

***“DIVERSITY AND SPECIFICITY OF THE MARINE SPONGE MICROBIOME AS
INSPECTED BY NEXT GENERATION SEQUENCING”***

André Rodrigues Soares
Mestrado em Biologia Molecular e Microbiana

Trabalho efetuado sob a orientação de:
Dr. Rodrigo da Silva Costa
Prof.^a Dr.^a Maria Margarida dos Prazeres Reis

Declaração de autoria de trabalho

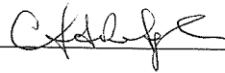
Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.



André Rodrigues Soares

©Copyright: André Rodrigues Soares

A Universidade do Algarve tem o direito, perpétuo e sem limites geográficos, de arquivar e publicitar este trabalho através de exemplares impressos reproduzidos em papel ou de forma digital, ou por outro qualquer meio conhecido ou que venha a ser inventado, de o divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.



André Rodrigues Soares

Acknowledgements

For the completion of the present work, the support of the Microbial Ecology and Evolution (MicroEcoEvo) research group at the CCMAR was crucial. Firstly, I thank Rodrigo Costa, my supervisor, who drew me into the field of microbial ecology of marine sponges and posed me the great challenge of tackling the EMP dataset. Furthermore, I am thankful for all the enthusiasm and opportunities provided and for the immense patience! Gianmaria Califano, now in Jena, Austria, and Asunción Lago-Lestón dispensed precious time and patience in the starting phase of this thesis. I further thank Elham Karimi for the help in the laboratory and for our relaxing coffee talks! Tina Keller-Costa, Telma Franco and more recently Miguel Ramos, along with the abovementioned, are all part of the MicroEcoEvo team, to whom I thank for a wonderful first medium-term experience in a laboratory!

I admittedly started my Biology Bsc. at the University of Algarve not being fond of any biological entity smaller than 2cm. It was Prof. Margarida Reis (Microbial and Molecular Ecology Laboratory, CIMA, UAlg) who showed me that ‘microbes can do anything’ and went on to accepting the challenge of supervising my Bsc. Technical and Scientific Project. I therefore thank her for introducing me to Microbiology in the best way possible.

For the completeness of this work, Lucas Moitinho-Silva (currently at the University of South Wales, Australia), was essential in introducing me to R scripting whilst finishing his PhD at Wüezburg University, Germany. I still cannot imagine how one is able to answer basic R questions the night before defending his PhD thesis, but Lucas did, and I thank him for his incredible availability and kindness throughout this year. Cymon Cox, head of the Plant Systematics and Bioinformatics group in CCMAR, very kindly introduced me to python scripting and handling of the Linux OS. Thoughtfully, Cole Easson and Robert Thacker from Stony Brook University, USA, let me know of various tools for 16S dataset analysis. All the help received from Pat Schloss and Sarah Wescott, mothur *software* developers at the University of Michigan was precious.

Out of the academic scope, ‘Associação Geonauta’, ‘my’ caving association in the Algarve has, for the last 4 years, given me the opportunity to discover the subterranean wonders of Algarve. I have learned greatly from all, but mainly from Cristiano Cavaco, Carlos Oliveira and Friedrich Zabel. Prof. Cristina Veiga-Pires, whom I met in a cave, has been a great supporter of our work in Associação Geonauta and a friend. I also want to thank all those I knew during my experience at the Biology Students Association at

the University of Algarve (NEBUA) and at the National Board of Biology Students (ANEBio). They are too many to list, but they know who they are!

Personally, I thank Andreia Nunes and Catarina Diniz, my friends and colleagues in the Molecular and Microbial Biology Msc., who accompanied me throughout all kinds of hurdles and coffee breaks during this thesis. Jessica Fróis and Beatriz Cruz are further credited for frequently giving out valuable wise advices which truly helped me for the last 4 years. Starting my Bsc. in 2009, I met Nadja Velez, who is to this day one of my best friends. Throughout these 6 years, she showed me how to get excited with Science and accompanied me in all sorts of events.

The support of my family has been crucial throughout my life, but mostly during its academic component and all external events occurring during that time. I would therefore like to dedicate this thesis to my late grandfather Fernando Rodrigues, who used to say that nothing in life is impossible if one works for it. He always supported my choices, and will always be a reference and source of inspiration to me regarding my personal or working future life. My mother Isabel and my sister Carolina have been essential throughout my life and both are therefore credited for having raised me to be who I am today.

Finally, I thank Sónia Serrão, who has been for the last three years my source of inspiration and joy. A lot is to come, but we will get through it all!

Abstract

Sponges are early-branching metazoans whose hosted microbial communities are currently seen as highly fruitful sources of microbial ecological, evolutionary and metabolic novelties.

This study aims to determine the composition and structure of the prokaryotic communities found in four marine sponges found off the coast of Algarve in 2012. These comprised 26 specimens belonging to the species *Phorbast fictitius* (n=12), *Dysidea fragilis* (n=3), *Cliona viridis* (n=4) and *C. celata* (n=7). Prior to this thesis, the latter specimens were subjected to DNA extraction, 16S rRNA gene-directed PCR, Illumina HiSeq sequencing and sequence pre-processing according to Earth Microbiome Project (EMP) standards.

A total of 3,215 prokaryotic Operational Taxonomic Units (OTUs), defined at 97% similarity, were identified from 291,278 sequence reads considered in the normalized dataset, which was rarefied for 11,203 sequences per sample, for statistical purposes. Rarefaction curves revealed that, in spite of the high sequencing effort employed, all sponge species, except *C. viridis*, would require further sequencing depth for complete coverage of their associated prokaryotic communities. Bray-Curtis dissimilarity-based non-metric multidimensional scaling (nMDS) showed distinct and host-specific communities that did not follow host phylogeny. High proteobacterial dominance (mostly α - and γ -Proteobacteria) was observed across all sponge-associated prokaryotic consortia, except in the case of *C. celata*, where unclassified bacteria prevailed. Mostly species-specific OTUs classified as α -, γ - and unclassified Proteobacteria were seen to predominate the core of highly abundant phylotypes, as seen by heatmap plotting of decreasing abundances. Further, using stringent phylogenetic assessments, it was possible to re-classify abundant bacterial phylotypes previously regarded as “unclassified” based on database-dependent taxonomic assignments.

Here, host-specific prokaryomes were found, meaning that other factors than host phylogeny must drive sponge prokaryome structure and composition. Moreover, striking prokaryotic diversity was noted within the surveyed sponge hosts, suggesting that further metagenome mining of these prokaryomes may unveil further novelties.

Keywords: marine sponges, 16S rRNA, OTUs, prokaryome, singletome.

Resumo

As esponjas marinhas (filo Porifera) vivem em simbiose com microrganismos que frequentemente apresentam alto interesse ecológico, evolutivo e metabólico. A descoberta da presença de procariotas simbioses em esponjas marinhas ocorreu nos anos 1970. No entanto, foi nos anos 2000 que o surgimento das tecnologias de sequenciação de última geração permitiu conhecer mais a fundo as interações simbiote-hospedeiro.

Atualmente, estas tecnologias são amplamente aplicadas e com grande êxito na exploração molecular das simbioses que ocorrem entre micróbios e esponjas. Ademais, por via destas, tem sido possível desvendar o nível de ‘intimidade’ molecular a que vivem estes organismos. Sabe-se hoje em dia que as comunidades microbianas simbioses de esponjas marinhas contribuem imensamente para o ‘bem-estar’ do seu hospedeiro por via da produção de moléculas antimicrobianas, por exemplo. Por outro lado, por meio do hospedeiro são ‘providenciados’ abrigo, produtos finais metabólicos como amónia, e produtos orgânicos que derivam da predação por filtração de plâncton unicelular. Todos os anteriores fatores contribuem de forma ainda não verdadeiramente quantificada para o estabelecimento e estruturação do microbioma destes metazoários. Neste estudo, foi abordado o procarioma (isto é, o consórcio de todas as bactérias e arqueias -procariotas- presentes num dado ambiente) de quatro esponjas marinhas. Estas foram amostradas ao largo da costa do Algarve e a baixa profundidade em 2012. No total, 26 espécimes pertencentes às espécies *Phorbastictitius* (n=12), *Dysidea fragilis* (n=3), *Cliona viridis* (n=4) and *C. celata* (n=7) foram recolhidos. Todas as amostras foram processadas no sentido de isolar o endossoma de cada indivíduo, isto é, o seu interior, coberto pelo ectossoma. Todas as amostras (indivíduos) foram ademais sujeitas à extração de ADN segundo protocolos padronizados pelo ‘Earth Microbiome Project’.

Posteriormente, o ARN ribossomal 16S destas amostras foi amplificado por via de PCR (reação de polimerase em cadeia) e sequenciado por via de tecnologia Illumina HiSeq na sede do Earth Microbiome Project, nos EUA. Os produtos de sequenciação foram então pré-processados por *software* específico no sentido de eliminar ou corrigir possíveis erros de diversas origens, criar unidades taxonómicas operacionais (OTUs) a um índice de similaridade de 97% e classificá-las de acordo com bases de dados taxonómicas de referência. Esta tese teve como objetivo inferir características ecológicas diversas dos procariomas encontrados em esponjas marinhas, focando

adicionalmente a componente rara destes: o “singletoma” (‘singletome’, na tradução em inglês). Para isto, foram utilizados diversos *software* e mesmo a linha de comandos do sistema operativo Ubuntu. Foi ademais caracterizada a filogenia das esponjas em estudo com base nos perfis de similaridade do gene da subunidade 1 da proteína citocromo oxidase (*cox1*). Assim, extraiu-se ADN de amostras conservadas das esponjas em estudo, ao que se amplificou, sequenciou e analisou o gene *cox1*.

Após filtração das amostras relativas a este estudo, seguiu-se a normalização das amostras, aplicada em função do menor número de sequências encontradas numa dada amostra. Assim, o limiar de normalização foi definido em 11.203 sequências por indivíduo, associadas a um total geral de 3.215 OTUs. Por via de análise de rarefação, foi possível estimar que apenas o procarioma de *C. viridis* foi sequenciado a uma ‘profundidade’ tomada como verdadeiramente representativa da comunidade procariótica da esponja (no sentido da sequenciação de todos os produtos de amplificação na amostra). Deste modo, o acesso a uma base de dados verdadeiramente representativa dos procariomas nestas esponjas dependerá de futuras rondas de sequenciação a maior profundidade. Ordenação da similaridade entre procariomas revelou perfis de grande intraespecificidade e notória diferenciação entre amostras correspondentes a diferentes esponjas (inter-especificidade).

No geral, α -, γ -, bem como clades não classificadas de Proteobacteria dominaram todos os procariomas, com a exceção de *C. celata*, em que dominaram OTUs bacterianas não classificadas (segundo classificação atribuída durante o pré-processamento no âmbito do EMP em 2013). Independentemente da filogenia dos hospedeiros, estes procariomas mostraram-se específicos de cada esponja, enquanto que os das esponjas do género *Cliona* se revelaram os mais dissimilares. Nos procariomas das restantes esponjas, *P. fictitius* e *D. fragilis*, notaram-se padrões de alguma similaridade, apesar de estas pertencerem a duas sub-classes diferentes (respetivamente *Heteroscleromorpha sensu* Cárdenas et al., 2012 e *Keratosa*, classe *Demospongiae*). Por meio de visualização das OTUs mais abundantes, observou-se também que cada esponja albergava um conjunto de dois a cinco filótipos específicos. Em *C. viridis*, cinco destas, muito abundantes, definiam o maior ‘núcleo’ de simbiontes específicos encontrados, possivelmente influenciando por demais os valores de equitabilidade atribuídos. *D. fragilis* revelou também um núcleo dominante, ainda que composto por OTUs menos abundantes que os restantes.

Por outro lado, a componente rara deste procarioma ('singletons', ou OTUs com apenas uma sequência atribuída) revelou diversidade maioritariamente conhecida. Ao contrário do esperado, o 'singletonoma', isto é, o conjunto de todos os singletons numa base de dados, foi na sua maioria eficazmente identificado. Além disso, não foi encontrada qualquer tendência para padrões de representação específica de singletons de determinado filo ao longo das esponjas marinhas amostradas.

Dada a vincada expressão de OTUs de taxonomia bacteriana não identificada, uma base de dados relativa apenas às 20 mais abundantes foi criada e seguidamente sujeita a reclassificação por meio da mais recente base de dados de referência taxonómica SILVA (versão 123, Julho 2015). Com esta abordagem, foi possível a reclassificação de todos os filótipos analisados em filós conhecidos. Nomeadamente, a segunda OTU mais abundante da base de dados, também a dominante no procarioma de *C. celata*, foi reclassificada como pertencendo a uma divisão parafilética da classe Clostridia, filo Firmicutes. Foi também reclassificado um filótipo como membro do filo candidato Nitrospinae.

Foi deste modo possível inferir que os presentes procariomas apresentam-se como específicos relativamente ao hospedeiro correspondente, independentemente do grau de parentesco filogenético entre hospedeiros. Tal implica que fatores ambientais ou características morfológicas, de estilo de vida ou metabólicas específicas de cada esponja desempenhem um papel decisivo relativamente à composição e estrutura destes procariomas. Assim, é possível que a natureza do substrato que sustenta a esponja, o seu estilo de vida (ereto ou incrustante, por exemplo) e mesmo diferentes ritmos metabólicos definam a composição dos microbiomas presentes. Uma análise da componente rara do procarioma presentemente analisado não revelou diversidade inesperada ou quaisquer padrões de especificidade ao nível de filo ao longo dos espécimes analisados. O isolamento e reclassificação das OTUs bacterianas não classificadas ao nível de filo resultou na identificação eficaz de diversidade previamente desconhecida.

Palavras-chave: esponjas marinhas, 16S ARNr, OTU, procarioma, singletonoma.

Contents

1. Introduction	
1.1. Aims.....	1
1.2. Importance	1
1.3. Morphological and phylogenetic diversity within Porifera	2
1.4. Sponge aquiferous system and sponge cell types	3
1.5. Ecology and lifestyles of poriferans	4
1.6. Sponge microbiology	5
1.6.1. Host-specificity of sponge microbiomes	9
1.6.2. Rare microbiota in sponges	11
1.7. Earth Microbiome Project.....	12
1.8. Targeted sponge species.....	13
2. Methodology	17
2.1. EMP dataset generation and processing.....	17
2.1.1. Sample collection	17
2.1.2. DNA extraction, sequencing and primary processing	17
2.1.3. Local dataset processing for Algarve samples.....	19
2.2. Ecological analysis	20
2.3. Phylogenetic cox1 analysis	23
3. Results	25
3.1. Ecological analysis of sponge prokaryomes	25
3.2. Exploration for prokaryotic ‘microbial dark matter’	43
3.3. Sponge host phylogeny	47
4. Discussion	49
4.1. Pre-processing methodologies	49
4.2. Richness and diversity of sponge prokaryomes.....	50

4.3.	Patterns of OTU presence across prokaryomes	54
4.4.	Taxonomy across prokaryomes.....	55
4.5.	Sponge- and species-specific associations.....	58
4.6.	Exploration for the ‘microbial dark matter’ within the sponge prokaryome ...	60
4.7.	Phylogenetic inference of abundant but unclassified OTUs.....	62
4.8.	Phylogenetic relationships among hosts and prokaryomes	64
6.	Bibliography	68

Annexes

I.	General R script for ecological and statistical analysis	I
II.	R script for generation of an averaged abundance matrix	II
III.	Ubuntu CLI script for Greengenes taxonomy extraction from EMP “.database” file	III
IV.	Mothur script for “singletome” dataset generation	IV
V.	DNA extraction and cox1 PCR figures	V
VI.	Non-normalized OTU and sequence numbers table	VI
VII.	Metadata table for “ALG” samples	VII

Figure Index

Figure 1.1 – General Leuconoid sponge morphology schematic picture (<i>in</i> Hentschel, 2012 ⁹)	4
Figure 1.2 – Worldwide distribution of phylum Porifera based on the World Porifera Database (http://www.marinespecies.org/porifera/).....	5
Figure 1.3 – <i>Cliona viridis</i> (A), <i>Cliona celata</i> (B), <i>Phorbas fictitius</i> (C) And <i>Dysidea fragilis</i> (D) pictured <i>in situ</i> , off the Algarve coast.. ..	14
Figure 1.4 – World distribution of the studied sponges, based on the World Porifera Database (http://www.marinespecies.org/porifera/).....	15
Figure 3.1 - Sequence number for all samples after singleton exclusion.....	25
Figure 3.2 - Boxplots depicting OTU richness, Shannon-Wiener diversity and Pielou's evenness indices averaged across all samples.....	26
Figure 3.3 - Rarefaction curves depicting cumulative (across all samples) richness for the normalized dataset.....	27
Figure 3.4 - Rarefaction curves depicting averaged (across all samples) richness for the normalized dataset.....	28
Figure 3.5 -Averaged observed and Chao1 estimated number of OTUs across all sponge species.	29
Figure 3.6 -Rarefaction curves depicting cumulative Chao1 estimated richness for the normalized dataset.....	30
Figure 3.7 - Clustered Bray-Curtis dissimilarity profiles for Algarve samples, coloured according to sponge species.	31
Figure 3.8 - non-metric Multidimensional scaling of Bray-Curtis dissimilarity profiles.	32
Figure 3.9 - Shepard's diagram for Bray-Curtis dissimilarity nMDS scaling, further showing the linear and non-metric fit for ordination distances.....	33
Figure 3.10 - Phylum-level relative abundance barchart for sequence libraries normalized by size (11.203 sequence reads per sample, singletons excluded).....	34
Figure 3.11 - Class-level relative abundance barchart for sequence libraries normalized by size (11.203 sequence reads per sample, singletons excluded).	37
Figure 3.12 - Order-level relative abundance barchart for sequence libraries normalized by size (11.203 sequence reads per sample, singletons excluded).....	38

Figure 3.13 - General profiles of OTU abundance set in ascending order across all samples in a fourth-root-transformed normalized dataset.....	39
Figure 3.14 – Heatmap showing the distributions of the 50 most abundant OTUs across all samples.....	40
Figure 3.15 – Venn diagram depicting the cumulative total number of shared and specific OTUs regarding each sponge species.	41
Figure 3.16 - Venn diagram depicting the cumulative number of shared and specific OTUs with a minimum of 50 assigned sequences regarding each sponge species.	42
Figure 3.17 - Venn diagram depicting the cumulative number of shared and specific OTUs with a minimum of 100 assigned sequences regarding each sponge species.	43
Figure 3.18 - Phylogenetic tree depicting the twenty most abundant OTUs unclassified at the phylum-level across the dataset, re-aligned using the SILVA 123 "Non-Redundant" (NR) SSU Reference dataset (July 2015).....	44
Figure 3.19 - Phylum-level barchart depicting singleton-specific taxonomic assignments across the non-normalized dataset.	46
Figure 3.20 - Unrooted phylogenetic tree depicting cox1-nased relationships among the sampled sponges.....	47

Acronym list

- OTU: operational taxonomic unit
- bp: base pair
- rRNA: ribosomal ribonucleic acid
- DNA: deoxyribonucleic acid
- DOM: dissolved organic matter
- POM: particulate organic matter
- WoRMS: World Register of Marine Species
- AR: ankyrin-repeat proteins
- TPR: tetratricopeptide-repeat proteins
- ELP: eukaryotic-like proteins
- HMA: high microbial abundance
- LMA: low microbial abundance
- SCC: sponge-specific cluster
- EMP: Earth Microbiome Project
- PCR: polymerase chain reaction
- NGS: next-generation sequencing
- RDP: Ribosomal Database Project
- CLI: command-line interface
- ANOVA: analysis of variance
- MRPP: multiple response permutation procedure
- nMDS: non-metric multidimensional scaling
- QIIME: Quantitative Insights Into Microbial Ecology
- COI/cox1: cytochrome oxidase subunit 1
- MB-CD: mean between-cluster dissimilarity
- W-CD: within-cluster dissimilarity
- ANOSIM: analysis of similarities
- NR: non-redundant
- SSU: small subunit
- ARISA: Automated Ribosomal Intergenic Spacer Analysis

1. Introduction

Poriferans (members of the ‘Porifera’ phylum, from the Latin ‘*porus*’, pore, and ‘*ferre*’, “to carry, to bear”) are the most ancient multicellular metazoans. They are therefore a pivotal group in the study of the metazoan transition from unicellularity to multicellularity. This very successful lifestyle encompasses a simple body plan, highly totipotent mobile cells, a characteristic aquiferous system and flexible reproduction strategies. Sponges are and have putatively been since pre-Cambrian times prevalent in oceans worldwide, ranging from coastal shores to abyssal depths and from tropical to polar settings¹. Although recent literature points towards there being little evidence for the veracity currently known “sponge” fossils, molecular clocks calculated for the phylogenetic divergence of poriferans and eumetazoans still support a pre-Cambrian origin^{2,3}. Apparently simple in its morphology and physiological traits, this phylum is regarded as an important model for animal phylogenetic, neuronal and immunological evolution, given its early-branching position in the metazoan tree of life.

1.1. Aims

This study aimed to characterize the prokaryotic communities of four species of marine sponge found off the coast of Algarve, southern Portugal. A dataset generated by means of high-throughput sequencing of 16S rRNA genes amplified from marine sponge total community DNA was thoroughly analysed using state-of-the-art bioinformatics and statistical tools.

The hypothesis that the taxonomic composition and structure of prokaryotic symbiont communities correlates with phylogenetic relatedness of their corresponding hosts was addressed.

Further, the diversity and magnitude of the “microbial dark matter” (that is, the pool of low-abundance prokaryotes) present in marine sponges was determined.

1.2. Importance

The marine sponge holobiont, that is, the ensemble of microbial communities inhabiting a marine sponge and their corresponding symbiosis, has now been researched for more than four decades. Recent advances in molecular biology techniques, namely next-generation sequencing (NGS), have allowed for a deeper understanding of the

structuring and composition of microbial communities within sponges. Besides the remarkable advances made in microbial ecology in general, studies of the marine sponge microbiome have been relevant from a biotechnological standpoint. It was long thought that sponges were the main producers of molecules of biotechnological interest. It is now widely accepted that microbes populating poriferans are the main producers of such compounds ⁴.

As such, it is of great interest to direct efforts towards a complete characterization of sponge “prokaryomes”: defined here as the ensemble of all bacterial and archaeal microbes present within a given setting. Previous explorations of poriferan microbiomes have consistently delivered novelties of biotechnological and ecological interest in the last two decades. For instance, new compounds ranging a plethora of bioactivities to previously unknown bacterial phyla have been found.

This study served as an exploratory analysis of the taxonomic and compositional characterization of the prokaryomes within four understudied marine sponges. Importantly, the present work’s data was the first high-throughput sequencing dataset to become available in the scope of molecular ecology of sponge prokaryomes at a worldwide scale. Given the extent of its size and depth, this dataset bore the potential to evaluate with higher confidence the ‘microbial dark matter’ within these organisms. This component of the prokaryome has recently been given greater attention because of its putative pivotal role in mediating important biochemical processes.

1.3. Morphological and phylogenetic diversity within Porifera

Accounting for all living sponges, it is estimated that up to 81% are taxonomically affiliated to the Demospongiae class ⁵. This class encompasses sponges whose skeleton, if present, is made up of siliceous or fibrous spicules, or both. When siliceous, spicules are either simple, or monaxonic, or present four axes with a common origin (tetraxonic). Further, the core of the aforementioned spicules, colloquially named axial filaments, are enclosed either in a triangular or in a hexagonal cavity ⁶.

Not all sponges possess a silica-based structural organization. In fact, some do not present any extracellular structure to provide a ‘skeleton’ of some sort. These sponges are part of a recently described class called Homoscleromorpha ⁵. This morphological characteristic has been the historical basis for sponge taxonomy and its reliance was

recently supported by molecular evidence ⁵. The remaining classes composing the phylum Porifera are Calcarea, made up of sponges bearing calcium carbonate spicules, Hexactinellida, generally known as ‘glass sponges’, which encompass siliceous spicules, and Demospongiae, in which spongin fibres, a kind of fibrous spicule, generally act as the main structural component.

Demosponges, the common denomination for members of the Demospongiae class, present a tremendous panoply of shapes, colours and sizes. Generally, sponge morphologies may be divided into massive (erect and large in size), encrusting (growing on top of a surface of either biological or geological origin) or excavating (by active erosion of the substrate on which the sponge lies).

1.4. Sponge aquiferous system and sponge cell types

Most marine sponges are benthic filter-feeders that rely on an aquiferous system in order to filter planktonic microorganisms ⁷. The external surface, or ectosome, as well as the outer lining of inner channels of poriferans are in general composed of a monolayer of pinacocytes, flattened cells which functionally act as ‘epithelial’. Along the ectosome, channels made up of one or more specialized cells, namely ostia or dermal pores, which allow for the entrance of water (therefore, incurrent channels) into the sponge endosome, that is, all that is surrounded by the ectosome. Seawater is further propelled through the sponge by a simple but effective mechanism ⁸. The pumping action of choanocytes, flagellated cells responsible for capture and intake of suspended particles, within choanocyte chambers, causes seawater to be driven further into the sponge by means of a continuous and uncoordinated beating of the flagella. A flow is then set from incurrent channels, connecting the external environment to the choanocyte chambers, leading to a central atrium and back towards the water column through the osculum and excurrent channels (see **Figure 1.1**) ^{9,10}. The aquiferous system in sponges is seen as early evidence for a metazoan circulatory system ¹¹. It allows, for example, effective gaseous exchanges, excretion of toxic metabolic end-products as ammonia, as well as easy take-up of dissolved and particulate organic matter (DOM, POM), the latter the subject of most attention since it leads to the retrieval of planktonic microorganisms from seawater.

The inner matrix encompassing the sponge ‘tissue’ in between the ectosome and all its inner surfaces is generally named mesohyl, or mesoglea (see **Figure 1.1**). This complex

framework is usually made up of specialized cells and a secreted skeleton of silicate or calcium carbonate spicules or collagen-based fibres, being that exceptions may occur in some sponges. Secretion of the aforementioned structural elements of the sponge occurs by action of sclerocytes, a cell type which biomineralizes the poriferan skeleton.

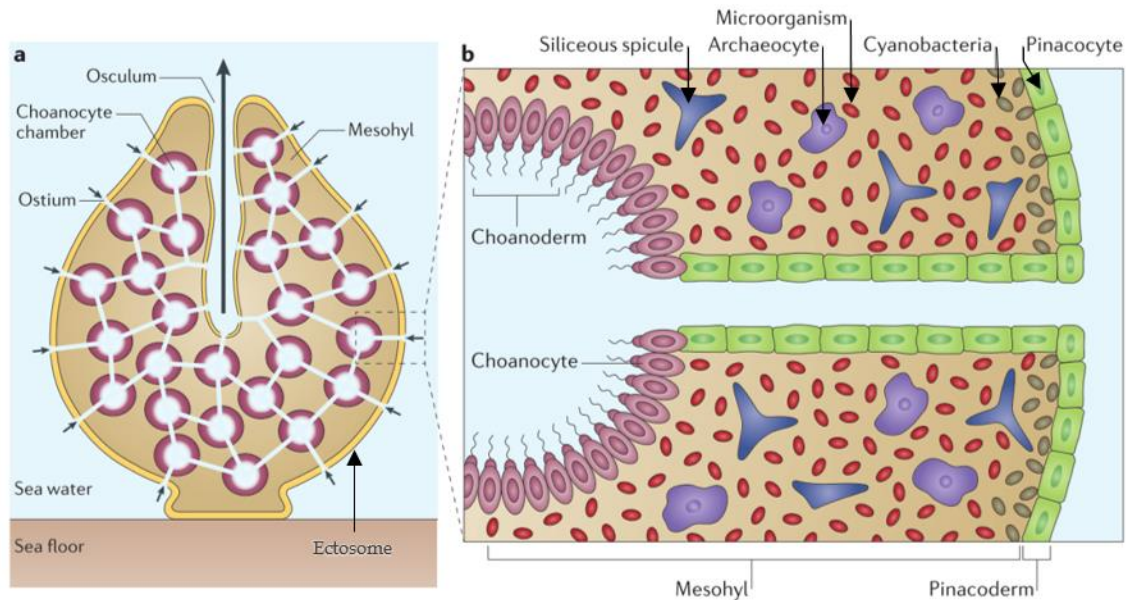


Figure 1.1 – General leuconoid sponge morphology schematic picture (adapted from Hentschel, 2012 ¹⁰)

Within the mesohyl, archaeocytes (see **Figure 1.1**) are characterized by being highly totipotent ameboid cells which proliferate, carrying out tasks which range from food digestion and transport to asexual reproduction. Archaeocytes are key to food digestion, as these cells are able to digest phagocytised material captured by choanocytes, as well as capture food particles by phagocytosis directly through inner walls of water channels.

1.5. Ecology and lifestyles of poriferans

Being extremely diverse in form, shape and colour, the morphology of sponges is determined by genotype and environmental conditions. Regarding the latter, sponges are mostly affected by underwater current velocity as well as turbidity, this is, the level of suspended planktonic detritus ¹².

Poriferans are widespread across all latitudes and longitudes, occurring mostly in seawater, but also in freshwater and in brackish environments (see **Figure 1.2**).

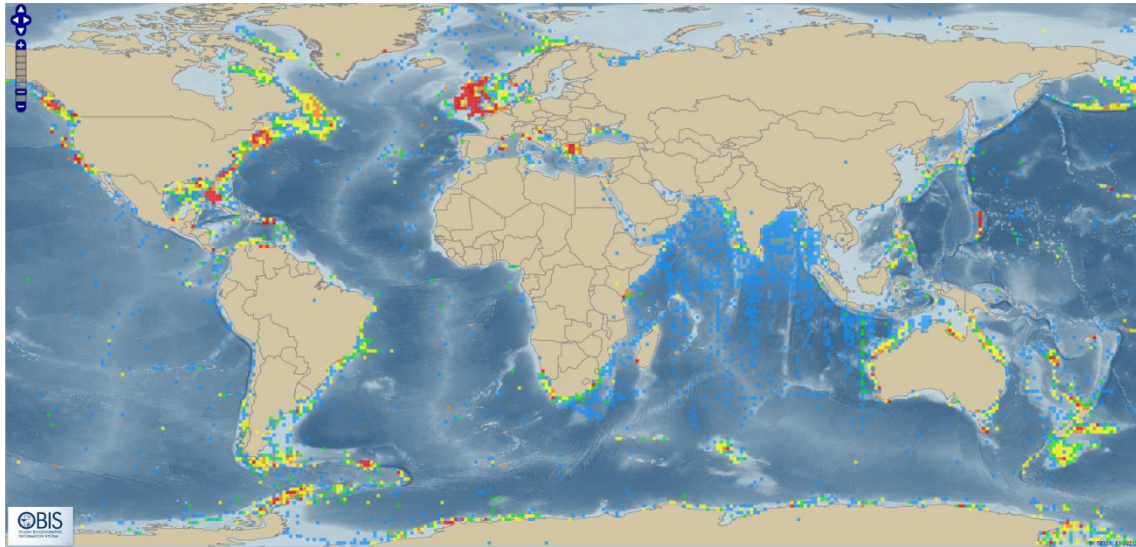


Figure 1.2 – Worldwide distribution of the phylum Porifera based on the World Porifera Database (<http://www.marinespecies.org/porifera/>). Warm colours depict high abundances whereas cold colours represent low levels of abundance. Accessed on the 28th of July 2015.

Freshwater sponges, although restricted to a single family formerly known as *Spongillina*, currently recognized as *Spongillida* Manconi & Pronzato, prevail across inland bodies of water across the world, ranging from Northern America to the Netherlands, India and China (WoRMS, 2015)⁵. Brackish sponges are also prevalent across the world, but remain understudied. The known studied exemplars are still poorly understood at a phylogenetic level, such as the *Malawispongiidae* and *Metschnikowiidae* families⁵.

Across oceans worldwide, demosponges populate several ecological niches, ranging from the intertidal regions to abyssal planes and marine caves^{13,14}. In general, sponges are sessile, benthic filter-feeders, but exceptions occur, as is the case of carnivorous sponges, which are able to predate selectively by means of a “sit-and-wait” strategy in generally oligotrophic and isolated settings such as marine caves¹⁵. These use specialized microscleres, or small spicules, which have evolved to a hook-like shape, for directed predation of microinvertebrates.

1.6. Sponge microbiology

It was in 1977 that Vacelet and Donadey published that demosponges were often densely populated by bacteria¹⁶. At the time, ‘tissue density’ of the hosts were set as defining factors of high and low observable microbial abundance. Although no

statistical evidence was provided, a linear relationship (e.g., high microbial abundance correlated to high sponge tissue density) was predicted. Further, it was noted that most of the microbes were present in the extracellular mesohyl matrix of the host animals.

