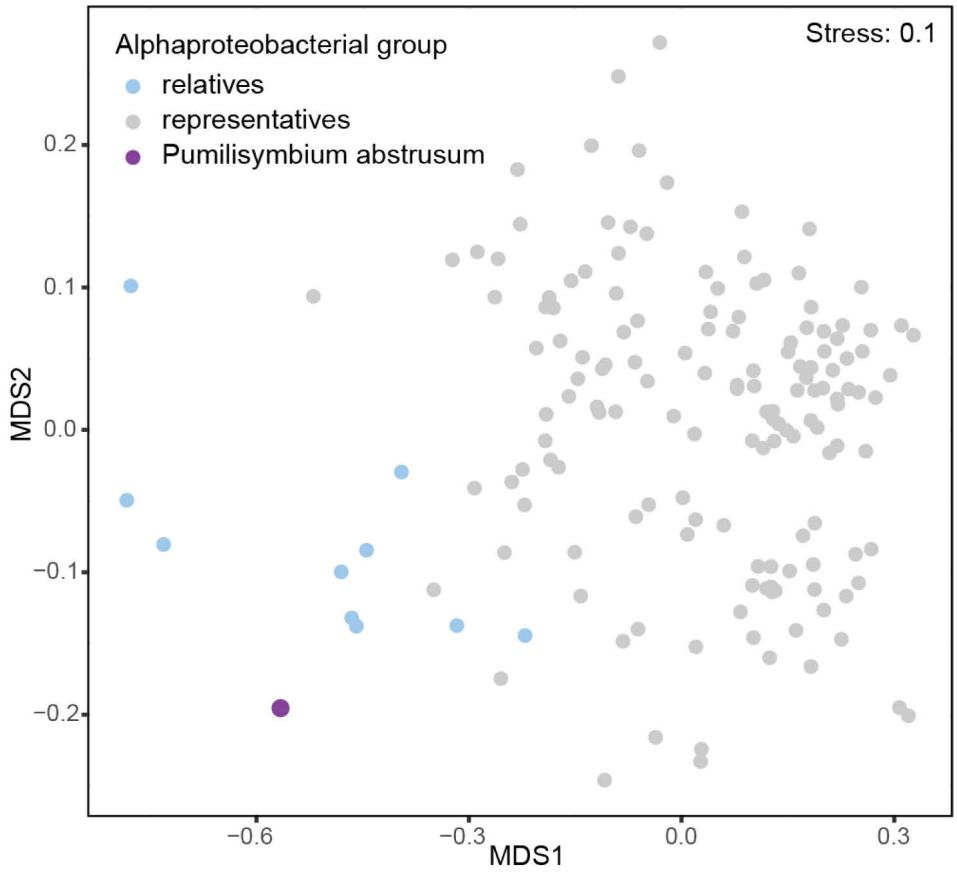


Figure 6: P. abstrusum (purple) has an outstanding COG profile compared to its phylogenomic relatives (light-blue) and representatives of alphaproteobacterial families (grey).
Non-metric multidimensional scaling (NMDS) of feature counts per COG category of the respective MAGs.

conversion to serine (Figure 5). The amino acid import from the environment, in this case likely the host, can be facilitated by a set of amino acid carriers, e.g. proline transporter (opuE and putP), amine/amino acid antiporter (cadB) and a dicarboxylate/amino acid:cation symporter. Various peptidases and proteases are encoded which would allow the efficient recycling of unused or damaged proteins. These include specific peptidases for methionine and proline, general (oligo)peptidases (pepA, pepF) and protein degradation complex like proteins (hslUV, clpPX).

Pumilisymbium abstrusum is capable of autonomous ATP synthesis

One way for a symbiont to minimize the metabolic impact on the host organisms is the ability to regenerate ATP from ADP. Our metabolic reconstruction of P. abstrusum revealed two options for ATP generation, glycolysis and an F0F1 type ATP synthase (Figure 5). Key enzymes for glycolysis are encoded in the genome of P. abstrusum. A fructokinase enables the start from D-fructose. The fructose-bisphosphate aldolase that would convert fructose 1,6-

bisphosphate to dihydroxyacetone phosphate and glyceraldehyde-3-phosphate could, however, not be annotated. Enzymes for all other canonical steps until the formation of pyruvate are encoded. It is unclear which fate the pyruvate formed by glycolysis has. Neither anoxic, fermentative processes nor pyruvate oxidation or carboxylation that would convert it to acetate and connect it to a TCA cycle is annotated. P. abstrusum does not encode for any genes that use acetyl-CoA or any kind of Coenzyme A connected substrates. The encoded glycolysis would lead to a net gain of 2 ATP per fructose molecule and 2 $NADH+H^+$.

A second important substrate likely used by P. abstrusum is malate, which is present in high relative abundance in the worm and likely generated during the host's anoxic metabolism[57]. The symbiont could import malate via a dicarboxylate/amino acid:cation symporter and on one hand generate $NADPH+H^+$ and pyruvate and on the other hand channel it into ATP *de novo* synthesis via fumarate and the conversion to aspartate. The charging of the ATP could be performed by an ATP synthase of the F0F1 type of which all eight subunits were detected. Such ATP formation requires a proton gradient but we could not detect a functional respiration chain. Only two subunits of relevant enzymes could be annotated, an NAD(P)H-quinone oxidoreductase subunit and a subunit of the cytochrome bc1 complex. Absence of a respiration chain in combination with presence of an ATP synthase has been reported for endosymbionts with small genomes, e.g. in the next relatives Holosporales and Rickettsiales as well as in unrelated Flavobacteriales[6, 58]. However, unlike other highly reduced endosymbionts, P. abstrusum possesses a pyrophosphate-energized proton pump that could generate the necessary proton gradient. This proton pump is K-stimulated and hydrolyzes pyrophosphate (PPi) to phosphate to transport H^+ across the inner membrane. PPi producing reactions in P. abstrusum are at least performed by DNA and RNA polymerases, aminoacyl-tRNA synthetases as well as the CDP-diacylglycerol synthase in its lipid metabolism. This link between several PPi generating processes, the proton pump and the ATP synthase would be an efficient way to conserve energy. A similar energy conservation process is also used by the main symbionts in the same host[57]. However, phylogenetic analyses suggest that P. abstrusum did not acquire the genes for this proton pump recently via horizontal gene transfer but possesses its own unique type (Supplement section 10, Figure S10). Taken together, PPi based energy conservation seems to play a general role in the *O. algarvensis* symbiosis.

Host modulation via T3SS and eukaryote like virulence factors

In addition to metabolic integration, P. abstrusum encodes a type III secretion system (T3SS) to actively interact with the host cells. The T3SS is one of the most complex bacterial protein apparatuses with a needle like structure to inject effector proteins via the host membrane[59]. We were able to annotate 17 of the 20 principal component proteins of the inner and outer membrane complex, the connector proteins, the needle and a variety of chaperones as well as the necessary Sec system components for early assembly[60]. We could not detect genes for needle length regulation, the needle tip and one of the two translocator pore proteins, but these three genes have a low degree of conservation or lack homologs in other characterized T3SS systems[59]. The microsynteny blocks encoding for the T3SS proteins are similar to what has been reported for other bacteria (Supplement section 10, Figure S11)[61]. Functional replacements for the missing genes could be encoded in the seven hypothetical proteins associated with the five microsynteny blocks. Phylogenetic analyses with the most conserved genes place P. abstrusum's T3SS in a diverged position without direct relatives (Supplement Figure S12-17)[59]. Given the very likely functional T3SS we screened the MAG for potentially secreted proteins[37]. Out of the 658 proteinogenic genes 114 proteins were predicted to contain a signal peptide for Type III secretion and 39 to serve as chaperone (annotation information is provided at zenodo.org/record/6515015)[37]. This abundance of potentially secreted proteins underlines the relevance of the T3SS for the interaction of P. abstrusum with its host.

In addition to T3SS based modulation, we found representatives of several gene families that could be involved in host modulation and immune response escape. Among them are proteins containing eukaryote-like domains such as a patatin-like phospholipase family protein. These enzymes have been shown to serve as a storage glycoprotein but also play a role in several biological processes like membrane cleavage, sepsis induction, host colonization or triglyceride metabolism[62]. A patatin-like protein in *P. aeruginosa* was shown to be secreted via a T3SS[62]. However, the patatin-like protein from P. abstrusum does not have a signal peptide for type III secretion. Other eukaryote-like proteins are an interleukin 8-like chemokine protein and a somatomedin B domain containing protein that both might play a role in pathogenicity, host recognition escape and internalization into the host[63, 64]. We detected another protein family potentially involved in cell surface mediated virulence, a PE/PPE dimer-like domain containing protein. Such PE/PPE proteins have been hypothesized to play a role in the virulence and intracellular survival of mycobacteria[65].

Membrane lipids and proteins

Given the trend in reduced endosymbionts to lose the capability to synthesize the peptidoglycan layer as well as their own membrane lipids, we investigated the completeness of the relevant pathways. P. abstrusum does not encode peptidoglycan synthesis which supports our hypothesis that P. abstrusum is an intracellular symbiont. In contrast to the missing peptidoglycan synthesis, P. abstrusum seems capable of phospholipid biosynthesis, despite few annotated genes involved in lipid metabolism. Intermediate products of glycolysis, dihydroxyacetone phosphate (glycerone phosphate), can be transformed into phosphatidylglycerol via six enzymatic reactions of which only one enzyme, the glycerol-3-phosphate acyltransferase, is missing from the annotation. An ACP-like protein could serve as acyl-group donor. We furthermore detected a phosphatidylserine decarboxylase proenzyme that likely is involved in biosynthesis of phosphatidylethanolamine. In summary, the capability of membrane biosynthesis would render P. abstrusum at an intermediate stage of reduction and host dependence.

The interface of intracellular symbionts to their hosts are membranes and membrane-bound proteins. We therefore systematically screened the predicted proteins for transmembrane domains to identify proteins located on the membranes. Around a third of all proteins in P. abstrusum contain a transmembrane domain and appear to be linked to membranes (annotation information is provided at zenodo.org/record/6515015). These include proteins from cell machineries, e.g. the Sec secretion system, T3SS proteins and the ATP synthase, transport systems, e.g. ABC transporter, multidrug efflux proteins or ion channels, and proteins involved in central cell metabolism. However, more than half of the proteins with a transmembrane domain remain without functional annotation. This ratio of hypothetical to annotated proteins is much higher than in the proteins without transmembrane domains (50.8% vs. 34.6%). We could detect several blocks of hypothetical genes that all feature transmembrane proteins. This high number of membrane bound proteins without annotation illustrates that a crucial part of the interaction and function of P. abstrusum in its host awaits biochemical characterization.

The importance of membrane bound proteins and transmembrane transfer in P. abstrusum is underlined by the variety of chaperones and proteases that play a role in folding and quality control of integral membrane proteins. Among these are cytoplasmic chaperones that fold proteins and prepare them for transmembrane transport, e.g groL and groES or metalloprotease

ftsH. In addition, we detected periplasmic proteins, e.g. surA with general chaperone activity and an OmpH family outer membrane protein that interacts with unfolded proteins secreted by the Sec system[66]. Furthermore, three of the five bam complex proteins for beta-barrel assembly in the outer membrane (bamABD) along with the beta-barrel assembly-enhancing protease bepA were annotated. Signal peptidases such as the lipoprotein signal peptidase (lspA) and signal peptidase I (lepB) are also present and could cleave signal sequences from secreted or periplasmic proteins marked for transport.

Vitamin and co-factor synthesis

As P. abstrusum does not have the combination of traits typical for intracellular parasites, e.g. reliance on host ATP and host membrane lipids, we searched for potentially beneficial metabolic contributions. Exploring vitamin and co-factor synthesis, we found that P. abstrusum could partially synthesize or salvage a selected set, e.g. the vitamins B1 and B9. Genes involved in Vitamin B1 (thiamine) salvage from thiamine fragments potentially available from the host metabolism are present in the genome. P. abstrusum encodes the majority of genes for Coenzyme F (tetrahydrofolate, vitamin B9) *de novo* biosynthesis and the one-carbon pool mediated by folate, except the tetrahydrofolate synthase (folC) itself. Folate mediated one-carbon transfer plays a role in amino acid biosynthesis and purin and pyrimidine *de novo* synthesis and other basic cell functions. Despite the patchy nature of tetrahydrofolate one-carbon metabolism, the presence of a serine hydroxymethyltransferase that is connected is another indication for the pathway's relevance to P. abstrusum. The symbiont has most genes for NAD *de novo* biosynthesis and salvage and is likely able to regenerate NAD and to convert NAD to NADP. In addition, it is capable of FAD *de novo* synthesis from riboflavin (vitamin B2). Other potential electron carriers are glutathione and glutaredoxin, putative glycerol-3-phosphate dehydrogenase 2 [NAD(P)] and a DsbA family protein (thioredoxin). Overall, P. abstrusum can salvage or synthesize several potentially beneficial co-factors and vitamins, but it remains unclear whether these can be provided to the host or are largely used to support its own metabolic independence.

Lysogenic phage in genomic island

We were surprised to detect a genomic region encoding a tailed dsDNA bacteriophage, given the small genome size of P. abstrusum. Seven phage genes were annotated within a sequence length of 9.9 Kbp and a total of 14 genes, the remaining seven are hypotheticals. The annotated genes include a major capsid protein, a HK97 prohead protease and a phage portal protein as

well as a putative tail and could form structural components of the lytic phage. Such components are typical features of *Caudovirales*, the largest group of viruses described [67-69]. We also detected a small terminase, the recognition subunit of the viral DNA, a large terminase phage packaging protein and an endonuclease/translocase subunit, which together with the portal protein could translocate one viral genome inside the phage capsid. A putative Holin involved in gene-transfer release could play a role in lysis of the host cell membrane during a lytic phase[70]. The presence of an apparently complete and functional phage suggests the possibility for infections at least at some stage in the symbionts' life cycle.

Discussion

Technological challenges hide low abundant and reduced bacteria

Symbionts with reduced genomes that only occur in low abundance per host individual have been rarely observed both in marine and terrestrial animals. Compared to nutritional symbionts, their low biomass hinders direct observation. These symbionts may evade detection when they do not invoke obvious strong detrimental or beneficial phenotypes. From a bioinformatics perspective, their low abundance is a technological challenge. Even if such symbionts are detected, commonly used cut-offs would label their MAGs that represent full genomes as incomplete. Technological innovations such as graph based screening and other marker sets, e.g. tRNAs and ribosomal proteins, are necessary as completeness indicator and detection measures. A second limitation for detection would be if not every member of a population is infected, as it appears to be the case with P. abstrusum in *O. algarvensis*. Detection despite low infection rates can be achieved with increased sample numbers and systematic and high-throughput screening approaches that are independent of prior database knowledge. The technical limitations that prevented the detection of this low abundant and likely intracellular symbiont have likely precluded the detection of taxa with a similar life-style and set of traits.

Few representatives of strongly reduced bacteria in marine animal-microbe symbioses

The genome size of 0.64 Mbp makes P. abstrusum one of the few bacteria with very small genomes reported for the marine realm. The apparently low number of marine strongly reduced bacteria is likely a biased distribution due to the overwhelming number of studies based on such genomes from insect symbioses. The smallest complete marine genome reported to date was sequenced from a hadal sea cucumber (Figure 1.1). *Ca.* Spiroplasma holothuricola has a genome of 0.42 Mpb distributed on two chromosomes and appeared to be the dominant species of hindgut samples[7]. Other marine reduced genomes were associated with diverse animal

hosts, i.e. bacteria of a novel phylum extracted from bryozoans, Neorickettsia from salmonid fish, and a variety of *Mycoplasma* spp. from larger marine animals (Figure 1). However, despite their strongly reduced genomes and their assumed host dependence, none of these animal-associated reduced bacteria has been characterized in detail and their roles for the host remain unknown.

Overall, only few bacterial genomes with a genome size below 1 Mbp are known from the marine habitat. More than 20 of the described 34 animal phyla occur predominantly if not exclusively in the marine realm. We expect the diversity of novel microbial lineages and particularly those with highly reduced genomes in marine animal-microbe symbioses to be much more versatile than anticipated today. We postulate a major increase in representative genomes once available and newly obtained sequencing data sets are systematically screened for them.

Pumilisymbium abstrusum combines a unique set of traits

We could show that P. abstrusum is an early branching Alphaproteobacterium and is part of the sister clade to Rickettsiales. Almost all known examples of early branching Alphaproteobacteria intracellularly colonize a variety of hosts from the animal, plant and the protist kingdom[71-73]. Despite their largely pathogenic role, some Rickettsiales were also shown to have commensalistic or mutualistic interactions[74, 75]. Almost all Rickettsiales have a reduced genome size of 1-1.5 Mbp and many of the described are intracellular energy parasites that rely on their hosts for the majority of nucleotide, amino acid and vitamin provisioning[76]. In contrast to typical energy parasites, P. abstrusum does not possess the ADP/ATP translocase that Rickettsiales commonly use to tap into the ATP pool of the host[76]. Instead, it generates its own ATP via glycolysis and an ATP synthase linked to a pyrophosphate-energized proton pump. Such a pyrophosphate-energized proton pump is an ancient mechanism to conserve energy and likely minimizes P. abstrusum's negative impact on the host[77]. However, despite its low necessary investment of a single gene, this type of energy conservation is unknown in Rickettsiales. Another outstanding feature of P. abstrusum compared to Rickettsiales is its T3SS. Members of Rickettsiales often possess a T4SS or T6SS for communication and invasion of the eukaryotic target. T3SSs have been observed in several major orders of Alphaproteobacteria, none of which are branching close to the root of the alphaproteobacterial diversity[61]. The deeply branching phylogenetic position of both the T3SS and the pyrophosphate-energized proton pump sequences from P. abstrusum make any recent

acquisition of these genes via horizontal gene transfer highly unlikely (Supplement Figure S10 and S12-S17). These single gene analyses combined with the deep phylogenetic position of P. abstrusum suggest that ATP synthesis linked to pyrophosphate and a T3SS for host-symbiont interaction might be an ancestral state and potentially more widespread among Alphaproteobacteria.

A more mutualistic than parasitic intracellular lifestyle?

Based on the high level of genome reduction, the lack of essential metabolic functions such as amino acids biosynthesis and lack of a peptidoglycan layer, P. abstrusum appears host dependent throughout its complete life cycle, occupying an intracellular niche. Despite this high host dependence, the impact on the host is likely very low as P. abstrusum's metabolism only uses commonly available substrates such as fructose and abundant intermediates such as malate[57]. In contrast to many strongly reduced obligate endosymbionts, P. abstrusum seems capable of its own energy conservation, partial membrane lipid biogenesis and even synthesis and salvaging of a selection of B vitamins and co-factors. These co-factors and vitamins are likely produced for its own needs, but overall, any generated biomass including vitamins and other valuable substrates will eventually be recycled by the host. Such minor benefits can only be net gains for the host if the symbionts do not evoke a costly immune reaction[78]. The symbiont combines several features that likely allows it to evade host detection or to modulate host immune response. P. abstrusum lacks a cell wall and therefore already avoids detection and host immune response triggered by cell wall components. In addition, the symbiont genome features a considerable amount of genes dedicated to host communication and modulation by the T3SS. Such an extensive genomic investment is rare in reduced intracellular bacteria and points to the critical role of host modulation for survival in the host population. In summary, the role of P. abstrusum could be commensal or even mutualistic depending on the importance of the provided substrates in a given environment.

An annelid reduced bacterial clade?

A major problem in the early branching clades of the Alphaproteobacteria is low taxonomic sampling. The closest relatives to P. abstrusum based on 16S rRNA gene sequences and MAGs are representatives from other family- to order-level clades. Based on 16S rRNA genes, a single close relative stands out which is associated with another annelid, the arctic ice worm *Mesenchytraeus solifugus*[51]. This observation makes it tempting to speculate that P. abstrusum and the *M. solifugus* associated bacterium are representatives of a much more widespread

bacterial clade that is coevolving with annelid hosts but has evaded detection. Since no genomic data is available, we cannot draw any conclusion on the genomic state of the *M. solifugus* symbiont. Broader taxon sampling of the staggering diversity of Annelida with state-of-the-art metagenomics techniques tailored to detect reduced genomes could provide additional hosts for this bacterial clade.

