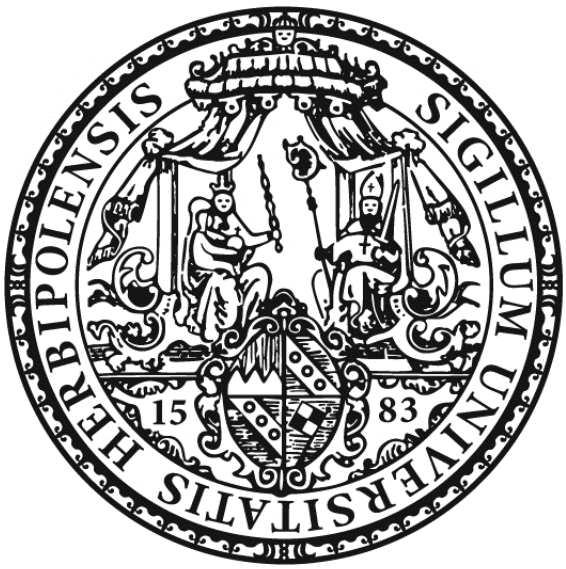


Exploring the microbiome of the Mediterranean sponge *Aplysina aerophoba* by single-cell and metagenomics

Untersuchungen am Mikrobiom des Mittelmeerschwamms *Aplysina aerophoba* mittels Einzelzell- und Metagenomik

Doctoral thesis for a doctoral degree
at the Graduate School of Life Sciences
Julius-Maximilians-Universität Würzburg
Section: Integrative Biology

Submitted by

Beate Magdalena Slaby

from
München

Würzburg, March 2017

Submitted on: ……

Members of the *Promotionskomitee*

Chairperson:	Prof. Dr. Thomas Müller
Primary Supervisor:	Prof. Dr. Ute Hentschel Humeida
Supervisor (Second):	Prof. Dr. Thomas Dandekar
Supervisor (Third):	Prof. Dr. Frédéric Partensky

Date of public defense: ……

Date of receipt of certificates: ……

Affidavit

I hereby confirm that my thesis entitled ‘Exploring the microbiome of the Mediterranean sponge *Aplysina aerophoba* by single-cell and metagenomics’ is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, Date

Signature

Acknowledgements

I received financial support for this thesis project by a grant of the German Excellence Initiative to the Graduate School of Life Sciences of the University of Würzburg through a PhD fellowship, and from the SponGES project that has received funding from the European Union's Horizon 2020 research and innovation program.

I would like to thank:

Dr. Ute Hentschel Humeida for her support and encouragement, and for providing so many extraordinary opportunities.

Dr. Thomas Dandekar and Dr. Frédéric Partensky for the supervision and a number of very helpful discussions.

my current and former colleagues in the Research Unit Marine Microbiology at the Division of Marine Ecology, GEOMAR Kiel, particularly Kristina, Hannes, Martin, Bettina, Tanja, Jutta, Regine, Sabrina, Giampiero, Laura, Lucía, Kathrin, Jule, Yu-Chen, Dr. Johannes Imhoff, Ignacio, and Alvaro for their support, their friendship, and making the relocation in the middle of the PhD project not only possible, but enjoyable.

all my former colleagues at the Department of Botany II in Würzburg, especially Janine, Lucas, Tine, Anni, Elli, Wilma, Moni, Andrea, Natascha, Usama, Cheng, and Ann-Janine for their friendship and support.

Dr. Gabriele Blum-Oehler, Jenny Braysher, Bianca Putz, and Franz-Xaver Kober of the Graduate School of Life Sciences for answering all my questions.

The US Department of Energy Joint Genome Institute, and especially Dr. Tanja Woyke, Dr. Susannah Tringe, Dr. Scott Clingenpeel and Dr. Alex Copeland for their support and sharing their expertise.

The Marine Biology Station Portorož and Piran, Slovenia, for the sampling possibility.

Dr. Monika Bright for her support with sampling.

Dr. Laura Steindler, Ilia Burgsdorf, and Dr. Thomas Hackl for great collaborations.

Dr. Frank Förster for access to the server and support.

The FACS and Cell Sorting Unit Würzburg for support with FACS sorting.

Dr. Autun Purser for proof-reading.

My family and friends for their unconditional support and encouragement, especially my parents, my brother, my 'moving crew' Markus and Willi, and Dom.

Table of contents

List of figures

List of tables

Summary

Sponges (phylum Porifera) are evolutionary ancient, sessile filter-feeders that harbor a largely diverse microbial community within their internal mesohyl matrix. Throughout this thesis project, I aimed at exploring the adaptations of these symbionts to life within their sponge host by sequencing and analyzing the genomes of a variety of bacteria from the microbiome of the Mediterranean sponge *Aplysina aerophoba*. Employed methods were fluorescence-activated cell sorting with subsequent multiple displacement amplification and single-cell / 'mini-metagenome' sequencing, and metagenomic sequencing followed by differential coverage binning. These two main approaches both aimed at obtaining genome sequences of bacterial symbionts of *A. aerophoba*, that were then compared to each other and to references from other environments, to gain information on adaptations to the host sponge environment and on possible interactions with the host and within the microbial community.

Cyanobacteria are frequent members of the sponge microbial community. My 'mini-metagenome' sequencing project delivered three draft genomes of "*Candidatus* Synechococcus spongiarum," the cyanobacterial symbiont of *A. aerophoba* and many more sponges inhabiting the photic zone. The most complete of these genomes was compared to other clades of this symbiont and to closely related free-living cyanobacterial references in a collaborative project published in Burgsdorf I*, Slaby BM* *et al.* (2015; *shared first authorship). Although the four clades of "*Ca.* Synechococcus spongiarum" from the four sponge species *A. aerophoba*, *Ircinia variabilis*, *Theonella swinhoei*, and *Carteriospongia foliascens* were approximately 99% identical on the level of 16S rRNA gene sequences, they greatly differed on the genomic level. Not only the genome sizes were different from clade to clade, but also the gene content and a number of features including proteins containing the eukaryotic-type domains leucine-rich repeats or tetratricopeptide repeats. On the other hand, the four clades shared a number of features such as ankyrin repeat domain-containing proteins that seemed to be conserved also among other microbial phyla in different sponge hosts and from different geographic locations. A possible novel mechanism for host phagocytosis evasion and phage resistance by means of an altered O antigen of the lipopolysaccharide was identified.

To test previous hypotheses on adaptations of sponge-associated bacteria on a broader spectrum of the microbiome of *A. aerophoba* while also taking a step forward in methodology, I developed a bioinformatic pipeline to combine metagenomic Illumina short-read sequencing data with PacBio long-read data. At the beginning of this project, no pipelines to combine short-read and long-read data for metagenomics were published, and at time of writing, there are still no projects published with a comparable aim of un-targeted assembly, binning and analysis of a metagenome. I tried a variety of assembly programs and settings on a simulated

test dataset reflecting the properties of the real metagenomic data. The developed assembly pipeline improved not only the overall assembly statistics, but also the quality of the binned genomes, which was evaluated by comparison to the originally published genome assemblies.

The microbiome of *A. aerophoba* was studied from various angles in the recent years, but only genomes of the candidate phylum Poribacteria and the cyanobacterial sequences from my above-described project have been published to date. By applying my newly developed assembly pipeline to a metagenomic dataset of *A. aerophoba* consisting of a PacBio long-read dataset and six Illumina short-read datasets optimized for subsequent differential coverage binning, I aimed at sequencing a larger number and greater diversity of symbionts. The results of this project are currently in review by *The ISME Journal*. The complementation of Illumina short-read with PacBio long-read sequencing data for binning of this highly complex metagenome greatly improved the overall assembly statistics and improved the quality of the binned genomes. Thirty-seven genomes from 13 bacterial phyla and candidate phyla were binned representing the most prominent members of the microbiome of *A. aerophoba*. A statistical comparison revealed an enrichment of genes involved in restriction modification and toxin-antitoxin systems in most symbiont genomes over selected reference genomes. Both are defense features against incoming foreign DNA, which may be important for sponge symbionts due to the sponge's filtration and phagocytosis activity that exposes the symbionts to high levels of free DNA. Also host colonization and matrix utilization features were significantly enriched. Due to the diversity of the binned symbiont genomes, a within-symbionts genome comparison was possible, that revealed three guilds of symbionts characterized by i) nutritional specialization on the metabolization of carnitine, ii) specialization on sulfated polysaccharides, and iii) apparent nutritional generalism. Both carnitine and sulfated polysaccharides are abundant in the sponge extracellular matrix and therefore available to the sponge symbionts as substrates. In summary, the genomes of the diverse community of symbionts in *A. aerophoba* were united in their defense features, but specialized regarding their nutritional preferences.

Zusammenfassung

Schwämme (Phylum Porifera) sind evolutionär alte, sessile Filtrierer, die eine äußerst vielfältige mikrobielle Gemeinschaft in ihrer internen Mesohylmatrix beherbergen. Das Ziel meiner Doktorarbeit war es, die Anpassungen dieser Symbionten an das Leben in ihrem Schwammwirt zu erforschen. Dazu habe ich die Genome einer Vielzahl von Bakterien aus dem Mikrobiom des Mittelmeer-Schwammes *Aplysina aerophoba* sequenziert und analysiert. Meine angewandten Methoden waren die fluoreszenzaktivierte Zellsortierung mit anschließender so genannter „multiple displacement amplification“ und Einzelzell- / „Mini-Metagenom“-Sequenzierung und metagenomischer Sequenzierung gefolgt von „differential coverage binning“. Diese beiden Ansätze zielten darauf ab, Genomsequenzen von bakteriellen Symbionten von *A. aerophoba* zu erhalten, die dann sowohl miteinander, als auch mit Referenzen aus anderen Habitaten verglichen wurden. So sollten Informationen gewonnen werden über Anpassungen an ein Leben im Wirtsschwamm und über mögliche Interaktionen mit dem Wirt und innerhalb der mikrobiellen Gemeinschaft.

Cyanobakterien sind häufig Mitglieder der bakteriellen Gemeinschaft in Schwämmen. Mein "Mini-Metagenom"-Sequenzierprojekt lieferte drei Genom-Entwürfe von „*Candidatus* Synechococcus spongiarum,“ dem cyanobakteriellen Symbionten von *A. aerophoba* und vieler weiterer Schwämme, die die photische Zone bewohnen. Das vollständigste dieser Genome wurden mit anderen Kladen dieses Symbionten verglichen und mit nah verwandten, freien lebenden Cyanobakterien-Referenzen in Burgsdorf I *, Slaby BM * et al. (2015; * geteilte Erstautorenschaft). Obwohl die vier Kladen von „*Ca.* Synechococcus spongiarum“ aus den vier Schwammarten *A. aerophoba, Ircinia variabilis*, *Theonella swinhoei* und *Carteriospongia foliascens* auf der Ebene der 16S-rRNA-Gensequenzen zu etwa 99% identisch waren, unterschieden sie sich deutlich auf Genom-Ebene. Nicht nur die Genomgrößen waren von Klade zu Klade verschieden, sondern auch der Gengehalt und eine Reihe von Merkmalen, einschließlich Proteinen mit genannten „eukaryotic-like domains,“ leucinreiche „repeats“ oder Tetratricopeptid-„repeats“. Auf der anderen Seite teilten die vier Kladen eine Reihe von Merkmalen wie Ankyrin-„repeat“-Domänen-haltige Proteine, die auch in anderen Phyla von Schwammsymbionten in verschiedenen Wirtsschwämmen und aus verschiedenen geografischen Orten konserviert zu sein schienen. Ein möglicher neuartiger Mechanismus zur Phagozytose-Vermeidung und zur Phagenresistenz mittels eines veränderten O-Antigens des Lipopolysaccharids wurde identifiziert.

Um vorherige Hypothesen über die Anpassung von Schwamm-assoziierten Bakterien auf ein breiteres Spektrum des Mikrobioms von *A. aerophoba* zu testen und gleichzeitig in der Methodik voran zu schreiten, entwickelte ich einen bioinformatischen Arbeitsablauf, um metagenomische Illumina-„short-read“-Sequenzdaten mit PacBio-„long-reads“ zu

kombinieren. Zu Beginn dieses Projektes gab es keine veröffentlichte Methodik zur Verknüpfung von „short-reads“ und „long-reads“ für die Metagenomik, und auch jetzt gibt es keine veröffentlichten Projekte mit einem vergleichbaren Ziel von nicht-gezieltem „Assembly“, „Binning“ und Analyse eines Metagenoms. Ich habe eine Auswahl von „Assembly“-Programmen und Einstellungen auf einem simulierten Testdatensatz getestet, der die Eigenschaften der realen metagenomischen Daten widerspiegelt. Die entwickelte „Assembly“-Methode verbesserte nicht nur die Gesamtstatistik, sondern auch die Qualität der einzelnen, „gebinnten“ Genome, die durch Vergleich zu den ursprünglich veröffentlichten Genom-Sequenzen evaluiert wurde.

Das Mikrobiom von *A. aerophoba* wurde in den letzten Jahren aus verschiedenen Blickwinkeln untersucht, aber nur Genome des Candidatus-Phylum Poribakterien und die Cyanobakteriensequenzen aus meinem oben beschriebenen Projekt wurden bisher veröffentlicht. Durch die Anwendung meiner neu entwickelten „Assembly“-Methodik auf einen metagenomischen Datensatz von *A. aerophoba* bestehend aus einem PacBio-„long-read“-Datensatz und sechs Illumina-„short-read“-Datensätzen, die für das anschließende „differential coverage binning“ optimiert waren, zielte ich darauf ab, eine größere Anzahl und Vielfalt von Symbionten zu sequenzieren. Die Ergebnisse dieses Projektes sind derzeit bei *The ISME Journal* in Review. Die Komplementierung von Illumina „short-read“ mit PacBio „long-read“-Sequenzdaten für das „binning“ dieses hochkomplexen Metagenoms hat die Gesamt-„assembly“-Statistik sowie die Qualität der „gebinnten“ Genome deutlich verbessert. Siebenunddreißig Genome aus 13 Bakterienphyla und Candidatus-Phyla wurden „gebinnt“, die die prominentesten Mitglieder des Mikrobioms von *A. aerophoba* darstellten. Ein statistischer Vergleich zeigte eine Anreicherung von Genen, die mit Restriktionsmodifikationen und Toxin-Antitoxin-Systemen zusammenhängen, in den meisten Symbionten-Genomen im Vergleich zu ausgewählten Referenzgenomen. Beides sind Mechanismen zur Verteidigung gegen eindringende Fremd-DNA, die für Schwamm-Symbionten aufgrund der Schwamm-Filtration und Phagozytose-Aktivität wichtig sein können, die die Symbionten hohen Konzentrationen von freier DNA aussetzen. Auch mögliche Wirtskolonisations- und Matrixnutzungsmechanismen waren signifikant angereichert. Wegen der Vielfalt der „gebinnten“ Symbionten-Genome war ein Genom-Vergleich innerhalb der Symbionten möglich, der drei Gilden von Symbionten zum Vorschein brachte, die gekennzeichnet waren durch i) Ernährungsspezialisierung auf die Metabolisierung von Carnitin, ii) Spezialisierung auf sulfatierte Polysaccharide und iii) scheinbaren Nahrungs-Generalismus. Sowohl Carnitin als auch sulfatierte Polysaccharide sind in der extrazellulären Schwammmatrix reichlich vorhanden und stehen so den Schwammsymbionten als Substrat zur Verfügung. Die Genome der diversen Symbionten-Gemeinschaft in *A. aerophoba* waren in ihren Verteidigungsmechanismen vereint, aber spezialisiert hinsichtlich ihrer Ernährung.

1 Introduction

1.1 Sponges (phylum Porifera)

Marine sponges (Porifera) are the oldest extant multicellular animals with a fossil record dating back to the Precambrian (Antcliffe *et al.*, 2014; Du *et al.*, 2015; Brain *et al.*, 2012; Gold *et al.*, 2016). Throughout Earth history, sponges played an important role as reef builders and even dominated reef communities at times (Heckel, 1974). To this day, they are present in a variety of marine ecosystems from shallow tropical reefs to the deep-sea, and they still dominate the community in specific deep sea regions known as sponge grounds (Maldonado *et al.*, 2016). Sponges are highly diverse spanning an estimated number of 15,000 species (Hooper, John and Van Soest, 2002). They differ in size from a few millimeters to meters, they show a range of shapes from bowl- or vase-shaped to encrusting and branching, and they can have a wide variety of colors. Taxonomically, sponges are divided into the four classes Demospongiae, Calcarea, Hexactinellida, and Homoscleromorpha that differ in the building materials for their spicules, the material – if present – of the exoskeleton, the presence or absence of spongin fibers, the cell type and the body form (Hooper, John and Van Soest, 2002; Bergquist, 1998). The majority of extant sponges are demosponges (Hentschel *et al.*, 2003).

Marine sponges are among the structurally simplest multicellular organisms on Earth. The sponge body (except for Hexactinellida) possesses two types of barrier-forming cell layers, namely pinacoderm and choanodem, that consist of pinacocyte and choanocyte cells, respectively (Simpson, 1984). The pinacoderm forms the outer surface of the sponge body and lines the aquiferous canal system, while the choanocyte cells are located in choanocyte chambers (Ereskovsky, 2010). Between the external pinacoderm and the canal system is the mesohyl matrix that is mainly composed of collagen, galectin and glycoconjugates (Ereskovsky, 2010). While sponges do not contain organs or tissues, they possess nonepithelian, totipotent cells, that are phagocytotically active and amoeboid, i.e. they can move freely through the mesohyl (Hentschel *et al.*, 2003). The skeleton of demosponges consists either of spongin fibers alone or of spongin fibers and siliceous spicules (Ereskovsky, 2010).

Based on the complexity of their canal system, sponges are categorized into three main types: asconoid, syconoid, and leuconoid (van Soest *et al.*, 2012). The structurally simplest form – only represented in a number of calcareous sponges today – is the ascon type, where pores in the thin wall enable waterflow into the central cavity, that is lined with choanocytes (Ghiold *et al.*, 1994). Likewise only extant in calcareous sponges, is the sycon type with radial

canals formed by folding of the body wall, and small choanocyte chambers (Ghiold *et al.*, 1994). The most widespread and complex form is the leucon type. Here, the body wall of the sponge is thickened and folded into a number of flagellate chambers lined with choanocytes and connected by a complex canal system (Ghiold *et al.*, 1994).

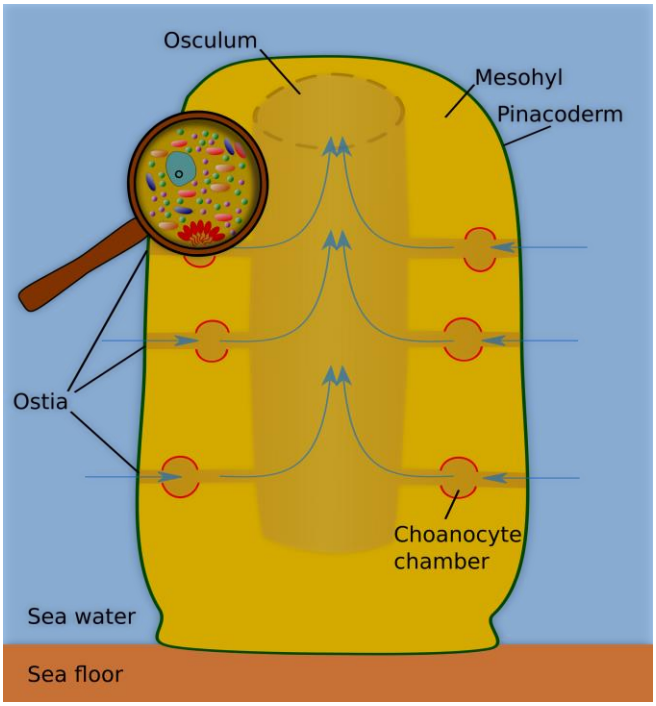


Figure 1-1 Schematic cross section through a leuconoid demosponge. Blue arrows indicate water flow produced by choanocyte cells lining choanocyte chambers (red). The magnifying glass indicates a zoom-in on the mesohyl, where the totipotent amoeboid cells (turquoise) and the symbiotic bacteria (various shapes and colors) are located. The array of flagellated cells (red) at the bottom of the magnification are choanocyte cells. Drawing: B. Slaby.

While recently carnivorous sponges were discovered in some deep-sea habitats (Hestetun *et al.*, 2016; Dressler-Allame *et al.*, 2016; Maldonado *et al.*, 2016), the vast majority of sponges are filter-feeders. They pump up to 24,000 liters of seawater per kg sponge per day through inhalant pores (ostia) in their outer pinacoderm layer and through a system of canals into choanocyte chambers. Specialized flagellated choanocyte cells create the water current for filtration by beating their flagellae and capture food particles out of the water (Figure 1-1) (Taylor *et al.*, 2007; Vogel, 1977). The nearly-sterile filtered water is pumped into the central cavity and emerges through an exhalant opening (osculum) (Hentschel *et al.*, 2012; Ghiold *et al.*, 1994). The food particles are phagocytosed by totipotent archaeocyte cells located in the sponge mesohyl matrix (Taylor *et al.*, 2007).

Sponges reproduce both sexually and asexually. In terms of asexual, clonal reproduction, sponges can fragment, bud, or produce gemmules (Webster and Thomas, 2016). For sexual reproduction, a sponge individual can possess either both male and female

reproductive parts (hermaphroditic) or only one (gonochoristic) (Webster and Thomas, 2016). Demosponges can be ovoviviparous, oviparous, and even viviparous: either fertilization and embryonic development take place internally in the mesohyl (ovoviviparity and viviparity), or – after external fertilization – larvae develop in the environment (oviparity) (Ereskovsky, 2010). The reproductive strategies are polyphyletic and even mainly oviparous orders like Astrophorida include ovoviviparous genera (Vacelet, 1999; Ereskovsky, 2010).

Sponges engage in a variety of ecological functions in marine ecosystems. They compete for space, but also positively interact with other organisms (Aerts, 2000; Rützler, 1970; Wulff, 2008). In carbonated coral reefs, they consolidate substrate that can then be used for corals to grow on, so sponges enhance reef growth (Wulff, 1984). But excavating sponges can also hinder reef growth by boring into the reef structure and thereby affecting the corals' structural integrity (Diaz and Rützler, 2001). Sponges create a trophic link between water column and benthos by coupling carbon fluxes via their filtering of food particles (Gili and Coma, 1998). Additionally, they are able to take up dissolved organic matter (DOM) such as carbon and nitrogen, and – by shedding large amounts of cells due to their rapid cell turnover rates – make them available to other heterotrophic organisms (de Goeij *et al.*, 2013; Alexander *et al.*, 2014). Large amounts of the taken-up DOM derive from other members of the coral reef community, namely corals and macroalgae (Rix *et al.*, 2016a, 2016b). This recycling process has been termed the 'sponge loop' in analogy to the established 'microbial loop' and explains how the biological hot spots of coral reefs can thrive in such oligotroghic environments (de Goeij *et al.*, 2013; Azam *et al.*, 1983). In oligotrophic tropical reef environments as well as the deep sea, sponges have been shown to take up dissolved organic matter (DOM) and to create detritus by cell renewal and shedding of old cells (de Goeij *et al.*, 2013; Maldonado, 2015; Rix *et al.*, 2016b). As the newly formed detritus serves as a food source for the associated fauna, sponges play a key role in these otherwise nutrient-poor ecosystems.

Already in Greek antiquity sponges were used for various purposes from cleaning to medical applications (Voultsiadou, 2007). The wound healing properties of sponges was already recognized then, a use probably explained today by the vast variety of bioactive compounds identified (Mehbub *et al.*, 2014; Flemer *et al.*, 2012; Horn *et al.*, 2015; Abdelmohsen *et al.*, 2010). These compounds can have various types of bioactivity, e.g. cytotoxicity, antiinfective, or anticancer activity (Belarbi, 2003). Sponges are also interesting for biotechnological applications e.g. for tissue engineering due to their natural skeleton structure, or for their collagen content, which has a plethora of applications from pharmaceutical use to cosmetics (Green *et al.*, 2003; Swatschek *et al.*, 2002).

1.2 Sponge-microbe symbiosis

Microbiomes have been a focus of much research in recent years. A wide variety of environments have been explored, from desert sand to permafrost soils (Rivkina *et al.*, 2016; Johnson *et al.*, 2017). Also the effects of the microbial communities on their environment, on the host in symbioses, and especially on humans have been intensely studied (Afshinnekoo *et al.*, 2015; Faist *et al.*, 2016; Heinsen *et al.*, 2016; Schröder and Bosch, 2016). Researchers have arrived at the conclusion, that animals and plants cannot be seen as isolated organisms any longer, but have to be studied as a holistic system comprising of the host itself and all its associated microorganisms, as a 'holobiont' (Bordenstein and Theis, 2015; Gordon *et al.*, 2013; Mayer *et al.*, 2014; Deines and Bosch, 2016; McFall-Ngai *et al.*, 2013).

1.2.1 Microbial diversity

In agreement with the holobiont concept, also sponges host highly diverse and distinct microbiomes that can constitute up to 40% of sponge volume and may be crucial for their evolutionary success (Vacelet, 1975; Easson and Thacker, 2014; Tian *et al.*, 2014; Webster and Thomas, 2016). Based on the abundance of microbes, two groups of sponges are observed: high microbial abundance (HMA) and low microbial abundance (LMA) sponges (Hentschel *et al.*, 2003). The microbial communities of high microbial abundance (HMA) sponges were hypothesized to play a crucial role in the sponges' success e.g. by supplying supplemental nutrition to the host (Tian *et al.*, 2014; Erwin and Thacker, 2008b). 16S rRNA gene amplicon studies discovered an unusually high phylum-level diversity and stability of microbial associations in marine sponges comprising phototrophic as well as heterotrophic symbionts (Schmitt *et al.*, 2012b; Thomas *et al.*, 2016; Easson and Thacker, 2014; Webster and Thomas, 2016). The sponge microbiome spans as many as 52 microbial phyla and candidate phyla with the diversity and abundance varying between sponge species (Webster and Thomas, 2016). The most dominant symbiont groups belong to the phyla Proteobacteria (mainly Gamma- and Alphaproteobacteria), Actinobacteria, Chloroflexi, Nitrospirae, Cyanobacteria, candidatus phylum Poribacteria, and Thaumarchaea (Webster and Thomas, 2016). Most of these symbionts seem to be sponge species-specific and vertically transmitted to the next generation of sponges via the larvae (Schmitt *et al.*, 2012b, 2008; Webster *et al.*, 2010; Usher *et al.*, 2001; Oren *et al.*, 2005). A comparison of bacterial community profiles derived from 16S rRNA and 16S rRNA genes revealed that a large part of the sponge-associated bacterial community is not only present, but also metabolically active (Kamke *et al.*, 2010).

Cyanobacteria are also common members of the sponge microbial community in tropical as well as temperate regions (Schmitt *et al.*, 2012b; Thacker and Freeman, 2012). The

group of cyanobacterial sponge symbionts is polyphyletic with symbiont species of the orders Chroococcales, Prochlorales, and Oscillatoriales, but it comprises mainly clade VI cyanobacteria of different *Synechococcus* spp. (Chroococcales) (Steindler *et al.*, 2005; Honda *et al.*, 1999; Usher, 2008). Within this group we find "*Candidatus* Synechococcus feldmanni" that is mainly inhabiting *Petrosia ficiformis* from the Mediterranean and eastern Atlantic oceans (Usher, 2008; Burgsdorf *et al.*, 2014). The more widespread cyanobacterial sponge symbiont is "*Candidatus* Synechococcus spongiarum," which comprises at least 12 different subclades that show up to 99% 16S rRNA gene sequence identity, but could be revealed by 16S to 23S rRNA internal transcribed spacer (ITS) sequence phylogeny (Erwin *et al.*, 2012a; Erwin and Thacker, 2008a). The separation into these clades seems to be driven mainly by geographic location and by host phylogeny (Erwin and Thacker, 2008a). This phototrophic symbiont has been shown to provide supplemental nutrition to its host sponges (Freeman and Thacker, 2011), while profiting from shelter and nutrition provided by the sponge (Erwin *et al.*, 2012a). "*Ca.* Synechococcus spongiarum" resides extracellularly and is vertically transmitted to the next generation of host sponges (Usher *et al.*, 2001; Oren *et al.*, 2005; Schmitt *et al.*, 2008; Webster *et al.*, 2010). Phylogenetically, it is equidistant from the above-described *Synechococcus*/*Prochlorococcus* subclade that spans marine as well as freshwater strains of the genera *Synechococcus*, *Prochlorococcus*, and *Cyanobium* (Gao *et al.*, 2014b; Steindler *et al.*, 2005).

1.2.2 Microbial function

The microbiome of marine sponges includes autotrophs as well as heterotrophs, which are involved in a number interactions with their host in terms of nutrient exchange – a supposedly mutualistically beneficial interaction (Webster and Thomas, 2016). The symbionts are supplied with nutrients and ammonia from the host, while the sponge benefits from waste removal and supplemental nutrition by the symbionts (Webster and Thomas, 2016). Cyanobacterial symbionts fix carbon and supply the host with photosynthesis products like glycerol (Webster and Thomas, 2016; Taylor *et al.*, 2007). Some sponges were even shown to obtain more than half of their required energy from their cyanobacterial symbionts (Wilkinson, 1983). Ammonia-oxidizing bacteria and archaea are also common members of the sponge microbial community, as well as sulfate-reducing and sulfur-oxidizing bacteria, microbes producing polyphosphate granules (possibly to store phosphate for times of deprivation), and symbionts producing essential vitamins, such as different B vitamins (Bayer *et al.*, 2008a; Fan *et al.*, 2012; Tian *et al.*, 2014; Colman, 2015; Thomas *et al.*, 2010).

Comparisons between metagenomes of sponge-associated and seawater microbial consortia identified gene features that might be of importance specifically to sponge-associated bacteria (Thomas *et al.*, 2010; Fan *et al.*, 2012; Hentschel *et al.*, 2012; Horn *et al.*,

2016). One recurring topic in sponge-microbial symbiosis are mobile genetic elements and genetic transfer with specific emphasis on transposases (Fan *et al.*, 2012; Thomas *et al.*, 2010; Gao *et al.*, 2014b). The abundance of mobile elements has been interpreted as crucial for symbiotic bacterial genome evolution as a means for genome reduction (Moran and Plague, 2004). Restriction modification systems and CRISPR-Cas systems, on the other hand, might be important protection mechanisms against incoming viruses and free DNA for the sponge symbionts, as they are hypothesized to be exposed to vastly higher quantities of viral particles in comparison to free-living seawater bacteria (Thomas *et al.*, 2010). Further recurring findings are metabolic adaptations of the symbionts, e.g. regarding vitamin B12 and ammonium assimilation (Kamke *et al.*, 2014; Thomas *et al.*, 2010; Bayer *et al.*, 2008a). Several studies have shown metabolic dependencies between sponge host and bacterial community (Bayer *et al.*, 2008a; Kamke *et al.*, 2013; Radax *et al.*, 2012a; Hoffmann *et al.*, 2009). For example, Poribacteria seem to be able to degrade complex carbohydrates produced by the host which are abundant in the mesohyl matrix (Kamke *et al.*, 2013). Further probable adaptations are the so-called eukaryotic-like protein domains, repeat proteins like ankyrins, that have been found enriched in sponge symbionts (Thomas *et al.*, 2010; Nguyen *et al.*, 2014; Liu *et al.*, 2012). These were hypothesized to play a role in the evasion of phagocytosis by the host (Thomas *et al.*, 2010).

1.3 *Aplysina aerophoba*

1.3.1 Geographic distribution and physical properties

This thesis project is divided into multiple parts that all study the host system of the marine HMA sponge *Aplysina aerophoba* SCHMIDT 1862 (class Demospongiae, subclass Verongimorpha, order Verongiida, family Aplysinidae), commonly known as ‘gold sponge’ (Bayer *et al.*, 2008b) with the aim of gaining genomic information on its microbial symbionts with state-of-the-art ‘omics’ and bioinformatics approaches. According to the World Porifera Database (www.marinespecies.org/porifera/), the bright yellow *A. aerophoba* (Figure 1-2) is common in the Mediterranean Sea, around the Azores and Cape Verde, the Saharan upwelling zone, the South European Atlantic Shelf, and the Southern Gulf of Mexico. The phylogeny of the family of Aplysinidae was resolved by ITS-2 and 18S rRNA gene trees, where *A. aerophoba* clusters with *Aplysina cavernicola*, which is likely due to geographic distribution (Schmitt *et al.*, 2005).

A. aerophoba has been intensely studied regarding its pumping behavior, chemistry, metabolism, microbiology, reactions to environmental change, and was even proposed as a model sponge (Friedrich *et al.*, 2001; Noyer *et al.*, 2010; Pfannkuchen *et al.*, 2009; Sacristan-

Soriano *et al.*, 2011; Sacristán-Soriano *et al.*, 2012; Schmitt *et al.*, 2012a). An *in situ* study showed that *A. aerophoba* is continuously pumping when healthy and undisturbed, independent from season and time of day, concluding that the sponge is always well oxygenated and that its own waste products are removed (Pfannkuchen *et al.*, 2009). Especially the chemistry of *A. aerophoba* has received attention, as it contains large amounts of brominated alkaloids that in turn play an ecological role in predatory protection, competition for space, protection against biofouling, and defense against pathogenic microorganisms (Sacristan-Soriano *et al.*, 2011; Sacristán-Soriano *et al.*, 2012; Turon *et al.*, 2000).

Figure 1-2 *Aplysina aerophoba*. Photo: B. Slaby

It has been demonstrated that *A. aerophoba* may take up food bacteria at rates of up to 2.76 x 10^6 bacteria per gram sponge wet weight per hour depending on the cell surface properties and size of the food bacteria (Wehrl *et al.*, 2007). At the same time, it is capable of differentiating between food bacteria and symbionts, taking up symbiont preparations at significantly lower rates of around 5.37 x 10^4 bacteria per g sponge wet weight per hour (Wehrl *et al.*, 2007). This supports the hypothesis of phagocytosis evasion mechanisms by the symbionts (Thomas *et al.*, 2010).

1.3.2 The *A. aerophoba* microbiome

As a HMA sponge, the microbial community associated with *A. aerophoba* is not only characterized by high numbers of 6.4 ± 4.6 x 10^8 bacteria per gram sponge tissue (Friedrich *et*

al., 2001), but also by an extraordinary diversity of bacteria. This diversity is already apparent at the phylum level: Acidobacteria, Actinobacteria, Chloroflexi, Nitrospira, Proteobacteria, Spirochaetae, Bacteroidetes, Cyanobacteria, Deinococcus-Thermus, Firmicutes, Gemmatimonadetes, and candidate phyla Poribacteria, OP10, OS-K, SAUL, and TM7 were discovered in the bacterial community of *A. aerophoba* by 16S rRNA gene amplicon sequencing (Schmitt *et al.*, 2012a). Quantitative PCR (qPCR) and fluorescence *in situ* hybridization (FISH) revealed the dominance of Chloroflexi and Poribacteria in *A. aerophoba* (Bayer *et al.*, 2014a). The microbial community is very stable, even when exposed to stress such as starvation and exposure to antibiotics (Friedrich *et al.*, 2001).

The functional gene repertoire of the *A. aerophoba* microbiome was assessed by GeoChip revealing increased numbers of nitrification and ammonification-related genes and archaeal autotrophic carbon fixation genes in comparison to seawater (Bayer *et al.*, 2014b). Stress-related genes, on the other hand, were reduced (Bayer *et al.*, 2014b). Targeting specific taxa and genes, evidence for the presence and activity of ammonia-oxidizing bacteria and archaea (AOB and AOA, respectively) was collected (Bayer *et al.*, 2007; Cardoso *et al.*, 2013; Bayer *et al.*, 2008a). The discovery of natural products is a research field in itself, with sponges a known and widely studied source. From *A. aerophoba*, a number of natural products have been described as well that are often produced by its microbial community (Hentschel *et al.*, 2001; Horn *et al.*, 2015; Bayer *et al.*, 2013; Siegl and Hentschel, 2010; Pimentel-Elardo *et al.*, 2012). Changes in the microbial as well as the chemical patterns of *A. aerophoba* were shown in diseased specimens (Webster *et al.*, 2008b).

In summary, a considerable amount of information on the microbiome of *A. aerophoba* has accumulated over the years. Yet, at the beginning of this thesis only a handful of genomes of representatives of the candidate phylum Poribacteria had been sequenced (Fieseler *et al.*, 2004, 2006; Siegl *et al.*, 2011; Kamke *et al.*, 2013, 2014). The microbial community was shown to be very stable even under conditions of stress, such as starvation or exposure to antibiotics (Friedrich *et al.*, 2001).

1.4 Sequence-based analyses of microbiomes

Increasing effort has been placed on gauging the diversity of the Earth's microbiome and we have come to understand that the vast majority of bacteria is still uncultivable, which limits our possibilities for determining their roles in the microbial community (Rinke *et al.*, 2013). The term 'microbial dark matter' comprises this large uncultivable part of the microbial community and mirrors the analogous 'dark matter' of astrophysics, as for both, proxies are needed from which to draw conclusions on their behavior and importance (Marcy *et al.*, 2007). The great majority of sponge symbionts are as yet uncultivable (Esteves *et al.*, 2016), and

therefore culture-independent approaches have to be applied to gain genomic and thereby functional information. Research thus far has focused mainly on general patterns by analyzing metagenomes, metaproteomes, and metatranscriptomes (Thomas *et al.*, 2010; Fan *et al.*, 2012; Liu *et al.*, 2012; Radax *et al.*, 2012b). Recently, via cultivation-independent methods such as single-cell genomics and metagenomic binning, a number of symbiont genomes were obtained and even new bacterial candidate phyla were described (Siegl *et al.*, 2011; Kamke *et al.*, 2014; Liu *et al.*, 2011). Nevertheless, the sequencing of uncultivated microbes is still in its infancy.

1.4.1 Recent developments in sequencing technologies

Sequencing technologies have come a long way from the early days of Sanger sequencing to USB stick-sized ultra-long read MinION sequencers (Oxford Nanopore Technologies, Oxford, UK) (Koren and Phillippy, 2015). Along the way, the diversity of uncultivable bacteria has been targeted by 16S rRNA gene diversity, first via polymerase chain reaction (PCR) followed by clone libraries and Sanger sequencing (Erwin *et al.*, 2012b), and later by amplicon sequencing targeting the same gene with high-throughput sequencing methods (Schmitt *et al.*, 2012a). As these approaches were merely delivering information on one gene to assess microbial diversity, they did not supply any further functional genomic information. Also in genomic sequencing, cultivable bacteria allow production of sufficient biomass for sequencing by growing them in culture media, whereas no sufficient biomass – and therefore the suitable DNA volume – of an individual bacterial species can be obtained for uncultivable bacteria.

In the early 2000s, new technologies emerged – single-cell genomics and metagenomics (Figure 1-3) – the former specifically targeting members of the microbial community after isolating them, the latter sequencing the whole microbial consortium at once (Woyke *et al.*, 2009; Gilbert and Dupont, 2011). To isolate bacteria for single-cell genomics, a variety of approaches were applied such as dilution to extinction, micropipetting, and fluorescence-activated cell sorting (FACS) (Lauro *et al.*, 2009; Macaulay and Voet, 2014). The DNA of the single isolated cell was then amplified in a whole genome amplification (WGA) reaction, commonly by multiple-displacement amplification (MDA) utilizing the phi29 polymerase (Dean *et al.*, 2001), to produce sufficient DNA of the target cell for genome sequencing (Woyke *et al.*, 2009). Single-cell genomics is a targeted approach that can be of great advantage if information on the target bacterium is at hand that enables or facilitates selective sorting, e.g. autofluorescence that can be used for FACS sorting. At the same time, this feature can be a disadvantage if no such information is available. In such situations, cells have to be isolated and whole genome amplified 'blindly' followed by possibly extensive PCR screening to identify the target bacterium. Additionally, the WGA reaction has some flaws, such as

uneven amplification, chimera formation, and co-amplification of contaminants (Blainey, 2013).

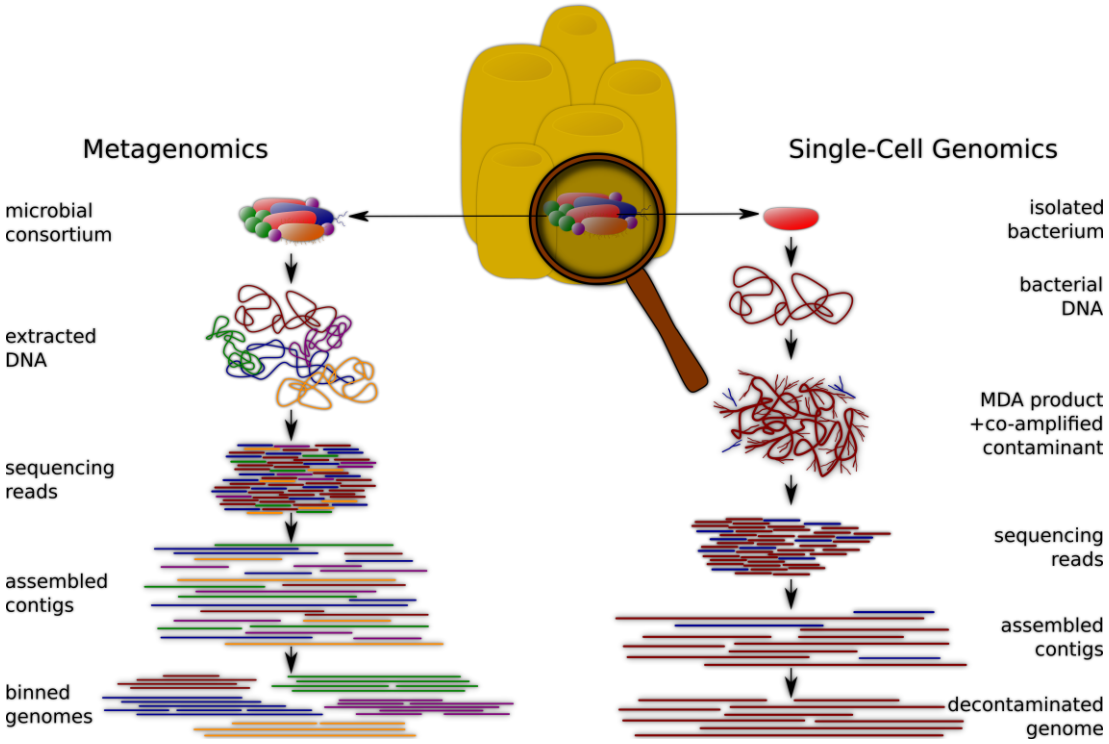


Figure 1-3 Workflow for metagenomic binning and single-cell genomics for retrieving genomes of sponge symbionts. Drawing: B. Slaby.

In metagenomics, the DNA of a microbial consortium is the basis for untargeted high-throughput sequencing. To obtain genomes from metagenomic data, either the reads or the assembled contigs need to be separated bioinformatically and ideally sorted into individual genomes by binning algorithms (Alneberg *et al.*, 2014; Albertsen *et al.*, 2013; Kang *et al.*, 2015). This way, many genomes out of the consortium are sequenced at the same time. But there are several flaws also to this approach. Due to the implemented short-read sequencing, the sequenced genomes can frequently not be closed but stay draft genomes. While it is rather straightforward to bin the dominating members of a bacterial community, increasing sequencing depth is required to reach sufficient coverage for bacteria of low abundance, which leads to increasing sequencing costs. Also, a metagenomic bin is never the genome of one bacterium but always of a community – a bin could be viewed as the genomic content of a species or strain at best.

As single-cell and metagenomics both have advantages and disadvantages, some studies have combined them to benefit from both techniques (Mason *et al.*, 2012; Wilson *et al.*, 2014). Yet, one common issue the approaches share, are the assembly gaps due to short read lengths.

Long-read sequencing has tackled this problem successfully for genomes (Huddleston *et al.*, 2014; Rhoads and Au, 2015; Koren and Phillippy, 2015; Shibata *et al.*, 2013).

1.4.2 Why short-reads fall short

In metagenomics, short-read sequencing (mostly Illumina HiSeq or MiSeq) has been the method of choice to obtain sufficient sequencing depth at reasonable costs (Koren and Phillippy, 2015). As short-reads cannot resolve repeat sequences that exceed the read length, these repeats cause ambiguities and ultimately break up the assembly into multiple contigs (Koren and Phillippy, 2015). Thus, genomes assembled from short-reads – binned from metagenomes or directly assembled in a genome sequencing project – will not be closed, but remain draft genomes.

