Eukaryotic metabarcoding pipelines for biodiversity assessment of marine benthic communities affected by ocean acidification

Ana Zaida Soto Valdés

2017

IPL
escola superior de turismo e tecnologia do mar
instituto politécnico de leiria

Eukaryotic metabarcoding pipelines for biodiversity assessment of marine benthic communities affected by ocean acidification

Ana Zaida Soto Valdés

Project Report for obtaining the Master's Degree in Marine Resources Biotechnology

Masters project carried out under the guidance of Doctor Owen S. Wangensteen, and co-supervision of Doctor Américo do Patrocínio Rodrigues

2017

Title: Eukaryotic metabarcoding pipelines for biodiversity assessment of marine benthic communities affected by ocean acidification

Acknowledgements

First and foremost, my special thanks goes to my supervisor Dr Owen S. Wangensteen for his unwavering guidance and help, in spite of being a person in great demand, he could dedicate a lot of his time to me. His knowledge and particular commitment hugely increased the speed of my progress, and made it possible to learn so much. I would also like to thank Professor Stefano Mariani who also supported me and my work in every way he could.

At this point, I warmly thank Clélia Afonso, who has always been available for any questions I had. I want to express my gratitude to the Internship Office at the School of Tourism and Maritime Technology (Polytechnic Institute of Leiria) for their patience during all the paperwork, and for the help they gave me to get a scholarship to do my dissertation abroad.

I would also like to thank the University of La Laguna (ULL), for welcoming us into their faculty and offering us a laboratory where we could work with the samples. A special mention must also go to Jose Carlos, since this work could not have been carried out without him.

Although I am always travelling (and every time I go a little bit further), I cannot forget to mention my dear friends. We may not see each other very often, but they are always close to me. Elk, miarma from another mother, muito obrigada por tudo, por seu apoio incondicional, por acreditar em mim, mesmo quando eu estava jogando a toalha. Oksana who has always been there supporting me and motivating me to fight for a future in marine biology, and my colleagues from faculty "Biologuitos", who are the first ones to understand that this career we have chosen is difficult in these times.

Finally, I must express my very profound gratitude to my parents, and to my *Cleverclog*s, for providing me with unfailing support and continuous encouragement throughout my years of study, and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you!

Abstract

The development of high-throughput sequencing technologies has provided ecologists with an efficient approach to assess biodiversity in benthic communities, particularly with the recent advances in metabarcoding technologies using universal primers. However, analyzing such high-throughput data is posing important computational challenges, requiring specialized bioinformatics solutions at different stages during the processing pipeline, such as assembly of paired-end reads, chimera removal, correction of sequencing errors, and clustering of obtained sequences into Molecular Operational Taxonomic Units (MOTUs). The inferred MOTUs can then be used to estimate species diversity, composition, and richness. Although a number of methods have been developed and commonly used to cluster the sequences into MOTUs, relatively little guidance is available on their relative performance.

We focused our study in the benthic community from a natural CO_2 vent present in the Canary Islands, as it can be used as a natural laboratory in which to investigate the impacts of chronic ocean acidification. Here, we propose a pipeline for studying this community using a fragment of the mitochondrial cytochrome c oxidase I (COI) sequence. We compared two DNA extraction methods, two clustering methods and validated a robust method to eliminate false positives.

We found that we can obtain optimal results purifying DNA from 0.3 g of sample. Using the step-by-step aggregation algorithm implemented in SWARM for clustering yields similar results as using the Bayesian clustering method of CROP, in much less time. We introduced the new algorithm MINT (Multiple Intersection of N Tags), in order to eliminate false positives due to random errors produced before or after the sequencing. Our results show that a fully-automated analysis pipeline can be used for assessing biodiversity of marine benthic communities using COI as a metabarcoding marker in an objective, accurate and affordable manner.