The larger relations of this potentially annelid associated clade remain elusive. Our analyses place the clade within the Rickettsiales *sensu lato*, a radiation of undersampled family- to order-level clades that associate with unicellular hosts and that also includes the only described Rickettsiales that can reproduce outside of eukaryote cells, *Ca.* Deianiraea vastatrix[55]. A typical sign for accelerated genome evolution in bacteria with highly reduced genomes would be long branches in phylogenetic reconstructions[2]. We, however, did not observe a specifically long-branch for P. abstrusum neither in the 16S rRNA gene based analysis nor in the phylogenomic analysis based on 81 marker genes. This apparently normal evolutionary rate is a surprising observation given the fact that most animal-associated symbionts with genome sizes below 0.7 Mbp exhibit accelerated evolution[2]. This suggests that P. abstrusum does not suffer from consequences that bacteria with small population sizes and population bottlenecks during transmission usually face.

Conclusion

Candidatus Pumilisymbium abstrusum is a bacterium with one of the most highly reduced genomes described from a marine animal-microbe symbiosis. It possesses a novel combination of traits with a type III secretion system to modulate its host and the ability to generate ATP that has not been described for Rickettsiales related bacteria before. Future research will shed light on its specific role in the *Olavius algarvensis* symbiosis. We expect P. abstrusum to be a representative of a yet undetected diversity of genome-reduced bacteria from the marine realm, not limited to annelid hosts. Expanding our knowledge on specific interactions with genome-reduced bacterial partners from marine habitats will help decipher evolutionary patterns of genome reduction linked to animal ecophysiology, which are to date almost exclusively researched in insect nutritional symbioses.

Data and code availability

Raw metagenomic sequences, symbiont marker genes and symbiont MAGs generated in this study will be deposited in the European Nucleotide Archive (ENA) upon peer-review submission and are currently available upon request.

The scripts that were used for data visualization concerning abundance estimation and COG profiling are available at github.com/TinaEnd/Alpha_reduced_Pumilisymbium.

Phylogenetic 16S rRNA gene and MAG based trees are available in project Alpha_reduced_Pumilisymbium at itol.embl.de/shared/tenders.

Acknowledgements

We thank Christian Lott, Miriam Weber and the HYDRA Marine Sciences team as well as the Elba field trip team 2020 for sampling and field assistance. This work was supported by the Max Planck society and sequencing was conducted at the Max Planck Genome Centre Cologne. We thank Marlene Jensen and Manuel Kleiner (NCSU, Raleigh, USA) for their thorough attempts to provide proteome data. We thank Martina Meyer, Silke Wetzel, Miriam Sadowski and Wiebke Ruschmeier for technical support in the laboratory.

References

1. Toft C, Andersson SG. Evolutionary microbial genomics: Insights into bacterial host adaptation. Nature Reviews Genetics. 2010;11(7):465-75.
2. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology. 2012;10(1):13-26.
3. Giovannoni SJ, Thrash JC, Temperton B. Implications of streamlining theory for microbial ecology. The ISME journal. 2014;8(8):1553-65.
4. Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: Perhaps more? Biochimie. 2012;94(1):110-6.
5. Moran NA, Bennett GM. The tiniest tiny genomes. Annual review of microbiology. 2014;68:195-215.
6. George EE, Husnik F, Tashyreva D, Prokopchuk G, Horák A, Kwong WK, et al. Highly reduced genomes of protist endosymbionts show evolutionary convergence. Current Biology. 2020;30(5):925-33. e3.
7. He L-S, Zhang P-W, Huang J-M, Zhu F-C, Danchin A, Wang Y. The enigmatic genome of an obligate ancient *Spiroplasma* symbiont in a hadal holothurian. Applied and environmental microbiology. 2018;84(1).
8. Miller IJ, Weyna TR, Fong SS, Lim-Fong GE, Kwan JC. Single sample resolution of rare microbial dark matter in a marine invertebrate metagenome. Scientific reports. 2016;6(1):1-10.
9. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature. 2006;443(7114):950-5.
10. Dubilier N, Mülders C, Ferdelman T, de Beer D, Pernthaler A, Klein M, et al. Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. Nature. 2001;411(6835):298-302.
11. Mankowski A, Kleiner M, Erséus C, Leisch N, Sato Y, Volland J-M, et al. Highly variable fidelity drives symbiont community composition in an obligate symbiosis. bioRxiv. 2021.
12. Sato Y, Wippler J, Wentrup C, Ansorge R, Sadowski M, Gruber-Vodicka H, et al. Fidelity varies in the symbiosis between a gutless marine worm and its microbial consortium. bioRxiv. 2021.

13. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012;19(5):455-77.
14. Nikolenko SI, Korobeynikov AI, Alekseyev MA, editors. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC genomics; 2013: Springer.
15. Gruber-Vodicka HR, Seah BK, Pruesse E. phyloFlash: Rapid small-subunit rRNA profiling and targeted assembly from metagenomes. Msystems. 2020;5(5).
16. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-6.
17. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, et al. MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3-11.
18. Bushnell B. BBmap. v38.34 ed2010.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
20. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
21. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
22. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of *de novo* genome assemblies. Bioinformatics. 2015;31(20):3350-2.
23. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: Circular and linear interactive genome visualization. Bioinformatics. 2009;25(1):119-20.
24. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015;32(1):268-74.
25. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10.
27. Floriano AM, Castelli M, Krenek S, Berendonk TU, Bazzocchi C, Petroni G, et al. The genome sequence of "*Candidatus* Fokinia solitaria": Insights on reductive evolution in Rickettsiales. Genome biology and evolution. 2018;10(4):1120-6.
28. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press; 2020.
29. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic acids research. 2019;47(W1):W256-W9.
30. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid annotations using subsystems technology. BMC genomics. 2008;9(1):1-15.
31. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic acids research. 2014;42(D1):D206-D14.
32. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific reports. 2015;5(1):1-6.
33. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
34. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, et al. SUPERFAMILY — Sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic acids research. 2009;37:D380-D6.
35. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of molecular biology. 2001;313(4):903-19.
36. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic acids research. 1997;25(17):3389-402.
37. Eichinger V, Nussbaumer T, Platzer A, Jehl M-A, Arnold R, Rattei T. EffectiveDB — Updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. Nucleic acids research. 2016;44(D1):D669-D74.

38. Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, et al. Pathway Tools version 23.0 update: Software for pathway/genome informatics and systems biology. Briefings in bioinformatics. 2021;22(1):109-26.
39. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic acids research. 2014;42(D1):D459-D71.
40. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Science. 2019;28(11):1947-51.
41. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.
42. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: Integrating viruses and cellular organisms. Nucleic acids research. 2021;49(D1):D545-D51.
43. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular biology and evolution. 2017;34(8):2115-22.
44. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. 2019;47(D1):D309-D14.
45. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics. 2010;11(1):1-11.
46. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015;12(1):59-60.
47. Junier T, Zdobnov EM. The Newick utilities: High-throughput phylogenetic tree processing in the Unix shell. Bioinformatics. 2010;26(13):1669-70.
48. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Scientific reports. 2016;6(1):1-9.
49. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: A software environment for sequence data. Nucleic acids research. 2004;32(4):1363-71.
50. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research. 2015;25(7):1043-55.
51. Murakami T, Segawa T, Bodington D, Dial R, Takeuchi N, Kohshima S, et al. Census of bacterial microbiota associated with the glacier ice worm *Mesenchytraeus solifugus*. FEMS microbiology ecology. 2015;91(3).
52. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology. 2014;12(9):635-45.
53. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nature biotechnology. 2020;38(9):1079-86.
54. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature biotechnology. 2018;36(10):996-1004.
55. Castelli M, Sabaneyeva E, Lanzoni O, Lebedeva N, Floriano AM, Gaiarsa S, et al. *Deianiraea*, an extracellular bacterium associated with the ciliate *Paramecium*, suggests an alternative scenario for the evolution of Rickettsiales. The ISME journal. 2019;13(9):2280-94.
56. Muñoz-Gómez SA, Hess S, Burger G, Lang BF, Susko E, Slamovits CH, et al. An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. Elife. 2019;8:e42535.
57. Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. Proceedings of the National Academy of Sciences. 2012;109(19):E1173-E82.
58. Bennett GM, Moran NA. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome biology and evolution. 2013;5(9):1675-88.
59. Diepold A, Wagner S. Assembly of the bacterial type III secretion machinery. FEMS microbiology reviews. 2014;38(4):802-22.
60. Lara-Tejero M, Galán JE. The injectisome, a complex nanomachine for protein injection into mammalian cells. Protein Secretion in Bacteria. 2019:245-59.
61. Hu Y, Huang H, Cheng X, Shu X, White AP, Stavrinides J, et al. A global survey of bacterial type III secretion systems and their effectors. Environmental microbiology. 2017;19(10):3879-95.

62. Banerji S, Flieger A. Patatin-like proteins: A new family of lipolytic enzymes present in bacteria? Microbiology. 2004;150(3):522-5.
63. Sansonetti P, Arondel J, Huerre M, Harada A, Matsushima K. Interleukin-8 controls bacterial transepithelial translocation at the cost of epithelial destruction in experimental shigellosis. Infection and immunity. 1999;67(3):1471-80.
64. Zhou A. Functional structure of the somatomedin B domain of vitronectin. Protein science. 2007;16(7):1502-8.
65. Fishbein S, Van Wyk N, Warren R, Sampson S. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. Molecular microbiology. 2015;96(5):901-16.
66. Behrens-Kneip S. The role of SurA factor in outer membrane protein transport and virulence. International Journal of Medical Microbiology. 2010;300(7):421-8.
67. Prevelige Jr PE, Cortines JR. Phage assembly and the special role of the portal protein. Current opinion in virology. 2018;31:66-73.
68. Duda RL, Martincic K, Hendrix RW. Genetic basis of bacteriophage HK97 prohead assembly. Journal of molecular biology. 1995;247(4):636-47.
69. Xie Z, Hendrix RW. Assemblyin Vitroof Bacteriophage HK97 Proheads. Journal of molecular biology. 1995;253(1):74-85.
70. Saier MH, Reddy BL. Holins in bacteria, eukaryotes, and archaea: multifunctional xenologues with potential biotechnological and biomedical applications. Journal of bacteriology. 2015;197(1):7-17.
71. Kaur R, Shropshire JD, Cross KL, Leigh B, Mansueto AJ, Stewart V, et al. Living in the endosymbiotic world of *Wolbachia*: A centennial review. Cell Host & Microbe. 2021.
72. Mahajan SK. Rickettsial diseases. J Assoc Physicians India. 2012;60:37-44.
73. Perlman SJ, Hunter MS, Zchori-Fein E. The emerging diversity of Rickettsia. Proceedings of the Royal Society B: Biological Sciences. 2006;273(1598):2097-106.
74. Fujishima M, Kawai M, Yamamoto R. *Paramecium caudatum* acquires heat-shock resistance in ciliary movement by infection with the endonuclear symbiotic bacterium *Holospora obtusa*. FEMS microbiology letters. 2005;243(1):101-5.
75. Fenn K, Blaxter M. *Wolbachia* genomes: Revealing the biology of parasitism and mutualism. Trends in parasitology. 2006;22(2):60-5.
76. Driscoll TP, Verhoeve VI, Guillotte ML, Lehman SS, Rennoll SA, Beier-Sexton M, et al. Wholly Rickettsia! Reconstructed metabolic profile of the quintessential bacterial parasite of eukaryotic cells. MBio. 2017;8(5).
77. Serrano A, Pérez-Castiñeira JR, Baltscheffsky M, Baltscheffsky H. H^+-PPases: Yesterday, today and tomorrow. IUBMB life. 2007;59(2):76-83.
78. Gerardo NM, Hoang KL, Stoy KS. Evolution of animal immunity in the light of beneficial symbioses. Philosophical Transactions of the Royal Society B. 2020;375(1808):20190601.

Chapter 3 | Oalg-alpha – a strongly reduced symbiont

Candidatus Pumilisymbium abstrusum, a highly reduced and deeply branching Alphaproteobacterium in symbiosis with marine invertebrate gutless oligochaetes (Oligochaeta, Annelida)

Supplementary information

Tina Enders[1], Alexander Gruhl[1], Manuel Kleiner[2], Nikolaus Leisch[1], Anna Mankowski[1,3], Yui Sato[1], Nicole Dubilier[1], Harald R. Gruber-Vodicka[1]

[1] Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

[2] Department of Plant & Microbial Biology, North Carolina State University, Raleigh 27695, North Carolina, USA

[3] Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Corresponding authors:
Nicole Dubilier, ndubilie@mpi-bremen.de,
Harald R. Gruber-Vodicka, hgruber@mpi-bremen.de

1 Supplementary methods

Protein trees for pyrophosphatase (Figure S10) and T3SS components (Figure S12-S17) were calculated with the according protein sequence of P. abstrusum, related protein sequences from BLAST implemented in geneious with the nr_nt, env_nr, refseq_protein and swissprot database, and sequences with the annotated function from *Olavius algarvensis* symbionts. Annotated sequences for secretion systems were obtained by searching *O. algarvensis* symbiont bins with hmmsearch (hmmer v3.1b2, hmmer.org/) against a database constructed with MacSyFinder (github.com/gem-pasteur/macsyfinder)[1]. Annotated sequences for the pyrophosphatase were obtained by annotating *O. algarvensis* symbiont bins with the custom annotation option in prokka (--proteins)[2]. Pyrophosphatase sequences from other *O. algarvensis* symbionts and their closest relatives from NCBI were used[3]. The alignments were produced with mafft-einsi and the trees calculated with IQ-TREE (-m MFP -alrt 1000 -bb 1000). Trees were annotated in iTOL.

Code used for Figure S7-S9 and Figure S2-S4 is available at: github.com/TinaEnd/Alpha_reduced_Pumilisymbium.

Phylogenomic trees based on 16S rRNA genes, MAGs and protein trees are available at itol.embl.de/shared/tenders under project Alpha_reduced_Pumilisymbium.

2 Software and tools used

Table S1: List of tools and software packages used with relevant publications and web access (continued on following pages).

Software	Reference	Availability
Adobe Illustrator v25.3	webpage	https://adobe.com/products/illustrator
ARB	[4]	http://www.arb-home.de/home.html
Bandage	[5]	https://rrwick.github.io/Bandage/
BayesHammer	[6]	http://bioinf.spbau.ru/en/spades/bayeshammer
BBtools	webpage	https://sourceforge.net/projects/bbmap/ https://jgi.doe.gov/data-and-tools/bbtools/
BLAST and psiBLAST	NCBI [7, 8]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
checkM	[9]	https://ecogenomics.github.io/CheckM/
diamond	[10]	https://github.com/bbuchfink/diamond
DNAplotter (artemis)	[11]	http://sanger-pathogens.github.io/Artemis/DNAPlotter/

effectiveDB	[12]	https://effectors.csb.univie.ac.at/
eggNOG-mapper v2.1.6	[13]	https://github.com/eggnogdb/eggnog-mapper
FastQC	webpage	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FinchTV	website	https://digitalworldbiology.com/FinchTV
Geneious v11.1.5	website	https://www.geneious.com
GTDB-Tk v1.5.0 GTDB R202	[14-16]	https://github.com/Ecogenomics/GTDBTk https://gtdb.ecogenomic.org/
Hmmer v3.1b2	webpage	http://hmmer.org/
Hmmscan (Pfam)	webpage	https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan http://hmmer.org/
IQ-TREE v1.6.10	[17, 18]	http://www.iqtree.org/
IMNGS	[19]	https://www.imngs.org/
iTOL	[20]	https://itol.embl.de/
KEGG	[21-23]	https://www.genome.jp/kegg/
MacSyFinder	[1]	https://github.com/gem-pasteur/macsyfinder
MAFFT v7.407	[24]	https://mafft.cbrc.jp/alignment/software/
MEGAHIT v1.2.8	[25, 26]	https://github.com/voutcn/megahit
MetaBAT	[27, 28]	https://bitbucket.org/berkeleylab/metabat/src/master/
MetaCyc	[29]	https://metacyc.org/
Pathway Tools	[30]	http://bioinformatics.ai.sri.com/ptools/
Pfam	[31]	http://pfam.xfam.org/
PhyloFlash	[32]	http://hrgv.github.io/phyloFlash/
Prodigal v2.6.3	[33]	https://github.com/hyattpd/Prodigal
Prokka v1.14.5	[2]	https://github.com/tseemann/prokka
RAST	[34-36]	http://rast.theseed.org/FIG/rast.cgi
RStudio v1.4.1106	webpage	https://rstudio.com/
R v4.0.4	[37]	https://www.R-project.org/
R: devtools	webpage	https://github.com/r-lib/devtools
R: data.table	webpage	https://github.com/Rdatatable/data.table
R: dplyr	webpage	https://dplyr.tidyverse.org/
R: ggfortify	webpage	https://github.com/sinhrks/ggfortify

R: ggplot2	webpage	https://github.com/tidyverse/ggplot2
R: htmltools	webpage	https://rstudio.github.io/htmltools/
R: plotly	webpage	https://plotly.com/r/
R: reshape	webpage	https://github.com/hadley/reshape
R: stringr	webpage	https://stringr.tidyverse.org/
R: tidyr	webpage	https://tidyr.tidyverse.org/
R: vegan	webpage	https://github.com/vegandevs/vegan
Samtools v1.9	[38]	http://www.htslib.org/
SimpleMappr	website	https://www.simplemappr.net
SPAdes v3.12.0	[6, 39]	http://cab.spbu.ru/software/spades/
SUPERFAMILY	[40, 41]	https://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/
TMHMM 2.0	website	https://services.healthtech.dtu.dk/service.php?TMHMM-2.0

3 Metadata to sampled individuals

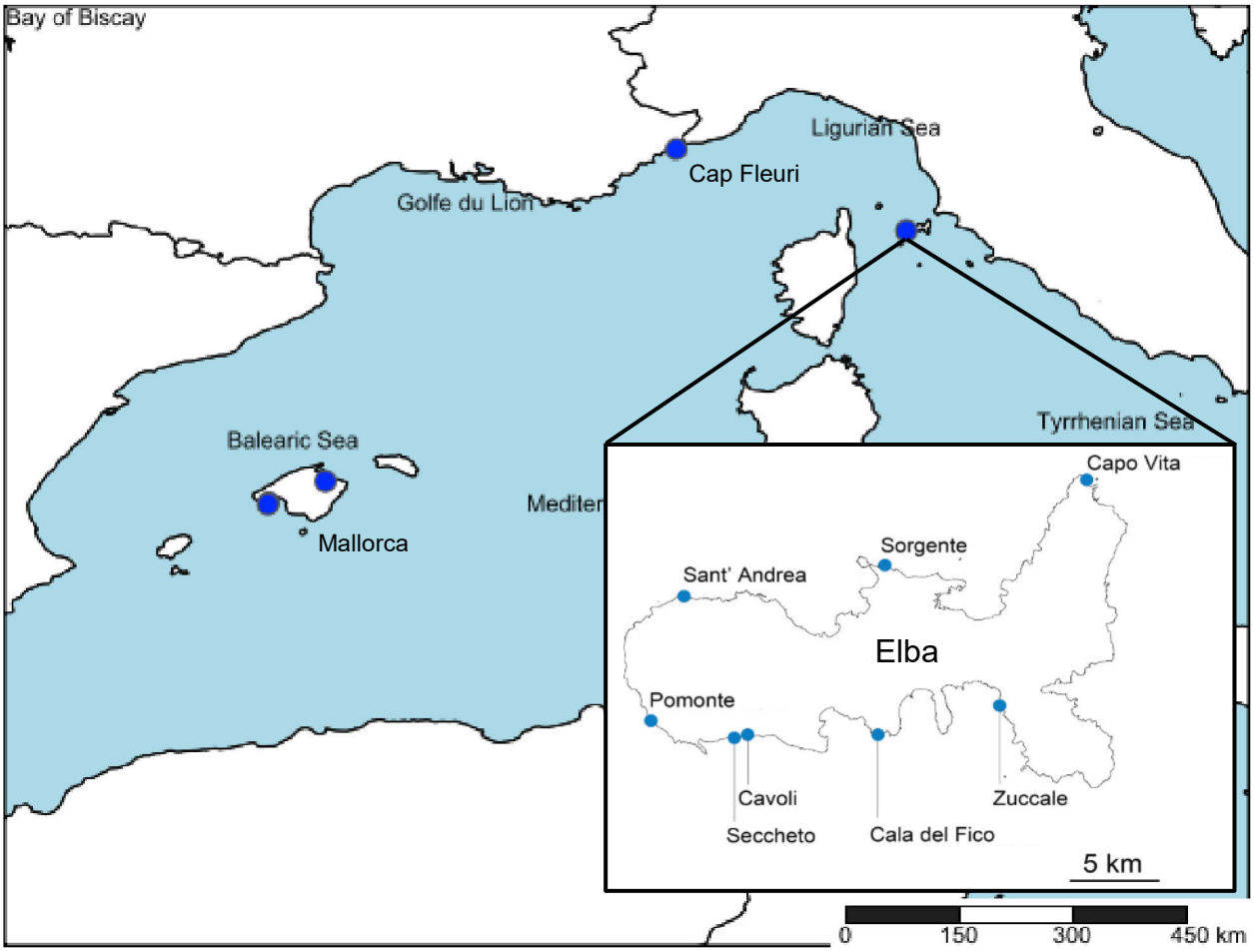


Figure S1: Sampling bays around Elba and Pianosa in Italy, Mallorca in Spain and Cap Fleuri in France in the Mediterranean Sea (www.simplemappr.net).