The advent of molecular techniques enabled an in-depth exploration of the underlying diversity and microbial abundance within the sponge host in subsequent studies. By means of cloning-and-sequencing of bacterial 16S rRNA gene fragments amplified from sponge communal DNA, Hentschel and colleagues (2002) investigated marine sponges from the Mediterranean region, Japan and the Palau Republic ¹⁷. They found evidence for a core of shared prokaryotic species among poriferans that was not correlated with host biogeography. These so-called “sponge-specific” groups of OTUs were recurrently found in the following years, and the uniqueness of the sponge microbiome was further recognized, usually through comparative analyses with the surrounding seawater microbial communities ^{18,19}. The most iconic sponge-specific taxa described to date are the Poribacteria, candidate bacterial phylum discovered by Hentschel and co-workers in demosponges ²⁰. This candidate phylum was found to be one of the most abundant across several sponge microbiomes, existing only scarcely in seawater ²¹. Further, the lifestyle of these vertically transmitted symbionts has now begun to be uncovered by single-cell genomics and other last generation molecular biology techniques ²². Special attention has been given, for example, to the discovery of genome-encoded eukaryotic-like proteins which could mediate sponge-Poribacteria symbiosis ²³. Recent evidence provided by Taylor *et al.* in 2013 has shown, however, that sponge-specific taxa may be much more widespread throughout diverse marine environments than previously thought. This study analysed more than 12 million 16S- rRNA gene reads (from the V6 hypervariable region) generated by 454 pyrosequencing in 42 studies worldwide and found that although about 55% of the previously set sponge-specific taxa were found only in sponges, the remaining taxa were noted in other hosts and marine environments in relative abundances between 0 and 1% ¹⁰. Poribacteria was among these and was detected in abundances reaching 0.25% in coral hosts, comprising up to 0.19% of other samples, which included seawater sampled near hydrothermal vents, for example.

The previous argument was not a rejection of the sponge-specific taxa theory, but a call for further proof of “true” symbiosis, in the sense that more complete information towards the lifestyle of true sponge symbionts must be provided. Analysing microbial communities from a functional point of view has allowed for valuable insights

regarding the metabolic diversity in these, and has further unveiled how complementary host and microbial metabolisms may be. For example, in 2010, Thomas and colleagues used shotgun metagenome sequencing to approach the mechanisms of microbe-host functionality in the light of the holobiont theory of evolution ²⁴. Special relevance was given to ankyrin-repeat (AR) and tetratricopeptide-repeat proteins (TPR), whose expression was found to be significantly augmented in sponges when compared to the surrounding seawater. These eukaryotic-like proteins (ELPs) were known to be present in genomes of prokaryotic symbionts of eukaryotic hosts, and were thought to play critical roles in the establishment of symbioses. Nguyen *et al.* in 2013 gave further evidence towards processes of colonization of symbionts by escaping digestion by phagocytosis²⁵. Ankyrin-repeat proteins are thought to interfere with phagosome maturation by protein-protein interaction and therefore avoid digestion of bacterial cells by the sponge host, allowing for settlement of the first in the mesohyl of the latter. The finding of ELPs in symbiont genomes could indicate the mechanisms behind primordial settlement of microbes in the sponge mesohyl and symbiosis establishment ²². These putative indicators of prokaryote-eukaryote symbiosis have also been found in intracellular amoebal symbionts ²⁵.

Functional profiling of the sponge microbiome has until now contributed for an understanding of how symbiont prokaryotic species take advantage of sponge excretion products, but also produce secondary metabolites of interest for the sponge, which may, in turn, become a selecting factor for symbiotic settlement ¹⁰. Indeed, it has been found by means of genomic analysis that sponge-associated microorganisms retain genes encoding for several of the carbon and nitrogen metabolism pathways. The assimilation of carbon dioxide via reductive tricarboxylic acid cycle and a modified 3-hydroxypropionate cycle has been reported for bacterial and archaeal sponge symbionts, for example ^{23,26,27}. Further, regarding nitrogen metabolism, current genomic evidence shows a microbial metabolic ensemble directed towards the assimilation of ammonia, naturally excreted by sponges as a metabolic waste product, and nitrite ²²⁻²⁴. This occurs either in archaeal and bacterial symbionts, being that at least in cold water sponges, the first seem to dominate ammonia assimilation either total or partially (3 to 4 orders of magnitude), as communitarian ammonia-monooxygenase activity profiling showed ²⁸.

Aside from primary metabolism, important discoveries were made regarding secondary metabolism, which mainly encompassed vitamin biosynthesis ²⁷. Thought to

be intimately associated with microcompartment genesis in Poribacteria, biosynthetic pathways encoding for vitamin B1, B2, B6, B7 and B12 were found through genome mining of these symbionts ^{10,26}.

Although mostly distinct across geography and phylogeny, the sponge microbiome seems to generally converge regarding metabolism ^{29,30}. Indeed, it was shown by means of functional profiling of sponge prokaryotic communities that core metabolic tasks are present across a broad host phylogenetic range ²⁹. This could mean that core prokaryotic functions are transversal to diverse sponge hosts disregarding geography, phylogeny and even the taxonomic composition of their associated microbial communities.

Abundance patterns for microbial communities have been noted since the early works by Vacelet and Donadey ¹⁶. The terms ‘high’ and ‘low microbial abundance’ (HMA, LMA) are currently set based on electron microscopy, as they have been since early sponge microbiology studies ³¹, but flow cytometry and epifluorescence microscopy methods have recently enabled the exact quantification of living prokaryotic cells in solution ³¹⁻³³. To date, there is no evidence for a correlation between microbial abundance and phylogeny in sponges. Indeed, it is known that hosts belonging to the same sponge taxonomic order, or even family, may harbour microbiomes of contrasting abundances ³¹.

It is known that HMA sponges tend to harbour microbiomes in abundances surpassing those of seawater by a magnitude of 3 or 4, while LMA tend to reach similar values as to those of the latter environment ³². Further, LMA sponges appear to host less diverse microbiomes, which often dominated by Proteobacteria, Cyanobacteria and *Nitrospira* ^{34,35}. Molecular profiling of asympatric sponges led Giles *et al.* to assert that LMA microbiomes further lack typical core-HMA lineages, such as Poribacteria ³⁴.

Present-day sequencing technologies fail to discern true microbial abundance, due to known PCR and sequencing biases ³⁶. These techniques are further limited by the essential nature of each analysis, which will not give true cell abundance in a sample. For these reasons, robust efforts are still made towards cataloguing microbial abundance in sponges by means of electron microscopy ³¹. Hopefully, by relating these with high-throughput sequencing data, it will be possible to generate standard thresholds of microbial abundance which could serve as reference for future descriptions of new sponge microbiomes.

With the advent of sequencing technologies in the last decade and a half, knowledge of microbial diversity in marine sponges has exponentially increased¹⁶. By generating phylogenetic data from 16S-rRNA clone libraries, up to eight bacterial phyla were found to be present in the marine sponge microbiome^{17,18,20,34}. Further, in-depth analysis of some consistently recovered sequences, transversely present across sponge species and geography, showed how some archaeal and bacterial species could putatively form sponge-specific clusters (SSC) in a phylogenetic tree³⁷. The advent of pyrosequencing applied to microbiology allowed for deeper insights into the sponge microbiome and is, at the current state-of-the-art, the most utilized NGS in sponge molecular microbiology research. This technique usually unveils from 20 to 28 bacterial phyla as being in association with sponges, commonly obtained from the mesohyl^{9,13,38,39}.

Characteristically high abundances of proteobacterial phylotypes are frequently noted in sponges, and closer inspection of these assemblages will often reveal varying abundances of each proteobacterial class in different marine sponges. The diversity in proteobacterial lineages is true for either HMA or LMA sponges, although other bacterial lineages, such as Chloroflexi, Acidobacteria, *Deferribacteres* and Poribacteria, are generally absent or in minimal proportions in the latter poriferans (LMA)^{35,40,41}. These phyla have thus been selected as “indicator phyla” for HMA sponges, while LMA, as stated before, are frequently dominated by Proteobacteria and Cyanobacteria³⁵. Contrastingly, the frequently observed cyanobacterial sponge-specific species *Synechococcus spongiarum* is known to be specifically maintained in LMA sponges throughout their geographical range³⁴.

1.6.1. Host-specificity of sponge microbiomes

Ever since Hentschel and colleagues suspected the sponge-associated microbiome to be uniform across geographic boundaries, studies addressing patterns of host-specific microbiome assemblage in these animals have been conducted¹⁷. In fact, it was further noted in 2005 that *Chondrilla nucula* (Verongimorpha, Chondrillida), a demosponge sampled in The Netherlands, could harbour a sponge-specific and uniform microbiome, following the concepts found in the latter work from Hentschel and colleagues⁴². A total of 21 OTUs (e.g., groups of sequences found to share similarities of 97% or more and therefore treated as putative species. If not stated otherwise, all mentioned OTUs

are assumed to have been created by sets of sequences with the latter similarity values – 97%) were found to be absent from seawater and highly similar to known sponge-specific microbes found in other studies. This microbiome was therefore considered sponge-specific with regard to the uniformity in the presence of several prokaryotic phyla found by profiling marine sponges across the world. However, no direct comparison was made with other species within the same genus or relatives. It was suggested in 2013 by Blanquer *et al.* that similarities across microbiomes seemed to be more correlated to the assigned microbial abundance (HMA, LMA) than to host phylogeny⁴¹. The latter study further found that Chloroflexi seemed to be enriched in HMA sponges and that some α -Proteobacteria taxa were often present in LMA sponges.

Analysing 20 species of tropical sponges, Easson and colleagues found host-species specificity relating prokaryomes to sponge phylogeny⁴³. They found up to 30 OTUs defining dissimilarities across sponge species⁴³. Likewise, evidence of host species-specificity across geographical borders and among species within a same genus has been given by Erwin and colleagues^{44,45}. Either sympatrically or across distances up to 800km across the north-west Mediterranean Sea, these sponges maintained stable and defined prokaryomes, whose structure was highly correlated to host phylogeny.

Sampling at a depth gradient and along the eastern and north Atlantic, Reveillaud *et al.* did not notice host species-specificity patterns among the microbiomes of seven species of the *Hexadella* (Verongimorpha, Verongida) genus and five exemplars of the *Mycale* (Heteroscleromorpha, Poecilosclerida) genus⁴⁶. *Nitrospira* (*Nitrospirae*) was observed to be one of the most abundant OTUs in the dataset, being present in small amounts in sponges of the *Mycale* genus and almost absent in seawater. Analysed among other two sponges of the Poecilosclerida order, an exemplar of the *Halichondria* genus (Heteroscleromorpha, Suberitida), a LMA sponge, was shown to be dominated by several α -proteobacterial OTUs³⁸. Moitinho-Silva *et al.* sampled two sponges each belonging to a microbial abundance category (i.e., LMA and HMA) in the Red Sea³⁵. Here, it was posed, in accordance with previous studies from the same research group, that contrary to prevailing ideas, LMA sponges seemed to harbour higher abundances of Proteobacteria, Cyanobacteria and *Nitrospirae*³⁴. Further, in sponges displaying high microbial abundance, Chloroflexi, Poribacteria and *Deferribacteres* were found to be the dominating phyla. Controversially, this study presented further evidence that suggested the specificity of the sponge microbiota was defined by the host's microbial

abundance. However, several studies in irciniids (this is, marine sponges affiliated to family Irciniidae, order Dactyoceratida) point towards a phylogeny-dependent microbial community structuring^{45,47}, and indeed, the inner morphology of both LMA and HMA sponge categories has already been shown to be significantly different⁴⁸. With regard to morphological dependency, Weisz *et al.* showed not only that HMA and LMA sponges differ in mesohyl densities, but also in retention time of pumped water volume. As such, the first were shown to hold a denser mesohyl and therefore a lower water pumping rate than the latter sponges. Microbial abundance also seemed to be the defining factor with regards to the functionality of the microbiome when these were analysed by means of a Geochip 4.2 gene array, a high-throughput functional gene array, which enables the hybridization of prokaryotic coding sequences to pre-set light-emitting probes⁴⁹.

As such, recent evidence seems to point towards host-specificity being less affected by sponge phylogeny but by LMA or HMA status. Further knowledge of the sponge microbiome across geography and sponge phylogeny is necessary before firm conclusions can be drawn.

1.6.2. Rare microbiota in sponges

It was posed in 2009 that the rare biosphere within seawater could be a stable source of sponge symbionts⁵⁰. The advent of “ultra-high-throughput” sequencing technologies such as Illumina HiSeq and MiSeq provided further possibilities regarding the depth at which investigation of sponge microbiomes was possible³⁶. With this, it has been possible to uncover more of the diversity within the rare fraction of the sponge microbiome, and to infer its possible functional roles^{35,51}. Giles *et al.* noted how up to 81% of the uncovered phylotypes of five LMA sponges originating from the Caribbean, Red Sea and South Pacific were indeed singleton (occurring only once across a given dataset) OTUs³⁴.

In 2014, Reveillaud and colleagues provided the first evidence towards host-specificity regarding ‘microbial dark matter’ within the sponge microbiome⁴⁶. Examining into the OTUs present in relative abundances below 0.01% and 0.001% (within a total of 100,000 sequence reads per sample), it was found that indeed their presence, as determined by accumulation of phylotypes of a corresponding phylum, was selective regarding the correspondent sponge host, therefore indicating host-specificity.

Further, 33 singletons were uncovered in the Illumina HiSeq-generated dataset, as well as 18 doubletons (OTUs containing only two sequence reads across the entire dataset).

The rare microbiota within sponges has therefore far been under-characterized, but evidence points towards the possibility that with increasing sequencing capabilities, it will be possible to further analyse the diversity and functionality of these so-far elusive but seemingly essential consortia. Currently, it is expected that studies utilizing NGS technologies will be able to uncover highly diverse ‘singletomes’, that is, singleton OTUs consortia, as well as other components of the microbial rare biosphere. As suggested by Reveillaud *et al.*, it could happen that the consortia of low-abundance OTUs in sponge microbiomes are host-specific and may provide important functional input for the holobiont ⁴⁶.

1.7. Earth Microbiome Project

The Earth Microbiome Project (www.earthmicrobiome.org³⁶) is an international network that aims to catalogue the biodiversity of microorganisms across several biomes on Earth. This project set out to sequence extensively the microbial communities from a multitude of environmental sources worldwide, in a collaborative sampling effort. So far, it has analysed 200,000 samples, being that approximately 30% of those are sponge-related ⁵².

Within the framework of the Earth Microbiome project, the Consortium for Sponge Microbiology organized a standardized sampling of marine sponges, which took place in every continent worldwide, further collecting sponge-associated marine sediments and seawater, as well as detailed metadata. The contribution of this international project to the sponge microbiology research community is yet to be quantified, since only one scientific article has so far been published ⁴³. It can, however, be expected to propel the development of state-of-the-art approaches towards core questions in this research area, and help set a threshold for low and high microbial abundance or the determination of true sponge-specific microbes worldwide.

This work was developed within the scope of the Earth Microbiome Project and is based on part of the first worldwide “ultra-high-throughput” 16S rRNA gene amplicon dataset generated from poriferans. Here, a set of 26 samples comprising two clonaidans (*Clionaida sensu* Morrow & Cárdenas, 2015 ⁵), one dyctioceratid (order Dyctioceratida)

and one poecilosclerid (order Poecilosclerida) sponge, were characterized to retrieve insights into the host species specificity of each prokaryome and the relationship between host phylogeny and microbial community composition. Further, the taxonomic composition and magnitude of the ‘microbial dark matter’ encountered in the surveyed sponges was interrogated. For this, in-depth analysis was pursued by means of specialized 16S dataset analysis software packages such as mothur and QIIME (Quantitative Insights Into Microbial Ecology) ^{53,54}. Further, ecological and statistical analysis packages within the R software were utilized. By using mothur, an amplicon-directed bioinformatics pipeline, normalization, filtering and generation of abundance matrices for the data was possible, whilst statistical and ecological information was processed by R. QIIME produced relative abundance barcharts.

1.8. Targeted sponge species

In this study, given its sampling effort and experimental design, only the Demospongiae class was investigated. In general terms, demosponges (sponges of class Demospongiae) are characterized by their possession of an extracellular matrix mostly made of spongin fibres, with some species showing added and frequently dominating silicate-based spicules. Demosponges are found in either marine, brackish or freshwater environments, near-polar, tropical and temperate climates and at varying depths. Exceptions occur, as is the case for order *Dictyoceratida*, where spongin is dominant and siliceous spicules are absent ⁵⁵. Demosponges are able to sexually or asexually reproduce by means of viviparity or ovoviparity ⁵⁶. The morphological plasticity observed in demosponges has led to recent efforts towards a clarification of the phylogeny of the class. The removal of Homoscleromorpha from the class and its elevation to a new class within Porifera led to a conclusive monophyletic status of *Demospongiae* ⁵⁷.

Two of the marine sponge species belonging to the genus *Cliona* (Grant, 1826) studied here had their taxonomy recently altered. Specifically, the order Hadromerida to which they were assigned to, is no longer formally recognized. They are now placed within the Heteroscleromorpha subclass of the Demospongiae, and are part of the newly created Clionaida order (therefore members of this orders are commonly denominated clionaidans) ⁵. This group of marine sponges is known to have the ability to erode calcium carbonate, frequently harming corals and other organisms depending on such

solid exoskeletons. Shellfish cultures are known to take measures against invading clionaidans, for example ⁵⁸. One of these, *Cliona viridis* (**Figure 1.3a**), is a small known coral-boring sponge, although it occasionally may significantly grow in size, with a distribution restricted to the Mediterranean Sea, and Azores and Madeira archipelagos (**Figure 1.4**). Morphologically, it generally shows a cerebroid smooth surface, interrupted by large oscula, and is predominantly of brownish colour (**Figure 1.3a**). This species has been recognized as an LMA sponge, according to Blanquer *et al.*, who found that an OTU assigned to α -Proteobacteria accounted for more than half of the recovered sequences belonging to those samples. A similar result was found with *H. columella* (Poecilosclerida) ⁴¹.

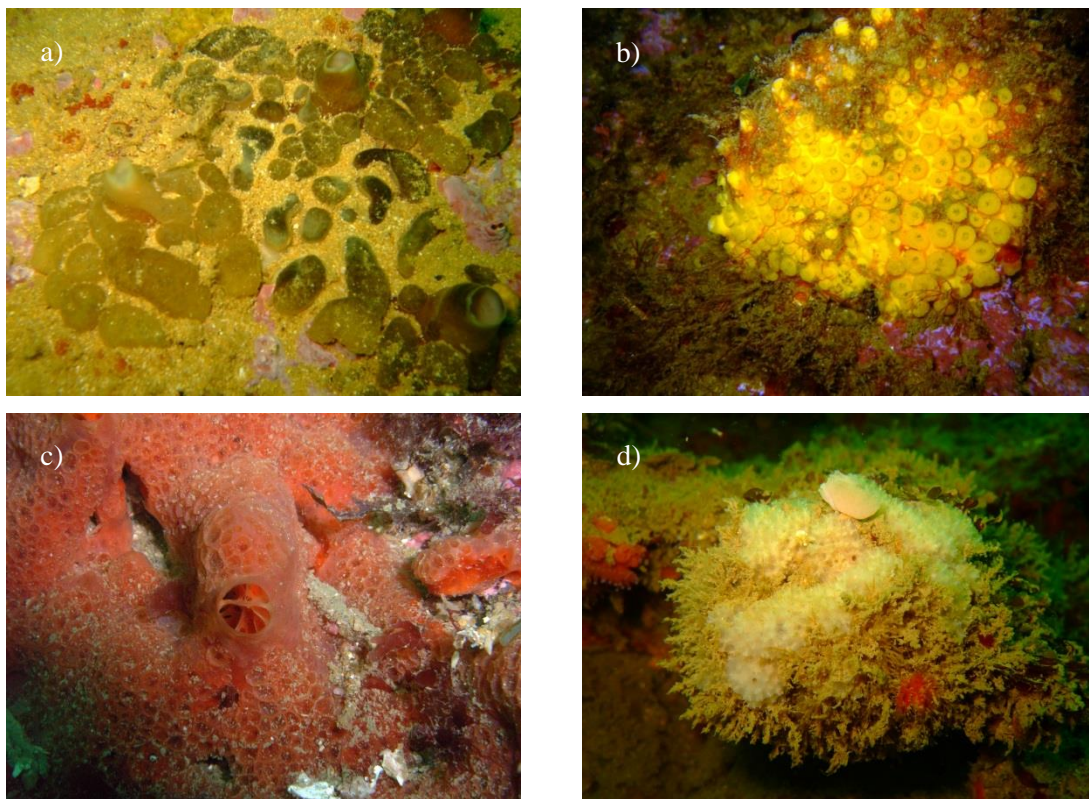


Figure 1.3 – *Cliona viridis* (a), *Cliona celata* (b), *Phorbas fictitius* (c) and *Dysidea fragilis* (d) pictured *in situ*, off the Algarve coast. Pictures by Francisco Pires, 2013.

Another member of this genus is *Cliona celata*, part of a cryptic species complex whose taxonomy is currently unsettled, reason why it is generally referred to as *Cliona celata* complex ⁵⁹. Since its first depiction in the 1900's, this has been one of the most studied sponge species, with its most defining characteristic being the “bright sulfur yellow” tone, as seen in **Figure 1.3b** ⁶⁰. Among the studied sponges, this species is,

according to the World Register of Marine Species, the one with the most extensive distribution. It is present in the Atlantic and Pacific Oceans, as well as in the Mediterranean Sea and in the South African coast. Evidence suggesting HMA in *C. celata* was found in 2015 when exemplars collected in Korea revealed relatively high diversity, when compared to prokaryomes of known LMA sponges ⁶¹.

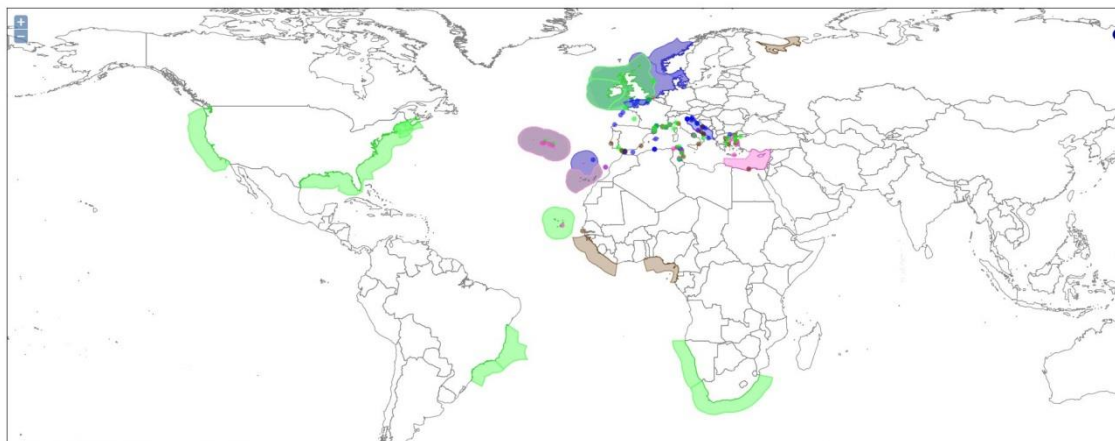


Figure 1.4 – World distribution of the studied sponges, based on the World Porifera Database (<http://www.marinespecies.org/porifera/>). Blue, green, pink and brown areas depict *P. fictitius*, *C. celata*, *C. viridis* and *D. fragilis*, respectively. Compound image of all distributions produced by Dr. Bart Vanhoorne (WoRMS) Data Management Team. Accessed on the 17th of April 2015.

Phorbas fictitius is externally characterized by its bright red to pinkish colouration, predominant oscula across the ectosome, and can either develop massively or stay incrusting. This sponge species has been shown to be the most abundant along the coast of Algarve by Pires, 2012, being its distribution influenced the most by depth and spatial variability ⁶². It ranges from the mid-Atlantic Madeira archipelago and Mediterranean Sea to the North Atlantic, occurring also in the Norwegian coast and North Sea (**Figure 1.4**). The genus *Phorbas* has not yet been studied regarding its microbial abundance, although, two studies have so far characterized 2 and 7 exemplars from the same order (Poecilosclerida) as LMA ^{31,63}. However, in 2012, Uriz *et al.* found sponges of the *Hemimycale* genus (Poecilosclerida) to be highly populated by “Calcibacteria”, which encompassed about 60% of the sponge’s microbiota total weight ⁶⁴.

As is typical of members of the order Dyctioceratida, *Dysidea fragilis* possesses an intricate, concentrically defined spongin skeleton. Typically encrusting but often found in massive form, this sponge’s superficial morphology shows a complex rugged

network of interconnecting conules. The distribution of *D. fragilis* ranges from the Atlantic African and Portuguese coasts to the Mediterranean Sea and the White Sea, in the northern coast of Russia (see **Figure 1.4**). Although no molecular studies have specifically targeted this species, other exemplars of the same genus were studied. The Red Sea sponge *Dysidea avara* has been considered to be a LMA sponge as determined by electron microscopy³¹.

2. Methodology

2.1. EMP dataset generation and processing

2.1.1. Sample collection

Twenty-six sponge specimens were collected off the coast of Algarve, southern Portugal by means of SCUBA diving, in October 2012. All were photographed *in situ*, before being collected. Each specimen was sampled in a hermetic plastic bag with ambient seawater, placed into cooling boxes and transported to the laboratory within 2 hours, where they were immediately processed. In the laboratory, all samples were photographed before being thoroughly rinsed in sterile calcium/magnesium free artificial seawater (ASW). They were further examined for the presence of obvious epibionts, which if present were removed. Sponge specimens were then cut into 0.25g pieces of sponge endosome using sterilized scalpel and tweezers. Backup pieces of sponge were kept at 96% ethanol for further taxonomical identification.

The specimens were then labelled with an “ALG” prefix indicating sampling location (Algarve), followed by “12”, indicating the sampling year (2012), and an individual code number. All samples were frozen at -80°C and sent on dry ice to Würzburg University for DNA extraction. Metadata were recorded while sampling took place, according to EMP standard procedures (www.earthmicrobiome.org).

2.1.2. DNA extraction, sequencing and primary processing

In order to achieve standardization regarding this step, the Sponge Microbiology Consortium decided that all samples should be processed in three laboratory hubs, located in Europe, Australia and United States of America (USA), with the same methodology. This way, all extractions were carried out under the same protocol, therefore minimizing the user biases associated to handling by several users, for example.

Nucleic acid extraction was performed with the PowerLyzer PowerSoil DNA Isolation Kit” (MoBio Laboratories, Inc.), allowing for high throughput extraction of metagenomic material from elevated numbers of environmental samples at a same time (www.earthmicrobiome.org).

Ribosomal ribonucleic acid (rRNA) forms complex RNA-protein domains, forming ribosomes, which play vital roles in microbial translation. The 16S rRNA (consisting of

about 1500 nucleotides) is part of the small subunit (SSU) of the prokaryotic 70S ribosome and is known to be highly conserved, for which it has in the last decades been widely used as a barcode for bacteria and archaea. The 16S rRNA gene is divided into 9 hypervariable regions, which are utilized either individually or in pairs, *e.g.* V7-V8, according to primer design. In this case, the V4 region of the 16S rRNA gene was utilized after intense scrutiny regarding its phylogenetic resolution by Caporaso and colleagues ⁶⁵. For such, paired-end 16S rRNA gene communitarian DNA sequencing was performed on an Illumina HiSeq (Illumina, Inc.) platform at the EMP headquarters in the USA, by means of a specifically designed primer set ⁶⁶ utilizing standard Illumina barcodes ⁶⁵. These primers (515F/806R ⁶⁵) were built in order to amplify both bacterial and archaeal ribosomal 16S-V4 rRNA gene fragments from metagenomic DNA samples.

Illumina-generated libraries were then processed by means of *mothur* v1.31.2 ⁵⁴. Sequence reads were subjected to quality control, whereby those shorter than 100bp or containing homopolymers larger than 8bp were removed. Further, sequence limits of the SILVA reference alignment were trimmed based on the V4 region, in order to create a custom database generated only for this region of the 16S rRNA gene, creating a temporary general taxonomy file which allowed for pre-exclusion of non-assignable sequences. This trimmed SILVA alignment (version 115, June 2013) served as a template to align all sequences. Using *uchime* ⁶⁷, chimeric sequences were removed (chimera-checking). Any sequences aligning outside of the 16S-V4 region were then excluded. Pairwise distances between sequences were calculated using a 95% cut-off, in order to then cluster sequences into OTUs, with similarities higher or equal to 97% based on the ‘furthest neighbour’ method ⁶⁸. This method allowed for OTUs to be created with less computation requirements, given the pre-calculation of pairwise distances, further making sure that all sequences within a same OTU presented 97% or more similarity from each other.

An abundance matrix was then generated, containing the number of sequences (putative individuals) within an OTU (putative species) present in a given sample. Custom EMP scripts were utilized in order to eliminate OTUs containing only one sequence and samples with less than 500 generated sequence reads (poor sequencing output). This outputted a “.shared” file, containing all Earth Microbiome Project samples, which was distributed among all partners and further processed *in-house* as

one of the base-files for the present study. This file is a regular abundance matrix, which in its structure includes columns depicting OTU cut-off level, sample identification, the total number of OTUs for each sample, followed by the abundance in sequences for each OTU.

Using SILVA as the chosen ‘gold’ standard for alignment of all sequences, a taxonomical assignment database was therefore created, with regard to other online 16S taxonomy-dedicated databases available at the time. The representative sequences of each OTU were then randomly picked and, after the V4 region-specific trimming of each database, the taxonomical classifications of SILVA, Greengenes and RDP databases were imputed into a “.database” file^{69–71}. This gave the end-user the ability to choose accordingly to their preferred online 16S taxonomy database, as all three differ in several aspects.

2.1.3. Local dataset processing for Algarve samples

To filter the general EMP “.shared” file, a “.accnos” file was generated by imputing the identification codes of samples collected in Algarve and separating them by a paragraph to create a list of samples for mothur to filter. Using the “get.groups” command of mothur it was then possible to generate a new filtered “.shared” file, containing only the samples corresponding to this study (“ALG”) and the associated OTUs and sequences. Further, OTUs not part of this study were automatically removed by mothur. The file was further processed in order to filter out OTUs to which only one sequence was assigned to (singletons) by means of a workaround (see **Annex IV**), given that mothur is not equipped for singleton-directed filtering. Therefore, singletons were removed from the filtered “.shared” file (“filter.shared”) and the recovered OTUs were listed (“list.otulabels”), so that, coming back to the original file they could be removed, using “remove.otulabels” (see **Annex IV**). This new “.shared” file was saved for posterior analysis.

The total number of sequences within each sample was then calculated by the “count.groups” command. Analysing the latter output, it was possible to know the lowest number of sequences across all samples (11.203 sequences), by which all samples were normalized, using the “sub.sample” command and its ‘size’ option. This command created the first major data input file (designated **P1**), which would later be subjected to ecological and statistical analysis, in which all “ALG” samples were

discriminated with no pooling of replicates. This was achieved by means of the “merge.groups” command, which used a “.design” file, which listed of the samples in this study associating each with their common designation, in this case the corresponding sponge taxonomical identification (*Cliona viridis*, *Cliona celata*, *Phorbas fictitius*, *Dysidea fragilis*). The samples that originated from the same sponge species were then pooled together, allowing for further species-specific analysis in a dedicated cumulative (pooled) dataset (this file was designated **P2**).

The “.database” file was edited using a custom script for the Ubuntu command line interface (CLI), in order to isolate the Greengenes taxonomy (see **Annex III**). The “.database” file was necessary for several steps within the mothur pipeline. This allowed the conversion of the “.shared” and “.database” files to “.biom” format, recognized by QIIME. This allowed the usage of specific QIIME commands.

2.2. Ecological analysis

Outputting ecological information related to the present dataset consisted of three general pathways: R-based outputs and direct outputs from mothur and QIIME software packages.

A custom R script (see **Annex 1**) was written with the objective to import the abundance matrix corresponding to the normalized unmerged and merged datasets (respectively P1 and P2), and a non-normalized dataset, as well as the filtered EMP metadata. All abundance matrices were previously formatted by means of the Ubuntu CLI, in order to safely be subjected to the subsequent processing steps (removal of ‘numOtus’ and ‘label’ columns). By using the “*vegan*” R package, the abundance matrix was in first place transformed following the Hellinger method (“decostand”, in *vegan*), and then matched with the metadata file ⁷². The Hellinger transformation involves taking the square root of relative abundance values. Other methods are known to be highly biased by large amounts of zeros in a given matrix ⁷³. This allowed for correspondence between the two files; therefore between sponge specimens and sponge identification. Further, an averaged normalized abundance matrix was generated, by means of a custom R script and the Ubuntu CLI (see **Annex II**).

Species richness, diversity (Shannon-Weaver index) and evenness (Pielou’s index) were calculated using to *vegan*-based functions. While the first is a direct measurement of the number of species (OTUs) present in each sample, the second accounts for

proportional abundance of species. Pielou's evenness calculates the ratio for diversity against the log-transformed richness, which can be interpreted as a dominance measure as well. They were then plotted together and aggregated into a summary table, as described in **Annex I**. Further, using analysis of variance (ANOVA), statistical differences for each of the aforementioned indices were tested, after which pairwise statistical differences in the resulting means were calculated with a *post-hoc* Tukey's range test. This resulted in the generation of a p-value describing significant differences among each pair of sponge species. The Chao1 estimation of richness was also calculated using *vegan*, being plotted *a posteriori* with *ggplot2*, a specialized R package directed towards plotting, against the observed number of OTUs in the dataset (observed richness)^{74,75}. This statistical technique estimates richness by means of a correction factor which takes into account the number of singletons and doubletons present in a given abundance matrix. To visualize community patterns, dissimilarity among samples was calculated with the *vegan* R package based on the Bray-Curtis algorithm. The statistical differences described by this dissimilarity measure were further tested with ADONIS (Permutational Multivariate Analysis of Variance Using Distance Matrices) and MRPP (Multiple Response Permutation Procedure), which analyse variance within distance matrices by means of F-value calculations and permutations, respectively. Using the Bray-Curtis similarity (which calculates distances between two samples in a dataset by dividing the sum of the absolute difference of OTU abundances by the sum of total OTU abundances in the samples) matrix as a basis, samples were then clustered using the average linkage algorithm and visualized as a dendrogram. Using the resulting distance matrix, non-metric Multidimensional Scaling (nMDS) was computed using two dimensions. Visualization of this output was facilitated by the previously set linkage of the abundance matrix to metadata, hence allowing the identification of samples by sponge species, generation of sponge species-specific ellipses and layering of a cluster dendrogram over the nMDS. A Shepard's stress plot depicted the scattering of data points associated to ordination distances versus original dissimilarities. Also, a linear regression served as a goodness of fit estimator regarding the scattering of data points around the regression line. Two useful statistical measures for nonmetric fit (based on stress) and simple correlation of fitted values against ordination distances were outputted. Further, nMDS was tested statistically by ANOSIM (Analysis of Similarities) and ADONIS.