Keywords: Enviromental DNA (eDNA), CO_2 vent, COI, pipeline, clustering, MINT

Resumo

O desenvolvimento de tecnologias de sequenciamento de alto rendimento proporcionou aos ecologistas uma abordagem eficiente para avaliar a biodiversidade nas comunidades bentônicas, particularmente com os recentes avanços nas tecnologias de *metabarcoding* utilizando *primers* universais. No entanto, a análise desses dados de alto rendimento apresenta desafios computacionais importantes, que requerem soluções de bioinformática especializadas em diferentes estágios durante o processamento da *pipeline*, tais como a montagem de *paired-end reads*, eliminação de quimeras, correção de erros de sequenciamento e agrupamento de sequências obtidas em Unidades Taxonômicas Operacionais Moleculares (MOTU). As MOTUs inferidas podem então serem usadas para estimar a diversidade, composição e riqueza das espécies. Atualmente, há pouca informação acerca do desempenho relativo com referência ao agrupamento de sequências em MOTUs, apesar do fato de que vários métodos foram desenvolvidos.

O estudo foi focado em uma comunidade bentônica a partir de uma fumarola de CO_2 natural presente nas Ilhas Canárias, pois pôde ser utilizado como um laboratório natural para investigar os impactos da acidificação crônica dos oceanos. Aqui, propomos um *pipeline* para estudar esta comunidade utilizando um fragmento da sequência mitocondrial de citocromo *c* oxidasa I (COI). Comparamos dois métodos de extração de DNA, dois métodos de agrupação e se validou um método robusto para eliminar os falsos positivos.

Descobrimos que podemos obter ótimos resultados purificando o DNA a partir de 0.3g de amostra. Usando o passo-a-passo do algoritmo de agregação implementado em SWARM, podemos comprovar que produz resultados similares quando comparados com o método de agrupamento Bayesiano de CROP, além disso, em muito menos tempo. Introduzimos um novo algoritmo MINT (Intersecção múltipla de N Tags), com a finalidade de eliminar os falsos positivos devido aos erros aleatórios produzidos antes ou após o sequenciamento. Nossos resultados mostram que se pode utilizar um sistema de análises completamente automatizado para avaliar a biodiversidade das comunidades bentônicas marinhas utilizando COI como marcador de uma maneira objetiva, precisa e acessível.

Table of Contents

List of Figures

Chapter 1. Introduction

Chapter 2. Material and Methods

Chapter 3. Results

Chapter 4. Discussion

List of Abbreviations

ppm – part(s) per million

DIC – dissolved inorganic carbon

My – million years

comDNA – community DNA

PCR – polymerase chain reaction

HTS – high-throughput sequencing

OTU – operational taxonomic unit

MOTU – molecular operational taxonomic unit

COI – mitochondrial cytochrome *c* oxidase I gene

bp – base pair

NCBI – National Center for Biotechnology Information

MINT – Multiple Intersection of N Tags

MDS – multidimensional scaling

CBOL – Consortium for the Barcode of Life

BOLD – Barcode of Life Data Systems

Chapter 1.

Introduction

1.1. Biodiversity monitoring

Biological monitoring techniques which allow us to obtain accurate data on species distributions and population sizes on ecologically relevant scales of time and space are crucial for a correct management of natural ecosystems. Species monitoring has traditionally relied on physical identification of species by means of visual surveys and counting individual abundances, in the field or in the laboratory, using distinct morphological characters. However, in many cases these techniques fall short of yielding efficient and standardized surveys, due to incorrect identifications arising from, among others, phenotypic plasticity, species with similar appearance in juvenile stages (Thomsen and Willerslev, 2015) or occurrence of cryptic species complexes (Knowlton, 1993). Additionally, traditional monitoring techniques have sometimes proven to be invasive on the species or ecosystems under study, such as marine surveys that have relied on highly destructive techniques (Baldwin *et al.*, 1996; Jones, 1992). Furthermore, morphological identification is heavily dependent on taxonomic expertise, which is often lacking or in rapid decline (Hopkins and Freckleton, 2002; Wheeler *et al.*, 2004), a scenario that is better known as "the taxonomic impediment" (Wheeler *et al.*, 2004). All such limitations of traditional biodiversity monitoring have created demand for alternative approaches.