Metdadata for all specimens used in this study is listed in at zenodo.org/record/6515035.

4 Abundance estimation

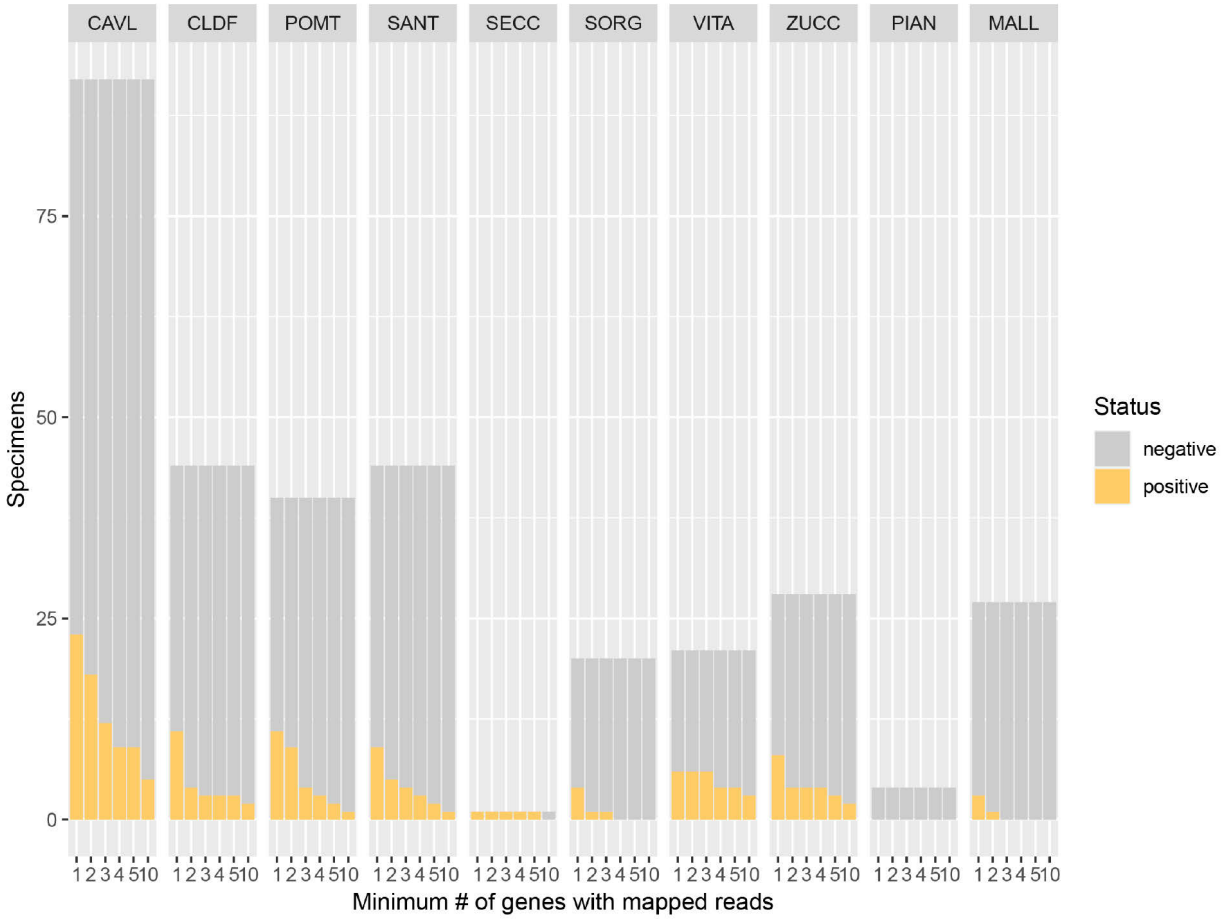


Figure S2: Occurrence of P. abstrusum in *Olavius algarvensis* around Elba (Italy), Pianosa (Italy) and Mallorca (Spain). Amount of *Olavius algarvensis* specimens per bay (Elba: Cavoli, Cala del Fico, Pomonte, Sant'Andrea, Seccheto, Sorgente, Capo Vita, Zuccale; Pianosa, Mallorca) that are positive (yellow) or negative (grey) for P. abstrusum. Specimens are considered positive when their quality filtered metagenomic reads mapped to a a minimum number (1-5 and 10) of non-ribosomal RNA coding sequences of the four main contigs of the P. abstrusum reference metagenome-assembled genome.

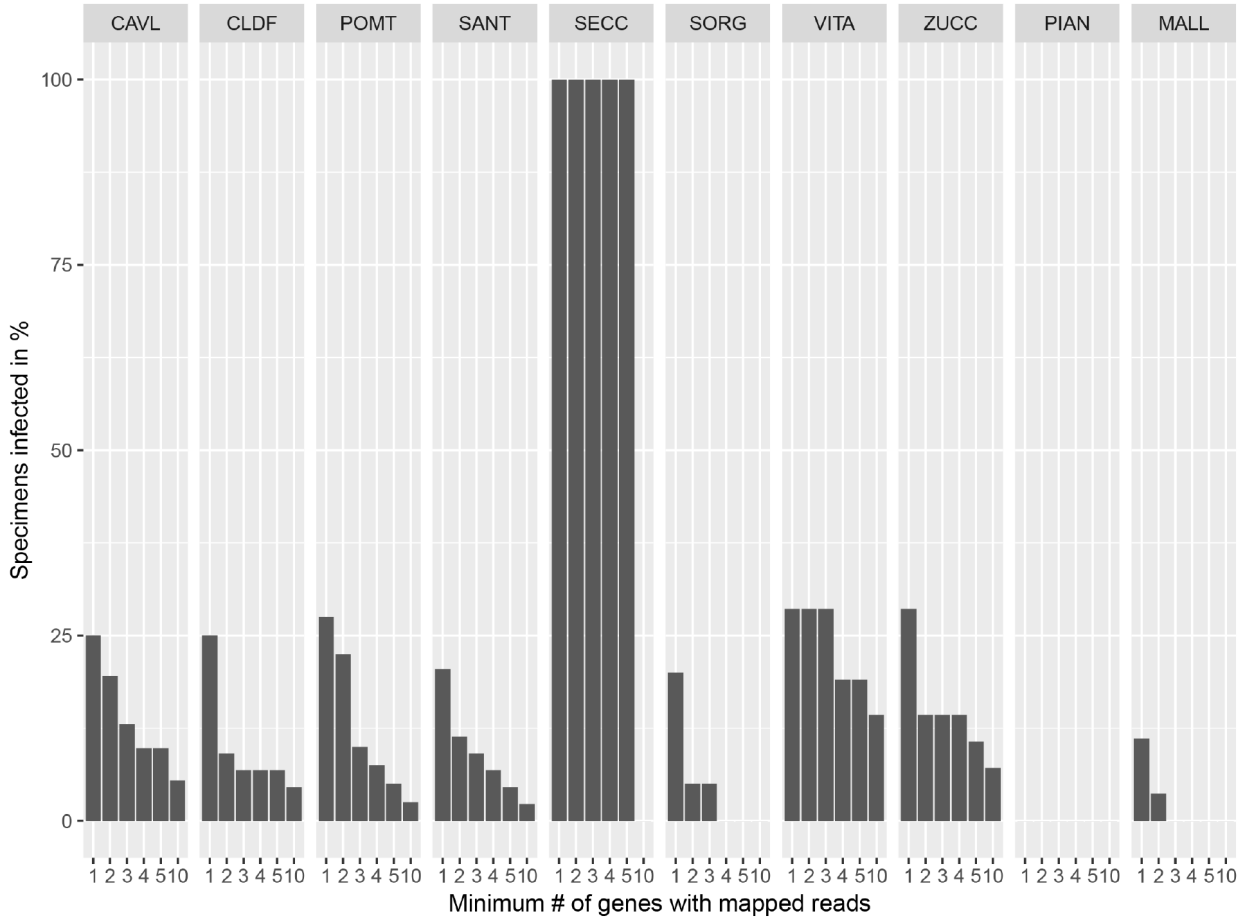


Figure S3: Occurrence of P. abstrusum in *Olavius algarvensis* around Elba (Italy), Pianosa (Italy) and Mallorca (Spain). Percent of *Olavius algarvensis* specimens per bay (Elba: Cavoli, Cala del Fico, Pomonte, Sant'Andrea, Seccheto, Sorgente, Capo Vita, Zuccale; Pianosa, Mallorca) that are positive for P. abstrusum. Specimens are considered positive when their quality filtered metagenomic reads mapped to a a minimum number (1-5 and 10) of non-ribosomal RNA coding sequences of the four main contigs of the P. abstrusum reference metagenome-assembled genome.

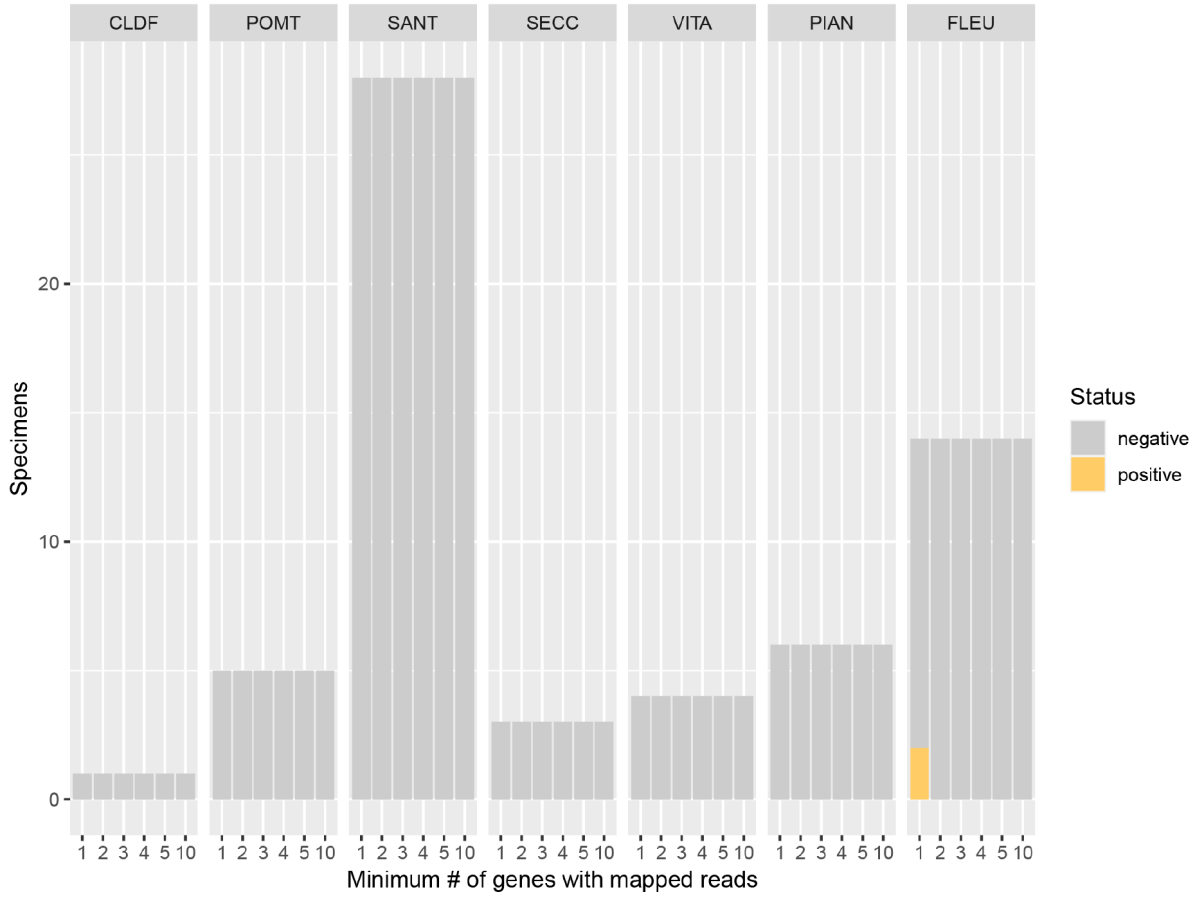


Figure S4: Occurrence of P. abstrusum in *Olavius ilvae* around Elba (Italy), Pianosa (Italy) and Cap Fleuri (France). Amount of *Olavius ilvae* specimens per bay (Elba: Cala del Fico, Pomonte, Sant'Andrea, Seccheto, Capo Vita; Pianosa, Cap Fleuri) that are positive (yellow) or negative (grey) for P. abstrusum. Specimens are considered positive when their quality filtered metagenomic reads mapped to a a minimum number (1-5 and 10) of non-ribosomal RNA coding sequences of the four main contigs of the P. abstrusum reference metagenome-assembled genome.

5 Genomic information

Table S3: Genomic information for P. abstrusum based on checkM.

Feature	Value
Completeness	68.86%
Contamination	0%
Strain heterogeneity	0%
Genome size	641 817 bp
Scaffolds/contigs	60
N50	179 649 bp
Mean contig length	10 696 bp
Longest scaffold	228 812 bp
GC content	34.74%
Coding density	79.66%
Predicted genes	570

6 Annotation information

Table S4: Metagenomic features of P. abstrusum from Prokka and RAST and the combined total evidence annotation.

Feature	Prokka	RAST	Total evidence annotation	Comments
Predicted genes	542	683	697	
Predicted proteins	502	645	658	
Predicted rRNAs	3	2	3	5S, 16S, 23S
Predicted tRNAs	36	36	36	Charging 20 AA
Predicted SSU ribosomal proteins			21	
Predicted LSU ribosomal proteins			30	Missing are non-essential

7 Description of *Candidatus* Pumilisymbium abstrusum

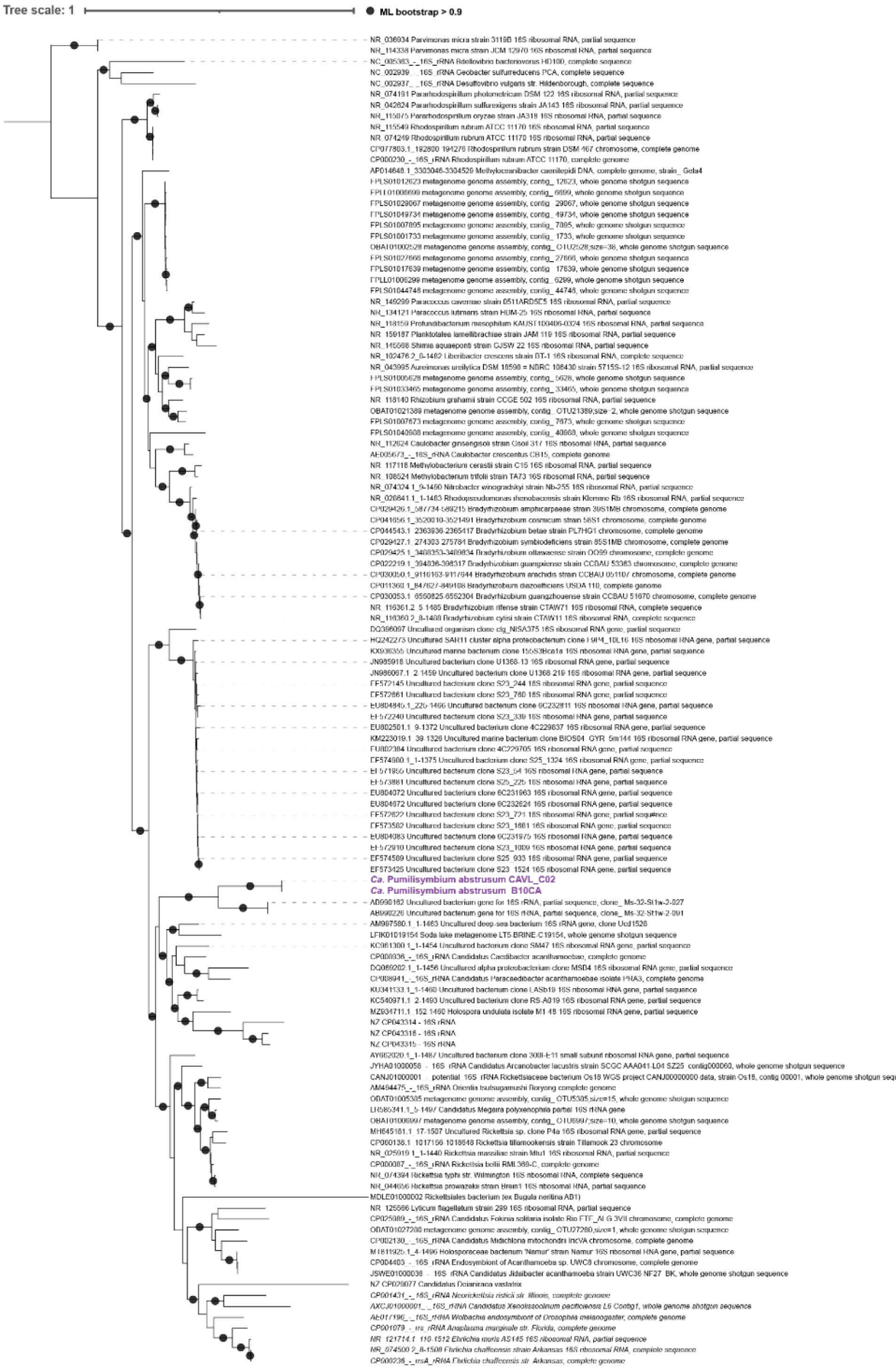
Candidatus Pumilisymbium abstrusum, gen. nov., sp. nov. (Pu.mi.li.sym´bi.um ab.stru´sum; L. masc. adj. *pumilus* dwarfish; Gr. pref. *sym-* together; Gr. masc. n. *bios* life; N.L. neut. n. *Pumilisymbium* a symbiotic dwarfish organism; L. neut. adj. *abstrusum* hidden, secluded). Symbiont of the marine gutless oligochaete *Olavius algarvensis* (Clitellata, Annelidae) from Elba, Italy. Basis of assignment: SSU rRNA gene sequence (Accession number XXX) and positive match with the specific 16S rRNA probes 379F (5'-GGAAACCTTGATCCGGTTATG-3') and 1026R (5'-AACACCTGTGATATGTATAGTG-3'). Forms a novel order taxon with *incertae sedis* status within the order of Rickettsiales in the class of Alphaproteobacteria in the phylum of Proteobacteria. Identified first by Yui Sato in *Olavius algarvensis* individual B10CA from Cavoli Bay on Elba, Italy. Remains uncultivated to date.

8 Phylogenetic placement

Figures are shown in order on the following pages.

Figure S5: Figure 3A with non-collaped clades including Holosporales: Phylogenetic placement of P. abstrusum based on full-length 16S rRNA gene sequences from NCBI (nr/nt, nr/nt type strain material, env_nt) and relevant relatives from GTDB-tk phylogeny. Bootstrap values greater than 90% are shown.

Figure S6: Figure 3A with non-collaped clades excluding Holosporales: Phylogenetic placement of P. abstrusum based on full-length 16S rRNA gene sequences from NCBI (nr/nt, nr/nt type strain material, env_nt) and relevant relatives from GTDB-tk phylogeny. Bootstrap values greater than 90% are shown.