A general view of community structure was achieved with the “*heatmap*” R package, for which the original abundance matrix was fourth-root transformed for visualization purposes. This package creates highly customisable heatmaps, which can then be coloured with a custom palette of colour gradients generated by the “*RColorBrewer*” R package ^{76,77}. Additionally, ‘single linkage’ clustering of samples based on community dissimilarities was plotted along with the heatmap. This analysis was performed using *mothur*’s merged and unmerged outputs. With R, it was possible to order the datasets by sum of sequences in an OTU across all samples. Extracting the 50 most abundant OTUs in the dataset, it was possible to plot a new heatmap depicting their abundance patterns across all samples.

Rarefaction curves were built based on *mothur*’s outputs for rarefaction of pooled samples. Using the R package ‘*stringr*’, it was possible to properly read the previous file, and plotting of rarefaction curves was performed using a personalized colour palette (Lucas Moitinho-Silva, personal communication) ⁷⁸. Using *mothur*’s implemented function for Venn diagram construction (command “*venn*”), it was possible to generate outputs with different cut-off levels for OTU abundances, as set by the “*filter.shared*” command. This caused Venn diagrams to show proportions of shared and unique prokaryotic species along an abundance gradient, which eventually set aside the ‘rare’ prokaryome, that is, the least abundant phylotypes. To plot relative abundance bar charts, the custom-filtered Greengenes “.database” file was inputted into *mothur*’s “*make.biom*” command, along with the final merged and unmerged “.shared” files, so that “.biom” files, adapted to analysis within QIIME software could be generated. As such, the “*summarize_taxa.py*” and “*plot_taxa_summary.py*” scripts provided cumulative bar charts across taxonomical levels (phylum to genus) ⁷⁹.

Within *mothur*, the ‘singletome’ (the dataset encompassing all OTUs with only one assigned sequence – singletons - across all samples) was extracted from the non-normalized dataset to explore the total singleton richness along all sampled sponges. Barcharts depicting cumulative abundance of singletons across phyla were produced using QIIME. The 20 most abundant OTUs unclassified to phylum-level were extracted from the non-normalized dataset independently from the given Greengenes’ domain-level similarity score. Firstly, the dataset was sorted by descending order of number of sequences per OTU to extract the labels of the 20 most abundant OTUs. This list served as a template for filtering of the taxonomy database within the open Galaxy web-server

(usegalaxy.org). By means of the Ubuntu CLI, columns depicting OTU labels and the representative “.fasta” sequence for each OTU were further filtered and edited for input in the SILVA Database Web Aligner version 1.2.11 (www.arb-silva.de/aligner/)⁸⁰. Here, the extracted sequences were realigned according to the latest SILVA Reference Database (version 123, July 2015) and further inputted to ARB (www.arb-home.de/), which allowed the generation of a phylogenetic tree using RAxML (Randomized Axelerated Maximum Likelihood -VI-HPC - version 7, High Performance Computing), using the rapid-hill climbing mode, and a GAMMA model of site rate heterogeneity with ML estimation of the alpha-parameter.

Summary tables depicting averaged OTU and sequence numbers by phyla and proteobacterial classes for both normalized and non-normalized datasets were built by means of a custom R script. Lucas Moitinho-Silva (Ute Hentschel’s laboratory at the University of Wüezburg in Germany) greatly assisted to this component.

2.3. Phylogenetic analysis of *cox1*

Each sponge specimen was taxonomically classified after sequencing of their cytochrome C oxidase subunit 1 gene (*cox1*, or COI), a general barcode (this is, a gene common to a set of organisms, e.g. Porifera, used to distinguish between species and inferring on phylogeny in nucleotide composition; also called genetic marker) for Porifera. DNA extraction, amplification by polymerase chain reaction (PCR) and Sanger sequencing were used. Regarding the first step, an UltraClean® Soil DNA Isolation Kit was used according to manufacturer’s protocols. The extracted genomic material was then subjected to electrophoresis (0.8% agarose, 60V for 45min), with a ratio of 5:1 DNA and loading buffer (6x) (see **Annex V** for corresponding figures).

“Folmer” primers were utilized in this task, producing a 650 base pairs (bp) sequence⁸¹. PCR conditions were identical to those of Xavier *et al*⁸². A total mix volume of 25µL contained 1,5µL DNA template, 10x reaction buffer (Bioline, London, UK), 50mM MgCl₂, 10mg/mL bovine serum albumin (BSA), 2mM deoxynucleoside triphosphates (dNTPs), 10mM of each primer (LCO1490 and HCO2198, Bioline, London, UK), and 5 U/µL BioTaq™ polymerase (Bioline, London UK). An amplification profile of 95°C for 3min was followed by 36 cycles of 94°C for 30s, 51°C for 1min and 72°C for 90s, after which extension followed for 10min at 72°C in a MyCycler thermal cycler (Bio-Rad, Hercules, CA, USA).

Agarose gel (1%) electrophoresis (110V for 50min) was then performed to verify the yield and size of the resulting amplicons. To this end, 3 μ L of DNA was used for each 4 μ L of loading buffer. To control electrophoresis run 4 μ L of Fast Ruler™ DNA ladder (Thermo Fischer Scientific, Inc., UK) were utilized at the edges of the gel. Purification of the generated amplicons was achieved by means of Sephadex® G-50 (Sigma-Aldrich, Taufkirchen, Germany) cleaning. Purified cox1 amplicons were then subjected to sequencing with the chain termination method (“Sanger sequencing”) in an Applied Biosystems 3130 genetic analyser, using the forward primer, at the CCMAR’s Sequencing Facility (see **Annex V** for corresponding figures). Sequences were manually inspected with Sequence Scanner Software v1.0, in which electropherograms (plots of nucleotide-specific signal intensities directly obtained from Sanger sequencing capillary electrophoresis) are easily analysed against the automatic interpretation of each base pair (Applied Biosystems, Thermo Fisher Scientific).

MEGA 6 software⁸³ was used for the creation of alignments and comprehensive phylogenetic assessments of the retrieved cox1 sequences. First, MUSCLE, a multiple sequence alignment program was utilized for aligning sequences⁸⁴. This software runs in three phases in which a draft multiple alignment is created, after which Kimura distances (a measure of mutation distance, this is the probability of multiple mutations occurring at a single site) improve the initial tree. The final step involves a refinement of the first multiple alignment based on the improvement made by the second step. A MEGA6-specific tool for search of the best DNA models found the Tamura-3 parameter (a model set to differentially correct different types of transitions and transversions) with proportion of invariant sites (I) to be the best fitting model for the generated cox1 sequences. Further aesthetical refinement of the generated tree took place within MEGA6.

The Inkscape open-source graphics editor was used for editing of all R, mothur, QIIME, ARB and MEGA6 outputs.

3. Results

3.1. Ecological analysis of sponge prokaryomes

Directed filtering of 1,226 sponge samples and 45,981 OTUs detected in these samples, originated from the processing of Earth Microbiome Project's raw data, led to the generation of a sub-sampled dataset corresponding to the sampling event which took place in Algarve 2012 (26 sponge samples). The total number of sequences analysed per sample, after singleton exclusion (i.e. removal of all OTUs with only one assigned sequence), is depicted in **Figure 3.5**.

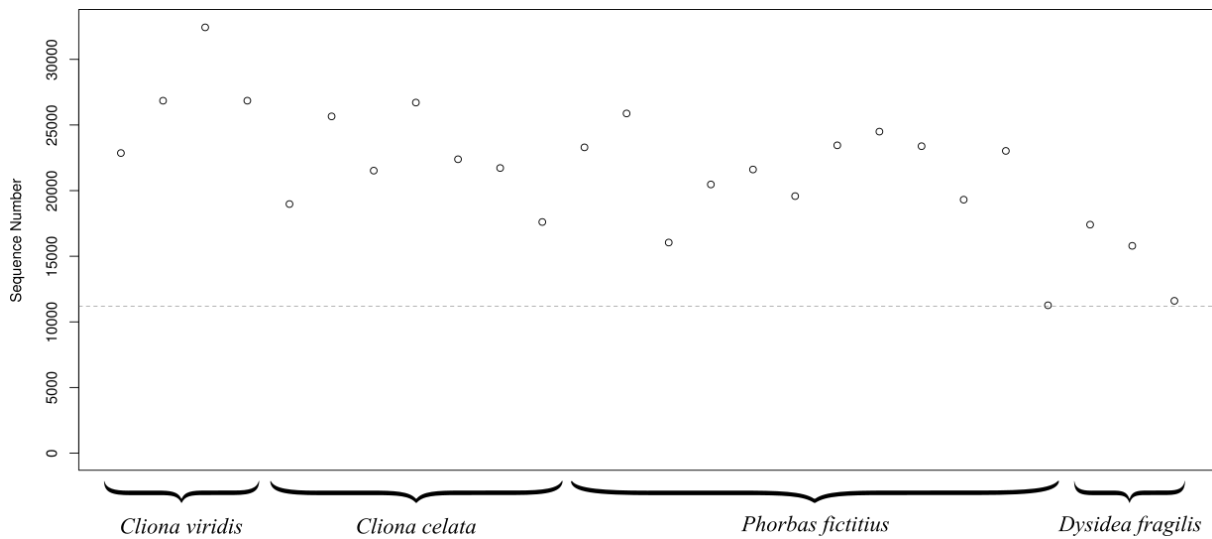


Figure 3.5 - Sequence number for all samples after singleton exclusion. The light grey line shows the normalization threshold value (11,203 sequences) used in downstream alpha and beta diversity analyses.

Normalization of the sequence effort (e.g., numbers of sequences obtained by sample) employed per sample was further achieved based on the lowest found sequence number among all samples (11,203 sequences). The resulting singleton-filtered normalized dataset comprised 26 samples and 3,214 OTUs in total. Using morphological criteria, 3 sponge samples were identified as *Dysidea fragilis*, 7 as *Cliona celata*, 4 as *Cliona viridis* and 12 as *Phorbas fictitius*.

Overall, the dataset presented medians which varied between 286 and 523 OTUs for each sponge species, as seen in **Figure 3.2**.

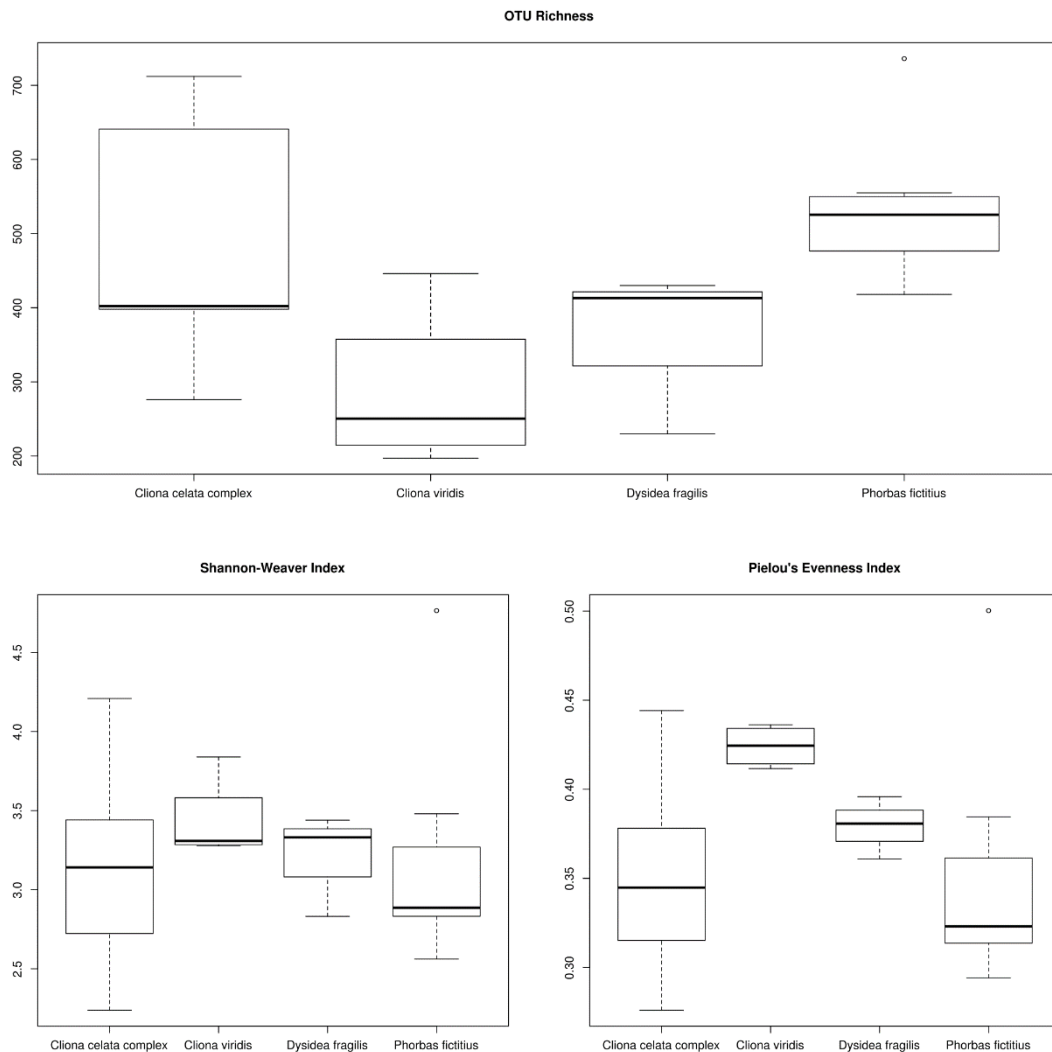


Figure 3.6 -Boxplots depicting OTU richness, Shannon-Wiener diversity and Pielou's evenness indices averaged across all samples. Upper and lower limits of solid line boxes depict upper and lower quartiles, whereas the solid bold line represents the median. Horizontal lines and points outside of boxes depict medians and outliers, respectively.

The Shannon-Wiener diversity (natural log base) was calculated to be between 3 and 3.5 across all sponge species, and no significant differences were found (ANOVA, $P=0.78$). While *P. fictitius* hosted the richest prokaryome, it presented the least diverse and even one. *Cliona viridis* displayed the highest Pielou's evenness values among all species followed by *Dysidea fragilis*. This suggests a larger co-dominance of the fewer members that constitute this prokaryome in comparison with the prokaryomes of the remainder sponges. The core of highly and equally abundant OTUs in *C. viridis* therefore seemed to be made up of more phylotypes than the other sponges, although not significantly (ANOVA, $P<0.1$). This sponge further showed about half the number

of prokaryotic species (richness) as the number found in *P. fictitius* (Tukey's Honest Significant Differences, $P < 0.01$).

Rarefaction analysis of the normalized dataset showed unique profiles for each of the prokaryomes in the study and evidence for a relatively well-sampled prokaryome albeit curve plateaus could not be fully achieved for any of the species (**Figure 3.3**).

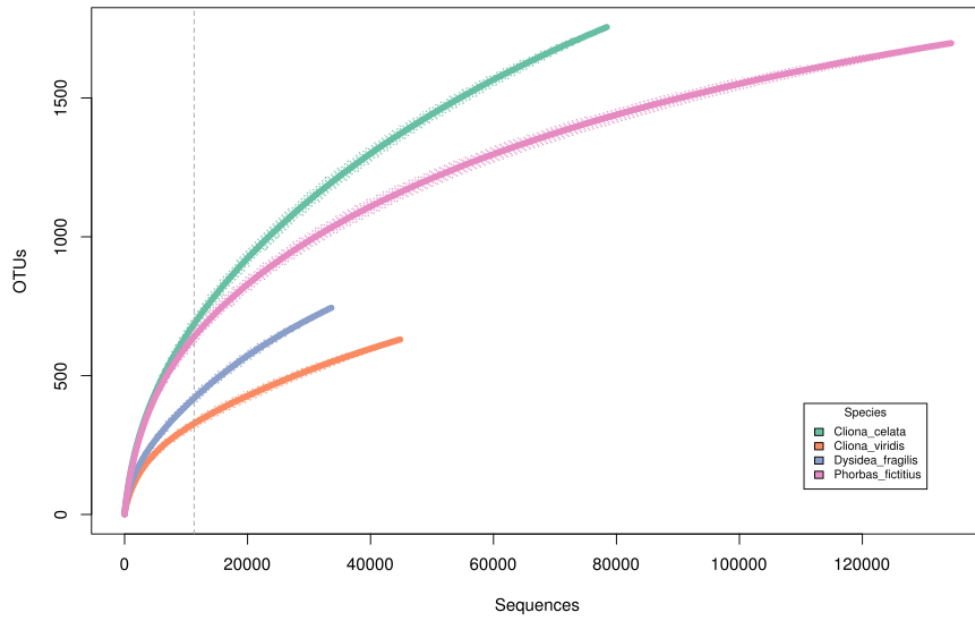


Figure 3.7 - Rarefaction curves depicting cumulative (across all samples) richness for the normalized dataset. Semi-transparent dotted shadows show high and low confidence intervals for each curve. Dashed grey line shows the sequence number threshold per sample (11.203 sequences).

The lowest set curve was observed for *Cliona viridis*. Indeed, this sponge species required the highest number of sampled sequence reads to reach an arbitrary threshold of 500 OTUs, useful for the comparison of all samples (**Figure 3.3**). An exponential growth trend was found for the curves depicting *Phorbas fictitius* and *Cliona celata* complex, both showing high richness values. Whereas the prokaryome within *C. celata* reached 1,500 OTUs after about 80,000 sequences had been analysed, the same happens in *P. fictitius* at around 100,000 sequences, showing greater levels of richness and diversity per sequence effort in the first sponge. At the normalization threshold, all the prokaryomes were found to most likely be undersampled.

By averaging rarefaction values by sponge species using the normalization threshold (11,203 sequence reads per sample), similar trends as the ones depicted in **Figure 3.3** could be observed (**Figure 3.4**).

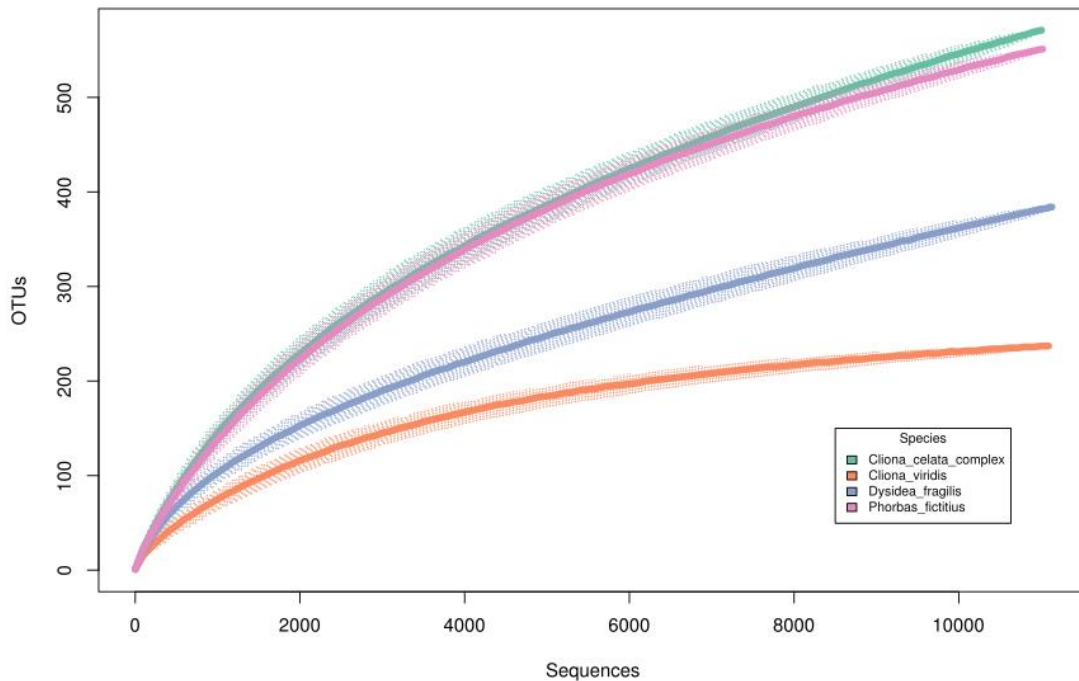


Figure 3.8 - Rarefaction curves depicting averaged (across all samples) richness for the normalized dataset. Semi-transparent dotted shadows show high and low confidence intervals for each curve.

However, a more adequate evaluation of whether individual samples within each species were satisfactorily covered by a sequence effort of 11,203 sequences can be achieved. Although smaller differences were noted between *C. celata* and *P. fictitius* for the same threshold of normalization (11,203), the previous relationship where the first surpasses the latter in magnitude of richness was maintained. These sponges were plotted still in an exponential growth phase, unlike *C. viridis*, which now tends to reach a plateau where it is relatively safe to state that further sampling (sequencing effort) should not reveal many new phylotypes. As *C. celata* and *P. fictitius*, the *D. fragilis* curve maintained a relatively steady increase as more sample reads were analysed, although less pronouncedly.

Observed and estimated richness values retrieved for all sponge hosts, depicted in **Figure 3.5**, indicated further richness was to be sampled. Both *C. celata* and *P. fictitius* were estimated to hold more than 750 OTUs, on average.

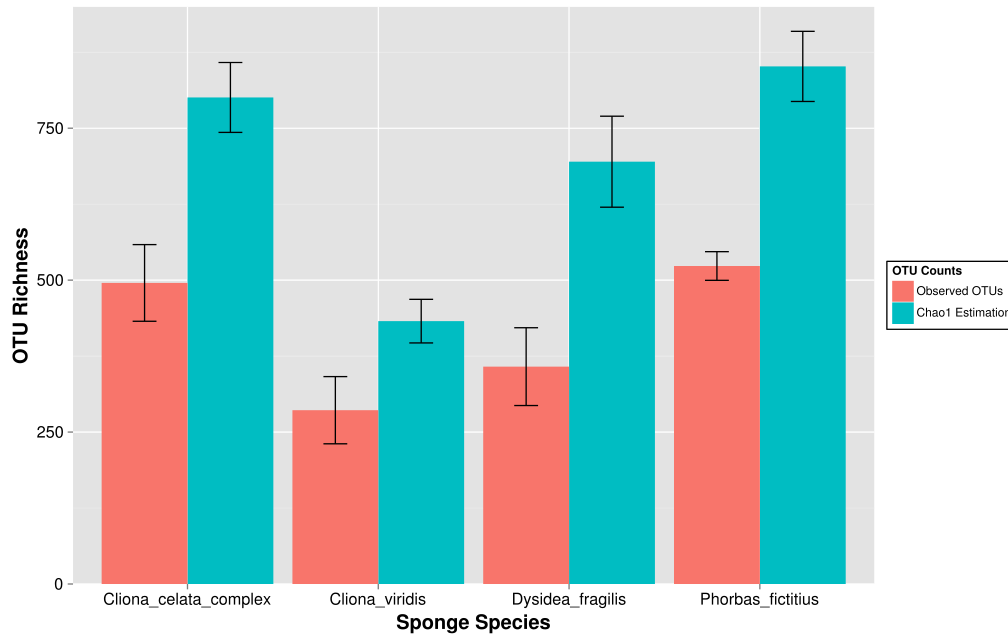


Figure 3.9 - Averaged observed and Chao1 estimated number of OTUs across all sponge species.

Moreover, the two sponges with the lowest number of phylotypes across the dataset – *D. fragilis* and *C. viridis* – were attributed strikingly different richness estimations. While the richness in *C. viridis* was not expected to rise above about 450 OTUs, *D. fragilis* was shown to potentially hold about 700 OTUs. The biggest difference regarding observed and estimated richness was noted in *P. fictitius*. Further, differences of observed richness values were significant across samples (ANOVA, $P < 0.01$).

Overall, Chao1 estimates showed how further richness remains to be sampled, supporting the information shown in **Figure 3.4**.

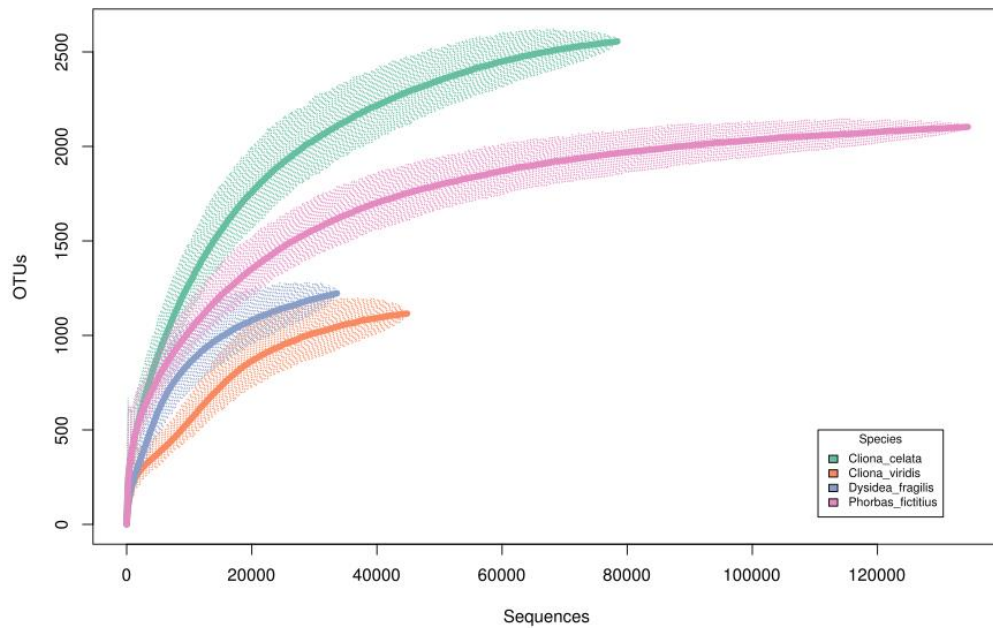


Figure 3.10 - Rarefaction curves depicting cumulative Chao1 estimated richness for the normalized dataset. Shaded polygons show high and low confidence intervals.

Rarefaction analysis using Chao1 diversity estimates suggests theoretically much richer prokaryomes (**Figure 3.6**) across all sponges than observed with the available data (**Figure 3.4**). Nevertheless, similar trends were maintained across the observed (**Figure 3.4**) and Chao1 (**Figure 3.6**) rarefactions, although *C. celata* and *P. fictitius* were seen to potentially hold much more differentiated richness patterns.

Four discernible groups of samples were plotted after Bray-Curtis dissimilarity was calculated for the normalized dataset and further clustered by Multi Response Permutation Procedure (MRPP, $P < 0.001$), each corresponding to a different sponge species. Therefore, given the dendrogram profiles (**Figure 3.7**), each sponge host harbours its own unique prokaryotic community, but sponge phylogenetic relatedness is not likely to be an influencing factor in the differentiation among prokaryomes, as *C. viridis* and *C. celata* are not clustered the closest.

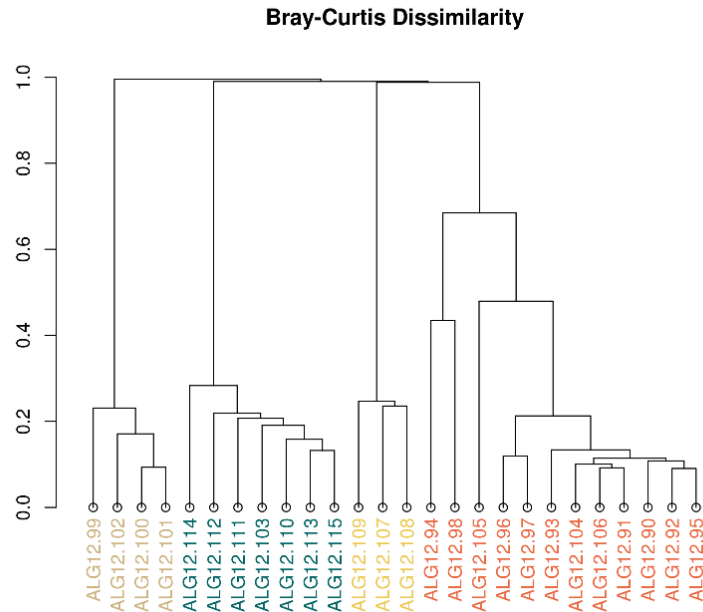


Figure 3.11 - Cluster analysis of sponge prokaryomes based on Bray-Curtis dissimilarity measures. In light brown, *C. viridis*, in green *C. celata*, in yellow *D. fragilis* and in red *P. fictitius*. The two most similar groups were made up of samples belonging to *P. fictitius* and *D. fragilis*, phylogenetically assigned to two different orders⁵. Very little similarity was seen for the congeneric species *C. viridis* and *C. celata*, as these clustered the furthest apart. Clear intra-group dissimilarity was shown for the samples belonging to *P. fictitius*, which does not relate to geographical origin, but may indicate intra-specific differentiation (see **Annex VII** for detailed geographical information).

It was therefore possible to discern that host-specificity was defined not by higher taxonomic ranks but by host species. Indeed, two sponges of the same genus (*Cliona*), possessed the most distant prokaryomes with a mean between-cluster dissimilarity (MB-CD) of 0.996, whereas *P. fictitius* and *D. fragilis* shared slightly lower MB-CD of 0.9879 (MRPP). *P. fictitius* presented the highest within-cluster dissimilarity (W-CD, 0.3633), compared to a general average of 0.2541.

Clear intra- and inter-differentiation of sponge prokaryomes was discernible with nMDS visualization of the Bray-Curtis dissimilarities plotted in **Figure 3.8**.

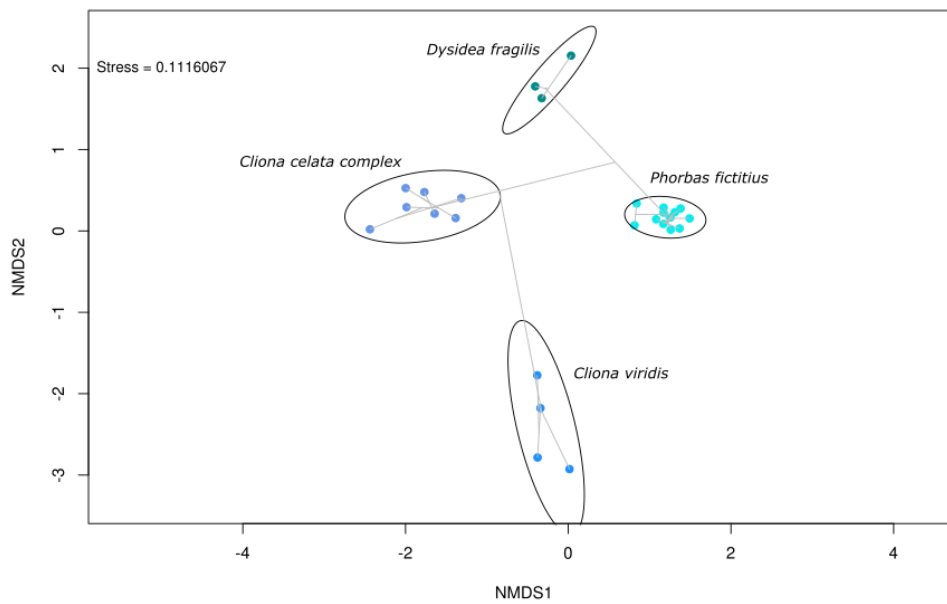


Figure 3.12 - non-metric Multidimensional Scaling of Bray-Curtis dissimilarity profiles. Ellipses around groups of samples represent a 95% confidence interval for clustering based on standard deviation.

Intra-group distances were the most noticeable in *C. viridis* and *C. celata*, in contrast with the dendrogram shown in **Figure 3.7** which highlights higher intra-group variation within *P. fictitius* specimens. Interesting patterns therefore emerge within the *P. fictitius* dataset, as three samples seem to outlie from the general core (see **Figure 3.7**). The limited number of samples belonging to *D. fragilis* restricts robust assessments of intra-variability within these prokaryomes. It was, however, possible to discern that they formed a well delineated group separated from all others.

The fit between the original Bray-Curtis dissimilarity values and the scaled distances used in nMDS ordination was pictured in **Figure 3.9**, showing strong correlation between both measures and indicating that dispersion was not seen for the analysed data points.