This issue could be resolved by long-read sequencing, when the reads exceed the repeat sequences in length (Koren and Phillippy, 2015). One commonly used long-read sequencing technique is single-molecule real-time (SMRT) sequencing developed by PacBio (Pacific Biosciences of California, USA). While an anchored polymerase replicates the template DNA by incorporating fluorescent-labeled nucleotides, their emission spectra are recorded in sequencing movies that can then be interpreted and translated into a sequence read (reviewed in Rhoads and Au 2015). The sequencing template is called a SMRTbell, which is a double-stranded DNA molecule closed into a single-stranded circular DNA by hairpin adaptors on both ends (Rhoads and Au, 2015). Therefore, depending on the lifetime of the polymerase, both strands of the template DNA can be sequenced multiple times in a single run, that will then be split into so-called subreads at the adaptor sequence locations (Rhoads and Au, 2015). While early SMRT reads were still relatively short and had a high error rate, later changes in chemistry improved sequencing length to up to 50 kbp and read accuracy to ~87% (Koren *et al.*, 2013; Lee *et al.*, 2014).

The error-prone PacBio reads need to be corrected either with themselves - provided sufficient sequencing depth is available, e.g. in the form of subreads - or with (Illumina) short-reads of far lower error-rate (Hackl *et al.*, 2014; Rhoads and Au, 2015; Koren and Phillippy, 2015). The combination of PacBio and Illumina sequencing data in hybrid assemblies is able to close the described assembly gaps by spanning over long repeats, merge contigs and thereby reconstruct the genome architecture. This has even enabled the *de novo* assembly of closed genomes (Liao *et al.*, 2015).

1.4.3 Long-read sequencing in metagenomics

In metagenomics, commonly high-throughput short-read sequencing is applied, e.g. on an Illumina HiSeq platform. Hybrid assembly projects combining long-reads and short-reads

for isolate genomes were widely used in isolate genomics as a stand-alone tool or in combination with Illumina short-read sequencing (Shibata *et al.*, 2013; Ricker *et al.*, 2016; Bashir *et al.*, 2012; Utturkar *et al.*, 2014; Beims *et al.*, 2015). As long-read sequencing has become progressively more affordable (Koren *et al.*, 2013), obtaining sufficient sequencing depth to improve a metagenomic assembly is now within reach on a manageable budget. Considering the large improvements in genome sequencing with greater read-lengths, the drawbacks of short-reads in metagenomic assembly, and at the same time the vast improvements in bioinformatics in the recent years, the attempt of complementing short-read sequences with long-reads also in metagenomics seemed to be the next logical step.

Yet, no studies on a hybrid assembly for metagenomics were published at the beginning of this project. Also at present, only a handful of publications on this topic are available (Frank *et al.*, 2016; Tsai *et al.*, 2016; Beckmann *et al.*, 2014). By the time of completion of this thesis project, PacBio-Illumina hybrid assembly approaches have been proven useful for a variety of applications (Frank *et al.*, 2016; Tsai *et al.*, 2016; Beckmann *et al.*, 2014). In recent targeted binning approaches, superior assembly quality has been demonstrated (Frank *et al.*, 2016; Tsai *et al.*, 2016). Yet, un-targeted binning and performance for less abundant members of the microbial communities have not been assessed.

To improve a metagenomic dataset already deeply sequenced and optimized for differential coverage binning by six Illumina HiSeq datasets, a complementary metagenomic PacBio dataset was obtained. Due to the lack of publications on the topic at the time, one of my goals in this thesis was the development of an assembly pipeline aiming to combine metagenomic Illumina short-read and PacBio long-read data in a hybrid assembly.

2 Material and methods

2.1 Research questions and aims

This thesis comprised of biological as well as methodological projects in the context of sponge-microbial symbiosis:

- Isolation and single-cell / mini-metagenome sequencing of the cyanobacterial symbiont "*Ca.* Synechococcus spongiarum" from the sponge host *A. aerophoba*, comparison of its genome to other clades of this species and to free-living cyanobacterial references, to explore adaptations to a life in sponges in general as well as specific adaptations to each host sponge and environment.

- Development of a bioinformatic pipeline for the metagenomic hybrid assembly of Illumina short-reads and PacBio long-reads using a test dataset of simulated reads and explore to what extent the addition of metagenomic long-reads improves the assembly and subsequent binning.

- Application of the developed assembly pipeline to the metagenome of *A. aerophoba* followed by untargeted binning and comparison of the sponge symbiont genomes to selected non-sponge-associated references to explore common symbiont-enriched genomic features as well as to identify divisions of labor between the symbionts.

Selected contents of this thesis were published in Burgsdorf I*, Slaby BM*, Handley KM, Haber M, Blom J, Marshall CW, Gilbert JA, Hentschel U, Steindler L. (2015). Lifestyle evolution in cyanobacterial symbionts of sponges. *mBio* **6**: e00391-15. *shared first authorship and submitted to The ISME Journal for publication under the reference Slaby BM, Hackl T, Horn H, Bayer K, Hentschel U. Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. *ISME J*, in review. All further sequenced "*Ca.* Synechococcus spongiarum" genomes were submitted to Genome Announcements for publication under the reference Slaby BM, Hentschel U. Draft genome sequences of "*Candidatus* Synechococcus spongiarum," cyanobacterial symbionts of the Mediterranean sponge *Aplysina aerophoba*. *Genome Announc*, accepted.

2.2 Sponge collection and enrichment of prokaryotic cells

Specimens of *A. aerophoba* were retrieved in May 2013 and 2014 in Piran, Slovenia (45°31' N, 13°34' E) from a depth of five to seven meters by SCUBA diving by a diver of the Marine Biology Station Portorož and Piran, and transported to the laboratory in ambient sea water, where they were placed in flow-through aquaria to recover. The sponges of the 2013 collection were transported to Würzburg in natural sea water at ambient temperature and placed in a Mediterranean aquarium for recovery, then sampled within one week upon arrival. The sponges of the 2014 collection were processed in the laboratory facilities of the Marine Biology Station Portorož and Piran within three days following collection.

Pinacoderm and mesohyl, visually distinguishable by color due to the pigments of the cyanobacterial symbionts concentrated in the outer pinacoderm (Figure 2-1), were separated with a sterile scalpel blade. The sponge-associated prokaryotes (SAPs) were enriched for both fractions following a previously published protocol (Fieseler *et al.*, 2004). A number of freshly prepared pinacoderm SAPs were used for the enrichment of cyanobacterial symbionts by fluorescence-activated cell sorting (FACS). SAPs not used for FACS sorting were frozen with 15% glycerin at -80°C.

Figure 2-1 Longitudinal section of *A. aerophoba* revealing the clearly distinguishable pinacoderm (reddish brown) and mesohyl (yellow). Photo: K. Bayer.

2.3 Sequencing and analysis of cyanobacterial sponge symbionts

2.3.1 Laboratory methods: From sample to sequence

"*Ca.* Synechococcus spongiarum" cells were enriched by sorting freshly prepared SAPs on a FACS Aria III (BD Biosciences, San Jose, CA, USA) of the core facility of the University of Würzburg. For cell sorting, a 488nm laser was used to excite the chlorophyll *a* and phycoerythrin autofluorescence of this cyanobacterium. Single-cells were sorted onto 96-well plates and multiple cells were bulk sorted into one tube to create an enrichment of the target organism ('mini-metagenome'). The cell sorts were transported on ice and stored at -80°C until further processing.

I simultaneously screened the mini-metagenome for "*Ca.* Synechococcus spongiarum" and tested whether the concentration of cells was high enough to serve directly as a template for a PCR by targeting the cyanobacterial 16S-23S ITS region with the primers 16S-1247f and ITS-Alar (Rocap *et al.*, 2002) in the following reaction. The PCR was performed in a volume of 50µl containing 10ng of each primer (16S-1247f: 5'-CGT ACT ACA ATG CTA CGG-3', ITS-Alar: 5'-CTC TAC CAA CTG AGC TAW A-3') (Sigma-Aldrich, Merck, Darmstadt, Germany), 10nmol total deoxynucleotides (dNTPs), (Fermentas, Thermo Fisher Scientific, Waltham, MA, USA), 1.25U DreamTaq (Fermentas, Thermo Fisher Scientific, Waltham, MA, USA), and 1x conc. DreamTaq buffer (green, containing $MgCl_2$ and loading dye). The PCR was performed three times containing different volumes of template: 2µl, 4µl, and 5µl. In the PCR an initial denaturation of 5 minutes at 95°C was followed by 30 cycles of 30 seconds denaturation at 95°C, 30 seconds of annealing at 49°C and 1:30 minutes of elongation at 72°C, followed by a final elongation of 5 minutes at 72°C.

To assess the purity of the cell sort in the mini-metagenome, a 16S rRNA gene PCR with the universal primers 27f and 1492r (Lane, 1991) (27f: 5'-GAG TTT GAT CCT GGC TCA-3', 1492r: 5'-TAC GGY TAC CTT GTT ACG ACT T-3') was performed followed by a clone library, and RFLP. A choice of PCR products of clones were sent for Sanger sequencing based on the RFLP patterns. The concentrations in the PCR mixture were as described above with 5µl of mini-metagenome as an insert. In the reaction, an initial denaturation of 5 minutes 95°C was followed by 30 cycles of 45 seconds of denaturation at 95°C, 1 minute of annealing at 54°C and 1:30 minutes of elongation at 72°C, followed by a final elongation of 5 minutes at 72°C. The PCR was performed in triplicate, the products were pooled and cleaned with the NucleoSpin Gel and PCR Cleanup Kit (Macherey-Nagel, Düren, Germany).

The CloneJET PCR Cloning Kit (Thermo Fisher Scientific, Waltham, MA, USA) was used to clone the PCR products of both PCRs described above into competent *Escherichia coli*

cells following the manufacturer's protocol. In the case of the universal 16S rRNA gene primers the clone library served as a contamination screening. For the cyanobacterial 16S-23S ITS primers it aimed to confirm that only the target cyanobacterium "*Ca*. Synechococcus spongiarum" was present in the mini-metagenome and no other cyanobacterial symbionts or non-symbionts. For the colony PCR, all 12 16S rRNA gene clones were picked, and 50 16S-23S ITS clones. The colony PCR was performed in 40µl of volume with 8nmol dNTPs, 0.625U DreamTaq and 8pmol of each primer (pJET forward and reverse sequencing primers provided with the cloning kit). In the PCR reaction 5 minutes of initial denaturation at 95°C were followed by 35 cycles of 30 seconds of denaturation at 94°C, 30 seconds of annealing at 60°C and 1:20 minutes elongation at 72°C, and a final elongation of 5 minutes at 72°C.

Clones with an insert at the expected length were assessed in a restriction fragment length polymorphism (RFLP) assay using both the MSPI and HAEIII FastDigest restriction enzymes (Fermentas, Thermo Fisher Scientific, Waltham, MA, USA). In a total reaction volume of 20µl contained 1µl of each restriction enzyme and 5µl of colony PCR product in 1x concentrated reaction buffer. The mixture was incubated for a minimum of 30 minutes at 37°C. Based on the fragment pattern, clones were selected for sequencing. For these, three colony PCRs were performed as described above, the PCR products were pooled and cleaned with the NucleoSpin Gel and PCR Cleanup Kit, and the DNA concentration was measured by NanoDrop. If necessary, the DNA was diluted to 20-80 ng/µl, and 5µl of PCR product mixed with 5µl of 5µM pJet forward primer were sent for Sanger sequencing at GATC. Low-quality ends of the sequences were trimmed with the Chromatogram Explorer Lite v5.0.2 (http://www.dnabaser.com/download/chromatogram-explorer/) automatically, and the closest relative for each sequence was determined by a BLASTn search (https://blast.ncbi.nlm.nih.gov) (Altschul *et al.*, 1990).

For single-cells as well as aliquots of the mini-metagenome, multiple-displacement amplification (MDA) was performed with REPLI-g Single Cell Kit (QIAGEN, Venlo, Netherlands) following the manufacturer's protocol with halved reagent volumes. In the case of the mini-metagenomes, 2µl of FACS-sorted cells were used as insert for the MDA reaction. The MDA products were diluted 1:10 with sterile water. From cell sorting until the end of MDA, all work was conducted on a clean bench.

For PCR screening, a serial dilution was used as an insert (1:25 → 1:10 → 1:5 in PCR mix) to obtain a final dilution of 1:12,500 of the MDA product in the PCR. The screening PCR was performed in a volume of 50µl containing 10µl of diluted MDA product as insert, 10ng of each primer (27f: 5'-GAG TTT GAT CCT GGC TCA-3', 1492r: 5'-TAC GGY TAC CTT GTT ACG ACT T-3'), (Sigma-Aldrich, Merck, Darmstadt, Germany), 10nmol dNTPs, 1.25U DreamTaq, and 1x conc. DreamTaq buffer (green, containing MgCl2 and loading dye). In the PCR an initial denaturation of 10 minutes at 95°C was followed by 35 cycles of 1 minute

denaturation at 95°C, 30 seconds of annealing at 54°C and 1:30 minutes of elongation at 72°C, followed by a final elongation of 5 minutes at 72°C. The PCR products were cleaned up with the NucleoSpin Gel and PCR Cleanup Kit (Macherey-Nagel, Düren, Germany) and the purity and concentration of the DNA was measured by NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA). If necessary, the PCR products were diluted to meet the sequencing company's requirements, and 5µl of cleaned PCR product were sent to Sanger sequencing with 5µl of 5µM primer (341f: 5'-CCT ACG GGA GGC AGC AG-3') at GATC Biotech (Konstanz, Germany). The 16S rRNA gene sequences were phylogenetically identified by BLAST (Altschul *et al.*, 1990).

Three MDA products of mini-metagenome 15 (L, M, and N) were selected for sequencing based on the 16S rRNA gene PCR screening. DNA concentrations were additionally measured by Qubit (Thermo Fisher Scientific, Waltham, MA, USA) using the high sensitivity assay. The MDA products were sequenced at the DOE JGI on an Illumina HiSeq2000 platform (150 bp paired-end reads) with the sample IDs 1033526, 1033529, and 1033532.

2.3.2 Bioinformatic methods: From sequencing reads to genome comparison

The sequencing reads were quality filtered and then *de novo* assembled with SPAdes 3.0.0 as part of JGI's assembly pipeline (Bankevich *et al.*, 2012). Assembly quality was assessed with QUAST 3.1 (Gurevich *et al.*, 2013). Decontamination of the assemblies was performed with the binning software CONCOCT v. 0.4.0 at default settings (Alneberg *et al.*, 2014). The bin containing the cyanobacterial target genome was identified with PhyloSift v1.0.1 (Darling *et al.*, 2014). A local version of rRNA prediction of WebMGA (Wu *et al.*, 2011) was used at default settings on the whole assembly to identify the rRNA genes, that were then compared to the 16S rRNA gene sequences obtained by Sanger sequencing and then added to the genome bin (Ollivier *et al.*, 2008). Open reading frames (ORFs) were called with prodigal v2.6.1 (Hyatt *et al.*, 2010) and genome completeness was assessed using a set of 111 single-copy essential genes (Albertsen *et al.*, 2013) that were annotated with hmmsearch against a hmm database of these genes with hmmer 3.1b1 (Eddy, 2009). To assess the similarity of the binned genomes, they were aligned with BRIG version 0.95 (Alikhan *et al.*, 2011) using BLAST+ version 2.5.0 (Altschul *et al.*, 1990). 15L was selected for comparison to other clades of "*Ca.* Synechococcus spongiarum" based on estimated genome completeness. Its 16S rRNA gene and 16S-23S ITS sequence was deposited under the GenBank accession KP763586, the draft genome sequence was deposited under accession JYFQ00000000. The genome sequences of 15M and 15N were deposited under the accession numbers MWLD00000000 and MWLE00000000, respectively.

Genome 15L was compared to "*Ca.* Synechococcus spongiarum" SP3, 142, and SH4 , from *Theonella swinhoei*, *Ircinia variabilis*, and *Carteriospongia foliascens*, respectively, and to selected free-living cyanobacterial references (Burgsdorf *et al.*, 2015; Gao *et al.*, 2014b). SP3 and 142 were sampled, sequenced and binned by Ilia Burgsdorf and Laura Steindler of the University of Haifa, Israel (see Burgsdorf et al. (2015) for details). SH4 was previously published (Gao *et al.*, 2014b).

To obtain information about genome architecture of the symbiont draft genomes, they were aligned to *Cyanobium gracile* PCC6377 due to its high mean amino acid similarity to the symbionts and close relatedness using Mauve version 20120303 (build 645). The contigs were reordered by first aligning SP3 to *C. gracile* PCC6377 with Mauve's reordering tool, and then the other symbiont draft genomes to SP3 with BLASTn and Artemis (Altschul *et al.*, 1990; Carver *et al.*, 2005). RAST was used to predict open reading frames (ORFs) and annotate the genomes of the four symbionts and six closely related free-living cyanobacteria (Aziz *et al.*, 2008; Overbeek *et al.*, 2014). In WebMGA, clusters of orthologous groups (COGs) and 'Kyoto Encyclopedia of Genes and Genomes' (KEGG) pathways were annotated using RPSBLAST with an e-value cutoff of 0.001 (Tatusov *et al.*, 2003; Kanehisa *et al.*, 2004; Wu *et al.*, 2011).

Genome completeness was estimated as described above omitting 11 genes based on their absence or presence in multiple copies in the closed reference genomes. The EDGAR platform was utilized to obtain genes found in all four symbiont genomes while absent from all six reference genomes by a reciprocal best-BLAST-hit approach (Blom *et al.*, 2009). COG annotation of this gene set was performed via WebMGA, then the obtained COGs were compared to those of the free-living references (Wu *et al.*, 2011). To consider a COG unique to "*Ca.* Synechococcus spongiarum," it had to be absent from all six analyzed references. If no COG annotation was available for an ORF, the KEGG annotation was used. Interactions between COGs were predicted with the STRING, member lists of COGs of interest were obtained from eggnog version 3.0, and protein domains from NCBI's refseq_protein database (http://blast.be-md.ncbi.nlm.nih.gov/Blast.cgi) (Snel *et al.*, 2000; Powell *et al.*, 2012). CRISPRFinder was used to detect clustered regularly interspaced short palindromic repeat (CRISPR) arrays (Grissa *et al.*, 2007). CRISPR-Cas modules were identified in the whole-genome alignments for 'confirmed CRISPRs' and associated proteins obtained from SEED and COG annotations.

STAMP v2.0.9 was implemented to create a heatmap of COG class abundance accompanied with an average neighbor clustering (UPGMA) dendrogram (Parks *et al.*, 2014). With Welch's *t* test, statistically significant differences were determined between "*Ca.* Synechococcus spongiarum" and free-living cyanobacteria on COG class level using Bonferroni correction for multiple testing and a *P* value cutoff of 0.05. The EDGAR platform was utilized to identify the pangenome of "*Ca.* Synechococcus spongiarum", i.e. the sum of

all genes in these four genomes, as well as its core genome, i.e. the intersection of genes common to all four symbionts (Blom *et al.*, 2009). From the mean percent identity values of the core genome genes, the amino acid identity matrix of the symbionts was calculated. A phylogenomic tree was additionally constructed by EDGAR based on a core genome of the "*Ca.* Synechococcus spongiarum" genomes and 15 free-living cyanobacterial references. Two strains of *Synechococcus elongatus* were used as an outgroup. First, a reciprocal BLAST search with *Cyanobium gracile* PCC6307 as reference was implemented to determine the amino acid sequences of the core genes, then homologous genes were aligned by MUSCLE (Edgar, 2004). The alignments were merged into one, which was subsequently used for phylogenomic neighbor-joining tree calculation by PHYLIP with 100 bootstrap replications using Kimura distance matrix (Blom *et al.*, 2009; Felsenstein, 1995).

The 16S-23S internal transcribed spacer (ITS) regions of SP3 and 142 were obtained with EMIRGE and confirmed by PCR and clone libraries by Ilia Burgsdorf (Burgsdorf *et al.*, 2015). For 15L, PCR and Sanger sequencing were used to obtain the nucleotide sequence of this region as described in 2.2.1. The sequences for SH4 and references were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). From a MUSCLE alignment created in MEGA 6.0, a maximum-likelihood tree was constructed with Kimura 2-parameter substitution model with gamma-distributed rate variation and a proportion of invariant sites (+G+I) and 1,000 bootstrap replications (Edgar, 2004; Tamura *et al.*, 2013).

2.4 Development of a hybrid assembly pipeline for PacBio long-reads and Illumina short-reads

2.4.1 Simulating sequencing reads for a test dataset

A selection of sequencing data preparation steps in combination with assemblers was tested to develop an assembly pipeline that could deal with the uneven coverage of an Illumina short-read-sequenced metagenome and at the same time implement PacBio long-reads. A test dataset was used that simulated the key conditions of a typical sponge-derived metagenome (e.g. uneven sequencing coverage). At the same time, it served as an assembly quality check, as the test dataset was composed of known, fully sequenced genomes. The test dataset was created by Thomas Hackl by simulating PacBio and Illumina sequencing reads (100bp paired-end reads with 180bp insert) from nine fully sequenced bacterial genomes: *Acidobacterium capsulatum* ATCC51196, *Bacteroides vulgatus* ATCC8482, *Clostridium thermocellum* ATCC27405, *Desulfovibrio vulgaris* DP4, *Fusobacterium nucleatum* ATCC25586, *Nitrosomonas europaea* ATCC19718, *Porphyromonas gingivalis* ATCC33277, *Sulfolobus tokodaii* 7, *Thermoanaerobacter pseudoethanolicus* ATCC33223 at sequencing coverages

200, 140, 100, 70, 50, 32, 20, 12, and 8, respectively. The software pbsim-1.0.2 was used to simulate PacBio reads (Ono *et al.*, 2013) and art-2.1.8 to simulate Illumina reads (Huang *et al.*, 2012). Thomas Hackl corrected the PacBio long-reads with all available Illumina short-reads using his newly developed tool proovread-meta, a version of proovread that he adapted for metagenomic datasets (Hackl *et al.*, 2014; Hackl, 2016).

2.4.2 Testing assemblers and settings

At the beginning of this project, no bioinformatic pipeline was published for a hybrid assembly of metagenomic long-reads and short-reads. Yet, several assemblers were available, some focusing on metagenomics, some on genomics, some accounting for uneven coverage of single-cell sequencing projects, and some enabling hybrid assembly of long-reads and short-reads for genome assemblies. As my aim in this project was not an exhaustive comparison of assembly methods but to find a working pipeline to be then tested on real data, I focused on three *de novo* assemblers, namely Omega, IDBA-UD, and SPAdes (Haider *et al.*, 2014; Peng *et al.*, 2012; Bankevich *et al.*, 2012). Omega is a metagenomic assembler that uses overlap-graphs for assembly (Haider *et al.*, 2014). IDBA-UD is optimized for sequencing data of uneven depth, namely single-cell and metagenomic datasets (Peng *et al.*, 2012). SPAdes on the other hand is specifically programmed for single-cell assemblies and therefore also optimized for uneven sequencing depth (Bankevich *et al.*, 2012).

Testing a variety of assemblers, settings and data preparation steps was necessary to develop a pipeline capable of assembling metagenomic long-reads and short-reads together. To safe computation time and resources, a test dataset simulating the features of the real *A. aerophoba* data was used consisting of simulated Illumina and PacBio reads from nine fully sequenced bacterial genomes at different coverages. The tested *de novo* assemblers were omega v.1.0.2, IDBA-UD of IDBA v.1.1.1 and SPAdes v.3.5.0 at a variety of settings (Haider *et al.*, 2014; Peng *et al.*, 2012; Bankevich *et al.*, 2012), and with and without prior Illumina sequencing read normalization by the bbnorm algorithm of bbmap v. 34 (https://sourceforge.net/projects/bbmap/) (Bushnell, 2015) (Figure 2-2). In summary, three *de novo* assembly programs were compared, all programmed to handle uneven sequencing coverage, with one optimized for metagenomics, one for single-cell genomics, and one for both. I also tested if bbnorm (Bushnell, 2015) could account for the uneven coverage of the metagenomic dataset beforehand and thereby improve the assembly or even enable an assembler to work with the data.

In a first step, assemblies of only the Illumina reads were created to assess how well the three assembly algorithms handled the uneven metagenomic data. At this step, also the effect of read normalization with bbnorm was also tested. Because bbnorm includes a read

correction, it may not be necessary to use an additional read correction by SPAdes. Therefore, also the differences were assessed when turning off read correction by SPAdes with the –only-assembler option. In a second step, different PacBio-Illumina hybrid assemblies were attempted with SPAdes and IDBA_UD. For SPAdes, possible improvements by assembly with longer kmers were also tested, and again the effect of the read correction by SPAdes was assessed. Either the read correction was turned off in SPAdes (--only-assembler) or the right encoding had to be supplied (--phred-offset 33). As SPAdes emerged superior to omega and IDBA_UD very early in the comparison, I only tested the default settings for the latter two and rather focused on optimizing the settings for SPAdes.

2.4.3 Comparing and evaluating of assemblies and bins

The assemblies were compared with QUAST version 2.3 (Gurevich *et al.*, 2013) and by aligning the contigs to the original published assemblies of the nine bacterial genomes by the script wgaDrawingPipeline.pl available via AliTV (Ankenbrand *et al.*, 2016). As SPAdes produced the best results, further comparisons were made between the Illumina-only assembly and the PacBio-Illumina hybrid assembly calculated with this program to assess also a possible binning of the genomes from the metagenomic assemblies. Reads were mapped back to the contigs with bowtie2 v. 2.2.2 at default settings (Langmead and Salzberg, 2012) and samtools v. 0.1.18 (Li *et al.*, 2009) was implemented for sorting, indexing and depth calculation of the resulting mapping files. An in-house python script was then used to calculate the average coverage of each contig from the depth files (https://github.com/bslaby/scripts/avgcov_from_samtoolsout.py). 16S rRNA genes were annotated with a local version of rRNA prediction at default settings (Wu *et al.*, 2011) and their phylogenetic identity was determined with RDPclassifier (Cole *et al.*, 2014) and BLASTn (Altschul *et al.*, 1990). The contigs were manually binned with a previously published R pipeline (Albertsen *et al.*, 2013) coloring contigs containing the 16S rRNA genes based on their phylogeny when plotted according to their coverage values determined by bowtie2 mapping and calculated by SPAdes during assembly. The completeness was estimated according to a previously published R pipeline based on 111 essential single-copy genes (Albertsen *et al.*, 2013) using prodigal v2.6.1 (Hyatt *et al.*, 2010) and hmmer3.1b1 (Finn *et al.*, 2011). The same completeness estimation was also used on the original references for comparison.

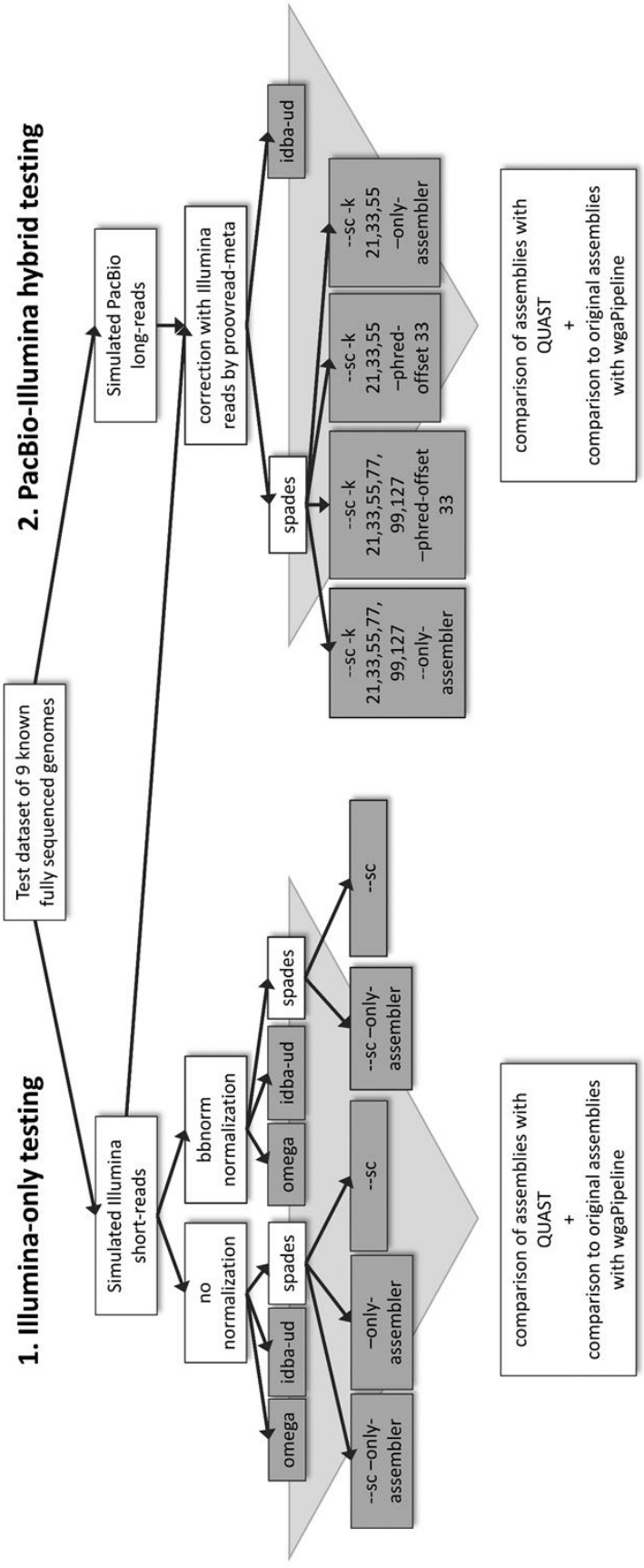


Figure 2-2 Overview of tested assemblies of 1) only the Illumina reads, and 2) PacBio and Illumina reads.

2.5 *Apylsina aerophoba* metagenomics

2.5.1 Laboratory methods: DNA extraction and sequencing

DNA of sponge-associated prokaryotes (SAPs) obtained from either pinacoderm or mesohyl tissue (three replicates each) was extracted with the FastDNA SPIN Kit for Soil (MP Biomedicals, Santa Ana, CA, USA). Different cell lysis protocols were applied for each triplicate to obtain differential sequencing coverage for downstream binning as previously described (Albertsen *et al.*, 2013; Alneberg *et al.*, 2014): (i) bead beating, following the manufacturer's protocol, (ii) freeze-thaw cycling (3 cycles of 20 minutes at -80 °C and 20 minutes at 42 °C), (iii) proteinase K digestion for 1 hour at 37°C (TE buffer with 0.5% SDS and proteinase K at 100 ng/ml final conc.). Quantity and quality of the extracted DNA were assessed by Nanodrop and Qubit high sensitivity assay, and agarose gel electrophoresis, respectively. Respective DNA from two exctraction rounds was pooled and metagenomic DNA was sequenced on an Illumina HiSeq2000 platform (150-bp paired-end reads) and quality filtered at the DOE Joint Genome Institute (Walnut Creek, CA, USA) following the JGI sequencing and data processing pipeline (Markowitz *et al.*, 2012). For the PacBio dataset, DNA was extracted with the above-mentioned kit following the manufacturer's protocol (cell lysis by bead beating) and sequenced on a PacBio RS II platform using 8 SMRT cells by GATC Biotech (Konstanz, Germany)**.**

2.5.2 Bioinformatic methods: From assembly to annotation

Illumina reads were coverage-normalized with bbnorm of BBMap v. 34 (https://sourceforge.net/projects/bbmap/) at default settings. PacBio reads were corrected with all (non-normalized) Illumina reads using proovread (Hackl *et al.*, 2014) optimized for handling metagenomic data (Hackl, 2016). Only corrected PacBio reads longer than 1 000 bp were used for further analyses. To assess the improvement of the assembly by adding PacBio long-reads compared to only Illumina short-reads, I assembled two sets of data: i) only the Illumina reads (Illumina-only assembly) and ii) Illumina and PacBio reads together (hybrid assembly). The two independent assemblies were calculated with SPAdes v. 3.5.0 (Bankevich *et al.*, 2012) for kmers 21, 33, 55, 77, 99, and 127 and with the single-cell and only-assembler options enabled. Only contigs of at least 1 000 bp length were used for further analyses.

Binning was performed with CONCOCT v. 0.4.0 at default settings (Alneberg *et al.*, 2014). Before binning, contigs longer than 20 000 bp were split into sub-contigs of at least 10 000 bp length with the script cut_up_fasta.py (Alneberg *et al.*, 2014). The non-normalized Illumina reads of the six Illumina datasets were mapped to the sub-contigs with bowtie2 v. 2.2.2 at default settings (Langmead and Salzberg, 2012). The resulting SAM files were

converted to BAM, sorted, and indexed with samtools v. 0.1.18 (Li *et al.*, 2009), and duplicates were marked according to the script map-bowtie2-markduplicates.sh provided with the CONCOCT package (Alneberg *et al.*, 2014). Samtools v. 0.1.18 was also used for sorting, indexing, and depth calculation (Li *et al.*, 2009). The in-house python script avgcov_from_samtoolsout.py (https://github.com/bslaby/scripts/) was used to calculate the average coverage of each sub-contig. The coverage tables for each mapping were merged into one for binning with CONCOCT v. 0.4.0 (Alneberg *et al.*, 2014) at default settings. A fasta file for each bin was created with the in-house python script mkBinFasta.py (https://github.com/bslaby/scripts/). Sub-contigs were merged into the original contigs again. If sub-contigs of one contig were assigned to different bins, the contig was placed in the bin by majority-vote. Assembly statistics were obtained from QUAST v. 3.1 (Gurevich *et al.*, 2013). To assess similarity of Illumina-only and hybrid assembly as well as assembly improvements by adding of PacBio long-reads on the genome level, the contigs of an Illumina-only bin were mapped to the contigs of the corresponding hybrid assembly bin with nucmer of MUMmer 3.0 (Kurtz *et al.*, 2004) and visualized with AliTV (Ankenbrand *et al.*, 2016).

Open reading frames (ORFs) were called with prodigal v. 2.6.1 (Hyatt *et al.*, 2010) with -m and -p meta options enabled, and the completeness of genomic bins was estimated by hmmsearch (HMMER 3.1b1) against a database of 111 essential genes with –cut_tc and –notextw options (Albertsen *et al.*, 2013; Finn *et al.*, 2011). Only reference genomes > 90% and bins > 70% completeness were used in further analyses.

The Illumina-only and the PacBio-Illumina hybrid assemblies were deposited on MG-RAST (Meyer *et al.*, 2008) (Table 3-17). Additionally, raw Illumina sequencing data was deposited under GOLD Study ID Gs0099546 (Reddy *et al.*, 2014). Uncorrected and corrected PacBio reads were deposited on MG-RAST (Meyer *et al.*, 2008) with the IDs mgm4670967.3 and mgm4670966.3, respectively. The accession numbers for all bins > 70% completeness are listed in Table 3-18. The Illumina-only assembly is also deposited on GenBank with the accession MKWU00000000.

2.5.3 Statistical analysis: Comparison to references and within symbionts

Twenty-seven reference genomes were chosen based on phylogeny and environment. Close taxonomic relatedness to the symbiont genomes, closed genomes, as well as marine (or at least aquatic) environments were preferably selected. In order to be able to validate the binning process, we included the sponge symbiont genomes "*Ca.* S. spongiarum" 15L (Burgsdorf *et al.*, 2015) and "*Ca.* Poribacterium" WGA3G (Kamke *et al.*, 2013) in the analyses. We retrieved nucleic acid fasta files for all selected references from GenBank and MG-RAST (Benson *et al.*, 2007; Meyer *et al.*, 2008) which were then processed like the

symbiont bins with respect to ORF prediction and annotation. Five additional references were added for 16S rRNA gene tree calculation for better phylogenetic resolution. The annotation of rRNA genes was performed with a local version of rRNA prediction at default settings (Wu *et al.*, 2011). The 16S rRNA genes were taxonomically placed with RDP classifier at a 80% confidence cutoff (Wang *et al.*, 2007) and the classification tool of SINA 1.2.11 (Pruesse *et al.*, 2012) using the SILVA and Greengenes databases (DeSantis *et al.*, 2006; Quast *et al.*, 2013). Gap-only sites were removed from the SINA alignment of both, bins and references, in SeaView 4.5.2 (Gouy *et al.*, 2010). A Neighbor Joining tree (GTR+G+I), which was determined to be the most suitable DNA/protein model for the data, was calculated in MEGA7 with 100 bootstrap replications (Kumar *et al.*, 2016). Additionally, a concatenated gene tree of 29 essential genes was created (see Appendix 3-1 for a list of genes). Alignments for every gene individually using the muscle algorithm in MEGA7 (Edgar, 2004; Kumar *et al.*, 2016) were merged with a sequence of 20 Ns between the genes. After identifying the most suitable DNA/protein model for the data, a maximum likelihood tree (LG+G+I) was calculated in MEGA7 with 100 bootstrap replications (Kumar *et al.*, 2016). Bins lacking 16S rRNA genes or with an ambiguous classification of this gene were phylogenetically classified according to their placement in the concatenated tree.

ORFs were annotated with rpsblast+ of BLAST 2.2.28+ against a local version of the COG database (ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/, download on 2015-05-28) (Tatusov *et al.*, 2003; Wu *et al.*, 2011). Only annotations with an e-value $\leq$ 1e-6 were used for further analyses, and only one annotation per ORF was kept ranked by e-value, length and bitscore. Because many sponge-symbiont lineages, in some cases whole phyla, are not abundant in seawater, we have opted for an approach different from previous publications, where only seawater metagenomes were used for comparison (Thomas *et al.*, 2010). We selected reference genomes based on phylogenetic similarity and on genome completeness. Marine sources were preferred over other sources.

To discover statistically significant differences between the sponge symbiont genomes and reference genomes, Welch's t-test was performed in STAMP 2.0.9 (Parks *et al.*, 2014) with Storey FDR and a q-value cut-off of 0.01. This was performed on the COG class level, double-counting COGs that belong to multiple classes, as well as on the COG level. Interactions between the significantly sponge-enriched COGs were explored using STRING v10 networks (Szklarczyk *et al.*, 2015) and a heatmap was created in R version 3.2.3 (https://www.r-project.org). The phylo.heatmap function of phytools package version 0.5.30 (Revell, 2012) was used to complement the heatmap with phylogeny. The phylogenetic tree accompanying the heatmap is a simplified version (bins only) of the concatenated gene phylogeny. The symbiont genomes were compared by applying a principle component analysis (PCA) in R with FactoMineR package version 1.33 (Lê *et al.*, 2008), factoextra

package version 1.0.3 (https://cran.r-project.org/web/packages/factoextra/index.html), and ggplot2 version 2.2.0 (http://ggplot2.org).

3 Results

3.1 Assessing the genome of the "*Ca.* Synechococcus spongiarum" group

3.1.1 Assessment of clade F genomes from *A. aerophoba*

3.1.1.1 Cell sort purity assessment

For sorting, double-positive signals for both types of autofluorescence of "*Ca.* Synechococcus spongiarum" were identified in the sorting plot in the FACS software with the chlorophyll a fluorescence in the APC-Cy7-A channel and the phycoerythrin fluorescence in the PE-Texas Red-A channel (Figure 3-1). The sorting window was set by hand around the cells showing the strongest of both signals. Single cells were sorted onto a total of nine 96-well plates and multiple cells were bulk sorted into one tube.

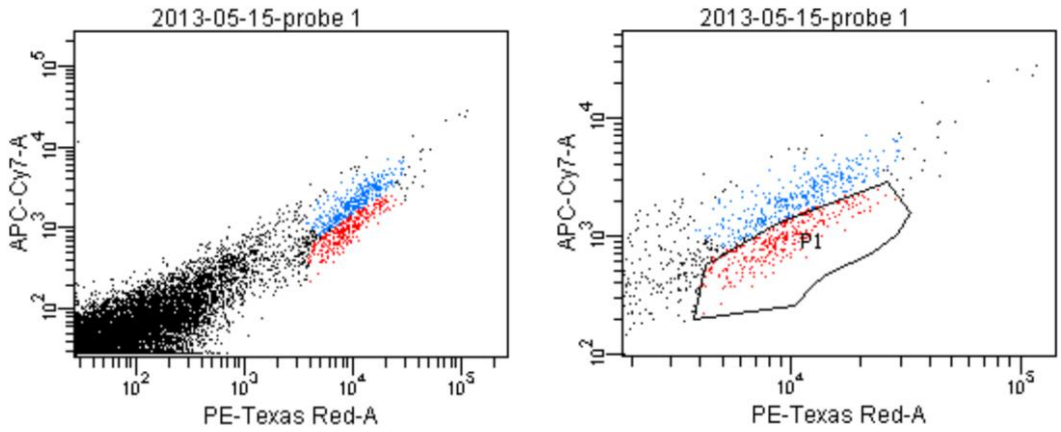


Figure 3-1 FACS sorting plot of "Ca. Synechococcus spongiarum" cells with positive APC-Cy7-A / chlorophyll a and PE-Texas Red-A / phycoerythrin signals. On the right: Zoom-in on the target signals. Two different sorting windows were selected (signals of target cells in blue and red; P1: selection for cells with signals in red). The cells for the mini-metagenome derived from the selection shown in red.

To assess the concentration and purity of the mini-metagenomes, a cyanobacterial 16S-23S ITS region PCR as well as a 16S rRNA gene PCR were performed. For mini-metagenome 15, three PCRs of the 16S-23S ITS region with different insert volumes were tested, all producing an amplification product (Figure 3-2). The mini-metagenome contained a sufficient concentration of cyanobacterial cells for PCR amplification. In the next screening step, the 16S rRNA gene was amplified in a PCR with universal bacterial primers (Figure 3-3). The PCR products for each primer pair were pooled and cleaned from PCR buffers and reagents, and then cloned into *E. coli*.

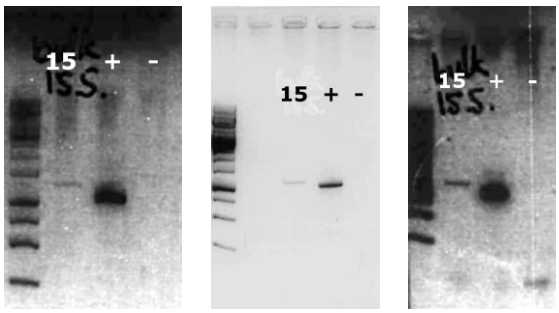


Figure 3-2 Agarose gel pictures of 16S-23S ITS region PCRs on mini-metagenome 15 with positive and negative controls in comparison to the 1kb DNA ladder. A) 2µl insert, B) 4µl insert, C) 5µl insert.

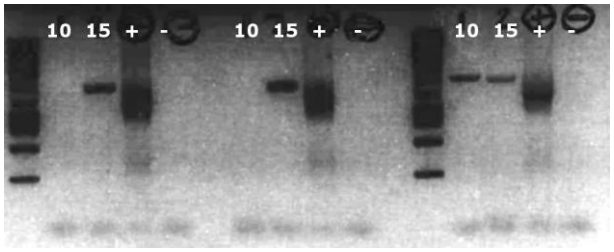


Figure 3-3 Agarose gel pictures of three replicates of the 16S rRNA gene PCR on mini-metagenome 15 (and 10) with positive and negative controls in comparison to the 1kb DNA ladder.

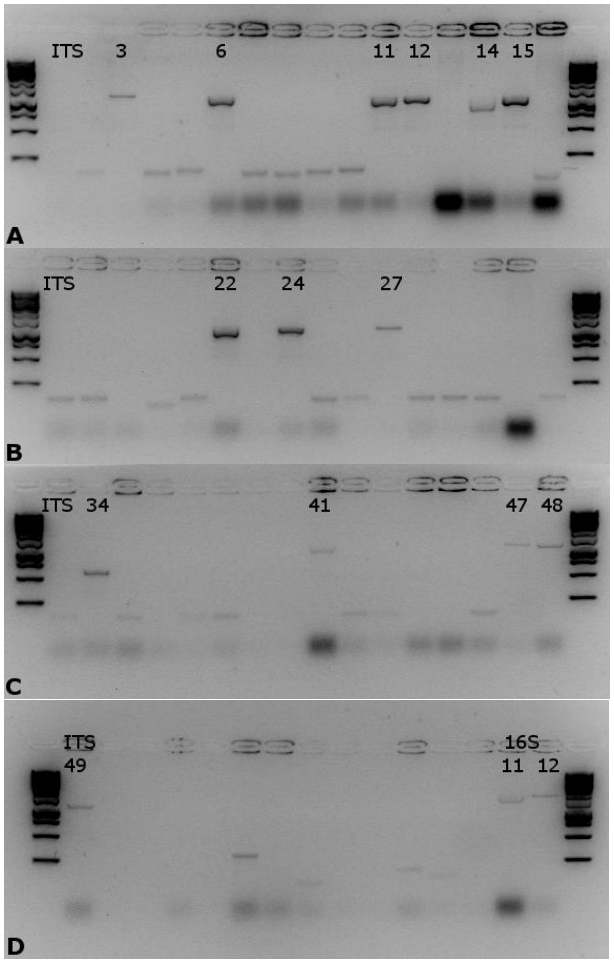


Figure 3-4 Agarose gel pictures of the colony PCR products with the 1kb DNA ladder. A)-C) and clone 49 in D) derive from the 16S-23S ITS region PCR, clones 3, 11, and 12 in D) from the 16S rRNA gene PCR. Only clones with the correct insert size were labeled.