Tree scale: 1
ML bootstrap > 0.9
NR_036934 Parvimonas micra strain 31190 16S ribosomal RNA, partial sequence
NR_114338 Parvimonas micra strain JCM 12970 16S ribosomal RNA, partial sequence
NC_005363_-_16S_rRNA Bdellovibrio bacteriovorus HD100, complete sequence
NC_002939_-_16S_rRNA Geobacter sulfurreducens PCA, complete sequence
NC_002937_-_16S_rRNA Desulfovibrio vulgaris str. Hildenborough, complete sequence
NR_074191 Pararhodospirillum photometricum DSM 122 16S ribosomal RNA, partial sequence
NR_042624 Pararhodospirillum sulfurexigens strain JA143 16S ribosomal RNA, partial sequence
NR_115975 Pararhodospirillum oryzae strain JA318 16S ribosomal RNA, partial sequence
NR_115549 Rhodospirillum rubrum ATCC 11170 16S ribosomal RNA, partial sequence
NR_074249 Rhodospirillum rubrum ATCC 11170 16S ribosomal RNA, partial sequence
CP077803.1_192800 194276 Rhodospirillum rubrum strain DSM 467 chromosome, complete genome
CP000230_-_16S_rRNA Rhodospirillum rubrum ATCC 11170, complete genome
AP014648.1_3303046-3304529 Methylocaniibacter caenitepidi DNA, complete genome, strain: Gela4
FPLS01012623 metagenome genome assembly, contig_12623, whole genome shotgun sequence
FPLL01006699 metagenome genome assembly, contig_6699, whole genome shotgun sequence
FPLS01029067 metagenome genome assembly, contig_29067, whole genome shotgun sequence
FPLS01019734 metagenome genome assembly, contig_19734, whole genome shotgun sequence
FPLS01007895 metagenome genome assembly, contig_7895, whole genome shotgun sequence
FPLS01001733 metagenome genome assembly, contig_1733, whole genome shotgun sequence
OBAT01002528 metagenome genome assembly, contig_OTU2528;size=38, whole genome shotgun sequence
FPLS01027866 metagenome genome assembly, contig_27866, whole genome shotgun sequence
FPLS01017639 metagenome genome assembly, contig_17639, whole genome shotgun sequence
FPLL01006299 metagenome genome assembly, contig_6299, whole genome shotgun sequence
FPLS01044748 metagenome genome assembly, contig_44748, whole genome shotgun sequence
NR_149299 Paracoccus cavernae strain 0511ARD5E5 16S ribosomal RNA, partial sequence
NR_134121 Paracoccus lutimaris strain HUM-25 16S ribosomal RNA, partial sequence
NR_118159 Prolundibacterium mesophilum KAUST100406-0324 16S ribosomal RNA, partial sequence
NR_159187 Planktotalea lamellibrachiae strain JAM 119 16S ribosomal RNA, partial sequence
NR_145568 Shimia aquaeponti strain GJSW 22 16S ribosomal RNA, partial sequence
NR_102476.2_8-1482 Liberibacter crescens strain BT-1 16S ribosomal RNA, complete sequence
NR_043995 Aureimonas ureilytica DSM 18598 = NBRC 108430 strain 5715S-12 16S ribosomal RNA, partial sequence
FPLS01005628 metagenome genome assembly, contig_5628, whole genome shotgun sequence
FPLS01033465 metagenome genome assembly, contig_33465, whole genome shotgun sequence
NR_118140 Rhizobium grahamii strain CCGE 502 16S ribosomal RNA, partial sequence
OBAT01021389 metagenome genome assembly, contig_OTU21389;size=2, whole genome shotgun sequence
FPLS01007873 metagenome genome assembly, contig_7873, whole genome shotgun sequence
FPLS01040908 metagenome genome assembly, contig_40908, whole genome shotgun sequence
NR_112624 Caulobacter ginsengisoli strain Gsoil 317 16S ribosomal RNA, partial sequence
AE005673_-_16S_rRNA Caulobacter crescentus CB15, complete genome
NR_117118 Methylobacterium cerastii strain C15 16S ribosomal RNA, partial sequence
NR_108524 Methylobacterium trifolii strain TA73 16S ribosomal RNA, partial sequence
NR_074324.1_9-1460 Nitrobacter winogradskyi strain Nb-255 16S ribosomal RNA, partial sequence
NR_026841.1_1-1483 Rhodopseudomonas rhenobacensis strain Klemme Rb 16S ribosomal RNA, partial sequence
CP029426.1_5817734-5819215 Bradyrhizobium amphicarpeae strain 39S1MB chromosome, complete genome
CP041656.1_3520010-3521491 Bradyrhizobium cosmicum strain 58S1 chromosome, complete genome
CP044543.1_2363936-2365417 Bradyrhizobium betae strain PL7HG1 chromosome, complete genome
CP029427.1_274303-275784 Bradyrhizobium symbiodeficiens strain 85S1MB chromosome, complete genome
CP029425.1_3408353-3409834 Bradyrhizobium ottawaense strain OO99 chromosome, complete genome
CP022219.1_394036-395517 Bradyrhizobium guangxiense strain CCBAU 53363 chromosome, complete genome
CP030040.1_9116163-9117644 Bradyrhizobium arachidis strain CCBAU 05110 chromosome, complete genome
CP011360.1_847627-849108 Bradyrhizobium diazoefficiens USDA 110, complete genome
CP030053.1_6550825-6552304 Bradyrhizobium guangzhouense strain CCBAU 51670 chromosome, complete genome
NR_116361.2_5-1485 Bradyrhizobium rifense strain CTAW71 16S ribosomal RNA, complete sequence
NR_116360.2_8-1488 Bradyrhizobium cytisi strain CTAW11 16S ribosomal RNA, complete sequence
DQ396097 Uncultured organism clone cfg_NISA375 16S ribosomal RNA gene, partial sequence
HQ242273 Uncultured SAR11 cluster alpha proteobacterium clone FBP4_10L16 16S ribosomal RNA gene, partial sequence
KX936355 Uncultured marine bacterium clone 1b553fbce1a 16S ribosomal RNA gene, partial sequence
JN985918 Uncultured bacterium clone U1368-13 16S ribosomal RNA gene, partial sequence
JN986097.1_2-1459 Uncultured bacterium clone U1368-219 16S ribosomal RNA gene, partial sequence
FR572145 Uncultured bacterium clone S23_244 16S ribosomal RNA gene, partial sequence
FR572661 Uncultured bacterium clone S23_760 16S ribosomal RNA gene, partial sequence
EU804849.1_226-1466 Uncultured bacterium clone 6C232811 16S ribosomal RNA gene, partial sequence
EF572240 Uncultured bacterium clone S23_339 16S ribosomal RNA gene, partial sequence
EU802501.1_9-1372 Uncultured bacterium clone 4C229637 16S ribosomal RNA gene, partial sequence
KM223019.1_39-1326 Uncultured marine bacterium clone BIOS04_GYR_5m144 16S ribosomal RNA gene, partial sequence
EU802384 Uncultured bacterium clone 4C229705 16S ribosomal RNA gene, partial sequence
FR574900.1_1-1375 Uncultured bacterium clone S25_1324 16S ribosomal RNA gene, partial sequence
EF571956 Uncultured bacterium clone S23_54 16S ribosomal RNA gene, partial sequence
EF573881 Uncultured bacterium clone S25_225 16S ribosomal RNA gene, partial sequence
EU804072 Uncultured bacterium clone 6C231963 16S ribosomal RNA gene, partial sequence
EU804672 Uncultured bacterium clone 6C232624 16S ribosomal RNA gene, partial sequence
FR572822 Uncultured bacterium clone S23_771 16S ribosomal RNA gene, partial sequence
FR573582 Uncultured bacterium clone S23_1661 16S ribosomal RNA gene, partial sequence
EU804083 Uncultured bacterium clone 6C231975 16S ribosomal RNA gene, partial sequence
EF572810 Uncultured bacterium clone S23_1009 16S ribosomal RNA gene, partial sequence
EF574589 Uncultured bacterium clone S25_933 16S ribosomal RNA gene, partial sequence
EF573425 Uncultured bacterium clone S23_1524 16S ribosomal RNA gene, partial sequence
Ca. Pumilisymbium abstrusum CAVL_C02
Ca. Pumilisymbium abstrusum B10CA
AB990182 Uncultured bacterium gene for 16S rRNA, partial sequence, clone: Ms-32-St1w-2-027
AB990226 Uncultured bacterium gene for 16S rRNA, partial sequence, clone: Ms-32-St1w-2-091
AM997580.1_1-1463 Uncultured deep-sea bacterium 16S rRNA gene, clone Ucd1526
LFIK01019154 Soda lake metagenome LT5-BRINE.C19154, whole genome shotgun sequence
KC981300.1_1-1454 Uncultured bacterium clone SM47 16S ribosomal RNA gene, partial sequence
CP000936_-_16S_rRNA Candidatus Cardidatus acanthamoebae, complete genome
DQ086202.1_1-1456 Uncultured alpha proteobacterium clone MSD4 16S ribosomal RNA gene, partial sequence
CP000941_-_16S_rRNA Candidatus Paracaedibacter acanthamoebae isolate PRA3, complete genome
KU341133.1_1-1460 Uncultured bacterium clone LASb19 16S ribosomal RNA gene, partial sequence
KC540971.1_2-1493 Uncultured bacterium clone RS-A019 16S ribosomal RNA gene, partial sequence
MZ834711.1_152-1460 Holospora undulata isolate M1 48 16S ribosomal RNA gene, partial sequence
NZ CP043314 - 16S rRNA
NZ CP043318 - 16S rRNA
NZ CP043315 - 16S rRNA
AY662020.1_1-1487 Uncultured bacterium clone 300II-E11 small subunit ribosomal RNA gene, partial sequence
JYHA01000058_-_16S_rRNA Candidatus Arcanobacter lacustris strain SCGC AAA041-L04 SZ25 contig000060, whole genome shotgun sequence
CANJ01000001_-_potential 16S rRNA Rickettsiaceae bacterium Os18 WGS project CANJ00000000 data, strain Os18, contig 00001, whole genome shotgun sequence
AM494475_-_16S_rRNA Orientia tsutsugamushi Boryong complete genome
OBAT01005335 metagenome genome assembly, contig_OTU5335;size=15, whole genome shotgun sequence
LR586341.1_5-1497 Candidatus Megaira polyxenophila partial 16S rRNA gene
OBAT01000997 metagenome genome assembly, contig_OTU0997;size=10, whole genome shotgun sequence
MH645161.1_17-1507 Uncultured Rickettsia sp. clone P4a 16S ribosomal RNA gene, partial sequence
CP000138.1_1017156-1018648 Rickettsia tillamookensis strain Tillamook 23 chromosome
NR_025019.1_1-1440 Rickettsia massiliae strain Mtu1 16S ribosomal RNA, partial sequence
CP000087_-_16S_rRNA Rickettsia bellii RML369-C, complete genome
NR_074394 Rickettsia typhi str. Wilmington 16S ribosomal RNA, complete sequence
NR_044656 Rickettsia prowazekii strain Brein1 16S ribosomal RNA, partial sequence
MDLE01000002 Rickettsiales bacterium (ex Bugula neritina AB1)
NR_125596 Lyticum flagellatum strain 299 16S ribosomal RNA, partial sequence
CP025889_-_16S_rRNA Candidatus Fokinia solitaria isolate Rio FTF_AI G 3VII chromosome, complete genome
OBAT01027280 metagenome genome assembly, contig_OTU27280;size=1, whole genome shotgun sequence
CP002130_-_16S_rRNA Candidatus Midichloria mitochondrii IricVA chromosome, complete genome
M1811925.1_4-1486 Holosporaceae bacterium 'Namur' strain Namur 16S ribosomal RNA gene, partial sequence
CP004403_-_16S rRNA Endosymbiont of Acanthamoeba sp. UWC8 chromosome, complete genome
JSWE01000036_-_16S rRNA Candidatus Jidaibacter acanthamoeba strain UWC36 NF27 BK, whole genome shotgun sequence
NZ CP029077 Candidatus Deianiraea vastatrix
CP001431_-_16S_rRNA Neorickettsia risticii str. Illinois, complete genome
AXCJ01000001_-_16S_rRNA Candidatus Xenolissoclinum pacificiensis L6 Contig1, whole genome shotgun sequence
AE017196_-_16S_rRNA Wolbachia endosymbiont of Drosophila melanogaster, complete genome
CP001079_-_rrs_rRNA Anaplasma marginale str. Florida, complete genome
NR_121714.1_110-1512 Ehrlichia muris AS145 16S ribosomal RNA, partial sequence
NR_074500.2_8-1508 Ehrlichia chaffeensis strain Arkansas 16S ribosomal RNA, complete sequence
CP000236_-_rrsA_rRNA Ehrlichia chaffeensis str. Arkansas, complete genome

Tree scale: 0.1
ML bootstrap > 0.9
NR_036934 Parvimonas micra strain 3119B 16S ribosomal RNA, partial sequence
NR_114338 Parvimonas micra strain JCM 12970 16S ribosomal RNA, partial sequence
NC_005363_-_16S_rRNA Bdellovibrio bacteriovorus HD100, complete sequence
NC_002939_-_16S_rRNA Geobacter sulfurreducens PCA, complete sequence
NC_002937_-_16S_rRNA Desulfovibrio vulgaris str. Hildenborough, complete sequence
NR_074191 Pararhodospirillum photometricum DSM 122 16S ribosomal RNA, partial sequence
NR_042624 Pararhodospirillum sulfurexigens strain JA143 16S ribosomal RNA, partial sequence
NR_115075 Pararhodospirillum oryzae strain JA318 16S ribosomal RNA, partial sequence
NR_115549 Rhodospirillum rubrum ATCC 11170 16S ribosomal RNA, partial sequence
NR_074249 Rhodospirillum rubrum ATCC 11170 16S ribosomal RNA, partial sequence
CP077803.1_192800 194276 Rhodospirillum rubrum strain DSM 467 chromosome, complete genome
CP000230_-_16S_rRNA Rhodospirillum rubrum ATCC 11170, complete genome
FPLS01012823 metagenome genome assembly, contig_ 12823, whole genome shotgun sequence
FPLL01006699 metagenome genome assembly, contig_ 6699, whole genome shotgun sequence
FPLS01029067 metagenome genome assembly, contig_ 29067, whole genome shotgun sequence
FPLS01049734 metagenome genome assembly, contig_ 49734, whole genome shotgun sequence
FPLS01007895 metagenome genome assembly, contig_ 7895, whole genome shotgun sequence
FPLS01001733 metagenome genome assembly, contig_ 1733, whole genome shotgun sequence
OBAT01002528 metagenome genome assembly, contig_ OTU2528,size=38, whole genome shotgun sequence
FPLS01027666 metagenome genome assembly, contig_ 27666, whole genome shotgun sequence
FPLS01017639 metagenome genome assembly, contig_ 17639, whole genome shotgun sequence
FPLL01006299 metagenome genome assembly, contig_ 6299, whole genome shotgun sequence
FPLS01044746 metagenome genome assembly, contig_ 44746, whole genome shotgun sequence
AP014648.1_3303046-3304529 Methyloceanibacter caenitepidi DNA, complete genome, strain_ Gela4
FPLS01040988 metagenome genome assembly, contig_ 40988, whole genome shotgun sequence
NR_149299 Paracoccus cavernae strain 0511ARD5E5 16S ribosomal RNA, partial sequence
NR_134121 Paracoccus lutimaris strain HDM-25 16S ribosomal RNA, partial sequence
NR_118159 Profundibacterium mesophilum KAUST100406-0324 16S ribosomal RNA, partial sequence
NR_159187 Planktotalea lamellibrachiae strain JAM 119 16S ribosomal RNA, partial sequence
NR_145568 Shimia aquaeponti strain CJSW-22 16S ribosomal RNA, partial sequence
NR_102478.2_6-1482 Liberibacter crescens strain BT-1 16S ribosomal RNA, complete sequence
NR_043995 Aureimonas ureilytica DSM 18598 = NBRC 106430 strain 5715S-12 16S ribosomal RNA, partial sequence
FPLS01005628 metagenome genome assembly, contig_ 5628, whole genome shotgun sequence
FPLS01033465 metagenome genome assembly, contig_ 33465, whole genome shotgun sequence
NR_118140 Rhizobium grahamii strain CCGE 502 16S ribosomal RNA, partial sequence
OBAT01021389 metagenome genome assembly, contig_ OTU21389;size=2, whole genome shotgun sequence
FPLS01007673 metagenome genome assembly, contig_ 7673, whole genome shotgun sequence
NR_112624 Caulobacter ginsengisoli strain Gsoil 317 16S ribosomal RNA, partial sequence
AE005673_-_16S_rRNA Caulobacter crescentus CB15, complete genome
NR_117118 Methylobacterium cerastii strain C15 16S ribosomal RNA, partial sequence
NR_108524 Methylobacterium trifolii strain TA73 16S ribosomal RNA, partial sequence
NR_074324.1_9-1490 Nitrobacter winogradskyi strain Nb-255 16S ribosomal RNA, partial sequence
NR_028641.1_1-1483 Rhodopseudomonas rhenobacensis strain Klemme Rb 16S ribosomal RNA, partial sequence
CP029426.1_587734-589215 Bradyrhizobium amphicarpaeae strain 39S1MB chromosome, complete genome
CP041656.1_3520010-3521491 Bradyrhizobium cosmicum strain 58S1 chromosome, complete genome
CP044543.1_2363936-2365417 Bradyrhizobium betae strain PL7HG1 chromosome, complete genome
CP029427.1_274303-275784 Bradyrhizobium symbiodeficiens strain 85S1MB chromosome, complete genome
CP029425.1_3488353-3489834 Bradyrhizobium ottawaense strain OO99 chromosome, complete genome
CP022219.1_394836-396317 Bradyrhizobium guangxiense strain CCBAU 53363 chromosome, complete genome
CP030050.1_9116163-9117644 Bradyrhizobium arachidis strain CCBAU 051107 chromosome, complete genome
CP011360.1_847627-849108 Bradyrhizobium diazoefficiens USDA 110, complete genome
CP030053.1_6550825-6552304 Bradyrhizobium guangzhouense strain CCBAU 51670 chromosome, complete genome
NR_116361.2_5-1485 Bradyrhizobium rifense strain CTAW71 16S ribosomal RNA, complete sequence
NR_116360.2_0-1480 Bradyrhizobium cytisi strain CTAW11 16S ribosomal RNA, complete sequence
DQ396097 Uncultured organism clone ctg_NISA375 16S ribosomal RNA gene, partial sequence
JN986067.1_2-1459 Uncultured bacterium clone U1368-219 16S ribosomal RNA gene, partial sequence
EF572145 Uncultured bacterium clone S23_244 16S ribosomal RNA gene, partial sequence
EF572661 Uncultured bacterium clone S23_760 16S ribosomal RNA gene, partial sequence
EU804845.1_225-1468 Uncultured bacterium clone 6C232611 16S ribosomal RNA gene, partial sequence
EF572240 Uncultured bacterium clone S23_339 16S ribosomal RNA gene, partial sequence
JN985918 Uncultured bacterium clone U1368-13 16S ribosomal RNA gene, partial sequence
KX936355 Uncultured marine bacterium clone 155S3Bca1a 16S ribosomal RNA gene, partial sequence
EU802501.1_9-1372 Uncultured bacterium clone 4C229837 16S ribosomal RNA gene, partial sequence
EU802384 Uncultured bacterium clone 4C229705 16S ribosomal RNA gene, partial sequence
EF574980.1_1-1375 Uncultured bacterium clone S25_1324 16S ribosomal RNA gene, partial sequence
EF571955 Uncultured bacterium clone S23_54 16S ribosomal RNA gene, partial sequence
HQ242273 Uncultured SAR11 cluster alpha proteobacterium clone F9P4_10L16 16S ribosomal RNA gene, partial sequence
EF573881 Uncultured bacterium clone S25_225 16S ribosomal RNA gene, partial sequence
EU804072 Uncultured bacterium clone 6C231963 16S ribosomal RNA gene, partial sequence
EU804672 Uncultured bacterium clone 6C232624 16S ribosomal RNA gene, partial sequence
EF572622 Uncultured bacterium clone S23_721 16S ribosomal RNA gene, partial sequence
EF573582 Uncultured bacterium clone S23_1681 16S ribosomal RNA gene, partial sequence
KM223019.1_39-1326 Uncultured marine bacterium clone BIOS04_GYR_5m144 16S ribosomal RNA gene, partial sequence
EU804083 Uncultured bacterium clone 6C231975 16S ribosomal RNA gene, partial sequence
EF572910 Uncultured bacterium clone S23_1009 16S ribosomal RNA gene, partial sequence
EF574589 Uncultured bacterium clone S25_933 16S ribosomal RNA gene, partial sequence
EF573425 Uncultured bacterium clone S23_1524 16S ribosomal RNA gene, partial sequence
Ca. Pumilisymbium abstrusum CAVL_C02
Ca. Pumilisymbium abstrusum B10CA
AB990162 Uncultured bacterium gene for 16S rRNA, partial sequence, clone_ Ms-32-St1w-2-027
AB990226 Uncultured bacterium gene for 16S rRNA, partial sequence, clone_ Ms-32-St1w-2-091
AY682020.1_1-1487 Uncultured bacterium clone 300I-F11 small subunit ribosomal RNA gene, partial sequence
JYHA01000058_-_16S_rRNA Candidatus Arcanobacter lacustris strain SCGC AAA041 L04 SZ25_contig000060, whole genome shotgun sequence
CANJ01000001_-_potential_16S_rRNA Rickettsiaceae bacterium Os18 WGS project CANJ00000000 data, strain Os18, contig 00001, whole genome shotgun sequence
AM494475_-_16S_rRNA Orientia tsutsugamushi Boryong complete genome
OBAT01005385 metagenome genome assembly, contig_ OTU5385,size=15, whole genome shotgun sequence
LR585341.1_5-1497 Candidatus Megaira polyxenophila partial 16S rRNA gene
OBAT01006997 metagenome genome assembly, contig_ OTU6997;size=10, whole genome shotgun sequence
MH845181.1_17-1507 Uncultured Rickettsia sp. clone P4a 16S ribosomal RNA gene, partial sequence
CP060138.1_1017156-1018648 Rickettsia tillamookensis strain Tillamook 23 chromosome
CP000087_-_16S_rRNA Rickettsia bellii RML369-C, complete genome
NR_025919.1_1-1440 Rickettsia massiliae strain Mtu1 16S ribosomal RNA, partial sequence
NR_074394 Rickettsia typhi str. Wilmington 16S ribosomal RNA, complete sequence
NR_044656 Rickettsia prowazekii strain Brein1 16S ribosomal RNA, partial sequence
NR_125568 Lyticum flagellatum strain 299 16S ribosomal RNA, partial sequence
CP025989_-_16S_rRNA Candidatus Fokinia solitaria isolate Rio ETE_ALG 3VII chromosome, complete genome
OBAT01027280 metagenome genome assembly, contig_ OTU27280;size=1, whole genome shotgun sequence
CP002130_-_16S_rRNA Candidatus Midichloria mitochondrii IricVA chromosome, complete genome
MT811925.1_4-1496 Holosporaceae bacterium 'Namur' strain Namur 16S ribosomal RNA gene, partial sequence
CP004403_-_16S_rRNA Endosymbiont of Acanthamoeba sp. UWC8 chromosome, complete genome
JSWE01000036_-_16S_rRNA Candidatus Jidaibacter acanthamoeba strain UWC36 NF27_BK, whole genome shotgun sequence
CP001431_-_16S_rRNA Neorickettsia risticii str. Illinois, complete genome
AXCJ01000001_-_16S_rRNA Candidatus Xenolissoclinum pacificiensis L6 Contig1, whole genome shotgun sequence
AE017196_-_16S_rRNA Wolbachia endosymbiont of Drosophila melanogaster, complete genome
CP001079_-_rrs_rRNA Anaplasma marginale str. Florida, complete genome
NR_121714.1_110-1512 Ehrlichia muris AS145 16S ribosomal RNA, partial sequence
NR_074500.2_8-1508 Ehrlichia chaffeensis strain Arkansas 16S ribosomal RNA, complete sequence
CP000236_-_rrsA_rRNA Ehrlichia chaffeensis str. Arkansas, complete genome