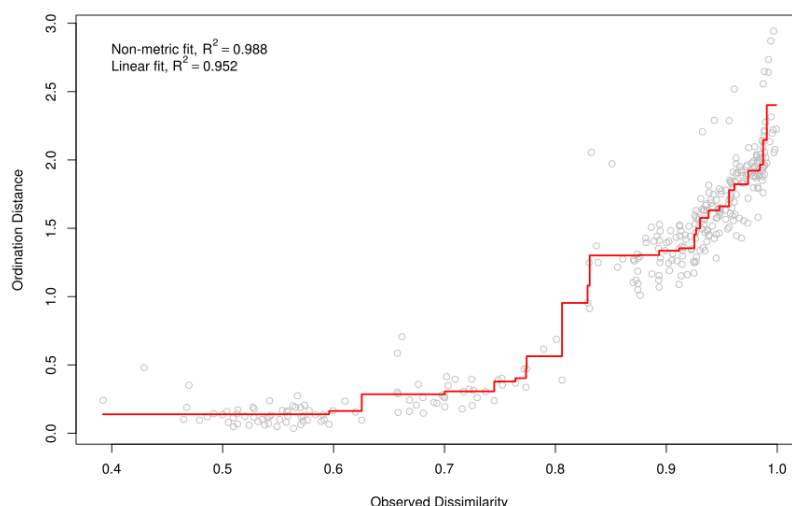


Figure 3.13 - Shepard's Diagram for Bray-Curtis dissimilarity nMDS scaling, further showing the linear and non-metric fit for ordination distances.

Overall, a large percentage of the Bray-Curtis dissimilarities (between 0.8 and 1.0 on the x-axis) was represented by growing ordination distances. Low values for scaling stress (0.1117) and linear fit (or the found correlation between dissimilarity values and ordination distances) show a well-constructed nMDS. Further, an R value of 1 for ANOSIM shows high statistical significance for the plotted groups of samples. Indeed, within-cluster distances are much shorter than between-cluster distances ($P < 0.001$). The result of the latter test, which aims towards discerning significant distances between two or more groups of samples was further confirmed by means of ADONIS. This test revealed highly significant values ($P < 0.001$) for the generated nMDS configuration, after 999 permutations, and that 85% of the variation observed in dissimilarity values occurs between (not within) clusters.

At the phylum level, contrasting prokaryomes were found for all sponge species, with an overall dominance of proteobacterial phylotypes (**Figure 3.10**).

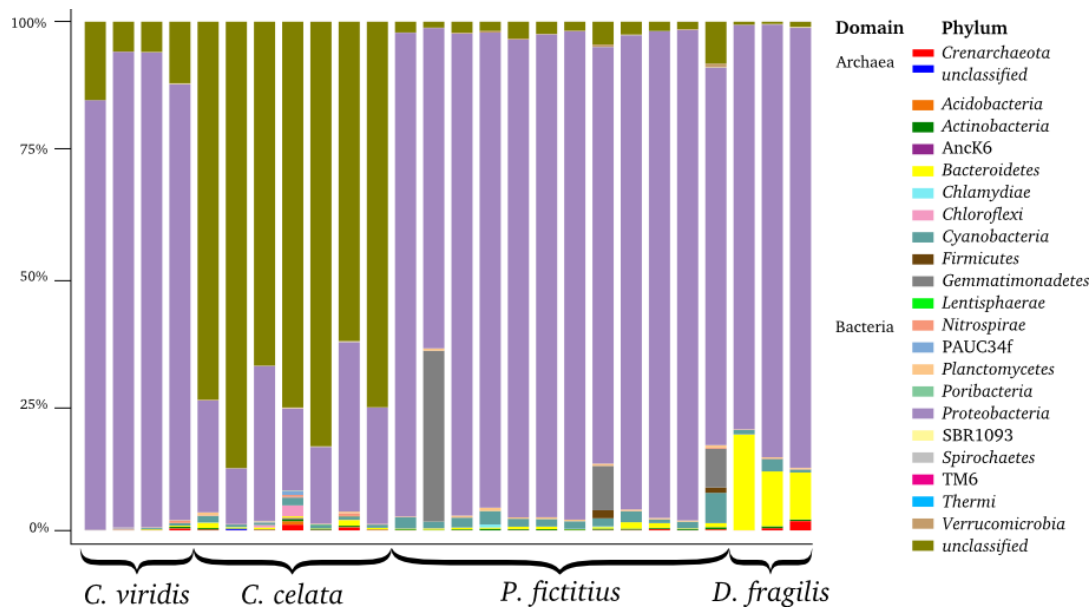


Figure 3.14 - Phylum-level relative abundance barchart for sequence libraries normalized by size (11,203 sequence reads per sample, singletons excluded).

Of note was also a large group of unclassified OTUs at the phylum level, observed in all *C. celata* profiles. Regarding *D. fragilis*, Bacteroidetes were seen in abundances of up to 18.79%. Further, *Gemmatimonadetes* was noted along some *P. fictitius* samples, although inconsistently, while Cyanobacteria represented about 5% of the prokaryome of this species.

In **Table 3.1**, values correspond to averages of OTUs and sequences (paired with standard deviations) found per specimen within each sponge species, when library sizes were normalized at *c.* 11,203 sequence reads per sample. Comprising values resulting not from the cumulative pooling of samples but from averages representing all replicates (specimens) within each sponge, this Table shows how the composition of the sampled prokaryomes vary in a quantitative manner.

Table 3.1 – Average number of sequences and OTUs as detected in the normalized dataset per prokaryotic phylum across sponge species.

Phylum	<i>n=7</i> <i>Cliona celata complex</i>		<i>n=4</i> <i>Cliona viridis</i>		<i>n=3</i> <i>Dysidea fragilis</i>		<i>n=12</i> <i>Phorbac fictitius</i>	
	OTUs	Sequences	OTUs	Sequences	OTUs	Sequences	OTUs	Sequences
<i>Crenarchaeota</i>	15 ± 5.68	53 ± 55.36	7 ± 8.85	30 ± 49.71	18 ± 8.08	186 ± 189.21	7 ± 5.42	17 ± 19.03
<i>Acidobacteria</i>	4 ± 6.18	15 ± 29.10	3 ± 1.71	4 ± 2.16	1 ± 1.53	2 ± 2.00	1 ± 1.38	2 ± 3.36
<i>Actinobacteria</i>	9 ± 6.19	21 ± 14.74	7 ± 8.54	12 ± 16.57	9 ± 4.93	25 ± 16.20	10 ± 4.94	27 ± 9.58
AncK6	0 ± 0.76	3 ± 7.18	0 ± 0.50	0 ± 0.50	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
<i>Bacteroidetes</i>	28 ± 20.39	57 ± 51.57	7 ± 8.08	9 ± 10.13	27 ± 17.52	51 ± 31.76	18 ± 5.90	52 ± 42.89
<i>Chlamydiae</i>	1 ± 0.95	1 ± 1.86	0 ± 0.50	1 ± 1.00	0 ± 0.00	0 ± 0.00	1 ± 0.79	7 ± 17.51
<i>Chloroflexi</i>	11 ± 19.92	46 ± 96.53	5 ± 7.14	6 ± 10.59	1 ± 1.00	1 ± 1.00	1 ± 0.67	1 ± 0.67
<i>Cyanobacteria</i>	41 ± 25.29	67 ± 51.67	14 ± 13.61	24 ± 25.18	54 ± 15.50	150 ± 117.23	91 ± 28.65	371 ± 397.85
<i>Firmicutes</i>	3 ± 2.44	4 ± 4.16	1 ± 0.96	1 ± 0.96	1 ± 1.73	1 ± 1.73	1 ± 1.14	1 ± 1.14
<i>Fusobacteria</i>	0 ± 0.53	1 ± 0.79	0 ± 0.50	1 ± 1.00	0 ± 0.58	0 ± 0.58	0 ± 0.00	0 ± 0.00
<i>Gemmatimonadetes</i>	3 ± 3.72	6 ± 8.46	1 ± 1.15	2 ± 1.73	1 ± 1.53	1 ± 1.53	1 ± 1.16	1 ± 1.44
<i>Lentisphaerae</i>	1 ± 1.41	2 ± 2.23	0 ± 0.50	1 ± 1.00	1 ± 1.00	1 ± 1.53	1 ± 0.90	1 ± 1.15
<i>Nitrospirae</i>	5 ± 4.93	21 ± 31.65	4 ± 6.73	14 ± 26.18	4 ± 3.61	10 ± 9.50	3 ± 3.55	5 ± 7.51
PAUC34f	2 ± 3.83	16 ± 31.58	1 ± 1.00	1 ± 1.00	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
<i>Planctomycetes</i>	11 ± 5.72	23 ± 20.75	6 ± 4.79	9 ± 6.24	12 ± 1.73	18 ± 1.53	17 ± 6.44	38 ± 19.65
<i>Poribacteria</i>	1 ± 1.51	1 ± 1.51	0 ± 0.50	1 ± 1.00	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
<i>Alpha Proteobacteria</i>	51 ± 27.99	163 ± 130.41	73 ± 19.26	8619 ± 457.18	46 ± 21.57	2238 ± 1095.18	38 ± 14.53	645 ± 1109.52
<i>Beta Proteobacteria</i>	5 ± 1.60	45 ± 35.15	3 ± 2.65	13 ± 15.50	8 ± 1.00	1401 ± 600.62	3 ± 1.51	7 ± 9.69
<i>Delta Proteobacteria</i>	14 ± 12.32	79 ± 141.97	5 ± 7.23	11 ± 10.97	14 ± 7.00	45 ± 52.54	8 ± 4.91	10 ± 8.08
<i>Epsilon Proteobacteria</i>	0 ± 0.49	5 ± 12.04	0 ± 0.50	0 ± 0.50	0 ± 0.58	0 ± 0.58	0 ± 0.39	0 ± 0.39
<i>Gamma Proteobacteria</i>	110 ± 65.55	1179 ± 630.52	41 ± 20.49	93 ± 34.31	83 ± 37.47	317 ± 189.86	191 ± 40.66	3546 ± 1961.19
<i>Unclassified Proteobacteria</i>	25 ± 6.52	883 ± 408.86	98 ± 6.19	2329 ± 264.26	60 ± 12.06	6685 ± 889.46	94 ± 20.59	6285 ± 2636.63
SBR1093	3 ± 3.21	5 ± 5.83	2 ± 2.06	4 ± 3.74	3 ± 2.52	9 ± 8.54	1 ± 1.08	2 ± 2.23
<i>Spirochaetes</i>	0 ± 0.76	0 ± 1.13	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
<i>Thermi</i>	0 ± 0.00	0 ± 0.00	0 ± 0.50	0 ± 0.50	0 ± 0.58	0 ± 0.58	0 ± 0.39	0 ± 0.62
<i>unclassified</i>	148 ± 4.47	8502 ± 983.01	6 ± 1.73	19 ± 4.99	11 ± 4.51	55 ± 16.26	33 ± 10.07	165 ± 71.08
<i>Verrucomicrobia</i>	5 ± 3.65	8 ± 5.89	3 ± 3.59	4 ± 5.48	4 ± 1.53	5 ± 2.52	5 ± 4.03	22 ± 23.44
Total Averages per Sample	480.71	11150.29	278.75	11173.25	340.00	11017.33	523.25	11203.00

By addressing phylum-level taxonomical data from all prokaryomes (**Table 3.1**), it was possible to observe the dominance of proteobacterial sequences, which in *C. viridis* and *D. fragilis* total about 10,000 sequences per specimen, going down to 2,300 in *C. celata*. All sponges presented about 200 OTUs assigned to Proteobacteria per specimen, except for *P. fictitius*, in which an average of 333 OTUs per specimen was found. Within the latter phylum, γ -Proteobacteria predominates in OTU numbers across all sponges but *C. viridis*, in which unclassified proteobacterial members dominate. Concerning sequence number, unclassified Proteobacteria accounted for more than half of the average proteobacterial numbers in *D. fragilis* and *P. fictitius*, while *C. celata* and *C. viridis* were dominated by γ - and α -Proteobacteria.

Further, Poribacteria was represented by no more than one exemplar (OTU) per sponge species, with seen absences in *D. fragilis* and *P. fictitius*. On average, *P. fictitius* held the most well developed consortium of Cyanobacteria, with 91 OTUs, totalling 371 sequences per specimen. This phylum was further observed in *D. fragilis*, with about half the OTU and sequence numbers per specimen. Only 11 OTUs and 46 sequences were assigned to Chloroflexi in *C. celata*, being almost absent in the remainder prokaryomes. The unclassified bacterial phylotypes across the dataset seemed to be evenly distributed across sponge species, except in the case of *C. celata*, in which 148 OTUs have had more than 8,500 assigned sequences per specimen, therefore being the dominant group in this species.

Further insight into the composition of Proteobacteria in the present dataset was gained, as two classes, α - and γ -Proteobacteria were shown to predominate along with an unclassified cluster of proteobacterial lineages (**Figure 3.11**).

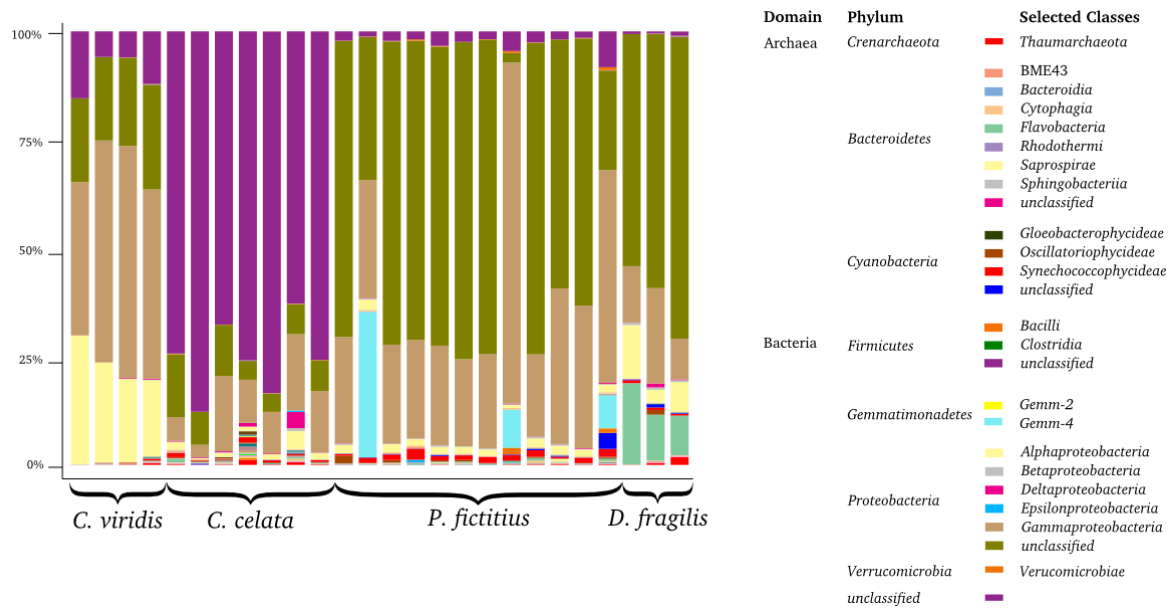


Figure 3.15 - Class-level relative abundance barchart for sequence libraries normalized by size (11,203 sequence reads per sample, singletons excluded).

Proteobacterial unclassified lineages were predominant across the dataset (except for *C. celata* samples where sequences unclassifiable at the phylum level - purple bars - prevailed) reaching relative abundances as high as 78.6%, whereas the maximum observed relative abundance of α -Proteobacteria was of about 30%. Further, the classes *Gemm-4* (*Gemmatimonadetes*) and *Flavobacteria* (*Bacteroidetes*) could be identified at reasonable numbers in *P. fictitius* and *D. fragilis*, respectively.

Examining taxonomic compositional data more closely, it was possible to notice how gamma-proteobacterial diversity was defined mostly by *Pseudomonadales* and unclassified phylotypes (**Figure 3.12**).

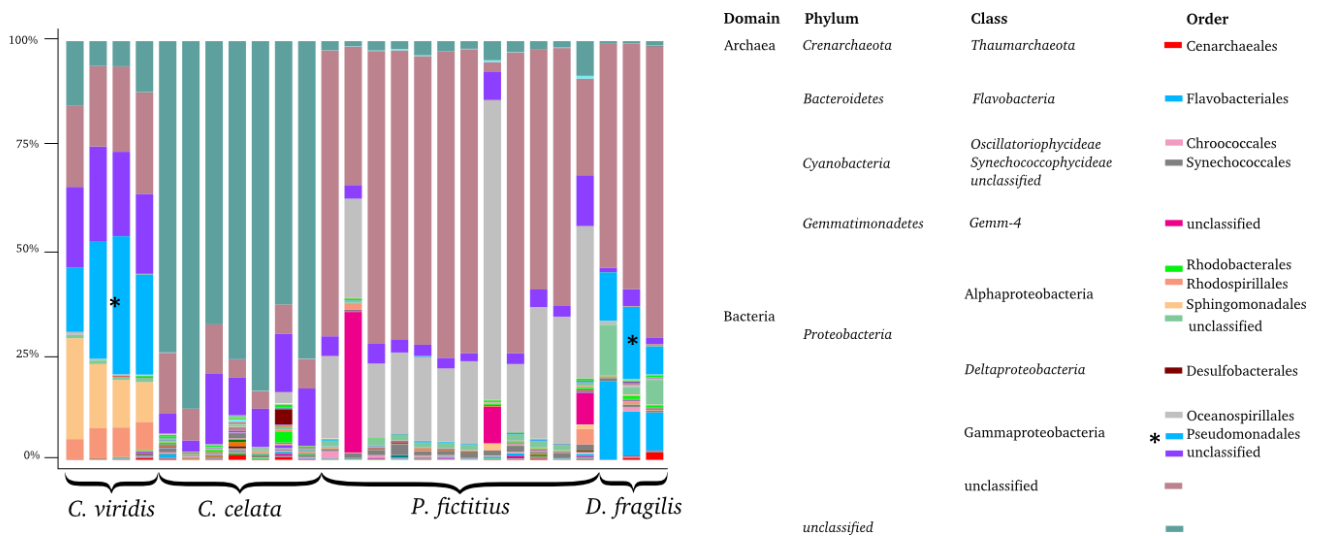


Figure 3.16 - Order-level relative abundance barchart for sequence libraries normalized by size (11,203 sequence reads per sample, singletons excluded).

Further, α -Proteobacteria were primarily defined by *Sphingomonadales*, *Rhodospirillales*, *Rhodobacterales* and a group of unclassified OTUs. Some classes were shown to be mainly composed of sequences assigned to one order, as was the case for *Gemm-4*, mostly dominated by unclassified OTUs, δ -Proteobacteria, with *Desulfobacterales* as the main order, and *Flavobacteria* showing only the *Flavobacteriales* order.

As seen in **Figure 3.13**, a pattern of total ordered OTU abundances across the dataset shows little information with regard to either inter- or intra-specific similarities among sponges. Summarily, it was noted that the most abundant OTUs prevailed within *P. fictitius* and *C. celata*.



Figure 3.17 - General profiles of OTU abundance set in ascending order across all samples in a fourth-root-transformed normalized dataset. Clustering was set by single linkage.

C. viridis presented a different pattern in OTU abundance profile, towards a small number of highly abundant phylotypes. It therefore became apparent that a focus on these OTU profiles was needed.

An in-depth analysis of the distribution of the 50 most abundant OTUs across the data was useful in depicting clear trends concerning the composition (or not) of a core and/or specific microbiota in each sponge species (**Figure 3.14**).

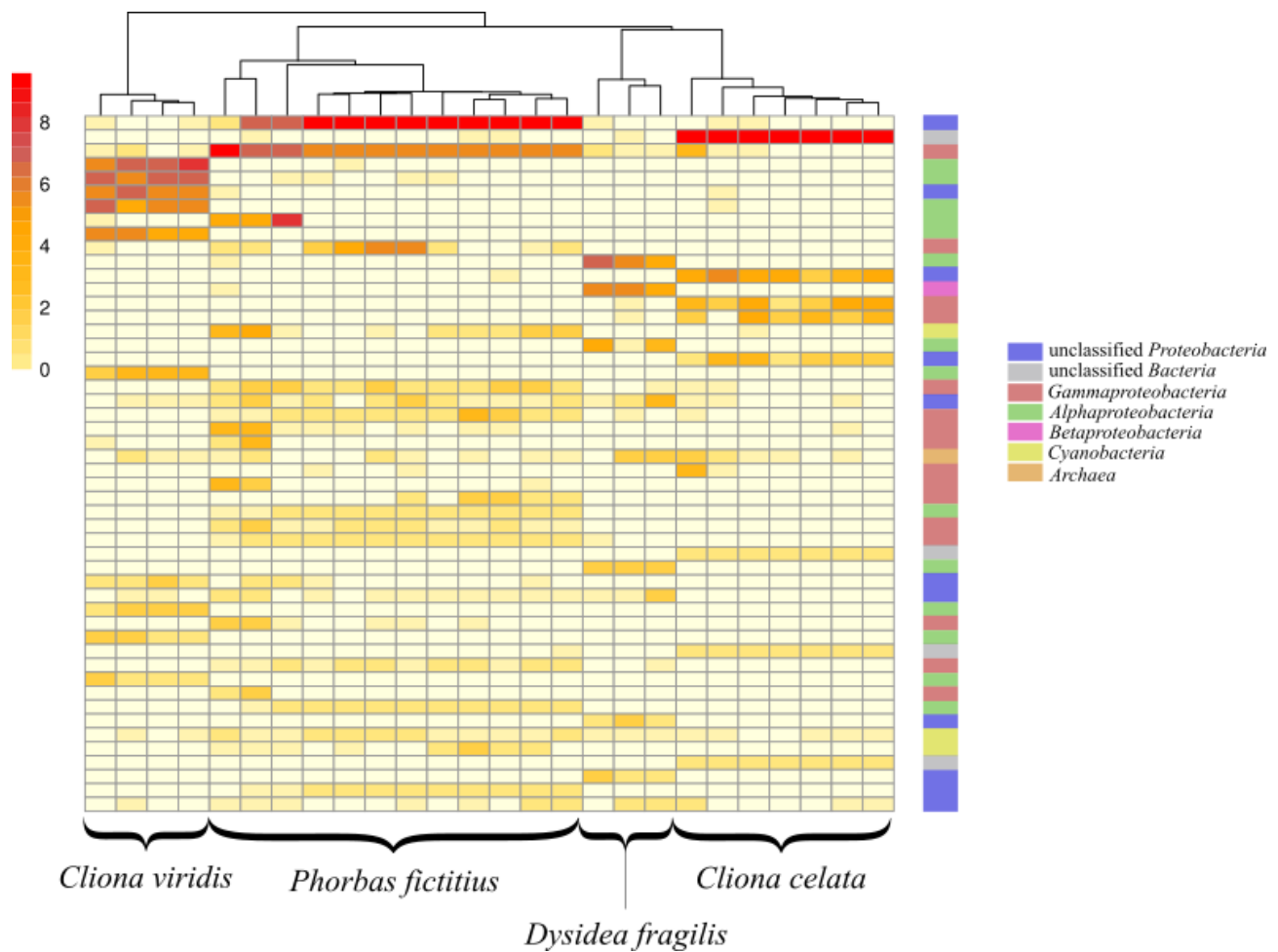


Figure 3.18 – Heatmap showing the distributions of the 50 most abundant OTUs across all samples. Clustering of samples was performed using the single linkage calculated with fourth-root transformed abundance values, for an improved visualization of OTU abundance (normalized dataset). Greengenes taxonomy to phylum- and class-level (for Proteobacteria) within *Bacteria*, and for Archaea in general, is depicted in the coloured squares placed right to the heatmap. *C. viridis* seemed to possess, across all collected specimens, a set of OTUs identified as two unclassified members of the *Rhodospirillaceae* family (α -Proteobacteria), one unclassified member of the *Rhodospirilalles* order (α -Proteobacteria) and four unclassified proteobacterial OTUs (Otu000032, Otu001733, Otu000163, Otu000004). The latter phylotypes accounted, in cumulative numbers across all *C. viridis* replicates, for 11,302, 8,958, 7,029 and 6,631 sequences, respectively.

The most well sampled sponge species, *P. fictitius*, was populated by an unclassified Proteobacteria (Otu001789) which, although absent from one sample, was the most abundant phylotype across the whole dataset, accounting for 68,809 sequences in total.

An OTU belonging to the *Endozoicimonaceae* family (γ -Proteobacteria, Oceanospirillales, (Otu029487) was the second most prevalent OTU (26,995 sequences) associated to *P. fictitius*. Further, an unclassified bacterium (Otu000602) totalling 50.951 sequences was stably present across all *C. celata* samples. No obvious dominances were observable in the *D. fragilis*-associated dataset, albeit two phylotypes (Otu000078 and Otu000047) occurred in slightly higher abundances. An unclassified α -Proteobacteria and a member of the EC94 order (β -Proteobacteria) were the most prevalent OTUs within the latter consortium, both accounting for about 4,000 sequences each. The 20 most abundant OTUs in the overall dataset all possessed sequence numbers above 1,000 (not shown in **Figure 3.14**).

By exploring the whole dataset (singletons excluded) in **Figure 3.15**, it was possible to determine a high number of phylotypes specific to *C. celata* and *P. fictitius*.

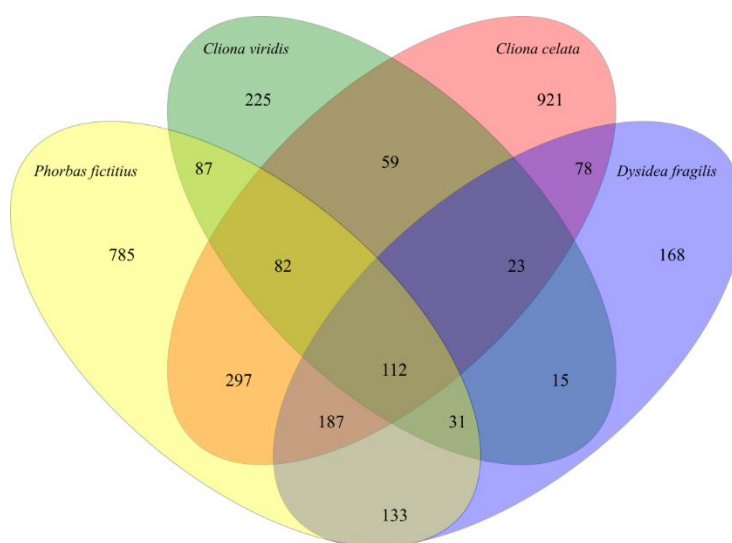


Figure 3.19 – Venn diagram depicting the cumulative total number of OTUs shared by and specific to each sponge species.

C. celata and *P. fictitius* hosted 1,759 and 1,714 prokaryotic species in 78,855 and 135,180 sequences, respectively. Species-specific OTUs were also observed in *C. viridis* and *D. fragilis*, although less pronouncedly probably because of the lower replication number, and consequently sequence effort, applied in the characterization of these species.

A set cut-off of 50 sequences per OTU allowed for the exclusion of the rarer components of the analysed dataset. As shown in **Figure 3.16**, a drastic decrease in the number of species-specific phylotypes was seen, namely in the set attributed to *C. celata*.

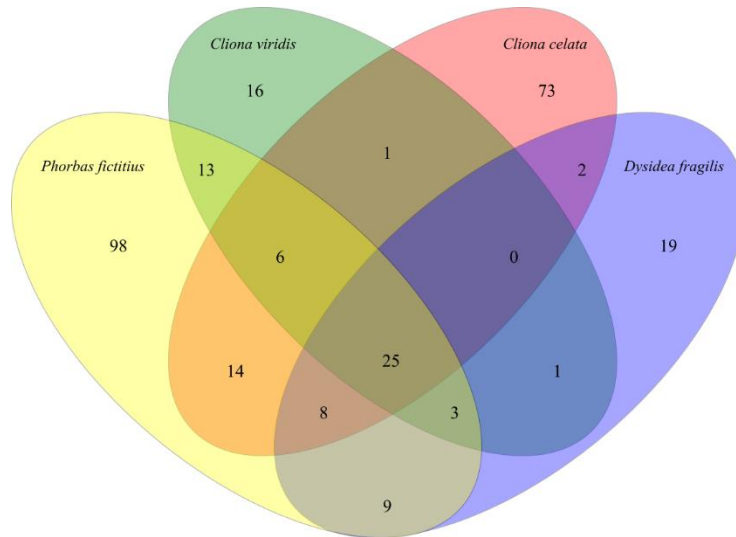


Figure 3.20 - Venn diagram depicting the cumulative number of OTUs, with a minimum of 50 assigned sequences, shared by and specific to each sponge species.

Shared OTUs across sponge species were further seen to decrease. Again comparing **Figure 3.16** to the latter Venn diagram (depicted in **Figure 3.15**), the set of OTUs shared by all sponges was reduced to less than a quarter.

Setting the minimum number of assigned sequences to 100 per OTU, it was possible to obtain further insights into a dataset where planktonic prokaryotes putatively not belonging to the true sponge prokaryome and the rare fraction of this consortium were excluded. This allowed for a more robust depiction of what may compose the “true” core prokaryome across this dataset (that is, the pool of OTUs common to all analysed sponge species), depicted in **Figure 3.17** with 15 assigned OTUs.

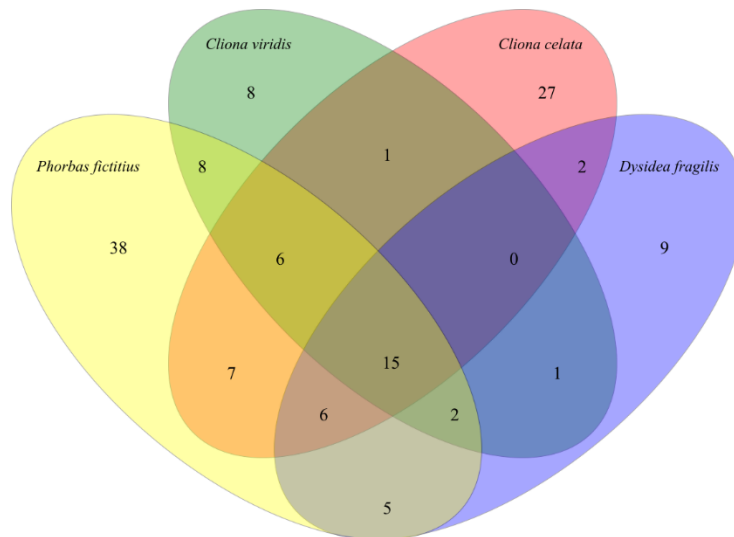


Figure 3.21 - Venn diagram depicting the cumulative number of OTUs, with a minimum of 50 assigned sequences, shared by and specific to each sponge species.

Starting with 112 OTUs, the richness of this “core” consortium dropped to 13.4% of its original values when filtered for OTUs containing more than 100 sequences, finally accounting for 15 OTUs.

Proportionally, the largest drop of species-specific OTUs relative to a sponge’s total OTU count was seen in *C. celata*. Starting with 921 OTUs, filtering led to only 8 of those being present in abundances higher than 100 sequences. In the case of *C. viridis*, the same process caused 225 OTUs to be reduced to 8, this proportionally being the second largest drop in species-specific OTU count. The most diverse, “species-specific” pool of abundant OTUs (> 100 sequences) was observed in *P. fictitius*, with 38 such phylotypes detected exclusively in this sponge (**Figure 3.17**).

3.2. Exploration for prokaryotic ‘microbial dark matter’

Re-classification of the 20 most abundant bacterial OTUs unclassified at the phylum level with the Greengenes database was attempted using sequence alignments in the SILVA database and posterior phylogenetic assessments with the ARB software. To this end, the SILVA Aligner tool was employed (www.arb-silva.de/aligner/), and the resulting alignment imported into the ARB software environment containing the SILVA reference database release 123 (July 23, 2015). After manually correcting the alignments whenever deemed necessary, a Maximum Likelihood tree was constructed (see **2.2**) containing all 20 previously unclassified sequences and reference sequences from the SILVA database (**Figure 3.18**).

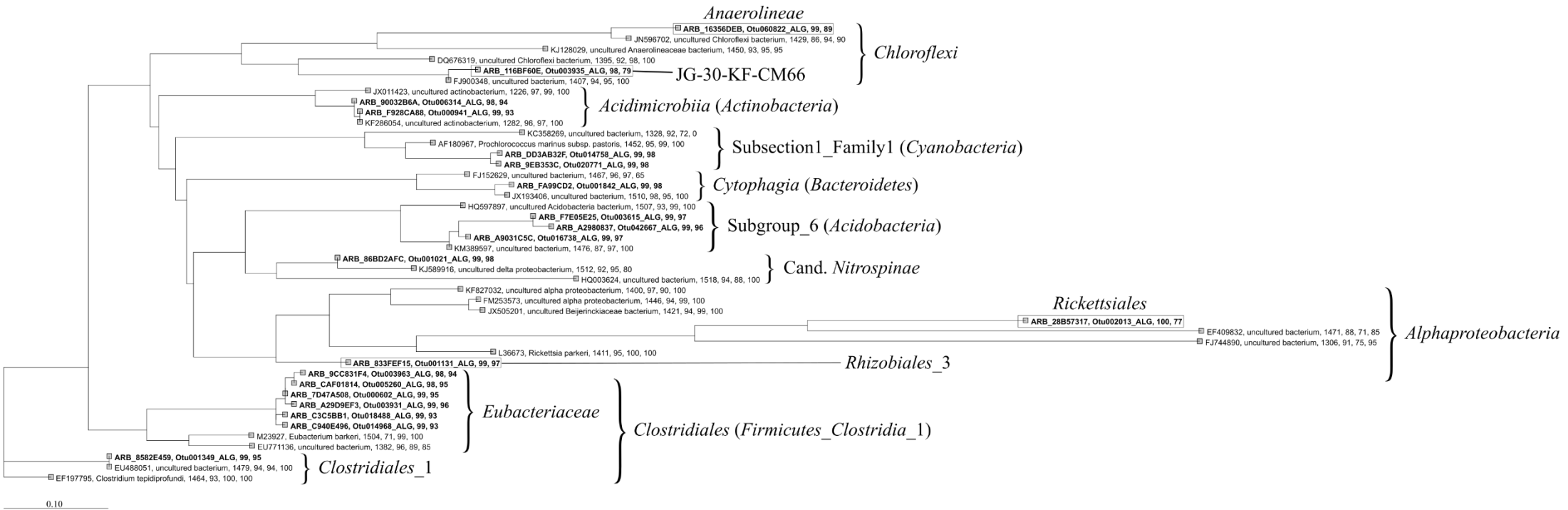


Figure 3.22 - Phylogenetic tree depicting the twenty most abundant OTUs (highlighted in black) unclassified at the phylum-level across the dataset, re-aligned using the SILVA 123 "non-redundant" (NR) SSU Reference dataset (July 2015). The tree was built in ARB⁸⁵, based on the RAxML⁸⁶ model, under the Gamma model of rate heterogeneity.