The PCR products of clones with the correct insert size were compared to each other in a RFLP assay. For the cyanobacterial 16S-23S ITS region, the 14 clones showed five slightly different patterns (Figure 3-5). Representatives were selected for each RFLP pattern, namely clones number 6, 11, 15, 34, and 41 (Figure 3-4). For the 16S rRNA gene, only two clones had the correct insert size, both showing the same RFLP pattern. Both of them were processed further. Three more colony PCRs were performed for the selected clones, the PCR products were pooled for each clone, cleaned, and Sanger sequenced. The best BLAST hits for the sequences are all "*Ca.* Synechococcus spongiarum" (Table 3-1), confirming the purity of the FACS sorted mini-metagenome. Based on these results and on the hypothesis that the ratio of contaminating free DNA to target DNA in the mini-metagenome would be smaller than in the single cells, I focused MDA and screening efforts on the mini-metagenome.

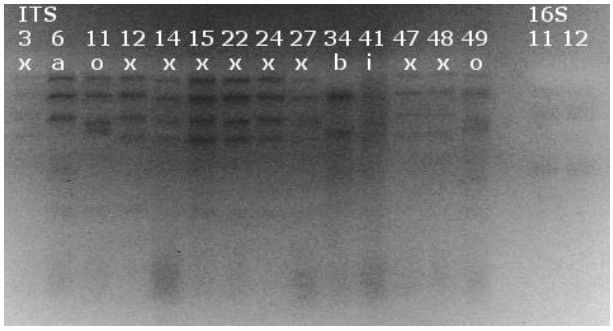


Figure 3-5 RFLP analysis of clones with inserts from the cyanobacterial 16S-23S ITS PCR and the universal bacterial 16S rRNA gene PCR . Columns are labeled according to clone numbers in the colony PCR (Figure 3-4) and the same RFLP patterns are indicated by the letters below.

Table 3-1 Best BLAST hits for colony PCR products of the clones selected based on the RFLP assay.

	Description	Accession	Query cov. (%)	Ident (%)
16S 11	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	99	99
	Candidatus Synechococcus spongiarum clone MB031NC3	EU307484.1	99	99
	Candidatus Synechococcus spongiarum clone MB031C3	EU307482.1	99	99
16S 12	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	100	99
	Candidatus Synechococcus spongiarum clone MB031NC3	EU307484.1	100	99
	Candidatus Synechococcus spongiarum clone MB031C3	EU307482.1	100	99
ITS 6	Candidatus Synechococcus spongiarum clone MB035C6	EU307487.1	68	99
	Candidatus Synechococcus spongiarum clone MB035C3	EU307486.1	68	99
	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	68	99
ITS 11	Candidatus Synechococcus spongiarum clone MB035C6	EU307487.1	99	98
	Candidatus Synechococcus spongiarum clone MB035C3	EU307486.1	99	98
	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	99	98
ITS 15	Candidatus Synechococcus spongiarum clone MB035C6	EU307487.1	66	99
	Candidatus Synechococcus spongiarum clone MB035C3	EU307486.1	66	99
	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	66	99
ITS 34	Candidatus Synechococcus spongiarum clone MB035C6	EU307487.1	88	98
	Candidatus Synechococcus spongiarum clone MB035C3	EU307486.1	88	98
	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	88	98
ITS 41	Candidatus Synechococcus spongiarum clone MB035C6	EU307487.1	99	99
	Candidatus Synechococcus spongiarum clone MB035C3	EU307486.1	99	99
	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	99	99

3.1.1.2 Identification of suitable sequencing candidates

A total of 15 MDA reactions were performed on 4µl aliquots of the mini-metagenome (named 15A-O). By 16S rRNA gene PCR, the MDA products were screened for "*Ca*. Synechococcus spongiarum." Only candidates with high 16S rRNA gene BLAST identities to "*Ca*. Synechococcus spongiarum" sequences were selected as sequencing candidates (Table 3-2). Their DNA concentrations were very similar, ranging between 10.2 ng/µl and 10.6 ng/µl, and also their absorbance ratios were in similar ranges (Table 3-3). The constant deviation from the ideal 260/280 and 260/230 ratios for pure DNA may be due to MDA reagents and buffers that are still present in the MDA products. Three candidates (15L, 15M, 15N) were chosen for sequencing from the MDA products of the mini-metagenome based on the 16S rRNA gene PCR screening.

Table 3-2 Best BLAST hits for 16S rRNA gene sequences of MDA products 15L-N.

	Description	Accession	Query cov. (%)	Ident (%)
15La	Candidatus Synechococcus spongiarum clone MB035C3	EU307486.1	100	99
	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	100	99
	Candidatus Synechococcus spongiarum clone MB031NC3	EU307484.1	100	99
15Lb	Candidatus Synechococcus spongiarum clone MB031NC1	EU307483.1	99	99
	Candidatus Synechococcus spongiarum clone MB031NC4	EU307485.1	99	99
	Candidatus Synechococcus spongiarum clone MB031NC3	EU307484.1	99	99
15M	Candidatus Synechococcus spongiarum clone 45Fr	AY190185.1	100	97
	Candidatus Synechococcus spongiarum clone MB031NC1	EU307483.1	100	97
	Uncultured cyanobacterium clone AnCha232f	EF076240.1	100	97
15N	Candidatus Synechococcus spongiarum clone 45Fr	AY190185.1	100	99
	Candidatus Synechococcus spongiarum clone MB035C3	EU307486.1	100	99
	Candidatus Synechococcus spongiarum clone MB031NC1	EU307483.1	100	99

Table 3-3 DNA concentrations measured by Qubit HS and absorbance ratios measured by NanoDrop for the 1:10 dilutions of MDA products 15L-N.

	15L	15M	15N
Qubit HS conc. (ng/µl)	10.6	10.2	10.2
NanoDrop 260/280	1.66	1.62	1.68
NanoDrop 260/230	1.93	1.85	1.99

3.1.1.3 Within-clade F comparison

The three decontaminated genomes "*Ca*. Synechococcus spongiarum" 15L-N were of very similar quality. They had between 187 and 229 contigs (>= 1000 bp) summing up to between 2.2 Mbp and 2.4 Mbp (Table 3-4). Also in N50 values, GC content, and estimated genome completeness, the three genomes were very similar. An alignment of the genomes to each other showed that they were also largely identical on nucleotide sequence level (Figure 3-6). This leads to the conclusion that the three datasets in fact represent the same genome of "*Ca*. Synechococcus spongiarum" associated to *A. aerophoba*. Therefore, no further comparisons were carried out between the three datasets, and the most complete genome 15L

was chosen for comparison to "*Ca*. Synechococcus spongiarum" clade genomes derived from other marine sponges.

Table 3-4 Assembly statistics and completeness estimation for "Candidatus Synechococcus spongiarum" genomes 15L, 15M, and 15N after decontamination by binning. ESC genes – essential single copy genes; est. – estimated.

	15L	**15M**	**15N**
Assembly statistics			
# contigs	229	187	208
# contigs (>= 1000 bp)	229	187	208
# contigs (>= 5000 bp)	136	133	147
# contigs (>= 10000 bp)	79	84	88
Largest contig (bp)	42,660	69,209	41,605
Total length (bp)	2,209,101	2,350,399	2,245,489
N50 (bp)	14,814	19,178	15,402
N75 (bp)	8,190	11,195	9,164
GC (%)	59.16	59.25	59.01
Completeness estimation			
# ESC genes (total: 111)	101	101	98
# duplicate ESC genes	2	4	2
# unique ESC genes	99	97	96
% est. completeness (111 genes)	89.19	87.39	86.49
Est. Genome size (bp)	2,476,848	2,689,551	2,596,241
Deposition in public databases			
JGI Project ID	1033525	1033528	1033531

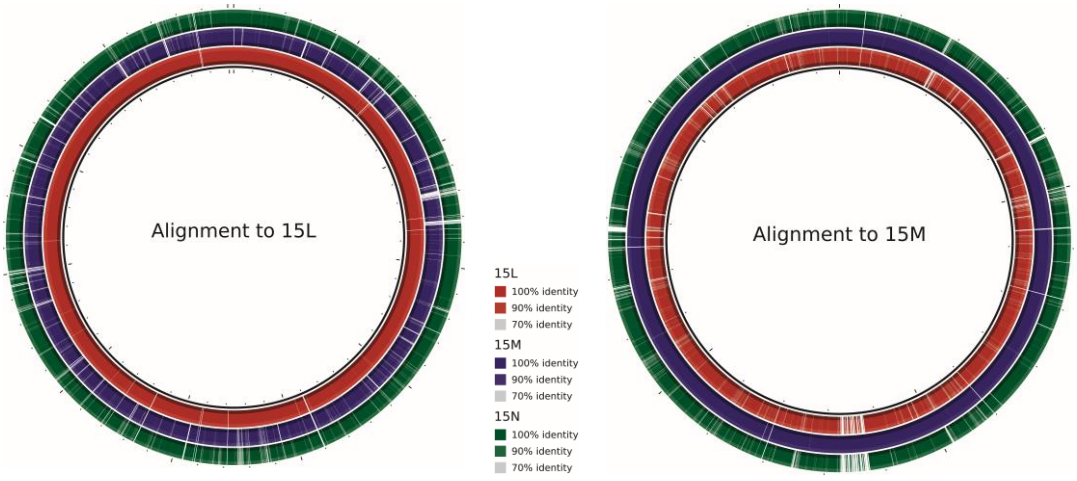


Figure 3-6 Alignment of 15L, 15M, and 15N to the most complete genome 15L and to the largest genome 15M. The three genomes are agreeing well with each other.

3.1.2 Comparison within the "*Ca.* Synechococcus spongiarum" group and to free-living references

3.1.2.1 Intraspecies phylogeny, genome recovery, and reordering

The 16S-23S ITS region phylogeny determined, that the compared "*Ca.* Synechococcus spongiarum" genomes from *A. aerophoba, I. variabilis,* and *T. swinhoei* belong to different clades of this symbiont. For "*Ca.* Synechococcus spongiarum" from *A. aerophoba* and *I. variabilis*, 16S-23S ITS sequences were published in earlier studies and the newly sequenced symbionts fell into the regarding clades, as expected (Figure 3-7). The remaining two phylotypes probably represent novel clades.

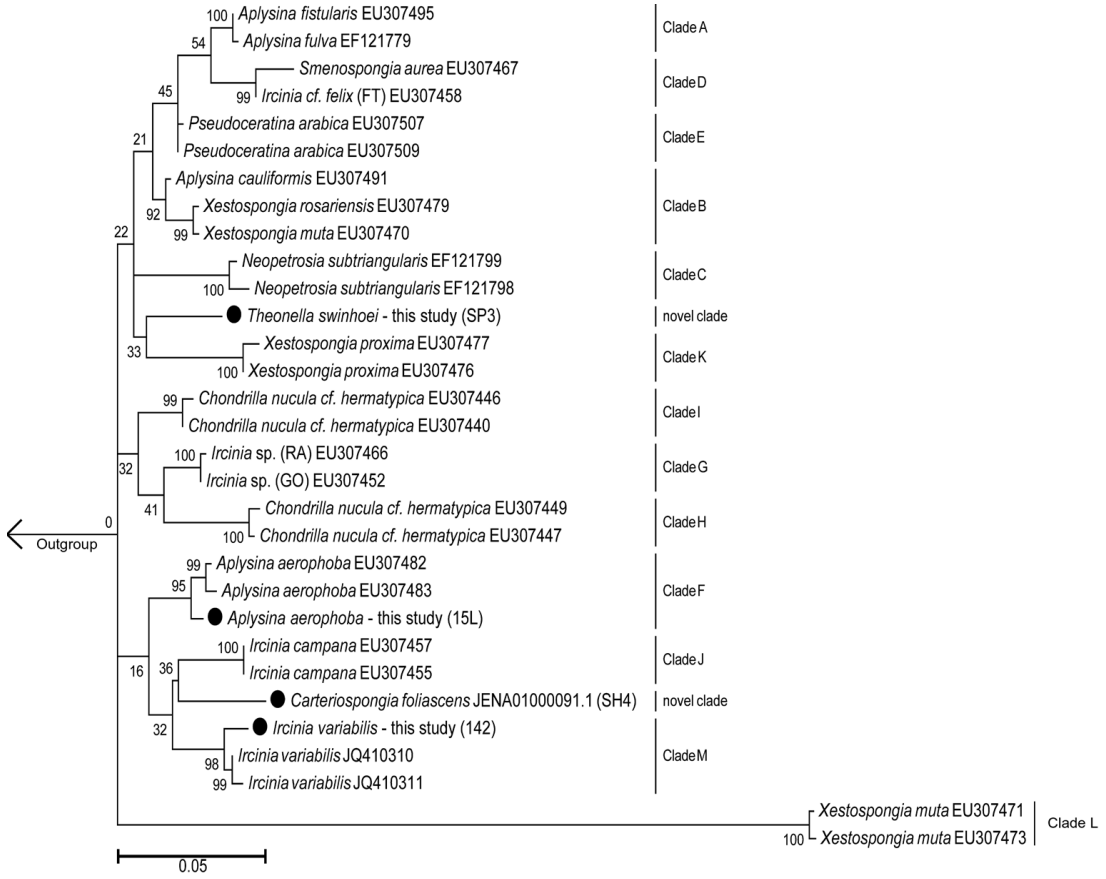


Figure 3-7 Phylogeny of the 16S-23S ITS region (and partial 16S rRNA gene) of the sponge-associated symbiont "*Ca.* Synechococcus spongiarum." Names on the tree are those of the host sponge species. Black circles mark sequences of genomes analyzed in this study. Maximum-likelihood criteria and distance estimates were calculated with the Kimura 2-parameter substitution model (+G+I). Bootstrap values at branch nodes derive from 1,000 replications.

The "*Ca.* Synechococcus spongiarum" draft genomes SP3, 142, and 15L were assembled in 117, 327, and 229 contigs representing an estimated completeness of 96%, 91%, and 95%, respectively. The previously published genome SH4 reached 89% estimated completeness (Gao *et al.*, 2014b). Genome sizes were predicted to range from ~1.9 Mbp for SH4 to ~2.5 Mbp for 142 with GC percentages of 63.1% and 58.7%, respectively (Table 3-5). Different assembly and binning approaches were used for SH4, SP3, and 142, and a single-cell sequencing approach for 15L. Also different parts of the sponge (pinacoderm with and without mesohyl) were used for DNA extraction. Despite these methodological differences, the genomes were very similar in size, contig number, completeness, and GC percentage. Methodological approach seemed to have no significant effect on the outcome.

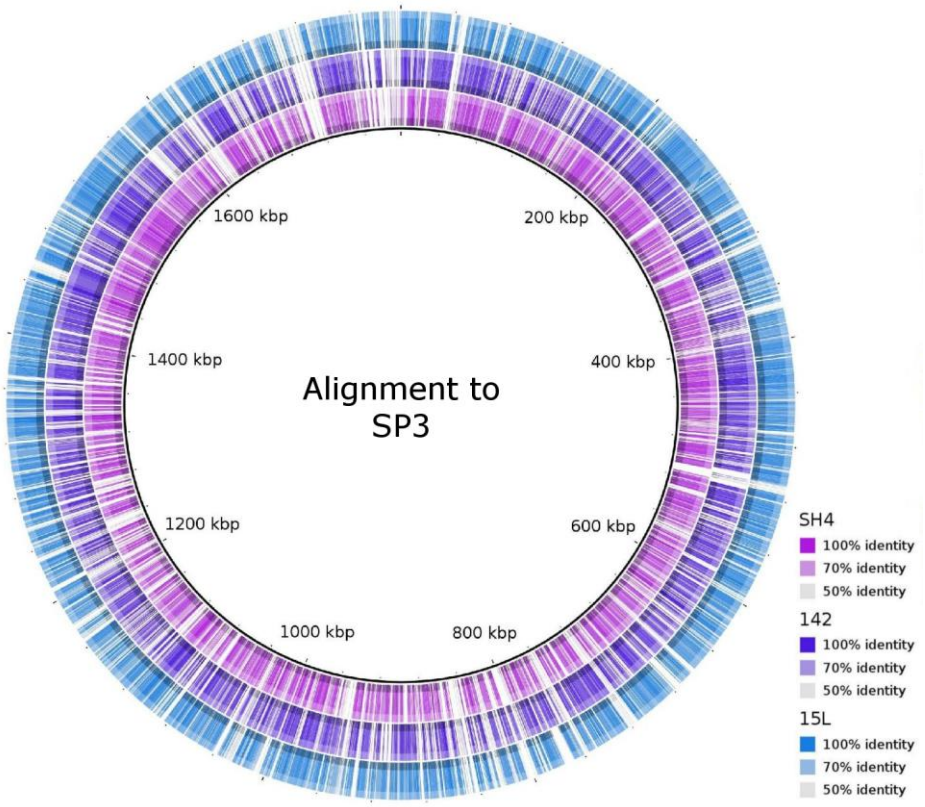


Figure 3-8 BLASTp-based alignment of four "*Ca.* Synechococcus spongiarum" genomes . The genomes of SH4, 142, and 15L were aligned to that of SP3, which showed the highest completeness and the fewest contigs.

Alignment and reordering of the draft genomes' contigs to the reference *Cyanobium gracile* PCC6307 slightly increased the number of open reading frames (ORFs) as well as annotated SEED subsystems. While this step seemed to have improved the annotation yield, the reordering does not necessarily mirror true ordering of the genomes. After initial reordering of SP3 against *Cyanobium gracile* PCC6307, the other three symbiont genomes were aligned to SP3 based on BLASTp and BLASTn (Figure 3-8 and Figure 3-9, respectively). A plot of reordered genomes SH4, 15L, and 142 against SP3 showed a high degree of gene synteny within contigs (Figure 3-10).

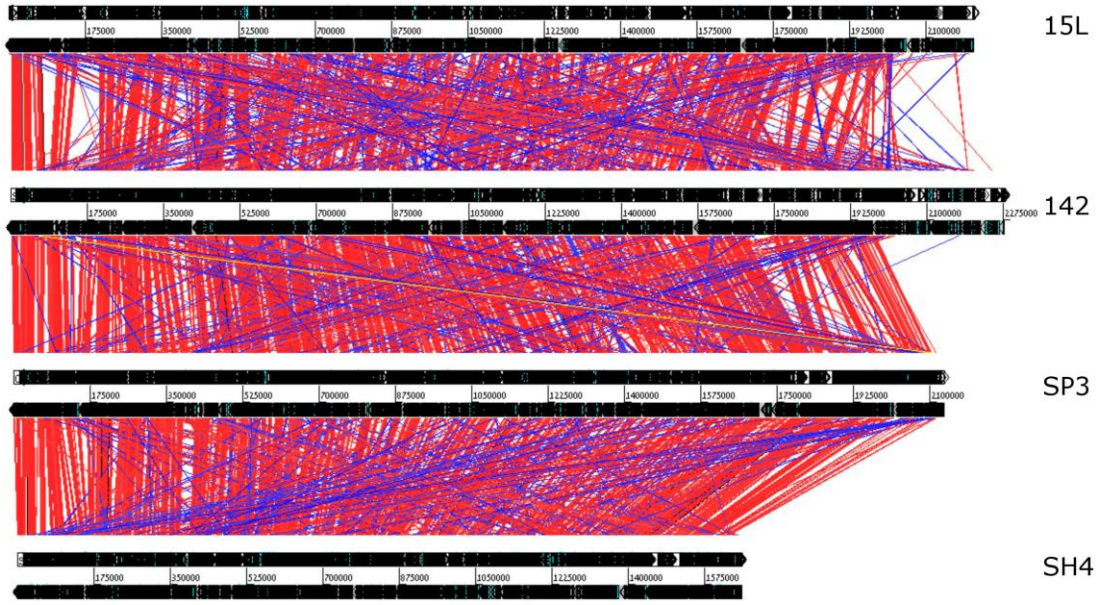


Figure 3-9 Pairwise BLASTn-based alignment of four draft genomes of "*Ca.* Synechococcus spongiarum." Bars indicate corresponding regions that are oriented in the same (red) and opposite (blue) directions.

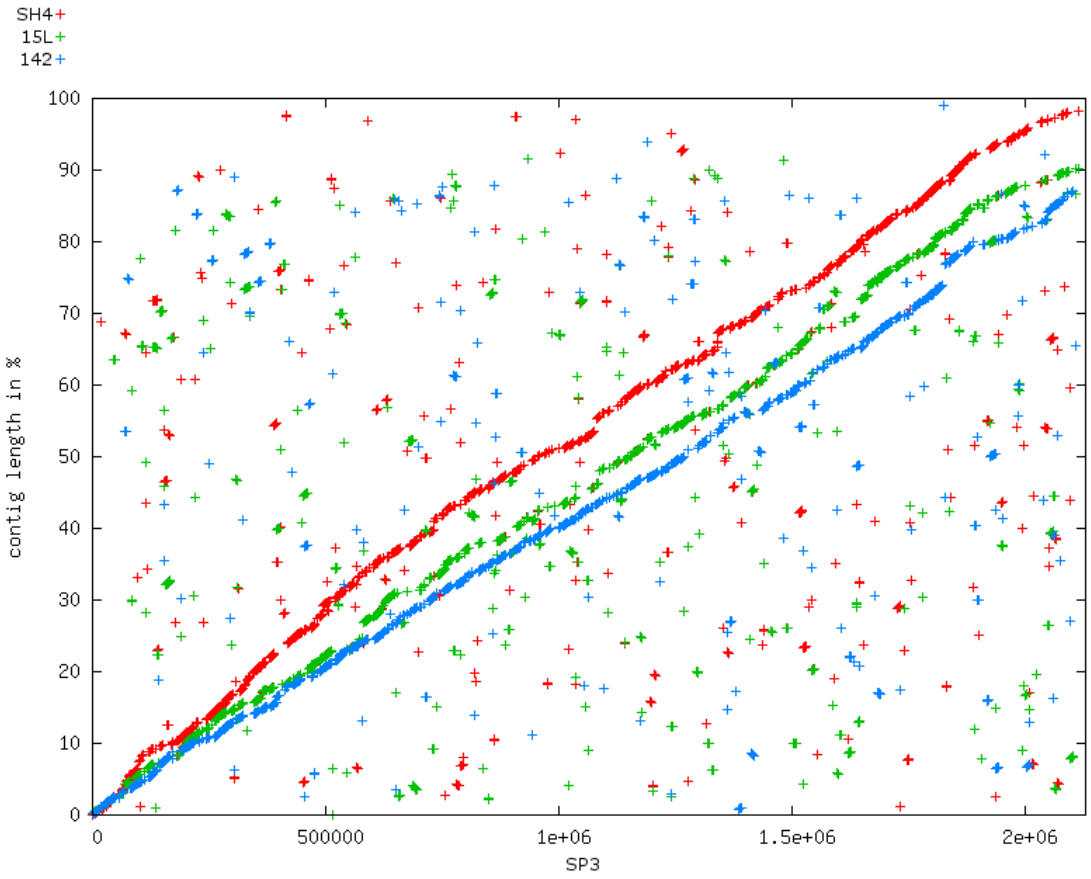


Figure 3-10 Synteny plot based on reciprocal best BLAST hits between each gene of "*Ca.* Synechococcus spongiarum" SP3 and one of the genomes SH4, 15L, and 142.

Table 3-5 General genomic information for the four "*Ca.* Synechococcus spongiarum" phylotypes 15L, SP3, 142, and SH4, and six free-living *Synechococcus* and *Cyanobium* species.

Taxon[a]	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Lifestyle (Dufresne *et al.*, 2008)	Sponge symbiont	Sponge symbiont	Sponge symbiont	Sponge symbiont	NA	Euryhaline	Coastal/ opportunist	Euryhaline	NA	Coastal/ opportunist
Salinity	NA	NA	NA	NA	Freshwater	Halotolerant	Marine	Marine	Marine	Marine
Predicted size (Mb)	2.3	2.2	2.5	1.9	3.3	3.1	2.2	2.6	2.8	2.4
Avg GC content (%)	59.2	60.9	58.7	63.1	68.7	66.0	60.8	64.5	68.7	60.2
No. of: ORFs	2260	2375	2268	1792	3220	2989	2535	2679	2756	2573
Hypothetical proteins	923	1006	994	630	1182	1125	1011	1017	992	1036
SEED functions	1337	1369	1274	1162	2038	1864	1524	1662	1764	1537
SEED subsystems	264	286	237	228	326	329	313	321	331	292
COGs	1338	1332	1230	1121	2142	1931	1542	1710	1830	1578

[a] Taxa: 1, "*Ca.* Synechococcus spongiarum" 15L (JYFQ00000000); 4, "*Ca.* Synechococcus spongiarum" SP3 (JXQG00000000), "*Ca.* Synechococcus spongiarum" 142 (JXUO00000000); 4, "*Ca.* Synechococcus spongiarum" SH4; 5, *Cyanobium gracile* PCC6307; 6, *Synechococcus* sp. strain WH5701; 7, *Synechococcus* sp. strain RCC307; 8, *Synechococcus* sp. strain RS9917; 9, *Cyanobium* sp. strain PCC7001; 10, *Synechococcus* sp. strain WH7803.

Table 3-6 Amino acid identity matrix. The mean percent identity values were based on BLAST hits between orthologous genes of the core genomes.

NAME/ID	SID_1	SID_19	SID_20	SID_18	SID_15	SID_21	SID_14	SID_16	SID_17	SID_22	SID_13	SID_9	SID_7	SID_6	SID_12	SID_11	SID_10	SID_8	SID_5
"Ca. S. spongiarum" 15L **(SID_1)**	100.00 (0.00)	92.04 (3.93)	91.45 (3.95)	90.96 (5.04)	72.42 (8.96)	72.21 (8.74)	72.18 (8.91)	71.78 (9.85)	71.68 (10.10)	71.37 (10.14)	70.76 (10.13)	70.57 (9.27)	70.35 (9.99)	67.12 (11.42)	65.25 (9.76)	65.17 (9.75)	63.57 (14.02)	63.48 (14.75)	63.47 (14.75)
"Ca. S. spongiarum" 142 **(SID_19)**	92.04 (3.93)	100.00 (0.00)	92.08 (3.62)	91.81 (4.80)	72.51 (8.88)	72.44 (8.63)	72.35 (8.84)	72.16 (8.96)	71.93 (9.91)	71.99 (9.27)	71.26 (9.16)	70.60 (9.97)	70.54 (9.92)	67.72 (10.86)	65.45 (9.72)	65.17 (10.41)	64.40 (12.45)	64.49 (12.89)	64.34 (13.35)
"Ca. S. spongiarum" SP3 **(SID_20)**	91.45 (3.95)	92.08 (3.62)	100.00 (0.00)	90.95 (5.15)	72.32 (8.95)	72.19 (8.81)	72.11 (9.00)	71.89 (9.07)	71.84 (9.99)	71.56 (9.36)	71.06 (9.36)	70.57 (9.27)	70.57 (9.24)	67.28 (10.81)	65.23 (9.81)	64.93 (10.48)	63.92 (12.98)	64.09 (13.43)	64.07 (13.41)
"*Ca.* S. spongiarum" SH4 **(SID_18)**	90.96 (5.04)	91.81 (4.80)	90.95 (5.15)	100.00 (0.00)	72.54 (8.97)	72.03 (9.71)	72.33 (9.04)	72.02 (9.07)	71.84 (9.97)	71.64 (9.36)	71.09 (9.44)	70.66 (10.05)	70.58 (10.05)	67.37 (11.59)	65.32 (9.89)	65.26 (9.89)	63.96 (13.57)	64.40 (13.08)	64.19 (13.52)
Cyanobium PCC7001 **(SID_15)**	72.42 (8.97)	72.32 (9.76)	72.32 (8.94)	72.54 (8.97)	100.00 (0.00)	83.40 (7.19)	84.59 (7.14)	79.46 (8.98)	81.21 (9.03)	80.62 (9.03)	79.59 (9.11)	78.94 (8.83)	78.90 (8.86)	73.42 (10.98)	68.21 (10.07)	67.90 (10.78)	68.49 (14.01)	68.60 (13.96)	68.49 (14.43)
Synechococcus WH5701 **(SID_21)**	72.20 (8.74)	72.21 (9.57)	72.19 (8.80)	72.03 (9.71)	83.40 (7.19)	100.00 (0.00)	83.86 (7.29)	79.23 (8.57)	80.70 (8.28)	80.05 (8.55)	79.36 (8.67)	78.95 (8.43)	78.89 (8.47)	73.15 (10.90)	68.13 (9.97)	67.82 (10.68)	68.36 (13.69)	68.34 (14.14)	68.40 (13.68)
C. gracile PCC6307 **(SID_14)**	72.18 (8.90)	72.10 (9.75)	72.11 (9.00)	72.33 (8.94)	84.59 (7.14)	83.86 (7.29)	100.00 (0.00)	79.15 (8.87)	80.74 (8.48)	80.18 (8.68)	79.22 (8.89)	78.75 (8.55)	78.67 (8.55)	73.04 (11.65)	68.02 (10.06)	67.93 (10.04)	68.21 (13.78)	68.49 (13.79)	68.48 (13.79)
Synechococcus RCC307 **(SID_16)**	71.92 (9.12)	72.16 (8.96)	71.89 (9.07)	72.02 (9.07)	79.46 (8.98)	79.23 (8.57)	79.15 (8.87)	100.00 (0.00)	79.23 (9.46)	78.97 (9.66)	78.41 (9.72)	77.35 (9.24)	77.24 (9.23)	72.49 (11.19)	67.60 (10.02)	67.53 (10.01)	68.34 (12.95)	68.50 (13.04)	68.53 (13.12)
Synechococcus RS9917 **(SID_17)**	71.91 (9.24)	72.18 (9.02)	72.06 (9.10)	72.07 (9.09)	81.21 (9.03)	80.70 (8.27)	80.73 (8.48)	79.23 (9.46)	100.00 (0.00)	87.66 (7.49)	86.17 (7.39)	83.38 (7.82)	83.46 (7.54)	75.71 (10.66)	67.73 (10.08)	67.66 (10.08)	70.32 (12.81)	70.53 (12.68)	70.44 (12.74)
Synechococcus WH7803 **(SID_22)**	71.42 (10.07)	71.85 (10.03)	71.42 (10.11)	71.48 (10.14)	80.62 (9.03)	80.05 (8.55)	80.18 (8.68)	78.97 (9.66)	87.66 (7.49)	100.00 (0.00)	86.92 (7.11)	82.44 (8.09)	82.59 (7.83)	75.41 (11.47)	67.44 (10.10)	67.37 (10.10)	70.18 (13.24)	70.48 (13.21)	70.42 (13.33)
Synechococcus CC9311 **(SID_13)**	70.76 (10.13)	71.26 (9.16)	71.06 (9.35)	71.09 (9.44)	79.59 (9.11)	79.36 (8.67)	79.22 (8.89)	78.41 (9.72)	86.17 (7.39)	86.92 (7.11)	100.00 (0.00)	82.08 (8.10)	82.15 (7.93)	75.44 (10.61)	67.01 (9.99)	66.93 (9.99)	70.20 (12.56)	70.35 (12.49)	70.30 (12.57)
P. marinus MIT9313 **(SID_9)**	70.57 (9.27)	70.56 (9.98)	70.57 (9.27)	70.79 (9.37)	78.94 (8.83)	78.95 (8.43)	78.75 (8.55)	77.35 (9.24)	83.38 (7.82)	82.44 (8.09)	82.08 (8.10)	100.00 (0.00)	98.14 (3.24)	76.82 (10.30)	66.38 (9.98)	66.30 (9.98)	70.49 (12.53)	70.79 (12.46)	70.74 (12.53)
P. marinus MIT9303 **(SID_7)**	70.48 (9.30)	70.49 (9.93)	70.57 (9.24)	70.72 (9.36)	78.90 (8.86)	78.89 (8.47)	78.68 (8.55)	77.24 (9.23)	83.46 (7.54)	82.59 (7.83)	82.15 (7.93)	98.14 (3.24)	100.00 (0.00)	76.85 (10.25)	66.39 (9.98)	66.31 (9.92)	70.48 (12.52)	70.77 (12.39)	70.71 (12.44)
P. marinus CCMP1375 **(SID_6)**	67.12 (11.42)	67.48 (11.52)	67.28 (10.81)	67.32 (11.66)	73.42 (10.98)	73.15 (10.90)	73.21 (10.94)	72.49 (11.19)	75.71 (10.66)	75.61 (10.66)	75.44 (10.61)	76.82 (10.30)	76.85 (10.25)	100.00 (0.00)	64.15 (10.83)	64.07 (10.82)	72.48 (12.34)	72.74 (12.28)	72.69 (12.27)
S. elongatus PCC7942 **(SID_12)**	65.25 (9.76)	65.45 (9.72)	65.23 (9.81)	65.32 (9.89)	68.21 (10.07)	68.13 (9.97)	68.02 (10.06)	67.60 (10.01)	67.73 (10.08)	67.44 (10.10)	67.01 (9.99)	66.38 (9.98)	66.39 (9.93)	64.15 (10.83)	100.00 (0.00)	99.88 (0.29)	62.16 (11.48)	62.31 (11.26)	62.34 (11.36)
S. elongatus PCC6301 **(SID_11)**	65.17 (9.75)	65.39 (9.72)	65.16 (9.80)	65.26 (9.89)	68.12 (10.05)	68.04 (9.95)	67.93 (10.04)	67.53 (10.00)	67.66 (10.07)	67.37 (10.10)	66.93 (9.99)	66.31 (9.97)	66.31 (9.92)	64.07 (10.82)	99.88 (0.29)	100.00 (0.00)	62.10 (11.48)	62.27 (11.29)	62.27 (11.35)
P. marinus MIT9515 **(SID_10)**	63.85 (13.15)	63.53 (14.78)	63.96 (12.62)	64.23 (12.68)	68.65 (13.50)	68.53 (13.16)	68.51 (12.76)	68.34 (12.95)	70.32 (12.81)	70.21 (13.19)	70.20 (12.56)	70.49 (12.53)	70.48 (12.52)	72.48 (12.34)	62.16 (11.48)	62.10 (11.48)	100.00 (0.00)	87.93 (7.15)	87.88 (7.37)
P. marinus MIT9312 **(SID_8)**	64.07 (13.01)	63.92 (14.34)	64.08 (13.44)	64.22 (13.57)	68.76 (13.44)	68.81 (12.58)	68.81 (12.75)	68.50 (13.04)	70.53 (12.68)	70.52 (13.15)	70.35 (12.49)	70.80 (12.46)	70.78 (12.39)	72.74 (12.28)	62.31 (11.26)	62.05 (11.84)	87.93 (7.15)	100.00 (0.00)	94.86 (4.47)
P. marinus AS9601 **(SID_5)**	63.60 (14.35)	63.58 (15.18)	63.75 (14.32)	63.83 (14.45)	68.79 (13.45)	68.72 (12.61)	68.67 (13.24)	68.53 (13.12)	70.44 (12.74)	70.46 (13.26)	70.30 (12.57)	70.74 (12.53)	70.71 (12.44)	72.69 (12.27)	62.34 (11.36)	61.89 (12.39)	87.88 (7.37)	94.86 (4.47)	100.00 (0.00)

3.1.2.2 Within-symbiont and symbiont-reference comparison

Six representative closely related, free-living *Synechococcus* and *Cyanobium* species were selected for comparison to the four "*Ca.* Synechococcus spongiarum" genomes (reference genomes marked in green in Figure 3-11). "*Ca.* Synechococcus spongiarum" is in the concatenated phylogenetic core genome tree equidistant from the *Synechococcus/Prochlorococcus* subclade consisting of marine and freshwater *Synechococcus*, *Prochlorococcus*, and *Cyanobium*, which is in agreement with earlier reports (Gao *et al.*, 2014b).

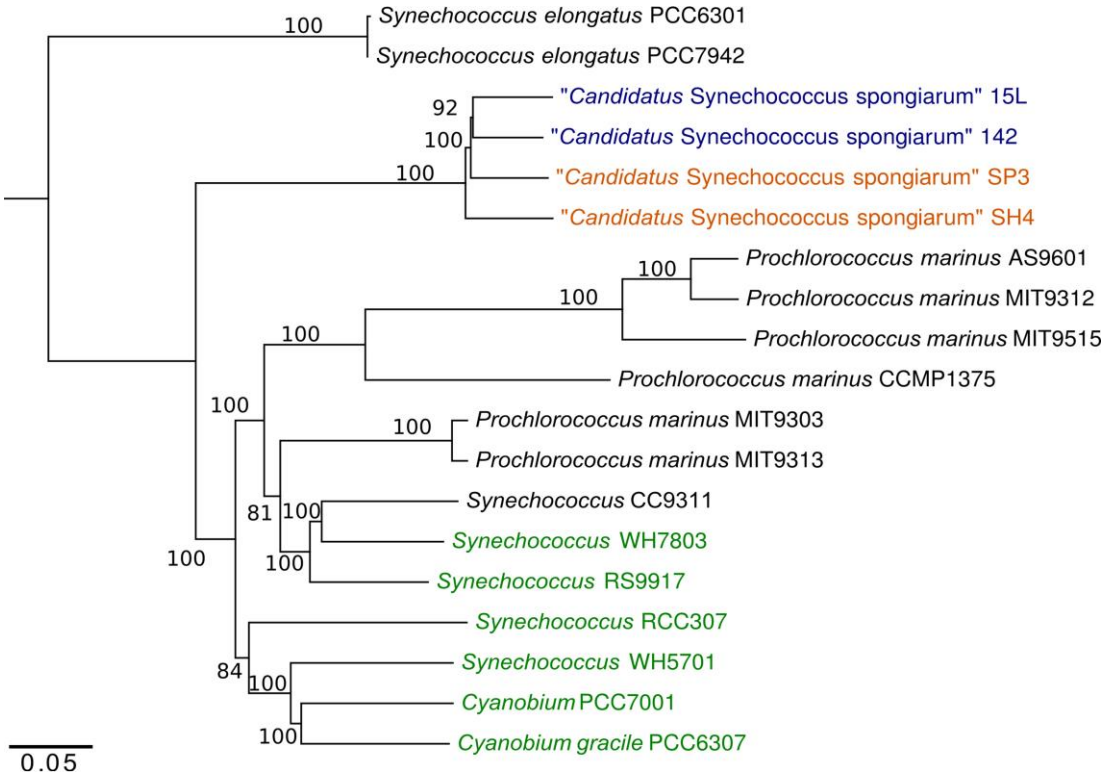


Figure 3-11 Concatenated phylogenetic core genome tree calculated by iterative pairwise comparison of genomes of the cyanobacteria analyzed here. Bootstrap values at branch nodes derive from 100 replications (Kimura distance matrix, neighbor joining algorithm). Names in orange and blue are "*Ca.* Synechococcus spongiarum" associated with Red Sea and Mediterranean sponges, respectively; those in green are free-living strains used for genomic comparisons.

An amino acid identity comparison between shared orthologous genes showed, that the four symbionts were between 91.0% and 92.1% identical regarding these shared genes, while they were between 63.6% and 72.5% similar to the six free-living cyanobacteria (Table 3-6). The symbionts were most similar to the marine *Cyanobium* PCC7001 and to the freshwater *Cyanobium gracile* PCC6307 with 72.4% and 72.2% mean amino acid identity, respectively.

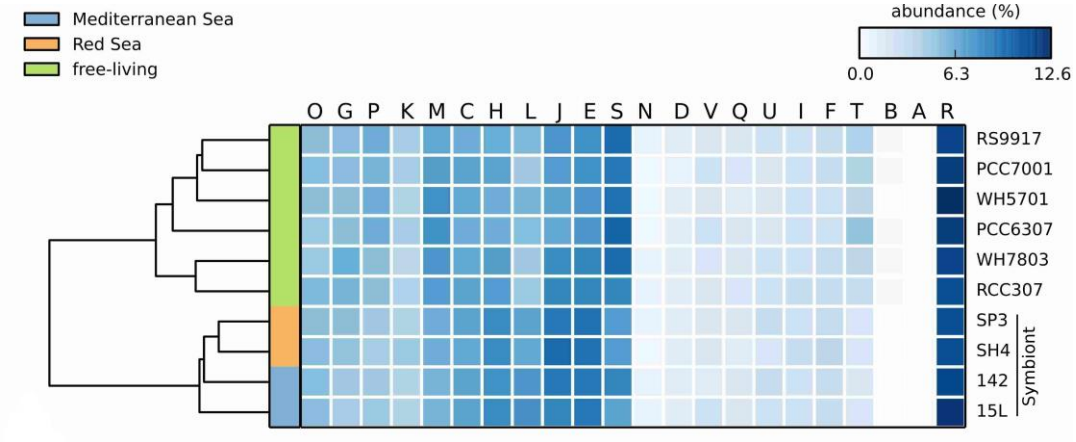
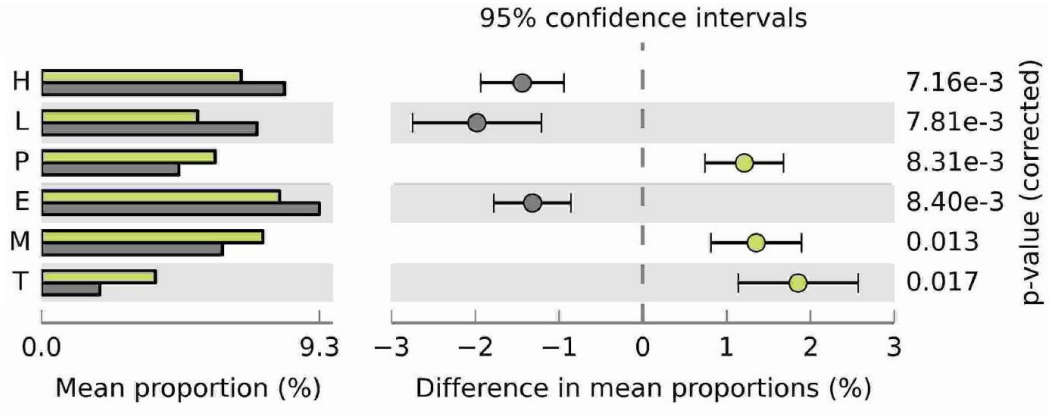


Figure 3-12 Heatmap of relative abundances of COG classes A to V. Two Mediterranean "*Ca.* Synechococcus spongiarum" genomes (blue), two Red Sea "*Ca.* Synechococcus spongiarum" genomes (orange), and six genomes of free-living cyanobacteria (green) were compared in this analysis. UPGMA clustering is presented to the left of the map.

A total of 1,759 COGs were annotated for the "*Ca.* Synechococcus spongiarum" genomes and references. Based on COG class abundances, the four symbionts were more similar to each other than to the free-living cyanobacteria, and also clustered by geographic location (Figure 3-12). The clear separation of symbionts and references was mainly due to six COG categories – three of which were significantly enriched in the symbionts, with three others depleted (Figure 3-13). Found in "*Ca.* Synechococcus spongiarum" at significantly higher proportions were COGs of the categories 'replication, recombination and repair' (L), 'coenzyme transport and metabolism' (H), and 'amino acid transport and metabolism' (E), whereas COGs of the categories 'signal transduction mechanisms' (T), 'cell wall/membrane/envelope biogenesis' (M), and 'inorganic ion transport and metabolism' (P) were depleted compared to the free-living relatives.



H - Coenzyme transport and metabolism
L - Replication, recombination and repair
P - Inorganic ion transport and metabolism
E - Amino acid transport and metabolism
M - Cell wall/membrane/envelope biogenesis
T - Signal transduction mechanisms

Figure 3-13 COG classes with statistically significant differences between "Ca. Synechococcus spongiarum" genomes (grey) and genomes of free-living cyanobacteria (green). Error bars indicate within-group standard deviations. Presented categories passed a corrected *P* value of <0.05 in Welch's *t* test.

Table 3-7 COGs unique to "*Ca.* Synechococcus spongiarum" – i.e. present in at least one of the four symbiont genomes but absent in all six free-living cyanobacteria. The 14 COGs present in all four "*Ca.* Synechococcus spongiarum" genomes are in bold.