Obtaining information of species, populations and communities by retrieving DNA from environmental samples holds the potential of combating many of these challenges associated with biodiversity monitoring (Baird and Hajibabaei, 2012; Kelly *et al.*, 2014). The fact that DNA from organisms of a complex community can be sampled, extracted and analyzed, has been a major technological and scientific breakthrough within the last decade. Within a single standardized sample, DNA from entire communities across taxonomic groups can potentially be analyzed simultaneously (Thomsen and Willerslev, 2015).

The content of a community DNA (comDNA) sample is typically analyzed by amplification using polymerase chain reaction (PCR) and subsequent DNA sequencing. The amplification is done by a multiple-species (multiple-taxon) approach using generic primers for a given focal group of organisms. Especially the fast advancing high-throughput sequencing (HTS) technologies have made comprehensive biodiversity surveys possible for limited effort and costs (Shokralla *et al.*, 2012). This has made the multiple-species comDNA approach especially powerful by DNA metabarcoding – mass DNA sequencing for the simultaneous molecular identification of multiple taxa in a complex sample (Taberlet

et al., 2012a). Although similar in principle to classical DNA barcoding of simple tissue DNA extracts (Hebert *et al.*, 2003), the practical approach and target sequences are very different.

1.2. DNA Metabarcoding

Taberlet *et al.* (2012b) introduced the term DNA metabarcoding to designate high-throughput multispecies identification using the total or typically degraded DNA extracted from an environment sample or from bulk samples of entire organisms. Metabarcoding differs from metagenomics in several ways as metagenomics describes the functional and sequence-based analysis of the collective genomes contained in an environmental sample (Riesenfeld *et al.*, 2004) whereas metabarcoding aims to study a subset of genes / gene. From methodology point of view, metagenomics approach includes preparation of shotgun (random) libraries for sequencing while metabarcoding is based on amplicon sequencing. Metagenomics approach generally used to get more insights about the interaction between species within an ecosystem (taxonomic and functional information). Metabarcoding approach is mainly used to document / characterize species diversity in the ecosystem and it can have better coverage to identify rare taxa within an ecosystem (Pavan-Kumar *et al.*, 2015).

The success of DNA-barcoding relies on the coexistence of two factors, one natural and other technological: (1) a genetic marker universally present in every species, which could be easily sequenced using standardized protocols. This marker should have enough sequence variability to allow distinction among related species but must be surrounded by regions conserved enough so that universal primers could be designed. And (2) a massive public database containing the known sequences of this marker for the maximum possible number of different species, which must be searchable by automated algorithms, so that unknown sequences could be matched to a known species. The identification by DNA-barcoding is evidently as good and reliable as complete and accurate this reference database is (Wangensteen and Turon, 2015).

The metabarcoding technology has now been successfully used for characterizing the microscopic biodiversity present in relatively homogeneous substrates, such as plankton (Pearman and Irigoien, 2015; de Vargas *et al.*, 2015), soil (Epp *et al.*, 2012; Schmidt *et al.*, 2013), marine sediments (Chariton *et al.*, 2010; Fonseca *et al.*, 2014;

Guardiola *et al.*, 2015; Lejzerowicz *et al.*, 2015) or gut contents (Leray *et al.*, 2013). However, its applications to characterize more complex and heterogeneous substrates such as marine rocky communities, for example, to identify differences produced by a natural CO_2 vent, are still little explored.

1.3. Bioinformatics pipelines for metabarcoding

Data processing is currently a bottleneck in metabarcoding projects. The number of reads per study has been continuously increasing since the introduction of high-throughput sequencing (HTS) methods, and it is expected to rise as sequencing technologies advance. Data processing must consider the peculiarities of the taxonomic marker, the sequencing instrument and chemistry, as well as the experimental needs, such as the requirements for sample multiplexing (Bálint *et al.*, 2014).