9 Gene lengths of predicted proteins

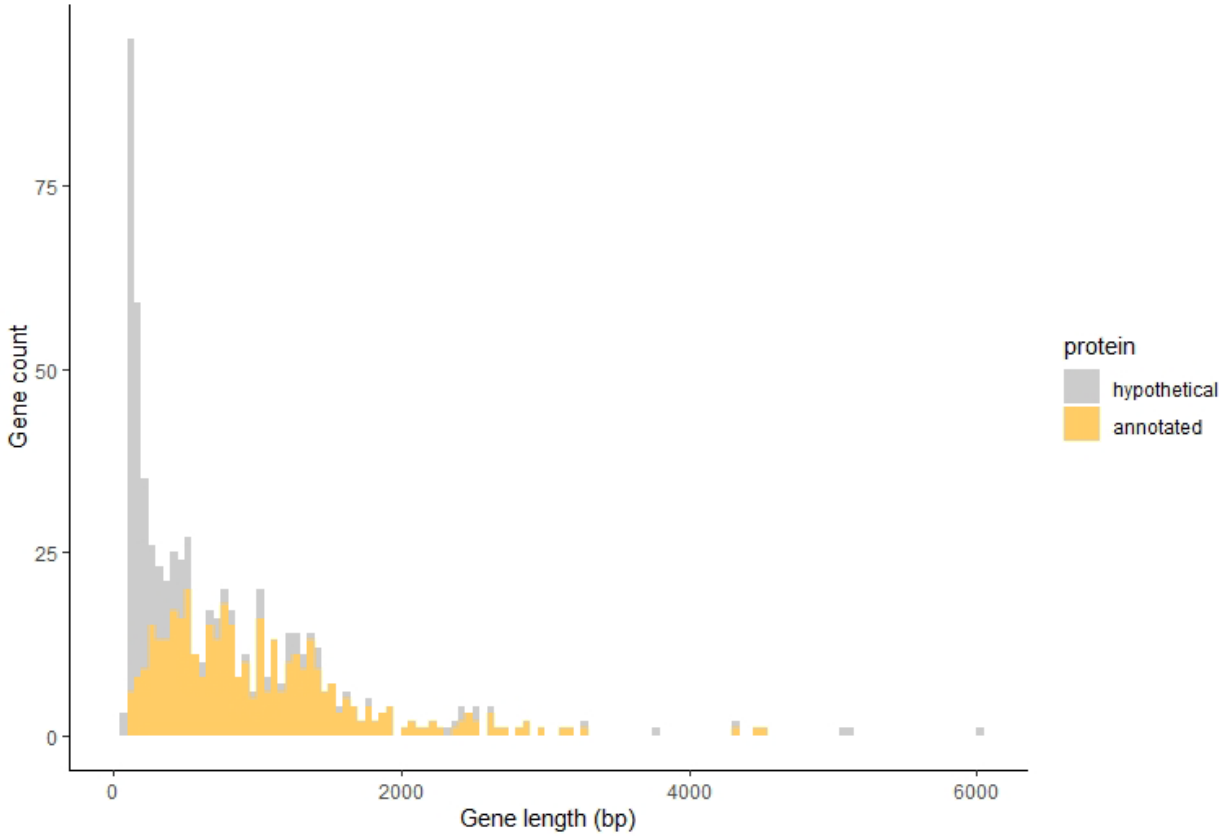


Figure S7: Histogram of the gene lengths of hypothetical (grey) and functionally annotated (yellow) proteins of P. abstrusum with a bin width of 50 bases length. Excluded are tRNA and rRNA genes. The peak in short genes consists mainly of hypothetical proteins. Functionally annotated protein lengths follow an expected bell shape. Few genes with up to 6003 bases are unusually long.

10 Extended information on the metabolic reconstruction of P. abstrusum

Information processing

P. abstrusum is capable of essential information processing with nucleotide *de novo* synthesis and salvage, DNA and RNA metabolism and processing and ribosomal protein biosynthesis. Correspondingly, main parts of the genome represent genes in the COG categories nucleotide metabolism and transport, translation, transcription, recombination and repair, ribosomal structure and biogenesis, replication, and cell wall/membrane/envelope biogenesis (COG categories F: 6.5%, J: 30.1%, L: 9.4%, K: 4.3% and M: 2.3%, Figure S6). P. abstrusum can generate nucleosides and nucleotides and has genes for desoxyribonucleotide salvage. It posses all genes to form ATP or dATP *de novo* from IMP and L-aspartate (via AMP) and GTP or dGTP from GMP and most of the genes for dCTP and dTTP synthesis from CTP and UMP.

P. abstrusum is capable of DNA replication, recombination and repair encoding for DNA polymerase I and III, a variety of helicases, ligases, topoisomerases, gyrases and other DNA modifying genes. DNA can be translated into RNA by a DNA-directed RNA polymerase and RNA molecules can be modified and degraded by a variety of ribonucleases.

As indicated in the genome completeness analysis, a large part of P. abstrusum genes are related to the ribosome and ribosomal translation. All relevant SSU and LSU proteins and rRNAs are present, along with a variety of ribosome related genes including maturation factors, trigger factors, rRNA methyltransferases, and translation initiation, elongation, peptide chain release and ribosome recycling factors. A total number of 36 tRNAs are encoded, capable of carrying all 20 proteinogenic aminoacids. At least 17 tRNA synthases are encoded, missing the tRNA synthase for charging proline and perhaps histamine and glutamine.

Central carbon metabolism and partial TCA cycle

Only three of the enzymes that could be part of the canonical TCA cycle are annotated. These are an NADP-dependent malic enzyme that decarboxylates malate to oxaloactetate or pyruvate under formation of $NADPH+H^+$ and CO_2 or vice versa, a fumarate hydratase class II that converts malate to fumarate or vice versa and an isocitrate/isopropylmalate dehydrogenase-like enzyme that could decarboxylate isocitrate into alpha-ketoglutarate. Malate is present in high relative abundance in the worm[3]. Potentially, malate could be used to form $NADPH+H^+$ and pyruvate. It could furthermore be converted into fumarate, which is needed to form aspartate from which ATP can be built *de novo*. Malate is a dicarboxylic acid and could hence be imported via the encoded dicarboxylate/amino acid:cation symporter.

Cross-membrane transport

P. abstrusum conserved a rather high variety of ion transporting enzymes compared to its relatives. These include a putative Na^+/H^+ antiporter, K^+ uptake proteins, Cd^{2+},Co^{2+},Zn^{2+}/H^+,Na^+ antiporter, Cd^{2+},Co^{2+},Zn^{2+} efflux proteins, a Mn^{2+} transporter, a Mg^{2+} transporter, a Ca^{2+}/Na^+ antiporter and an unspecified mechanosensitive ion channel (annotation information is provided at zenodo.org/record/6515015). Notably, P. abstrusum does not encode for iron transport and iron co-factor dependent proteins. Such a metabolic independence from iron has been shown in other bacteria, largely pathogens, and has been shown to be a means to escape host regulation via iron limitation[42, 43].

A variety of unspecified multidrug efflux transporters from the RND (resistence-nodulation-division), MFS (major facilitator superfamily) and DMT (drug/metabolite transporter) families are encoded as well as a bicyclomycin restistance protein[44-47].

P. abstrusum features a limited set of ABC transporters for specific membrane transport. We detected a complete phosphate ABC transporter. For all other ABC transporters we could only annotate a subset of the genes. One is a phospholipid transporter involved in the maintenance of outer membrane lipid asymmetry[48, 49]. A second incompletely annotated ABC transporter is specific for amines. Furthermore, single genes for methionine transport and unspecific modules of ABC transporters could be detected.

Comparative genomics of P. abstrusum based on COG categories

For details to reference samples see Table S5 and zenodo.org/record/6514058.

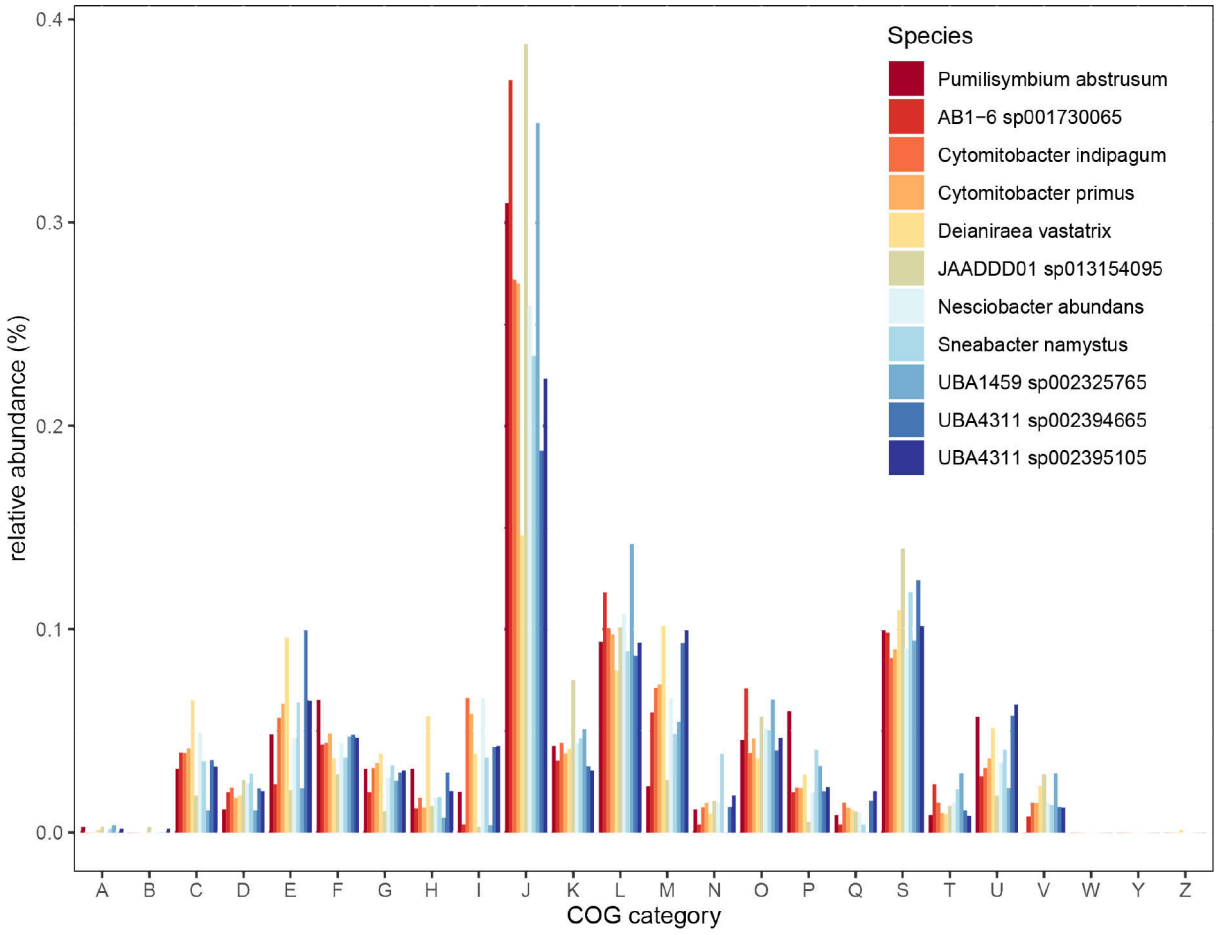


Figure S8: Relative abundance of feature counts of COG categories in the metagenome-assembled genome of P. abstrusum compared to its phylogenomic relatives from GTDB based phylogenetic analysis (Figure 3B).

Compared to its closest relatives, P. abstrusum encodes few genes for cell cycle control, cell division and chromosome partitioning (COG D), lipid transport and metabolism (COG I), cell wall/membrane/envelop biogenesis (COG M), and defense mechanisms (COG V). Instead, its genomic space is allocated for and increased amount of genes for amino acid transport and metabolism (COG E), nucleotide transport and metabolism (COG F), especially inorganic ion metabolism (COG P), and intracellular trafficking, secretion and vesicular transport (COG U).

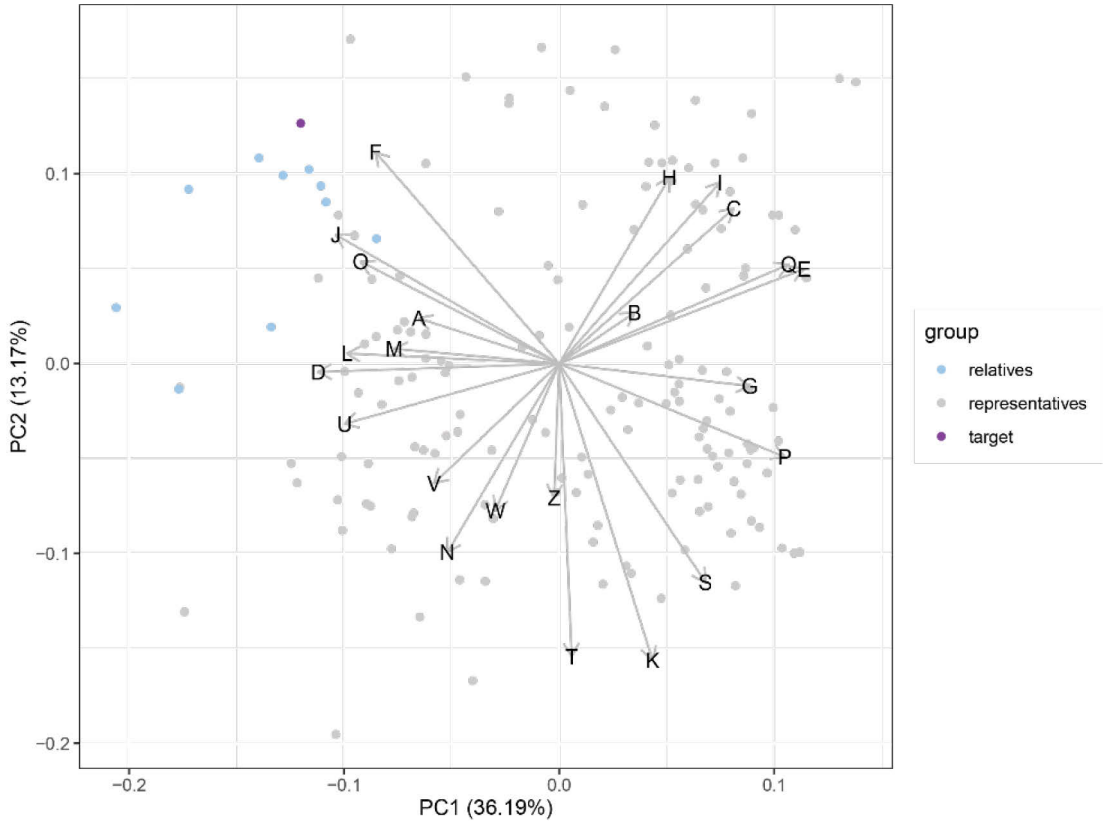


Figure S9: Driving COG categories for the COG profile of P. abstrusum compared to relatives and representatives of alphaproteobacterial families visualized by a principal component analysis (PCA).

Table S5: Relatives from GTDB taxonomy used for comparative genomics.

Accession (GTDBtk)	Accession (NCBI)	MAG (d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;)
GB_GCA_001730065.1	AB1_rickettsiales	;o__Rickettsiales_B;f__AB1-6;g__AB1-6;s__AB1-6 sp001730065
GCA_002325765.1	ASM232576v1	;o__Rickettsiales_B;f__UBA1459;g__UBA1459;s__UBA1459 sp002325765
GB_GCA_002394665.1	ASM239466v1	;o__Rickettsiales_A;f__Deianiraeaceae;g__UBA4311;s__UBA4311 sp002394665
GB_GCA_002395105.1	ASM239510v1	;o__Rickettsiales_A;f__Deianiraeaceae;g__UBA4311;s__UBA4311 sp002395105
GB_GCA_013154095.1	ASM1315409v1	;o__Rickettsiales_B;f__JAADDD01;g__JAADDD01;s__JAADDD01 sp013154095
RS_GCF_007993655.1	ASM799365v1	;o__Rickettsiales_A;f__Deianiraeaceae;g__Deianiraea;s__Deianiraea vastatrix
RS_GCF_008189285.1	ASM818928v1	;o__1604HC;f__1604HC;g__Cytomitobacter;s__Cytomitobacter indipagum
RS_GCF_008189405.1	ASM818940v1	;o__1604HC;f__1604HC;g__Cytomitobacter;s__Cytomitobacter primus
RS_GCF_008189525.1	ASM818952v1	;o__1604HC;f__1604HC;g__Nesciobacter;s__Nesciobacter abundans
RS_GCF_008189685.1	ASM818968v1	;o__Rickettsiales;f__Rickettsiaceae;g__Sneabacter;s__Sneabacter namystus

K-stimulated pyrophosphate-energized proton pump

Figure S10: Protein tree of pyrophosphate-energized proton pumps from P. abstrusum, related protein sequences from BLAST databases (nr_nt, env_nr, refseq_protein and swissprot) and other *Olavius algarvensis* symbionts (Gamma1, Delta1, Delta3, Delta4, Delta13). The tree was rooted with the Archaea/Eukaryota containing branch.