#COG	SP3	142	15L	SH4	COG description	COG class
COG0003	0	1	0	0	Oxyanion-translocating ATPase	P
COG0067	0	0	0	1	Glutamate synthase domain 1	E
COG0070	0	0	0	1	Glutamate synthase domain 3	E
COG0270	**1**	**1**	**2**	**1**	**Site-specific DNA methylase**	**L**
COG0323	0	0	1	0	DNA mismatch repair enzyme (predicted ATPase)	L
COG0338	1	2	1	0	Site-specific DNA methylase	L
COG0423	0	0	1	0	Glycyl-tRNA synthetase (class II)	J
COG0433	1	1	1	0	Predicted ATPase	R
COG0501	0	0	1	0	Zn-dependent protease with chaperone function	O
COG0517	**1**	**1**	**1**	**1**	**FOG: CBS domain**	**R**
COG0609	1	0	1	0	ABC-type Fe3+-siderophore transport system, permease component	P
COG0646	0	1	0	0	Methionine synthase I (cobalamin-dependent), methyltransferase domain	E
COG0666	**4**	**4**	**4**	**4**	**FOG: Ankyrin repeat**	**R**
COG0675	4	6	7	0	Transposase and inactivated derivatives	L
COG0716	1	1	1	0	Flavodoxins	C
COG0846	0	0	1	0	NAD-dependent protein deacetylases, SIR2 family	K
COG0849	1	0	0	1	Actin-like ATPase involved in cell division	D
COG0863	**5**	**4**	**9**	**3**	**DNA modification methylase**	**L**
COG1002	0	0	3	2	Type II restriction enzyme, methylase subunits	V
COG1106	3	3	6	0	Predicted ATPases	R
COG1111	0	1	0	0	ERCC4-like helicases	L
COG1120	1	0	1	0	ABC-type cobalamin/Fe3+-siderophores transport systems, ATPase components	PH
COG1146	0	1	0	0	Ferredoxin	C
COG1203	0	1	2	1	Predicted helicases	R
COG1204	0	1	1	0	Superfamily II helicase	R
COG1223	0	0	1	0	Predicted ATPase (AAA+ superfamily)	R
COG1304	0	1	1	1	L-lactate dehydrogenase (FMN-dependent) and related alpha-hydroxy acid dehydrogenases	C
COG1331	0	0	1	0	Highly conserved protein containing a thioredoxin domain	O
COG1336	0	1	0	0	Uncharacterized protein predicted to be involved in DNA repair (RAMP superfamily)	L
COG1343	0	1	1	0	Uncharacterized protein predicted to be involved in DNA repair	L
COG1360	1	0	0	0	Flagellar motor protein	N
COG1451	1	1	0	1	Predicted metal-dependent hydrolase	R
COG1468	1	0	0	0	RecB family exonuclease	L
COG1476	0	0	0	1	Predicted transcriptional regulators	K
COG1479	0	1	0	1	Uncharacterized conserved protein	S
COG1483	0	2	1	1	Predicted ATPase (AAA+ superfamily)	R
COG1518	1	2	1	0	Uncharacterized protein predicted to be involved in DNA repair	L
COG1604	0	1	0	0	Uncharacterized protein predicted to be involved in DNA repair (RAMP superfamily)	L
COG1629	**3**	**1**	**3**	**1**	**Outer membrane receptor proteins, mostly Fe transport**	**P**

COG1651	**1**	**1**	**1**	**1**	**Protein-disulfide isomerase**	**O**
COG1743	**1**	**1**	**2**	**1**	**Adenine-specific DNA methylase containing a Zn-ribbon**	**L**
COG1744	2	1	1	0	Uncharacterized ABC-type transport system, periplasmic component/surface lipoprotein	R
COG1769	0	1	0	0	Uncharacterized protein predicted to be involved in DNA repair (RAMP superfamily)	L
COG1879	0	1	0	0	ABC-type sugar transport system, periplasmic component	G
COG1893	1	0	1	0	Ketopantoate reductase	H
COG1943	0	1	3	0	Transposase and inactivated derivatives	L
COG2110	1	0	0	0	Predicted phosphatase homologous to the C-terminal domain of histone macroH2A1	R
COG2141	0	0	1	0	Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases	C
COG2189	**3**	**1**	**3**	**1**	**Adenine specific DNA methylase Mod**	**L**
COG2241	0	0	1	0	Precorrin-6B methylase 1	H
COG2253	0	0	0	1	Uncharacterized conserved protein	S
COG2340	1	1	0	0	Uncharacterized protein with SCP/PR1 domains	S
COG2520	0	1	0	0	Predicted methyltransferase	R
COG2608	0	0	1	0	Copper chaperone	P
COG2810	0	0	1	1	Predicted type IV restriction endonuclease	V
COG2832	0	1	0	0	Uncharacterized protein conserved in bacteria	S
COG2856	2	2	2	0	Predicted Zn peptidase	E
COG2910	0	0	1	0	Putative NADH-flavin reductase	R
COG2932	1	0	1	1	Predicted transcriptional regulator	K
COG3041	0	1	2	0	Uncharacterized protein conserved in bacteria	S
COG3064	1	0	1	0	Membrane protein involved in colicin uptake	M
COG3106	1	1	0	0	Predicted ATPase	R
COG3150	1	0	0	0	Predicted esterase	R
COG3290	0	0	1	0	Signal transduction histidine kinase regulating citrate/malate metabolism	T
COG3293	4	2	0	3	Transposase and inactivated derivatives	L
COG3337	0	1	0	0	Uncharacterized protein predicted to be involved in DNA repair	L
COG3344	0	0	1	0	Retron-type reverse transcriptase	L
COG3392	0	2	1	1	Adenine-specific DNA methylase	L
COG3464	0	0	1	0	Transposase and inactivated derivatives	L
COG3512	0	1	0	0	Uncharacterized protein conserved in bacteria	S
COG3513	0	1	0	0	Uncharacterized protein conserved in bacteria	S
COG3549	1	0	0	1	Plasmid maintenance system killer protein	R
COG3587	0	0	0	1	Restriction endonuclease	V
COG3607	0	0	1	0	Predicted lactoylglutathione lyase	R
COG3668	0	0	0	1	Plasmid stabilization system protein	R
COG3705	**1**	**1**	**1**	**1**	**ATP phosphoribosyltransferase involved in histidine biosynthesis**	**E**
COG3727	0	0	1	0	DNA G:T-mismatch repair endonuclease	L
COG3768	1	1	0	0	Predicted membrane protein	S
COG3848	0	1	0	0	Phosphohistidine swiveling domain	T
COG3881	0	0	1	0	Uncharacterized protein conserved in bacteria	S
COG3893	0	0	1	0	Inactivated superfamily I helicase	L
COG3898	0	0	0	1	Uncharacterized membrane-bound protein	S

COG3950	**2**	**6**	**4**	**1**	**Predicted ATP-binding protein involved in virulence**	**R**
COG4122	**1**	**1**	**2**	**1**	**Predicted O-methyltransferase**	**R**
COG4123	0	0	1	1	Predicted O-methyltransferase	R
COG4278	0	0	0	1	Uncharacterized conserved protein	S
COG4422	1	0	2	0	Bacteriophage protein gp37	S
COG4558	1	0	1	0	ABC-type hemin transport system, periplasmic component	P
COG4564	1	0	1	0	Signal transduction histidine kinase	T
COG4623	0	0	1	0	Predicted soluble lytic transglycosylase fused to an ABC-type amino acid-binding protein	M
COG4694	0	0	1	0	Uncharacterized protein conserved in bacteria	S
COG4717	0	1	0	1	Uncharacterized conserved protein	S
COG4725	0	0	1	1	Transcriptional activator, adenine-specific DNA methyltransferase	TK
COG4733	0	0	1	0	Phage-related protein, tail component	S
COG4748	**1**	**2**	**1**	**2**	**Uncharacterized conserved protein**	**S**
COG4771	0	1	0	0	Outer membrane receptor for ferrienterochelin and colicins	P
COG4823	0	0	1	0	Abortive infection bacteriophage resistance protein	V
COG4886	3	13	0	1	Leucine-rich repeat (LRR) protein	S
COG4889	**1**	**3**	**6**	**2**	**Predicted helicase**	**R**
COG4923	1	1	0	0	Uncharacterized conserved protein	S
COG4928	0	1	0	0	Predicted P-loop ATPase	R
COG4938	2	4	3	0	Uncharacterized conserved protein	S
COG4942	0	0	1	0	Membrane-bound metallopeptidase	D
COG4978	1	0	1	0	Transcriptional regulator, effector-binding domain/component	KT
COG4982	1	0	0	0	3-oxoacyl-[acyl-carrier protein] reductase	I
COG5009	1	0	0	0	Membrane carboxypeptidase/penicillin-binding protein	M
COG5011	0	0	1	0	Uncharacterized protein conserved in bacteria	S
COG5244	1	0	0	0	Dynactin complex subunit involved in mitotic spindle partitioning in anaphase B	D
COG5395	**1**	**1**	**1**	**1**	**Predicted membrane protein**	**S**
COG5480	0	1	0	0	Predicted integral membrane protein	S
COG5483	0	0	1	1	Uncharacterized conserved protein	S
COG5507	0	0	1	0	Uncharacterized conserved protein	S

Approximately one third of the annotated COGs were present in all ten analyzed genomes and are thereby interpreted as an essential functional core. The free-living cyanobacteria had a total of 581 COGs missing in the symbionts, of which 105 were found in all six genomes. On the other hand, the four "*Ca.* Synechococcus spongiarum" genomes had 112 COGs missing in their free-living relatives, 14 of which were shared by all four symbionts (Table 3-7). Four of these shared symbiont-specific genes were methylases (COG2189, COG1743, COG0863, and COG0270) and are assigned to COG class L. Two symbiont-enriched COGs were ankyrin and leucine-rich repeat proteins (COG0666 and COG4886, respectively). While all four "*Ca.* Synechococcus spongiarum" genomes contained four copies

of COG0666 each, they contained COG4886 in different amounts. COG4886 was not annotated in 15L. The outer membrane receptor protein COG1629 (K02014) was annotated in all four symbionts. This COG and also COG4771 annotated adjacent to it in 142 are related to TonB-dependent siderophore receptors. In the reference genomes, this iron-sensing pathway (K02014) was absent, whereas the symbionts had the potential for iron-sensing and contained large protein conglomerations related to the K02014 pathway (SP3 and 15L) comprising a number of ABC-type transport systems (COG4558, COG0609/K02015, and COG1120/K02013). The KEGG annotation confirmed the annotation of these genes by COG.

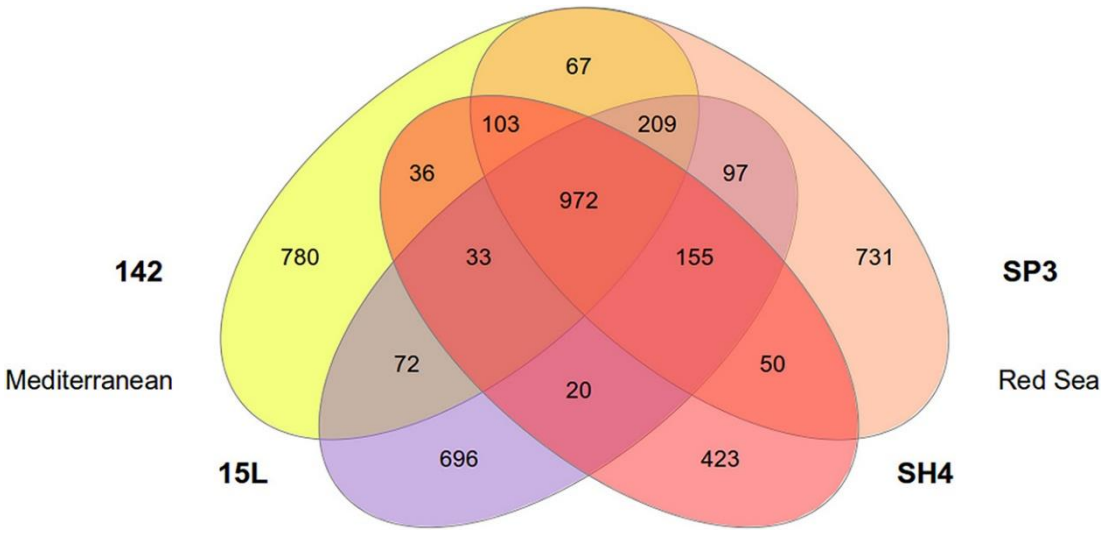


Figure 3-14 Venn diagram comparing the gene inventories of four "*Ca.* Synechococcus spongiarum" genomes computed by EDGAR (Blom *et al.*, 2009) based on reciprocal best BLAST hits of the coding sequences predicted by RAST (Aziz *et al.*, 2008). SH4 and SP3 are symbionts of Red Sea sponges, and 15L and 142 are symbionts of Mediterranean sponges.

The majority (40) of the 105 COGs missing in the symbionts but abundant in the free-living references belonged to COG classes 'general function prediction only' (R) and 'unknown function' (S). Nine, eight, six, and five COGs belonged to the classes 'replication, recombination, and repair' (L), 'cell wall/membrane/envelope biogenesis' (M), 'inorganic ion transport and metabolism' (P), and 'translation, ribosomal structure and biogenesis' (J), respectively. STRING networks revealed a possible link between five COGs of class M and one of class G encoding for genes involved in the production of L-Rhamnose which is an important residue of the O antigen of lipopolysaccharides (LPS) in Gram-negative bacteria (Snyder *et al.*, 2009). Also the RAST annotation confirmed the lack of these genes in the symbiont genomes. The genomes of the "*Ca.* Synechococcus spongiarum" clades were also characterized by smaller numbers of genes involved in several essential functions such as signal transduction (COG0642) or carbohydrate transport and metabolism (COG1175, COG9363, COG0366), (Table 3-8).

Table 3-8 Reduction in the number of genes related to essential COG functions in four genomes of "*Ca.* Synechococcus spongiarum" compared to six genomes of free-living cyanobacteria and the plastid of amoeboid *P. chromatophora*.

	Taxon[a]											
COG#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**COG annotation**
COG0642	1	1	1	1	1	8	6	6	5	6	7	Signal transduction histidine kinase
COG0745	5	5	5	4	3	10	10	9	10	10	10	Two component system response regulator, OMP-R family
COG0664	2	2	2	1	0	12	8	4	10	9	6	Transcriptional regulator, CPR family
COG0226	1	1	1	1	1	3	5	4	5	3	4	ABC-type phosphate transport
COG3239	1	1	1	1	1	5	4	3	3	3	4	Beta-carotene hydroxylase, carotenoid biosynthesis
COG0845	1	1	1	1	0	3	4	4	3	3	3	Secretion protein, HlyD family
COG0695	1	1	1	1	0	4	1	2	4	3	2	Glutaredoxin 3
COG0204	1	1	1	1	2	4	3	2	3	3	2	1-acyl-sn-glycerol-3-phosphate acyltransferase
COG0415	1	1	1	1	0	3	3	3	3	3	3	Deoxyribodipyrimidine photolyase
COG1233	1	1	1	1	1	3	3	3	3	3	3	Carotenoid isomerase, carotenoid biosynthesis
COG0366	1	1	1	1	0	3	4	2	4	2	2	Sucrose phosphorylase
COG0124	1	1	1	1	2	2	2	2	2	2	2	Histidyl-tRNA synthetase
COG0042	1	1	1	1	1	2	2	2	2	2	2	tRNA-dihydrouridine synthase A
COG0229	1	1	1	1	0	2	2	2	2	2	2	Peptide-methionine (R)-S-oxide reductase
COG0363	1	1	1	1	0	2	2	2	2	2	2	6-phosphogluconolactonase, Pentose phosphate pathway
COG0459	1	1	1	1	1	2	2	2	2	2	2	Chaperonin GroEL (HSP60 family), RNA degradation
COG0488	1	1	1	1	1	2	2	2	2	2	2	ATP-binding protein of ABC transporter
COG0843	1	1	1	1	1	2	2	2	2	2	2	Cytochrome c oxidase subunit I, Oxidative phosphorylation
COG1175	1	1	1	1	0	2	2	2	2	2	2	Lactose/L-arabinose transport system permease protein, ABC transporters
COG1186	1	1	1	1	1	2	2	2	2	2	2	Peptide chain release factor RF-2,
COG1187	1	1	1	1	1	2	2	2	2	2	2	Ribosomal small subunit pseudouridine synthase A
COG1622	1	1	1	1	1	2	2	2	2	2	2	Cytochrome c oxidase subunit II, Oxidative phosphorylation
COG1845	1	1	1	1	1	2	2	2	2	2	2	Cytochrome c oxidase subunit III, Oxidative phosphorylation

[a] Taxa: 1 – "*Ca.* Synechococcus spongiarum" SP3, 2- "*Ca.* Synechococcus spongiarum" 142, 3 – "*Ca.* Synechococcus spongiarum" 15L, 4 – "*Ca.* Synechococcus spongiarum" SH4, 5 – *P. chromatophora* plastid, 6 – *C. gracile* PCC6307, 7 – *Synechococcus* sp. strain WH5701, 8 – *Synechococcus* sp. strain RCC307, 9 – *Synechococcus* sp. strain RS9917, 10 – *Cyanobium* sp. PCC7001, 11 – *Synechococcus* sp. strain WH7803.

While the pangenome of all four symbionts determined with EDGAR spanned 3,746 genes, their core genome consisted of a mere 972 genes (Figure 3-14), 173 of which were absent from the reference genomes. These genes may represent symbiotic features unique to "*Ca.* Synechococcus spongiarum." In terms of the COG annotation of these genes, only three were unique to the symbionts, namely COG5395, COG1651, and COG2932 (Table 3-9). COG5395 is a predicted membrane protein of unknown function that belongs to superfamily DUF2306. COG1651, encoding a disulfide interchange protein, was annotated in freshwater *Synechococcus* strain JA33. The same gene, according to EDGAR, received different COG

annotations in the symbionts: COG2932 for 15L, SP3, and SH4, and COG0681 (signal peptidase I) for 142. After filtering out all genes with COG annotations not unique to the symbionts, 78 genes remained, 75 of which without COG. Fourty nine of them did also not get a KEGG annotation and were hypothetical genes according to SEED. The remaining genes contained a tetratricopeptide repeat (TPR, K07280), the phycoerythrin-associated proteins K05380 and K05279, the phycocyanobilin:ferredoxin oxidoreductase K05371, allophycocyanin subunit K02092, and the nickel-dependent superoxide dismutase EC 1.15.1.1.

Table 3-9 Potential symbiotic genes in "*Ca.* Synechococcus spongiarum" genomes. They were found to be orthologous and unique to the four symbiont genomes. Genes are described according to the SEED annotation in their respective genomes. NA – not available.

	15L	**SH4**	**142**	**SP3**	**COG**
1	27kDa outer membrane protein	27kDa outer membrane protein	27kDa outer membrane protein	27kDa outer membrane protein	COG1651
2	membrane protein, putative	hypothetical protein	membrane protein, putative	hypothetical protein	COG5395
3	hypothetical protein	hypothetical protein	hypothetical protein	hypothetical protein	COG2932*
4	hypothetical protein	hypothetical protein	hypothetical protein	hypothetical protein	NA
5	hypothetical protein	hypothetical protein	hypothetical protein	hypothetical protein	NA
6	hypothetical protein	hypothetical protein	hypothetical protein	hypothetical protein	NA
7	Outer membrane receptor proteins, mostly Fe transport	TonB-dependent receptor	TonB-dependent receptor	TonB-dependent receptor	NA
8	FIG048548: ATP synthase protein I2	FIG048548: ATP synthase protein I2	FIG048548: ATP synthase protein I2	FIG048548: ATP synthase protein I2	NA
9	FOG: TPR repeat	FOG: TPR repeat	hypothetical protein	FOG: TPR repeat	NA
10	Nickel-dependent superoxide dismutase (EC 1.15.1.1)	Nickel-dependent superoxide dismutase (EC 1.15.1.1)	Nickel-dependent superoxide dismutase (EC 1.15.1.1)	Nickel-dependent superoxide dismutase (EC 1.15.1.1)	NA
11	Phycobilisome phycoerythrin-associated linker polypeptide	Phycobilisome rod linker polypeptide, phycocyanin-associated	Phycobilisome phycoerythrin-associated linker polypeptide	Phycobilisome phycoerythrin-associated linker polypeptide	NA
12	Phycobilisome protein	Phycobilisome core component	Phycobilisome protein	Phycobilisome core component	NA
13	Cell division protein ZipN/Ftn2/Arc6, specific for cyanobacteria and chloroplast	Cell division protein ZipN/Ftn2/Arc6, specific for cyanobacteria and chloroplast	Cell division protein ZipN/Ftn2/Arc6, specific for cyanobacteria and chloroplast	Cell division protein ZipN/Ftn2/Arc6, specific for cyanobacteria and chloroplast	NA
14	hypothetical protein	Sll1884 protein	Sll1884 protein	hypothetical protein	NA
15	Mobile element protein	hypothetical protein	Mobile element protein	hypothetical protein	NA
16	possible Protein phosphatase 2C	possible Protein phosphatase 2C	possible Protein phosphatase 2C	possible Protein phosphatase 2C	NA
17	possible Zinc finger, C3HC4	possible Zinc finger, C3HC4	possible Zinc finger, C3HC4	possible Zinc finger, C3HC4	NA

	type (RING finger)	type (RING finger)	type (RING finger)	type (RING finger)	
18	putative phycobiliprotein linker	putative phycobiliprotein linker	Phycobilisome rod linker polypeptide, phycocyanin-associated	Phycobilisome rod linker polypeptide, phycocyanin-associated	NA
19	Rod shape-determining protein MreD	Rod shape-determining protein MreD	Rod shape-determining protein MreD	Rod shape-determining protein MreD	NA
20	Small primase-like proteins (Toprim domain)	Small primase-like proteins (Toprim domain)	Small primase-like proteins (Toprim domain)	Small primase-like proteins (Toprim domain)	NA
21	Two-component response regulator	Two-component response regulator	Two-component response regulator	Two-component response regulator	NA
22	All3116 protein	All3116 protein	hypothetical protein	hypothetical protein	NA
23	Chlorophyll a(b) binding protein, photosystem II CP43 protein (PsbC) homolog	Chlorophyll a(b) binding protein, photosystem II CP43 protein (PsbC) homolog	Chlorophyll a(b) binding protein, photosystem II CP43 protein (PsbC) homolog	Chlorophyll a(b) binding protein, photosystem II CP43 protein (PsbC) homolog	NA
24	FIG01150038: hypothetical protein	hypothetical protein	hypothetical protein	FIG01150038: hypothetical protein	NA
25	FIG01150241: hypothetical protein	FIG01150241: hypothetical protein	FIG01150241: hypothetical protein	FIG01150241: hypothetical protein	NA
26	Glycerol dehydrogenase related protein Slr0730	Glycerol dehydrogenase related protein Slr0730	Glycerol dehydrogenase related protein Slr0730	Glycerol dehydrogenase related protein Slr0730	NA
27	Possible restriction /modification enzyme	Possible restriction /modification enzyme	Possible restriction /modification enzyme	Possible restriction /modification enzyme	NA
28	Small GTP-binding protein domain	Small GTP-binding protein domain	Small GTP-binding protein domain	Small GTP-binding protein domain	NA
29	L-lactate permease	hypothetical protein	hypothetical protein	L-lactate permease	NA
30 - 78	hypothetical protein	hypothetical protein	hypothetical protein	hypothetical protein	NA

CRISPR-associated proteins were a common feature of all four "*Ca.* Synechococcus spongiarum" genomes and CRISPR regions were discovered in three of them (15L, 142, and SH4). Phylotype 142 had the most CRISPR-associated annotations. Eight CRISPR regions were identified including two large modules (Table 3-10). One of them had a spacer region of 66 spacers (module 1) and another module with three spacer regions each spanning 70 spacers (module 2), (Figure 3-15). A gap of more than 7.5 kb separated the two modules. Upstream of the CRISPR-associated protein conglomeration, a helicase (COG1200) was annotated with an adjacent phage-related regulatory protein cII gene (COG1192).

Table 3-10 Classification of CRISPR-associated proteins in 142. The names in brackets were added when the annotated gene names differed from those proposed according to the nomenclature by Makarova et al. (2011). NA – not available.

CRISPR-associated protein	**COG**	**Module**	**Classification**
*cas*1	COG1518	2	Type I, II and III
*cas*1	COG1518	1	Type I, II and III
*cas*2	NA	2	Type I, II and III
*cas*2	COG3512	1	Type I, II and III
*cas*2	COG3512	NA	Type I, II and III
*cas*3	COG1203	2	Type I
cas5e (*cas5*)	NA	2	Subtype I-A,B,C,E
cse1	NA	2	Subtype I-E
cse2	NA	2	Subtype I-E
cse3 (*cas6e*)	NA	2	Subtype I-E
cse4 (*cas7*)	NA	2	Subtype I-A,B,C,E
cmr3	COG1769	1	Subtype III-B
cmr4	COG1336	1	Subtype III-B
cmr5	COG3337	1	Subtype III-B
cmr6	COG1604	1	Subtype III-B
TM1812 (*csx1*)	NA	1	Subtype III-U
Potential CRISPR-associated protein			
ATP-dependent DNA helicase	COG1200	2	NA
phage-related regulatory protein cII	COG1192	2	NA

Two CRISPR regions with six spacers each were found in SH4 with the CRISPR-associated proteins Cse 4,2,1 and Cas1 forming a conglomeration. In contrast to 142, in SH4 no module was formed by CRISPR regions and CRISPR-associated proteins, which were located on different contigs. But as with 142, additional helicases (COG1247) and the phage-related regulatory protein cII (COG1192) were located upstream of the CRISPR-associated proteins. One CRISPR region was found in 15L containing 49 spacers and a conglomeration of Cas1, Cas4, Cas2 and two Cas3 proteins. While SP3 did not contain any CRISPR regions, a conglomeration of Cas1, Cas4, and two Cas3 proteins was found. The six free-living cyanobacteria in this study were devoid of any CRISPR regions or CRISPR-associated proteins.

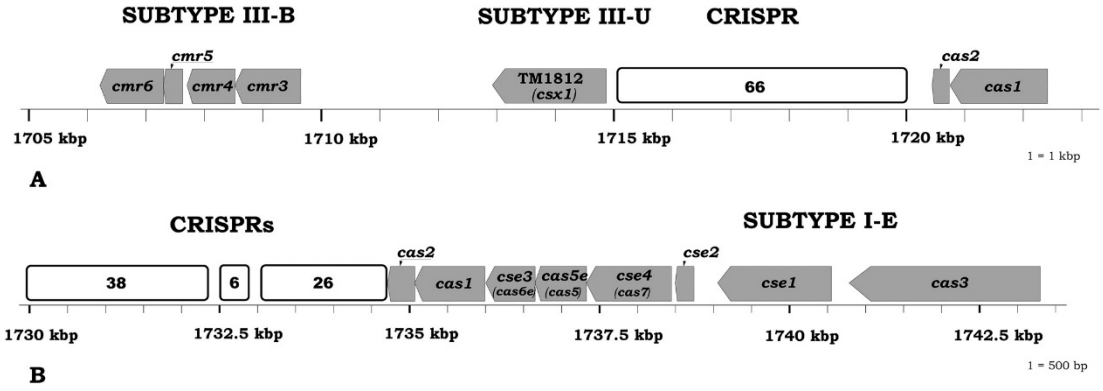


Figure 3-15 Schematic representation of the genomic architectures of two CRISPR-Cas of "*Ca.* Synechococcus spongiarum" 142. The number of spacers of the CRISPR regions and the closest CRISPR-Cas subtype according to Makarova *et al.* (2011) are shown. The names of genes are described as they were annotated in the analysis (see Materials and Methods). The names in parentheses were added when the annotated gene names differed from the nomenclature proposed by Makarova *et al.* (2011). (A) Module 1, consisting of proteins resembling subtype III-B and subtype III-U. (B) Module 2, showing proteins resembling subtype I-E.

Table 3-11 KEGG enzymes found to be missing among several distinctive metabolic pathways in "*Ca.* Synechococcus spongiarum" genomes. Enzymes were considered missing only if they were present in all six genomes of the free-living reference group.

KEGG pathway	SP3	142	SH4	15L
Pentose phosphate pathway		4.1.2.4		
Fructose and mannose metabolism (biosynthesis of GDP-D-rhamnose)	4.2.1.47	4.2.1.47	4.2.1.47	4.2.1.47
Aminosugars metabolism	3.5.99.6	3.5.99.6 3.5.1.25a	3.5.99.6	3.5.99.6
Pyruvate metabolism	3.6.1.7	3.6.1.7 4.4.1.5	3.6.1.7 4.4.1.5	3.6.1.7 4.4.1.5
Sulfur metabolism	2.7.1.25			
Biosynthesis of dTDP-L-rhamnose	2.7.7.24 5.1.3.13 1.1.1.133	2.7.7.24 5.1.3.13 1.1.1.133	2.7.7.24 5.1.3.13 1.1.1.133	2.7.7.24 5.1.3.13 1.1.1.133
Glycine and Serine	2.7.1.39a			
Methionine metabolism	4.1.1.50 5.3.1.23 1.13.11.53/4 4.2.1.109a	4.1.1.50 5.3.1.23 1.13.11.53/4 4.2.1.109a	4.1.1.50 5.3.1.23 1.13.11.53/4 4.2.1.109a 2.5.1.16 2.4.2.28 3.3.1.1	4.1.1.50 1.13.11.53/4 4.2.1.109a

[a]for one or two free-living cyanobacteria not supported by SEED annotation

According to RAST annotation, the key functional pathways were nearly complete in all four symbionts including glycolysis, the tricarboxylic acid (TCA) cycle, nitrogen metabolism, the pentose phosphate pathway, fatty acid biosynthesis, fructose and mannose metabolism, amino sugar metabolism, pyruvate metabolism, amino acid metabolism, sulfur metabolism, sucrose metabolism, and photosynthesis. The lack of two genes encoding for the enzymes adenosylhomocysteinase (EC 3.3.1.1) and O-acetyl-L-homocysteine acetate-lyase (EC 2.5.1.49) previously reported for SH4 was confirmed in this study (Gao *et al.*, 2014b). Both genes are part of the L-homocysteine synthesis. EC 2.5.1.49, that is involved in the

synthesis of L-homocysteine from L-homoserine, was also found missing in SP3 and 142, but EC 3.3.1.1, that is synthesizing homocysteine from S-adenosyl-L-homocysteine, was annotated in all three "*Ca.* Synechococcus spongiarum" draft genomes SP3, 142, and 15L. While the methionine salvage pathway was complete in all six free-living cyanobacteria, a number of the involved enzymes were missing in the symbionts (Table 3-11).

Table 3-12 Abundance of photosynthetic genes of PSI and PSII in "*Ca.* Synechococcus spongiarum" and free-living cyanobacterial references based on SEED annotations (and KEGG for PSII).

psa	**SP3**	**142**	**15L**	**SH4**	**PCC 6307**	**WH 5701**	**RS 9917**	**WH 7803**	**RCC 307**	**PCC 7001**
A	1	1	1	1	1	1	1	1	1	1
B	1	1	1	1	1	1	1	1	1	1
C	1	1	1	1	1	1	1	1	1	1
D	1	1	1	1	1	1	1	1	1	1
E	1	1	1	1	1	1	1	1	1	1
F	1	1	0	1	1	1	1	1	1	1
I	0	0	0	0	0	1	1	1	1	1
J	1	1	0	0	1	1	1	1	1	1
K	1	1	1	1	1	1	1	1	1	1
L	1	1	1	1	1	1	1	1	1	1
M	0	0	0	0	0	0	0	0	1	0
Sum	**9**	**9**	**7**	**8**	**9**	**10**	**10**	**10**	**11**	**10**
psb	**SP3**	**142**	**15L**	**SH4**	**PCC 6307**	**WH 5701**	**RS 9917**	**WH 7803**	**RCC 307**	**PCC 7001**
A	4	3	1	2	5	4	4	4	4	4
B	1	1	1	1	1	1	1	1	1	1
C	2*	2*	1	2*	1	1	1	1	1	1
D	0	0	0	0	2	2	2	2	2	2
E	1	1	1	1	1	1	1	1	1	1
F	1	1	1	1	1	1	1	1	1	1
H	1	1	1	0	1	1	1	1	1	1
I	1*	1	0	1*	1	1	1	1	0	1
J	1	1	1	1	1	1	1	1	1	1
K	0	1	0	0	1*	1*	1	1	0	1
L	1	1	1	1	1	1*	1	1	1*	1
M***	0	0	0	0	1	1	1	1	1	1
N	1**	1	1	0	1	1	1	1	1	1
O	1	1	1*	1	1	1	1	1	1	1
P	0	0	0	0	1	1	1	1	1	1
U	1	1	1	1	1	1	1	1	1	1
X	1	1	1	1	1	1	1	1	1	1
V	1	1	1	1	1	1	1	1	1	1
Y	1**	0	0	0	1	1	1	1	1	1*
Z	1	1	1	1	1	1	1	1	1	1
27	0	1	1	0	1	1	1	1	1	1
28	1	1	1	0	1	1	1	1	1	1
Sum	**21**	**21**	**16**	**15**	**27**	**26**	**26**	**26**	**24**	**26**

* only supported by SEED (if multiple genes: only one of the group not supported by KEGG), ** only supported by KEGG, *** found by BLAST of the genome sequence

A number of small peptides (*psb* genes) was lost not only in SH4 in comparison to the free-living relatives, as previously reported (Gao *et al.*, 2014b), but also in all three newly sequenced "*Ca.* Synechococcus spongiarum" genomes (Table 3-12). Some genes were only annotated by one or two of the three applied annotation methods SEED, KEGG and BLAST, e.g. *psbM* was only detected in the BLAST analysis. The number of oxidative stress resistance-

related genes was also reduced in "*Ca.* Synechococcus spongiarum" and glutathione peroxidase (EC 1.11.1.9) was completely missing in all four symbionts, while present in all references (Table 3-13).

Table 3-13 Resistance to oxidative stress, based on SEED annotation, is reduced in the genomes of "*Ca.* Synechococcus spongiarum" compared to the free-living cyanobacterial references.

SEED annotation	**SP3**	**142**	**15L**	**SH4**	**PCC 6307**	**RCC 307**	**WH 7803**	**WH 5701**	**PCC 7001**	**RS 9917**
Glutathione reductase (EC 1.8.1.7)	1	1	1	1	1	1	1	1	1	1
Glutathione peroxidase (EC 1.11.1.9)	0	0	0	0	1	1	1	1	1	1
Glutathione synthetase (EC 6.3.2.3)	1	1	1	1	1	1	2	1	1	1
Gamma-glutamyltranspeptidase (EC 2.3.2.2)	0	0	0	0	1	1	1	1	1	0
Methylhydantoinases A, B (EC 3.5.2.14)	0	0	0	0	0	1	1	0	0	0
Rubredoxin	2	1	1	2	2	2	1	3	2	2
Non-specific DNA-binding protein Dps	1	0	1	0	1	1	1	1	1	1
Metallothionein	0	0	0	0	0	0	1	0	0	0
Alkyl hydroperoxide reductase	2	3	2	1	5	5	4	4	4	4
Peroxide stress regulator	1	1	1	1	1	1	1	1	1	1
Transcriptional regulator, Crp/Fnr family	0	0	0	0	1	0	0	1	0	1
Zinc uptake regulation protein ZUR	0	1	0	0	1	1	1	1	1	1
Ferric uptake regulation protein FUR	1	1	1	1	1	1	1	1	1	1
Glutaredoxins	2	2	2	3	5	3	3	3	3	3
Phytochelatin synthase (EC 2.3.2.15)	0	0	0	0	0	0	0	1	0	0
Superoxide dismutase (total)	2	1	2	1	2	2	2	2	2	2
Glutathione S-transferase (total)	2	2	2	2	6	4	2	7	6	4
SUM	15	14	14	13	29	25	23	29	25	23

3.2 PacBio-Illumina hybrid assembly pipeline development

3.2.1 Statistics of the tested assembly strategies

A total of 14 assemblies were created from the test dataset and compared to each other, nine of them from only the Illumina reads and five hybrid assemblies of PacBio and Illumina reads together (Table 3-14). The two omega assemblies – with and without prior read normalization – had the lowest N50 and overall very small contigs. Therefore, no hybrid assembly was attempted with this assembler. IDBA_UD and SPAdes performed both well and produced similar output regarding contig numbers and lengths. While the prior read normalization resulted in higher N50 values for both assemblers, the overall assembly size as well as the largest contig were smaller. Also, the GC content of the assemblies of normalized reads were higher than the GC content of the assemblies of all reads.

Table 3-14 QUAST comparison of assemblies of the test dataset at various settings sorted by decreasing N50. Only contigs (not scaffolds) were compared and the assemblies do not contain Ns. In bold are the two assemblies with the highest N50 in the groups of hybrid assemblies and Illumina-only assemblies. Those two are further referred to as 'hybrid assembly' and 'Illumina-only assembly,' respectively.

Assembly	# contigs (>= 1000 bp)	Total length (>= 1000 bp)	Largest contig	GC (%)	N50
hybrid_spades-oa_k-127	103	25124367	4127107	47.1	3338481
hybrid_spades-o33-k-127	111	25082412	4127355	47.12	2741038
hybrid_spades-oa_k-55	145	24778010	3206856	47.21	1549056
hybrid_spades-o33_k-55	164	24711857	2571480	47.21	1270551
bbnorm_spades-oa	1071	21155562	489015	48.12	88639
bbnorm_spades	1095	21142641	489098	48.12	83307
bbnorm_idba-ud	1166	21076910	416899	48.12	74357
not-normalized_spades-oa	2012	27126067	790396	45.73	71332
not-normalized_spades-sc-oa	2111	27060928	636480	45.73	65744
hybrid_idba-ud	1276	26304267	534367	46.26	62615
not-normalized_idba-ud	2213	27109318	416899	45.74	54133
not-normalized_spades	27	63480	11477	39.67	1451
bbnorm_omega	194	243917	2618	46.03	612
not-normalized_omega	198	249488	2618	46	599

For the hybrid assemblies to run through, the Illumina reads had to be normalized first. Therefore, all hybrid assemblies compared here were created with normalized Illumina reads (Table 3-14). IDBA_UD seemed not to incorporate the PacBio long-reads well – this assembly listed even below the IDBA_UD assembly on the normalized Illumina reads alone. The hybrid assemblies created with SPAdes at any of the tested settings were clearly superior to all other tested assemblies. Using longer kmers (21, 33, 55, 77, 99, 127) improved the assembly over those with default kmer settings (21, 33, 55). The assembly with the highest N50 in the group

of Illumina-only assemblies was further compared to the one in the group of PacBio-Illumina hybrid assemblies. These two will be hereafter referred to as 'Illumina-only assembly' and 'hybrid assembly,' respectively.

3.2.2 Comparison back to original reference genomes

Both assemblies were aligned to the nine original, closed assemblies for evaluation. In the Illumina-only assembly the genomes remained split up into numerous contigs (Figure 3-16). The contigs were larger for genomes for which higher read coverage was simulated (e.g. 200x coverage for the *Acidobacterium*). In the hybrid assembly on the other hand, the reads were merged into large contigs by the addition of the PacBio long-reads (Figure 3-17). The *Acidobacterium*, the *Desulfovibrio*, and the *Nitrosomonas* genomes were even assembled into a single contig each, and *Clostridium* and *Fusobacterium* into two contigs. Also the genomes with low simulated read coverage were assembled into considerably larger contigs than in the Illumina-only assembly. The hybrid assembly is therefore clearly superior to the Illumina-only assembly.

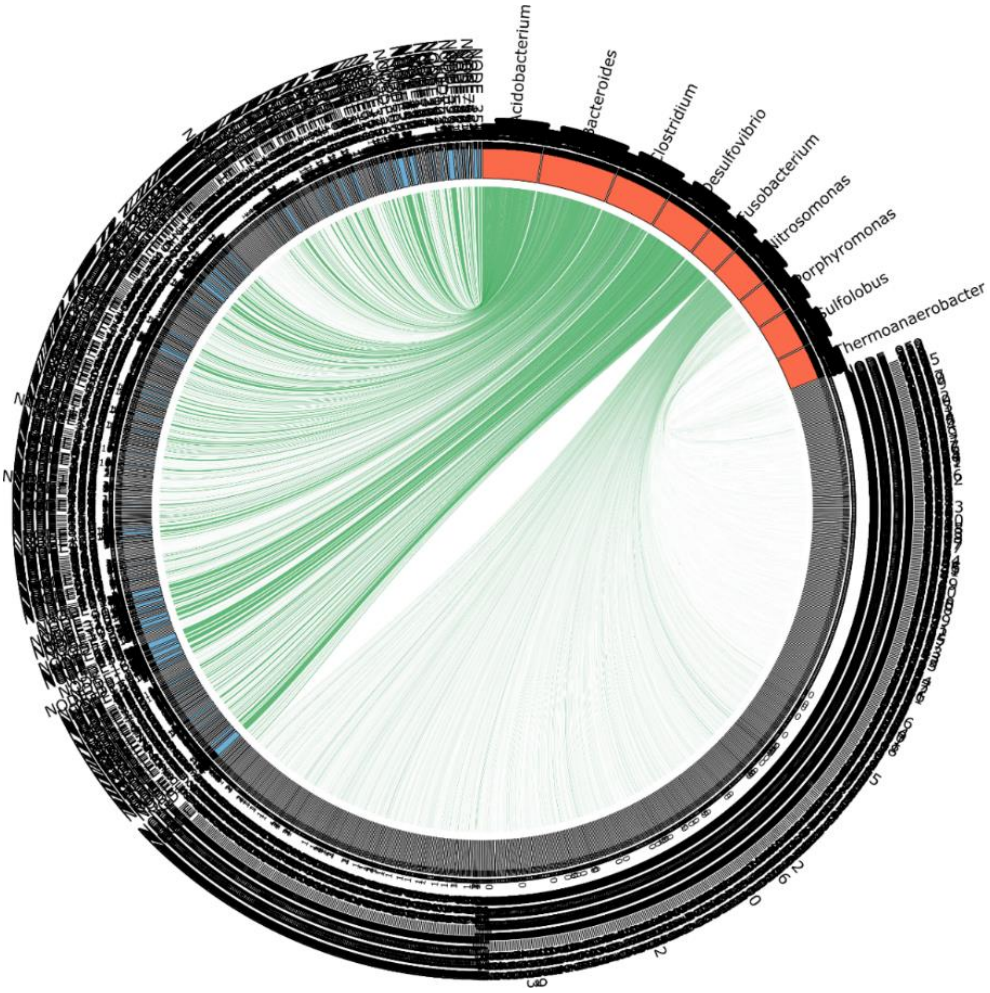


Figure 3-16 Alignment of contigs of the Illumina-only assembly (blue) to the nine original bacterial assemblies (red). Matching areas are connected in green.

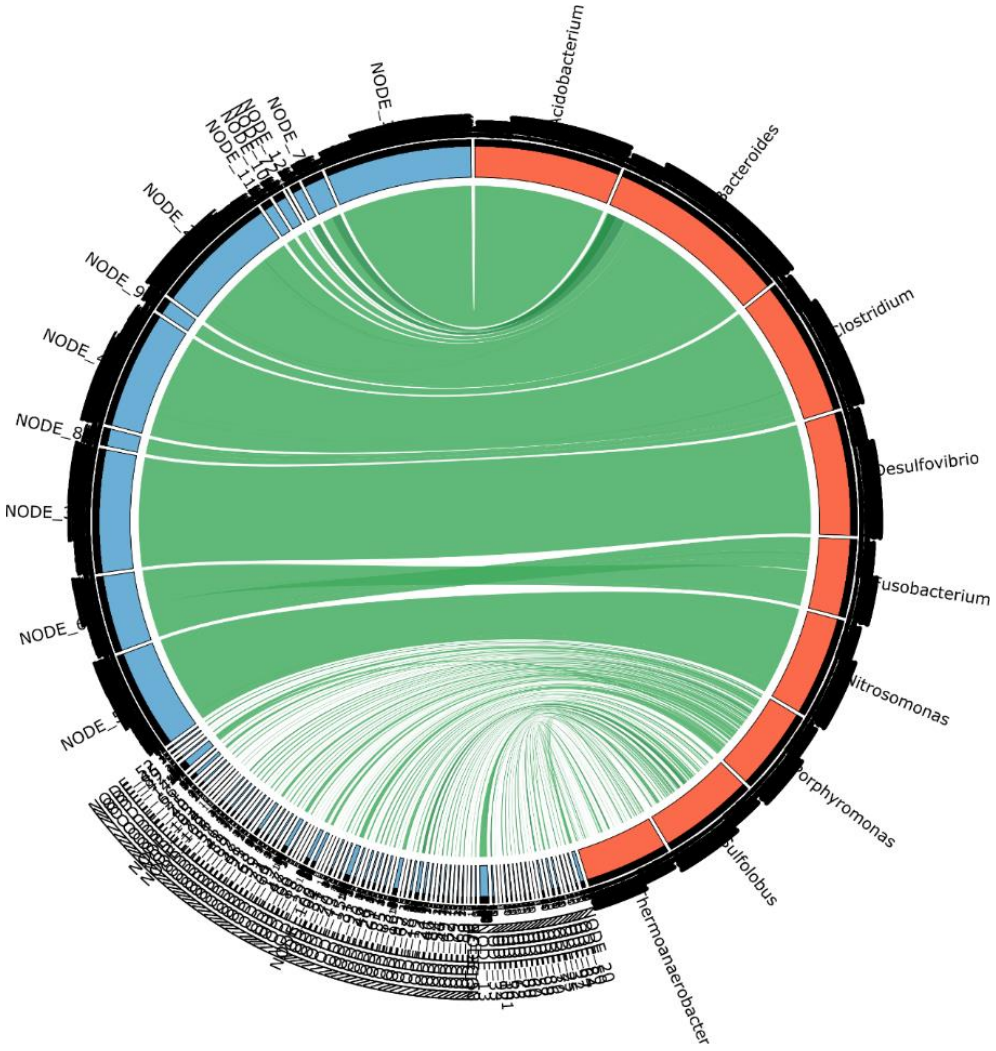


Figure 3-17 Alignment of contigs of the hybrid assembly (blue) to the nine original bacterial assemblies (red). Matching areas are connected in green.