The metabarcoding procedure will introduce random errors, both during the amplification and sequencing steps, which will generate an initial dataset of sequences considerably different from the sequences present in the original sample (Figure 1C). The analysis pipeline typically starts with a denoising stage, which is a stringent quality control step and has a strong effect on downstream analyses (Schloss *et al.*, 2011). In this first step, the raw data are corrected by clustering the flowgrams using frequency-based heuristics or approximate likelihood with empirically derived error distributions, or fasta-formatted files are corrected based on alignments (Reeder and Knight, 2010). The result is that all sequences with putative random errors will be removed from downstream analyses.

Denoising procedures may work well when using short metabarcoding markers with low natural diversity. However, they are not very useful when using longer markers with high natural variability (such as, for example, COI), because too high a proportion of the initial reads will be removed by the denoising algorithm. A better alternative is to cluster the obtained sequences into operational taxonomic units (OTUs) (Figure 1D), which will also have a strong impact on the number of observed and estimated community richness, but it will keep a higher number of useful reads, providing more accurate information about relative abundances. Whether two sequences are clustered together into the same OTU depends both on their similarity and other sequences in the studied dataset, as well as the selected clustering algorithm (Schloss *et al.*, 2011).

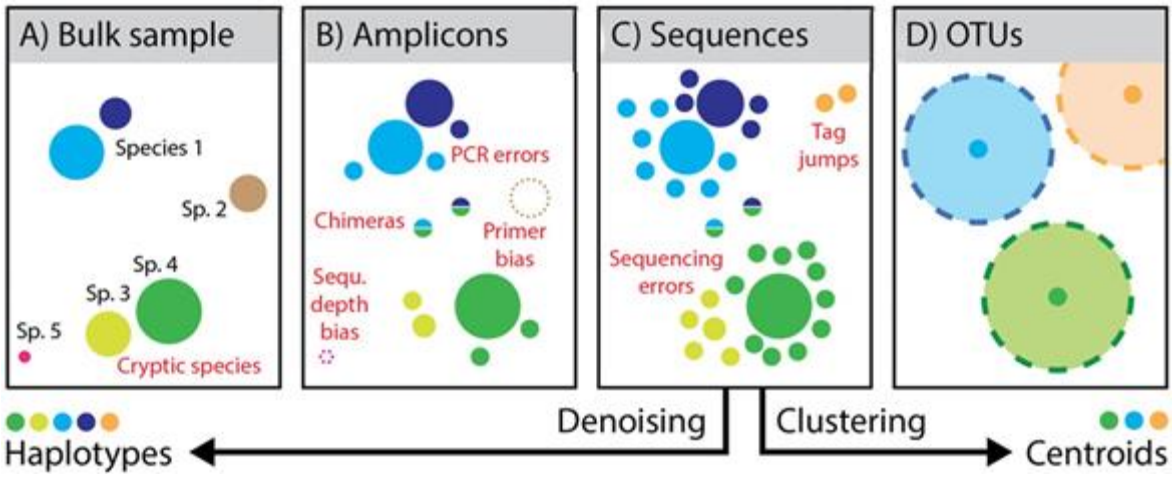

Figure 1. Representation of the analysis of a metabarcoding sequence dataset including random errors (C), using either denoising or clustering procedures

Heuristic algorithms have been developed to cluster sequence reads into OTUs, allowing analysis of large datasets without access to highly powerful computer resources. The clustering rate is typically much faster and the maximum memory requirement is considerably lower, but the heuristic algorithms do not work as accurately as the best of the classic hierarchical clustering tools (Schloss *et al.*, 2011). Additionally, different algorithms can produce surprisingly divergent results, particularly in OTU numbers and diversity estimates.

In this study, we compared two clustering algorithms to estimate which algorithms and parameters provide the most reliable results for characterizing marine benthic diversity: CROP and SWARM.