This approach resulted in the successful re-classification of all 20 OTUs into known bacterial phyla. By integrating the unknown sequences in the pre-aligned gold reference database for prokaryotic 16S phylogeny provided by SILVA, rooting of the tree was not deemed necessary. As such, following the explorative nature of this component of the work and for classification purposes, **Figure 3.22** shows an unrooted phylogenetic tree. In total, these were assigned to eight phyla: Bacteroidetes, Proteobacteria, Acidobacteria, candidate phylum Nitrospinae, Actinobacteria, Chloroflexi, Cyanobacteria and Firmicutes_Clostridia_1. The latter, currently considered a putative phylum by SILVA version 123, is a branch of Clostridia (Firmicutes), highly paraphyletic as depicted by SILVA 16S-based phylogeny, was associated to seven OTUs, being that six of these were classified as part of the Eubacteriales order (Clostridiales class) and one as Clostridiaceae, within the same class. Within the first order, four (Otu000602, Otu003963, Otu014968 and Otu018488) were part of the 50 most abundant OTUs in the normalized dataset, seen as ‘unclassified Bacteria’ in **Figure 3.14**.

Each of the two OTUs assigned to Chloroflexi were attributed a class: *Anaerolineae* and JG-30-KF-CM66. Two clearly defined clusters were formed, encompassing the closest matches of each phylotype. Un-transformed abundances of Otu060822 and Otu003935 were of 1,052 and 466 sequences. Two OTUs were assigned to *Acidimicrobiia*, a class within Actinobacteria. A total of 14,544 sequences were distributed across both phlotypes: Otu000941 and Otu006314. As for Cyanobacteria, two OTUs were classified as members of Subsection 1 Family 1 class. Both of these were members of the *Prochlorococcus* order, here totalling more than 800 sequences.

Regarding Bacteroidetes, one assigned OTU was classified as a member of the class *Cytophagia*, accounting 314 sequences, whereas in Acidobacteria, three OTUs belonged to Subgroup_6 class. These altogether summed up 15,614 sequences. Lastly, a member of the recently proposed Candidate phylum Nitrospinae, previously a clade assigned to δ -Proteobacteria, was also found⁸⁷. The corresponding OTU held 2,128 sequences and the generated cluster was seen to be close to α -Proteobacteria. The latter class was observed as two OTUs were assigned to the orders *Rickettsiales* and *Rhizobiales_3*. These presented 427 and 697 sequences, respectively.

Following OTU reclassification, it was possible to review the distribution of each of the previously unclassified OTUs across all samples for. As shown in **Figure 3.10**, the

largest concentration of unclassified OTUs was observed in *C. celata*. In this species, six previously unclassified OTUs could be affiliated with Firmicutes_Clostridia_1, a Clostridia branch paraphyletic to the remaining clades within phylum Firmicutes. The unclassified consortium of phylotypes found by the Greengenes 2013 reference database in *C. celata* was therefore successfully re-classified. Other samples belonging to *P. fictitius* and *C. viridis* were also shown to hold previously unclassified phylotypes re-classified as belonging to α -Proteobacteria.

The diversity within the singletome (**Figure 3.19**), that is, the ensemble of OTUs comprising solely one sequence across the non-normalized dataset, was relatively high, comprising 21 bacterial phyla and 2 archaeal phyla. In total, 2,561 singleton OTUs were observed.

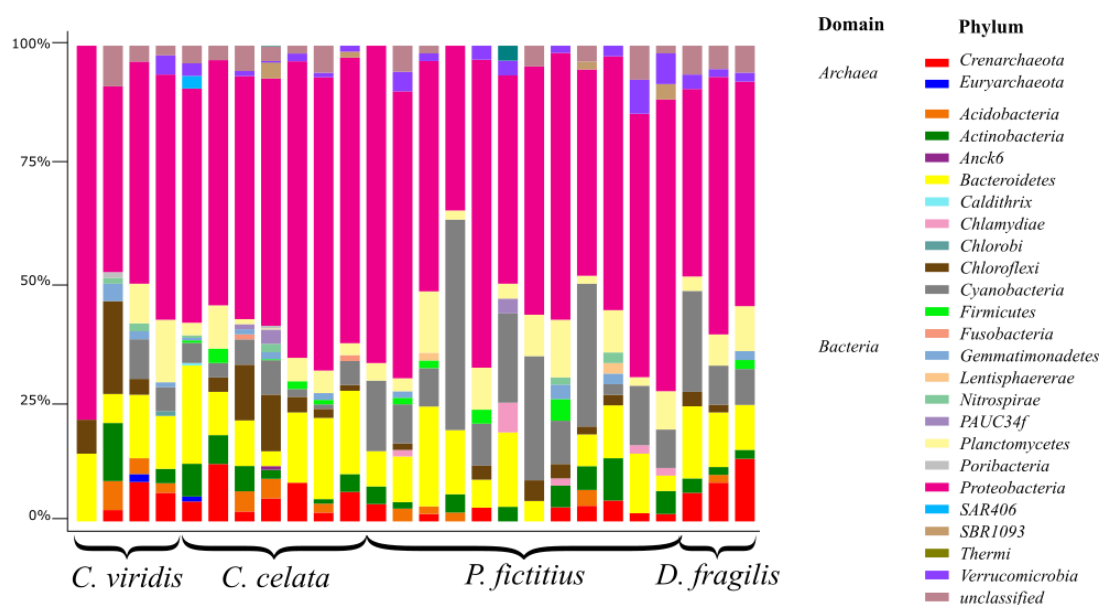


Figure 3.23 - Phylum-level bar chart depicting singleton-specific taxonomic assignments across the non-normalized dataset.

On average, 223.9 ± 115.2 singleton OTUs were found in *C. celata*, while in *C. viridis* 63.8 ± 18.4 singleton phylotypes were present, against 495.43 ± 63.07 and 286.0 ± 55.32 OTUs in the singleton-excluded dataset. The remainder sponges, *D. fragilis* and *P. fictitius*, showed 49 ± 8.3 and 49.3 ± 4.8 singleton OTUs on average. This compares to 357 ± 64.02 and 523.25 ± 23.58 OTUs in the singleton-excluded dataset, accounting for as much as a 10-fold decrease in richness.

In total, 22 bacterial and two archaeal phyla were found in the singletome. Comparing the singletome with the singleton-excluded dataset shown in **Table 3.1**, with 21 found

bacterial phyla and one archaeal phylum, general trends for greater abundance of Proteobacteria and Bacteroidetes-related taxa were maintained, while the fraction of unclassified bacteria was greatly diminished. Moreover, cyanobacterial OTUs as well as many exemplars of *Crenarchaeota* were noted.

These singleton phylotypes presented little taxonomical convergence in comparison with the singleton-excluded dataset, mainly due to the absence of large amounts of unclassified bacterial and proteobacterial taxa. However, in accordance with the singleton-excluded dataset, taxa typically found in sponges such as Poribacteria, Chloroflexi and Acidobacteria were frequently not found.

3.3. Sponge host phylogeny

Phylogenetic inference of sponge cytochrome oxidase (subunit 1) gene (*cox1*) sequences depicted little differentiation among most of the sampled clionids, since 8 individuals belonging to both species concisely formed one single phylogenetic cluster with little heterogeneity at the nucleotide sequence level (see **Figure 3.20**).

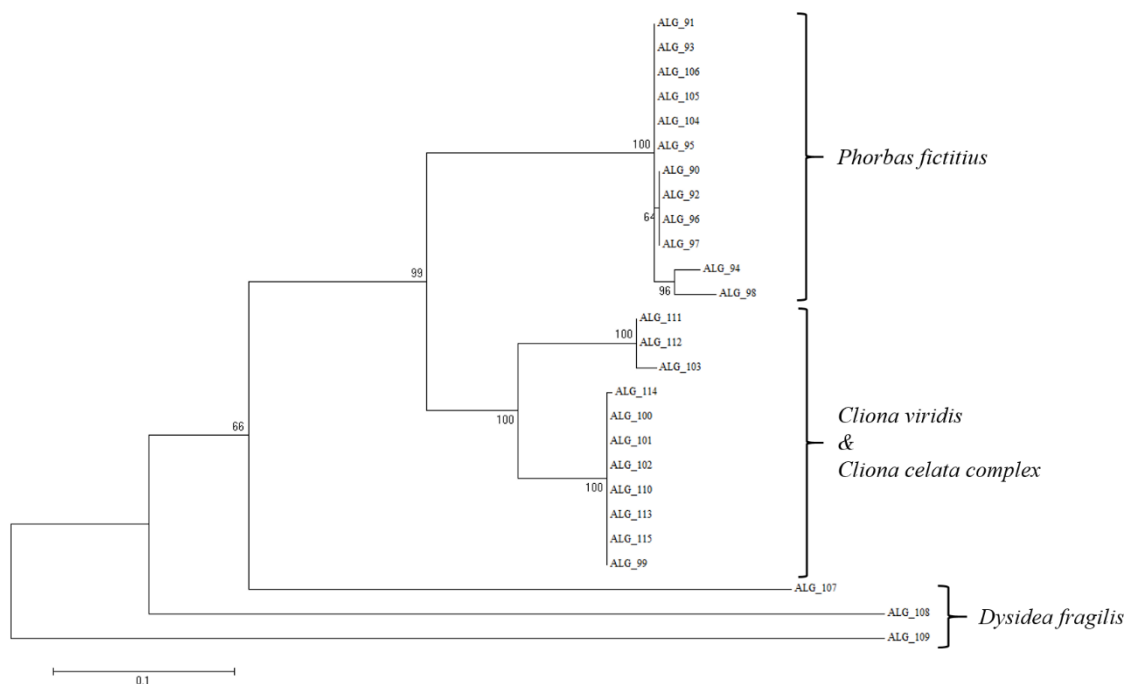


Figure 3.24 - Phylogenetic tree depicting *cox1*-based relationships among the sampled sponges. Maximum Likelihood method based on the Tamura 3-parameter model was used and tested with 500 bootstrap replicates.

Strikingly, however, a smaller group formed by three *C. celata* individuals, collected in both sampling sites, was formed and found to be clearly distinct from the previous

cluster. Among the individuals identified as *P. fictitius*, further intra-species variation was revealed by *cox1* phylogenetic assessments. In fact, three separated groups were revealed, with the most distinct cluster encompassing samples 94 and 98, followed by another group including samples 90, 92, 96 and 97. All of the latter were sampled in Galé Alta beach, and were therefore allopatric to the ones included in the last group, which contained *P. fictitius* individuals collected in Armação de Pêra. Lastly, sampled individuals assigned to *D. fragilis* were set apart from all other sample clusters.

4. Discussion

In this thesis, four marine sponges, sympatrically coexisting off the coast of Algarve, were sampled for their prokaryomes by means of high throughput sequencing. The hypothesised species-specific character of their prokaryotic assemblages was verified with a suite of bioinformatics tools coupled to state-of-the-art statistical assessments. Specifically, this thesis aimed to diagnose how (dis-)similarity between symbiotic consortia vary according to the level of phylogenetic relatedness among host species, and unveil the extent of bacterial diversity hidden in the “microbial dark matter” associated with marine sponges.

4.1. Pre-processing methodologies

Pre-processing of Earth Microbiome Project data by the Consortium for Sponge Microbiology led to a total of 22,848,828 sequences assigned to 45,981 OTUs from 1,226 sponge-derived, seawater and sediment samples. The present study comprised about 7% of the abovementioned OTU richness across its assigned 26 samples, encompassing the analysis of 557,618 16S rRNA gene sequence reads in total. Of the 26 samples studied in detail here, 17 presented more than 20,000 sequences, while 6 presented more than 25,000 sequences and one surpassed 30,000 sequences. Therefore, normalization of such values was needed for further statistical analyses. As such, a threshold of 11,203 sequences, corresponding to the lowest sequence read number obtained within this set of samples, was defined, causing for large amounts of information to be lost, therefore reducing the dataset to 291,278 sequences. This technique randomly and proportionally sub-samples a set value from within each sample to equalize sample size across the whole dataset. As such, loss of information is compensated by statistical validity for further downstream analyses, as ordination and alpha diversity (e.g, within-sample diversity, as opposed to beta diversity, concerned to among-sample diversity) indices calculations. The full dataset was nevertheless used for exploration of diversity at a general scale, determination of shared and specific OTUs by means of Venn diagram construction, and for permitting a visualization of ‘microbial dark matter’. This way, the whole sequence information generated in this study could be fully exploited using both ecologically-driven statistical assessments along with exploratory surveys of the genetic resources present in the analysed samples.

Illumina sequencing is currently growing in usage and sequencing depth, subsequently decreasing its associated cost of operation ⁶⁶. The high computational demand for the pre-processing of nucleotide (e.g. 16S rRNA gene) sequence data, namely chimera-checking, OTU clustering and taxonomical assignments, but also for the post-processing of the generated information, via for instance robust statistical assessments, is the main concern when analysing NGS-derived (Next Generation Sequencing) outputs. Helpfully, since 2009 and 2010 software packages such as mothur and QIIME, trained in the analysis of 16S rRNA gene fragments, have become available and transitioned from 454-pyrosequencing pipelines to Illumina-directed ones, thus adapting to not only much larger datasets but to new methodologies as well ^{53,54}. These are not, however, complete analysis packages and the resulting outputs are inherently dependent on high-end statistical software such as R, which are not constrained by dataset size ⁸⁸.

In the case of this study, 11,203 sequences per sample in the context of sponge molecular microbiology surpassed those of some recent studies using 454-pyrosequencing ^{35,47,49} and even Illumina ⁴⁶. It can therefore be stated that the normalized number of sequences employed here compares to those used in the most advanced studies of microbial community diversity in marine sponges, allowing for robust statistical and ecological assessments to be made.

4.2. Richness and diversity of sponge prokaryomes

The sheer number of phylotypes, OTUs, or “species” within a dataset is by itself an important measure which intrinsically defines diversity. In the case of the four sampled sponges, no relationship was found in the matter of OTU richness (total numbers of OTUs). That is, all sponges presented significantly different prokaryome sizes, indicating instead a putative host-specific composition of the prokaryomes (ANOVA, $P < 0.01$). The highest value among these was attributed to *P. fictitius* (order *Poecilosclerida*). Apart from one recent study exploring the culturable bacterial community in *Phorbas tenacior*, knowledge of unculturable prokaryotic community diversity in species of the genus *Phorbas* is virtually non-existent, unlike in other genera of *Poecilosclerida* ⁸⁹. Indeed, Uriz and colleagues proposed that species in the genus *Hemymicale* are HMA sponges, whereas Hentschel *et al.* and Gloeckner *et al.* noted several LMA genera (2 and 7, respectively) within *Poecilosclerida* by means of

electronic microscopy^{31,63,64}. Here, *Phorbas fictitius* hosted about 520 OTUs and was estimated to host about twice as much as the observed prokaryotic species (**Figure 3.2**).

No generally applicable bacterial richness threshold has been proposed to distinguish among high and low microbial abundances in marine sponges. In fact, total OTU values tend to vary within and between LMA and HMA sponge categories. So far, consensus regarding OTU richness for HMA sponges is little, varying from 138 to 263 off the coast of Cataluña, Spain, to 3,942 for deep water marine sponges, and from 90 to 140 OTUs detected by ARISA (Automated Ribosomal Intergenic Spacer Analysis)^{13,41,90}. In LMA sponges, OTU numbers have been seen to predominantly be hundreds of OTUs lower¹³. It is important to notice that such values may vary sharply depending on the sequence output, fingerprinting techniques, and analytical pipelines employed in each study. In this regard, information generated by the Sponge Microbiology Consortium in the scope of the EMP, will be of invaluable assistance in future studies aiming to compare diversity values between LMA and HMA sponges in a standardized manner. The observed values of OTU richness in *P. fictitius* resemble HMA sponges when compared to other studies, as do the ones for *C. celata* (500 OTUs), which is further estimated to hold as much as 750 OTUs (**Figure 3.5**)^{13,43}. Compared to numbers given by LMA sponges in other studies, OTU richness and Chao1 estimations in prokaryomes associated with *P. fictitius* and *C. celata* were strikingly high^{34,41}.

Rarefactions of the cumulative and average numbers of species (**Figures 3.3 and 3.4**) serve as estimates of the coverage of the microbial diversity found in any given sample. It is known that PCR-based methods may disproportionally amplify rare and common DNA templates or selectively favour particular taxonomic groups, creating constraints towards the reliability of richness estimations from environmental samples⁴⁰. Further, rarefaction methods for richness analysis are known to be biased in the sense that sampling of a complete community is greatly affected by the current inability of finding rarer species by means of sequencing⁹¹. However, and acknowledging such biases currently transversal to DNA sequencing-dependent molecular microbial ecology, rarefaction was applied in this study not to predict richness but to assess the adequacy of sampling.

Indeed, rarefaction analysis rendered very informative results for the purposes of this study. Owing to sample numbers affecting the rarefaction curve, it served as a proxy to analyse sampling depth for each group of sponges. The inferred depth of sampling was,

however, considered acceptable only for *C. viridis*, whereas the remainder groups of samples presented growing rarefaction curves. The absence of a decrease in slope along the curve served to infer the extent of unsampled diversity. It was therefore possible to classify *C. celata*, *P. fictitius* and *D. fragilis* as undersampled groups of sponges in spite of the very high sequencing effort applied in this study, with *P. fictitius* being the one with a curve closest to a 'plateau', or to reach a maximum in detection of new species.

The cumulative nature of the rarefaction analysis depicted in **Figure 3.3** enabled valuable conclusions to be made regarding the requirements for sampling size in the context of future Illumina-based analyses of sponge-associated prokaryomes. Indeed, where the cumulative increments in richness within each sponge species were avoided by averaging OTU presence across all samples (**Figure 3.4**), it can be stated that at the present depth of sequencing, all sponges but *C. viridis* are undersampled. The four individuals assigned to *C. viridis* reach a plateau at around 200 OTUs in **Figure 3.4**, thereby indicating a stabilization of richness detection.

A normalized rarefaction analysis allowed for further insights to be gained regarding richness across sponges. Namely, the two *Cliona* species differed greatly in prokaryome richness; *C. celata* being the one with the steepest cumulative curve. This trend was contradicted when the latter was compared with an averaged rarefaction curve, which eliminated in part the 'pan-microbiome' effect of OTUs specific of each sample. *C. celata* therefore presented similar exponential trends along with *P. fictitius* in the normalized and averaged analysis. By contrast, both *D. fragilis* and *C. viridis* seemed to have reached a plateau for exponential growth, and was shown to be very stable in the latter sponge. This likely indicates that the applied sequencing effort was effective for *C. viridis*, but further generation of information is needed for the remaining sponges.

It is known that in most cases HMA and LMA sponges differ significantly and proportionally in diversity and OTU richness, although a few exceptions are known where no correlation between microbial abundance and both the latter indices is noted^{13,35,41,43}. The present results therefore seemed to indicate the presence of LMA prokaryomes in all four sponges, assuming present richness values could depict true richness. Although the trending growth in OTU and sequence number seen for *P. fictitius* and *C. celata* may indicate possible high microbial abundance, the present normalized data shows no evidence towards such a status. Future microscopy surveys are required to adequately classify the species studied here as LMA or HMA sponges.

While evenness among the sponge prokaryomes proved to not significantly vary (ANOVA, $P < 0.1$), values from 0.34 to 0.42 indicate medium to low Pielou's evenness, and coincide with the assessment by Poppel *et al.*⁹². In this study, Pielou's evenness was shown not to vary among 25 species of HMA and LMA sponges. The results presented here hint towards some slight dominance, or a number of few species abounding in higher number than others, forming a larger (values near 0) or smaller (values near 1) core of dominant species. This becomes evident upon inspection of **Figure 3.14**, where all sponges are shown to hold slightly variable cores of dominant OTUs.

All sponges presented Shannon diversity values with no significant differences (ANOVA, $P = 0.78$), in spite of the observed significant differences across OTU richness values for all. It is seen in the evenness values (**Figure 3.6**) and illustrated in **Figure 3.14** that higher richness values correspond to the sponges with the most dominating OTUs, which in turn lowers the diversity measures. HMA sponges usually harbour microbiomes with significantly higher values of diversity, especially when compared to LMA profiles^{35,43,92}. Cumulatively, *C. celata* proved to hold the highest diversity across normalized samples. However, when using averaged rarefaction for the same dataset, the *C. celata* prokaryome did not differ as much from that of *P. fictitius* in terms of alpha diversity. Across samples, *C. celata* holds the largest pan-prokaryome (see **Figure 3.3**), meaning that a more diverse prokaryome within the sponge is to be discovered if further samples are analysed. An increased permissiveness for unique or generally not shared OTUs, seemingly 'transient', across *C. celata* replicates was therefore noted. |

4.3. Patterns of OTU presence across prokaryomes

In this study, ordination of dissimilarities among prokaryomes served to show that host-specificity of OTU consortia was observed, independently of poriferan phylogeny. Each consortium was viewed as highly specific (see **Figure 3.8**) and formed robust clades (MRPP, $P < 0.001$). This was mainly due to the fact that the prokaryotic consortia within *C. viridis* were identified as the most dissimilar with regard to *C. celata*. This sponge's prokaryotic communities clustered the closest with the remainder sponges, namely *P. fictitius* and *D. fragilis*, which are phylogenetically more distant from the former two species, as seen in **Figure 3.20**.

Host-specificity (e.g. organism-specific as opposed to species-specific; independent of host phylogeny) of sponge microbiomes is a topic for which a great amount of contradicting evidences have been published ^{34,46,90,95}. In one part, this hypothesis has been proven for a large number of sponges sampled by means of NGS technologies in sponges inhabiting coral reefs in Norway, as well as coastal and deep waters in the Mediterranean and Northern Atlantic ^{45,46,90}. Oppositely, solid research found no association between species host phylogeny and microbiome structure, thus disproving host-specificity ^{34,95}. This is still a topic under investigation, as proof for each hypothesis (host-specificity and species-specificity) have been found. In the present study, the differences seen across prokaryomes arise from species host-specific occurrences of OTUs. As seen in **Figure 3.14**, the seven most abundant OTUs in the dataset further suggest host-specificity, as their abundance in the remainder sponges is close to zero. Closely related sponge species harbouring dissimilar prokaryomes were noted in the non-metric MDS analysis (**Figure 3.8**) (scaling stress = 0.1116067, ANOSIM, R=1; ADONIS, P<0.001). As inter-specificity was notably seen, intra-species dissimilarities were noted as well, mainly in *P. fictitius* (MRPP MW-CD = 0.3633), as three replicates were placed apart from the core cluster of samples of this species. Indeed, as depicted in a Bray-Curtis dendrogram (**Figure 3.11**), samples 94 and 98 were clustered outside of the main group as did sample 105, although closer to the latter. This variability was not of a biogeographical nature: sample 105 was the only one collected in Armação de Pêra, among the closest prokaryomes (96 and 97) and the ‘outlier’ samples 94 and 98. In the case of the remaining *P. fictitius* samples, no such dissimilarity was noted, indicating that most probably this was due to the unique presence of *Gemmatimonadetes* in the three outlying replicates, further assessed in **4.4**. With no differentiating methodological or environmental parameters, these seem like interesting patterns of prokaryotic diversity among *P. fictitius* samples. Altogether, beta diversity analyses performed here point towards host species-specific shaping of sponge prokaryomes, independently of host phylogenetic relationships.

4.4. Taxonomy across prokaryomes

The known limitations of 16S rRNA gene profiling are mainly attributed to the biases in classifying not-yet cultured microbes and the fact that this gene is hypervariable and frequently inconsistent throughout prokaryotic phylogeny (contradicting phylogenomic, this is, whole-genome-based phylogeny, evidence) ^{65,68}. For more than 10 years,

however, this has been, and currently is, the gold-standard for prokaryotic diversity barcoding^{63,66}.

16S-based NGS techniques were first applied in the context of the vertebrate gut microbiome by means of clone libraries in 2008^{96,97}. Caporaso *et al.* laid the foundations towards an universal bacterial and archaeal primer pair (515F/806R) by adapting the latter primers to the context of Illumina sequencing technologies, ceasing its increasing sequencing power (this is, the so-called depth of sequencing, or the total number of read-in sequences)⁶⁶. Thought to initially be an effective barcode in differentiating taxonomic diversity at the phylum level, this primer pair was recently shown to have a limited ability to detect certain bacterial clades, such as SAR11 (α -Proteobacteria), SAR86 (γ -Proteobacteria) and other proteobacterial lineages, as well as of some lineages in the archaeal phyla *Crenarchaeota* and *Thaumarchaeota*^{52,98}.

In the present study, each of the sponge species was found to host a taxonomically distinct prokaryotic community. Starting with *C. viridis*, a prokaryome of high α -, γ - and unclassified proteobacterial abundance was noted. Whereas the first proteobacterial class was associated mainly to the orders *Sphingomonadales* and then to *Rhodospirillales*, the phylotypes within γ -Proteobacteria were shown to belong mainly to the *Pseudomonadales* order (see **Figure 3.12**), rarely found in marine sponges^{61,99}. However, species of the genus *Pseudomonas* have already been isolated from marine sponges and demonstrated to have high antibacterial and cytotoxic activities¹⁰⁰. Members of the order *Sphingomonadales* have been identified as dominating the prokaryome within *Hymeniacidon sinapium* (Family Halichondriidae) with only two OTUs, it was also documented in near-zero abundances off the coast in Indonesia, which contrasted with a 20% abundance of the order in seawater⁶¹.

Present in abundances reaching about 30% on average across all sponges studied here, γ -Proteobacteria is frequently found in LMA sponges in abundances below or near 25%^{41,93,101,102}. Regarding α -proteobacterial diversity, *Rhodospirillales* has been cultured in 1976 by Imhoff and Stöhr from *Euspongia officinalis* (currently accepted as genus *Spongia* in WoRMS, consulted on the 7th of August 2015, Order Irciniidae) collected in coastal waters of former Yugoslavia¹⁰³. *Rhodospirillales* was one of the most dominant orders in *Stylissa carteri* (Order Halichondrida), an LMA sponge collected in the Red Sea¹⁰⁴. This order is present in varying abundances throughout the world: in the Red Sea (*Haliclona tubifera*, Order Haplosclerida) it reached similar abundances as those

found here. It also dominated the prokaryomes of *Cliona sp.*, *Halichondria okadai* (Order Suberitida) and *Spirastrella panis* (Order Clionaida) sampled in the South Korean Sea ^{61,105}.

The prokaryome here depicted for *C. viridis* coincides, at the phylum level, with the one described by Blanquer *et al.* regarding the same species, where Proteobacteria dominated with a relative abundance of almost 100% ⁴¹. At the class level, it is seen that the sponge herein depicted is much more diverse, as two proteobacterial classes are seen, whereas the previous study demonstrated an α -Proteobacteria-dominated prokaryome. Not only was each class seen to hold two dominating orders, but further diversity was seen to be encrypted in unclassified proteobacterial and bacterial phylotypes. Adjoined with unclassified γ -proteobacterial OTUs, the latter clades account to as much as 50% of the total *C. viridis* prokaryome.

In the case of *C. celata*, a strikingly large consortium of unclassified bacterial phylotypes was found, and most of the remaining taxonomic composition was attributed to unclassified Proteobacteria and γ -Proteobacteria. The only published prokaryome composition for a sponge of the same species was authored by Jeong and colleagues and describes the sponge as being dominated (around 55%) by an unclassified member of the EC94 order (β -Proteobacteria) ⁶¹. Although the latter study utilized different primers, targeting the V1-V3 region of the 16S rRNA gene, the most recent version of the Greengenes database was used as reference, making the taxonomic assignment output comparable with the one presented here ⁷⁰. The remaining classes (γ -Proteobacteria and unclassified Proteobacteria) were observed in similar proportions as in Jeong *et al.*, with exception to the unclassified bacterial consortia herein found ⁶¹.

The prokaryome within *P. fictitius* presented abundances of unclassified Proteobacteria in magnitudes from around 50% to almost 70%, with three exceptions. The presence of γ -Proteobacteria was noted in relative abundances from around 15% to 40%, with the same exceptions as before. Other taxa, including α -Proteobacteria, were seen in abundances lower than 5%, therefore suggesting a highly specific and possibility specialized prokaryome. This is in agreement with previous observations of the culturable prokaryome of *Phorbas tenacior*, where 96% of the isolates were classified as Proteobacteria ^{14,89}. In the present study, the order *Oceanospirillales* was found to abound among the classified γ -Proteobacteria. This order was also among the most dominant in *S. carteri*, collected in the Red Sea, as was the case for *Sphingomonadales*

in *C. viridis*¹⁰⁴. It is known to be prevalent in seawater and to respond positively to hydrocarbon presence in seawater by degrading it¹⁰⁶. Further, a role in nitrogen cycling for *Oceanospirillales* phylotypes has been suggested, as denitrification key enzymes like nitrite reductase, nitrous-oxide reductase and nitric oxide reductase subunit B were found to be produced by pelagic members of this order off of West Java, Indonesia¹⁰⁷.

In the three outlying *P. fictitius* samples an augmented presence of *Gemmatimonadetes* (from 7.53% to 33.54%) was noted. Indeed, *Oceanospirillales* was noted in proportions reaching three times as the remainder two outlying samples. Order *Gemm-4* was observed in notable abundances in one of the outlying *P. fictitius* samples, accounting for about 10% of the prokaryome in the remaining two. Noted in abundances usually never higher than 10% in several sponges, the ecology of this recently described order remains largely unknown^{47,102,107–109}.

Regarding the prokaryome found within *D. fragilis*, proteobacterial abundances were slightly diminished and Bacteroidetes were noted to increase to as much as 18.79%. The proteobacterial component was mostly composed of unknown members, followed by α - and γ -Proteobacteria. While the latter was present in as much as 21.99% of the general profile, α -Proteobacteria corresponded to 2.97% to 12.4% of the total community. Whereas the α -Proteobacteria were dominated by unclassified members, *Pseudomonadales* accounted for most of the γ -proteobacterial diversity in *D. fragilis*, followed by a small consortium of unclassified OTUs in the same class, reaching up to 4.04% in relative abundance. Regarding the functional potential of *Pseudomonadales* presence in sponges, as mentioned before, antibacterial and cytotoxic activities may putatively be the cause for such association¹⁰⁰. As for the diversity within Bacteroidetes, it was fully composed of the order *Flavobacteriales*, a pelagic group of bacteria with a worldwide distribution¹¹⁰. Marine *Flavobacteria* possess aerobic photo-organoheterotrophic metabolism, coupling proteorhodopsin with peptidase production, as well as polysaccharide synthesis and degradation, which could serve as basis for a symbiotic relationship^{111,112}.

In the present work, highly distinct and quite diverse prokaryotic consortia were found for all surveyed sponge species. Starting with *C. viridis*, an almost equally balanced consortium of unclassified Proteobacteria and γ -Proteobacteria, as well as *Sphingomonadales* was found, with lesser abundances of unclassified *Bacteria* and *Rhodospirillales*. *C. celata* presented a strikingly dominant presence of unclassified

bacterial phylotypes followed by descending abundances of unclassified Proteobacteria and γ -Proteobacteria. In *P. fictitius*, a large consortium of Proteobacteria was shown to be dominated by unclassified proteobacterial OTUs, followed by *Oceanospirillales* phylotypes. Within the same sponge, three samples differed greatly from the remainder, as abundances in *Gemmatimonadetes* rose unexpectedly. Similarly to *P. fictitius*, high abundances of unclassified Proteobacteria were found in *D. fragilis*, followed by *Pseudomonadales* and unclassified α -Proteobacteria. Further, a small consortium of *Flavobacteriales* was found in considerable amounts in this sponge. The exploration of poorly studied sponges' prokaryomes thus revealed a host-specific enrichment of certain bacterial classes beforehand rarely noticed in marine sponges.