3.2.3 Reference-independent binning

While the above analysis makes use of the original, closed genomes that this data is based on, this assembly pipeline was intended to be applied for metagenomic communities with a multitude of unknown members at unknown fractions and therefore read coverages. To account for this while testing the assembly pipeline, the 16S rRNA genes were annotated and phylogenetically identified with the RDPclassifier. Phylogenetic identity was confirmed by BLASTn (Table 3-15). The contigs of both assemblies were plotted according to their coverages derived from bowtie2 read mapping and SPAdes (Figure 3-18 and Figure 3-19). These kind of plots are commonly used for manual binning of members of the microbial community (Albertsen *et al.*, 2013). Plotting the Illumina-only assembly (Figure 3-18), the contigs build a line with no clear borders between possible bins, especially in the low-coverage regions. Also in regions of high coverage, bins are overlapping: the *Desulfovibrio* 16S rRNA gene-containing contig (pink) plots on top of the contig containing the *Acidobacterium* 16S rRNA gene (red). In the hybrid assembly on the other hand, the bins are more clearly

distinguishable also in low-coverage areas (Figure 3-19). This suggests that the assembly into larger contigs enabled by the PacBio long-reads also eases, and even enables, binning of single members of the microbial community.

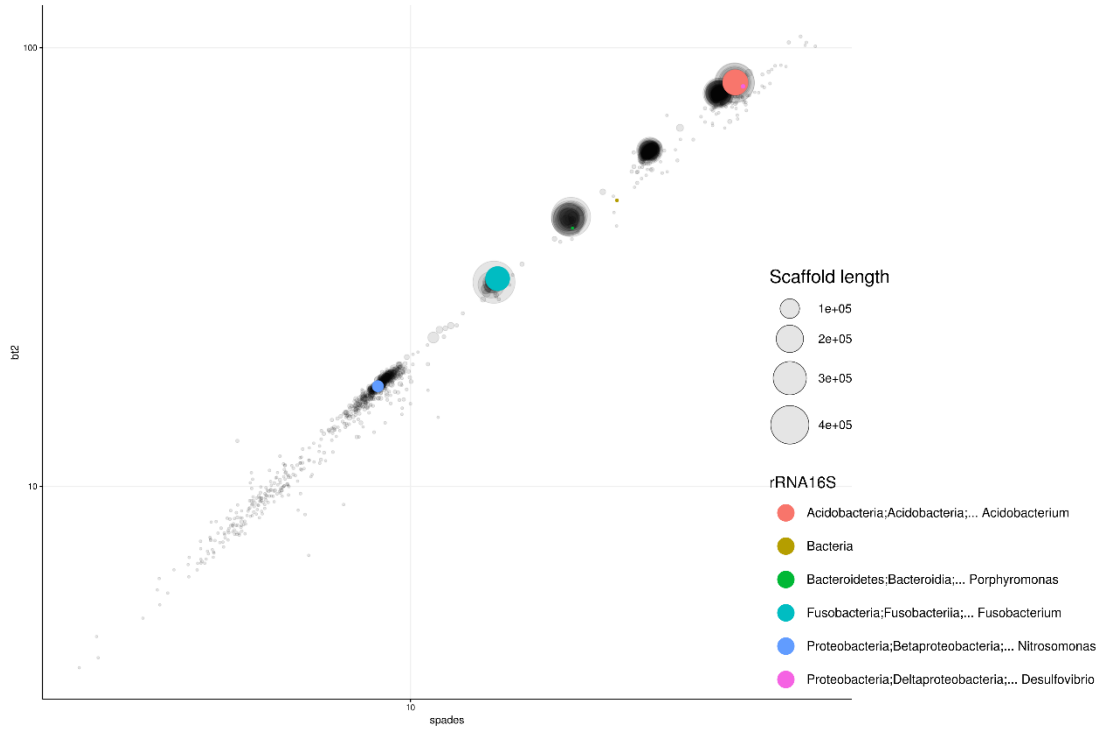


Figure 3-18 Contigs of the Illumina-only assembly plotted according to their coverage determined by mapping the Illumina reads back to the assembly with bowtie2 (bt2, y-axis) and the coverage determined by SPAdes (spades, x-axis). Contigs containing a 16S rRNA gene are colored according to phylogeny.

Figure 3-19 Contigs of the hybrid assembly plotted according to their coverage determined by mapping the Illumina reads back to the assembly with bowtie2 (bt2, y-axis) and the coverage determined by SPAdes (spades, x-axis). Contigs containing a 16S rRNA gene are colored according to phylogeny.

Comparing the genomes recovered by binning from the hybrid assembly to the references based on their content of essential single-copy genes (Albertsen *et al.*, 2013) shows that genome recovery by binning worked very well for the majority of genomes (Table 3-16). Only the two genomes with the lowest sequence coverage, namely *Sulfolobus tokodaii* and *Thermoanaerobacter pseudethanolicus* could not be sorted into individual bins.

Table 3-15 Phylogenetic identification of hybrid assembly bins determined by a BLASTn search of the rRNA genes (preferably 16S; 23S only if no 16S rRNA gene available).

query	gene	hit	query cov	ident	accession
binA_1.1	16S	*Acidobacterium capsulatum* ATCC 51196	100	100	CP001472.1
binB_2.6	16S	*Bacteroides vulgatus* ATCC 8482	100	99	CP000139.1
binB_2.7	16S	*Bacteroides vulgatus* ATCC 8482	100	99	CP000139.1
binB_2.8	16S	*Bacteroides vulgatus* ATCC 8482	100	99	CP000139.1
binB_2.9	16S	*Bacteroides vulgatus* ATCC 8482	100	99	CP000139.1
binB_2.10	16S	*Bacteroides vulgatus* ATCC 8482	100	99	CP000139.1
binC_4.2	16S	*Clostridium thermocellum* ATCC 27405	100	100	CP000568.1
binC_8.4	16S	*Clostridium thermocellum* ATCC 27405	100	100	CP000568.1
binC_8.5	16S	*Clostridium thermocellum* ATCC 27405	100	100	CP000568.1
binC_8.6	16S	*Clostridium thermocellum* ATCC 27405	100	100	CP000568.1
binD_3.1	23S	*Desulfovibrio vulgaris* DP4	100	100	CP000527.1
binD_3.2	23S	*Desulfovibrio vulgaris* DP4	100	100	CP000527.1
binD_3.3	23S	*Desulfovibrio vulgaris* DP4	100	100	CP000527.1
binD_3.4	23S	*Desulfovibrio vulgaris* DP4	100	100	CP000527.1
binD_3.5	23S	*Desulfovibrio vulgaris* DP4	100	100	CP000527.1
binE_6.6	16S	*Fusobacterium nucleatum subsp. nucleatum* ATCC 25586	100	100	AE009951.2
binE_6.7	16S	*Fusobacterium nucleatum subsp. nucleatum* ATCC 25586	100	100	AE009951.2
binE_6.8	16S	*Fusobacterium nucleatum subsp. nucleatum* ATCC 25586	100	100	AE009951.2
binE_6.9	16S	*Fusobacterium nucleatum subsp. nucleatum* ATCC 25586	100	100	AE009951.2
binE_6.10	16S	*Fusobacterium nucleatum subsp. nucleatum* ATCC 25586	100	100	AE009951.2
binF_5.2	16S	*Nitrosomonas europaea* ATCC 19718	100	100	AL954747.1
binG_47.2	16S	*Porphyromonas gingivalis* ATCC 33277	100	100	AP009380.1
binH_23.2	16S	*Sulfolobus tokodaii* str. 7	100	100	BA000023.2
binH_50.3	16S	*Thermoanaerobacter pseudethanolicus* ATCC 33223	100	99	CP000924.1

Table 3-16 Comparison of the hybrid assembly bins to the references regarding the single-copy essential gene content (Albertsen *et al.*, 2013).

genome	sum	#duplicates	sum-dupl	% of 111
Acidobacterium capsulatum				
binA	108	3	105	94.59%
Reference	108	3	105	94.59%
Bacteroides vulgatus				
binB	101	3	98	88.29%
Reference	109	3	106	95.50%
Clostridium thermocellum				
binC	109	4	105	94.59%
Reference	109	4	105	94.59%
Desulfovibrio vulgaris				
binD	107	3	104	93.69%
Reference	107	3	104	93.69%
Fusobacterium nucleatum				
binE	106	4	102	91.89%
Reference	107	4	103	92.79%
Nitrosomonas europaea				
binF	108	3	105	94.59%
Reference	107	2	105	94.59%
Porphyromonas gingivalis				
binG	70	1	69	62.16%
Reference	107	1	106	95.50%
Sulfolobus tokodaii and *Thermoanaerobacter pseudethanolicus*				
binH	24	2	22	19.82%
S. tokodaii	29	0	29	26.13%
T. pseudethanolicus	108	2	106	95.50%

3.3 Binning 37 symbiont genomes from the metagenome of *A. aerophoba*

3.3.1 Assessment of metagenomic DNA extraction and sequencing

For Illumina sequencing, metagenomic DNA was extracted from six SAPs of *A. aerophoba*. Three were derived from pinacoderm, and three from mesohyl tissue. For both tissue types, the three replicates differed in cell lysis method (bead beating, proteinase K digestion, and freeze-thaw cycling). Quality of the extracted DNA was generally high (Figure 3-20). Concentrations differed between extraction methods (Table 3-17). Highest yields were obtained from bead beating as included in the DNA extraction kit. Proteinase K digestion delivered comparably high DNA concentrations, and freeze-thaw cycling produced the lowest DNA concentrations.

Metagenomic Illumina HiSeq sequencing resulted in between 82,698,080 and 111,951,445 reads (Mft and Mpk, respectively). In total, 567,206,927 Illumina reads were sequenced. PacBio sequencing delivered 235,016 sequences and read correction with proovread resulted in 101,530 corrected PacBio long-reads.

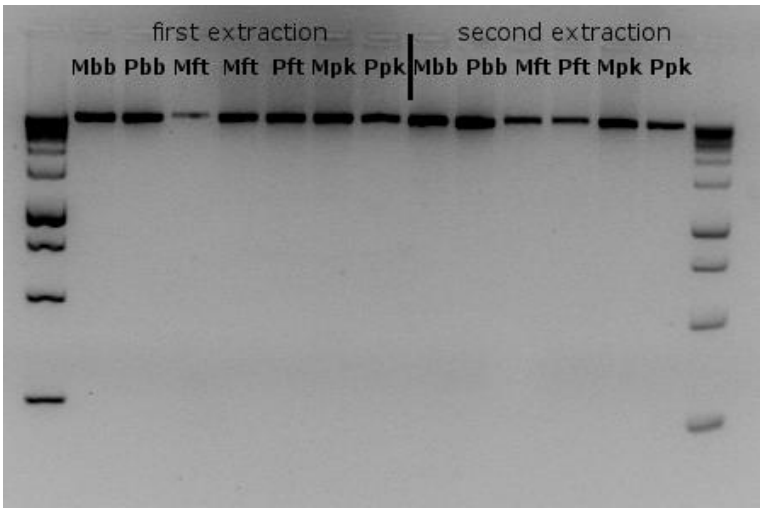


Figure 3-20 Agarose gel picture of metagenomic DNA from different extractions. Labeling refers to the respective tissue type and cell lysis method: capital M or P for mesohyl or pinacoderm, respectively; bb – bead beating, ft – freeze-thaw cycling, pk – proteinase K digestion.

Table 3-17 DNA concentrations of metagenomic DNA from different extractions. Labeling refers to the respective tissue type and cell lysis method: capital M or P for mesohyl or pinacoderm, respectively; bb – bead beating, ft – freeze-thaw cycling, pk – proteinase K digestion.

Extraction round	Extract	Qubit (ng/µl)
First round	Mbb	18.6
	Pbb	18.9
	Mft	3.7
	Pft	15.2
	Mpk	16.5
	Ppk	13.1
Second round	Mbb	16.1
	Pbb	17.9
	Mft	5.7
	Pft	6.0
	Mpk	16.0
	Ppk	12.3

3.3.2 Comparison of Illumina-only and PacBio-Illumina hybrid assemblies

Two metagenome assemblies were obtained, one only from Illumina HiSeq short-reads (Illumina-only assembly), and one from the same Illumina short-reads set, but combined with pre-corrected PacBio long-reads (hybrid assembly). The two assemblies differed notably in number of contigs and total size (Table 3-18). The Illumina-only assembly comprised >100 000 contigs with a total length of 490 Mbp, the hybrid assembly consisted of >30 000 contigs with a total length of 301 Mbp. Only contigs >=1 000 bp were considered. The addition of the PacBio reads to the assembly increased the N_{50} value 3.8-fold, from about 9 kbp to 34 kbp. While the number of highly complete genome bins (> 70% completeness) decreased (42 Illumina-only bins vs 37 hybrid bins), the portion of full-length 16S rRNA gene containing bins doubled from 16 in the Illumina-only assembly to 32 in the hybrid assembly. To assess if contigs from the Illumina-only assembly were reappearing in the hybrid assembly and if the

PacBio reads merged them into larger contigs, an Illumina-only bin was mapped to the corresponding hybrid bin. This allowed a visual comparison of the assemblies (Figure 3-21). This mapping shows that the two assemblies corresponded well because contigs that had been constructed out of the Illumina data reappeared upon addition of the PacBio reads. Moreover, they were merged into even larger contigs, thus resulting in a higher-quality bin.

Table 3-18 Comparison of Illumina-only and PacBio-Illumina hybrid assemblies

	Illumina-only	**PacBio-Illumina hybrid**
MG-RAST ID	mgm4671062.3	mgm4671058.3
Contig number (≥ 1 000 bp)	110 609	31 187
Size (Mb)	490	301
N50	8 958	33 831
N75	2 873	12 184
L50	8 886	1 980
L75	34 979	5 726
CDSs	509 054	289 685
Bin number	217	137
> 90% completeness (with 16S rRNA gene)	25 (12)	26 (22)
85-90% completeness (with 16S rRNA gene)	12 (4)	6 (6)
70-85% completeness (with 16S rRNA gene)	5 (0)	5 (4)

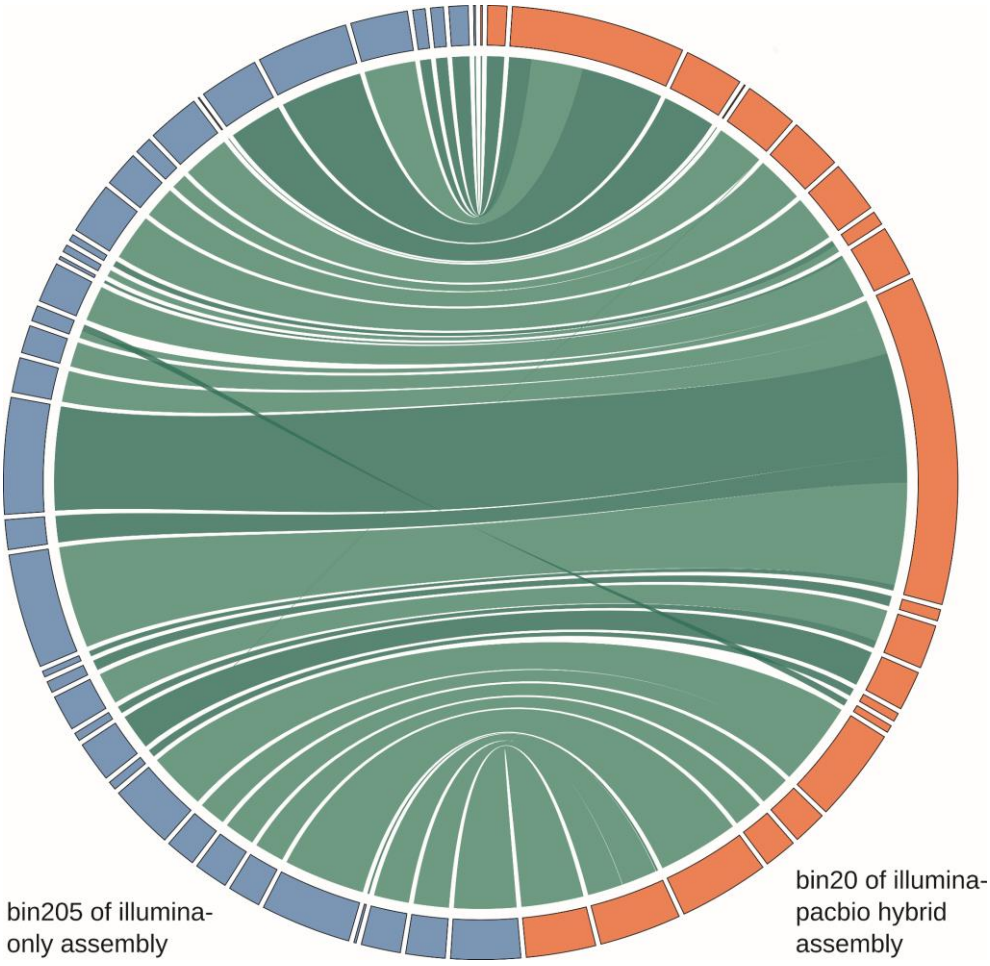


Figure 3-21 Mapping of Illumina-only assembly bin205 (blue) to PacBio-Illumina hybrid assembly bin20 (red). Corresponding areas are connected in green.

To obtain short-read data optimized for differential coverage binning, six DNA samples from the same sponge specimen were extracted with varied lysis protocols, and deeply sequenced on an Illumina HiSeq2000 instrument (see Figure 3-22 of JGI Project ID 1024999 for additional ribosomal 16S rRNA V4 iTag data of this sequencing project). Although we already obtained a large number of high-completeness bins from the Illumina-only assembly, only 38% of the binned genomes contained a 16S rRNA gene. Contrasting, in the PacBio-Illumina hybrid assembly 86% of the bins contained a 16S rRNA gene (Table 3-18). Furthermore, with a 3.8-fold higher N_{50} hybrid assembly was more contiguous. For these reasons, all downstream analyses were carried out with the genomes binned from the PacBio-Illumina hybrid assembly.

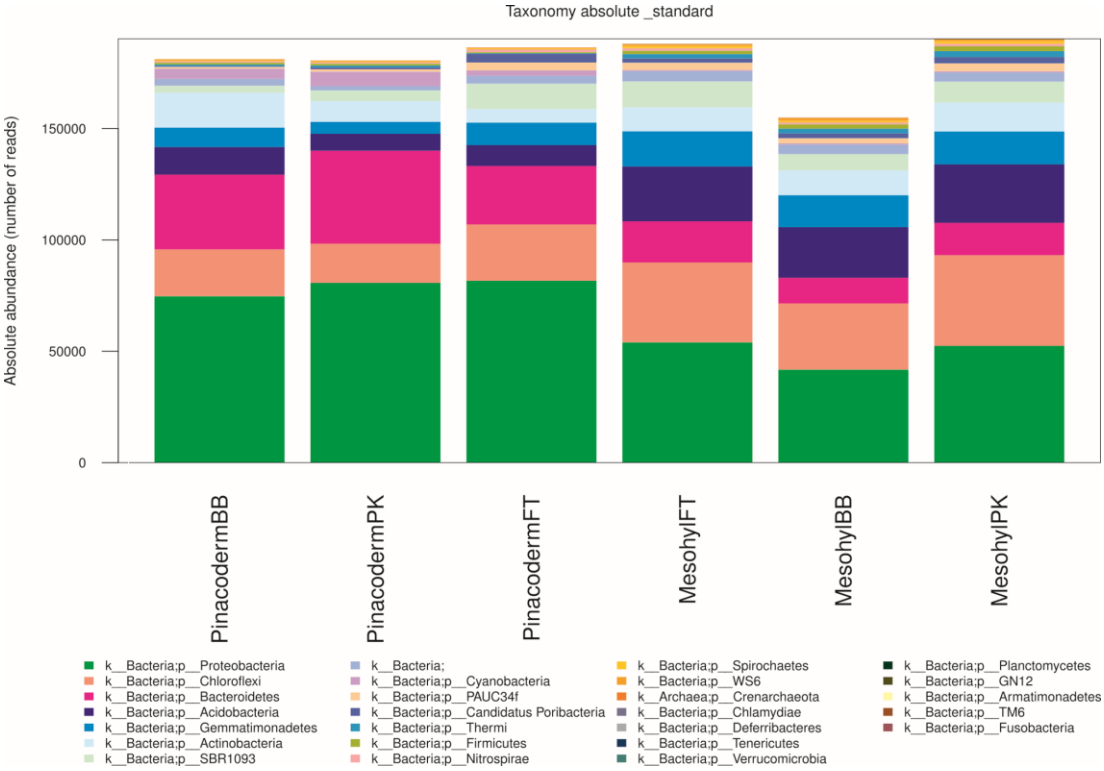


Figure 3-22 iTag analysis of the six DNA extracts for Illumina sequencing differing in cell lysis. Abbreviations: BB – bead beating, PK – proteinase K digestion, FT – freeze-thaw cycles.

3.3.3 Bacterial genomes binned from hybrid assembly

The 37 binned genomes belonged to 11 bacterial phyla and 2 candidate phyla, which are representative of the sponge symbiont consortium: Proteobacteria (Alpha, Gamma, and Delta), Chloroflexi, Acidobacteria, Actinobacteria, Bacteroidetes, Gemmatimonadetes, Deinococcus-Thermus, Nitrospirae, Nitrospinae, Cyanobacteria, Spirochaetes and the candidate phyla Poribacteria and SBR1093 (Table 3-19). The bins varied in total number of contigs from 21 to 758. Large numbers of contigs did not correlate with low sequence coverage: the bin with lowest coverage (bin18 with 38-times coverage), for example, was composed of as few as 83 contigs and was 87% complete. Estimated genome sizes, based on total length and estimated genome completeness, ranged from 1.9 Mbp (Alphaproteobacterium bin98) to 7.9 Mbp (Acidobacterium bin110). With respect to GC content, the genomes ranged from 36% (Bacteroidetes bin25) to nearly 70% (Alphaproteobacterium bin129). Overall, the sponge symbionts had genomes of high GC-content: 13 were between 50% and 60%, 17 of symbiont genomes comprised >60% of GC-bases. Comparably high average GC contents are a known feature of sponge metagenomes (Horn *et al.*, 2016). The N_{50} values also showed variability, with the smallest being 6 974 bp for Alphaproteobacterium bin95 and the largest being 309 970 bp for Chloroflexi bin127. The number of coding sequences (CDSs) in the symbiont genomes ranged from 1 455 (Alphaproteobacterium bin98) to 6 288 (Ca. Poribacterium bin44). The number of COGs annotated for each genome ranged between 490

(bin98) and 3 450 (Alphaproteobacterium bin129) which translates to 34% (bin98) and 76% (Alphaproteobacterium bin65) CDSs in COGs (see Appendix 3-2 for detailed COG annotations).

Table 3-19 Binned genomes of PacBio-Illumina hybrid assembly . Only duplicate genes other than PF00750, PF01795, and TIGR00436 were counted, as these genes are known to occur in multiple copies (Albertsen *et al.*, 2013).

Accession	Bin	Phylogeny	% est. com.	Contig no.	times cov.	est. size (Mb)	% GC	N50	CDS no.	% in COGs	dupl.
MPNP00000000	bin131	Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;[c]	93,69	416	612	4	41,99	19 278	3 392	59,58	1
MPMP00000000	bin36	Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Albidovulum;[c]	89,19	201	189	6,3	58,04	44 410	5 122	64,02	0
MPMX00000000	bin65	Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;uncultured;[a]	91,89	94	396	4,7	66,16	72 338	4 036	76,39	0
MPNO00000000	bin129	Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;Defluviicoccus;[a]	82,88	122	237	5,8	69,54	56 772	4 742	72,75	0
MPMV00000000	bin56	Proteobacteria;Alphaproteobacteria;Rickettsiales;EF100-94H03;[a]	93,69	102	54	4,8	63,69	78 682	4 292	70,36	0
MPMT00000000	bin52	Proteobacteria;Alphaproteobacteria;[d]	92,79	120	234	4,5	66,54	52 938	3 989	72,75	1
MPNG00000000	bin98	Proteobacteria;Alphaproteobacteria;[c]	85,59	105	96	1,9	40,65	46 493	1 455	33,68	0
MPNF00000000	bin95	Proteobacteria;Alphaproteobacteria;[d]	75,68	582	152	4,5	66,27	6 974	3 890	64,16	0
MPMI00000000	bin18	Proteobacteria;Deltaproteobacteria;Desulfurellales;Desulfurellaceae;uncultured;[a]	87,39	83	38	6	57,83	103 191	5 238	65,65	0
MPMQ00000000	bin40	Proteobacteria;Gammaproteobacteria;Oceanospirillales;Hahellaceae;Kistimonas;[c]	93,69	215	76	4	57,27	38 525	2 848	46,91	0
MPMU00000000	bin55	Proteobacteria;Gammaproteobacteria;[c]	84,68	183	47	3,5	47,27	24 711	2 562	67,49	1
MPNI00000000	bin106	Proteobacteria;[c]	90,09	148	53	2,9	39,61	55 882	2 088	41,62	0
MPNJ00000000	bin107	Nitrospinae/Tectomicrobia group;Nitrospinae;[d+]	90,09	60	440	4,9	59,46	165 774	4 046	69,7	1
MPMZ00000000	bin75	Nitrospirae;Nitrospira;Nitrospirales;Nitrospiraceae;Nitrospira;[a]	91,89	115	65	3,3	56,24	44 884	3 093	56,58	2
MPMH00000000	bin63	SBR1093;EC214;[b]	88,29	150	479	2,6	50,41	30 980	2 180	68,21	4
MPNK00000000	bin110	Acidobacteria;Holophagae;Subgroup 10;TK85;[a]	79,28	758	549	7,9	67,45	12 332	5 726	55,43	4
MPMW00000000	bin61	Acidobacteria;Acidobacteria;[c]	70,27	207	117	4,1	67,65	19 828	2 561	55,92	0
MPMY00000000	bin70	Candidatus Poribacteria;[d+]	91,89	106	351	5,5	40,34	70 347	4 254	59,07	6
MPMS00000000	bin44	Candidatus Poribacteria;Poribacteria genera incertae sedis;[c]	91,89	465	265	7,7	47,18	23 989	6 288	54,28	10
MPNB00000000	bin80	Bacteroidetes;Cytophagia;Order II;Rhodothermaceae;uncultured;[a]	92,79	192	453	4,4	50,96	34 696	3 555	50,44	0
MPMN00000000	bin25	Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;[c]	90,09	124	589	3,3	36,18	40 599	2 420	51,03	0
MPMR00000000	bin43	Gemmatimonadetes;Gemmatimonadetes;BD2-11 terrestrial group;[a]	92,79	65	633	4,8	67,96	132 700	3 664	60,84	1
MPNE00000000	bin94	Gemmatimonadetes;[d+]	91,89	83	190	4,8	66,9	89 414	3 702	59,62	1
MPNH00000000	bin103	Spirochaetae;Spirochaetes;Spirochaetales;Spirochaetaceae;[c]	87,39	96	66	5,5	67,36	71 825	4 338	67,75	0
MPNA00000000	bin76	Actinobacteria;Acidimicrobiia;Acidimicrobiales;OM1 clade;[a]	90,09	82	102	3,9	61,59	108 948	3 269	65,65	0
MPNQ00000000	bin134	Actinobacteria;Acidimicrobiia;Acidimicrobiales;Sva0996 marine group;[a]	90,09	77	224	4,1	64,29	91 761	3 487	68,71	1
MPNL00000000	bin119	Deinococcus-Thermus;Deinococci;Deinococcales;Trueperaceae;Truepera;[c]	91,89	91	62	3,5	62,23	62 429	2 876	69,47	1
MPMK00000000	bin9	Cyanobacteria;Cyanobacteria;Subsection I;Family I;uncultured;[a]	89,19	391	157	3	58,71	12 771	2 808	50,68	2
MPML00000000	bin5	Chloroflexi;Caldilineae;Caldilineales;Caldilineaceae;uncultured;[a]	92,79	120	81	6	58,5	64 429	4 593	68,26	1
MPMO00000000	bin34	Chloroflexi;Caldilineae;Caldilineales;Caldilineaceae;uncultured;[a]	90,99	111	46	5,1	63,15	63 615	3 982	63,01	0
MPMM00000000	bin22	Chloroflexi;SAR202 clade;[d+]	90,99	58	94	4,8	59,2	163 655	4 049	57,08	4
MPNN00000000	bin127	Chloroflexi;SAR202 clade;[d+]	90,09	21	74	3,3	56,35	309 970	2 976	59,98	0
MPND00000000	bin90	Chloroflexi;SAR202 clade;[a]	89,19	213	61	5,2	57,14	50 603	4 453	54,91	5
MPMJ00000000	bin16	Chloroflexi;SAR202 clade;[a]	88,29	101	70	3,7	65,63	62 928	3 253	60,62	0
MPNC00000000	bin87	Chloroflexi;SAR202 clade;[d]	90,99	67	331	5,4	62,79	269 076	4 711	55,47	1
MPNM00000000	bin125	Chloroflexi;[c]	92,79	66	757	4	62,27	125 355	3 410	61,73	1
MPMG00000000	bin20	Chloroflexi;[c]	91,89	22	245	4	59,31	250 998	3 218	72,5	2

Abbreviations: est. – estimated; com. – completeness; cov. – coverage; no. – number; dupl. – duplicates; **Phylogenetic information**: [a] LCA SILVA (SINA); [b] LCA greengenes (SINA); [c] RDPclassifier; [d] concatenated gene tree; [d+] concatenated gene tree + 16S rRNA gene tree

Table 3-20 Reference genomes for comparison with binned genomes of the PacBio-Illumina hybrid assembly.

Accession	Phylogeny	Source	% est. com.	Contig no.	Length (Mb)	% GC	N50	CDS no.	COG no.	% in COGs
NC_012483.1	Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);Acidobacterium;*Acidobacterium capsulatum* ATCC51196	acidic mine drainage	94.59	1	4.1	60.50	4 127 356	3 343	2 290	68.50
NC_008536.1	Acidobacteria;Acidobacteria;Subgroup 3;unknown;Candidatus Solibacter;*Candidatus* Solibacter usitatus Ellin6076	pasture (soil)	93.69	1	10.0	61.90	9 965 640	8 102	4 755	58.69
CP001631	Actinobacteria;Acidimicrobiia;Acidimicrobiales;Acidimicrobiaceae;Acidimicrobium;*Acidimicrobium ferrooxidans* DSM10331	hot spring runoff	86.49	1	2.2	68.29	2 158 157	2 029	1 419	69.94
NC_009953.1	Actinobacteria;Actinobacteria;Micromonosporales;Micromonosporaceae;Salinispora;*Salinispora arenicola* CNS-205	tropical marine sediment	96.40	1	5.8	69.52	5 786 361	4 977	3 074	61.76
NC_013501.1 - NC_013502.1	Bacteroidetes;Cytophagia;Order II;Rhodothermaceae;Rhodothermus;*Rhodothermus marinus* DSM4252	shallow marine hot spring	95.50	2	3.4	64.30	3 261 604	2 918	2 053	70.36
NZ_ABVV01000000.1	Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;NS5 marine group;*Flavobacteria bacterium* MS024-2A	coastal marine bacterium	88.29	17	1.9	35.71	371 257	1 762	1 221	69.30
NZ_CH724148.1 NZ_AAOG01000007.1	Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;Polaribacter;*Polaribacter irgensii* 23-P	Antarctic surface water	92.79	2	2.8	34.52	2 753 184	2 423	1 583	65.33
NC_014960.1	Chloroflexi;Anaerolineae;Anaerolineales;Anaerolineaceae;Anaerolinea;*Anaerolinea thermophila* UNI-1	thermophilic granular sludge	93.69	1	3.5	53.85	3 532 378	3 161	2 227	70.45
NC_017079.1	Chloroflexi;Caldilineae;Caldilineales;Caldilineaceae;Caldilinea;*Caldilinea aerophila* DSM14535	NA	95.50	1	5.1	58.77	5 144 873	4 127	3 059	74.12
AQTZ01000000.1	Chloroflexi;SAR202 clade;SAR202 cluster bacterium SCGC-AAA240-N13	HOT station ALOHA	25.23	211	1.5	55.02	12 400	1 567	933	59.54
JYFQ01000000.1	Cyanobacteria;Cyanobacteria;Subsection I;Family I;Candidatus Synechococcus spongiarum 15L	*Aplysina aerophoba*	89.19	229	2.2	59.16	14 814	2 276	1 290	56.68
CH724158.1	Cyanobacteria;Cyanobacteria;Subsection I;Family I;Synechococcus;Synechococcus sp. RS9917	Eliat, Red Sea, 10m depth	94.60	1	2.6	64.46	2 584 918	2 738	1 630	59.53
NC_014221.1	Deinococcus-Thermus;Deinococci;Deinococcales;Trueperaceae;Truepera;*Truepera radiovictrix* DSM17093	hot spring runoff	92.79	1	3.3	68.14	3 260 398	2 995	2 174	72.59
NC_015387.1	Deinococcus-Thermus;Deinococci;Thermales;Thermaceae;Marinithermus;*Marinithermus hydrothermalis* DSM14884	deep-sea hydrothermal vent chimney	93.69	1	2.3	68.08	2 269 167	2 249	1 695	75.37
AXVX01000000.1	Gemmatimonadetes;Gemmatimonadetes;BD2-11 terrestrial group;Gemmatimonadetes bacterium SCGC-AAA240-J22	dark ocean below 200m	49.55	58	0.6	51.60	17 295	623	477	76.57
NC_012489.1	Gemmatimonadetes;Gemmatimonadetes;Gemmatimonadales;Gemmatimonadaceae;Gemmatimonas;*Gemmatimonas aurantiaca* T-27	anaerobic-aerobic sequential batch reactor	94.59	1	4.6	64.27	4 636 964	3 968	2 593	65.35
NC_014355.1	Nitrospirae;Nitrospira;Nitrospirales;Nitrospiraceae;Nitrospira;Candidatus Nitrospira defluvii	municipal wastewater treatment plant	93.69	1	4.3	59.03	4 317 083	4 119	2 514	61.03
ASZN01000000.1	Candidatus Poribacteria;Poribacteria genera incertae sedis;Candidatus Poribacteria WGA3G	*Aplysina aerophoba*	82.88	283	5.4	47.79	51 110	4 716	2 693	57.10
NZ_JASE01000000.1	Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Sulfitobacter;*Sulfitobacter geojensis*	coastal seawater	94.59	5	4.2	57.84	3 790 153	3 990	3 101	77.72
NC_016617.1	Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;Azospirillum;*Azospirillum brasilense* Sp245	rhizosphere and plant tissues	91.89	7	7.5	68.49	1 766 028	6 985	4 852	69.46
NC_016803.1	Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;Desulfovibrio;*Desulfovibrio desulfuricans* ND132	found in soil, freshwater, saltwater, intestinal tract of animals	93.69	1	3.9	65.21	3 858 580	3 515	2 578	73.34
NZ_JOMG01000000.1	Proteobacteria;Deltaproteobacteria;Desulfuromonadales;Desulfuromonadaceae;Pelobacter;*Pelobacter seleniigenes* DSM18267	wetlands, freshwater	91.89	6	5.1	54.19	3 076 292	4 448	3 354	75.40
NC_008738.1 – NC_008740.1	Proteobacteria;Gammaproteobacteria;Alteromonadales;Alteromonadaceae;Marinobacter;*Marinobacter hydrocarbonoclasticus* VT8	hydrocarbon polluted marine environments, survives in open seawater	95.50	3	4.8	56.93	4 326 849	4 453	3 161	70.99
NC_008260.1	Proteobacteria;Gammaproteobacteria;Oceanospirillales;Alcanivoraceae;Alcanivorax;*Alcanivorax borkumensis* SK2	seawater sediment, North Sea	94.59	1	3.1	54.73	3 120 143	2 793	2 239	80.16
NC_015436.1	Spirochaetae;Spirochaetes;Spirochaetales;Spirochaetaceae;Sphaerochaeta;*Sphaerochaeta coccoides* DSM17374	termite hindgut	89.19	1	2.2	50.56	2 227 296	1 913	1 425	74.49
NZ_HG422173.1 - NZ_HG422176.1	Nitrospinae/Tectomicrobia group;Nitrospinae;Nitrospinales;Nitrospinaceae;Nitrospina;*Nitrospina gracilis* 3/211	seawater	92.79	4	3.1	56.21	3 072 611	2 934	1 940	66.12
mgm4539207.3	SBR1093;clade I;SBR1093 HKSP	industrial wastewater treatment plant	89.19	184	2.7	56.49	26 372	2 585	1 802	69.71

Abbreviations: est. - estimated; com. - completeness; no. - number

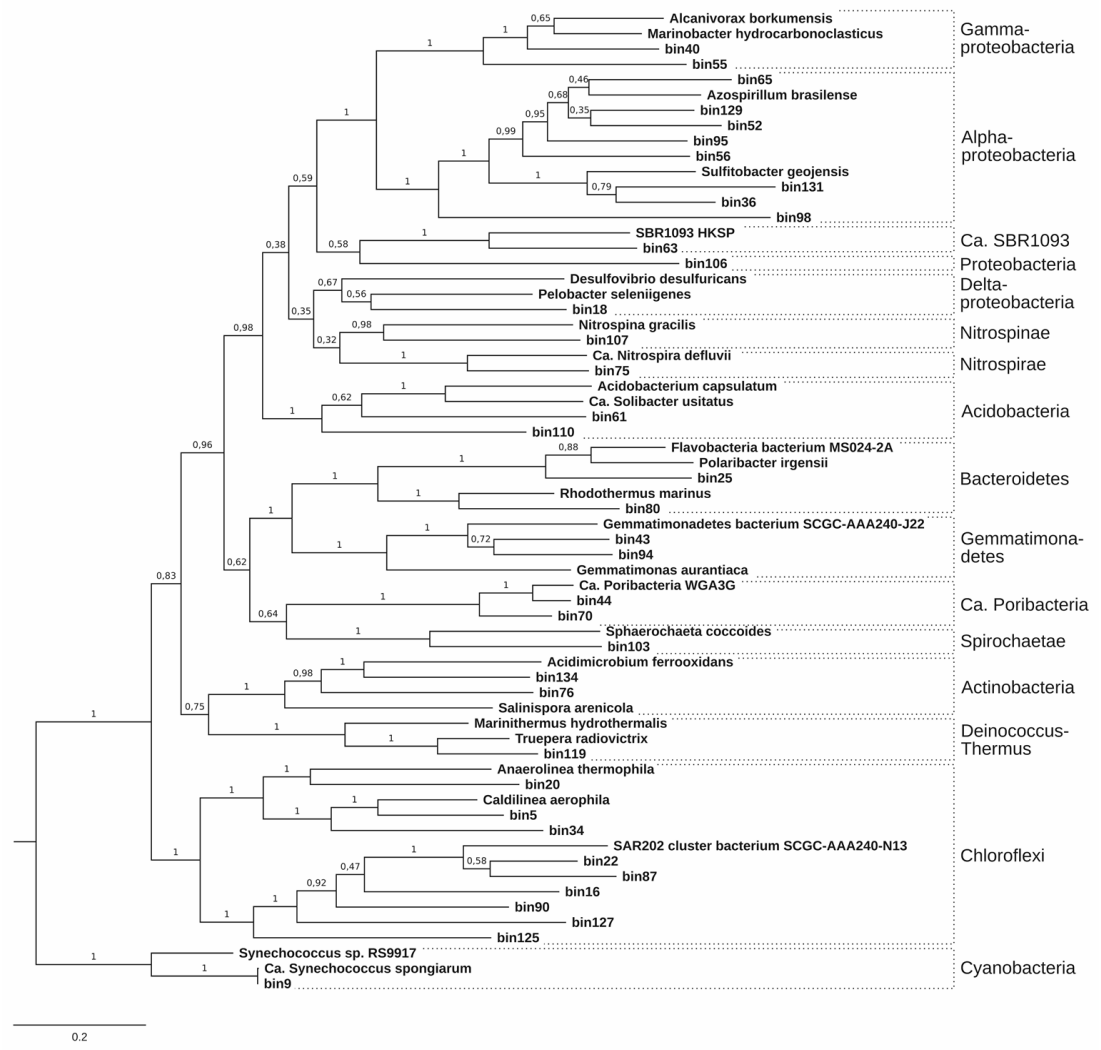


Figure 3-23 Maximum likelihood (LG+G+I) phylogenetic tree based on the amino acid sequences of 29 essential genes, calculated in MEGA7 with 100 bootstrap replications. Cyanobacteria were used as outgroup, because they were closest to the Archaeal outgroup in the 16S rRNA gene phylogeny (Figure 3-24).

In order to resolve the phylogenies of the recovered bins, a concatenated tree (Figure 3-23) of 29 essential single-copy genes (Table 3-20) as well as a 16S rRNA gene tree were constructed (Figure 3-24). Overall, the phylogeny of the binned bacterial genomes reflected the major phylogenetic lineages known to inhabit sponges (Thomas *et al.*, 2016). This finding suggests that the sequenced lineages are prevalent in *A. aerophoba*, as more abundant taxa were more likely sequenced than rare lineages from this diverse metagenome. Our hypothesis that the binned genomes derive from symbionts and not from environmental bacteria was further supported by the 16S rRNA gene data. The best BLAST hits for all 34 bin-derived 16S rRNA genes were from sponge-associated or sponge/coral-associated bacteria (Appendix 3-1). As the remaining three bins did not contain a 16S rRNA gene, their identity could not be confirmed by BLAST alone.

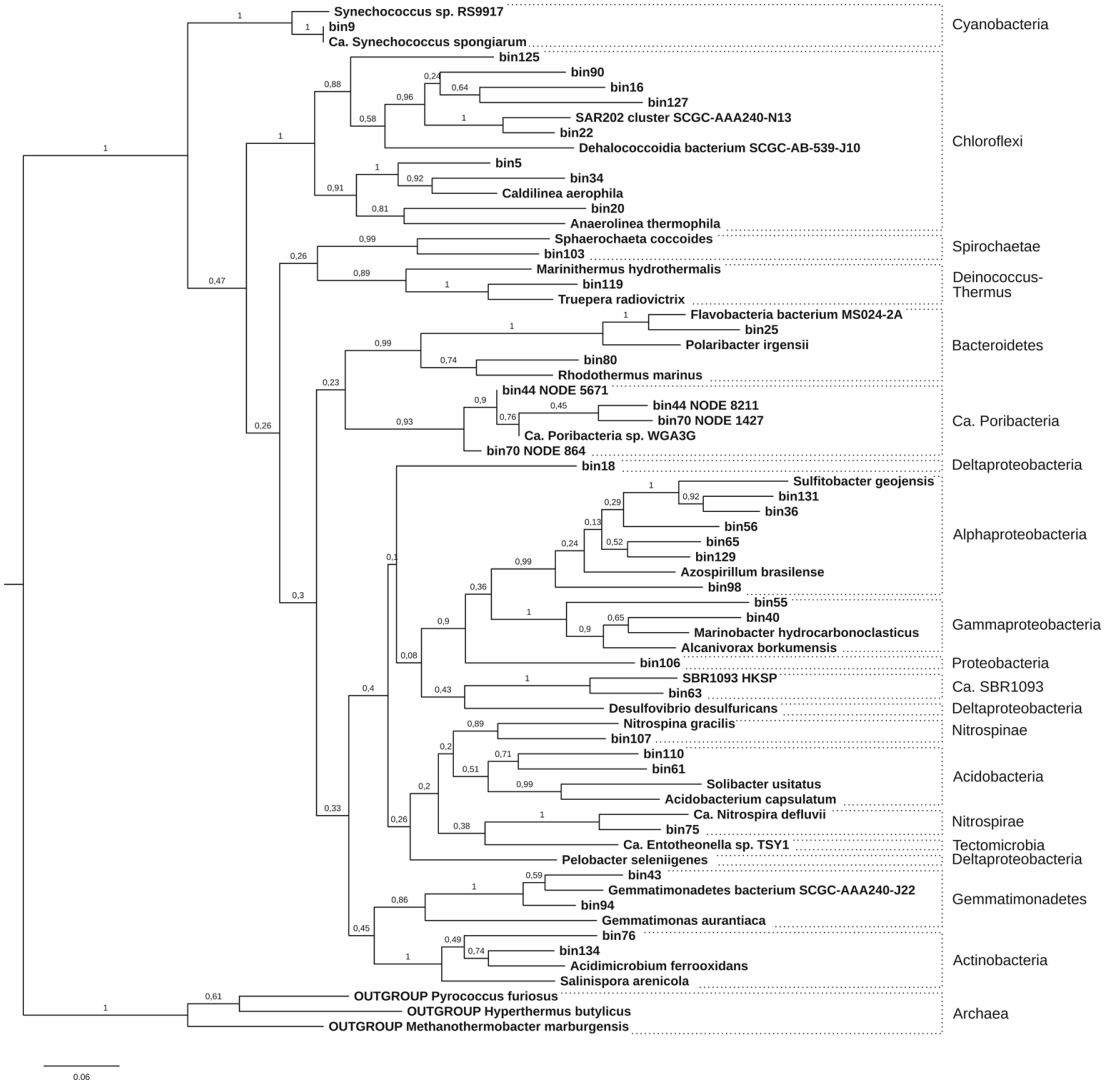


Figure 3-24 Neighbor Joining tree (GTR+G+I) of 16S rRNA genes from hybrid assembly bins and their references with 100 bootstrap replications. The following references were added to this tree only for better phylogenetic resolution: Dehalococcoidia bacterium SCGC-AB-539-J10 (ARPL01000017.1), "Candidatus Entotheonella" sp. TSY1 (KF926817.1), *Pyrococcus furiosus* (NR_074375.1), *Hyperthermus butylicus* (NR_102938.1), *Methanothermobacter marburgensis* (NR_028241.1).