T3SS microsynteny and conserved genes

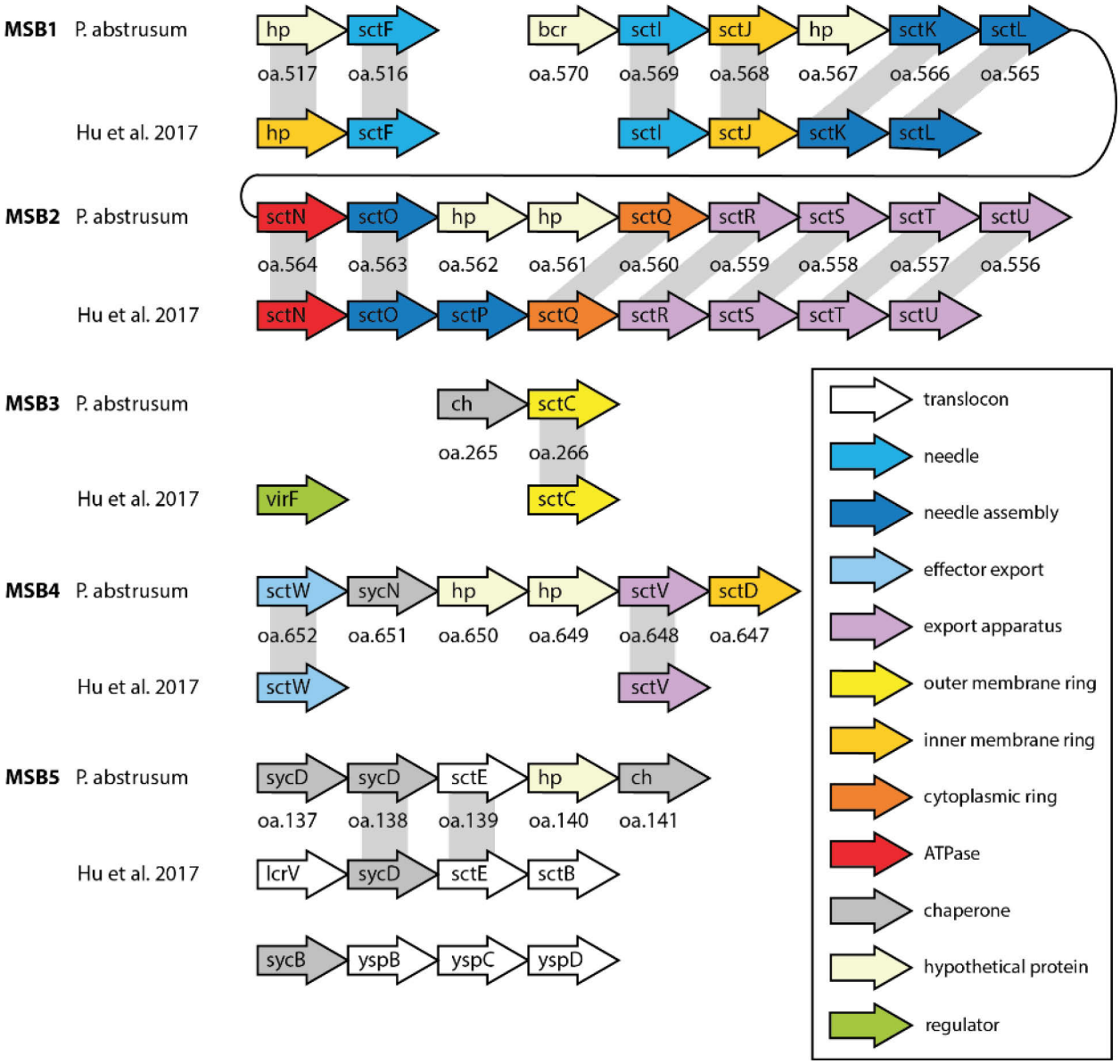


Figure S11: Organisation of T3SS components in the genome of P. abstrusum compared to microsynteny blocks of T3SS components conserved in other bacteria adapted from Hu *et al.* 2017[50]. Arrows represent genes of T3SS components. Colours represent components of the T3SS. Grey bars connect components with conserved clustering in P. abstrusum. Presence of hypothetical genes could hint to yet unannotated functional genes, e.g. oa.561 and oa.562 might be sctP and oa.140 might be sctB.

Figure S12: Unrooted protein tree of T3SS component sctC from P. abstrusum, related protein sequences from BLAST databases (nr_nt, env_nr, refseq_protein and swissprot) and other *Olavius algarvensis* symbionts.

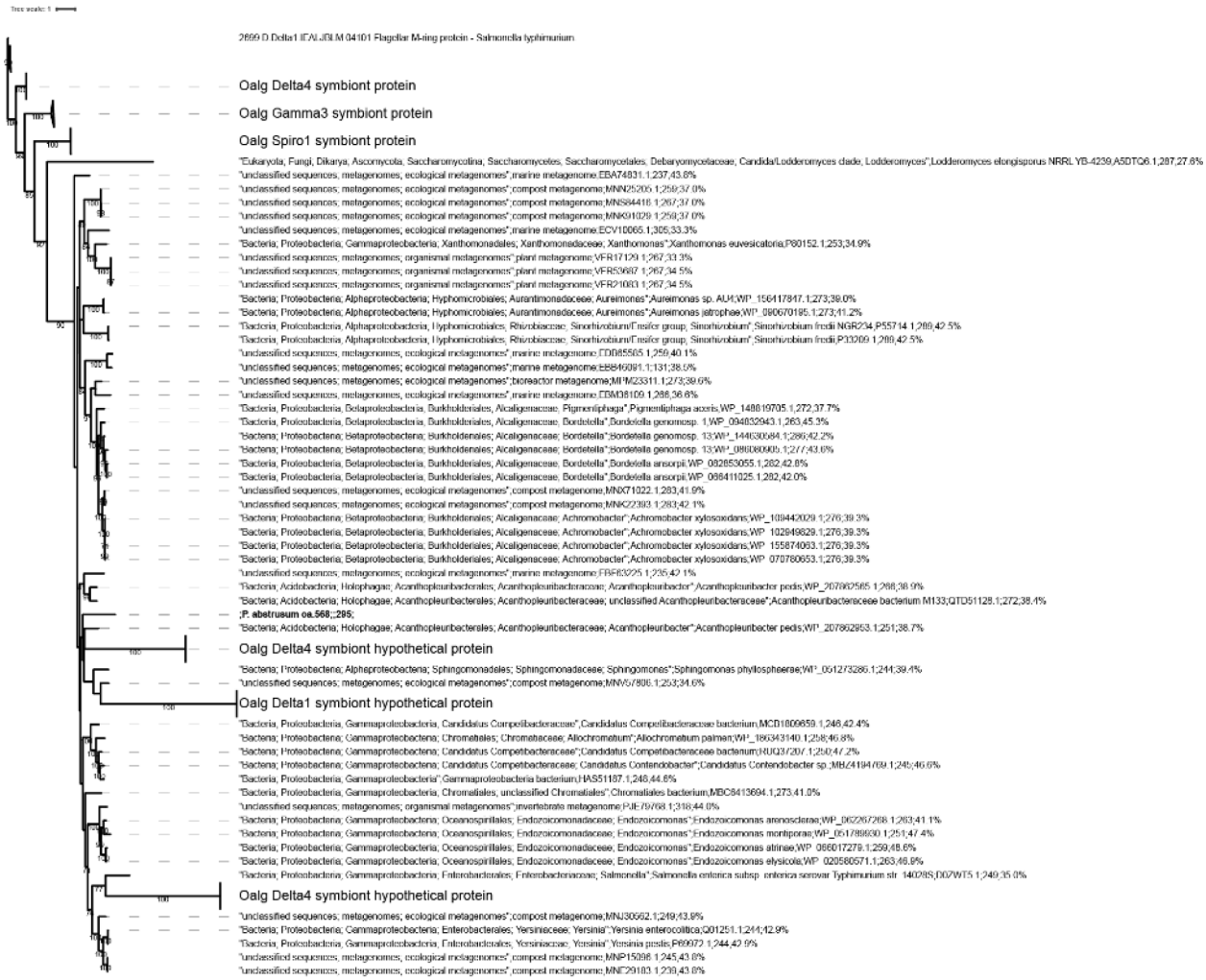


Figure S13: Unrooted protein tree of T3SS component sctJ from P. abstrusum, related protein sequences from BLAST databases (nr_nt, env_nr, refseq_protein and swissprot) and other *Olavius algarvensis* symbionts.

Figure S14: Unrooted protein tree of T3SS component sctN from P. abstrusum, related protein sequences from BLAST databases (nr_nt, env_nr, refseq_protein and swissprot) and other *Olavius algarvensis* symbionts.

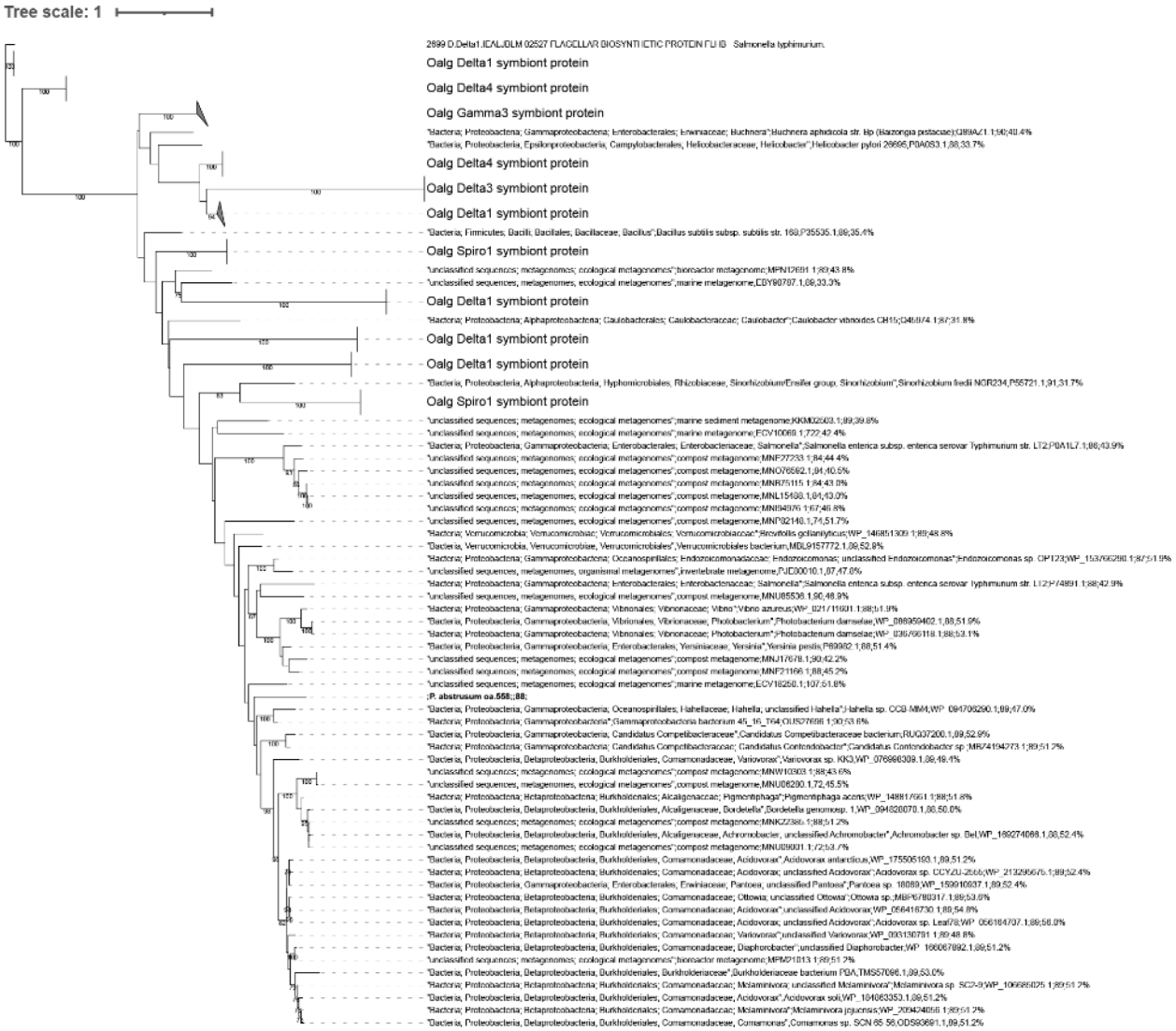


Figure S15: Unrooted protein tree of T3SS component sctS from P. abstrusum, related protein sequences from BLAST databases (nr_nt, env_nr, refseq_protein and swissprot) and other *Olavius algarvensis* symbionts.

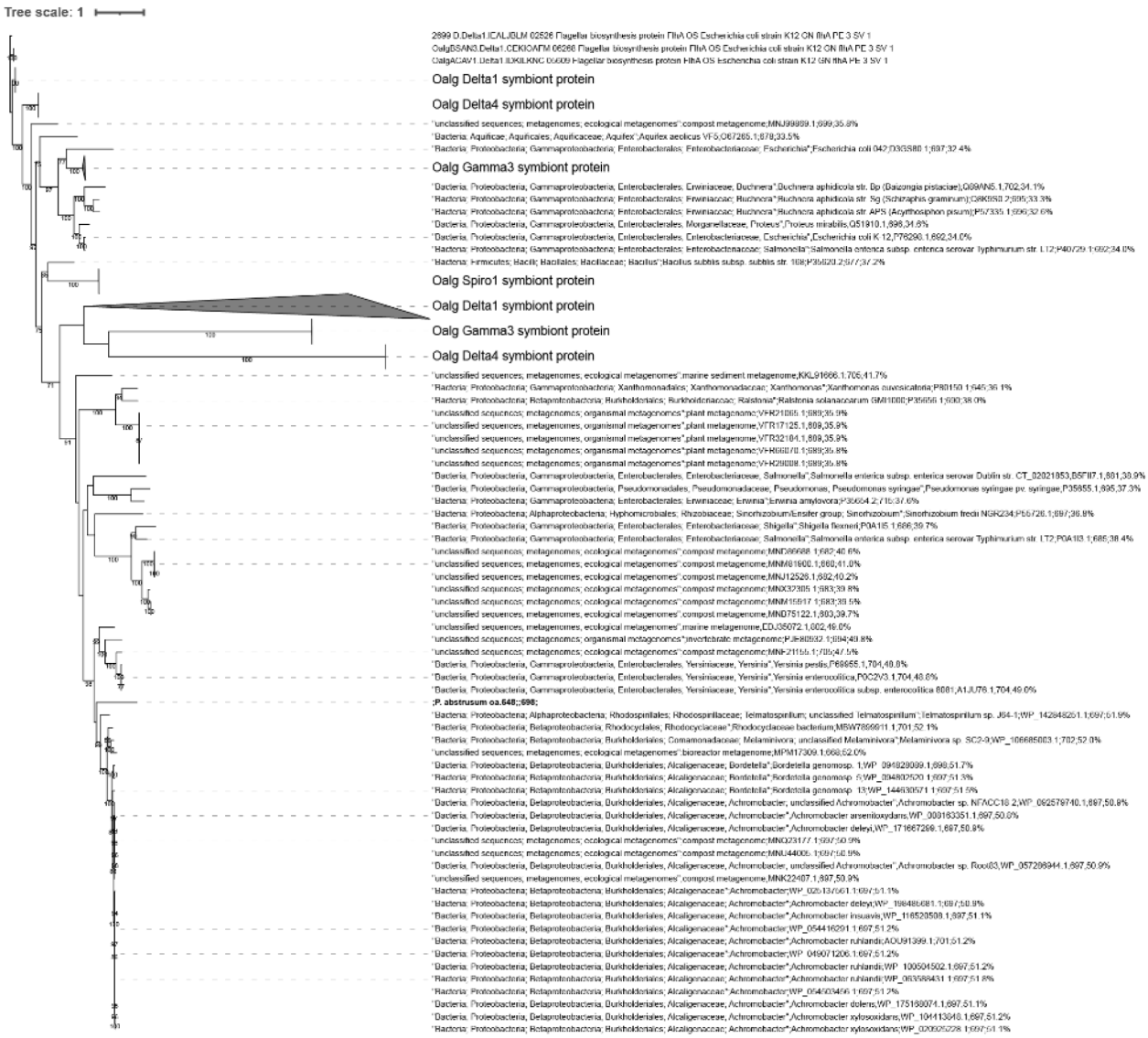


Figure S16: Unrooted protein tree of T3SS component sctV from P. abstrusum, related protein sequences from BLAST databases (nr_nt, env_nr, refseq_protein and swissprot) and other *Olavius algarvensis* symbionts.

Figure S17: Unrooted protein tree of T3SS component sctE from P. abstrusum and related protein sequences from NCBI BLAST databases (nr_nt, env_nr, refseq_protein and swissprot).

11 Current insight on the location inside the host

We used PCR to acquire information of P. abstrusum's location in the worm individuals. P. abstrusum appears to reside in the front half of the worm, as PCR was only positive for the front part when worms were cut in half prior screening. The reproductive organs and the head region are outstanding morphological features in the front half of the animals compared to the rather uniform tail. This could indicate that P. abstrusum either is connected to the reproduction of the worm or is located in the head segments where less of the primary symbionts are present. Further investigation is needed to reveal P. abstrusum's location in the host tissue.

References

1. Abby SS, Néron B, Ménager H, Touchon M, Rocha EP. MacSyFinder: A program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PloS one. 2014;9(10):e110726.
2. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
3. Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. Proceedings of the National Academy of Sciences. 2012;109(19):E1173-E82.
4. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: A software environment for sequence data. Nucleic acids research. 2004;32(4):1363-71.
5. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of *de novo* genome assemblies. Bioinformatics. 2015;31(20):3350-2.
6. Nikolenko SI, Korobeynikov AI, Alekseyev MA, editors. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC genomics; 2013: Springer.
7. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic acids research. 1997;25(17):3389-402.

8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10.
9. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research. 2015;25(7):1043-55.
10. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015;12(1):59-60.
11. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: Circular and linear interactive genome visualization. Bioinformatics. 2009;25(1):119-20.
12. Eichinger V, Nussbaumer T, Platzer A, Jehl M-A, Arnold R, Rattei T. EffectiveDB — Updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. Nucleic acids research. 2016;44(D1):D669-D74.
13. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. 2019;47(D1):D309-D14.
14. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nature biotechnology. 2020;38(9):1079-86.
15. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature biotechnology. 2018;36(10):996-1004.
16. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press; 2020.
17. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015;32(1):268-74.
18. Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature methods. 2017;14(6):587-9.
19. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Scientific reports. 2016;6(1):1-9.
20. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic acids research. 2019;47(W1):W256-W9.
21. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.
22. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: Integrating viruses and cellular organisms. Nucleic acids research. 2021;49(D1):D545-D51.
23. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Science. 2019;28(11):1947-51.
24. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.
25. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-6.
26. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, et al. MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3-11.
27. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
28. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
29. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic acids research. 2014;42(D1):D459-D71.
30. Karp PD, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, et al. Pathway Tools version 19.0 update: Software for pathway/genome informatics and systems biology. Briefings in bioinformatics. 2016;17(5):877-90.
31. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: The protein families database in 2021. Nucleic Acids Research. 2021;49(D1):D412-D9.
32. Gruber-Vodicka HR, Seah BK, Pruesse E. phyloFlash: Rapid small-subunit rRNA profiling and targeted assembly from metagenomes. Msystems. 2020;5(5).

33. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics. 2010;11(1):1-11.
34. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic acids research. 2014;42(D1):D206-D14.
35. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid annotations using subsystems technology. BMC genomics. 2008;9(1):1-15.
36. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific reports. 2015;5(1):1-6.
37. R Core Team. R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
39. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012;19(5):455-77.
40. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, et al. SUPERFAMILY — Sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic acids research. 2009;37:D380-D6.
41. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of molecular biology. 2001;313(4):903-19.
42. Gomes AC, Moreira AC, Mesquita G, Gomes MS. Modulation of iron metabolism in response to infection: Twists for all tastes. Pharmaceuticals. 2018;11(3):84.
43. Posey JE, Gherardini FC. Lack of a role for iron in the Lyme disease pathogen. Science. 2000;288(5471):1651-3.
44. Colclough AL, Alav I, Whittle EE, Pugh HL, Darby EM, Legood SW, et al. RND efflux pumps in Gram-negative bacteria; regulation, structure and role in antibiotic resistance. Future microbiology. 2020;15(2):143-57.
45. Pasqua M, Grossi M, Zennaro A, Fanelli G, Micheli G, Barras F, et al. The varied role of efflux pumps of the MFS family in the interplay of bacteria with animal and plant cells. Microorganisms. 2019;7(9):285.
46. Kumar S, Lekshmi M, Parvathi A, Ojha M, Wenzel N, Varela MF. Functional and structural roles of the major facilitator superfamily bacterial multidrug efflux pumps. Microorganisms. 2020;8(2):266.
47. Zakataeva N, Kutukova E, Gronskiy S, Troshin P, Livshits V, Aleshin V. Export of metabolites by the proteins of the DMT and RhtB families and its possible role in intercellular communication. Microbiology. 2006;75(4):438-48.
48. Malinverni JC, Silhavy TJ. An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane. Proceedings of the National Academy of Sciences. 2009;106(19):8009-14.
49. Powers MJ, Trent MS. Intermembrane transport: Glycerophospholipid homeostasis of the Gram-negative cell envelope. Proceedings of the National Academy of Sciences. 2019;116(35):17147-55.
50. Hu Y, Huang H, Cheng X, Shu X, White AP, Stavrinides J, et al. A global survey of bacterial type III secretion systems and their effectors. Environmental microbiology. 2017;19(10):3879-95.