4.5. Sponge- and species-specific associations

An initial analysis of total numbers of OTU shared among sponge species and those specific to sponge species showed high numbers of species-specific phylotypes. Indeed more than 900 and 800 species-specific phylotypes were unique to *C. celata* and *P. fictitius*, respectively. Interestingly, species belonging to the genus *Cliona* showed the least number of shared OTUs (**Figure 3.15**). When the least abundant OTUs were excluded from the analysis, a different abundance pattern was noted as, for example, the species-specific consortium of *P. fictitius* decreased to 4.8% of its initial richness (**Figure 3.17**). Each sponge species was notably populated by a pool of specific and quite abundant OTUs (**Figure 3.14**). This preference is likely indicative of species-specificity of a microbe towards a sponge host. *C. viridis* showed the largest consortium of OTUs with high abundances (**Figure 3.2**). These highly abundant OTUs included three *Rhodospirillales* (α -Proteobacteria) phylotypes, one unclassified member of the latter class and an unclassified Proteobacteria. Frequently found with a high abundance in seawater, *Rhodospirillales* has not been observed as part of a species-specific sponge prokaryome^{104,113,114}.

Two main OTUs were seen to stand out in the *P. fictitius* prokaryome. These were classified as one *Oceanospirillales* phylotype and an unclassified Proteobacteria, which accounted for the most sequences assigned to one OTU across the dataset (68,809 sequences). Although γ -Proteobacteria has already been detected in the cultured prokaryome of *Phorbas tenacior*, *Oceanospirillales* was only detected in high amounts in *S. carteri*^{14,104}.

In *D. fragilis*, two unclassified α -Proteobacteria and a member of the EC94 order (β -Proteobacteria) were the most prevalent phylotypes. Whereas β -Proteobacteria was seen to populate no more than 0.5% of the *D. fragilis* prokaryome, this phylotype presents here more than 4,000 assigned sequences, which are further mostly exclusive to this sponge. Members of the EC94 order were found to predominate in *Haliclona cinerea* (Order Haplosclerida) and reach up to 56.4% in relative abundance in *C. celata*⁶¹. At rather low abundances, a member of this group populated deep-sea irdniids (*Irciniidae*), totalling almost 100% of the observed β -proteobacterial abundance in the analysed specimens¹¹⁵. Members of the EC94 order were found in the deep-sea sponge *Inflatella pellicula* by Jackson and colleagues and explained almost all the β -proteobacterial abundance across the two analysed individuals¹¹⁵. In the South Korean Sea, high abundances of this order were noted in *Haliclona cinerea* and *C. celata* (92.2% and 56.4% of the total prokaryotic communities)⁶¹. The lack of statistical robustness in sponge sampling was however a considerable drawback in this study, in which regions V5-V6 were used along with the most recent version of the Greengenes database. No insights have been gained thus far into the metabolic potential within this order, which has not yet been described outside of the sponge microenvironment.

The second most abundant OTU in the dataset, accounting for 50,951 sequences, was found in *C. celata* and its classification was not possible beyond the domain level using the Greengenes database. Phylogenetic inference using the latest SILVA database (version 123, dated to July 2015) showed this highly abundant OTU may belong to a paraphyletic branch of the Clostridia class within Firmicutes (**Figure 3.18**). Also lacking taxonomical classification at the same level, two more OTUs were found almost exclusively in this sponge. Furthermore, two unclassified proteobacterial phylotypes, as well as an unclassified γ -Proteobacteria were found in this prokaryome. One study describing prokaryotic communities within *C. celata* found a high abundance of β -Proteobacteria, whereas one unclassified proteobacterial phylotype was seen to reach abundances of 6.4% and an exemplar of the *Pseudomonadales* order (γ -Proteobacteria) reached 9%.

Compelling evidence was found for species-specificity among the 50 most abundant OTUs, as these were notably not dominant in more than one sponge species. All OTUs present in this dataset have been assigned more than 191 sequences. *C. viridis* possessed the largest set of species-specific OTUs with high abundances (5,645 to 11,328

sequences), corresponding to slightly higher evenness values. The two most abundant OTUs within *D. fragilis* were found to have 4,224 and 4,019 assigned sequences. The prokaryome within *P. fictitius* encompassed the first and third most abundant OTUs across the dataset, which altogether accounted for 95,804 sequences. A set of less abundant phylotypes (ranging from about the 20th to 30th most abundant OTU, as seen in **Figure 3.14**) are further seen to only be present in samples belonging to *P. fictitius*. Furthermore, for *C. celata*, a core of four OTUs was also found to present similar patterns of host-specificity along with the second most abundant OTU (Otu000602).

Within the 50 most abundant OTUs, 51,531 sequences assigned to 4 OTUs were found to be unclassifiable at phylum-level. Further 10 OTUs accounted for 83,503 proteobacterial sequences unclassifiable at the class level. Altogether, these 14 OTUs represent 46.36% of the sequences present in the normalized dataset. It has been demonstrated by Apprill and colleagues that the primers utilized by the Earth Microbiome Project tend to suppress abundances of some clades within α - γ - and other proteobacterial classes⁹⁸. Classification of OTUs is further inherently dependent on amplicon size, here at about 100bp (base pairs), which might be another factor contributing to the noted limitation in classifying several OTUs using the Greengenes database.

4.6. Exploration for the ‘microbial dark matter’ within the sponge prokaryome

The generation of a singleton-only dataset, is usually a controversial technique. Inaccurate base calling during sequencing, or misses in error detection by pre-processing tools are among the most common bioinformatics problems with regard to microbial ecology. Singletons and sometimes doubletons are therefore mostly excluded from amplicon datasets during pre-processing steps⁴⁷. However, singletons have been noted to represent up to 81% of the prokaryome within some sponges³⁴.

In the scope of the present study, singleton OTUs were considered given the nature of their generation. Here, OTU generation was accomplished taking all 1,226 samples belonging to the Consortium for Sponge Microbiology into account, where true singletons found in this entire dataset had been previously excluded. As such, OTUs seen in the present singleome, that is the consortium of all singleton OTUs in the dataset, were indeed not technical singletons. These were in fact single reads of known

OTUs present in other samples in unknown amounts, making the present singletome more robust concerning the reliability found for its enclosed taxonomical information.

Here, previously observed species-specific profiles for phylum-level prokaryotic taxonomy were not noted. (**Figure 3.19**). The total singleton abundances (on average) in each sponge represented here almost half of the total OTU richness (44.3%, 2,561 OTUs) found across the non-normalized data. Indeed, Proteobacteria was seen to predominate in relative abundance across all samples, usually comprising more than 50% of the total singletome, while Bacteroidetes and Cyanobacteria were reached 20.97% and 44.23%. Further, *Planctomycetes* and the archaeal phylum *Crenarchaeota* were also represented in a greater magnitude regarding OTU richness. On the opposite, the percentage of unclassified bacterial phylotypes was seen to decrease heavily across the entire singletome in comparison with the whole dataset.

Exploring the microbial ‘dark matter’ within a wide range of environments has therefore far provided new information regarding previously hidden phylogenetic and metabolic potential ¹¹⁶. With the unusual depths of sequencing hereby reached, and regarding the panorama of published works in this matter, it was possible to take the analysis of the non-transformed singletome into account as a significant repository of diversity for the sponge prokaryome, as previously suggested ²¹. Within the total dataset, this is non-normalized to 11,203 sequences, the singletome presented high numbers of phylotypes, totalling 2,561 OTUs, despite not following the phylum-level taxonomical trends presented in the normalized dataset or showing any signs of host-specificity at phylum-level.

4.7. Phylogenetic inference of abundant but unclassified OTUs

Given the time-lapse between the present study and the pre-processing of EMP datasets, and the hence possible increased number of novelties in the available 16S-based online databases, an attempt was made towards positioning the 20 most abundant representative sequences of unclassified bacterial taxa. These OTUs were originally classified according to the 2013 version of the Greengenes reference database, which could have been the cause for the unexpected abundances of unknown bacterial phylotypes. As such, these were realigned with the latest SILVA SSU (small sub-unit) and LSU (large sub-unit) database, version 123, released on July 2015, for phylogenetic re-assessments.

Within α -Proteobacteria, Otu001131 was found to be an exemplar of the order *Rhizobiales_3*, placed close to the family *Beijerinckiaceae*. Members of this order were already seen in sponges of the genus *Discodermia* (Order Tetractinellida) that are found at depths ranging from 24m to 161m¹¹⁷. Another α -proteobacterial OTU was assigned to the *Rickettsiales* order, which is known to dominate the prokaryomes within deep-sea sponges of the *Inflatella* genus¹¹⁵. This OTU (Otu002013) was further found to be related to mitochondrial ancestors (*Mitochondria* family, not shown in **Figure 3.18**). This classification could, however, have been due to the failure of mitochondria-excluding algorithms within the mothur pipeline.

One OTU was found to be a member of the *Cytophagia* order within Bacteroidetes. Whereas Haroim & Costa⁴⁷ found two OTUs and two associated sequences derived from the analysis of two sympatric ircniid sponges and Gladkikh *et al.* found one phylotype with 67 corresponding sequences in two endemic sponges from Lake Baikal, a total of 314 sequences were found for a phylotype assigned to the *Cytophagia* order in the present study¹¹⁸.

As for Acidobacteria, Subgroup 6 has already been noted as part of the prokaryome within the deep-sea sponge *Stelletta normani* (Order Tetractinellida), predominating along with phylotypes classified as belonging to the SAR202 class of Chloroflexi¹³. In *Xestospongia* spp. (Order Haplosclerida), evidence for host-specific speciation within this acidobacterial order was found³⁵.

One OTU was classified as belonging to a recently proposed phylum, Cand. Nitrospinae⁸⁷. This former paraphyletic branch of δ -Proteobacteria was proven not to belong to that phylum, following genomic evidence¹¹⁹. It was therefore proposed that the '*Nitrospinaceae*' clade should now be considered as a phylum⁸⁷. This clade has so far not been observed in poriferans and is here noted with the presence of 2,128 sequences.

Currently known to be highly paraphyletic, the class Clostridia of Firmicutes is now considered as a putative phylum, with unknown position in the general prokaryotic phylogenetic tree¹²⁰. A member of the *Clostridiaceae_1* family (*Clostridiales* order), Otu001349, was found in abundances of 414 sequences. The remaining six members of *Clostridiales* were classified as exemplars of the *Eubacteriaceae* family. This order has been detected in sympatric marine sponges off the coast of the Algarve and in *Theonella*

swinhoei (Order Tetractinellida), in the Red Sea ¹²¹. Among the six OTUs assigned to the Eubacteriaceae is the second most abundant OTU of the normalized dataset (Otu000602, 99,553 sequences in the non-normalized dataset).

As for Chloroflexi, one OTU was assigned to the class *Anaerolineae*, known to be frequently present in sponges. In *Xestospongia testudinaria* (Order Haplosclerida), this class was noted to occupy a ‘significant portion’ of the total number of reads obtained from the sponge ¹⁰⁴. Montalvo *et al.* noted patterns of speciation of this phylum across geographical settings for the same genus of sponge ¹⁰². Off the coast of Algarve, southern Portugal, this clade was seen to populate the ‘pan bacteriome’ (i.e., the microbial ensemble of all OTUs found across individuals of a certain sponge species, as opposed to the core bacteriome, which includes only the ones believed to be the true symbionts, as they are present across all individuals) of *Sarcotragus spinosulus* (Order Dictyoceratida), accounting for a third of the total Chloroflexi OTU count ⁴⁷. The remaining OTU was classified as a member of JG-30-KF-CM66, an undetected class in poriferan prokaryomes. Hug and colleagues have published the only observation of this class, where it was detected in anaerobic sediment extracted from an aquifer adjacent to the Colorado River in the USA¹²².

Two previously unclassified phylotypes were reclassified as Actinobacteria and further as belonging to class *Acidimicrobiia*. An unclassified member of the Sva0996 Marine Group order within this class was shown to be present in about 5% of the prokaryome within *X. testudinaria*, on average ³⁵. Here, the observed phylotypes were seen to total 14,544 sequences within the non-transformed dataset, achieving a relative abundance of about 2.61%.

Lastly, two cyanobacterial OTUs associated with the genus *Prochlorococcus* (Subsection 1 Family 1) were found. Corresponding to 809 sequences in total in the present study, the genus *Prochlorococcus* is known to be a widespread marine microbe, highly similar to “Cand. *Synechococcus spongiarum*”, an uncultured sponge-specific species ¹²³.

A reclassification of the Greengenes-unclassified consortia among the sampled sponges according to the latest SILVA Reference Database enabled further insights into the taxonomic profiles of these prokaryomes. Specifically, the previously dominant clades of unclassified OTUs in *C. celata*, reaching up to 87.72% of the total prokaryotic

abundance, are most likely related to the Clostridia (Firmicutes) ¹²⁰. An unclassified bacterial phylotype (Otu000602, represented by 99,553 sequences in the whole dataset) prevailed in high abundances in *C. celata* (**Figure 3.14**), surely defining the phylum-level relative abundance profiles in **Figure 3.10**. The search for a reclassification of these unclassified sequences was therefore successful, mainly due to the newly provided insights into the phylum-level taxonomical profiles of these prokaryomes. Further insight is needed with regard to the most abundant OTUs, attributed for now to the order *Clostridia* (Firmicutes). This situation should be resolved upon upcoming releases of the SILVA reference database.

4.8. Phylogenetic relationships among hosts and prokaryomes

Both clionaidan sponges show an unresolved *cox1*-based taxonomy, as does *D. fragilis* (**Figure 3.21**). Both the first sponges have previously exhibited patterns of intra-specific, high phylogenetic complexity ^{124,125}. However, although the Hadromerida clade was recently shown to be highly polyphyletic and thus eliminated as a taxonomical group, Clionaida is a well-established new monophyletic order ⁵. Here, a *cox1* phylogeny resulted in poor differentiation between *C. celata* and *C. viridis*, contrasting the relative ease with which both species can be recognized using simple morphological criteria (**Figure 3.20**). The prokaryomes associated to these sponges were strikingly set apart, eliminating the possibility of a positive linear relationship across prokaryome structuring and host phylogeny (**Figure 3.7**). The three sampled exemplars of *D. fragilis* did not cluster, but were set the furthest apart from the remaining sponges (**Figure 3.21**). *D. fragilis* belongs to the subclass Keratosa, whereas both clionids and *P. fictitius* are part of the subclass Heteroscleromorpha *sensu* Cárdenas *et al.* ^{5,126}. The phylogenetic tree found in the present study follows a similar trend, in spite of the observed *cox1* variability within *D. fragilis*. As such, and considering the dissimilarities among prokaryomes in **Figure 3.7**, no evidence was found of a host phylogeny-driven structuring of prokaryomes among the studied sponges. Specific prokaryomes can be identified within each sponge species regardless of phylogeny. As suggested by Blanquer *et al.*, such differences, further noted between LMA and HMA sponges, may occur due to morphological differences in each sponge's mesohyl ⁴¹. The density and composition (e.g. presence of spongin fibres or calcium carbonate spicules) may therefore be a stronger selection factor for prokaryomes than sponge phylogeny, which frequently is not related to the latter characters.

C. celata and *C. viridis* follow strikingly different life styles (**Figure 1.3**). While the former is an excavating sponge, the latter is known to live erect. The vertical height attained by a sponge and its degree of contact with planktonic microbes, or the level of contact with marine sediments, may therefore define prokaryome composition and structuring. Further, the rate of filtration of seawater could differ between both, given their lifestyles. As such, the strikingly different prokaryotic consortia found in *C. celata* may be the product of a more stable colonization process, in which the decreased water velocity allows for an easier attachment of microorganisms.

Aside from host morphology and lifestyle, metabolism and physiology also play a role in microbiome specificity. Primary metabolic end-products such as ammonia are known to serve as a nitrogen source for prokaryotic communities and therefore shape sponge symbiont communities⁴. Other primary metabolic processes provide additional organic compounds that may further contribute to such processes, but in-depth investigations of these host-symbiont interactions are needed.

The mechanisms defining symbioses between sponges and microbes include the ‘mimicking’ of eukaryotic-like proteins²², a phenomenon confirmed by means of single-cell genomics as being present among members of symbiotic phyla Poribacteria^{22,25}. The shaping of prokaryomes therefore depends on complex processes, but this symbiont selection process may depend on further unknown mechanisms. Also, sponge feeding trends may further drive symbioses by providing prokaryotic communities within the host with important organic compounds derived from preyed-upon plankton.

In summary, a *cox1*-based host phylogeny provided valuable evidence for a lack of correlation between host evolution and prokaryome structuring. Therefore, here, prokaryome structuring follows a host-specific specificity trend, not established by gradual increase/decrease in host phylogenetic relatedness. The establishment of symbiosis between poriferans and microbes is a complex and largely unknown process. Given the results observed here, further in-depth perspectives onto the molecular and chemical factors modulating this host-specificity are needed. Valuable insights into the mechanisms contributing to the sponge host-microbe symbiotic relationship will likely provide an answer to whether host-specificity of the microbiome is a factor driving holobiont evolution or is merely an evolutionary neutral result of sponge morphological evolution.

5. Conclusion

In the present work, four sympatric sponges with different phylogenetic levels of relatedness were analysed for their prokaryomes. Host-specific patterns of prokaryome composition were found, with no correspondence with *cox1*-based host phylogeny. Further, host-specific cores of highly abundant OTUs varying in size were found.

General patterns of proteobacterial dominance were found, although with class-level relative abundance differences among sponges. Namely, while in *C. viridis* a co-dominance of γ -, α - and unclassified Proteobacteria phylotypes was seen, *P. fictitius* and *D. fragilis* were notably populated firstly by unclassified and then by γ -Proteobacteria. However, for *C. celata*, high proportions of unclassified bacterial phylotypes were found. Posteriorly, these were re-aligned and reclassified to known phyla. As such, it was found that these dominating OTUs may mostly be assigned to a so-far putative phylum derived from a branch of *Clostridia* (Firmicutes). Conjoining alpha-diversity analyses and a taxonomy overview, it was posed that all surveyed sponges may be LMA taxa, due to low Shannon diversity indices and a general absence of HMA-indicative phyla.

By isolating the singletome within these prokaryomes, no evidence was found for host-specificity, as previously noted in the singleton-excluded dataset. Surprisingly, most of the diversity found within the rare prokaryome was known, as opposed to previous reports on taxonomically exotic microbial dark matter. Finally, rarefaction analysis indicates that sequencing depth was insufficient for all sponges but *C. viridis*, as this was the only sponge to reach a coverage 'plateau'. Further coverage of these prokaryomes may therefore unveil added taxonomic and host-specificity patterns, for example. Trends of prokaryome composition and structure were found for four previously understudied marine sponges. Further investigation of these sponges would benefit from metagenomics, as this technique may help to determine the observed host-specificity patterns at the functional level. Moreover, this technique would possibly validate the re-classification step used here to identify previously unknown bacterial phylotypes. Understudied sponges therefore hold the potential for unlocking further taxonomic diversity of the global sponge microbiome and for testing relevant sponge prokaryome-specific ecological hypotheses.

6. Bibliography

1. Hooper, JN A and Soest, RWM Van (2002). *Systema Porifera: A Guide to the Classification of Sponges. Syst. Porifera . A Guid. to Classif. Sponges*, Kluwer Academic/Plenum Publishers, New York.
2. Muscente, AD, Marc Michel, F, Dale, JG and Xiao, S (2015). Assessing the veracity of Precambrian ‘sponge’ fossils using in situ nanoscale analytical techniques. *Precambrian Res.* **263**: 142–156.
3. Peterson, KJ and Butterfield, NJ (2005). Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record. *Proc. Natl. Acad. Sci. U. S. A.* **102**: 9547–9552.
4. Hentschel, U, Usher, KM and Taylor, MW (2006). Marine sponges as microbial fermenters. *FEMS Microbiol. Ecol.* **55**: 167–77.
5. Morrow, C and Cárdenas, P (2015). Proposal for a revised classification of the Demospongiae (Porifera). *Front. Zool.* **12**: 1–27.
6. Simpson, TL, Langenbruch, P-F and Scalera-Liaci, L (1985). Silica spicules and axial filaments of the marine sponge *Stelletta grubii* (Porifera, Demospongiae). *Zoomorphology* **105**: 375–382.
7. Brusca, RC and Brusca, GJ (2003). Perspectives on Invertebrate Phylogeny. *Invertebrates*: 16.
8. Leys, SP, Yahel, G, Reidenbach, M a, Tunnicliffe, V, Shavit, U and Reiswig, HM (2011). The sponge pump: the role of current induced flow in the design of the sponge body plan. *PLoS One* **6**: e27787.
9. Hardoim, C and Costa, R (2014). Microbial Communities and Bioactive Compounds in Marine Sponges of the Family Irciniidae—A Review. *Mar. Drugs* **12**: 5089–5122.
10. Hentschel, U, Piel, J, Degnan, SM and Taylor, MW (2012). Genomic insights into the marine sponge microbiome. *Nat. Rev. Microbiol.* **10**: 641–54.
11. Marshall, CR (2006). Explaining the Cambrian ‘Explosion’ of Animals. *Annu. Rev. Earth Planet. Sci.* **34**: 355–384.
12. Taylor, MW, Radax, R, Steger, D and Wagner, M (2007). Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. *Microbiol. Mol. Biol. Rev.* **71**: 295–347.
13. Kennedy, J, Flemer, B, Jackson, S a, Morrissey, JP, O’Gara, F and Dobson, ADW (2014). Evidence of a putative deep sea specific microbiome in marine sponges. *PLoS One* **9**: e91092.

14. Dupont, S, Carré-Mlouka, A, Descarrega, F, Ereskovsky, A, Longeon, A, Mouray, E, *et al.* (2013). Diversity and biological activities of the bacterial community associated with the marine sponge *Phorbastenia tenacior* (Porifera, Demospongiae). *Letts. Appl. Microbiol.* **58**: 42–52.
15. Chevaldonné, P, Pérez, T, Crouzet, J-M, Bay-Nouailhat, W, Bay-Nouailhat, A, Fourt, M, *et al.* (2014). Unexpected records of ‘deep-sea’ carnivorous sponges *Asbestopluma hypogea* in the shallow NE Atlantic shed light on new conservation issues. *Mar. Ecol.* **36**.
16. Vacelet, J and Donadey, C (1977). Electron Microscopy Study of the Association Between Some Sponges and Bacteria. *J. Exp. Mar. Biol. Ecol.* **30**: 301–314.
17. Hentschel, U, Horn, M, Friedrich, AB, Wagner, M and Moore, BS (2002). Molecular Evidence for a Uniform Microbial Community in Sponges from Different Oceans. *Appl. Environ. Microbiol.* **68**: 4431–4440.
18. Webster, NS, Negri, AP, Munro, MMHG and Battershill, CN (2004). Diverse microbial communities inhabit Antarctic sponges. *Environ. Microbiol.* **6**: 288–300.
19. Thacker, W and Starnes, S (2003). Host specificity of the symbiotic cyanobacterium *Oscillatoria spongelliae* in marine sponges, *Dysidea* spp. *Mar. Biol.* **142**: 643–648.
20. Fieseler, L, Horn, M, Wagner, M and Hentschel, U (2004). Discovery of the Novel Candidate Phylum ‘Poribacteria’ in Marine Sponges. *Appl. Environ. Microbiol.* **70**: 3724–3732.
21. Taylor, MW, Tsai, P, Simister, RL, Deines, P, Botte, E, Ericson, G, *et al.* (2013). ‘Sponge-specific’ bacteria are widespread (but rare) in diverse marine environments. *ISME J.* **7**: 438–43.
22. 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.
23. 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.
24. 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**: 1557–67.
25. Nguyen, MTHD, Liu, M and Thomas, T (2013). Ankyrin-repeat proteins from sponge symbionts modulate amoebal phagocytosis. *Mol. Ecol.* **23**: 1635–45.

26. 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**: 2287–300.
27. Hallam, SJ, Konstantinidis, KT, Putnam, N, Schleper, C, Watanabe, Y, Torre, D, *et al.* (2006). Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *PNAS* **103**.
28. Radax, R, Hoffmann, F, Rapp, HT, Leininger, S and Schleper, C (2012). Ammonia-oxidizing archaea as main drivers of nitrification in cold-water sponges. *Environ. Microbiol.* **14**: 909–923.
29. Fan, L, Reynolds, D, Liu, M, Stark, M, Kjelleberg, S and Webster, NS (2012). Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *PNAS* **109**.
30. Ribes, M, Jiménez, E, Yahel, G, López-Sendino, P, Diez, B, Massana, R, *et al.* (2012). Functional convergence of microbes associated with temperate marine sponges. *Environ. Microbiol.* **14**: 1224–1239.
31. Gloeckner, V, Wehrl, M, Moitinho-silva, L, Gernert, C, Schupp, P, Pawlik, JR, *et al.* (2014). The HMA-LMA Dichotomy Revisited: an Electron Microscopical Survey of 56 Sponge Species. *Biol. Bull.* **227**: 78–88.
32. Hardoim, CCP, Esteves, AIS, Pires, FR, Gonçalves, JMS, Cox, CJ, Xavier, JR, *et al.* (2012). Phylogenetically and spatially close marine sponges harbour divergent bacterial communities. *PLoS One* **7**: e53029.
33. Schmitt, S, Angermeier, H, Schiller, R, Lindquist, N and Hentschel, U (2008). Molecular microbial diversity survey of sponge reproductive stages and mechanistic insights into vertical transmission of microbial symbionts. *Appl. Environ. Microbiol.* **74**: 7694–708.
34. Giles, EC, Kamke, J, Moitinho-Silva, L, Taylor, MW, Hentschel, U, Ravasi, T, *et al.* (2012). Bacterial community profiles in low microbial abundance sponges. *FEMS Microbiol. Ecol.* **83**: 232–41.
35. Moitinho-Silva, L, Bayer, K, Cannistraci, C V, Giles, EC, Ryu, T, Seridi, L, *et al.* (2014). Specificity and transcriptional activity of microbiota associated with low and high microbial abundance sponges from the Red Sea. *Mol. Ecol.* **23**: 1348–63.
36. Caporaso, JG, Lauber, CL, Walters, W A, Berg-Lyons, D, Huntley, J, Fierer, N, *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**: 1621–1624.

37. Simister, RL, Deines, P, Botté, ES, Webster, NS and Taylor, MW (2012). Sponge-specific clusters revisited: a comprehensive phylogeny of sponge-associated microorganisms. *Environ. Microbiol.* **14**: 517–24.
38. Naim, MA, Morillo, J A, Sørensen, SJ, Waleed, AA-S, Smidt, H and Sipkema, D (2014). Host-specific microbial communities in three sympatric North Sea sponges. *FEMS Microbiol. Ecol.* **90**: 390–403.
39. Cárdenas, C, Bell, JJ, Davy, SK, Hoggard, M and Taylor, MW (2014). Influence of environmental variation on symbiotic bacterial communities of two temperate sponges. *FEMS Microbiol. Ecol.* **88**: 516–527.
40. Webster, NS, Taylor, MW, Behnam, F, Lückner, S, Rattei, T, Whalan, S, *et al.* (2009). Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. *Environ. Microbiol.* **12**: 2070–82.
41. Blanquer, A, Uriz, MJ and Galand, PE (2013). Removing environmental sources of variation to gain insight on symbionts vs. transient microbes in high and low microbial abundance sponges. *Environ. Microbiol.* **15**: 3008–3019.
42. Hill, M, Hill, A, Lopez, N and Harriott, O (2005). Sponge-specific bacterial symbionts in the Caribbean sponge, *Chondrilla nucula* (Demospongiae, Chondrosida). *Mar. Biol.* **148**: 1221–1230.
43. Easson, CG and Thacker, RW (2014). Phylogenetic signal in the community structure of host-specific microbiomes of tropical marine sponges. *Front. Microbiol.* **5**: 1–11.
44. Pita, L, Turon, X, López-Legentil, S and Erwin, PM (2013). Host rules: spatial stability of bacterial communities associated with marine sponges (*Ircinia* spp.) in the Western Mediterranean Sea. *FEMS Microbiol. Ecol.* **86**: 268–76.
45. Erwin, PM, López-Legentil, S, González-Pech, R and Turon, X (2012). A specific mix of generalists: bacterial symbionts in Mediterranean *Ircinia* spp. *FEMS Microbiol. Ecol.* **79**: 619–37.
46. Reveillaud, J, Maignien, L, Eren, M a, Huber, J a, Apprill, A, Sogin, ML, *et al.* (2014). Host-specificity among abundant and rare taxa in the sponge microbiome. *ISME J.* **8**: 1198–209.
47. Hardoim, CCP and Costa, R (2014). Temporal dynamics of prokaryotic communities in the marine sponge *Sarcotragus spinosulus*. *Mol. Ecol.* doi:10.1111/mec.12789.
48. Weisz, JB, Lindquist, N and Martens, CS (2008). Do associated microbial abundances impact marine demosponge pumping rates and tissue densities? *Oecologia* **155**: 367–76.

49. Bayer, K, Moitinho-Silva, L, Brümmer, F, Cannistraci, C V, Ravasi, T and Hentschel, U (2014). GeoChip-based insights into the microbial functional gene repertoire of marine sponges (high microbial abundance, low microbial abundance) and seawater. *FEMS Microbiol. Ecol.*
50. Webster, NS and Blackall, LL (2009). What do we really know about sponge-microbial symbioses? *ISME J.* **3**: 1–3.
51. Webster, NS and Taylor, MW (2012). Marine sponges and their microbial symbionts: love and other relationships. *Environ. Microbiol.* **14**: 335–46.
52. Gilbert, J a, Jansson, JK and Knight, R (2014). The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**: 69.
53. Caporaso, JG, Kuczynski, J, Stombaugh, J, Bittinger, K, Bushman, FD, Costello, EK, *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**: 335–6.
54. Schloss, PD, Westcott, SL, Ryabin, T, Hall, JR, Hartmann, M, Hollister, EB, *et al.* (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**: 7537–7541.
55. Cook, SDC and Bergquist, PR (2002). Order Dictyoceratida Minchin , 1900. *Order A J. Theory Ordered Sets Its Appl.* **2**: 92019.
56. Van Soest, RWM and Hooper, JN a (2002). Order Poecilosclerida Topsent , 1928. *Syst. Porifera A Guid. to Classif. Sponges*: 403–408.
57. Hill, MS, Hill, AL, Lopez, J, Peterson, KJ, Pomponi, S, Diaz, MC, *et al.* (2013). Reconstruction of family-level phylogenetic relationships within Demospongiae (Porifera) using nuclear encoded housekeeping genes. *PLoS One* **8**: e50437.
58. Rosell, D, Uriz, M-J and Martin, D (1999). Infestation by excavating sponges on the oyster (*Ostrea edulis*) populations of the Blanes littoral zone (north-western Mediterranean Sea). *J. Mar. Biol. Assoc. UK* **79**: 409–413.
59. Xavier, JR, Rachello-Dolmen, PG, Parra-Velandia, F, Schönberg, CHL, Breeuwer, JAJ and van Soest, RWM (2010). Molecular evidence of cryptic speciation in the ‘cosmopolitan’ excavating sponge *Cliona celata* (Porifera, Clionidae). *Mol. Phylogenet. Evol.* **56**: 13–20.
60. Rosell, D and Uriz, M-J (2002). Excavating and endolithic sponge species (Porifera) from the Mediterranean: species descriptions and identification key. *Org. Divers. Evol.* **2**: 55–86.

61. Jeong, J-B, Kim, K-H and Park, J-S (2015). Sponge-Specific Unknown Bacterial Groups Detected in Marine Sponges Collected from Korea Through Barcoded Pyrosequencing. *J. Microbiol. Biotechnol.* **25**: 1–10.
62. Pires, F (2007). Padrões de Distribuição e Taxonomia para os Porifera da Região Central do Algarve: 158.
63. Hentschel, U, Fieseler, L, Wehrl, M, Gernert, C, Steinert, M, Hacker, J, *et al.* (2003). Microbial diversity of marine sponges. *Prog. Mol. Subcell. Biol.* **37**: pp 59–88.
64. Uriz, MJ, Agell, G, Blanquer, A, Turon, X and Casamayor, EO (2012). Endosymbiotic Calcifying Bacteria : A New Cue To The Origin Of Calcification In Metazoa? *Evolution.* 2993–2999
65. Caporaso, JG, Lauber, CL, Walters, WA, Berg-lyons, D, Lozupone, CA, Turnbaugh, PJ, *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *PNAS*
66. Caporaso, JG, Lauber, CL, Walters, W a, Berg-Lyons, D, Huntley, J, Fierer, N, *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**: 1621–1624.
67. Edgar, RC, Haas, BJ, Clemente, JC, Quince, C and Knight, R (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–200.
68. Schloss, PD and Westcott, SL (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* **77**: 3219–3226.
69. 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.
70. 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.
71. Cole, JR, Wang, Q, Cardenas, E, Fish, J, Chai, B, Farris, RJ, *et al.* (2009). The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**: 141–145.
72. Oksanen, J, Kindt, FGBR, Legendre, P, Minchin, PR, O’Hara, RB, Simpson, GL, *et al.* (2015). vegan: Community Ecology Package. *R Packag. version 2.3-0at* <<http://cran.r-project.org/package=vegan>>.