The concatenated tree shows the phylogenetic placement of all 37 bins and their references which had been selected based on genome completeness, phylogenetic similarity, and habitat (marine preferred over other habitats), (Table 3-21). It was in overall agreement with the 16S rRNA gene tree regarding the phylogenetic placement of the bins containing this gene and furthermore provides placement for the three bins missing the 16S rRNA gene.

Table 3-21 Best BLAST hits for the 16S rRNA genes of the PacBio-Illumina hybrid assembly bins.

Genome	Query coverage	Identity	BLAST hit accession	description	Sponge-derived	Sponge species
bin131	94%	94%	JX206636.1	Uncultured bacterium clone TO10-919_C31	yes	*Ircinia oros*
bin56	95%	99%	JX206526.1	Uncultured bacterium clone AF10-915_C9	yes	*Ircinia fasciculata*
bin52	no 16S rRNA gene available					
bin65	95%	99%	JX206637.1	Uncultured bacterium clone TO10-919_C32	yes	*Ircinia oros*
bin36	94%	97%	EF076136.1	Uncultured alpha proteobacterium clone AD015	yes	*Agelas dilatata*
bin98	86%	92%	EU350888.1	Uncultured alpha proteobacterium clone HAL-T-35	yes	*Haliclona simulans*
bin129	98%	99%	HE817825.1	Uncultured bacterium partial 16S rRNa gene, clone B293/GW947	yes	*Vaceletia crypta*
bin95	no 16S rRNA gene available					
bin107	94%	99%	JX206591.1	Uncultured bacterium clone TO10-97_C15	yes	*Ircinia oros*
bin63	97%	99%	JN002375.1	Uncultured microorganism clone WGA_alt_1-4E-16S(clone3-3)	yes	*Aplysina aerophoba*
bin18	97%	99%	JX206694.1	Uncultured bacterium clone TV10-97_C2	yes	*Ircinia variabilis*
bin40	99%	92%	CP013251.1	Endozoicomonas montiporae CL-33	no (coral)	-
	97%	92%	KC669143.1	Uncultured bacterium clone 15E04	no (coral)	-
	97%	92%	AB205011.1	Spongiobacter nickelotolerans gene for 16S rRNA	yes	-
bin55	90%	93%	KF286003.1	Uncultured gamma proteobacterium clone CtgComparison_34	yes	*Aplysina cauliformis*
bin106	83%	98%	AM259914.1	Uncultured proteobacterium partial 16S rRNA gene, clone CN28	yes	*Chondrilla nucula*
bin75	98%	99%	JQ359623.1	Uncultured bacterium clone bac37	yes	*Xestospongia testudinaria*
bin110	95%	99%	JX206593.1	Uncultured bacterium clone TO10-97_C17	yes	*Ircinia oros*
bin61	97%	99%	FJ269286.1	Uncultured Acidobacteria bacterium clone XA2H05F	yes	*Xestospongia testudinaria*
bin70 (NODE_864)	100%	96%	AY713479.1	Uncultured Poribacteria bacterium 64K2	yes	-
bin70 (NODE_1427)	97%	96%	AY713479.1	Uncultured Poribacteria bacterium 64K2	yes	-
bin44 (NODE_5671)	97%	99%	AY713479.1	Uncultured Poribacteria bacterium 64K2	yes	-
bin44 (NODE_8211)	99%	97%	AY713479.1	Uncultured Poribacteria bacterium 64K2	yes	-
bin80	97%	99%	JX206706.1	Uncultured bacterium clone TV10-97_C25	yes	*Ircinia variabilis*
bin25	96%	98%	JN655253.1	Uncultured bacterium clone AF10-3-9_C14	yes	*Ircinia fasciculata*
bin43	95%	98%	HQ270243.1	Uncultured bacterium clone XA2F08F	yes	*Xestospongia testudinaria*
bin94	94%	98%	JX280155.1	Uncultured bacterium clone BA01-C14-seq	yes	*Ircinia felix* tan morph
bin103	95%	98%	JQ612254.1	Uncultured bacterium clone GBc085	yes	*Geodia barretti*
bin76	98%	99%	KC669080.1	Uncultured bacterium clone 14H01	no (coral)	-
	98%	98%	FJ229928.1	Uncultured actinobacterium clone XA3F02F	yes	*Xestospongia testudinaria*
bin134	95%	99%	JX206600.1	Uncultured bacterium clone TO10-97_C25	yes	*Ircinia oros*
bin119	88%	99%	HQ270284.1	Uncultured Truepera sp. Clone XE1D04	yes	*Xestospongia muta*
bin9	99%	99%	KJ174471.1	Candidatus Synechococcus spongiarum SH4	yes	*Carteriospongia foliascens*
bin5	99%	99%	FJ560485.1	Uncultured Chloroflexi bacterium li19	yes	*Aplysina aerophoba*
bin34	97%	99%	JX206705.1	Uncultured bacterium clone TV10-97_C23	yes	*Ircinia variabilis*
bin22	95%	98%	EF076083.1	Uncultured Chloroflexi bacterium clone PK016	yes	*Plakortis* sp.
bin127	95%	98%	JQ612181.1	Uncultured bacterium clone GBc144	yes	*Geodia barretti*
bin90	97%	98%	HE985083.1	Uncultured bacterium partial 16S rRNA gene, clone A48/GW950	yes	*Astrosclera willeyana*
bin16	95%	99%	JX206718.1	Uncultured bacterium clone TV10-912_C6	yes	*Ircinia variabilis*
bin125	95%	98%	FJ481340.1	Uncultured Chloroflexus sp. Clone XB3G04F	yes	*Xestospongia muta*
bin20	87%	99%	AJ347043.1	Uncultured bacterium 16S rRNa gene, clone TK35	yes	*Aplysina aerophoba*
bin87	no 16S rRNA gene available					

3.3.4 Symbiont-reference comparison

In order to identify the gene functions that are enriched in the genomes of sponge symbionts, we compared the pool of symbiont genomes against the pool of selected reference genomes. Significant differences were identified between the symbiont genomes and reference genomes on the level of COG classes. While COG classes R (‘General function prediction only’), E (‘amino acid transport and metabolism’), L (‘replication, recombination and repair’), and Q (‘secondary metabolites biosynthesis, transport, and catabolism’) are enriched in the

symbionts, the classes T (‘signal transduction mechanisms’), K (‘transcription’), M (‘cell wall/membrane/envelope biogenesis’), and N (‘cell motility’) were depleted in comparison to the reference genomes (Figure 3-25).

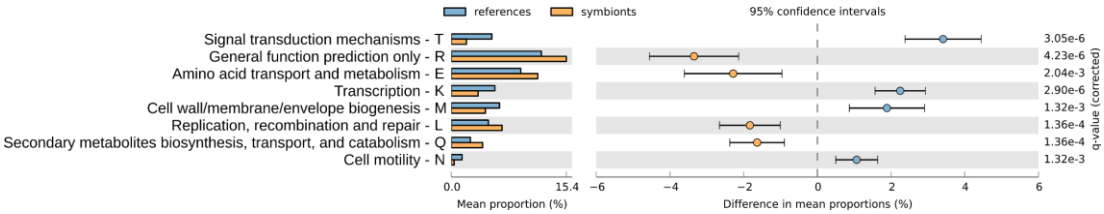


Figure 3-25 Welch’s *t* test on COG classes with Storey FDR at a *q*-value cutoff of 0.01 and a confidence interval of 95%.

When comparing on the level of individual COGs, 42 symbiont-enriched genes were identified (Figure 3-26). Most of these (43%) belonged to COG classes R and S (‘general function prediction only’ and ‘function unknown’), a large fraction (19%) belonged to class V (‘defense mechanisms’), and five (12%) to class L (‘replication, recombination and repair’). According to the STRING database, many of these significantly symbiont-enriched COGs were likely interacting (Figure 3-27). At a high confidence cut-off (0.700 minimum required interaction score), five networks (A-E) comprising 17, 6, 3, 2, and 2 COGs were obtained. The remaining 12 symbiont-enriched COGs did not interact with any other COGs in the list. The set includes a restriction endonuclease (COG2810) and a bacteriophage protein gp37 (COG4422). The largest STRING network was built of sponge-enriched COGs related to restriction-modification (RM) with endonucleases, helicases and methylases (cluster A in Figure 3-27, see Appendix 3-3 for COG counts). It was present in all sponge symbiont phyla in this study (Figure 3-28).

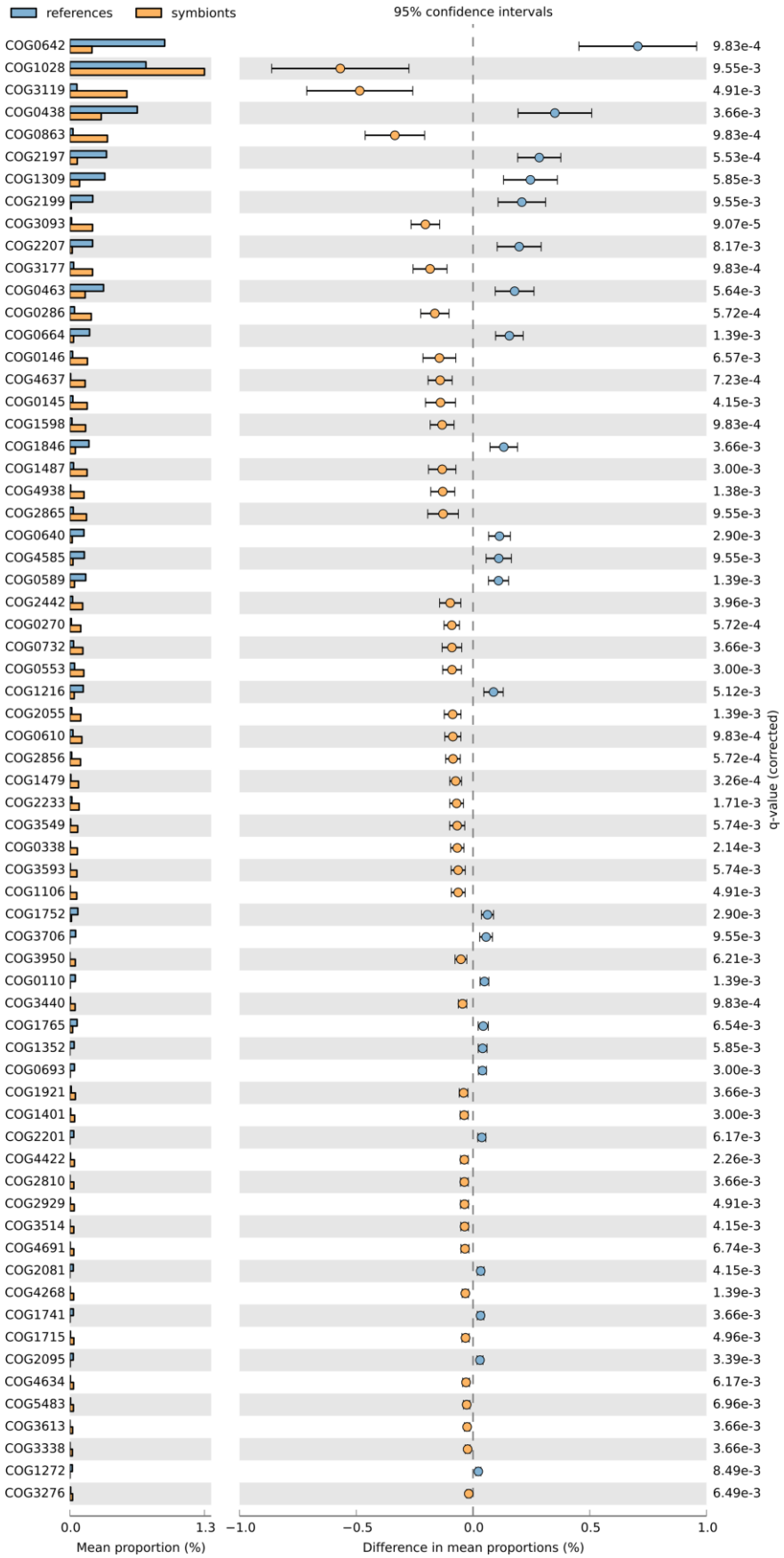


Figure 3-26 Welch's *t* test on COGs with Storey FDR at a *q*-value cutoff of 0.01 and a confidence interval of 95%.

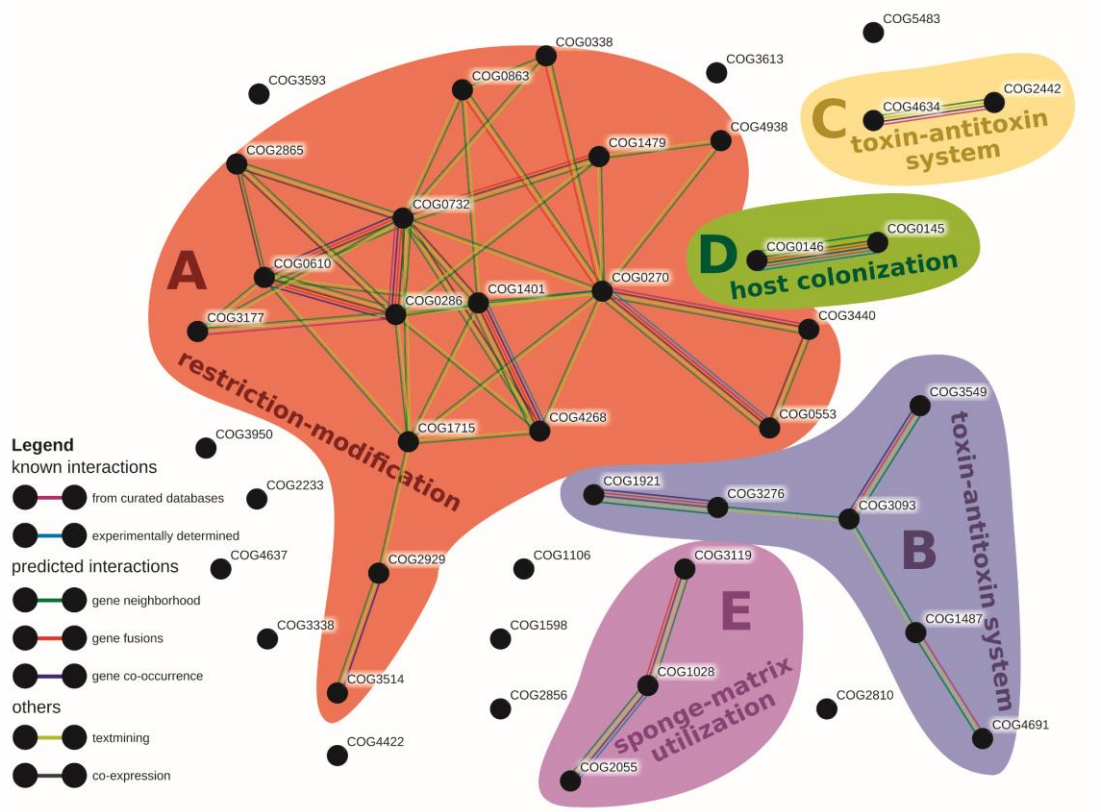


Figure 3-27 STRING network of significantly sponge symbiont-enriched COGs. Colored areas mark COGs that belong to the same network (A-E). Colors of the connectors indicate the type of evidence of the predicted interaction between the two connected COGs. Only connections of 'high confidence' (minimum required interaction score: 0.700) are shown.

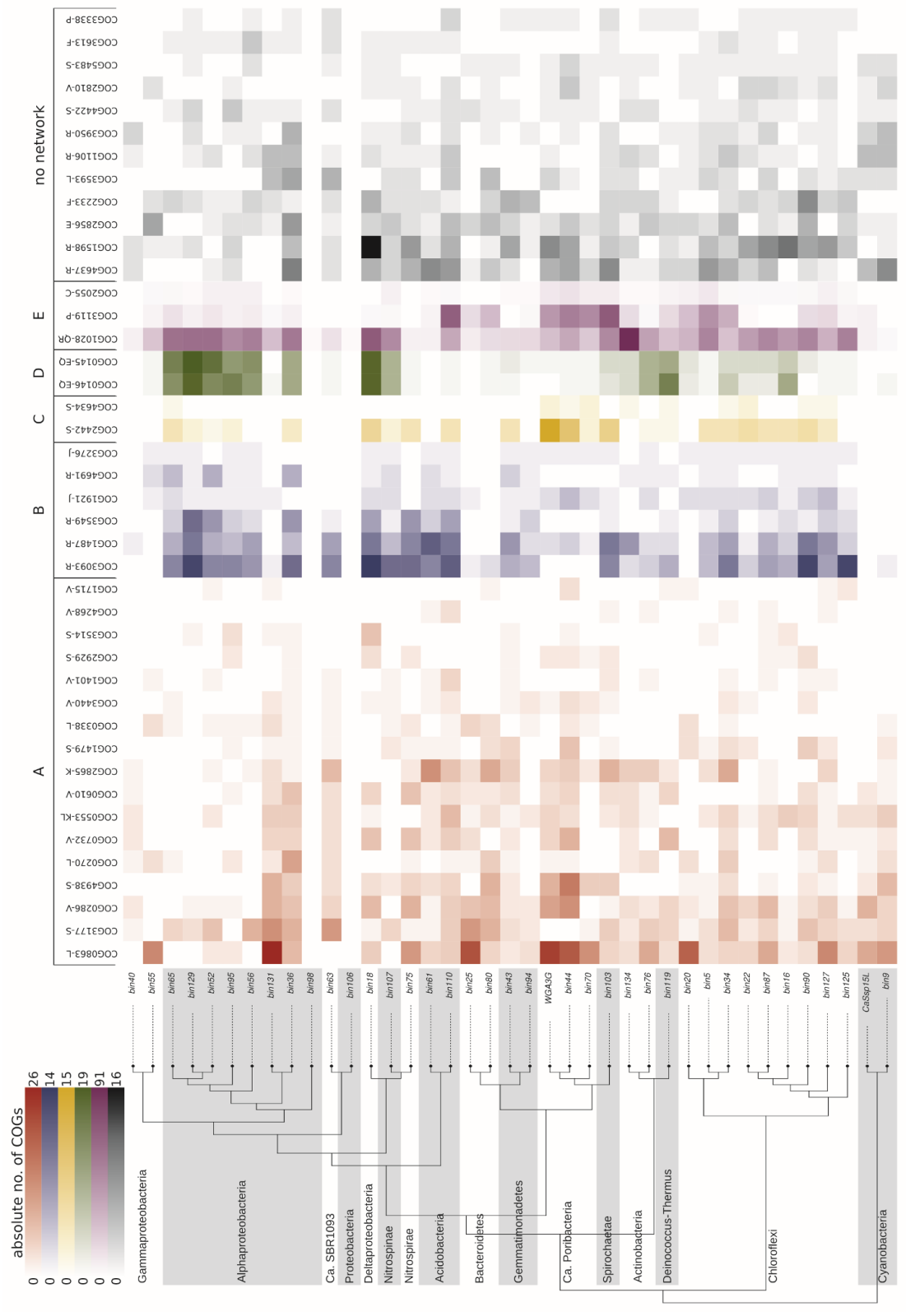


Figure 3-28 Heatmap of significantly sponge symbiont-enriched COGs (absolute counts) in the genomes binned from the PacBio-Illumina hybrid assembly. Phylogenetic relationships of the genomes are indicated by a simplified version of the tree in Figure 3-23 (only sponge symbionts are shown here). Possibly interacting COGs as shown in Figure 3-27 are grouped and colored accordingly and marked by the letters A-E. The letters next to each COG indicate the according COG class.

All COGs of STRING network B were related to toxin-antitoxin (TA) systems. COG3549 and COG3093 form the HigAB TA plasmid maintenance system, and COG1487 encodes for the toxin in a TA system of the VapBC family (Makarova *et al.*, 2009; Sberro *et al.*, 2013). COG4691 is a plasmid stability protein and encodes for a proposed antitoxin of a VapBC TA system (Chen, 2007). COG1921 (SelA) and COG3276 (SelB) co-occurred in the majority of symbiont bins of various phyla but were missing in the majority of their closely related references (Appendix 3-3). STRING network C consists of COG4634 and COG2442, two uncharacterized conserved proteins according to the NCBI annotation. COG4634 is hypothesized to be a fine-tuning modulator in conjugative plasmid transfer (López-Fuentes *et al.*, 2015), and COG2442 is a PIN-associated antitoxin in a widespread TA system most abundant in Cyanobacteria and Chloroflexi (Makarova *et al.*, 2009). Furthermore, COG2929 and COG3514, which are part of network A, were predicted to form a TA system as well (Makarova *et al.*, 2009). Both COGs co-localize on a plasmid of the cyanobacterium *Synechococcus elongatus* PCC7942 where this TA system plays a crucial role in plasmid maintenance (Chen, 2007). In our dataset, both COGs co-occurred in 16 sponge symbiont bins of various bacterial phyla, but only once in the reference group, in the acidobacterium *Solibacter usitatus*.

Symbiont-enriched STRING networks D and E are related to colonization of the host and possibly utilization of the host matrix. COG0145 (hyuA) and COG0146 (hyuB) of network D have been hypothesized to play an important role for *Helicobacter pylori* in the colonization of mice (Zhang *et al.*, 2009). The abundance and distribution of network D across various phyla of sponge-associated bacteria in our study suggests that it may also be of importance for the colonization of sponge hosts. COG1028 (FabG) and COG3119 (arylsulfatase A) of network E displayed the highest counts within the sponge-enriched COGs. Arylsulfatase A might allow the symbionts to metabolize sulfated polysaccharides from the sponge extracellular matrix, where their abundance has been documented (Vilanova *et al.*, 2009; Zierer and Mourão, 2000).

Additional recurring topics in sponge-microbial symbioses are CRISPR-Cas systems and eukaryotic-like domains, the former related to bacterial defense against foreign DNA, the latter related to host interaction. Both features are enriched in the symbiont group, albeit not to a statistically significant degree. More common and also more abundant in sponge symbionts are Cas1, Cas2, a Cas2 homolog, Cas3 and a predicted CRISPR-associated nuclease (COG1518, COG1343, COG3512, COG1203, and COG3513, respectively). Eukaryote-like repeat domain containing proteins, such as Leucine-rich repeat proteins (COG4886) are present in 56% of the sponge symbionts and in 12% of the references, with up to 30 copies per genome in the symbionts and a maximum of 5 copies in the references. Likewise, ankyrin repeats (COG0666) show up to 29 copies per genome in the sponge symbionts and a maximum

of 11 copies in the references. Both of these eukaryotic-like proteins are more common and more abundant in the symbiont group and therefore likely represent a sponge-symbiosis specific feature facilitating escape from phagocytosis by the sponge host (Fan *et al.*, 2012; Liu *et al.*, 2012; Thomas *et al.*, 2010).

3.3.5 Within-symbiont comparison

In order to compare the symbiont genomes among each other and to identify functional groups, a principle component analysis (PCA) was performed. The functional grouping is only partly coherent with phylogeny (Figure 3-29). While Gemmatimonadetes cluster closely together, Chloroflexi are split up in two groups: i) Caldilineae that built a group with Poribacteria and Spirochaetae, and ii) SAR202 clustering with a group of Alphaproteobacteria, Deltaproteobacteria, Nitrospinae, and Actinobacteria.

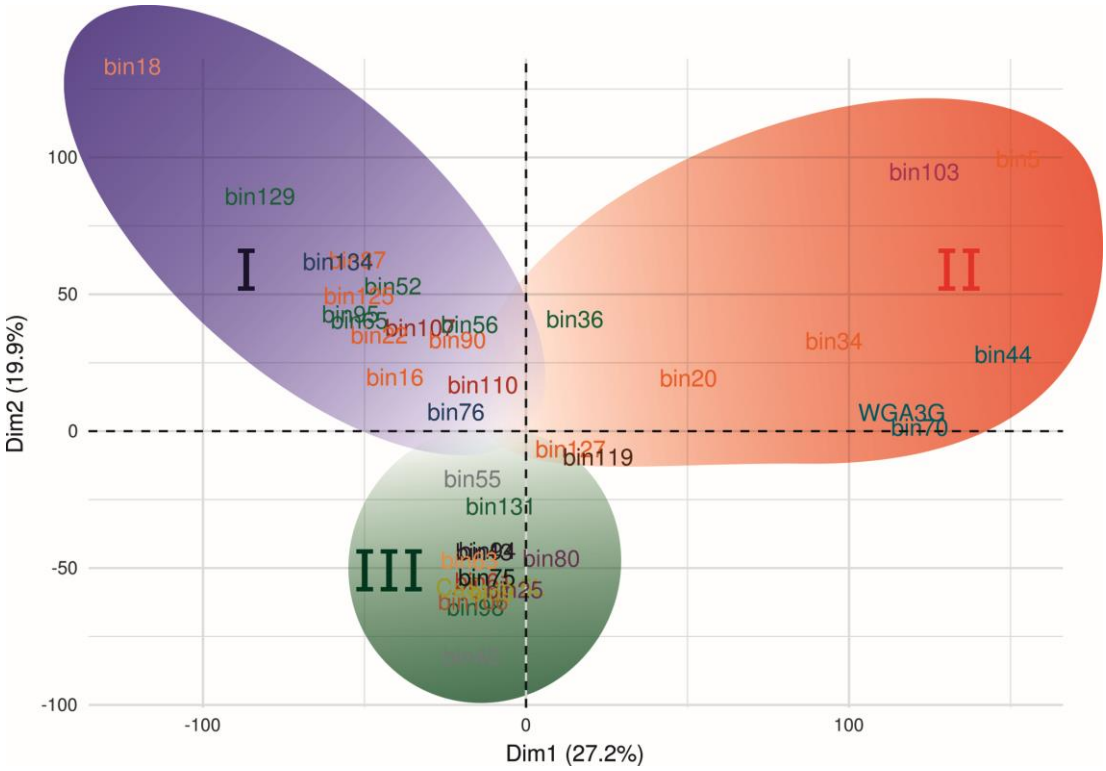


Figure 3-29 PCA plot comparing the genomes of the sponge-symbionts to each other based on their COG annotation. Phylogenetic affiliation is indicated by font colors (see Table 3-19 for details). The symbionts build three groups I-III marked by bacskground color (blue, red, and green, respectively).

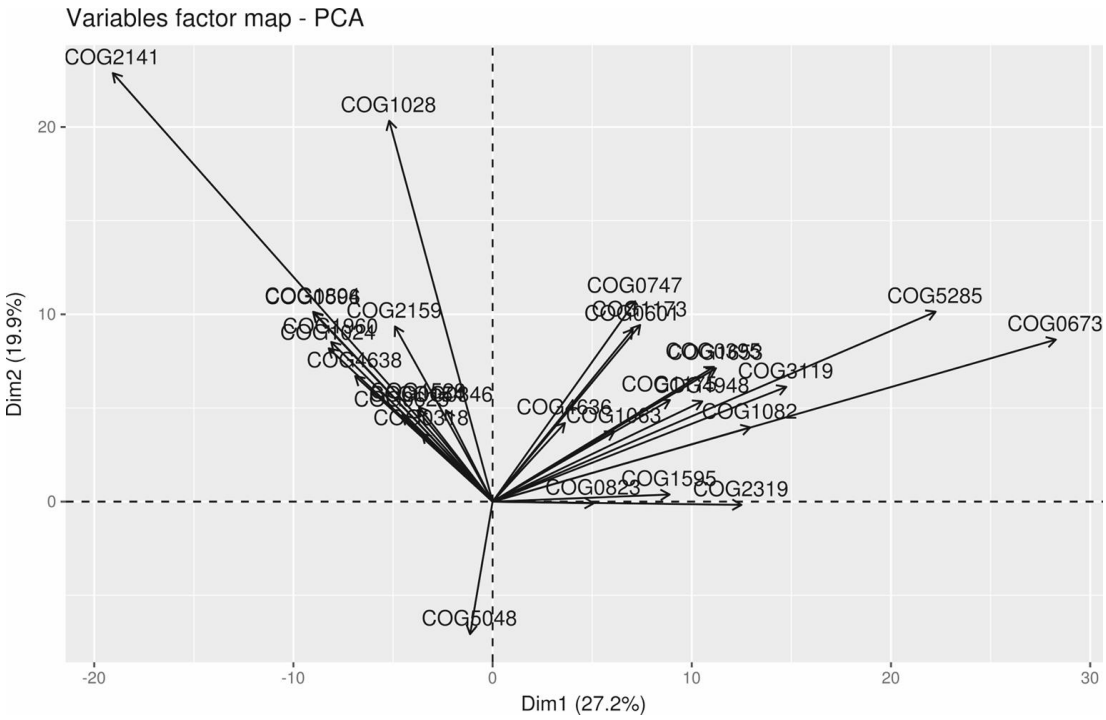


Figure 3-30 The 30 COGs with the strongest influence on the PCA grouping of the sponge symbiont genomes.

The 30 COGs with the greatest influence on the functional grouping of the sponge symbionts are shown in Figure 3-30. According to this analysis, the COGs enriched in symbiont group I are mainly involved in metabolism and energy production. Most enriched in this group are COGs related to carnitine metabolism. Carnitine is an organic compatible solute that some bacteria can use as a source for carbon, nitrogen, and energy (Meadows and Wargo, 2015).

Symbiont group II is characterized by high numbers of arylsulfatase A genes (COG3119), various ABC transporters, and dehydrogenases. This phylogenetically heterogeneous guild of microorganisms seems to be specialized on the utilization of sulfated polysaccharides, as described above for symbiont-enriched COG network E. Inspection of the genomic context on the bin-level shows that the arylsulfatase repeatedly clusters with the ABC transporters and the dehydrogenase that are likewise enriched in symbiont group II (Figure 3-31). This further supports our hypothesis that this gene cluster is of importance for sponge symbionts, and especially for the members of symbiont group II.

Figure 3-31 Typical gene cluster around the arylsulfatase A gene (AslA, shown in red).

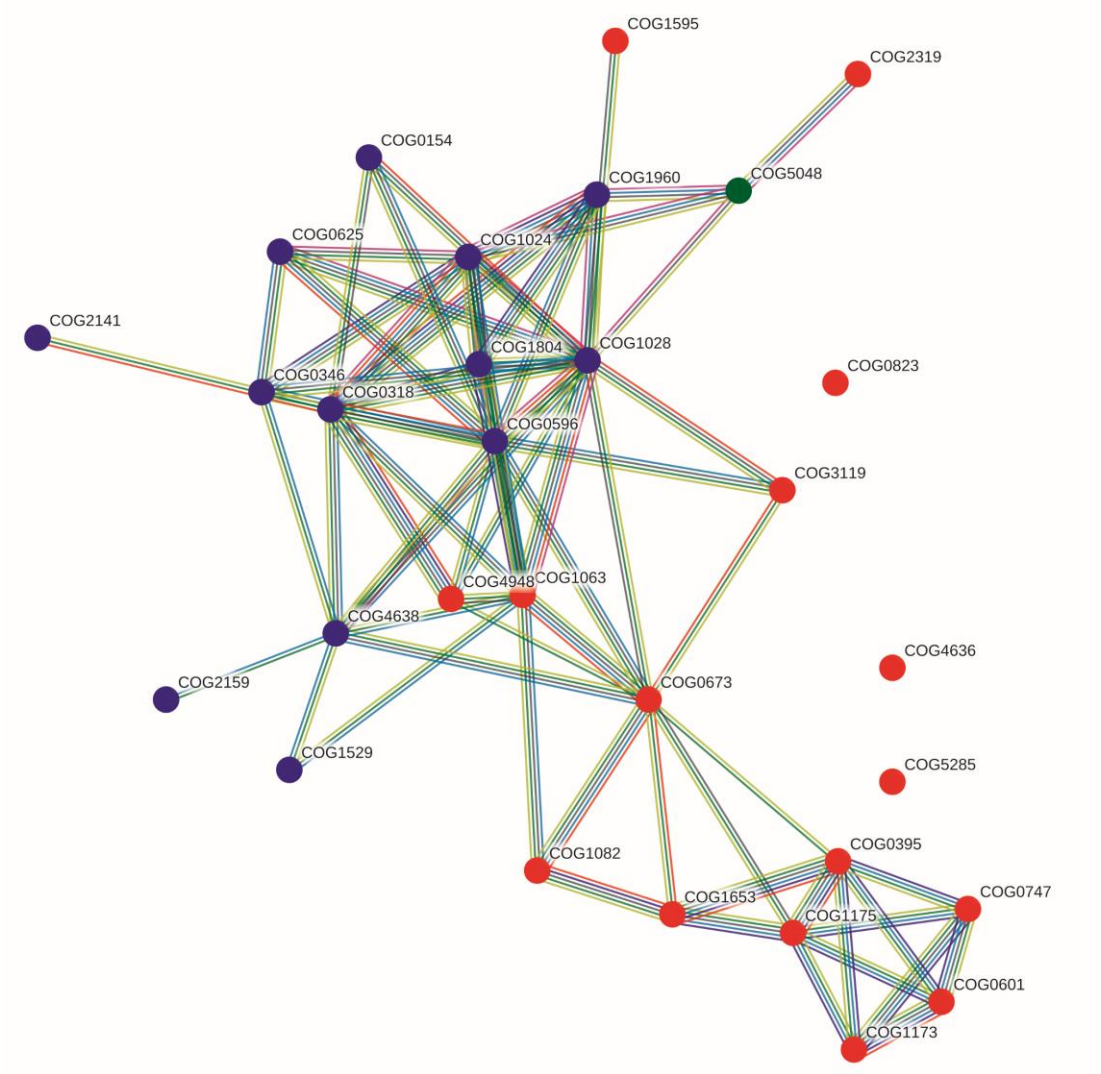


Figure 3-32 STRING network of the 30 COGs contributing most to the grouping of the sponge-symbionts in Figure 3-29. Circles representing the COGs' position in the network are colored according to the symbiont group where they are overrepresented. Colors of the connectors indicate the type of evicence of the predicted interactions between the two connected COGs as shown in Figure 3-27. Only connections of 'high confidence' (minimum required interaction score: 0.700) are shown.

The genomes of symbiont group III did not show an enrichment of any particular COGs. They also contained the COGs of symbiont groups I and II, but not in as high numbers. Thus, we posit that symbiont group III is not metabolically specialized and may represent a group of metabolic generalists. Within the 30 COGs most responsible for the grouping, only COG5048 (FOG: Zinc-finger) was enriched in bin40 of this group with a total of 159 copies. Zn-fingers are small structural protein motifs commonly found in eukaryotes, but also present in prokaryotes where they are likely involved in virulence or symbiosis (Malgieri *et al.*, 2015).

Most COGs of symbiont groups I, II, and III are strongly connected according to a STRING network with the COGs enriched in groups I and II clustering on different sides of the network (Figure 3-32). The symbionts of group III are able to perform the same metabolic pathways as the two specialized groups, however without possessing such high numbers of the

corresponding genes (Figure 3-33). They may be considered as nutritional generalists in the microbial consortium.

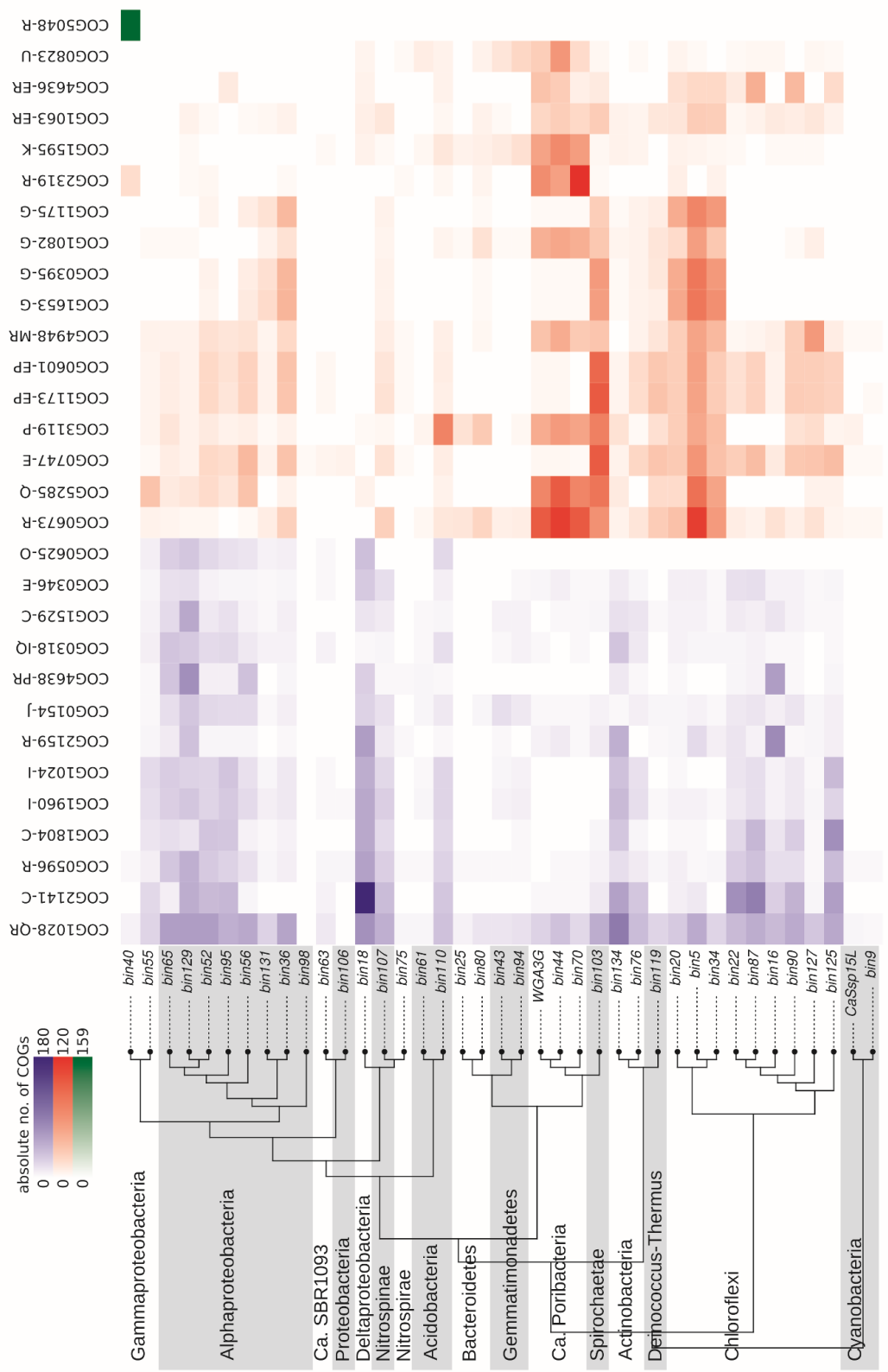


Figure 3-33 Heatmap of the 30 COGs contributing most to the grouping (absolute counts) of the sponge-symbionts as shown in Figure 3-29. Phylogenetic relationships of the genomes are indicated by a simplified version of the tree in Figure 3-23 (only sponge symbionts are shown here). Colors represent the symbiont group where the regarding COGs is overrepresented. The letters next to each COG indicate the according COG class.

4 Discussion

4.1 "*Ca.* Synechococcus spongiarum" group – closely related but different in gene content

4.1.1 An optimal candidate for 'mini-metagenomics'

Due to the autofluorescence of "*Ca.* Synechococcus spongiarum," FACS sorting was a perfectly suitable approach for this cyanobacterial sponge symbiont and a high level of purity could be achieved as the screening results showed (Table 3-1, Table 3-2). The purity assessment only aimed at the 16S rRNA gene and ITS region. Therefore, other free DNA fragments could not be detected before sequencing, which is a common issue in single-cell sequencing projects and necessitates decontamination of the genome assembly (Woyke *et al.*, 2010). Even the MDA reagents themselves have been shown to be possible sources of contamination and protocols have been developed to avoid co-amplification of contaminants beforehand (Woyke *et al.*, 2011). When symbionts are as abundant in the host as "*Ca.* Synechococcus spongiarum," genomes of comparable quality could also be retrieved by metagenomic binning as it was done for the three other strains in the following comparison (Burgsdorf *et al.*, 2015). For low-abundance members of the microbial community, on the other hand, FACS sorting may be of great advantage to sort single cells or mini-metagenomes. One disadvantage of this approach is that, only if a distinct physical property of the target cell – e.g. autofluorescence – is available, is it possible to sort a pure mini-metagenome, because the sorting window could not otherwise be set sufficiently narrow just sorting by cell size. The right target cell would need to be identified by PCR screening after MDA instead. One way to avoid this issue is using in-solution fixation-free fluorescence in situ hybridization (FISH) to label the target cells for FACS sorting, if the target cells and the sample are suitable for this method (Haroon *et al.*, 2013). A disadvantage of this method is a likely decrease in genome quality (Clingenpeel *et al.*, 2014). Given that the MDA comes with flaws such as uneven amplification, chimera formation, and co-amplification of contaminants (Blainey, 2013), may also be worthwhile to look for ways to avoid this amplification step e.g. by FACS sorting a sufficient amount of target cells to be used directly for sequencing.

In comparison to metagenomic binning and 'mini-metagenomics', genomics on real single-cells has the advantage of a comparison on the level of individual cells. In this project, opting for the mini-metagenomes, I did not take advantage of this aspect and rather focused on clade-level and species-level comparisons. Yet, with the advance in knowledge about sponge symbionts in general, increasing effort may be put in the study of individual members of the microbial community to study genome evolution, niche differentiation, and speciation,

as has been conducted in similar fields of research, e.g. for coexisting subpopulations of free-living marine *Prochlorococcus* (Kashtan *et al.*, 2014).

4.1.2 Lifestyle evolution in cyanobacterial symbionts of sponges

The cyanobacterium "*Ca.* Synechococcus spongiarum" is a common member of sponge microbial communities in a variety of host species and geographic locations. This raises the question of how conserved its genome is, taking recent studies into account that have reported different productivity and carbon assimilation and transfer abilities for genetically distinct "*Ca.* Synechococcus spongiarum" clades (Freeman *et al.*, 2013). Here, four clades of this symbiont species are compared, which are associated to four different host sponge species from two geographic locations. Despite a 16S rRNA gene identity above 98.6%, they shared only around half of their protein coding genes per genome. The clades may be highly variable and adapted to their particular host sponge and environment. This great difference in gene content is surprising, considered that two strains (coastal and off-shore) of diazotrophic cyanobacterial symbionts (UCYN-A) of prymnesiophyte algae have 96.6% genes in common at a 16S rRNA gene identity of 98.7% (Bombar *et al.*, 2014). On the other hand, the average amino acid sequence identity between orthologous genes within core genomes is higher among the four "*Ca.* Synechococcus spongiarum" clades (>91%) than between the two UCYN-A strains (86%). Interpretation of the significance of this genome divergence is unfortunately limited by the high number of genes of unknown function in these sponge symbionts.

Most of the genomic traits postulated for "*Ca.* Synechococcus spongiarum" SH4 of Red Sea sponge *C. foliascens* (Gao *et al.*, 2014b) were confirmed for three more clades of this symbiont species in this study, and novel, supposedly clade specific features were discovered.