Discussion, future perspectives and concluding remarks

Symbiosis is one of the most fascinating fields of research as it is ubiquitous and essential for life as we know it. Yet, we have only started to understand the mechanisms of how symbiotic partners recognize, colonize and support each other. Gutless oligochaetes and their symbionts are a valuable system to study such host-microbe interactions. While the symbiont community offers a complexity that allows to study different lifestyles and mechanisms, it is still constrained enough for us to understand the roles of individual members of the symbiosis[1-3]. Many of these members in the gutless oligochaetes are Alphaproteobacteria. Five clades of alphaproteobacterial symbionts in gutless oligochaetes have been previously identified in studies on individual host species. They have been investigated concerning their 16S rRNA gene phylogeny and potentially their location inside the host based on fluorescence *in situ* hybridization[4-6]. However, little research has focused on their genomic potential concerning their metabolism and their roles in the symbiosis[5]. Only recent untargeted metagenomic studies on a large set of globally sampled host specimens revealed their full diversity and abundance[3]. In my thesis, I focused on this neglected bacterial clade of Alphaproteobacteria that makes up the most abundant and most diverse group of secondary symbionts in the gutless oligochaetes. I aimed to understand and describe the variety of Alphaproteobacteria in symbiosis with gutless oligochaetes (Chapter I) and give examples for their mode of interaction with and their metabolism inside the host (Chapter II and III). In total, I described 16 clades of Alphaproteobacteria that associate with gutless oligochaete hosts. Of these, 15 clades are abundant in the host specimens and I expect all of these 15 to be subcuticular, extracellular and mutualistic symbionts based on my findings. However, one Alphaproteobacterium that I could identify is only low abundant in its community as well as its host populations and stands out by its highly reduced genome. I do not expect this bacterium to live in the subcuticular symbiotic layer and despite indications against a parasitic lifestyle, I cannot pinpoint its function in the host to date. Alphaproteobacteria that interact with gutless oligochaetes belong to six distinct orders. Other bacteria in these orders, especially from the Rhodospirillales, Rhizobiales and Rickettsiales, are well known for their ability to intrude eukaryotes and the fact that they can have both beneficial and harmful impacts on their hosts[7-10]. Alphaproteobacteria in other eukaryotes have developed distinct tools to infect host organisms and possess a large set of tools to either support their host's metabolism, e.g. via nitrogen or carbon fixation or vitamin provisioning, or to harm them, e.g. by being energy parasites or causing cytoplasmic

incompatibility[11-14]. In gutless oligochaetes, the detailed investigation of two contrasting examples of alphaproteobacterial symbionts presented in chapters II and III already led to the discovery of specific patterns of abundance and distribution in the hosts, tools for infection and communication, and distinct metabolic interactions. Based on these results, I expect that Alphaproteobacteria in symbiosis with gutless oligochaetes do not serve one single purpose but contribute with a variety of functions, depending on the environment that their specific host species thrives in. Overall, Alphaproteobacteria seem to take on diverse roles and future research has to disclose the distinct metabolic traits that make these Alphaproteobacteria such successful members of the gutless oligochaete symbiosis.

Olavius algarvensis from Elba, Italy, is the best-studied host species and is the only species of which in-depth physiological analyses have been conducted[1, 2, 15]. It does not host an Alphaproteobacterium in its subcuticular symbiont layer, which might be the reason why the role of Alphaproteobacteria in the symbiosis of gutless oligochaetes has not been in focus so far. Interestingly, the co-occurring and numerous host species *Olavius ilvae* does host an Alphaproteobacterium. Despite 16S rRNA gene-based investigations prior to this thesis, it has not been detected. Chapter II details the taxonomy and metabolic potential of this Alphaproteobacterium and presents microscopy data that localizes this symbiont in the subcuticular symbiont layer together with other symbionts. The presence of different gutless oligochaete host species with similar but distinguishable symbiont communities in quite overlapping but not identical habitats pose the question which roles the individual partners play and how these differentially distributed symbionts shape the environmental niches that the worms can thrive in[3, 16]. Efforts to better characterize abundances and distributions of host species and biochemical characteristics of the environment in small-scale sampling campaigns could help unveil diverging abundance patterns and the underlying causes of apparently co-occurring host species.

One factor that likely shapes the distributions of Alphaproteobacteria in gutless oligochaetes is their metabolism and the resulting roles within the symbiont community. The symbionts' metabolism can highly influence the niche that their hosts can live in. In a complex bacterial community such as the gutless oligochaete symbiosis, different bacteria in the different hosts might fulfill similar functional roles. *O. algarvensis* could either live in a slightly different niche than *O. ilvae* where it does not need an Alphaproteobacterium, or it could host another symbiont, supplying the same or similar function as the Alphaproteobacteria in *O. ilvae*. Our

current insights in the metabolism of the diverse alphaproteobacterial symbiont clades suggest that they are mainly heterotrophs, potentially redundant to the Deltaprotebacteria, but some also have the potential to contribute autotrophically to the symbiosis comparable to the Gammaproteobacteria[15]. In my studies, I found indications that Alphaproteobacteria provide access to external carbon sources and help their hosts to scavenge available substrates from their environments. Widely distributed clades like Alpha10, Alpha5 and Alpha7 (Chapter I) likely sustain a heterotrophic lifestyle as they express metabolic traits to provide access to external carbon sources like saccharides or aromatic compounds. In comparison, the role of a heterotroph accessing external carbon sources for *O. algarvensis* might be taken over by other bacteria such as Deltaproteobacteria or the Spirochaetia[1, 2, 17]. Examples for potential autotrophy are Alpha3 and Alpha1 from *I. leukodermatus* (Chapter I). Proteomic expression analysis in this host species revealed that ribulose bisphosphate carboxylases and sulfur oxidation enzymes are among the most highly expressed proteins. However, particularly enzymes involved in sulfur oxidation can catalyze the same chemical reaction in a different direction. To further investigate and consolidate the automated annotations that these bacteria are indeed sulfur oxidizing autotrophs, more detailed sequence-based comparison of enzyme families, e.g. in a protein tree, will be necessary.

A beneficial but not obligatory symbiont

One observation from my research that still puzzles me is the fact that Alphaproteobacteria do not necessarily seem to be obligatory to their hosts. Some species such as *O. algarvensis* are successful fauna in their habitat without being associated with a subcuticular Alphaproteobacterium. Furthermore, host species that regularly associate with Alphaproteobacteria do not necessarily host these in all individuals as in the case of *O. ilvae*. As described, I expect the Alphaproteobacteria to have beneficial contributions to the symbiosis but these do not cause an obligate association - at least on the host side. I can only speculate whether the symbiosis is obligate for the Alphaproteobacteria, but indications for *Candidatus* Saccharisymbium (Chapter II) and also other widely distributed alphaproteobacterial symbionts (Chapter I) point towards a potential for a free-living stage. Flagellar genes are encoded in the genome and also the rest of their genomic potential indicates that they could survive independent of host supplies[18]. This gives rise to a couple of questions: How do Alphaproteobacteria benefit from their association with gutless oligochaetes? What are transmission modes that enable a stable association of gutless oligochaetes with the symbiont

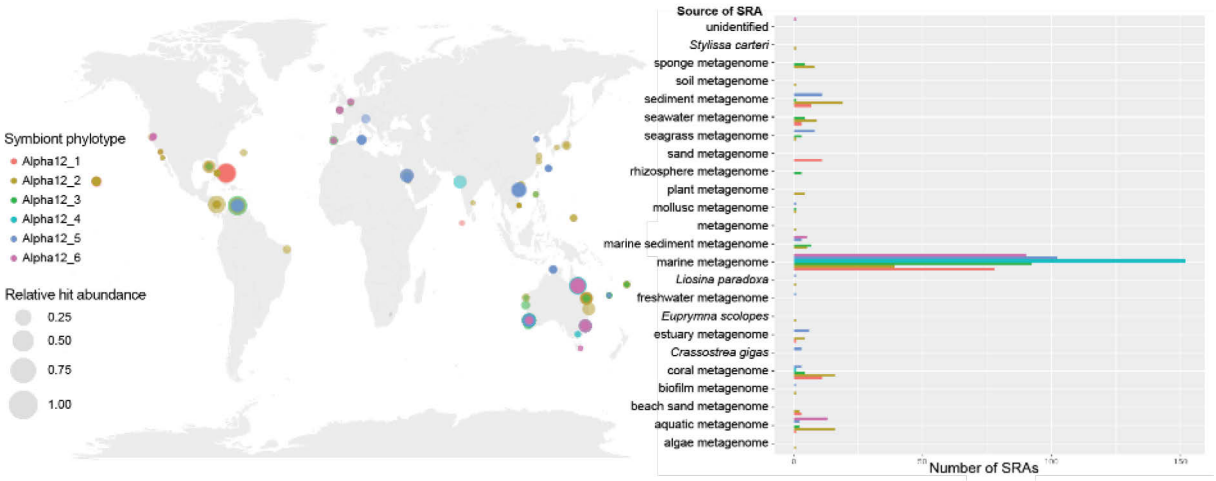


Figure 1: Insight on global distribution, habitats and abundances of gutless oligochaete symbionts can be obtained by screening public sequence read archives for related 16S rRNA gene sequences.
The plots show exemplary mapping of the global distribution (left) and relative hit abundances of sequences (right) from sequence read archives (SRAs) with 99% sequence similarity to six Alpha12 phylotypes. Raw data was obtained with the IMNGS platform and analyzed using custom R-scripts[19]. Preliminary results indicate that species within alphaproteobacterial symbiont clades can have distinct distribution patterns and were mainly but not exclusively sampled in marine sediments.

community? Can symbionts be taken up from the environment or from other host specimen during fertilization? Or is imperfect vertical transmission the cause why we see patchy community patterns?

Mankowski *et al.* and Sato *et al.* have started to find answers to these questions concerning symbiont transmission, evolutionary history and the resulting distribution patterns[3, 20] They already learned that one mechanism alone cannot explain the diverse gutless oligochaete symbioses. Instead, each individual symbiont has its own level of fidelity and specific modes for host colonization and interaction[3, 20]. Further population studies on individual host-symbiont systems would be necessary to answer the question how symbiont transmission shapes the individual associations.

To address questions concerning free-living stages of the symbionts, sequence read archives provide a searchable knowledge base for further investigation. As done for *Candidatus* Saccharisymbium (Chapter II), it is possible to screen public databases comprising global sequencing data sets such as the TARA ocean data for closely related 16S rRNA gene sequences. By plotting the resulting sampling locations, sample sources such as marine sediment or alpine spring water, or related read hit abundances in the individual sequence read

archives, we can obtain insights in the global distribution and potential specificity of symbionts to gutless oligochaetes on a global scale (Figure 1).

One approach to investigate the dependency of the alphaproteobacterial symbionts on the host could be cultivation experiments. Cultivating the individual partners of intimate animal-microbe symbioses is a generally challenging task. It can be even more difficult if the partners come from chemosynthetic symbioses, which is why this has not been achieved so far. However, with knowledge of the symbionts' metabolism based on genomic and expression data and potentially host-independent symbionts such as Saccharisymbium, we might be in the best state to go back to the laboratory and try to cultivate the secondary symbionts. Such "meta-omics" informed enrichment or cultivation attempts could be promising first steps to yield independent cultures of secondary symbionts. As gutless oligochaete symbionts share many metabolic functions, carefully selected cultivation conditions specific to the target symbionts would be crucial in these attempts. A good starting point for *Candidatus* Saccharisymbium could be an experimental setup with anaerobic to microaerobic conditions, carbon sources like saccharides and aromatic compounds, as available from seagrass meadows, and a nitrous oxide containing atmosphere to serve as terminal electron acceptor. Other genera such as Alpha5 and Alpha1 could be targeted with culture conditions selecting for their potential capabilities in lipid metabolism.

Easy to sample, challenging to investigate – technical limitations in the research of gutless oligochaete symbioses

A great advantage of studying gutless oligochaetes is that they are rather easily accessible. One bucket of sediment from the knee-deep waters in close vicinity to the shoreline can be enough to retrieve hundreds of gutless oligochaete worms. The chances to find sites with high abundances of gutless oligochaetes can be increased with simple to gain knowledge of the sampling site such as granulometry, amounts and types of organic input or depth of oxygen penetration and e.g. then be used by diving teams to screen seagrass meadow-associated sediments. This has enabled researchers, and still does, to investigate the multipartite and intricate symbiosis of gutless oligochaetes, their speciation patterns, hosted symbionts, and population genetics, among others. When new knowledge is gained, also new questions arise. The beauty of the gutless oligochaete symbiosis is that we can go out sampling without fear to exhaust specimens at a certain site as long as we collect reasonable amounts of worms for specific follow-up experiments. This is not only true for our common sampling sites around the

island of Elba in the Mediterranean Sea and in the Caribbean, where host diversity is even higher, but also at less frequently sampled sites around the globe. It will not be a legal or ecological problem to obtain enough material to further investigate this symbiotic system, especially the abundant alphaproteobacterial symbiont clades that have been found at most sampling sites investigated. Since gutless oligochaete host species seem rather location-specific, I expect to find novel host species when we go out and explore new sampling sites. Major symbiont clades have stayed rather consistent on a global scale but I expect to find even more symbiont lineages the more host species we explore.

Although samples are easily accessible, research on gutless oligochaetes comes with challenges. Given the small host size, investigation of the symbiont community is happening on the edge of several of the key technologies that drive molecular ecology, despite recent advances in the fields of sequencing and imaging technology. Decreased costs for large-scale sequencing have enabled us to see the variety of gutless oligochaete symbionts and to study the metabolic potential of the association, for both hosts and symbionts. However, expression data based on the transcriptome and proteome are still limiting and can hardly provide insight into the complete expressed metabolism. Researchers still need to pool individuals or obtain enriched symbiont fractions to get a handle on expression data for the low abundant symbionts. Reasons for this are the technological limitations such as a lack of specificity for bacteria rRNA for which e.g. novel rRNA removal tools suggest progress in the near future. With the small size, the nature of our study objects aggravates such mixture based problems as overall biomass of the worms is low, and host tissue and the abundant gammaproteobacterial symbiont outnumber the diverse secondary symbionts by far (Figure 2). Preliminary results of both transcriptome and proteome expression of a selected set of host species showed that secondary symbionts together make up less than 5% of the biomass of single worm individuals whereas the main Gammaproteobacterium can make up 30-50%. 2D LC-MS approaches revealed the most highly expressed metabolic pathways of *Candidatus* Saccharisymbium (Chapter II) but could only cover expression data for about 10% of the total genomic capabilities. Available proteome expression data, even from symbiont fractions of pooled worm individuals, was not sufficient to calculate characteristic stable isotope fingerprints for the alphaproteobacterial symbionts which would have helped to identify whether sources of carbon come from within the host system or from external sources[2]. In the case of *Candidatus* Pumilisymbium, no expression-based data could be obtained due to the even lower abundance of the symbiont. Prescreening of specimens to verify symbiont presence and screening and targeted extraction

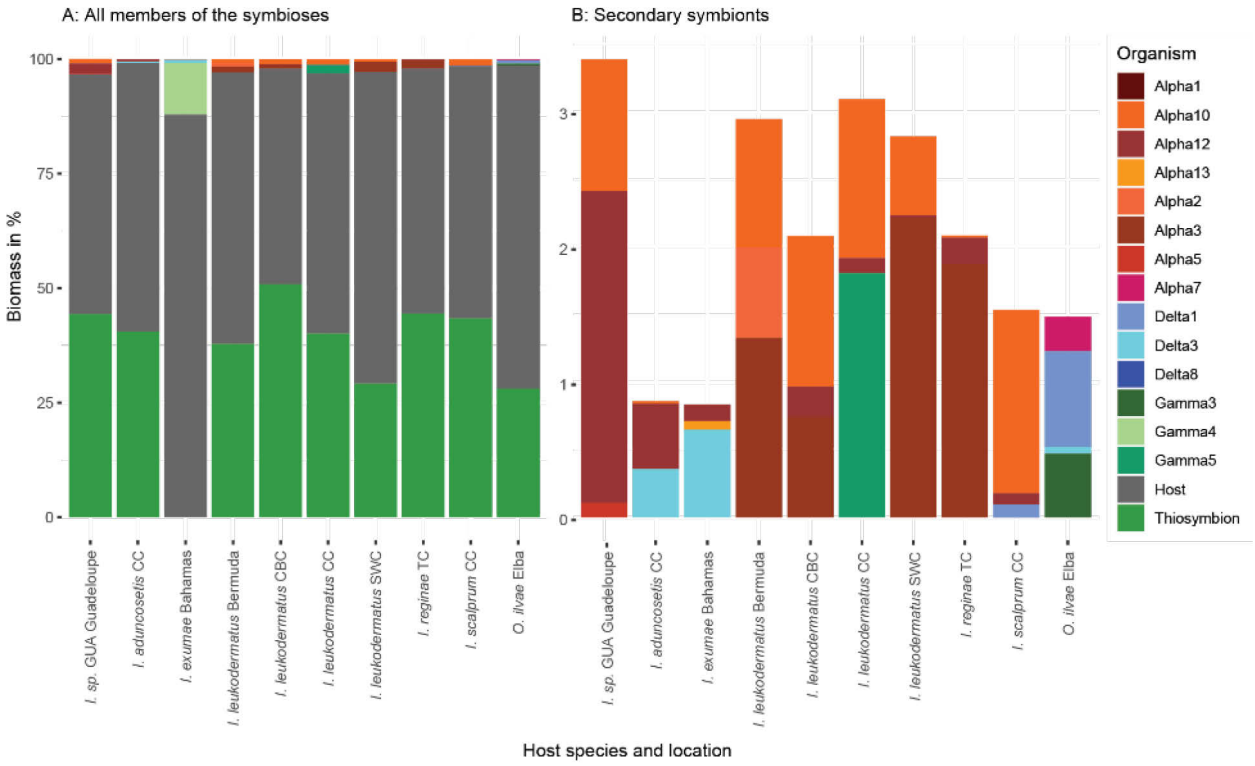


Figure 2: Secondary symbionts make up only 1-3% of the biomass of their gutless oligochaete hosts.
Preliminary calculations of biomasses based on proteome expression of a selected set of gutless oligochaete hosts show that the main Gammaproteobacterium can make up 30-50% of the whole community whereas secondary symbionts make up only 1-3% of the biomass (A: all symbionts; B: focus on the secondary symbionts). Sampling sites in Belize: CC: Curlew Cay, CBC: Carrie Bow Cay, SWC: South Water Caye, TC: Twin Cays.

of body parts with a high probability to be colonized by the symbionts could help to overcome this gap, e.g. by using 16S rRNA or functional gene based assays in quantitative polymerase chain reaction (qPCR).

The low abundance of the secondary symbionts does not only affect the feasibility of analyses but also poses biological questions on the overall function of the gutless oligochaete symbiosis. How can the secondary symbionts fulfill their crucial metabolic functions in such low abundances? One obvious assumption would be that the worms spend more time in the sediment layers favored by the secondary symbionts so that their metabolic rates suffice to cope with the resources and needs imposed by the host and the Gammaproteobacterium. Research based on further physiological knowledge and guided by metabolic modelling will be necessary to answer these types of questions.

Besides the limitations due to low input biomass for our analyses, we have been and are still partially blind sighted on the imaging level of investigation of the secondary gutless oligochaete symbionts. Bacteria do not reveal their taxonomy based on their morphology but necessitate molecular tools for their identification. This has hampered the identification of the diverse symbiont community of gutless oligochaetes in the first place[21]. It also let us miss novel types of symbionts over and over again based on sole microscopic and also too targeted investigation, e.g. the identification of Alpha7 in *O. ilvae*[16]. Furthermore, we have not used the full potential of imaging technologies to study the breadth of gutless oligochaete symbionts. One powerful development is coupling transmission electron microscopy (TEM) to FISH, which could help to obtain better high-resolution information on the local patterns of host-symbiont interactions. Another promising approach is the use of spatially resolved mass spectrometry imaging to gain insight on symbiont-specific metabolites and with that pinpoint metabolic relevant pathways of the symbionts[22].