73. Ramette, A (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* **62**: 142–60.
74. Chao, A (1984). Non-parametric estimation of the number of classes in a population. *Scand. J. Stat.* **11**.
75. Wickham, H and Chang, W (2015). ggplot2: An Implementation of the Grammar of Graphics.
76. Kolde, R (2012). Pheatmap: pretty heatmapsat <<http://cran.r-project.org/web/packages/pheatmap/>>.
77. Neuwirth, E (2014). RColorBrewer: ColorBrewer Palettes.
78. Wickham, H (2015). stringr: Simple, Consistent Wrappers for Common String Operations.
79. Caporaso, JG, Kuczynski, J, Stombaugh, J, Bittinger, K, Bushman, FD, Costello, EK, *et al.* (2010). QIIME allows analysis of high- throughput community sequencing data. *Nat. Methods* **7**: 335–336.
80. Pruesse, E, Peplies, J and Glöckner, FO (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.
81. Folmer, O, Black, M, Hoeh, W, Lutz, R and Vrijenhoek, R (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* **3**: 294–299.
82. Xavier, JR, Soest, RWM Van, Breeuwer, JAJ, Martins, AMF and Menken, SBJ (2010). Phylogeography , genetic diversity and structure of the poecilosclerid sponge *Phorbas fictitius* at oceanic islands. *Contrib. to Zool.* **79**: 119–129.
83. Tamura, K, Stecher, G, Peterson, D, Filipski, A and Kumar, S (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**: 2725–9.
84. Edgar, RC (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**: 1792–1797.
85. Ludwig, W, Strunk, O, Westram, R, Richter, L, Meier, H, Yadhukumar, *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**: 1363–71.
86. Stamatakis, A (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

87. Spieck, E, Keuter, S, Wenzel, T, Bock, E and Ludwig, W (2014). Characterization of a new marine nitrite oxidizing bacterium, *Nitrospina watsonii* sp. nov., a member of the newly proposed phylum 'Nitrospinae'. *Syst. Appl. Microbiol.* **37**: 170–176.
88. Team, RDC (2008). R: A language and environment for statistical computing. at <<http://www.r-project.org>>.
89. Dupont, S, Carre-Mlouka, A, Domart-Coulon, I, Vacelet, J and Bourguet-Kondracki, M-L (2014). Exploring cultivable Bacteria from the prokaryotic community associated with the carnivorous sponge *Asbestopluma hypogea*. *FEMS Microbiol. Ecol.* **88**: 160–74.
90. Schöttner, S, Hoffmann, F, Cárdenas, P, Rapp, HT, Boetius, A and Ramette, A (2013). Relationships between host phylogeny, host type and bacterial community diversity in cold-water coral reef sponges. *PLoS One* **8**: e55505.
91. Haegeman, B, Hamelin, J, Moriarty, J, Neal, P, Dushoff, J and Weitz, JS (2013). Robust estimation of microbial diversity in theory and in practice. *ISME J.* **7**: 1092–101.
92. Poppell, E, Weisz, J, Spicer, L, Massaro, A, Hill, A and Hill, M (2013). Sponge heterotrophic capacity and bacterial community structure in high- and low-microbial abundance sponges. *Mar. Ecol.*: n/a–n/doi:10.1111/maec.12098.
93. Moitinho-Silva, L, Seridi, L, Ryu, T, Voolstra, CR, Ravasi, T and Hentschel, U (2014). Revealing microbial functional activities in the Red Sea sponge *Stylissa carteri* by metatranscriptomics. *Environ. Microbiol.* **16**: 3683–3698.
94. Ribes, M, Dziallas, C, Coma, R and Riemann, L (2015). Microbial diversity and putative diazotrophy in high and low microbial abundance Mediterranean sponges. *Appl. Environ. Microbiol.*: AEM.01320–15doi:10.1128/AEM.01320-15.
95. Schmitt, S, Tsai, P, Bell, J, Fromont, J, Ilan, M, Lindquist, N, *et al.* (2012). Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J.* **6**: 564–76.
96. Ley, RE, Lozupone, C A, Hamady, M, Knight, R and Gordon, JI (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* **6**: 776–788.
97. Ley, RE, Hamady, M, Lozupone, C, Turnbaugh, PJ, Ramey, RR, Bircher, JS, *et al.* (2008). Evolution of mammals and their gut microbes. *Science* **320**: 1647–1651.
98. Apprill, A, McNally, S, Parsons, R and Weber, L (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* **75**: 129–137.

99. Jackson, S A., Kennedy, J, Morrissey, JP, O’Gara, F and Dobson, ADW (2012). Pyrosequencing Reveals Diverse and Distinct Sponge-Specific Microbial Communities in Sponges from a Single Geographical Location in Irish Waters. *Microb. Ecol.* **64**: 105–116.
100. Imhoff, JF and Stöhr, R (2003). Sponge-associated bacteria: general overview and special aspects of bacteria associated with *Halichondria panicea*. *Prog. Mol. Subcell. Biol.* **37**: 35–57.
101. Hardoim, CCP, Cardinale, M, Cúcio, ACB, Esteves, AIS, Berg, G, Xavier, JR, *et al.* (2014). Effects of sample handling and cultivation bias on the specificity of bacterial communities in keratose marine sponges. *Front. Microbiol.* **5**: 1–15.
102. Montalvo, NF and Hill, RT (2011). Sponge-associated bacteria are strictly maintained in two closely related but geographically distant sponge hosts. *Appl. Environ. Microbiol.* **77**: 7207–16.
103. Imhoff, JF and Trüper, HG (1976). Marine Sponges as Habitats of Anaerobic Phototrophic Bacteria. *Microb. Ecol.* **2**.
104. Lee, OO, Wang, Y, Yang, J, Lafi, FF, Al-Suwailem, A and Qian, P-Y (2011). Pyrosequencing reveals highly diverse and species-specific microbial communities in sponges from the Red Sea. *ISME J.* **5**: 650–64.
105. Erwin, PM, López-Legentil, S, González-Pech, R and Turon, X (2011). A specific mix of generalists: Bacterial symbionts in Mediterranean *Ircinia* spp. *FEMS Microbiol. Ecol.* **79**: 619–637.
106. Yergeau, E, Maynard, C, Sanschagrín, S, Champagne, J, Juck, D, Lee, K, *et al.* (2015). Microbial community composition, functions and activities in the Gulf of Mexico, one year after the Deepwater Horizon accident. *Appl. Environ. Microbiol.*: AEM.01470–15doi:10.1128/AEM.01470-15.
107. De Voogd, NJ, Cleary, DFR, Polonia, a. RM and Gomes, NCM (2015). Bacterial community composition and predicted functional ecology of sponges, sediment and seawater from the thousand-islands reef complex, West-Java, Indonesia. *FEMS Microbiol. Ecol.*: 1–12doi:10.1093/femsec/fiv019.
108. Bayer, K, Moitinho-Silva, L, Brümmer, F, Cannistraci, C V, Ravasi, T and Hentschel, U (2014). GeoChip-based insights into the microbial functional gene repertoire of marine sponges (HMA, LMA) and seawater. *FEMS Microbiol. Ecol.*doi:10.1111/1574-6941.12441.
109. Kamke, J, Taylor, MW and Schmitt, S (2010). Activity profiles for marine sponge-associated bacteria obtained by 16S rRNA vs 16S rRNA gene comparisons. *ISME J.* **4**: 498–508.

110. Alonso, C, Warnecke, F, Amann, R and Pernthaler, J (2007). High local and global diversity of Flavobacteria in marine plankton. *Environ. Microbiol.* **9**: 1253–1266.
111. Taggart, TL, Shapiro, N, Woyke, T and Chistoserdova, L (2015). Draft Genomes of Two Strains of Flavobacterium Isolated from Lake Washington Sediment. *Genome Announc.* **3**: 2014–2015.
112. González, JM, Pinhassi, J, Fernández-Gómez, B, Coll-Lladó, M, González-Velázquez, M, Puigbò, P, *et al.* (2011). Genomics of the proteorhodopsin-containing marine flavobacterium *Dokdonia* sp. strain MED134. *Appl. Environ. Microbiol.* **77**: 8676–8686.
113. Pham, VD, Konstantinidis, KT, Palden, T and DeLong, EF (2008). Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical profile in the North Pacific Subtropical Gyre. *Environ. Microbiol.* **10**: 2313–2330.
114. Cuvelier, ML, Blake, E, Mulheron, R, McCarthy, PJ, Blackwelder, P, Thurber, RLV, *et al.* (2014). Two distinct microbial communities revealed in the sponge *Cinachyrella*. *Front. Microbiol.* **5**: 1–12.
115. Jackson, S A, Flemer, B, McCann, A, Kennedy, J, Morrissey, JP, O’Gara, F, *et al.* (2013). Archaea appear to dominate the microbiome of *Inflatella pellicula* deep sea sponges. *PLoS One* **8**: e84438.
116. 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**: 431–7.
117. Brück, WM, Reed, JK and McCarthy, PJ (2012). The bacterial community of the lithistid sponge *Discodermia* spp. as determined by cultivation and culture-independent methods. *Mar. Biotechnol. (NY)*. **14**: 762–73.
118. Gladkikh, AS, Kalyuzhnaya, O V., Belykh, OI, Ahn, TS and Parfenova, V V. (2014). Analysis of bacterial communities of two Lake Baikal endemic sponge species. *Microbiology* **83**: 787–797.
119. Lücker, S, Nowka, B, Rattei, T, Spieck, E and Daims, H (2013). The genome of *Nitrospina gracilis* illuminates the metabolism and evolution of the major marine nitrite oxidizer. *Front. Microbiol.* **4**: 1–19.
120. Pfeiffer, S, Pastar, M, Mitter, B, Lippert, K, Hackl, E, Lojan, P, *et al.* (2014). Improved group-specific primers based on the full SILVA 16S rRNA gene reference database. *Environ. Microbiol.* **16**: 2389–2407.
121. Hardoim, CCP (2013). Microbiome diversity and composition in the phylogenetically related marine sponges *S. spinosulus* and *I. variabilis*.

122. Hug, L A, Castelle, CJ, Wrighton, KC, Thomas, BC, Sharon, I, Frischkorn, KR, *et al.* (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1**: 22.
123. Burgsdorf, I, Slaby, BM, Handley, KM, Haber, M, Blom, J, Marshall, CW, *et al.* (2015). Lifestyle Evolution in Cyanobacterial Symbionts of Sponges. *MBio* **6**: 1–14.
124. Escobar, D, Zea, S and Sánchez, J A. (2012). Phylogenetic relationships among the Caribbean members of the *Cliona viridis* complex (Porifera, Demospongiae, Hadromerida) using nuclear and mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* **64**: 271–284.
125. Erpenbeck, D, Sutcliffe, P, Cook, SDC, Dietzel, A, Maldonado, M, van Soest, RWM, *et al.* (2012). Horny sponges and their affairs: On the phylogenetic relationships of keratose sponges. *Mol. Phylogenet. Evol.* **63**: 809–816.
126. Cárdenas, P, Pérez, T and Boury-Esnault, N (2012). Sponge Systematics Facing New Challenges. *Adv. Sponge Sci. Phylogeny, Syst. Ecol.* **61**: pp 79–209.

Annex I – General R script for ecological and statistical analysis

- All commentaries are preceded by a “#” sign, so as to not be recognized as code lines by R.

```
##For P1.shared (all samples) and P2.shared (samples grouped by Sponge Species)
```

```
##Making sure label and numOtus columns are deleted by means of the Ubuntu CLI  
(cut -f2,4- XXX.shared > YY.shared). Any spreadsheet programs won't fully open the  
tables.
```

```
#Set working directory
```

```
setwd("~/Desktop/EMP12.v24_final")
```

```
#Load R packages being used (this can be done further along as well)
```

```
library(picante)
```

```
library(pheatmap)
```

```
#Import EMP Dataset
```

```
ALG <- read.table(file.choose(~Desktop/EMP12.v24_final/P1_forR.shared), header =  
TRUE, row.names = 1)
```

```
class(ALG) # checks the file type (should be "data.frame")
```

```
dim(ALG) # gives you the dimensions of your table
```

```
rownames(ALG) # prints the row names
```

```
head(colnames(ALG)) # prints the first 5 column names
```

```
apply(ALG,1,sum) #total number of reads in each sample
```

```
#if sums are all equal, no need for transformation
```

```
#if not, use decostand in vegan, method="total" or "hellinger", as follows
```

```
ALG_Hellinger <- decostand(ALG, method="total")
```

```
#Importing the metadata file into the R console
```

```
#Don't forget SCIE_NAME is the taxonomical ID of sponges in metadata.
```

```
metadata <- read.table(file.choose(~Desktop/EMP12.v24_final/ALG_metadata.tsv),  
header = TRUE, row.names = 1, sep='\t')
```

```

#Remove underscores from sponge taxonomy, if needed
metadata$SCIE_NAME <- gsub(" ", "_", metadata$SCIE_NAME)

#Sorting rows in "ALG" to match the row order of "metadata"
ALG <- ALG[rownames(metadata), ]

#Check to make sure the row names for each file match.
all.equal(rownames(ALG), rownames(metadata))

#Define Standard Error function (for later)
se<-function(x) sqrt(var(x)/length(x))

#Plot of Sequence Number across the dataset.
#svg format is used so as to posteriorly edit vectorised images in Inkscape
ALG_seqs <-
read.table(file.choose(~Desktop/EMP12.v24_final/Seqcount_beforeSubSampling_ALG.
summary), header = TRUE, row.names = 1)
par(xpd = NA, mar=par()$mar + c(2, 0, 0, 0))
svg("SeqNumbers_ALG.svg", width=14,height=7)
plot(ALG_seqs$Sequence_Number, ylim=c(0,32500), xaxt = 'n', ylab="Sequence
Number", xlab="")
axis(1, at=1:26, labels=rownames(ALG_seqs), par(las=2), tick=FALSE)
#Plot a line depicting the lowest sequence number in a sample i. e., normalization
threshold
abline(h=11203, lty=2, lwd=0.5,col="darkgray")
dev.off()

#Calculate species richness
richness.ALG <- specnumber(ALG)
#Visualize species richness as averaged by sponge species
boxplot(specnumber(ALG) ~ metadata$SCIE_NAME, cex.axis=0.65, ylab = "OTU
Richness")
#Prepare species richness data for diversity summary table
rich.ALG <- as.matrix(specnumber(ALG))
rich.mean<- aggregate(rich.ALG, by=list(metadata$SCIE_NAME), FUN=mean)

```

```

rich.se <- aggregate(rich.ALG, by=list(metadata$SCIE_NAME), FUN=se)
#Test for statistical differences in species richness using ANOVA
richness.aov <- aov(specnumber(ALG) ~ SCIE_NAME, data = metadata)
summary(richness.aov)
#Tukey multiple comparisons of means
#95% family-wise confidence level
TukeyHSD(richness.aov, ordered=TRUE)

#Calculate the Shannon diversity
diversity.ALG <- diversity(ALG, index = "shannon")
#Visualize Shannon diversity as averaged by sponge species
boxplot(diversity.ALG ~ metadata$SCIE_NAME, cex.axis=0.65, ylab = "Shannon-
Weaver Index")
#Prepare Shannon diversity data for diversity summary table
shannon.ALG <- as.matrix(diversity.ALG)
shannon.mean<- aggregate(diversity.ALG, by=list(metadata$SCIE_NAME),
FUN=mean)
shannon.se <- aggregate(diversity.ALG, by=list(metadata$SCIE_NAME), FUN=se)
#Test for statistical differences in Shannon index using ANOVA
shannon.aov <- aov(diversity.ALG ~ SCIE_NAME, data = metadata)
summary(shannon.aov)
#Tukey multiple comparisons of means
#95% family-wise confidence level
tukey.aov<-TukeyHSD(shannon.aov, ordered=TRUE)

#Calculate Pielou index (not directly included in vegan package)
H <- diversity(ALG)
S <- specnumber(ALG)
pielou.ALG <- H/log(S)
#Visualize species dominance as averaged by sponge species
boxplot(pielou.ALG ~ metadata$SCIE_NAME, cex.axis=0.65, ylab = "Pielou's
Evenness Index")
#Prepare species dominance data for diversity summary table

```

```

pielou.ALG <- as.matrix(pielou.ALG)
pielou.mean<- aggregate(pielou.ALG, by=list(metadata$SCIE_NAME), FUN=mean)
pielou.se <- aggregate(pielou.ALG, by=list(metadata$SCIE_NAME), FUN=se)
#Test for statistical differences in species dominance using ANOVA
pielou.aov <- aov(pielou.ALG ~ SCIE_NAME, data = metadata)
summary(pielou.aov)
#Tukey multiple comparisons of means
#95% family-wise confidence level
TukeyHSD(pielou.aov, ordered=TRUE)

###Multiple Plot for Richness, Shannon and Pielou indices
svg("RichDivPielou.svg", width=14, height=14)
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
#Calculate species richness
richness.ALG <- specnumber(ALG)
#Visualize species richness by sponge species
boxplot(specnumber(ALG) ~ metadata$SCIE_NAME, cex.axis=1.1, main = "OTU
Richness")
#Calculate the Shannon diversity
diversity.ALG <- diversity(ALG, index = "shannon", base=2)
#Visualize Shannon diversity by sponge species
boxplot(diversity.ALG ~ metadata$SCIE_NAME, cex.axis=1.1, main = "Shannon-
Weaver Index")
#Calculate Pielou index
H <- diversity(ALG)
S <- specnumber(ALG)
pielou.ALG <- H/log(S)
#Visualize species dominance by sponge species
boxplot(pielou.ALG ~ metadata$SCIE_NAME, cex.axis=1.1, main = "Pielou's
Evenness Index")
dev.off()
plot.new()

```

```

#Construct and output diversity summary table

div.summary <- cbind(rich.mean, rich.se$V1, shannon.mean$x, shannon.se$x,
invsimpson.mean$V1, invsimpson.se$V1, pielou.mean$V1, pielou.se$V1)

names(div.summary) <- c("Species", "Mean.Richness", "se.Richness",
"Mean.Shannon", "se.Shannon", "Mean.InvSimpson", "se.InvSimpson", "Mean.Pielou",
"se.Pielou")

write.csv(div.summary, file = "DiversityMetricSummary.csv")

#Richness vs. Chao1 Estimated Richness

#Using ALG, so as to obtain sponge species averages, as opposite to pooled results
est.plot<-estimateR(ALG)
nomes=colnames(est.plot)
est.plot.inv=t(est.plot)
rownames(est.plot.inv)= nomes
est.plot_df<-as.data.frame(est.plot.inv)
media=aggregate(cbind(S.obs=est.plot_df$S.obs, S.chao1=est.plot_df$S.chao1,
se.chao1=est.plot_df$se.chao1), by=list(metadata$SCIE_NAME), mean)
media_se=aggregate(cbind(se.obs=est.plot_df$se.obs), by=list(metadata$SCIE_NAME),
se)
agg_media=c(row.names=media$row.names, S.obs=media$S.obs,
se.obs=media_se$se.obs, S.chao1=media$S.chao1, se.chao1=media$se.chao1)
agg_media=merge(media, media_se)
agg_media <- agg_media[c("Group.1", "S.obs", "se.obs", "S.chao1", "se.chao1")]
write.csv(agg_media, file="Obs_vs_Est_Richness.csv")

#Plot with ggplot2.

#This part of the script has been developed with the help of Stack Overflow forum
members

library(tidyr)
library(ggplot2)

mdat <- gather(agg_media, S, value, -Group.1)

# I want the S variable data split on the period (.) into two variables which I'll call type
and var.

# type contains values S or se and var contains obs or chao1

mdat <- separate(mdat, S, c("type","var"))

```

```
#spread out the currently compact data so that we have columns S and se, which we do with spread()
```

```
mdat <- spread(mdat, type, value)
```

```
#Reorder mdat by obs followed by chao1
```

```
mdat <- transform(mdat, var = relevel(factor(var), "obs"))
```

```
blue.bold.italic.16.text <- element_text(color = "black", size = 13)
```

```
red.bold.italic.text.x <- element_text(face = "bold", color = "black", size=16, vjust=0)
```

```
red.bold.italic.text.y <- element_text(face = "bold", color = "black", size=16, vjust=1)
```

```
g<-ggplot(mdat, aes(x = Group.1, y = S, fill = var))+
```

```
  geom_bar(position = "dodge", stat = "identity") +
```

```
  geom_errorbar(mapping = aes(ymax = S + se, ymin = S - se),
```

```
    position = position_dodge(width=0.9), width = 0.25) +
```

```
  scale_y_continuous(expand = c(0, 0), limits=c(0,950)) +
```

```
  scale_fill_discrete(labels=c("Observed OTUs", "Chao1 Estimation"), name="OTU Counts") +
```

```
  theme_bw()+
```

```
  theme(legend.background = element_rect(colour = "black"), axis.text = blue.bold.italic.16.text,
```

```
    axis.title.x = red.bold.italic.text.x, axis.title.y = red.bold.italic.text.y) +
```

```
  xlab("Sponge Species") +
```

```
  ylab("OTU Richness")
```

```
ggsave(g, file = "Obs_Est_OTUs.svg")
```

```
#Assessing dissimilarity profiles along samples
```

```
#Bray-Curtis distance among samples
```

```
#No previous transformation (Hellinger) is required
```

```
ALG.bc.dist <- vegdist(ALG, method = "bray")
```

```
#Test for differences in Bray-Curtis dissimilarity among host species
```

```
braycurtis.adonis<-adonis(comm.bc.dist ~ SCIE_NAME, data = metadata, perm=1e3)
```

```
#Cluster the communities using average-linkage algorithm
```

```
ALG.bc.clust <- hclust(ALG.bc.dist, method = "average")
```

```
#Visualize community dissimilarity with a cluster dendrogram
```

```
plot(ALG.bc.clust, cex = 0.9, ylab = "Bray-Curtis dissimilarity", hang = -1, lwd = 2)
```



```

## Bray-Curtis Dendrogram w/ ALG_clean_names (sample names arranged prettier)
ALG_clean_names <-
read.table(file.choose(~Desktop/EMP12.v24_final/P1_forR.shared), header = TRUE,
row.names = 1)

png("BC_Dissimilarity.png", width=2200, height=1800, res=300)
ALG.bc.dist_dendro <- vegdist(ALG_clean_names, method = "bray")
ALG.bc.clust_dendro <- hclust(ALG.bc.dist_dendro, method = "average")
# vector of colors labelColors = c('red', 'blue', 'darkgreen', 'darkgrey', 'purple')
hcd = as.dendrogram(ALG.bc.clust_dendro)
labelColors = c("#CDB380", "#036564", "#EB6841", "#EDC951")
# cut dendrogram in 4 clusters
clusMember = cutree(hcd, 4)
# function to get color labels
colLab <- function(n) {
  if (is.leaf(n)) {
    a <- attributes(n)
    labCol <- labelColors[clusMember[which(names(clusMember) == a$label)]]
    attr(n, "nodePar") <- c(a$nodePar, lab.col = labCol)
  }
  n
}
# using dendrapply
clusDendro = dendrapply(hcd, colLab)
# make plot
plot(clusDendro, main = "Bray-Curtis Dissimilarity")
dev.off()

# #MRPP Test for BC Clustering
mrpp_ALG<-mrpp(ALG, group=metadata$SCIE_NAME, distance="bray")
scie_names<- as.factor(metadata$SCIE_NAME)

#nMDS plot with Bray-Curtis distance
ALG.bc.mds <- metaMDS(ALG, dist = "bray", k = 2, trymax = 50)

```

```

##Customize MDS visualization
svg("BC_nMDS.svg", width=10,height=7)
ordiplot(ALG.bc.mds, type = "none")
points(mds.fig, "sites", pch = 19, col = "dodgerblue", select = metadata$SCIE_NAME
== "Cliona viridis")
points(mds.fig, "sites", pch = 19, col = "cornflowerblue", select =
metadata$SCIE_NAME == "Cliona celata complex")
points(mds.fig, "sites", pch = 19, col = "cyan2", select = metadata$SCIE_NAME ==
"Phorbas fictitius")
points(mds.fig, "sites", pch = 19, col = "darkcyan", select = metadata$SCIE_NAME ==
"Dysidea fragilis")
ALG.bc.mds
text(-5,2, "Stress = 0.1116067", cex = .9) ##As checked by gof, below
ordiellipse(ALG.bc.mds, metadata$SCIE_NAME, conf = 0.95, label = TRUE)
ordicluster(ALG.bc.mds, ALG.bc.clust, col = "gray")
dev.off()

# Assess goodness of ordination fit (stress plot), nonlinear fit
ALG.anosim<-anosim(ALG, metadata$SCIE_NAME, distance ="bray")
summary(ALG.anosim)

ALG.adonis<-adonis(ALG.bc.dist ~ SCIE_NAME, data = metadata)
summary(ALG.adonis)
svg("BC_nMDS_stressplot.svg", width=10,height=7)
stressplot(ALG.bc.mds, pch=1, p.col="gray", lwd=2, l.col="red")
dev.off()
Hellinger_nMDS_Stress <- metaMDS(comm = ALG_Hellinger, trace = FALSE)
Goodness_Hell_nMDS<-goodness(ALG.bc.mds)

gof <- goodness(ALG.bc.mds)
gof
plot(ALG.bc.mds, display = "sites", type = "n")
points(ALG.bc.mds, display = "sites", cex = gof/2)

```

```
# "good rule of thumb: stress > 0.05 provides an excellent representation in reduced dimensions, > 0.1 is great, >0.2 is good/ok,
```

```
#and stress > 0.3 provides a poor representation."
```

```
http://jonlefcheck.net/2012/10/24/nmds-tutorial-in-r/
```

```
# matrix of mean within-cluster dissimilarities (diagonal) and
```

```
# between-cluster dissimilarities (off-diagonal elements), and an attribute n of grouping counts.
```

```
ALG.md <- with(metadata, meandist(vegdist(ALG), scie_names))
```

```
ALG.md
```

```
mrpp_ALG$Pvalue
```

```
summary(ALG.md)
```

```
plot(ALG.md)
```

```
##Checking the Singletome
```

```
Singletome <-
```

```
read.table(file.choose(~Desktop/EMP12.v24_final/Singletome/mothur/Singletome_forR.shared), header = TRUE, row.names = 1)
```

```
Singletome <- Singletome[rownames(metadata), ]
```

```
all.equal(rownames(Singletome), rownames(metadata))
```

```
dim(Singletome)
```

```
rich.Singletome <- as.matrix(specnumber(Singletome))
```

```
rich.Sing.mean<- aggregate(rich.Singletome, by=list(metadata$SCIE_NAME), FUN=mean)
```

```
rich.Sing.se <- aggregate(rich.Singletome, by=list(metadata$SCIE_NAME), FUN=se)
```

```
##Construct heatmap ordered by abundance for P1
```

```
#load pheatmap and RColorBrewer packages
```

```
library(pheatmap)
```

```
library(RColorBrewer)
```

```
#set color palette for heatmap
```

```
ALG_palette <- brewer.pal(9, "YlOrRd")
```

```
pal<-colorRampPalette(c("lightyellow", "lightgoldenrod1", "goldenrod1", "orange", "indianred", "red"))(10)
```

```
#Transposing data frame
```

```

ALG_t <- as.data.frame(t(ALG_clean_names[,-1]))
ALG_t <- ALG_t[order(rowSums(ALG_t), decreasing=T),]
ALG_t <- as.data.frame(t(ALG_t))
#Fourth-Root Transformation for optimal abundance gradient visualization
ALG.sqrt <- sqrt(ALG_t)
ALG.ftrt <- sqrt(ALG.sqrt)
#Plot heatmap
svg("gen_order_pheatmap.svg")
pheatmap(ALG.ftrt, color = pal, cluster_cols=FALSE, show_rownames=T,
show_colnames=F, clustering_method="single")
dev.off()

##Filter P1 for the top 50 Otus and plot into a heatmap
ALG_t <- as.data.frame(t(ALG[, -1]))
ALG_t <- ALG_t[order(rowSums(ALG_t), decreasing=T),]
ALG_top50 <- ALG_t[1:50,]
ALG_top50_Sums <- as.data.frame(c(row.names(rownames(ALG_top50)),
rowSums(ALG_top50)))
ALG_top50_sqrt <- sqrt(ALG_top50)
ALG_top50_ftrt <- sqrt(ALG_top50_sqrt)
svg("top50_pheatmap.svg")
pheatmap(ALG_top50_ftrt, color = pal, cluster_cols=T, cluster_rows=F,
show_rownames=T, show_colnames=T, clustering_method="single")
dev.off()

##Construct heatmap for P2
#Import file (no numOtus, label columns)
ALG_P2 <- read.table(file.choose(~Desktop/EMP12.v24_final/P2_forR.shared), header
= TRUE, row.names = 1)
ALG_P2.sqrt <- sqrt(ALG_P2)
ALG_P2.ftrt <- sqrt(ALG_P2.sqrt)
ALG_P2_palette <- brewer.pal(9, "GnBu")

```

```
pheatmap(ALG_P2.ftrt, color = ALG_P2_palette, cluster_cols=FALSE,
show_rownames=T, show_colnames=F, clustering_method="single")
```

```
##Top50 OTUs for P2
```

```
ALG_P2_t <- as.data.frame(t(ALG_P2[,-1]))
```

```
ALG_P2_t<-ALG_P2_t[order(rowSums(ALG_P2_t),decreasing=T),]
```

```
ALG_P2_top50<-ALG_P2_t[1:50,]
```

```
ALG_P2_top50_sqrt<-sqrt(ALG_P2_top50)
```

```
ALG_P2_top50_ftrt<-sqrt(ALG_P2_top50_sqrt)
```

```
pheatmap(ALG_P2_top50_ftrt, color = ALG_palette, cluster_cols=T, cluster_rows=F,
show_rownames=T, show_colnames=T, clustering_method="single")
```

```
##Construct Rarefaction Curves for P2 using mothur outputs
```

```
#Lucas Moitinho-Silva helped greatly in this component
```

```
##Command used in mothur: rarefaction.shared(shared=EMP12_ALG_P1.shared,
##design=ALG_EMP12_groups.design, iters=1500)
```

```
library("stringr")
```

```
library("RColorBrewer")
```

```
rare=read.table("EMP12_ALG_P1.groups_per_sample.rarefaction",
stringsAsFactors=F, header=T)
```

```
cols_methods=c()
```

```
for (i in c(1:ncol(rare))) {
```

```
  if (str_split(colnames(rare)[i], "method")[[1]][1] == "")
  { cols_methods=append(cols_methods, 1) }
```

```
rare_0=rare[,2:ncol(rare)]
```

```
rare_0[is.na(rare_0)]=0
```

```
yl=seq(from=min(rare_0), to=max(rare_0), length.out=nrow(rare)) #ylim
```

```
num_of_samps=(ncol(rare)-1)/3
```

```
#Colors
```

```
ncol=num_of_samps
```

```
cols <- RColorBrewer:::brewer.pal(ncol, "Set2") # OR c("purple", "white", "orange")
```

```
rampcols <- colorRampPalette(colors = cols, space="Lab")(ncol)
```

```
colors=data.frame(c("0"), rampcols, stringsAsFactors=FALSE)
```

```
svg("RareF_ALG_P2.svg")
```

```

plot(rare[,1], yl, type="n", ylab="OTUs", xlab="Samples")
list_of_columns=seq(from=2, to=ncol(rare), by=3)
col=0
for (i in list_of_columns){
  col=col+1
  species= str_split(colnames(rare)[i], "03.")[[1]][2]
  a=rare[,i:(i+2)]
  lines(rare[,1], a[,1], col=colors[col,2], lwd=6)
  #Plotting the error
  polygon(c(rev(rare[,1]), rare[,1]), c(rev(a[,2]), (a[,3])), col=colors[col,2], density=60,
  lty="dashed", border = NA)
  colors[col,1] = species
}
legend(9000, 140, title="Species", colors[1:col,1], fill=colors[,2], horiz=F, cex=0.7)
dev.off()

## (same as above) Rarefaction by sequence for P2
##rarefaction.single(shared=EMP12_ALG_P2.shared, iters=1500)
#Lucas Moitinho-Silva helped greatly in this component
library("stringr")
library("RColorBrewer")
rare=read.table(file.choose(~Desktop/EMP12.v24_final/EMP12_ALG_P2.groups_per_s
equence.rare_nTfaction), stringsAsFactors=F, header=T)
cols_methods=c()
for (i in c(1:ncol(rare))) {
  if (str_split(colnames(rare)[i], "method")[[1]][1] == "")
  {cols_methods=append(cols_methods,1)}
}
rare_nT_0=rare[,2:ncol(rare)]
rare_nT_0[is.na(rare_0)]=0
yl=seq(from=min(rare_0), to=max(rare_0), length.out=nrow(rare)) #ylim
num_of_samps=(ncol(rare)-1)/3
#Colors
ncol=num_of_samps

```

```

cols <- RColorBrewer:::brewer.pal(ncol,"Set2") # OR c("purple","white","orange")
rampcols <- colorRampPalette(colors = cols, space="Lab")(ncol)
colors=data.frame(c("0"), rampcols, stringsAsFactors=FALSE)
svg("rare_ALG_P2_by_seq.svg", width=10,height=7)
plot(rare[,1], yl, type="n", ylab="OTUs", xlab= "Sequences", cex.axis=1, cex.lab=1)
list_of_columns=seq(from=2, to=ncol(rare), by=3)
col=0
for (i in list_of_columns){
  col=col+1
  species= str_split(colnames(rare)[i],"03.")[[1]][2]
  a=rare[,i:(i+2)]
  lines(rare[,1], a[,1], col=colors[col,2], lwd=6)
  #Plotting the error
  polygon(c(rev(rare[,1]), rare[,1]), c(rev(a[,2]), (a[,3])), col=colors[col,2], density=50,
lty="dotted", border = NA)
  colors[col,1] = species
  abline(v=11265, lty=2, lwd=0.5,col="darkgray")
}
legend(110500, 400, title="Species", colors[1:col,1], fill=colors[,2], horiz=F, cex=0.7)
dev.off()