4.1.2.1 Sponge-specific functional genomic signatures

Previous metagenomic comparisons of sponge and seawater microbiomes revealed a clear separation of the sponge bacterial communities from the surrounding seawater (Thomas *et al.*, 2010; Fan *et al.*, 2012; Liu *et al.*, 2012). One of the sponge symbiont-enriched traits confirmed for "*Ca.* Synechococcus spongiarum," is the significantly higher proportion of COGs related to 'recombination and repair' (L). This may enable a stable insertion of mobile DNA into the symbionts' chromosomes by repairing the flanking regions of the newly inserted DNA (Thomas *et al.*, 2010; Fan *et al.*, 2012). Transposable insertion elements present in high numbers in bacterial symbionts have been reported for a variety of host types including the intracellular *Drosophila melanogaster* symbiont *Wolbachia pipientis* wMel (Wu *et al.*, 2004). They may be a driver of microbial adaptation to specific niches (Moliner *et al.*, 2010; Smillie *et al.*, 2011). Among the analyzed "*Ca.* Synechococcus spongiarum" genomes, three of four possess the transposase COG3293, that has been reported as enriched in sponge microbiomes

over planktonic microbiomes before (Fan *et al.*, 2012). This horizontal gene transfer feature and the symbiont-enriched site-specific DNA methylase COG0270 are highly conserved sequences within the symbionts, which suggests the importance of horizontal gene transfer for sponge symbionts.

Previous studies have also shown an enrichment in proteins containing eukaryotic-type domains like ankyrin and tetratricopeptide repeats (TPR), and leucine-rich repeat (LRR) domains for sponge microbiomes in general (Thomas *et al.*, 2010; Fan *et al.*, 2012). Ankyrin and TPR repeats are involved in protein-protein interactions in eukaryotes, LRR proteins are essential for virulence in the pathogen *Yersinia pestis* and for host cell invasion by *Listeria monocytogenes* (Thomas *et al.*, 2010; Evdokimov *et al.*, 2001; Marino *et al.*, 1999). In a model system resembling sponge amoebocytes, sponge symbiont-derived ankyrin repeat proteins have the capacity to modulate phagocytosis of amoebas, when expressed in *Escherichia coli* (Nguyen *et al.*, 2014). Ankyrin repeat protein gene COG0666 was present in four copies in each of the four "*Ca.* Synechococcus spongiarum" genomes, but it was not annotated in any of the free-living cyanobacterial references. Also a sulfur-oxidizing bacterial symbiont of *Haliclona cymaeformis* contains a large number of ankyrin repeat domains (Tian *et al.*, 2014). This leads to the conclusion that ankyrin repeat domains are likely an obligatory feature also for sponge bacterial symbionts, as for other symbiotic systems such as *W. pipientis* in *D. melanogaster* (Wu *et al.*, 2004).

CRISPRs have also been identified as an abundant feature of sponge microbiomes in previous studies (Fan *et al.*, 2012). Together with their associated proteins they are forming adaptive immunity systems that are common among most archaea and many bacteria, acting against invading genetic elements like viruses and plasmids (Makarova *et al.*, 2011). Also in cyanobacteria, CRISPR-Cas systems have been found in the majority of sequenced genomes except the *Synechococcus/Prochlorococcus* subclade (Cai *et al.*, 2013). It has been hypothesized that either its genetic load is too high for the small genomes of the *Synechococcus/Prochlorococcus* subclade, or that the viral diversity outruns the CRISPR-Cas immune system (Weinberger *et al.*, 2012). The latter was suggested by a mathematical model and currently lacks empirical proof. In this study, CRISPR-Cas systems were present in the small-genome-sized and highly phage-exposed "*Ca.* Synechococcus spongiarum" 142, which suggests that the absence of these defense systems in the free-living *Synechococcus/Prochlorococcus* subclade as an alternative explanation. The presence of the CRISPR-based immune system may be the ancestral state that the *Synechococcus/Prochlorococcus* ancestor has lost after the divergence from "*Ca.* Synechococcus spongiarum," or alternatively the sponge symbiont may have acquired it by horizontal gene transfer, likely from other sponge symbionts. A high selective pressure for acquiring phage resistance inside sponges may serve as an explanation for the prevalent

CRISPR-Cas systems in the sponge symbionts. Considering the sponges' high water pumping rates, the associated bacteria, being exposed to approximately 1,000 viral particles per bacterial cell per day (Thomas *et al.*, 2010), likely encounter a multiple of the viral particles their free-living relatives are exposed to. This may explain the retention of CRISPR-Cas systems in sponge symbionts.

4.1.2.2 Common genomic features

Like in mitochondria and chloroplasts, genomic streamlining may eventually lead to a complete dependence of symbionts on the host and to the evolution of organelles (Tripp *et al.*, 2010; Kwan *et al.*, 2012). In "*Ca.* Synechococcus spongiarum," the reduction of genes involved in essential functions is similar to the pattern recently described for the plastid of the amoeba *P. chromatophora* (Nowack *et al.*, 2008). Cytochrome *c* oxidase, carotenoid biosynthesis, and signal transduction regulators, for example, were reduced in all four "*Ca.* Synechococcus spongiarum" genomes as well as the plastid *P. chromatophora* (Nowack *et al.*, 2008) (Table 3-8). It has to be stated that this is only an observed trend, whereas no conclusions can be drawn on the basis of missing genes in unclosed genomes. Yet, the trend is rather similar among all four different clades of "*Ca.* Synechococcus spongiarum," which supports this notion.

A comparably less stable PSII complex has served as explanation for the loss of a number of *psb* genes in SH4 in comparison with free-living cyanobacteria, probably representing an adaptation of the photosynthetic system to low-light conditions (Gao *et al.*, 2014b). This finding was confirmed for three additional "*Ca.* Synechococcus spongiarum" clades. The genes *psbD* and *psbP* were absent in all four symbiont genomes (Table 3-12). As *psbP* may optimize the water-splitting reaction, its absence may lead to a decreased efficiency of the photosynthetic system and a lower competitive potential (Sveshnikov *et al.*, 2007). In *I. variabilis*, there may be competition between different cyanobacterial species due to the abundance of more than one symbiont species (Usher *et al.*, 2006). However, it was shown that the different species are spatially separated, "*Ca.* Synechococcus spongiarum" residing in the pinacoderm, and "*Ca.* Synechococcus feldmanni" and "*Ca.* Aphanocapsa raspaigellae" in the mesohyl matrix (Usher *et al.*, 2006). The gene *psbY* that was missing in three "*Ca.* Synechococcus spongiarum" genomes, has been shown not to be essential for *Synechocystis* sp. PCC6803 for oxygenic photosynthesis (Meetam *et al.*, 1999). In symbiotic cyanobacteria, the loss of nonessential photosynthetic genes may be due to a tradeoff between smaller genome sizes and thus a reduction in genome replication cost, which is payed by a reduced competitive potential (Larsson *et al.*, 2011; Kwan *et al.*, 2012).

A byproduct of aerobic metabolism are reactive oxygen species (ROS) that can cause oxidative damage to photosynthetic organisms like cyanobacteria, that counter this oxidative

stress with antioxidant enzymes (Latifi *et al.*, 2009). Several of these antioxidant enzymes were missing in SH4 (Gao *et al.*, 2014b) as well as the three "*Ca.* Synechococcus spongiarum" genomes analyzed here. Due to their location within the sponge, only reduced amounts of light radiation, and thereby decreased amounts of ROS may reach the symbionts. Also the heterotrophic part of the sponge microbiome in close proximity to the cyanobacteria may reduce the amount of ROS in the sponge tissue by respiration of oxygen immediately after production by the photosymbionts.

A loss of genes involved in the formation of the cell wall in SH4 has been reported previously (Gao *et al.*, 2014b). Furthermore, also the loss of genes responsible for dTDP-L-rhamnose production is common to all analyzed "*Ca.* Synechococcus spongiarum" clades. dTDP-L-rhamnose, a residue of the O antigen of LPS, has been found in free-living marine *Synechococcus* (Snyder *et al.*, 2009). A variation of O antigens alters the Gram-negative bacterial cell wall. For host-microbe interactions the correct structures of the LPS and its O antigen are essential, because they are important to establish disease in pathogens or beneficial outcomes in symbiosis (Lerouge and Vanderleyden, 2001). Planktonic cyanobacteria are part of the typical sponge diet (Pile *et al.*, 1996). Thus, O antigens like dTDP-L-rhamnose and GDP-D-rhamnose may be used for 'food recognition' by the sponge. Already in the 1970s, studies have proposed mechanisms for differentiation between symbionts and food bacteria by the sponge. Either the symbionts would be recognized as such, e.g. via the ankyrin repeats, as recently suggested (Nguyen *et al.*, 2014), or they may evade host phagocytosis by using masking coatings (Wilkinson, 1978b). The masking hypothesis is supported by *in situ* feeding experiments with potential symbionts isolated from sponges versus free-living seawater bacteria in combination with electron radioautography (Wilkinson *et al.*, 1984). Chemical compounds surrounding the bacteria functioning as protective capsules were proposed as a masking mechanism for the symbionts (Wilkinson *et al.*, 1984). Further evidence for a food-symbiont discrimination was provided by later studies on *A. aerophoba* (Wehrl *et al.*, 2007). In "*Ca.* Synechococcus spongiarum," the missing dTDP-L-rhamnose and GDP-D-rhamnose O antigens on the LPS, implied by the absence of the respective biosynthetic genes, may be a mechanism of host phagocytosis resistance, as the symbionts may not be recognized as food bacteria by the sponge host. Contrasting the previously proposed masking mechanism with a protective capsule covering the recognition element on the bacterial cell wall (Wilkinson *et al.*, 1984), a lack of the recognition element itself would achieve host evasion. Supporting this hypothesis, freshwater *S. elongatus* PCC7942 mutants that are deficient in O antigen synthesis, resist amoebal grazing (Simkovsky *et al.*, 2012). Yet, further experiments are needed to test this hypothesis. Additionally, mutations in genes involved in dTDP-L-rhamnose production and transport in the marine *Synechococcus* sp. strain WH7803 have also been shown to be responsible for phage resistance (Marston *et al.*, 2012). This suggests a protective function

against cyanophages for the lack of the O antigen in "*Ca.* Synechococcus spongiarum," a potentially important mechanism due to an enrichment of cyanophages resulting from the sponge pumping activity. In free-living cyanobacteria, the lack of O antigen promotes autoflocculation (Marston *et al.*, 2012), which may not concern a symbiont in the sponge mesohyl while selecting against free-living *Synechococcus* with this characteristic, as they may sink into nonphotic zones.

4.1.2.3 Divergent genomic features

In all four "*Ca.* Synechococcus spongiarum" genomes the methionine salvage pathway (MSP) was only partially present, through which methionine is recycled from 5-methylthioadenosine (Albers, 2009). This suggests that methionine is obtained from other, external sources such as the sponge host or the heterotrophic fraction of the sponge microbiome. In SH4 more MSP genes are missing, which may be explained by the comparably lower completeness of the draft genome. An alternative explanation is a higher rate of genome reduction for SH4. Additionally, the predicted genome sizes for the four symbionts of different "*Ca.* Synechococcus spongiarum" clades varied between 16% and 25% with SH4 as the smallest. This suggests, that the different clades may follow different symbiotic trajectories, which leaves them with differing degrees of genomic streamlining and host dependencies.

The low-molecular-weight compound siderophores are secreted to the environment to bind Fe(III) and get transported back into the cell, which is an energy-dependent mechanism that can include TonB receptors. An extracellular substrate binding protein, an integral membrane protein, and ATPase (ATP hydrolases) build the transport component of ABC-type siderophore systems (Köster, 2001). *Cyanobium* sp. strain PCC7002 is the sole marine cyanobacterium reported so far, that harbors genes for siderophore synthesis and transport (Hopkinson and Morel, 2009). In cyanobacteria that are phylogenetically distant from "*Ca.* Synechococcus spongiarum," this iron uptake system is more common, e.g. in the freshwater cyanobacteria *Synechococcus* sp. strain JA23 and *Synechocystis* sp. strain PCC6803, whereas the phylogenetically closer, free-living *Synechococcus/Prochlorococcus* subclade lacks this siderophore transport ability (Hopkinson and Morel, 2009). COG1629, coding for a membrane iron receptor likely related to siderophores, is a common gene for all four "*Ca.* Synechococcus spongiarum" genomes. However, all components of an active ABC-type iron transport system related to siderophores was only found in SP3 and 15L, suggesting that SH4 and 142 either use a nonactive siderophore transport system or that their COG1629 senses a different type of available iron.

Eukaryotic-type domains have been shown to be common features of microbial sponge symbionts (Thomas *et al.*, 2010; Fan *et al.*, 2012). In "*Ca.* Synechococcus spongiarum," ankyrin domain proteins were a typical genomic signature, while other eukaryotic-type

domains (e.g. TPR and LRR) varied in number between the different clades. SP3 had comparably more proteins with TPR domains, whereas 142 had more proteins containing LRRs. A varying number of proteins containing LRR and TPR domains may be a type of host-specific fingerprint with a certain combination of proteins containing eukaryotic-type domains according to their host. Yet, further research is required to shed light on the role of these domains and also more genomes of sponge-associated bacteria derived from the same host sponge species need to be analyzed to support this hypothesis.

The presence of CRISPRs in cyanobacteria from the *Synechococcus/Prochlorococcus* subclade was surprising. The genome of 142 had two large CRISPR-Cas modules, the genomes of SH4 and 15L harbored dissociated CRISPR-associated proteins and CRISPR regions, SP3 only CRISPR-associated proteins. Also alternative antiviral defense mechanisms may be available to "*Ca.* Synechococcus spongiarum." For example, two unique endonucleases (COG2810 and COG3587) were found in SH4. Restriction-modification systems or genes preventing phage attachment to the cell surface could be alternative immune system features against bacteriophages (Stoddard *et al.*, 2007). The latter may be represented by the lack of a typical *Synechococcus* O antigen on the symbionts' LPS as discussed above. The great differences between "*Ca.* Synechococcus spongiarum" clades regarding antiviral defense mechanisms may be due to the different host sponge associations. A variety of parameters such as water pumping behavior of the host along with different levels of exposure of the symbionts to incoming water may influence their exposure to foreign free DNA and phages. The virus types that the symbionts are exposed to may also differ due to biogeographic location. It has been described, that 'old' CRISPR sequences are maintained against persistent or reemerging viruses (Weinberger *et al.*, 2012). Localized virus-host coevolution may thus explain the "*Ca.* Synechococcus spongiarum" intraspecies genomic divergence.

4.1.2.4 Conclusions

Despite nearly identical 16S rRNA gene sequences, the "*Ca.* Synechococcus spongiarum" group is characterized by a number of intraspecies genomic differences, such as different genome sizes, gene content, immune system mechanisms, methionine *de novo* synthesis patterns, and eukaryotic-type domain-containing proteins (LRR and TPR). Ankyrin repeats, on the other hand, seem to be a conserved feature that is common among different sponge microbial phyla in a variety of sponge host species and geographic locations. This suggests, that ankyrin domain proteins may be involved in sponge bacterial recognition as symbionts.

Enriched and depleted functions in the genomes of "*Ca.* Synechococcus spongiarum" in comparison to the phylogenetically closest free-living cyanobacterial relatives are summarized in Table 4-1. COGs assigned to class 'replication, recombination and repair' (L)

are represented at significantly higher proportion in the symbionts, which matches well with earlier findings from metagenomic studies and likely relates to horizontal gene transfer. COG class ‘signal transduction mechanisms’ (T) is represented in lower proportions than in the free-living relatives, which may reflect a more stable environment provided by the sponge host in comparison to the surrounding seawater. The type of the O antigen of the LPS in “*Ca.* Synechococcus spongiarum” will be affected by the lack of biosynthesis genes for dTDP-L-rhamnose, which possibly represents a novel mechanisms for host phagocytosis evasion and phage resistance in a niche characterized by possibly largely elevated phage pressure.

Table 4-1 Functions enriched and depleted in "*Ca.* Synechococcus spongiarum" compared to members of the closely related free-living marine *Synechococcus/Prochlorococcus* subclade.

Function	Context or interpretation (reference[s])
Enriched	
Recombination and repair	Insertion of mobile DNA into chromosomes (Thomas *et al.*, 2010; Fan *et al.*, 2012)
Transposable insertion elements	Horizontal gene transfer (Wu *et al.*, 2004)
Eukaryotic-type domains	Ankyrin repeat domains possibly obligatory feature of sponge symbionts (Nguyen *et al.*, 2014)
CRISPR-Cas systems	Selective pressure to acquire phage resistance (higher viral exposure) (Thomas *et al.*, 2010; Fan *et al.*, 2012; Cai *et al.*, 2013)
ABC-type iron transport system	Retained ancestral function (lost in free-living subclade) (Köster, 2001; Hopkinson and Morel, 2009)
Depleted	
Cell wall biogenesis	Symbiotic minimalism (Larsson *et al.*, 2011)
Signal transduction mechanism	Symbiotic minimalism (Tripp *et al.*, 2010; Kwan *et al.*, 2012)
Transcriptional regulation and (post)translational modification	Symbiotic minimalism (Nowack *et al.*, 2008)
ABC-type phosphate transport	Symbiotic minimalism (Nowack *et al.*, 2008)
Carbohydrate transport and metabolism and subunits of cytochrome *c*	Symbiotic minimalism (Nowack *et al.*, 2008)
Biosynthesis of LPS O antigen	Defense against phagocytosis by the sponge and anti-phage defense (Lerouge and Vanderleyden, 2001; Marston *et al.*, 2012)
Antioxidant enzymes	Reduced light radiation in sponge tissue (Latifi *et al.*, 2009)
Peptides of photosystem II and carotenoid biosynthesis	More stable light environment in the sponge tissue (Larsson *et al.*, 2011; Sveshnikov *et al.*, 2007; Meetam *et al.*, 1999)
Methionine salvage pathway	Methionine obtained from external sources (Albers, 2009)

4.2 PacBio-Illumina hybrid assembly pipeline development

The aim of this study was the development of an assembly pipeline that would be able to combine Illumina HiSeq short-reads and PacBio long-reads in a metagenomic assembly thereby improving the outcome over an Illumina-only assembly. The test dataset consisted of reads simulating the features of the real *A. aerophoba* metagenomic data to be assembled in the next step. As sponge microbiomes consist of unknown taxa that are expected to be rather different from their closest sequenced relatives on a genome basis, reference-independent *de novo* assembly and binning should be applied. My comparison revealed that a hybrid assembly of corrected PacBio reads and normalized Illumina reads created in SPAdes 3.5.0 with the only-assembler option enabled and with a *kmer* range of 33 to 127 was superior to all other tested assemblers and settings. Assembly statistics as well as bin quality was greatly improved by incorporating the PacBio long-reads.

At the beginning of the project, no publications were available on metagenomic hybrid assemblies of long-reads and short-reads. By the end of the project, a few approaches have been published. Beckmann and colleagues developed a tool for the detection of epigenetic motifs in bacterial genomes at low coverage and metagenomic settings (Beckmann *et al.*, 2014). As in my project, they simulated PacBio and Illumina read data from fully sequenced genomes to test their approach. Yet, for their aim, a metagenome of very low complexity consisting of only three bacterial genomes was sufficient. They used Meta-Velvet and Velvet for assembly (Namiki *et al.*, 2012; Zerbino and Birney, 2008). Details about which reads were assembled with which algorithm were not provided and the focus of the article clearly lies on the developed tool rather than the quality of the assembled metagenome(s) (Beckmann *et al.*, 2014).

Frank and colleagues compared a number of assembly approaches to each other: Illumina HiSeq only, PacBio circular consensus sequencing only, and a hybrid assembly using both read types (Frank *et al.*, 2016). They used different assembly strategies based on sample complexity and read type, and co-assembled only phylotypes-specific reads in a hybrid assembly, that were extracted by mapping after binning of the initial assemblies and focusing only on the two dominant phylotypes (Frank *et al.*, 2016).

Tsai and colleagues also aimed only for a dominant bacterium of the human skin microbiome and its bacteriophage in their long-read short-read hybrid approach (Tsai *et al.*, 2016). They compared a PacBio-only assembly, an Illumina-only assembly, and a hybrid assembly to each other. Similar to my approach, they used SPAdes-3.5.0 (Bankevich *et al.*, 2012) for a *de novo* hybrid assembly. Then, they focused on their target bacterium for further

mapping and re-assembly steps taking advantage of an available reference genome (Tsai *et al.*, 2016).

In summary, the three studies had approaches very different to each other and also to my un-targeted and reference-independent approach. Yet, all come to the same conclusion, that the integration of PacBio long-reads in a metagenomic assembly provided clear advantages for each respective project (Beckmann *et al.*, 2014; Frank *et al.*, 2016; Tsai *et al.*, 2016). As my project aimed for no specific member of the microbial community and tested the implications for un-targeted genome binning, it added an entirely novel approach to the recently emerging series of pipelines for the integration of PacBio long-reads into metagenomic projects. Also in this un-targeted approach, the addition of PacBio long-reads proved valuable for overall assembly statistics, bin reconstruction, and phylogenetic identification.

4.3 Metagenomic bins from the microbiome of *A. aerophoba* reveal unity in defense but metabolic specialization

4.3.1 Breaking new ground in assembly strategy and choice of references

Complementing six datasets of Illumina short-read data optimized for differential coverage binning with PacBio long-read data in a metagenomic assembly enabled the fully automated, un-targeted binning of 37 high-quality bacterial genomes from a highly diverse and complex sponge microbiome. The genomes derive from 13 bacterial phyla, two of which are candidate phyla, and represent the microbial community typically found to be abundant in *A. aerophoba* (Schmitt *et al.*, 2012a). The approach was validated by including two *A. aerophoba*-derived symbiont genomes in the analysis that were sequenced in previous studies by single-cell genomics (Poribacterium WGA3G) and 'mini-metagenomics' (cyanobacterium "*Ca.* Synechococcus spongiarum" 15L) after fluorescence-activated cell sorting (Kamke *et al.*, 2013; Burgsdorf *et al.*, 2015).

The choice of reference genomes differed from previous studies, taking advantage of the (now) known identity of the symbionts enabled by the binning approach. On the one hand, this decision was at the expense of comparability to previous studies which used seawater microbiomes as references. On the other hand, seawater metagenomes do not offer the microbial diversity that is abundant in the sponge microbiome and thus seemed no suitable comparison when using this binning approach. Yet, it has to be noted that the similarity of the results to previous studies, comparing sponge to seawater microbiomes, is remarkable in many aspects, e.g. TA and RM systems, and possible matrix utilization. Thus, the statistical signal proves to be a strong one only underlining that these frequently encountered features

discovered with very different approaches at different times and places truly play a biological role within the sponge microbiome.

4.3.2 Unity in defense

In a comparison of the 37 binned bacterial sponge symbionts and two previously published symbionts with closely related references from other environments, we revealed networks of COGs involved in a number of symbiont-enriched functions. RM as well as TA systems are significantly enriched in sponge symbionts. RM systems represent one major line of defense against incoming, foreign DNA, a feature frequently referred to as bacterial immunity (Vasu and Nagaraja, 2013). RM systems are also known to play a role in symbioses (Zheng et al. 2016) and have recently also been described in sponge symbionts (Gauthier *et al.*, 2016; Horn *et al.*, 2016; Tian *et al.*, 2016). Many of the COGs of network A in Figure 3-27 were previously described as sponge-enriched (Burgsdorf *et al.*, 2015; Fan *et al.*, 2012; Gao *et al.*, 2014b; Thomas *et al.*, 2010). This recurring finding of RM in symbionts of a variety of sponges from different geographic locations, and the abundance of RM in all 13 bacterial phyla in our dataset underscore the apparent significance for sponge symbioses. TA systems supposedly play a role in phage defense, stress response, and programmed cell death (Sberro *et al.*, 2013). The abundance and distribution of multiple RM and TA systems in the genomes of *A. aerophoba* symbionts suggests that defense against foreign DNA is an important feature of sponge symbionts confirming the previously stated concept of their convergent evolution (Fan *et al.*, 2012; Thomas *et al.*, 2010). Defense mechanisms such as RM and TA were previously found to be enriched in sponge symbionts (Fan *et al.*, 2012; Horn *et al.*, 2016) and are possibly a necessary countermeasure against the exposure to free DNA resulting from the sponge's extensive filtration and phagocytosis activity (Reiswig, 1974).

A second commonly sponge symbiont-enriched feature is the *hyuA-hyuB* gene pair (COG0145 and COG0146) that likely enables *H. pylori* to colonize its mouse host (Zhang *et al.*, 2009). This gene pair is significantly enriched in sponge symbionts in a variety of bacterial phyla suggesting a significance also for sponge host colonization. Furthermore, genes are symbiont-enriched, that are likely involved in the metabolization of components of the sponge extracellular mesohyl matrix, confirming a hypothesize previously published for the candidate phylum Poribacteria (Kamke *et al.*, 2013) and extending it to diverse members of the sponge microbiome.

CRISPR-Cas systems as well as eukaryotic-like protein domains have both been hypothesized to play crucial roles for sponge symbionts in previous studies (Thomas *et al.*, 2010; Fan *et al.*, 2012; Liu *et al.*, 2012; Barrangou *et al.*, 2007; Burgsdorf *et al.*, 2015; Horn *et al.*, 2016). Both features were more common and also more abundant in the sponge symbiont

group, although not at statistically significant levels. This is likely due to the approach of choosing reference genomes primarily by phylogenetic relatedness. Thus, the references derive from a multitude of environments including other 'dense' bacterial communities where those defense mechanisms may be comparably important. Additionally, CRISPR-Cas systems and eukaryotic-like proteins in general do not show as high gene counts as other symbiont-enriched features. Therefore, a statistically significant effect is less likely to be reached.

4.3.3 Metabolic specialization

While defense mechanisms emerged as the main topic when comparing sponge symbionts to references, metabolic specialization was the main driver for a grouping within the symbionts. Three symbiont guilds were observed, one specialized on carnitine metabolism, one on the catabolism of sulfated polysaccharides, and one group of generalists. Carnitine is produced by most eukaryotes including sponges (Fraenkel, 1954) and we posit that it may be taken up by symbiotic bacteria from the readily available sponge-derived detritus consisting largely of shed sponge cells (Alexander *et al.*, 2014; de Goeij *et al.*, 2009). Uptake of carnitine by bacteria can also serve as protection against environmental stressors, such as variation in water content, salinity, or temperature (Meadows and Wargo, 2015). Sulfated polysaccharides are likely metabolized utilizing arylsulfatase A. While this enzyme was enriched in the symbionts over the references in general and is distributed across a variety of symbiont phyla, it is largely enriched in symbiont group II together with a number of ABC transporters. Both carnitine and sulfated polysaccharides are possibly components of the extracellular matrix of the sponge and/or components of cells shed by the sponge as a consequence of cell renewal (Alexander *et al.*, 2014; de Goeij *et al.*, 2009; Fraenkel, 1954; Vilanova *et al.*, 2009; Zierer and Mourão, 2000). The members of symbiont group III also possessed many of the COGs that are enriched in groups I and II, but in far lower numbers. We thus posit, that same metabolic pathways are utilized but less extensively, and that group III represents a group of metabolic generalists.

4.3.4 Conclusions

The complementation of Illumina short-read with PacBio long-read sequencing for metagenomic binning of highly complex environmental samples greatly improves the overall assembly statistics. It also improves the quality of binned genomes and eases, and often newly enables phylogenetic classification of the binned genomes. The statistical comparison revealed an enrichment of genes related to RM and TA systems in most symbiont genomes over the reference genomes. This implies that the defense against incoming foreign DNA is of high importance for a symbiotic existence within the sponge mesohyl. This finding is particularly

relevant in the context of the extensive animal's filtration and phagocytosis activities, with the resultant ample exposure of the symbionts to free DNA. Secondly, host colonization and host matrix utilization were identified as significantly enriched features in sponge symbionts. The within-symbiont genome comparison revealed a nutritional specialization, where one guild of symbionts appears to metabolize carnitine, while the other appears to metabolize sulfated polysaccharides, both of which are abundant molecules of the sponge extracellular matrix. We hypothesize that the sponge symbionts feed on the sponge cells that are shed as part of the cell turnover, and on components of the sponge extracellular matrix. A third guild of symbionts may be viewed as nutritional generalists, whose precise function within this consortium remains to be identified. The unprecedented resolution of the genomic repertoire was enabled by binning of a metagenomic hybrid assembly of hitherto unprecedented depth for sponge symbioses.

4.4 General conclusions and future directions

In recent years, methodology has developed significantly in the field of microbiology, both in the areas of technology as well as bioinformatics. With increasing interest in and acknowledgement of the role of the microbiome of the Earth and its inhabitants (http://www.earthmicrobiome.org; The MetaSUB International Consortium 2016), major efforts have begun to shed light on the 'microbial dark matter,' the yet uncultivable, but major fraction of most microbial communities (Rinke *et al.*, 2013; Marcy *et al.*, 2007). Metagenomic approaches have developed from targeting just one gene, first by PCR and clone libraries (Ahlgren and Rocap, 2006; Rotthauwe *et al.*, 1997; Webster *et al.*, 2008a), and later by amplicon sequencing (Bourne *et al.*, 2013; Schmitt *et al.*, 2012b; Vik *et al.*, 2013; Thomas *et al.*, 2016). Next in line were methodologies for the interpretation and comparison of the genomic content of whole microbial communities (Pimentel-Elardo *et al.*, 2012; Li *et al.*, 2015; Martín-Cuadrado *et al.*, 2007; Thomas *et al.*, 2010). Now, the field has arrived at the point where binning genomes of single community members from the metagenomic data is possible (Albertsen *et al.*, 2013; Gauthier *et al.*, 2016; Gao *et al.*, 2014b).

Sequencing technologies themselves have also improved and greatly reduced in cost, thereby allowing them to aid in presenting genomics and metagenomics as widely available standard methodologies (Koren *et al.*, 2013). The most recent developments in the realm of sequencing have aimed at improving assembly contiguity by increasing read length (Koren *et al.*, 2013; Koren and Phillippy, 2015). While hybrid assemblies of short-reads and long-reads have become a standard procedure in genomics in the last years (Bashir *et al.*, 2012; Madoui *et al.*, 2015; Liao *et al.*, 2015), the implementation of long-reads in metagenomics is still in its infancy. A number of metagenomic studies have applied long-reads for targeted approaches,

e.g. binning a specific dominant taxon (Frank *et al.*, 2016; Tsai *et al.*, 2016), but so far, no studies on un-targeted hybrid assembly and binning approaches have been conducted. In contrast to this upcoming technology, single-cell genomics has become a standard procedure for genomic studies of uncultivable bacteria (Woyke *et al.*, 2009; Kamke *et al.*, 2014; Yoon *et al.*, 2011). Automation of laboratory procedures and even assembly and decontamination have enabled high-throughput single-cell sequencing (Rinke *et al.*, 2013; Tennessen *et al.*, 2015; Swan *et al.*, 2013; Lasken and McLean, 2014). As of today (February 10, 2017), as many as 1,267 contamination-screened single cell genome analysis projects are listed in the Genomes OnLine Database (GOLD, https://gold.jgi.doe.gov).

This thesis aimed not only at answering specific biological questions, but also on methodological development. A multitude of approaches were applied to eventually obtain genomes from different members of the uncultivable microbial community of the marine HMA sponge *A. aerophoba* (Figure 4-1). One of the target symbionts was the cyanobacterium "*Ca.* Synechococcus spongiarum," which was – due to its autofluorescence – perfectly suited for FACS sorting followed by single-cell genomics. The amplification reactions did not contain real single-cells as an insert but aliquots of FACS sorted cell enrichments, termed 'mini-metagenomes.' Here, contrasting the original definition of 'mini-metagenomes' describing pools of randomly sorted single cells (McLean *et al.*, 2013), the aim was a pure cell enrichment of the target cells. With this approach I aimed to increase the overall yield by balancing low-coverage amplification regions from one cell by the amplification products from another cell. Additionally, I expected lower fractions of contaminants than in 'true' single-cells. The JGI single-cell pipeline has incorporated an automated decontamination step (Tennessen *et al.*, 2015) that is rather strictly filtering out any possible contamination, thereby also elimination genes acquired by horizontal gene transfer, phages, and other potentially interesting genomic features leaving only a core likely free of any contaminants (Dr. Tanja Woyke, DOE JGI, personal communication). As this decontamination method would likely also exclude features previously hypothesized as sponge symbiont-specific or –enriched, such as horizontal gene transfer features or eukaryotic-like protein domains (Thomas *et al.*, 2010), I discarded this step and replaced it with the above-described less radical binning approach. The most complete genome from this single-cell amplification approach was comparable in quality to genomes binned from the metagenomes of other sponges, also varying in methodology (Burgsdorf *et al.*, 2015; Gao *et al.*, 2014b). The similarity of the results from fundamentally different approaches leads to the conclusion that methodology did not essentially influence the outcome in this study.

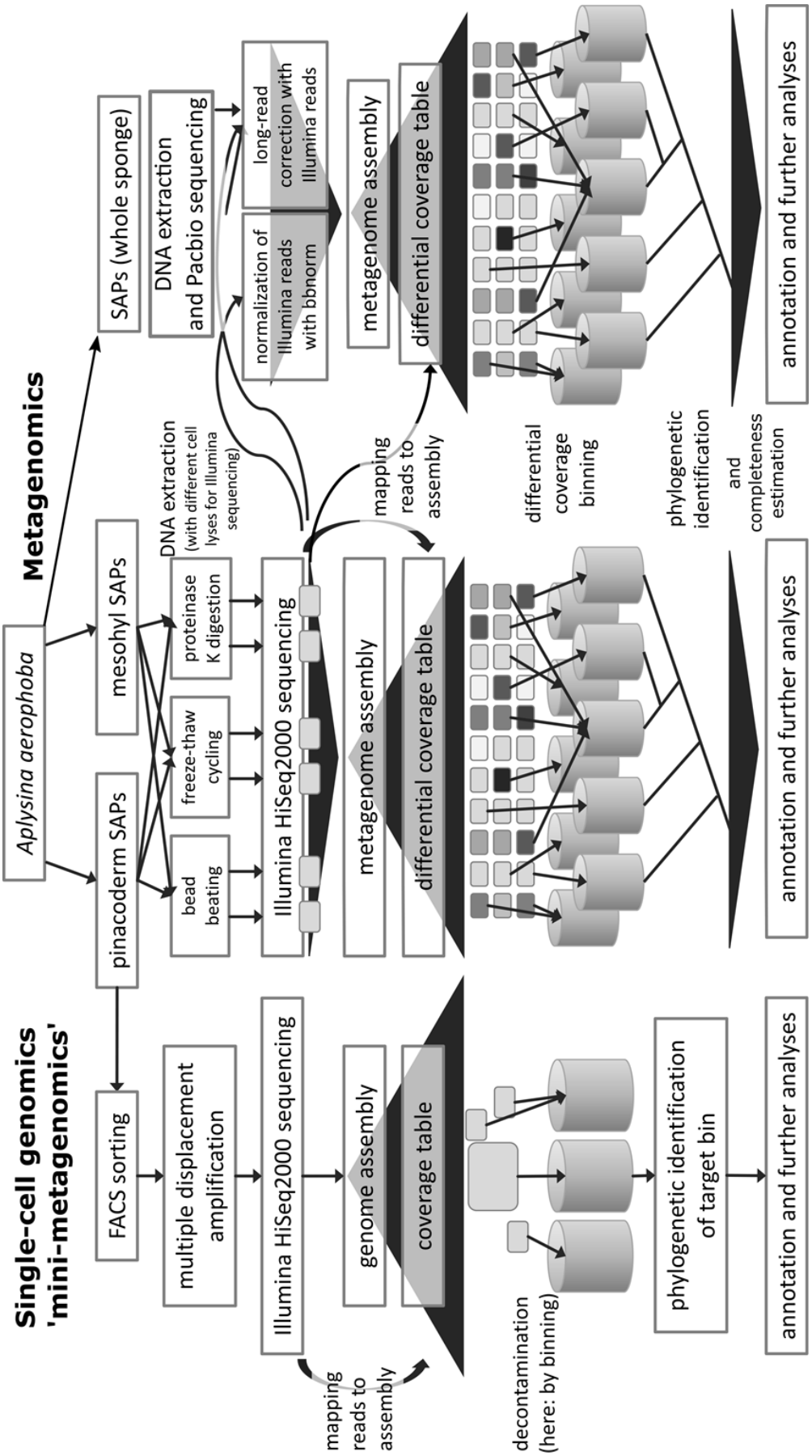


Figure 4-1 Overview of the applied sequencing and bioinformatics strategies to obtain genomes of sponge symbionts.

My second major project aimed at binning as many sponge-symbionts as possible from the metagenome of *A. aerophoba*, utilizing PacBio long-read data complementing Illumina short-read data optimized for differential coverage binning. As there was no assembly pipeline published to achieve this, I developed an assembly pipeline combining simulated PacBio long-reads and Illumina short-reads of a test dataset reflecting the most prominent properties of the real metagenomic data such as uneven coverage and sequencing errors according to the respective sequencing machine. I then applied the developed assembly pipeline to the real *A. aerophoba* dataset and compared the results to an assembly consisting of only the Illumina data to evaluate the improvement by addition of PacBio long-reads.

One issue working with uncultivated bacteria is that in many cases only the 16S rRNA gene or no information at all is available beforehand (Lasken and McLean, 2014). A common issue in single-cell sequencing projects is the lack of this gene due to amplification bias or primer mismatches (Clingenpeel *et al.*, 2015). With "*Ca.* Synechococcus spongiarum" as the only cyanobacterial symbiont of *A. aerophoba* a later addition of this gene to the genome was possible. In contrast, for an unknown bacterium isolated with less restrictive sorting approaches, e.g. FACS sorting by cell size, or a symbiont binned from a metagenome, a later addition of the right 16S rRNA gene to the genome would not be possible. Without the 16S rRNA gene information and in most cases no sequenced reference at hand, a taxonomic classification of the genome would not be possible. Considering that the first genomes for whole bacterial candidate phyla were often sequenced from yet uncultivable representatives (Lasken and McLean, 2014), such discoveries may be missed entirely, solely due to the lack of information to identify them as such. This highlights the improvement of the metagenomic assembly and subsequent binning by the addition of PacBio long-reads. Although, in the presented study, many genomes were binned at high completeness from the Illumina data only, just 38% of them contained a 16S rRNA gene. In the hybrid assembly, in contrast, 86% of the >70%-completeness bins contained this gene (Table 3-18). Assuming that 16S rRNA gene phylogeny was the only means to phylogenetically place a certain bin, the implementation of long-reads greatly improved the portion of 'usable' bins.

A great volume of knowledge has been collected on sponge-microbe symbioses since the 1970s/80s (Reiswig, 1974; Wilkinson, 1978a, 1980; Wilkinson *et al.*, 1984). Symbiont-enriched features have been discovered, such as horizontal gene transfer, restriction modification systems, CRISPR-Cas systems, ammonium assimilation, eukaryotic-like domains, and metabolic adaptations possibly enabling the symbionts to metabolize parts of the sponges' extracellular mesohyl matrix (Thomas *et al.*, 2010; Fan *et al.*, 2012; Kamke *et al.*, 2014; Bayer *et al.*, 2008a; Kamke *et al.*, 2013).

One focus of this thesis was the cyanobacterial symbiont "*Ca.* Synechococcus spongiarum." Although the genomic content of this symbiont species differed largely between

clades – despite nearly 99% 16S rRNA gene sequence identity, a number of the previously hypothesized sponge symbiont-specific features were confirmed in this species in comparison to free-living relatives. Horizontal gene transfer features, such as transposable insertion elements and COGs involved in recombination and repair, were enriched, as well as eukaryotic-type ankyrin repeat domains that may be obligatory in sponge symbionts to evade host phagocytosis. Also CRISPR-Cas systems were enriched likely due to phage pressure caused by the sponges' pumping activity. ABC-type iron transport system features may represent an ancestral function retained by the symbionts, while it was lost in the free-living relatives. On the other hand, a number of features were depleted in the symbionts which can be interpreted as symbiotic minimalism. Those features were cell wall biogenesis, signal transduction mechanisms, transcriptional regulation and (post)translational modification genes, ABC-type phosphate transport, and carbohydrate transport and metabolism. Another possible means of defense against phagocytosis and phages was an altered O antigen of the LPS. A reduction in antioxidant enzymes and in peptides of photosystem II and carotenoid biosynthesis was due to the reduced, and more stable light radiation within the sponge tissue. As also genes involved in methionine salvage were depleted in the symbionts, they may obtain methionine from external sources.

In the large-scale, un-targeted binning approach, the list of *A. aerophoba*-associated symbiont genomes was expanded by 37 genomes from the 13 bacterial phyla and candidate phyla *Proteobacteria* (Alpha, Delta, and Gamma), *Nitrospinae*, *Nitrospirae*, candidate phylum SBR1093, *Acidobacteria*, candidate phylum Poribacteria, *Bacteroidetes*, *Gemmatimonadetes*, *Spirochaetae*, *Actinobacteria*, *Deinococcus-Thermus*, *Cyanobacteria*, and *Chloroflexi*. This dataset thereby provides genomes of nearly all main symbiont phyla known for *A. aerophoba* (Schmitt *et al.*, 2012a). This enabled an analysis at unprecedented resolution, comparing these not only to the bulk reference community of e.g. seawater, as in previous studies (Thomas *et al.*, 2010), but to selected references of the regarding phyla. As the statistical comparison confirmed an enrichment of defense features in the symbionts, this only underlines the importance of this feature in the sponge host environment, that has now been identified via a multitude of very different approaches (Thomas *et al.*, 2010; Gao *et al.*, 2014a; Burgsdorf *et al.*, 2015; Horn *et al.*, 2016). Also host colonization and matrix utilization features were symbiont-enriched (Figure 3-27).

Due to the diversity of the binned symbiont genomes, for the first time, also a within-symbiont genome comparison was possible, which revealed three guilds of symbionts that did not necessarily coincide with phylogeny (Figure 3-29). These three guilds were characterized by enrichments – or lack of enrichment – of certain genes involved in nutrition. While one group seems to be specialized on the metabolization of carnitine, the second group apparently

specializes on sulfated polysaccharides, and the third on no particular metabolism (Figure 3-33). The members of this third group thus seem to be nutritional generalists.

The microbiome of *A. aerophoba* has been the focus of this thesis. A great number of bacterial genomes have been added to the pool of available sponge symbiont genomes and methods have been developed to improve the yield in binning approaches. Future studies should consider the findings of this thesis and test them in experimental approaches.

Perspectives are:

- Further analyses of the sequenced symbiont genomes focusing on specific taxa within the microbial community to discover taxon-specific adaptations like nutritional mode, defense strategies, dependencies on microbial partners and the host sponge
- Application of imaging techniques to localize and quantify symbionts in the host sponge, construct networks and reveal interactions between symbionts and with the host
- Targeted cultivation efforts implementing information deducted from the genomic information and interaction networks
- Application of the developed sequencing and binning strategy to other sponge species to obtain more symbiont genomes and test if the features discovered in *A. aerophoba* reflect a general pattern in sponge microbiomes

5 Bibliography

Abdelmohsen UR, Pimentel-Elardo SM, Hanora A, Radwan M, Abou-El-Ela SH, Ahmed S, *et al.* (2010). Isolation, phylogenetic analysis and anti-infective activity screening of marine sponge-associated actinomycetes. *Mar Drugs* **8**: 399–412.

Aerts LAM. (2000). Dynamics behind standoff interactions in three reef sponge species and the coral *Montastraea cavernosa*. *Mar Ecol* **21**: 191–204.

Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, *et al.* (2015). Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst* **1**: 72–87.

Ahlgren NA, Rocap G. (2006). Culture isolation and culture-independent clone libraries reveal new marine *Synechococcus* ecotypes with distinctive light and N physiologies. *Appl Environ Microbiol* **72**: 7193–7204.

Albers E. (2009). Metabolic characteristics and importance of the universal methionine salvage pathway recycling methionine from 5'-methylthioadenosine. *IUBMB Life* **61**: 1132–1142.

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533–538.

Alexander BE, Liebrand K, Osinga R, Van Der Geest HG, Admiraal W, Cleutjens JPM, *et al.* (2014). Cell turnover and detritus production in marine sponges from tropical and temperate benthic ecosystems. *PLoS One* **9**: e109486.

Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**: 402.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, *et al.* (2014). Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144–1150.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–10.

Ankenbrand MJ, Hohlfeld S, Hackl T, Förster F. (2016). AliTV – interactive visualization of whole genome comparisons. *PeerJ Prepr*. e-pub ahead of print, doi: 10.7287/peerj.preprints.2348v1.

Antcliffe JB, Callow RHT, Brasier MD. (2014). Giving the early fossil record of sponges a squeeze. *Biol Rev* **89**: 972–1004.

Azam F, Fenchel T, Field JG, Gray JS, Meyer-Reil LA, Thingstad F. (1983). The ecological role of water-column microbes in the sea. *Mar Ecol Prog Ser* **10**: 257–263.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, *et al.* (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455–477.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, *et al.* (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709–1712.

Bashir A, Klammer AA, Robins WP, Chin C-S, Webster D, Paxinos E, *et al.* (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol* **30**: 701–707.

Bayer K, Kamke J, Hentschel U. (2014a). Quantification of bacterial and archaeal symbionts in high and low microbial abundance sponges using real-time PCR. *FEMS Microbiol Ecol* **89**: 679–690.

Bayer K, Moitinho-Silva L, Brümmer F, Cannistraci C V., Ravasi T, Hentschel U. (2014b). GeoChip-based insights into the microbial functional gene repertoire of marine sponges (high microbial abundance, low microbial abundance) and seawater. *FEMS Microbiol Ecol* **90**: 832–843.