Directions for future research

During my studies, I investigated and described the diverse Alphaproteobacteria that are associated with gutless oligochaetes. Besides generating new knowledge about this neglected symbiont clade, I also set the basis for further research. Based on my findings, I recommend studying at least one of the more abundant Alphaproteobacteria, which are Alpha3, Alpha10 and Alpha12. This will help to understand main functions of the Alphaproteobacteria in the gutless oligochaete symbiosis, potentially their mode of host colonization and contribution to and benefit from the symbiosis. Studying the abundant Alphaproteobacteria will help to explain how Alphaproteobacteria can be so widespread symbionts of the gutless oligochaete symbiosis. If I had to pick one clade to further investigate, I would focus future research on Alpha3 for several reasons. This bacterium is one of the most widely distributed, has the most diverse subspecies, is also present in one of the closely related gutbearing oligochaetes and might play a special role as a sulfur-oxidizing autotroph[3].

Next to the general investigation of the diversity of Alphaproteobacteria and their role in the gutless oligochaete symbiosis, there are also more specific research questions that came up during my research and are worthwhile to follow-up on.

One interesting research area is the bacterial communication with their host. Bacteria have diverse modes to communicate with eukaryote hosts, colonize them or evade detection.

Examples are lipopolysaccharide coatings, which might be used by *Candidatus* Saccharisymbium to interact with its host *O. ilvae*, or secretion systems as observed in *Candidatus* Pumilisymbium in *O. algarvensis* that has a T3SS. Gutless oligochaetes are a suitable system to study different modes of interaction between bacteria and their hosts as we can visualize and investigate expression data of distinct members that interact with the same host cells. The bacterial clades might employ distinct systems for the same purposes of communication and colonization. Other options that we have not yet detected in the gutless oligochaete symbiosis but that are present in other shallow-water or deep-sea chemosynthetic symbioses are outer membrane vesicles or the use of toxins[23, 24]. Utilization of proteins containing eukaryote-like domains at their membrane surface would be another option for the symbionts to communicate with their hosts. This is an interesting point to follow-up as I observed such proteins to be expressed in *Candidatus* Saccharisymbium species. However, they are often difficult to characterize and would need quite some time and technological investement. Other interaction mechanisms of interest would be the presence of antibiotic and antibiotic resistance genes in both the hosts and the symbionts. To date, we have only little insight in the control of population numbers, how the hosts can distinguish between different colonizing bacteria and how the symbionts communicate with co-localized bacteria in the same constrained space. To tackle these questions, imaging-based analyses could improve our understanding of these modes of communication. By imaging molecules such as lipopolysaccharides or potential outer membrane vesicles with targeted visualization approaches, we could learn about their mode of action and better understand how symbionts locate on a small-scale level compared to the other subcuticular symbionts in the same system.

Another exciting field of research would be genome reduction. In symbiosis, genome reduction can ultimately lead to a stage where the symbionts are no longer independent entities but rather form organelle-like structures that might not even be cell shaped compartments anymore[25]. There are diverse examples form terrestrial habitats such as intensively studied insect symbionts with highly reduced genomes[13, 26]. However, there is little knowledge on marine bacteria with a strongly reduced genome. The discovery of the highly reduced Alphaprotebacterium in the gutless oligochaete *O. algarvensis* also brings up questions about different evolutionary trajectories within the symbiont community. In this case, the questions is whether the *Candidatus* Pumilisymbium has been part of the symbiosis alongside the other subcuticular symbionts for a long time but has gone through completely different processes of development and genome evolution. Or has it been around for much a longer than the other symbionts and

has had a distinct role in its gutless oligochaete host leading to such a small genome over time? My results show that it would not be surprising to find more of these reduced genomes in future studies as it cannot be serendipity alone that we screened metagenomes from only one host species – *Olavius algarvensis* – and found such a reduced symbiont. I expect to find more of these if we start to systematically screen more host species as intensively as *O. algarvensis*. Discoveries of more reduced symbiont genomes in the gutless oligochaete symbiosis would then enable a more detailed analysis of why genome reduction has happened in these cases. Previous investigations have already detected rickettsial 16S rRNA genes from the host species *Inanidrilus leukodermatus*[27]. However, we have so far not been able to obtain a metagenome-assembled genome for this bacterium. I expect that not only screening of other gutless oligochaete host species but also annelids or other marine animals in general, would bring more strongly reduced genomes to light. Future long-read sequencing approaches will yield more genomic information on low abundant symbionts and genome bins of small size that often lead to artificially low completeness scores should be studied with care and not be neglected. In this regard, it will be important to reconsider quality thresholds for metagenome-assembled genomes and classify e.g. small sets of contigs that can be circularized in the assembly graph, but that get low completeness scores as potentially reduced genomes. Other markers such as presence and number of rRNAs and tRNAs, and ribosomal genes should then be systematically consulted to detect bins for bacteria with reduced genomes.

Another promising research area arises from the detection of phage-related genes already in the smallest of the alphaproteobacterial symbionts. Similarly, COG category analyses of all alphaproteobacterial symbionts pointed to the presence of mobilome related genes such as prophages and transposons. To date, we do not know the roles that these traits might play in the gutless oligochaete symbiosis but we cannot exclude that phages or transposons might play a regulatory or secondary functional role. Cases, where these “matrjoschka” -like mechanisms occur in symbiosis have been for example shown in *Wolbachia* symbionts or *Hamiltonella defensa* from pea aphids[28, 29]. There, these mechanisms can have a crucial role for the interaction and functioning of the partners. No investigations in this direction have been conducted on the gutless oligochaete symbiosis so far. However, with the vast metagenomic data at hand, one could start to investigate small-scale intricacies like the role of (bacterio-)phages or species and strain diversity patterns within single host species and even specimens.

Concluding remarks

This work is a piece of curiosity driven and descriptive science. I was astonished by the diversity of Alphaproteobacteria that the investigation of a large metagenomic dataset of 64 gutless oligochaete host species revealed[3]. With this thesis, I aimed to improve our knowledge on a previously almost neglected aspect of the gutless oligochaete symbiosis. I could show that the 16 alphaproteobacterial symbionts that we know to date belong to six distinct orders of Alphaproteobacteria. The two symbiont genera described here in detail illustrate how different association patterns and metabolic capabilities of the symbionts can be. The proposed *Candidatus* Saccharisymbium (Chapter II) is a Rhodospirillales related bacterium with a genome size of ~4.4 Mbp capable of a variety of heterotrophic metabolic pathways potentially providing access to carbon from surrounding seagrass meadows. It resides in the subcuticular symbiont layer with other members of the symbiotic consortia in a range of globally distributed host species. In contrast, the proposed *Candidatus* Pumilisymbium (Chapter III) has a strongly reduced genome size of only 0.64 Mpb and is highly dependent on nutrient provisioning by its host, and appears to be endemic in the Mediterranean species *Olavius algarvensis*. Based on first insights into the metabolism of other alphaproteobacterial symbiont clades, I expect them to bring even more diverse functions to their gutless oligochaete host (Chapter I). Indicated by the different distribution patterns of alphaproteobacterial symbiont clades that can be widely distributed, such as Alpha3, Alpha12 and Alpha10, and symbiont clades that seem location-specific, such as Alpha2 or Alpha16, I also expect a range from general metabolic functions that are beneficial to many host species to very specific ones that only provide benefit in specific environmental settings. With the large diversity of Alphaproteobacteria in symbiosis with gutless oligochaetes, this thesis can only be the starting phase in our quest to understand their role and function in the gutless oligochaete symbiosis. Future research will take much more than a dedicated PhD to dive into the metabolism, or into mechanisms for host colonization and communication and will very likely discover even more fascinating properties of the diverse and abundant Alphaproteobacteria in symbiosis with gutless oligochaetes.

References

1. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature. 2006;443(7114):950-5.
2. Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. Proceedings of the National Academy of Sciences. 2012;109(19):E1173-E82.
3. Mankowski A, Kleiner M, Erséus C, Leisch N, Sato Y, Volland J-M, et al. Highly variable fidelity drives symbiont community composition in an obligate symbiosis. bioRxiv. 2021.
4. Dubilier N, Amann R, Erséus C, Muyzer G, Park S, Giere O, et al. Phylogenetic diversity of bacterial endosymbionts in the gutless marine oligochete *Olavius loisae* (Annelida). Marine Ecology Progress Series. 1999;178:271-80.
5. 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. Applied and Environmental Microbiology. 2006;72(8):5527-36.
6. Bergin C, Wentrup C, Brewig N, Blazejak A, Erséus C, Giere O, et al. Acquisition of a novel sulfur-oxidizing symbiont in the gutless marine worm *Inanidrilus exumae*. Applied and environmental microbiology. 2018;84(7):e02267-17.
7. Gruber-Vodicka HR, Dirks U, Leisch N, Baranyi C, Stoecker K, Bulgheresi S, et al. *Paracatenula*, an ancient symbiosis between thiotrophic Alphaproteobacteria and catenulid flatworms. Proceedings of the National Academy of Sciences. 2011;108(29):12078-83.
8. Poole P, Ramachandran V, Terpolilli J. Rhizobia: From saprophytes to endosymbionts. Nature Reviews Microbiology. 2018;16(5):291-303.
9. Driscoll TP, Verhoeve VI, Guillotte ML, Lehman SS, Rennoll SA, Beier-Sexton M, et al. Wholly Rickettsia! Reconstructed metabolic profile of the quintessential bacterial parasite of eukaryotic cells. MBio. 2017;8(5).
10. Giannotti D, Boscaro V, Husnik F, Vannini C, Keeling PJ. The “other” Rickettsiales: An overview of the family “*Candidatus* Midichloriaceae”. Applied and Environmental Microbiology. 2022;88(6):e02432-21.
11. Green ER, Mecsas J. Bacterial secretion systems: An overview. Microbiology spectrum. 2016;4(1):4.1. 13.
12. Sogin EM, Kleiner M, Borowski C, Gruber-Vodicka HR, Dubilier N. Life in the dark: Phylogenetic and physiological diversity of chemosynthetic symbioses. Annual review of microbiology. 2021;75:695-718.
13. Bennett GM, Moran NA. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome biology and evolution. 2013;5(9):1675-88.
14. Kaur R, Shropshire JD, Cross KL, Leigh B, Mansueto AJ, Stewart V, et al. Living in the endosymbiotic world of *Wolbachia*: A centennial review. Cell Host & Microbe. 2021.
15. Dubilier N, Mülders C, Ferdelman T, de Beer D, Pernthaler A, Klein M, et al. Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. Nature. 2001;411(6835):298-302.
16. Ruehland C, Blazejak A, Lott C, Loy A, Erséus C, Dubilier N. Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. Environmental Microbiology. 2008;10(12):3404-16.
17. Kleiner M, Wentrup C, Holler T, Lavik G, Harder J, Lott C, et al. Use of carbon monoxide and hydrogen by a bacteria–animal symbiosis from seagrass sediments. Environmental microbiology. 2015;17(12):5023-35.
18. Mankowski A. From genomes to communities - Evolution of symbionts associated with globally distributed marine invertebrates: Universität Bremen; 2021.
19. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Scientific reports. 2016;6(1):1-9.
20. Sato Y, Wippler J, Wentrup C, Ansorge R, Sadowski M, Gruber-Vodicka H, et al. Fidelity varies in the symbiosis between a gutless marine worm and its microbial consortium. bioRxiv. 2021.
21. Dubilier N, Giere O, Distel DL, Cavanaugh CM. Characterization of chemoautotrophic bacterial symbionts in a gutless marine worm (Oligochaeta, Annelida) by phylogenetic 16S rRNA sequence analysis and *in situ* hybridization. Applied and Environmental Microbiology. 1995;61(6):2346-50.
22. Geier B, Oetjen J, Ruthensteiner B, Polikarpov M, Gruber-Vodicka HR, Liebeke M. Connecting structure and function from organisms to molecules in small-animal symbioses through chemo-histo-tomography. Proceedings of the National Academy of Sciences. 2021;118(27).

23. Jäckle O, Seah BK, Tietjen M, Leisch N, Liebeke M, Kleiner M, et al. Chemosynthetic symbiont with a drastically reduced genome serves as primary energy storage in the marine flatworm Paracatenula. Proceedings of the National Academy of Sciences. 2019;116(17):8505-14.
24. Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, et al. Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. elife. 2015;4:e07966.
25. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology. 2012;10(1):13-26.
26. Moran NA, Bennett GM. The tiniest tiny genomes. Annual review of microbiology. 2014;68:195-215.
27. Blazejak A. Phylogenetic and functional characterization of symbiotic bacteria in gutless marine worms (Annelida, Oligochaeta): Universität Bremen Bremen; 2005.
28. Masui S, Kamoda S, Sasaki T, Ishikawa H. Distribution and evolution of bacteriophage WO in *Wolbachia*, the endosymbiont causing sexual alterations in arthropods. Journal of molecular evolution. 2000;51(5):491-7.
29. Moran NA, Degnan PH, Santos SR, Dunbar HE, Ochman H. The players in a mutualistic symbiosis: Insects, bacteria, viruses, and virulence genes. Proceedings of the National Academy of Sciences. 2005;102(47):16919-26.

Acknowledgements

I started to study biology to learn how life on this planet works. 10 years later, I gained some insight but have more questions than answers. Part of my biology journey was this PhD and I am grateful that I had so many friendly people on my side that provided me with so much support.

Prof. Dr. Nicole Dubilier, I am very grateful for the opportunity to write my PhD thesis in the Department of Symbiosis. You gave me the time and space to grow both as a scientist and as a person. Thank you for all your support concerning conferences, fieldwork, sequencing, discussing science and developing my critical thinking. Thank you!

Dr. Pierre Offre, I was delighted that you agreed to review my thesis. I want to thank you for your time and effort reviewing this thesis, being part of my defense committee, and the fruitful discussions you sparked during thesis committee meetings!

Prof. Dr. Michael Friedrich, Dr. Luis H. 'Coto' Orellana Retamal, Dr. Harald R. Gruber-Vodicka, and Sebastián G. Silva Solar, thank you for your time and willingness to be members of my thesis defense committee!

My great thanks go to **Christiane Glöckner** and the whole **IMPRS MarMic** for all the organization and the great support in every imaginable situation.

I would like to thank everybody whose work supported me at the MPIMM: the **IT** and the **personal department**, **Bernd** for finding literature, the **reception desk, Fanni and the press office** and all the **worm sorters**. I am also more than thankful for the great sequencing and bioinformatics infrastructure provided by the **Max Planck Genome Centre Cologne**.

I thank the whole **Department of Symbiosis**, all people that came and went, for this great atmosphere in the group! I enjoyed the daily lab and office life, all the cakes and BBQs, our retreats and group meetings with tons of constructive feedback. I thank **team blue** (Harald, Oli, Målin, Anna, Jero *et al.*) for our weekly meetings, our fruitful discussions and all your support. A shout-out goes especially to Niko, Alex, Yui, Dolma, Grace, Max, Bene and Miriam for great discussions on (not only) gutless oligochaetes, their morphology and how to visualize symbionts with FISH and microscopy.

Miriam, Wiebke, Martina, Janine and Silke, you made my life in the lab a lot easier by finding an answer to almost all my questions and providing me with everything I needed. You were a great help with FISH, extractions, worm sorting and much more. It was great knowing that I can count on you!

My deep thank you goes to all **HYDRA** people, Miriam and Christian, Bo and Doro, Andi and others. Without you, I would not have had worms to work on and would have missed the experience of wonderful Elba and Mallorca field trips. Thank you for all your support!

Marlene and Manuel, thank you for the great collaboration and all the time and thoughts that you invested in the proteomics of the alphaproteobacterial symbionts! You helped me get much closer to figure out what the Alphas are actually doing.

Isidora and **Lisa**, I enjoyed having you two as lab rotation students. I hope this was such a good learning experience for your as it was for me!

Thank you **Merle, Anna, Caro, David, Oli and Målin** for proofreading parts of my thesis – much appreciated!

Harald, it has been a pleasure to be your student and will be a pleasure to continue working on the projects with you. Thank you for the great discussions, your time and thinking that you invested in me and my science, teaching me to learn all this command line stuff (quite useful!) and the moments when we found time to play some table tennis!

Anna, I never promised to write a page of thank you for you, but you would deserve it for sure! Your research laid the foundation for my PhD project and your eager work attitude inspired me on a regular basis. Even more so, I want to thank you for the great time of sharing an office with you, going together through our scientific and private highs and lows, and being a great friend!

Caro, I am glad to have you as an office mate! I enjoy our discussions, our friendly work atmosphere, and I am looking forward to rocking the next Elba fieldtrip together with you!

My dear **PhD representatives** (Elisa, Chiara, Bledina, David), I feel so lucky that I got elected together with you! That year was much socializing, much organizing outside of academia, much laughing! I am grateful for the long-lasting bonds it created between us. “Trust the process!”

Dear **Tea Gang** (Ben, Miri, Merle, Julia, Inma), without you my PhD would not have been the same! Thank you for all the lunch breaks, tea breaks, gym nights and Unisee runs, the regular scientific exchanges and fun discussions, and most of all, thank you for your friendship!

Merle, words are not enough to express how grateful I am for the time with you. I enjoyed all our activities, was it scientific (discussing, collaborating, revising ...) or private (talking, cooking, cycling, swimming, Rückensport …). I hope for more times like these to come!

Elisa, we made it till the end! And I am convinced that all the hiking, and the chicken and especially the Bisi did their share. Thanks to your parents again! Also, thanks for all the spontaneous German field trips and some insight in the world and biogeochemistry of diatoms.

I thank my dear **friends** who supported me along the way and regularly reminded me that there are things besides worms and bacteria: Hanna, Susi with Edgar, Jannick with Tung and Levke, Leander and Mattus, Hendrik, Erik, and Niss. I am so grateful to have all of you in my life!

My special thanks go to **Fredy**. Thanks for choosing me as a flat mate way back when I moved to Bremen. You walked with me from the first to the last day of this journey, no matter the condition I was in. Thank you for your friendship!

Thank you, **Hans,** for hours of discussion on bacteria in worms, providing an outside view and supporting me emotionally for the most part of the way.

I thank Ecco, Jónsi, Tycho, Emancipator, Bombay Bicycle Club, Enter Shikari and many more for their **music** that led me through data analysis, lab work, frustration and joy, writing and finishing this thesis.

I am grateful to the TSC Bremen e. V., the Grienenbergsee, the Stadtwaldsee and the Unibad. My marine biologist heart and me enjoyed **diving** and swimming in those waters a lot! Breathing (or rather not breathing) was my form of meditation and helped me relax and focus on the essential.

Zum Schluss möchte ich herzlichst meiner **Familie** danken, allen voran **Mu und Julia** und Ralph, Papa und Biene und Thomas, Oma und Jürgen, Anett und Stefan, Rike und Florian. Ihr habt mich mein Leben lang begleitet, mich unterstützt, meine Neugier verstärkt und gestillt, mit mir diskutiert, mir neue Orte und Dinge gezeigt, immer ein offenes Ohr für mich gehabt und euch von den Würmern begeistern lassen. Danke für alles!

„We should always know that we can do everything. Go do!”

Jónsi – Go do

Contribution to manuscripts

Chapter 1 | Alphaproteobacteria are the most abundant and diverse class of symbionts across the marine gutless oligochaete (Oligochaeta, Annelida) diversity

Conceptual design: 50%

Data acquisition and experiments: 30%

Analysis and interpretation of the results: 70%

Preparation of figures and tables: 90%

Writing the manuscript: 70%

Chapter 2 | A novel family of Rhodospirillales (Alphaproteobacteria) in symbiosis with globally distributed gutless marine worms (Oligochaeta, Annelida)

Conceptual design: 60%

Data acquisition and experiments: 50%

Analysis and interpretation of the results: 80%

Preparation of figures and tables: 90%

Writing the manuscript: 90%

Chapter 3 | Candidatus Pumilisymbium abstrusum, a highly reduced and deeply branching Alphaproteobacterium in symbiosis with marine invertebrate gutless oligochaetes (Oligochaeta, Annelida)

Conceptual design: 50%

Data acquisition and experiments: 60%

Analysis and interpretation of the results: 70%

Preparation of figures and tables: 90%

Writing the manuscript: 80%

Universität Bremen

Versicherung an Eides Statt

Ich, Tina Enders,

versichere an Eides Statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich dem Sinne nach aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe.

Ich versichere an Eides Statt, dass ich die vorgenannten Angaben nach bestem Wissen und Gewissen gemacht habe und dass die Angaben der Wahrheit entsprechen und ich nichts verschwiegen habe.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

Erklärungen zur elektronischen Version und zur Überprüfung einer Dissertation

Hiermit betätige ich gemäß §7, Abs. 7, Punkt 4, dass die zu Prüfungszwecken beigelegte elektronische Version meiner Dissertation identisch ist mit der abgegebenen gedruckten Version.

Ich bin mit der Überprüfung meiner Dissertation gemäß §6 Abs. 2, Punkt 5 mit qualifizierter Software im Rahmen der Untersuchung von Plagiatsvorwürfen einverstanden.

_____ _____

(Ort und Datum/Place and Date) (Unterschrift/Signature)