CROP (Hao *et al.*, 2011) adopts an unsupervised probabilistic Bayesian clustering algorithm and uses a soft threshold (different similarity for the diverging branches of the phylogenetic tree) for defining the OTUs, which is probably more accurate for reflecting real species diversity, and it also reduces the effects of PCR and sequencing errors in inferring OTUs (Wei *et al.*, 2016), by clustering the sequences affected by these errors to their mother sequences, instead of removing them altogether. CROP has the big disadvantage of requiring time-consuming calculations, even when using powerful multiprocessor computing clusters. Moreover, given the heuristic nature of Bayesian algorithms, the reproducibility of CROP is somehow low, so that different runs on the same input dataset might result in variable numbers of resulting molecular operational taxonomic units (MOTUs) (Wangensteen and Turon, 2015).

SWARM (Mahé *et al.*, 2014) was the other clustering algorithm used, which is based in a step-by-step aggregation procedure. This method has the advantage of being deterministic, so it yields robust and repeatable results. Every resulting cluster is a network of sequences, somehow comparable to the haplotype networks used in population genetics. However, it also bears the arbitrariness of having to choose a value of the distance threshold for including a sequence into a growing cluster. SWARM is an iterative process which also requires large amounts of computing time for completion (Wangensteen and Turon, 2015).

We compared different metabarcoding analysis pipelines, based on these two clustering algorithms, for selecting the best method to assess the biodiversity present in marine benthic communities affected by ocean acidification.

1.4. Ocean acidification

Increasing atmospheric CO_2 is causing unprecedented changes in seawater chemistry (Guinotte and Fabry, 2008). Atmospheric CO_2 concentrations are projected to increase by 0.5 % per year throughout the XXI century, this rate of change is around 100 times faster than has occurred in the past 650,000 years (Meehl *et al.*, 2007). The current concentration of atmospheric CO_2 is around 400 ppm, having increased from pre-industrial levels of 280 ppm, and emissions scenarios predict a continued increase to ~750-1000 ppm by 2100 (Meehl *et al.*, 2007; Feely *et al.*, 2009). The rising atmospheric CO_2 levels drive changes in seawater chemistry and lower pH (Gattuso and Buddemeier, 2000). The oceans play a crucial role in the global carbon cycle, forming an important sink for anthropogenic CO_2. The carbonate equilibrium in seawater is the balance between dissolved inorganic carbon (DIC) species; $CO_2(aq)$, HCO_3^- and C ions. Changes in pH control this equilibrium; therefore, increasing CO_2 will increase the total DIC and shift the proportion of DIC speciation in seawater, with widespread biological consequences. When CO_2 dissolves in seawater it reacts with water molecules to form carbonic acid:

$$[CO_2] + [H_2O] \rightleftarrows [H_2CO_3]$$

Carbonic acid then dissociates to bicarbonate and hydrogen ions:

$$[H_2CO_3] \rightleftarrows [H^+] + [HCO_3^-]$$

This leads to a lowering of pH and other chemical changes collectively termed as ocean acidification (Caldeira and Wickett, 2003). The increase in hydrogen ion concentrations causes some carbonate ions to react with hydrogen to become bicarbonate:

$[H^+] + CO_3^= \rightleftarrows [HCO_3^-]$

Therefore, the dissolution of CO_2 in seawater increases concentrations of hydrogen ions, carbonic acid and bicarbonate whilst decreasing the concentration of carbonate.

The dissolution of CO_2 in seawater decreases carbonate ion concentrations, shifting the equation to the right, impeding the formation of carbonate minerals and promoting dissolution. This is likely to negatively impact a wide range of calcifying marine biota (Leclercq *et al.*, 2000; Riebesell *et al.*, 2000; Gazeau *et al.*, 2007)).

This dissolution of carbonate minerals produces carbonate ions that can react to consume hydrogen ions which in turn counteracts some of the hydrogen generating effects of CO_2 enrichment (Morse *et al.*, 2007). However, as CO_2 is being absorbed so rapidly, it is unlikely that this natural buffering capacity of the ocean surface will be able to prevent a substantial reduction in ocean pH (Raven *et al.*, 2005).