Bayer K, Scheuermayer M, Fieseler L, Hentschel U. (2013). Genomic mining for novel $FADH_2$ -dependent halogenases in marine sponge-associated microbial consortia. *Mar Biotechnol* **15**: 63–72.

Bayer K, Schmitt S, Hentschel U. (2007). Microbial nitrification in Mediterranean sponges: possible involvement of ammonia-oxidizing Betaproteobacteria. In: *Porifera Research: Biodiversity, Innovation and Sustainability*. pp 165–171.

Bayer K, Schmitt S, Hentschel U. (2008a). Physiology, phylogeny and *in situ* evidence for bacterial and archaeal nitrifiers in the marine sponge *Aplysina aerophoba*. *Environ Microbiol* **10**: 2942–2955.

Bayer K, Siegl A, Schmitt S, Hoffmann F, Hentschel U. (2008b). Unravelling microbial diversity and metabolism in marine sponges. *Nova Acta Leopoldina* **96**: 71–78.

Beckmann ND, Karri S, Fang G, Bashir A. (2014). Detecting epigenetic motifs in low coverage and metagenomics settings. *BMC Bioinformatics* **15**: S16.

Beims H, Wittmann J, Bunk B, Spröer C, Rohde C, Günther G, *et al.* (2015). *Paenibacillus*

larvae-directed bacteriophage HB10c2 and its application in American foulbrood-affected honey bee larvae Goodrich-Blair H (ed). *Appl Environ Microbiol* **81**: 5411–5419.

Belarbi E. (2003). Producing drugs from marine sponges. *Biotechnol Adv* **21**: 585–598.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. (2007). GenBank. *Nucleic Acids Res* **35**: D21–D25.

Bergquist PR. (1998). Porifera. In: Anderson DT (ed). *Invertebrate Zoology*. Oxford University Press, pp 10–27.

Blainey PC. (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* **37**: 1–29.

Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter F-J, Zakrzewski M, *et al.* (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* **10**: 154.

Bombar D, Heller P, Sanchez-Baracaldo P, Carter BJ, Zehr JP. (2014). Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *ISME J* **8**: 2530–2542.

Bordenstein SR, Theis KR. (2015). Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. *PLOS Biol* **13**: e1002226.

Bourne DG, Dennis PG, Uthicke S, Soo RM, Tyson GW, Webster N. (2013). Coral reef invertebrate microbiomes correlate with the presence of photosymbionts. *ISME J* **7**: 1452–1458.

Brain CK 'Bob', Prave AR, Hoffmann KH, Fallick AE, Botha A, Herd DA, *et al.* (2012). The first animals: Ca. 760-million-year-old sponge-like fossils from Namibia. *S Afr J Sci* **108**: 1–8.

Burgsdorf I, Erwin PM, López-Legentil S, Cerrano C, Haber M, Frenk S, *et al.* (2014). Biogeography rather than association with cyanobacteria structures symbiotic microbial communities in the marine sponge Petrosia ficiformis. *Front Microbiol* **5**: 529.

Burgsdorf I, Slaby BM, Handley KM, Haber M, Blom J, Marshall CW, *et al.* (2015). Lifestyle evolution in cyanobacterial symbionts of sponges. *MBio* **6**: e00391-15.

Bushnell B. (2015). BBMap. https://sourceforge.net/projects/bbmap/.

Cai F, Axen SD, Kerfeld CA. (2013). Evidence for the widespread distribution of CRISPR-Cas system in the phylum Cyanobacteria. *RNA Biol* **10**: 687–693.

Cardoso JFMF, van Bleijswijk JDL, Witte H, van Duyl FC. (2013). Diversity and abundance of ammonia-oxidizing Archaea and Bacteria in tropical and cold-water coral reef sponges. *Aquat Microb Ecol* **68**: 215–230.

Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J. (2005). ACT: the Artemis comparison tool. *Bioinformatics* **21**: 3422–3423.

Chen Y. (2007). Functional genomics of the unicellular cyanobacterium *Synechococcus elongatus* PCC 7942. http://www.springerlink.com/index/G2P966R842P4H317.pdf.

Clingenpeel S, Clum A, Schwientek P, Rinke C, Woyke T. (2015). Reconstructing each cell's genome within complex microbial communities - dream or reality? *Front Microbiol* **5**: 1–6.

Clingenpeel S, Schwientek P, Hugenholtz P, Woyke T. (2014). Effects of sample treatments on genome recovery via single-cell genomics. *ISME J* **8**: 2546–2549.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, *et al.* (2014). Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**: D633–D642.

Colman AS. (2015). Sponge symbionts and the marine P cycle. *Proc Natl Acad Sci* **112**: 4191–4192.

Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243.

Dean FB, Nelson JR, Giesler TL, Lasken RS. (2001). Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**: 1095–1099.

Deines P, Bosch TCG. (2016). Transitioning from microbiome composition to microbial community interactions: The potential of the metaorganism *Hydra* as an experimental model. *Front Microbiol* **7**: 1610.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Diaz MC, Rützler K. (2001). Sponges: An essential component of Caribbean coral reefs. *Bull Mar Sci* **69**: 535–546.

Dressler-Allame M, Göcke C, Kersken D, Plotkin A, Janussen D. (2016). Carnivorous sponges (Cladorhizidae) of the deep Weddell Sea, with descriptions of two new species. *Deep Sea Res Part II Top Stud Oceanogr*. e-pub ahead of print, doi: 10.1016/j.dsr2.2016.08.006.

Du W, Wang XL, Komiya T. (2015). Potential Ediacaran sponge gemmules from the Yangtze Gorges area in South China. *Gondwana Res* **28**: 1246–1254.

Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik BP, *et al.* (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**: R90.

Easson CG, Thacker RW. (2014). Phylogenetic signal in the community structure of host-specific microbiomes of tropical marine sponges. *Front Microbiol* **5**: 1–11.

Eddy SR. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**: 205–211.

Edgar RC. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Ereskovsky A V. (2010). The comparative embryology of sponges. Springer Netherlands: Dordrecht.

Erwin PM, López-Legentil S, Turon X. (2012a). Ultrastructure, molecular phylogenetics, and chlorophyll *a* content of novel cyanobacterial symbionts in temperate sponges. *Microb Ecol* **64**: 771–783.

Erwin PM, Pita L, López-Legentil S, Turon X. (2012b). Stability of sponge-associated bacteria over large seasonal shifts in temperature and irradiance. *Appl Environ Microbiol* **78**: 7358–7368.

Erwin PM, Thacker RW. (2008a). Cryptic diversity of the symbiotic cyanobacterium *Synechococcus spongiarum* among sponge hosts. *Mol Ecol* **17**: 2937–2947.

Erwin PM, Thacker RW. (2008b). Phototrophic nutrition and symbiont diversity of two Caribbean sponge–cyanobacteria symbioses. *Mar Ecol Prog Ser* **362**: 139–147.

Esteves AI, Amer N, Nguyen M, Thomas T. (2016). Sample processing impacts the viability and cultivability of the sponge microbiome. *Front Microbiol*. e-pub ahead of print, doi: 10.3389/fmicb.2016.00499.

Evdokimov AG, Anderson DE, Routzahn KM, Waugh DS. (2001). Unusual molecular architecture of the *Yersinia pestis* cytotoxin YopM: A leucine-rich repeat protein with the shortest repeating unit. *J Mol Biol* **312**: 807–821.

Faist H, Keller A, Hentschel U, Deeken R. (2016). Grapevine (*Vitis vinifera*) crown galls host distinct microbiota. *Appl Environ Microbiol* **82**: 5542–5552.

Fan L, Reynolds D, Liu M, Stark M, Kjelleberg S, Webster NS, *et al.* (2012). Functional equivalence and evolutionary convergence in complex communities of microbial

sponge symbionts. *Proc Natl Acad Sci U S A* **109**: E1878–E1887.

Felsenstein J. (1995). PHYLIP: phylogeny inference package, version 3.57c. *Seattle Dep Genet Univ Washingt*.

Fieseler L, Horn M, Wagner M, Hentschel U. (2004). Discovery of the novel candidate phylum Poribacteria'' in marine sponges. *Appl Environ Microbiol* **70**: 3724–3732.

Fieseler L, Quaiser A, Schleper C, Hentschel U. (2006). Analysis of the first genome fragment from the marine sponge-associated, novel candidate phylum Poribacteria by environmental genomics. *Environ Microbiol* **8**: 612–624.

Finn RD, Clements J, Eddy SR. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res* **39**: W29-37.

Flemer B, Kennedy J, Margassery LM, Morrissey JP, O'Gara F, Dobson ADW. (2012). Diversity and antimicrobial activities of microbes from two Irish marine sponges, *Suberites carnosus* and *Leucosolenia* sp. *J Appl Microbiol* **112**: 289–301.

Fraenkel G. (1954). The distribution of vitamin BT (carnitine) throughout the animal kingdom. *Arch Biochem Biophys* **50**: 486–495.

Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, *et al.* (2016). Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci Rep* **6**: 25373.

Freeman CJ, Thacker RW. (2011). Complex interactions between marine sponges and their symbiotic microbial communities. *Limnol Oceanogr* **56**: 1577–1586.

Freeman CJ, Thacker RW, Baker DM, Fogel ML. (2013). Quality or quantity: Is nutrient transfer driven more by symbiont identity and productivity than by symbiont abundance? *ISME J* **7**: 1116–1125.

Friedrich AB, Fischer I, Proksch P, Hacker J, Hentschel U. (2001). Temporal variation of the microbial community associated with the mediterranean sponge *Aplysina aerophoba*. *FEMS Microbiol Ecol* **38**: 105–115.

Gao Z-M, Wang Y, Lee OO, Tian R-M, Wong YH, Bougouffa S, *et al.* (2014a). Pyrosequencing reveals the microbial communities in the Red Sea sponge *Carteriospongia foliascens* and their impressive shifts in abnormal tissues. *Microb Ecol* **68**: 621–632.

Gao Z-M, Wang Y, Tian R-M, Wong YH, Batang ZB, Al-Suwailem AM, *et al.* (2014b). Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont '*Candidatus* Synechococcus spongiarum'. *MBio* **5**: e00079-14.

Gauthier M-EA, Watson JR, Degnan SM. (2016). Draft genomes shed light on the dual bacterial symbiosis that dominates the microbiome of the coral reef sponge *Amphimedon queenslandica*. *Front Mar Sci* **3**: 196.

Ghiold J, Rountree GA, Smith SH. (1994). Common sponges of the Cayman Islands. In: Brunt MA, Davies JE (eds). *The Cayman Islands: Natural History and Biogeography*. Kluwer Academic Publishers, pp 131–138.

Gilbert JA, Dupont CL. (2011). Microbial metagenomics: Beyond the genome. *Ann Rev Mar Sci* **3**: 347–371.

Gili JM, Coma R. (1998). Benthic suspension feeders: Their paramount role in littoral marine food webs. *Trends Ecol Evol* **13**: 316–321.

de Goeij JM, de Kluijver A, van Duyl FC, Vacelet J, Wijffels RH, de Goeij AFPM, *et al.* (2009). Cell kinetics of the marine sponge *Halisarca caerulea* reveal rapid cell turnover and shedding. *J Exp Biol* **212**: 3892–3900.

de Goeij JM, van Oevelen D, Vermeij MJA, Osinga R, Middelburg JJ, de Goeij AFPM, *et al.* (2013). Surviving in a marine desert: The sponge loop retains resources within coral reefs. *Science* **342**: 108–110.

Gold DA, Grabenstatter J, de Mendoza A, Riesgo A, Ruiz-trillo I, Summons RE. (2016). Sterol and genomic analyses validate the sponge biomarker hypothesis. *Proc Natl Acad Sci* **113**: 2684–2689.

Gordon J, Knowlton N, Relman DA, Rohwer F, Youle M. (2013). Superorganisms and holobionts. *Microbe* **8**: 152–153.

Gouy M, Guindon S, Gascuel O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**: 221–224.

Green D, Howard D, Yang X, Kelly M, Oreffo ROC. (2003). Natural marine sponge fiber skeleton: A biomimetic scaffold for human osteoprogenitor cell attachment, growth, and differentiation. *Tissue Eng* **9**: 1159–1166.

Grissa I, Vergnaud G, Pourcel C. (2007). CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52–W57.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Hackl T. (2016). A draft genome for the Venus flytrap, *Dionaea muscipula*.

Hackl T, Hedrich R, Schultz J, Förster F. (2014). proovread: Large-scale high-accuracy

PacBio correction through iterative short read consensus. *Bioinformatics* **30**: 1–8.

Haider B, Ahn T-H, Bushnell B, Chai J, Copeland A, Pan C. (2014). Omega: An overlap-graph *de novo* assembler for metagenomics. *Bioinformatics* **30**: 2717–2722.

Haroon MF, Skennerton CT, Steen JA, Lachner N, Hugenholtz P, Tyson GW. (2013). In-solution fluorescence in situ hybridization and fluorescence-activated cell sorting for single cell and population genome recovery. In: DeLong EF (ed) Vol. 531. *Methods in Enzymology*. Elsevier Inc., pp 3–19.

Heckel PH. (1974). Carbonate buildups in the geologic record: A review. *Soc Econ Paleontol Mineral*.

Heinsen F-A, Fangmann D, Müller N, Schulte DM, Rühlemann MC, Türk K, *et al.* (2016). Beneficial effects of a dietary weight loss intervention on human gut microbiome diversity and metabolism are not sustained during weight maintenance. *Obes Facts* **9**: 379–391.

Hentschel U, Fieseler L, Wehrl M, Gernert C, Steinert M, Hacker J, *et al.* (2003). Microbial diversity of marine sponges. In: Müller W (ed). *Sponges (Porifera)*. Springer Verlag: Berlin/Heidelberg/New York, pp 59–88.

Hentschel U, Piel J, Degnan SM, Taylor MW. (2012). Genomic insights into the marine sponge microbiome. *Nat Rev Microbiol* **10**: 641–654.

Hentschel U, Schmid M, Wagner M, Fieseler L, Gernert C, Hacker J. (2001). Isolation and phylogenetic analysis of bacteria with antimicrobial activities from the Mediterranean sponges *Aplysina aerophoba* and *Aplysina cavernicola*. *FEMS Microbiol Ecol* **35**: 305–312.

Hestetun JT, Vacelet J, Boury-Esnault N, Borchiellini C, Kelly M, Ríos P, *et al.* (2016). The systematics of carnivorous sponges. *Mol Phylogenet Evol* **94**: 327–345.

Hoffmann F, Radax R, Woebken D, Holtappels M, Lavik G, Rapp HT, *et al.* (2009). Complex nitrogen cycling in the sponge *Geodia barretti*. *Environ Microbiol* **11**: 2228–2243.

Honda D, Yokota A, Sugiyama J. (1999). Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains. *J Mol Evol* **48**: 723–739.

Hooper, John NA, Van Soest RWM. (2002). Systema Porifera: A guide to the classification of sponges. Kluwer Academic/Plenum Publishers: New York.

Hopkinson BM, Morel FMM. (2009). The role of siderophores in iron acquisition by photosynthetic marine microorganisms. *Biometals* **22**: 659–669.

Horn H, Hentschel U, Abdelmohsen UR. (2015). Mining genomes of three marine sponge-associated actinobacterial isolates for secondary metabolism. *Genome Announc* **3**: e01106-15.

Horn H, Slaby BM, Jahn MT, Bayer K, Moitinho-Silva L, Förster F, *et al.* (2016). An enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes. *Front Microbiol* **7**: 1751.

Huang W, Li L, Myers JR, Marth GT. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics* **28**: 593–594.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, *et al.* (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24**: 688–696.

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Johnson RM, Ramond J-B, Gunnigle E, Seely M, Cowan DA. (2017). Namib Desert edaphic bacterial, fungal and archaeal communities assemble through deterministic processes but are influenced by different abiotic parameters. *Extremophiles*. e-pub ahead of print, doi: 10.1007/s00792-016-0911-1.

Kamke J, Rinke C, Schwientek P, Mavromatis K, Ivanova N, Sczyrba A, *et al.* (2014). The candidate phylum Poribacteria by single-cell genomics: New insights into phylogeny, cell-compartmentation, eukaryote-like repeat proteins, and other genomic features. *PLoS One* **9**: e87353.

Kamke J, Sczyrba A, Ivanova N, Schwientek P, Rinke C, Mavromatis K, *et al.* (2013). Single-cell genomics reveals complex carbohydrate degradation patterns in poribacterial symbionts of marine sponges. *ISME J* **7**: 2287–2300.

Kamke J, Taylor MW, Schmitt S. (2010). Activity profiles for marine sponge-associated bacteria obtained by 16S rRNA vs 16S rRNA gene comparisons. *ISME J* **4**: 498–508.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**: D277–D280.

Kang DD, Froula J, Egan R, Wang Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**: e1165.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, *et al.* (2014).

Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**: 416–420.

Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, *et al.* (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* **14**: R101.

Koren S, Phillippy AM. (2015). One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol* **23**: 110–120.

Köster W. (2001). ABC transporter-mediated uptake of iron, siderophores, heme and vitamin B 12. *Res Microbiol* **152**: 291–301.

Kumar S, Stecher G, Tamura K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**: 1870–1874.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, *et al.* (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Kwan JC, Donia MS, Han AW, Hirose E, Haygood MG, Schmidt EW. (2012). Genome streamlining and chemical defense in a coral reef symbiosis. *Proc Natl Acad Sci U S A* **109**: 20655–20660.

Lane DJ. (1991). 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic Acid Techniques in Bacterial Systematics*. Wiley: Chichester, pp 115–175.

Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Larsson J, Nylander JAA, Bergman B. (2011). Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evol Biol* **11**: 187.

Lasken RS, McLean JS. (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* **15**: 577–584.

Latifi A, Ruiz M, Zhang C-C. (2009). Oxidative stress in cyanobacteria. *FEMS Microbiol Rev* **33**: 258–278.

Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci U S A* **106**: 15527–15533.

Lê S, Josse J, Husson F. (2008). FactoMineR : An R package for multivariate analysis. *J Stat Softw* **25**: 1–18.

Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M. (2014). Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv*. e-pub ahead

of print, doi: 10.1101/006395.

Lerouge I, Vanderleyden J. (2001). O-antigen structural variation : Mechanisms and possible roles in animal/plant - microbe interactions. *FEMS Microbiol Rev* **26**: 17–47.

Li B, Yang Y, Ma L, Ju F, Guo F, Tiedje JM, *et al.* (2015). Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J* **9**: 2490–2502.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liao Y-C, Lin S-H, Lin H-H. (2015). Completing bacterial genome assemblies: strategy and performance comparisons. *Sci Rep* **5**: 8747.

Liu M, Fan L, Zhong L, Kjelleberg S, Thomas T. (2012). Metaproteogenomic analysis of a community of sponge symbionts. *ISME J* **6**: 1515–1525.

Liu MY, Kjelleberg S, Thomas T. (2011). Functional genomic analysis of an uncultured δ-proteobacterium in the sponge *Cymbastela concentrica*. *ISME J* **5**: 427–435.

López-Fuentes E, Torres-Tejerizo G, Cervantes L, Brom S. (2015). Genes encoding conserved hypothetical proteins localized in the conjugative transfer region of plasmid pRet42a from *Rhizobium etli* CFN42 participate in modulating transfer and affect conjugation from different donors. *Front Microbiol* **5**: Article 793.

Macaulay IC, Voet T. (2014). Single cell genomics: Advances and future perspectives. *PLoS Genet* **10**: e1004126.

Madoui M-A, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, *et al.* (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* **16**: 327.

Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, *et al.* (2011). Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**: 467–477.

Makarova KS, Wolf YI, Koonin E V. (2009). Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol Direct* **4**. e-pub ahead of print, doi: 10.1186/1745-6150-4-19.

Maldonado M. (2015). Sponge waste that fuels marine oligotrophic food webs: a re-assessment of its origin and nature. *Mar Ecol* 1–15.

Maldonado M, Aguilar R, Bannister RJ, James J, Conway KW, Dayton PK, *et al.* (2016). Sponge grounds as key marine habitats: A synthetic review of types, structure,

functional roles, and conservation concerns. In: *Marine Animal Forests*. e-pub ahead of print, doi: 10.1007/978-3-319-17001-5.

Malgieri G, Palmieri M, Russo L, Fattorusso R, Pedone P V., Isernia C. (2015). The prokaryotic zinc-finger: Structure, function and comparison with the eukaryotic counterpart. *FEBS J* **282**: 4480–4496.

Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, *et al.* (2007). Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A* **104**: 11889–11894.

Marino M, Braun L, Cossart P, Ghosh P. (1999). Structure of the InlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen *L. monocytogenes*. *Mol Cell* **4**: 1063–1072.

Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Grechkin Y, *et al.* (2012). IMG/M: The integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* **40**: D123–D129.

Marston MF, Pierciey FJ, Shepard A, Gearin G, Qi J, Yandava C, *et al.* (2012). Rapid diversification of coevolving marine *Synechococcus* and a virus. *Proc Natl Acad Sci* **109**: 4544–4549.

Martín-Cuadrado AB, López-García P, Alba JC, Moreira D, Monticelli L, Strittmatter A, *et al.* (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One* **2**. e-pub ahead of print, doi: 10.1371/journal.pone.0000914.

Mason OU, Hazen TC, Borglin S, Chain PSG, Dubinsky EA, Fortney JL, *et al.* (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J* **6**: 1715–1727.

Mayer EA, Knight R, Mazmanian SK, Cryan JF, Tillisch K. (2014). Gut microbes and the brain: Paradigm shift in neuroscience. *J Neurosci* **34**: 15490–15496.

McFall-Ngai M, Hadfield MG, Bosch TCG, Carey H V., Domazet-Lošo T, Douglas AE, *et al.* (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proc Natl Acad Sci* **110**: 3229–3236.

McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, *et al.* (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc Natl Acad Sci* **110**: E2390–E2399.

Meadows JA, Wargo MJ. (2015). Carnitine in bacterial physiology and metabolism.

Microbiology **161**: 1161–1174.

Meetam M, Keren N, Ohad I, Pakrasi HB. (1999). The PsbY protein is not essential for oxygenic photosynthesis in the cyanobacterium *Synechocystis* sp. PCC 6803. *Plant Physiol* **121**: 1267–1272.

Mehbub M, Lei J, Franco C, Zhang W. (2014). Marine sponge derived natural products between 2001 and 2010: Trends and opportunities for discovery of bioactives. *Mar Drugs* **12**: 4539–4577.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, *et al.* (2008). The metagenomics RAST server - a public resource for the automatic phylo- genetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Moliner C, Fournier P-E, Raoult D. (2010). Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS Microbiol Rev* **34**: 281–294.

Moran NA, Plague GR. (2004). Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* **14**: 627–633.

Namiki T, Hachiya T, Tanaka H, Sakakibara Y. (2012). MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* **40**: e155.

Nguyen MTHD, Liu M, Thomas T. (2014). Ankyrin-repeat proteins from sponge symbionts modulate amoebal phagocytosis. *Mol Ecol* **23**: 1635–1645.

Nowack ECM, Melkonian M, Glöckner G. (2008). Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol* **18**: 410–418.

Noyer C, Hamilton A, Sacristan-Soriano O, Becerro MA. (2010). Quantitative comparison of bacterial communities in two Mediterranean sponges. *Symbiosis* **51**: 239–243.

Ollivier PRL, Bahrou AS, Marcus S, Cox T, Church TM, Hanson TE. (2008). Volatilization and precipitation of tellurium by aerobic, tellurite-resistant marine microbes. *Appl Environ Microbiol* **74**: 7163–7173.

Ono Y, Asai K, Hamada M. (2013). PBSIM: PacBio reads simulator - toward accurate genome assembly. *Bioinformatics* **29**: 119–121.

Oren M, Steindler L, Ilan M. (2005). Transmission, plasticity and the molecular identification of cyanobacterial symbionts in the Red Sea sponge *Diacarnus erythraeus*. *Mar Biol* **148**: 35–41.

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, *et al.* (2014). The SEED and

the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* **42**: D206–D214.

Parks DH, Tyson GW, Hugenholtz P, Beiko RG. (2014). STAMP: Statistical analysis of taxonomic and functional profiles. *Bioinformatics* **30**: 3123–3124.

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: A *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Pfannkuchen M, Fritz GB, Schlesinger S, Bayer K, Brümmer F. (2009). *In situ* pumping activity of the sponge *Aplysina aerophoba*, Nardo 1886. *J Exp Mar Bio Ecol* **369**: 65–71.

Pile AJ, Patterson MR, Witman JD. (1996). *In situ* grazing on plankton *Mycale lingua*. *Mar Ecol Prog Ser* **141**: 95–102.

Pimentel-Elardo SM, Grozdanov L, Proksch S, Hentschel U. (2012). Diversity of nonribosomal peptide synthetase genes in the microbial metagenomes of marine sponges. *Mar Drugs* **10**: 1192–1202.

Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, *et al.* (2012). eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* **40**: D284–D289.

Pruesse E, Peplies J, Glöckner FO. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* **41**: 590–596.

Radax R, Hoffmann F, Rapp HT, Leininger S, Schleper C. (2012a). Ammonia-oxidizing archaea as main drivers of nitrification in cold-water sponges. *Environ Microbiol* **14**: 909–923.

Radax R, Rattei T, Lanzen A, Bayer C, Rapp HT, Urich T, *et al.* (2012b). Metatranscriptomics of the marine sponge *Geodia barretti*: Tackling phylogeny and function of its microbial community. *Environ Microbiol* **14**: 1308–1324.

Reddy TBK, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, *et al.* (2014). The Genomes OnLine Database (GOLD) v.5: A metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res* **43**: D1099–D1106.

Reiswig HM. (1974). Water transport, respiration and energetics of three tropical marine

sponges. *J Exp Mar Bio Ecol* **14**: 231–249.

Revell LJ. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* **3**: 217–223.

Rhoads A, Au KF. (2015). PacBio sequencing and its applications. *Genomics, Proteomics Bioinforma* **13**: 278–289.

Ricker N, Shen SY, Goordial J, Jin S, Fulthorpe RR. (2016). PacBio SMRT assembly of a complex multi-replicon genome reveals chlorocatechol degradative operon in a region of genome plasticity. *Gene* **586**: 239–247.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Rivkina E, Petrovskaya L, Vishnivetskaya T, Krivushin K, Shmakova L, Tutukina M, *et al.* (2016). Metagenomic analyses of the late Pleistocene permafrost - additional tools for reconstruction of environmental conditions. *Biogeosciences* **13**: 2207–2219.

Rix L, de Goeij JM, Mueller CE, Struck U, Middelburg JJ, van Duyl FC, *et al.* (2016a). Coral mucus fuels the sponge loop in warm- and cold-water coral reef ecosystems. *Sci Rep* **6**. e-pub ahead of print, doi: 10.1038/srep18715.

Rix L, de Goeij JM, van Oevelen D, Struck U, Al-Horani FA, Wild C, *et al.* (2016b). Differential recycling of coral and algal dissolved organic matter via the sponge loop Power S (ed). *Funct Ecol*. e-pub ahead of print, doi: 10.1111/1365-2435.12758.

Rocap G, Distel DL, Waterbury JB, Chisholm SW. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68**: 1180–1191.

Rotthauwe JH, Witzel KP, Liesack W. (1997). The ammonia monooxygenase structural gene *amoA* as a functional marker: Molecular fine-scale analysis of natural ammonia-oxidizing populations. *Appl Environ Microbiol* **63**: 4704–4712.

Rützler K. (1970). Spatial competition among Porifera: Solution by epizoism. *Oecologia* **5**: 85–95.

Sacristan-Soriano O, Banaigs B, Becerro MA. (2011). Relevant spatial scales of chemical variation in *Aplysina aerophoba*. *Mar Drugs* **9**: 2499–2513.

Sacristán-Soriano O, Banaigs B, Becerro M a. (2012). Temporal trends in the secondary metabolite production of the sponge Aplysina aerophoba. *Mar Drugs* **10**: 677–693.

Sberro H, Leavitt A, Kiro R, Koh E, Peleg Y, Qimron U, *et al.* (2013). Discovery of

functional toxin/antitoxin systems in bacteria by shotgun cloning. *Mol Cell* **50**: 136–148.

Schmitt S, Angermeier H, Schiller R, Lindquist N, Hentschel U. (2008). Molecular microbial diversity survey of sponge reproductive stages and mechanistic insights into vertical transmission of microbial symbionts. *Appl Environ Microbiol* **74**: 7694–7708.

Schmitt S, Hentschel U, Taylor MW. (2012a). Deep sequencing reveals diversity and community structure of complex microbiota in five Mediterranean sponges. *Hydrobiologia* **687**: 341–351.

Schmitt S, Hentschel U, Zea S, Dandekar T, Wolf M. (2005). ITS-2 and 18S rRNA gene phylogeny of Aplysinidae (Verongida, Demospongiae). *J Mol Evol* **60**: 327–336.

Schmitt S, Tsai P, Bell J, Fromont J, Ilan M, Lindquist N, *et al.* (2012b). Assessing the complex sponge microbiota: Core, variable and species-specific bacterial communities in marine sponges. *ISME J* **6**: 564–576.

Schröder K, Bosch TCG. (2016). The origin of mucosal immunity: Lessons from the holobiont *Hydra*. *MBio* **7**: e01184-16.

Shibata TF, Maeda T, Nikoh N, Yamaguchi K, Oshima K, Hattori M, *et al.* (2013). Complete genome sequence of *Burkholderia* sp. strain RPE64, bacterial symbiont of the bean bug *Riptortus pedestris*. *Genome Announc* **1**: e00441-13.

Siegl A, Hentschel U. (2010). PKS and NRPS gene clusters from microbial symbiont cells of marine sponges by whole genome amplification. *Environ Microbiol Rep* **2**: 507–513.

Siegl A, Kamke J, Hochmuth T, Piel J, Richter M, Liang C, *et al.* (2011). Single-cell genomics reveals the lifestyle of *Poribacteria*, a candidate phylum symbiotically associated with marine sponges. *ISME J* **5**: 61–70.

Simkovsky R, Daniels EF, Tang K, Huynh SC, Golden SS, Brahamsha B. (2012). Impairment of O-antigen production confers resistance to grazing in a model amoeba-cyanobacterium predator-prey system. *Proc Natl Acad Sci* **109**: 16678–16683.

Simpson TL. (1984). The cell biology of sponges. Springer-Verlag: New York.

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**: 241–244.

Snel B, Lehmann G, Bork P, Huynen MA. (2000). STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28**: 3442–3444.

Snyder DS, Brahamsha B, Azadi P, Palenik B. (2009). Structure of compositionally simple lipopolysaccharide from marine *Synechococcus*. *J Bacteriol* **191**: 5499–5509.

van Soest RWM, Boury-Esnault N, Vacelet J, Dohrmann M, Erpenbeck D, de Voogd NJ, *et al.* (2012). Global diversity of sponges (Porifera). *PLoS One* **7**: e35105.

Steindler L, Huchon D, Avni A, Ilan M. (2005). 16S rRNA phylogeny of sponge-associated cyanobacteria. *Appl Environ Microbiol* **71**: 4127–4131.

Stoddard LI, Martiny JBH, Marston MF. (2007). Selection and characterization of cyanophage resistance in marine *Synechococcus* strains. *Appl Environ Microbiol* **73**: 5516–5522.

Sveshnikov D, Funk C, Schröder WP. (2007). The PsbP-like protein (sll1418) of *Synechocystis* sp. PCC 6803 stabilises the donor side of photosystem II. *Photosynth Res* **93**: 101–109.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci U S A* **110**: 11463–8.

Swatschek D, Schatton W, Kellermann J, Müller WEG, Kreuter J. (2002). Marine sponge collagen: Isolation, characterization and effects on the skin parameters surface-pH, moisture and sebum. *Eur J Pharm Biopharm* **53**: 107–113.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, *et al.* (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**: D447–D452.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**: 2725–2729.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V., *et al.* (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

Taylor MW, Radax R, Steger D, Wagner M. (2007). Sponge-associated microorganisms: Evolution, ecology, and biotechnological potential. *Microbiol Mol Biol Rev* **71**: 295–347.

Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, *et al.* (2015). ProDeGe: A computational protocol for fully automated decontamination of genomes. *ISME J* **10**: 269–272.

Thacker RW, Freeman CJ. (2012). Sponge-microbe symbioses: Recent advances and new directions. *Adv Mar Biol* **62**: 57–111.

The MetaSUB International Consortium. (2016). The metagenomics and metadesign of the subways and urban biomes (MetaSUB) international consortium inaugural meeting report. *Microbiome* **4**. e-pub ahead of print, doi: 10.1186/s40168-016-0168-z.

Thomas T, Moitinho-Silva L, Lurgi M, Björk JR, Easson C, Astudillo-García C, *et al.* (2016). Diversity, structure and convergent evolution of the global sponge microbiome. *Nat Commun* **7**: 11870.

Thomas T, Rusch D, DeMaere MZ, Yung PY, Lewis M, Halpern A, *et al.* (2010). Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J* **4**: 1557–1567.

Tian R-M, Sun J, Cai L, Zhang W-P, Zhou G-W, Qiu J-W, *et al.* (2016). The deep-sea glass sponge *Lophophysema eversa* harbours potential symbionts responsible for the nutrient conversions of carbon, nitrogen and sulfur. *Environ Microbiol* **18**: 2481–2494.

Tian R-M, Wang Y, Bougouffa S, Gao Z-M, Cai L, Bajic V, *et al.* (2014). Genomic analysis reveals versatile heterotrophic capacity of a potentially symbiotic sulfur-oxidizing bacterium in sponge. *Environ Microbiol* **16**: 3548–3561.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**: 90–94.

Tsai Y-C, Conlan S, Deming C, NISC Comparative Sequencing Program, Segre JA, Kong HH, *et al.* (2016). Resolving the complexity of human skin metagenomes using single-molecule sequencing. *MBio* **7**: e01948-15.

Turon X, Becerro MA, Uriz MJ. (2000). Distribution of brominated compounds within the sponge *Aplysina aerophoba*: Coupling of X-ray microanalysis with cryofixation techniques. *Cell Tissue Res* **301**: 311–322.

Usher KM. (2008). The ecology and phylogeny of cyanobacterial symbionts in sponges. *Mar Ecol* **29**: 178–192.

Usher KM, Kuo J, Fromont J, Sutton DC. (2001). Vertical transmission of cyanobacterial symbionts in the marine sponge *Chondrilla australiensis* (Demospongiae). *Hydrobiologia* **461**: 15–23.

Usher KM, Kuo J, Fromont J, Toze S, Sutton DC. (2006). Comparative morphology of five species of symbiotic and non-symbiotic coccoid cyanobacteria. *Eur J Phycol* **41**: 179–188.

Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, *et al.* (2014). Evaluation and validation of *de novo* and hybrid assembly techniques to

derive high-quality genome sequences. *Bioinformatics* **30**: 2709–2716.

Vacelet J. (1975). Etude en microscopie electronique de l'association entre bacteries et spongiaires du genre Verongia (Dictyoceratida). *J Microsc Biol Cell* **23**: 271–288.

Vacelet J. (1999). Planktonic armoured propagules of the excavating sponge *Alectona* (Porifera: Demospongiae) are larvae: evidence from *Alectona wallichii* and *A. mesatlantica* sp.nov. *Mem Queensl Museum* **44**: 627–642.

Vasu K, Nagaraja V. (2013). Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev* **77**: 53–72.

Vik U, Logares R, Blaalid R, Halvorsen R, Carlsen T, Bakke I, *et al.* (2013). Different bacterial communities in ectomycorrhizae and surrounding soil. *Sci Rep* **3**: 3471.

Vilanova E, Coutinho CC, Mourão PAS. (2009). Sulfated polysaccharides from marine sponges (Porifera): An ancestor cell-cell adhesion event based on the carbohydrate-carbohydrate interaction. *Glycobiology* **19**: 860–867.

Vogel S. (1977). Current-induced flow through living sponges in nature. *Proc Natl Acad Sci U S A* **74**: 2069–2071.

Voultsiadou E. (2007). Sponges: An historical survey of their knowledge in Greek antiquity. *J Mar Biol Assoc United Kingdom* **87**: 1757–1763.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

Webster NS, Cobb RE, Negri AP. (2008a). Temperature thresholds for bacterial symbiosis with a sponge. *ISME J* **2**: 830–842.

Webster NS, Taylor MW, Behnam F, Lücker S, Rattei T, Whalan S, *et al.* (2010). Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. *Environ Microbiol* **12**: 2070–2082.

Webster NS, Thomas T. (2016). The sponge hologenome. *MBio* **7**: e00135-16.

Webster NS, Xavier JR, Freckelton M, Motti CA, Cobb R. (2008b). Shifts in microbial and chemical patterns within the marine sponge *Aplysina aerophoba* during a disease outbreak. *Environ Microbiol* **10**: 3366–3376.

Wehrl M, Steinert M, Hentschel U. (2007). Bacterial uptake by the marine sponge *Aplysina aerophoba*. *Microb Ecol* **53**: 355–365.

Weinberger AD, Sun CL, Pluciński MM, Denef VJ, Thomas BC, Horvath P, *et al.* (2012). Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput*

Biol **8**: e1002475.

Wilkinson CR. (1980). Cyanobacteria symbiotic in marine sponges. In: Schwemmler W, Schenk HEA (eds). *Endocytobiology: Endosymbiosis and Cell Biology. A Synthesis of Recent Research*. Walter de Gruyter: Berlin, New York, pp 553–563.

Wilkinson CR. (1978a). Microbial associations in sponges. I. Ecology, physiology and microbial populations of coral reef sponges. *Mar Biol* **49**: 161–167.

Wilkinson CR. (1978b). Microbial associations in sponges. III. Ultrastructure of the *in situ* associations in coral reef sponges. *Mar Biol* **49**: 177–185.

Wilkinson CR. (1983). Net primary productivity in coral reef sponges. *Science* **219**: 410–412.

Wilkinson CR, Garrone R, Vacelet J. (1984). Marine sponges discriminate between food bacteria and bacterial symbionts: Electron microscope radioautography and *in situ* evidence. *Proc R Soc London, Ser B Biol Sci* **220**: 519–528.

Wilson MC, Mori T, Rückert C, Uria AR, Helf MJ, Takada K, *et al.* (2014). An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**: 58–62.

Woyke T, Sczyrba A, Lee J, Rinke C, Tighe D, Clingenpeel S, *et al.* (2011). Decontamination of MDA reagents for single cell whole genome amplification. *PLoS One* **6**: e26161.

Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, *et al.* (2010). One bacterial cell, one complete genome. *PLoS One* **5**: e10314.

Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, *et al.* (2009). Assembling the marine metagenome, one cell at a time. *PLoS One* **4**: e5299.

Wu M, Sun L V., Vamathevan J, Riegler M, Deboy R, Brownlie JC, *et al.* (2004). Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: A streamlined genome overrun by mobile genetic elements. *PLoS Biol* **2**: E69.

Wu S, Zhu Z, Fu L, Niu B, Li W. (2011). WebMGA: A customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**: 444.

Wulff JL. (2008). Collaboration among sponge species increases sponge diversity and abundance in a seagrass meadow. *Mar Ecol* **29**: 193–204.

Wulff JL. (1984). Sponge-mediated coral reef growth and rejuvenation. *Coral Reefs* **3**: 157–163.

Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, *et al.* (2011).

Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**: 714–717.

Zerbino DR, Birney E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

Zhang M-J, Zhao F, Xiao D, Gu Y-X, Meng F-L, He L-H, *et al.* (2009). Comparative proteomic analysis of passaged *Helicobacter pylori*. *J Basic Microbiol* **49**: 482–490.

Zheng H, Dietrich C, Hongoh Y, Brune A. (2016). Restriction-modification systems as mobile genetic elements in the evolution of an intracellular symbiont. *Mol Biol Evol* **33**: 721–725.

Zierer MS, Mourão PAS. (2000). A wide diversity of sulfated polysaccharides are synthesized by different species of marine sponges. *Carbohydr Res* **328**: 209–216.

6 Appendix

The following appendices are submitted on a CD attached to this thesis:

Appendix 3-1 Genome completeness estimation based on 111 single-copy essential genes (Albertsen et al., 2013). Only genes abundant in at least one bin are shown. Printed in bold are genes that can occur in duplicates. Genes marked with an asterisk were used for the concatenated gene phylogeny in Figure 3-23.

Appendix 3-2 COG annotations for sponge symbionts and references after filtering: only annotations with an e-value ≤ 1e-6 were kept, and only one annotation per ORF was kept ranked by e-value, length and bitscore.

Appendix 3-3 Overview table of COG annotations of sponge symbiont and reference genomes.

7 Publications list

Slaby BM, Hentschel U. (2017). Draft genome sequences of "*Candidatus* Synechococcus spongiarum," cyanobacterial symbionts of the Mediterranean sponge *Aplysina aerophoba*. *Genome Announc*. Accepted.

Horn H, **Slaby BM**, Jahn MT, Bayer K, Moitinho-Silva L, Förster F, Abdelmohsen UR, Hentschel U. (2016). An enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes. *Front Microbiol* **7**: 1751.

Burgsdorf I*, **Slaby BM***, Handley KM, Haber M, Blom J, Marshall CW, Gilbert JA, Hentschel U, Steindler L. (2015). Lifestyle evolution in cyanobacterial symbionts of sponges. *mBio* **6**: e00391-15. *shared first authorship

Bayer K, **Slaby BM**, Hentschel U. (2015). Den Unkultivierbaren auf der Spur. *BIOspektrum* **21**: 17-19. Not peer-reviewed.

8 Curriculum vitae

Education

July 2015 – present	PhD studies continued at GEOMAR Helmholtz Centre for Ocean Research Kiel, Germany
February 2013 – present	PhD studies in Life Sciences (Dr. rer. nat.), Julius-Maximilians University of Würzburg, Germany
	PhD thesis "Exploring the microbiome of the Mediterranean sponge *Aplysina aerophoba* by single-cell and metagenomics," Supervisor: Prof. Dr. Ute Hentschel Humeida
2010 - 2012	Master in Geological Sciences, Ludwig-Maximilians University and Technical University Munich, Germany
	Master Thesis: "Assessment of sponge-associated symbionts and changes in the community under climate change scenarios," Supervisors: Prof. Dr. Gert Wörheide, Dr. Susanne Schmitt
2007 – 2010	Bachelor in Geosciences, Ludwig-Maximilians University and Technical University Munich, Germany
	Bachelor Thesis: "3D reconstruction of a living foraminifer," Supervisor: Prof. Dr. Alexander Altenbach

Field-going and research experience

September – October 2016	Cruise PS101 on research vessel FS Polarstern, chief scientist: Prof. Dr. Antje Boetius, Alfred-Wegener Institute for Polar and Ocean Research Bremerhaven, Germany
May 2014	Field excursion to Piran, Slovenia, with Dr. Kristina Bayer, University of Würzburg
January – March 2014	Research stay at the Joint Genome Institute of the Department of Energy, Walnut Creek, CA, USA, under the supervision of Dr. Tanja Woyke
May 2013	Field excursion to Rovinj, Croatia, with Dr. Kristina Bayer, University of Würzburg
October – November 2012	Cruise SO224 on research vessel FS Sonne, chief scientist: Prof. Dr. Gabriele Uenzelmann-Neben, Alfred-Wegener Institute for Polar and Ocean Research Bremerhaven, Germany
May – July 2009	Cruise MSM12/2 on research vessel FS Maria S. Merian, chief scientist: Prof. Dr. Gabriele Uenzelmann-Neben
May 2009	Cruise ANT XXV/5 on research vessel FS Polarstern, chief scientist: Prof. Dr. Saad El Dine El Naggar

Selected scientific conferences

July 2016	Essential Ocean Variables Workshop for monitoring and assessment of marine biodiversity and ecosystem health, Kiel, Germany, poster: 'SponGES: deep-sea sponge grounds ecosystems of the North Atlantic'
July 2015	Gordon Research Conference Applied & Environmental Microbiology, South Hadley, MA, USA, short-talk and poster: 'Lifestyle evolution in cyanobacterial symbionts of sponges'
March 2015	2nd Global Invertebrate Genomics Alliance (GIGA) Workshop, Munich, Germany, poster: 'Genomics of "*Candidatus* Synechococcus spongiarum", a Cyanobacterial Sponge Symbiont'
October 2014	2nd International Symposium on Sponge Microbiology, Baltimore, MD, USA, poster: 'Genomics of "*Candidatus* Synechococcus spongiarum", a Cyanobacterial Sponge Symbiont'
March 2014	Annual JGI User Meeting, Joint Genome Institute of the Department of Energy, Walnut Creek, CA, USA, poster: 'Genomics of "*Candidatus* Synechococcus spongiarum", a Cyanobacterial Sponge Symbiont'

Awards

February 2013 – January 2016	Grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg, Germany, including a PhD fellowship and coverage of laboratory and travel expenses.

Language Skills

German (native), English (fluent), Italian (basic), Latin

Place, Date

Signature