```

```
##Rarefaction Curve for Chao1 Estimation
```

```
##rarefaction.single(shared=EMP12_ALG_P2.shared, iters=1500, calc=chao)
```

```
#Lucas Moitinho-Silva helped greatly in this component
```

```
library("stringr")
```

```
library("RColorBrewer")
```

```
rare_chao=read.table(file.choose(~Desktop/EMP12.v24_final/EMP12_ALG_P2..groups
_per_sequence.r_chao), stringsAsFactors=F, header=T)
```

```
cols_methods=c()
```

```
for (i in c(1:ncol(rare_chao)))
```

```
  if (str_split(colnames(rare_chao)[i],"method")[[1]][1] == "")
  {cols_methods=append(cols_methods,1)}
```

```
rare_chao_0=rare_chao[,2:ncol(rare_chao)]
```

```

rare_chao_0[is.na(rare_chao_0)]=0

yl=seq(from=min(rare_chao_0), to=max(rare_chao_0), length.out=nrow(rare_chao))
#ylim

num_of_samps=(ncol(rare_chao)-1)/3

#Colors

ncol=num_of_samps

cols <- RColorBrewer::brewer.pal(ncol,"Set2") # OR c("purple","white","orange")
rampcols <- colorRampPalette(colors = cols, space="Lab")(ncol)

colors=data.frame(c("0"), rampcols, stringsAsFactors=FALSE)

svg("rare_chaoF_ALG_P2_by_seq.svg", width=10,height=7)

plot(rare_chao[,1], yl, type="n", ylab="OTUs", xlab= "Sequences")

list_of_columns=seq(from=2, to=ncol(rare_chao), by=3)

col=0

for (i in list_of_columns){

  col=col+1

  species= str_split(colnames(rare_chao)[i],"03.")[[1]][2]

  a=rare_chao[,i:(i+2)]

  lines(rare_chao[,1], a[,1], col=colors[col,2], lwd=6)

  #Plotting the error

  polygon(c(rev(rare_chao[,1]), rare_chao[,1]), c(rev(a[,2]), (a[,3])), col=colors[col,2],
density=60, lty="dotted", border = NA)

  colors[col,1] = species

}

legend(110500, 500, title="Species", colors[1:col,1], fill=colors[,2], horiz=F, cex=0.7)

dev.off()

##Rarefaction curve of the custom-generated averaged ALG dataset
##rarefaction.single(shared=ALG_averaged_formothur.shared, iters=1500)
#Lucas Moitinho-Silva helped greatly in this component

library("stringr")

library("RColorBrewer")

rare=read.table("ALG_averaged_formothur.groups.rarefaction", stringsAsFactors=F,
header=T)

```



```

cols_methods=c()
for (i in c(1:ncol(rare))) {
  if (str_split(colnames(rare)[i], "method")[[1]][1] == "")
  { cols_methods=append(cols_methods, 1) }
rare_0=rare[,2:ncol(rare)]
rare_0[is.na(rare_0)]=0
yl=seq(from=min(rare_0), to=max(rare_0), length.out=nrow(rare)) #ylim
num_of_samps=(ncol(rare)-1)/3
#Colors
ncol=num_of_samps
cols <- RColorBrewer:::brewer.pal(ncol, "Set2") # OR c("purple", "white", "orange")
rampcols <- colorRampPalette(colors = cols, space="Lab")(ncol)
colors=data.frame(c("0"), rampcols, stringsAsFactors=FALSE)
svg("RareF_ALG_averaged_by_seq.svg", width=10,height=7)
plot(rare[,1], yl, type="n", ylab="OTUs", xlab="Sequences")
list_of_columns=seq(from=2, to=ncol(rare), by=3)
col=0
for (i in list_of_columns){
  col=col+1
  species= str_split(colnames(rare)[i], "03.")[[1]][2]
  a=rare[,i:(i+2)]
  lines(rare[,1], a[,1], col=colors[col,2], lwd=6)
  #Plotting the error
  polygon(c(rev(rare[,1]), rare[,1]), c(rev(a[,2]), (a[,3])), col=colors[col,2], density=60,
  lty="dotted", border = NA)
  colors[col,1] = species
}
legend(8500, 150, title="Species", colors[1:col,1], fill=colors[,2], horiz=F, cex=0.7)
dev.off()

##Construct Untransformed rarefaction Curves for P2 using mothur outputs
##rare_nTfaction.single(shared=P1_nt.merge.shared, iters=1500)
#Lucas Moitinho-Silva helped greatly in this component

```

```

library("stringr")
library("RColorBrewer")
rare_nT=read.table(file.choose("EMP12_ALG_P1.groups_per_sample.rarefaction"),
stringsAsFactors=F, header=T)
cols_methods=c()
for (i in c(1:ncol(rare_nT))){
  if (str_split(colnames(rare_nT)[i],"method")[[1]][1] == "")
  {cols_methods=append(cols_methods,1)} }
rare_nT_0=rare_nT[,2:ncol(rare_nT)]
rare_nT_0[is.na(rare_nT_0)]=0
yl=seq(from=min(rare_nT_0), to=max(rare_nT_0), length.out=nrow(rare_nT)) #ylim
num_of_samps=(ncol(rare_nT)-1)/3
#Colors
ncol=num_of_samps
cols <- RColorBrewer:::brewer.pal(ncol,"Set2") # OR c("purple","white","orange")
rampcols <- colorRampPalette(colors = cols, space="Lab")(ncol)
colors=data.frame(c("0"), rampcols, stringsAsFactors=FALSE)
pdf("rare_nT_nTF_ALG_P2.pdf")
plot(rare_nT[,1], yl, type="n", ylab="OTUs", xlab= "Sequences")
list_of_columns=seq(from=2, to=ncol(rare_nT), by=3)
col=0
for (i in list_of_columns){
  col=col+1
  species= str_split(colnames(rare_nT)[i],"03.")[[1]][2]
  a=rare_nT[,i:(i+2)]
  lines(rare_nT[,1], a[,1], col=colors[col,2], lwd=6)
  #Plotting the error
  polygon(c(rev(rare_nT[,1]), rare_nT[,1]), c(rev(a[,2]), (a[,3])), col=colors[col,2],
density=60, lty="dashed", border = NA)
  colors[col,1] = species
}
legend(180000, 800, title="Species", colors[1:col,1], fill=colors[,2], horiz=F, cex=0.8)
# dev.off()

```

```

##Building Normalized OTU and Sequence Abundance Table per Phylum
#Lucas Moitinho-Silva helped greatly in this component
library(stringr)
library(reshape)
EMP=EMP_list$ALG
map=EMP_list$metadata
EMP=EMP[,colSums(EMP) !=0]
identical(rownames(ALG), rownames(metadata))

database=read.delim("../study_1740_wPE.final.filter.LwRmvd.RwdnoCpt.nm.SwRm
vd.Cln.d.database", stringsAsFactors=F, head=T,colClasses = "character", na.strings = F)
database= database[database$OTU %in% colnames(EMP),]
identical(database$OTU, colnames(EMP))
#select database level and phyogeny
database=database[, colnames(database) %in% c("OTU", "OTURepTaxGG")]
levels=colsplit(database[,2], split=";", names=c("k","p","c","o","f", "g", "s"))
#parse taxa names
for (i in 1:ncol(levels)){
  taxa=data.frame(as.character(levels[,i]))
  col1=data.frame(as.character(colsplit(taxa[,1], "\\(", names=c("a","b"))[,1]))
  taxa=data.frame(as.character(colsplit(col1[,1], "__", names=c("a","b"))[,2]))
  levels[,i]=taxa}
#Convert the factors to character
levels <- data.frame(lapply(levels, as.character), stringsAsFactors=FALSE)
##Break Proteobacteria into class
for (i in c(1:nrow(levels))){
  if (levels$p[i] == "Proteobacteria") levels$p[i] =
as.character(paste("Proteo",levels[i,3], sep=" "))}
database=data.frame(database[,1], levels$p, stringsAsFactors = F)
colnames(database)=c("OTU", "Phylum")
#transpose emp
EMP.t=t(EMP)

```

```

EMP.t.PA=tran(EMP.t, method="pa")
identical(rownames(EMP.t.PA), database$OTU)
#sum of OTUs per phylum
EMP.t.PA.sum=aggregate(EMP.t.PA, by=list(database$Phylum), sum)
EMP.PA.sum = setNames(data.frame(t(EMP.t.PA.sum[,-1])), EMP.t.PA.sum[,1])
means.EMP.PA.sum=aggregate(EMP.PA.sum, by=list(metadata$SCIE_NAME), mean)
sd.EMP.t.PA=aggregate(EMP.PA.sum, by=list(metadata$SCIE_NAME), sd)
write.csv(means.EMP.PA.sum, file="mean_OTU_Counts_perPhylum.csv")
write.csv(sd.EMP.t.PA, file="sd_OTU_Counts_perPhylum.csv")
#number of sequences per phylum
EMP.t.sum=aggregate(EMP.t, by=list(database$Phylum), sum)
EMP.seqs.sum = setNames(data.frame(t(EMP.t.sum[,-1])), EMP.t.sum[,1])
means.EMP.seqs.sum=aggregate(EMP.seqs.sum, by=list(metadata$SCIE_NAME),
mean)
sd.EMP.seqs.sum=aggregate(EMP.seqs.sum, by=list(metadata$SCIE_NAME), sd)
write.csv(means.EMP.seqs.sum, file="mean_Seq_Counts_perPhylum.csv")
write.csv(sd.EMP.seqs.sum, file="sd_Seq_Counts_perPhylum.csv")

##Building not-Normalized OTU and Sequence Abundance Table per Phylum
#Lucas Moitinho-Silva helped greatly in this component
library(stringr)
library(reshape)
EMP_nT <- read.table(file.choose(~Desktop/EMP12.v24_final/P1_nT_forR.shared),
header = TRUE, row.names = 1)
nT_EMP.list<-list(EMP_nt=EMP_nT, metadata=metadata)
EMP_nT=nT_EMP.list$EMP_nt
map=nT_EMP.list$metadata
EMP_nT=EMP_nT[,colSums(EMP_nT) !=0] #!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!1
identical(rownames(EMP_nT), rownames(metadata))
database=read.delim(file.choose("../..study_1740_wPE.final.filter.LwRmvd.RwdnoCpt.
nm.SwRmvd.CInd.database"),
stringsAsFactors=F, head=T,colClasses = "character", na.strings = F)
database= database[database$OTU %in% colnames(EMP_nT),]

```

```

identical(database$OTU, colnames(EMP_nT))
#select database level and phyogeny
database=database[, colnames(database) %in% c("OTU", "OTURepTaxGG")]
levels=colsplit(database[,2], split=";", names=c("k","p","c","o","f", "g", "s"))
#parse taxa names
for (i in 1:ncol(levels)){
  taxa=data.frame(as.character(levels[,i]))
  col1=data.frame(as.character(colsplit(taxa[,1], "\\(", names=c("a","b"))[,1]))
  taxa=data.frame(as.character(colsplit(col1[,1], "__", names=c("a","b"))[,2]))
  levels[,i]=taxa}
#Convert the factors to character
levels <- data.frame(lapply(levels, as.character), stringsAsFactors=FALSE)
##Break Proteobacteria into class
for (i in c(1:nrow(levels))){
  if (levels$p[i] == "Proteobacteria") levels$p[i] =
as.character(paste("Proteo",levels[i,3], sep=" "))}
database=data.frame(database[,1], levels$p, stringsAsFactors = F)
colnames(database)=c("OTU", "Phylum")
#transpose EMP_nt
EMP_nT.t=t(EMP_nT)
EMP_nT.t.PA=decostand(EMP_nT.t, method="pa")
identical(rownames(EMP_nT.t.PA), database$OTU)
#sum of OTUs per phylum
EMP_nT.t.PA.sum=aggregate(EMP_nT.t.PA, by=list(database$Phylum), sum)
EMP_nT.PA.sum = setNames(data.frame(t(EMP_nT.t.PA.sum[,-1])),
EMP_nT.t.PA.sum[,1])
means.EMP_nT.PA.sum=aggregate(EMP_nT.PA.sum, by=list(metadata$SCIE_NAME),
mean)
sd.EMP_nT.t.PA=aggregate(EMP_nT.PA.sum, by=list(metadata$SCIE_NAME), sd)
write.csv(means.EMP_nT.PA.sum, file="mean_nT_OTU_Counts_perPhylum.csv")
write.csv(sd.EMP_nT.t.PA, file="sd_nT_OTU_Counts_perPhylum.csv")
#number of sequences per phylum
EMP_nT.t.sum=aggregate(EMP_nT.t, by=list(database$Phylum), sum)

```

```

EMP_nT.seqs.sum = setNames(data.frame(t(EMP_nT.t.sum[,-1])), EMP_nT.t.sum[,1])
means.EMP_nT.seqs.sum=aggregate(EMP_nT.seqs.sum,
by=list(metadata$SCIE_NAME), mean)
sd.EMP_nT.seqs.sum=aggregate(EMP_nT.seqs.sum, by=list(metadata$SCIE_NAME),
sd)
write.csv(means.EMP_nT.seqs.sum, file="mean_nT_Seq_Counts_perPhylum.csv")
write.csv(sd.EMP_nT.seqs.sum, file="sd_nT_Seq_Counts_perPhylum.csv")

##SIMPER test for highlighting specific microbes that result in the differences among
samples

##new metadata file had to be generated, with added presence/absence columns
regarding each sponge species

##new .shared file also has to generated

##P1 was filtered for OTUs with at least 100 assigned sequences, using "filter.shared"
metadata_for_simper <-
read.table(file.choose(~Desktop/EMP12.v24_final/Ametadata_for_simper.txt), header =
TRUE, row.names = 1, sep='\t')

ALG_for_simper <-
read.table(file.choose(~Desktop/EMP12.v24_final/ALG_for_simper_morethan_100_se
qsbyOTU.shared), header = TRUE, row.names = 1)

dim(ALG_for_simper)

#SIMPER is working for
ncol(ALG_for_simper)

#species
ALG_for_simper <- ALG_for_simper[rownames(metadata_for_simper), ]
all.equal(rownames(ALG_for_simper), rownames(metadata_for_simper))

##Heatmap of OTUs w/ at least 100 sequences (top 132 OTUs)
library(pheatmap)
library(RColorBrewer)
ALG_simper_palette <- brewer.pal(9, "GnBu")
ALG_simper.sqrt <- sqrt(ALG_for_simper)
ALG_simper.ftrt <- sqrt(ALG_simper.sqrt)
ALG_simper.ftrt$rareOTUs <- NULL
dim(ALG_simper.ftrt)

```

```

svg("simper.heatmap.OTUsW100ormoreSeqs.svg")
simper.heatmap<-pheatmap(ALG_simper.ftrt, color = ALG_simper_palette,
cluster_cols=FALSE, show_rownames=T,
                        show_colnames=T, clustering_method="single")
dev.off()

# SIMPER tables
Ccel.simp<-simper(ALG_for_simper, metadata_for_simper$Ccel)
Dfra.simp<-simper(ALG_for_simper, metadata_for_simper$Dfra)
Pfic.simp<-simper(ALG_for_simper, metadata_for_simper$Pfic)
Cvir.simp<-simper(ALG_for_simper, metadata_for_simper$Cvir)
Ccel.simp.sum<-as.matrix(summary(Ccel.simp, ordered=TRUE, digits=3))
Dfra.simp.sum<-as.matrix(summary(Dfra.simp, ordered=TRUE, digits=3))
Pfic.simp.sum<-as.matrix(summary(Pfic.simp, ordered=TRUE, digits=3))
Cvir.simp.sum<-as.matrix(summary(Cvir.simp, ordered=TRUE, digits=3))
#Still huge, get first 10 lines of each
#cumsum is the ordered cumulative controbution
Ccel.simp.top<-do.call("rbind", Ccel.simp.sum)
Ccel.simp.top<-head(Ccel.simp.top,10)
write.csv(Ccel.simp.top, file="Top10_OTUs_for_Ccel.csv")

Dfra.simp.top<-do.call("rbind", Dfra.simp.sum)
Dfra.simp.top<-head(Dfra.simp.top,10)
write.csv(Dfra.simp.top, file="Top10_OTUs_for_Dfra.csv")
Pfic.simp.top<-do.call("rbind", Pfic.simp.sum)
Pfic.simp.top<-head(Pfic.simp.top,10)
write.csv(Pfic.simp.top, file="Top10_OTUs_for_Pfic.csv")
Cvir.simp.top<-do.call("rbind", Cvir.simp.sum)
Cvir.simp.top<-head(Cvir.simp.top,10)
write.csv(Cvir.simp.top, file="Top10_OTUs_for_Cvir.csv")
##André Soares 2015

```

Annex II – R Script for generation of an averaged abundance matrix

```

##Create P2, but as averaged values of grouped sequences
##mothur's merge.groups pools samples, instead of averaging them
ALG_merged<-aggregate(ALG, by = metadata.for.pcoa, FUN=mean)
ALG_merged_mean <- as.data.frame(lapply(ALG_merged[is.num], round, 0),
rownames(ALG_merged))

row.names(ALG_merged_mean)<-ALG_merged$SCIE_NAME
class(ALG_merged_mean) # checks the file type (should be "data.frame")
dim(ALG_merged_mean) # gives you the dimensions of your table

#Export .shared of averaged merged ALG dataset
ALG_merged_mean$label <- 0.03
otunum_grep<-grep("Otu", colnames(ALG_merged_mean))
ALG_merged_mean$numOtus <- length(otunum_grep)
ALG_merged_mean<-ALG_merged_mean[, colSums(ALG_merged_mean != 0) > 0]
grep("numOtus", colnames(ALG_merged_mean))
grep("label", colnames(ALG_merged_mean))
dim(ALG_merged_mean)

write.table(ALG_merged_mean, "ALG_averaged_formothur.shared", sep =
"\t",quote=FALSE)

#Then change label to 1st column and numOtus to 3rd
#keep everything tab-separated
#tab in the end of each row
#the latter are essential conditions for mothur to parse abundance matrices

```


Annex III – Ubuntu CLI script for Greengenes taxonomy extraction from EMP “.database” file

##All the following commands run only in Ubuntu’s command line interface (CLI)

```
#transpose study1740....shared file  
(in mothur) make.table(shared=study1740...shared)  
#extract OTU size from transposed shared file  
cut -f2 study1740....transpd.shared > OTUsize.list
```

```
#extract OTU_ID and GG_Tax from .database file  
cut -f 1,5 study1740....database > EMP_for_biom.database  
#insert OTUsize.list into .database  
paste OTUsize.list EMP_for_biom.database
```

```
#Swap first two columns  
awk ' { t = $1; $1 = $2; $2 = t; print; } ' EMP12_Database_wOTUsize.database >  
EMP12.database  
#Replace spaces for tabs (or fix the previous command, I am not sure how to...)  
awk -v OFS="\t" '$1=$1' EMP12.database > EMP12_tab.database  
#Change required column labels (joe is a in-UbuntuCLI text editor/viewer)  
joe EMP12_tab.database edit ----> ^KX (joe command to save)  
#Change database name to include a .cons.taxonomy suffix
```

#please check if while running filter.shared you requested makerare=f, as make.biom will not respond well to that

#It is now possible to run mothur’s “make.biom” command

```
#make.biom(shared=,constaxonomy=,metadata=)
```

Annex IV – Mothur script for “Singletome” dataset generation

#removes all OTUs with abundance < mintotal. ie. the singletons.

```
filter.shared(shared=, mintotal=2, makerare=f)
```

#list OTUs with abundance greater than 1.

```
list.otulabels(shared=current)
```

#remove all OTUs with abundance greater than 1 from original shared file.

```
remove.otulabels(shared=first_shared_file, accnos=list_created_in_last_step)
```

Annex V – DNA extraction and cox1 PCR figures

Both processes took place in different times, being that samples 90 to 106 and 110 to 115 were processed in April 2014 and the remainder in November of the same year.

Conditions as described under section 2.3 (Phylogenetic cox1 Analysis).

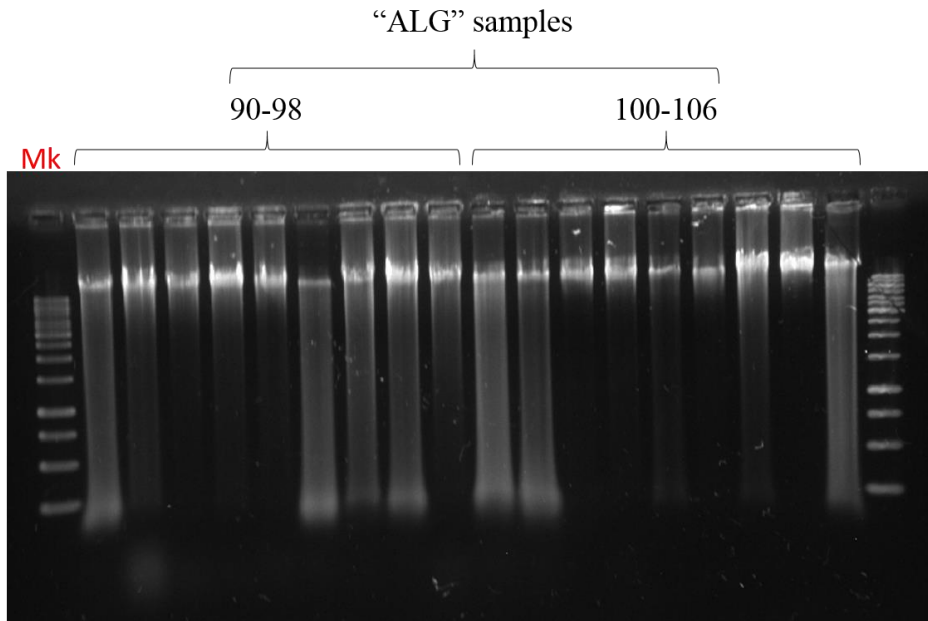


Figure V.1 – Extraction performed in April 2014. ‘Mk’ depicts 1kb marker lane.

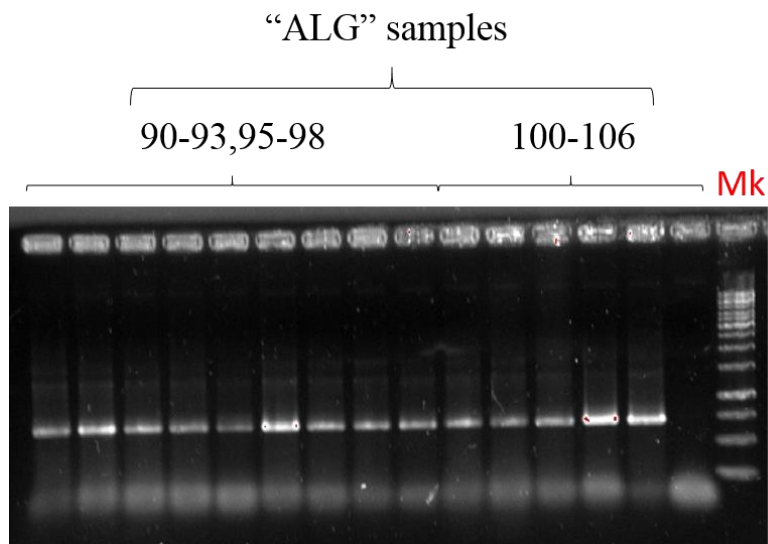


Figure V.2 – Electrophoresis of cox1 PCR products (April 2014). Samples 94 and 98 were subjected to PCR only in November 2014, as seen in **Figure IV.4**. ‘Mk’ depicts 1kb marker lane.

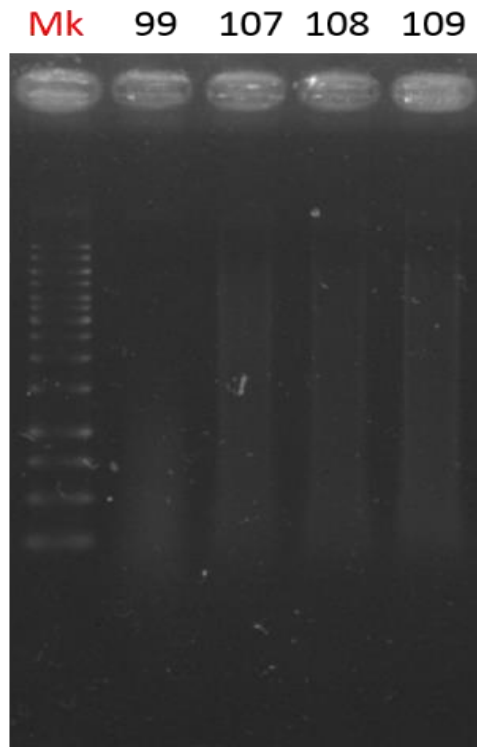


Figure V.3 – Electrophoresis of DNA extraction products (November 2014).
 ‘Mk’ depicts 1kb marker lane.

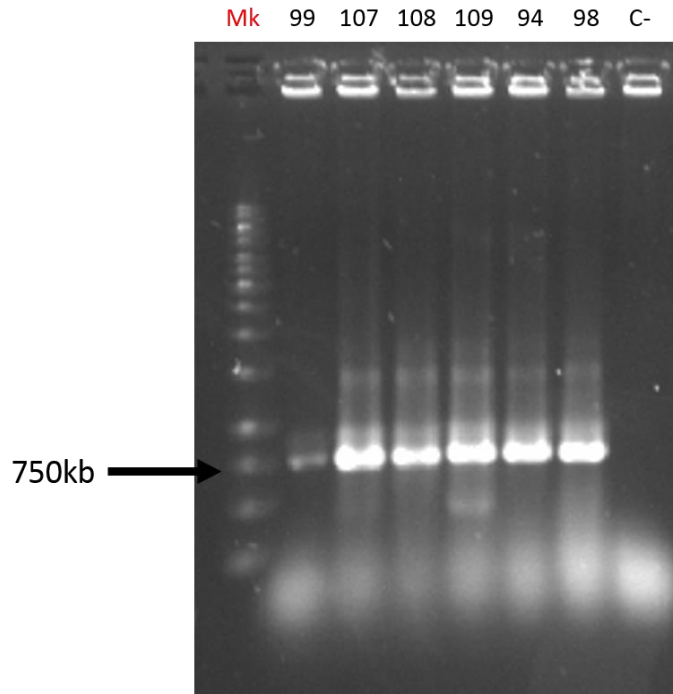


Figure V.4 – Electrophoresis of cox1 PCR products (November 2014). Samples 94 and 98 were re-processed due to low quality sequencing outputs. ‘Mk’ depicts 1kb marker lane and ‘C-’ a negative control (Ultra-pure sterile water).

Annex VI – Non-normalized OTU and sequence numbers table

Phylum	<i>n</i> =7		<i>n</i> =4		<i>n</i> =3		<i>n</i> =12	
	OTUs	Sequences	OTUs	Sequences	OTUs	Sequences	OTUs	Sequences
<i>Crenarchaeota</i>	15 ± 5.68	53 ± 55.36	7 ± 8.85	30 ± 49.71	18 ± 8.08	186 ± 189.21	7 ± 5.42	17 ± 19.03
<i>Acidobacteria</i>	4 ± 6.18	15 ± 29.10	3 ± 1.71	4 ± 2.16	1 ± 1.53	2 ± 2.00	1 ± 1.38	2 ± 3.36
<i>Actinobacteria</i>	9 ± 6.19	21 ± 14.74	7 ± 8.54	12 ± 16.57	9 ± 4.93	25 ± 16.20	10 ± 4.94	27 ± 9.58
<i>Anek6</i>	0 ± 0.76	3 ± 7.18	0 ± 0.50	0 ± 0.50	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
<i>Bacteroidetes</i>	28 ± 20.39	57 ± 51.57	7 ± 8.08	9 ± 10.13	27 ± 17.52	51 ± 31.76	18 ± 5.90	52 ± 42.89
<i>Chlamydiae</i>	1 ± 0.95	1 ± 1.86	0 ± 0.50	1 ± 1.00	0 ± 0.00	0 ± 0.00	1 ± 0.79	7 ± 17.51
<i>Chloroflexi</i>	11 ± 19.92	46 ± 96.53	5 ± 7.14	6 ± 10.59	1 ± 1.00	1 ± 1.00	1 ± 0.67	1 ± 0.67
<i>Cyanobacteria</i>	41 ± 25.29	67 ± 51.67	14 ± 13.61	24 ± 25.18	54 ± 15.50	150 ± 117.23	91 ± 28.65	371 ± 397.85
<i>Firmicutes</i>	3 ± 2.44	4 ± 4.16	1 ± 0.96	1 ± 0.96	1 ± 1.73	1 ± 1.73	1 ± 1.14	1 ± 1.14
<i>Fusobacteria</i>	0 ± 0.53	1 ± 0.79	0 ± 0.50	1 ± 1.00	0 ± 0.58	0 ± 0.58	0 ± 0.00	0 ± 0.00
<i>Gemmatimonadetes</i>	3 ± 3.72	6 ± 8.46	1 ± 1.15	2 ± 1.73	1 ± 1.53	1 ± 1.53	1 ± 1.16	1 ± 1.44
<i>Lentisphaerae</i>	1 ± 1.41	2 ± 2.23	0 ± 0.50	1 ± 1.00	1 ± 1.00	1 ± 1.53	1 ± 0.90	1 ± 1.15
<i>Nitrospirae</i>	5 ± 4.93	21 ± 31.65	4 ± 6.73	14 ± 26.18	4 ± 3.61	10 ± 9.50	3 ± 3.55	5 ± 7.51
<i>PAUC34F</i>	2 ± 3.83	16 ± 31.58	1 ± 1.00	1 ± 1.00	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
<i>Planctomycetes</i>	11 ± 5.72	23 ± 20.75	6 ± 4.79	9 ± 6.24	12 ± 1.73	18 ± 1.53	17 ± 6.44	38 ± 19.65
<i>Poribacteria</i>	1 ± 1.51	1 ± 1.51	0 ± 0.50	1 ± 1.00	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
<i>Alpha Proteobacteria</i>	51 ± 27.99	163 ± 130.41	73 ± 19.26	8619 ± 457.18	46 ± 21.57	2238 ± 1095.18	38 ± 14.53	645 ± 1109.52
<i>Beta Proteobacteria</i>	5 ± 1.60	45 ± 35.15	3 ± 2.65	13 ± 15.50	8 ± 1.00	1401 ± 600.62	3 ± 1.51	7 ± 9.69
<i>Delta Proteobacteria</i>	14 ± 12.32	79 ± 141.97	5 ± 7.23	11 ± 10.97	14 ± 7.00	45 ± 52.54	8 ± 4.91	10 ± 8.08
<i>Epsilon Proteobacteria</i>	0 ± 0.49	5 ± 12.04	0 ± 0.50	0 ± 0.50	0 ± 0.58	0 ± 0.58	0 ± 0.39	0 ± 0.39
<i>Gamma Proteobacteria</i>	110 ± 65.55	1179 ± 630.52	41 ± 20.49	93 ± 34.31	83 ± 37.47	317 ± 189.86	191 ± 40.66	3546 ± 1961.19
<i>Unclassified Proteobacteria</i>	25 ± 6.52	883 ± 408.86	98 ± 6.19	2329 ± 264.26	60 ± 12.06	6685 ± 889.46	94 ± 20.59	6285 ± 2636.63
<i>SBR1.093</i>	3 ± 3.21	5 ± 5.83	2 ± 2.06	4 ± 3.74	3 ± 2.52	9 ± 8.54	1 ± 1.08	2 ± 2.23
<i>Sphrochaetes</i>	0 ± 0.76	0 ± 1.13	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
<i>Thermi</i>	0 ± 0.00	0 ± 0.00	0 ± 0.50	0 ± 0.50	0 ± 0.58	0 ± 0.58	0 ± 0.39	0 ± 0.62
<i>unclassified</i>	148 ± 4.47	8502 ± 983.01	6 ± 1.73	19 ± 4.99	11 ± 4.51	55 ± 16.26	33 ± 10.07	165 ± 71.08
<i>Verrucomicrobia</i>	5 ± 3.65	8 ± 5.89	3 ± 3.59	4 ± 5.48	4 ± 1.53	5 ± 2.52	5 ± 4.03	22 ± 23.44
Total Averages per Sample	480.71	11150.29	278.75	11173.25	340.00	11017.33	523.25	11203.00

Table VI.2 – Average number of sequences and OTUs as detected in the non-normalized dataset per prokaryotic phylum across sponge species.

Annex VII – Metadata table for “ALG” samples

Sample Name	Sponge ID	Sample Location	Latitude(WGS)	Longitude (WGS)	Temperature
ALG12/90	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	20.4
ALG12/91	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	20.4
ALG12/92	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	20.4
ALG12/93	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	20.4
ALG12/94	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/95	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/96	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/97	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/98	<i>Phorbas fictitius</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/99	<i>Cliona viridis</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/100	<i>Cliona viridis</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/101	<i>Cliona viridis</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/102	<i>Cliona viridis</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/103	<i>Cliona celata</i>	Galé alta	37° 04' 09.6	8° 19' 52.1	16.7
ALG12/104	<i>Phorbas fictitius</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	16.7
ALG12/105	<i>Phorbas fictitius</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	16.7
ALG12/106	<i>Phorbas fictitius</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/107	<i>Dysidea fragilis</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/108	<i>Dysidea fragilis</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/109	<i>Dysidea fragilis</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/110	<i>Cliona celata</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/111	<i>Cliona celata</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/112	<i>Cliona celata</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/113	<i>Cliona celata</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/114	<i>Cliona celata</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4
ALG12/115	<i>Cliona celata</i>	Armação baixa	37° 05' 16.7	8° 20' 33.3	20.4