Over the past 200 years the oceans have absorbed about one third of the total human CO_2 emissions (Sabine *et al.*, 2004) resulting in gradual acidification of seawater. Ocean pH has dropped by 0.1-0.2 pH units (corresponding to a 30% increase in hydrogen ion concentration) since the Industrial Revolution and, under predicted emission scenarios, is expected to drop another 0.3-0.4 units, from pH 8.2-8.1 to 7.8-7.6, by the end of this century (Caldeira and Wickett, 2003; Orr *et al.*, 2005; Meehl *et al.*, 2007).The current rate of CO_2 release into the atmosphere is capable of driving a magnitude of ocean geochemical changes potentially unparalleled in at least ~300 My of Earth history (Hönisch *et al.*, 2012). Current anthropogenic trends in ocean acidification already exceed the level of natural variability by up to 30 times on regional scales and are detectable in many areas of the world's ocean (Friedrich *et al.*, 2012).

Near-future acidification is predicted to have dramatic impacts on some marine species with cascading biological consequences for marine ecosystems (Abbasi and Abbasi, 2011). These studies indicate that variation in the sensitivity of organisms to ocean acidification is likely to disrupt the species balance of communities and influence species interactions which could potentially lead to unforeseen impacts on marine ecosystems.

1.5. Natural CO_2 Vents

Natural CO_2 vents can be used as natural laboratories in which to investigate the impacts of chronic ocean acidification. They have advantages (e.g. populations studied in a natural environment, biotic interactions taken into consideration, consequences of long-term acidification) and limitations (e.g. high pH variation, open system, obscuration of direct versus indirect effects) in comparison with experiments based in a laboratory setting (Barry *et al.*, 2010). Ecological consequences of ocean acidification remain unexplored and, until now, the vast majority of studies have been performed in laboratories and are characterized by short-term and univariate experimental approaches (Hernández *et al.*, 2016).

Although, laboratory manipulation experiments simulating current and future pCO_2 concentrations are a crucial tool to ensure causality, a great proportion of them, do not make use of the appropriate experimental design (Cornwall and Hurd, 2016). Other problem of these experimental approaches is that they do not incorporate the understanding of the carbon chemistry environment that is naturally experienced by the study organism (McElhany and Shallin Busch, 2013; Hernández *et al.*, 2015). All these make it very difficult to draw an ecological relevant interpretation from some of the current data, and the need for truly rigorous experimental designs has been recently highlighted (Cressey, 2015).

An alternative way to improve results of laboratory experiments is to study naturally occurring CO_2 vents, however field observations at such spots are scarce and very few studies investigating the biological effects of these natural experiments have been performed to date (Hernández *et al.*, 2016). There are several studies on the effects of *in situ* CO_2 leakage on benthic community (Raulf *et al.*, 2015; Hall-Spencer *et al.*, 2008; Meadows *et al.*, 2015), microbial communities in costal sediments from volcanic shores (Oppermann *et al.*, 2010), macroalgal communities from volcanic vents (Porzio *et al.*, 2013) or even on specific organisms (Rodolfo-Metalpa *et al.*, 2010; Lucey *et al.*, 2016). However, changes in the composition of the whole eukaryotic communities have never been studied in these natural laboratories using an integrative approach based on metabarcoding.

1.6. Aims of the study

In the present work, we introduce an enhanced metabarcoding methodology for characterizing complex communities inhabiting natural CO_2 vents. We tested the suitability of this methodology to study the eukaryotic biodiversity from different communities sampled from four marine locations in the southeast coast of La Palma Island, Canary Islands (Spain), along a gradient of distances to a natural CO_2 vent. The main objectives of this study are:

- To validate a robust method to eliminate false positives from the initial metabarcoding sequence dataset.
- To compare two types of clustering algorithms: CROP and SWARM
- To compare two DNA extraction methods which use different sample weights

Chapter 2.

Material and Methods

2.1. Area of study

The study site is located in the southeast coast of La Palma Island, Canary Islands, in a location called “Fuencaliente”. Besides the vent site (1) and the nearby area (2), we also selected two more control locations (3 and 4), representing the most common and characteristic coastal ecosystems of the subtropical Eastern Atlantic Archipelagos (Fig. 2).

At each location, samples from the communities were collected during September 2017 by scraping the surface with a trowel from a 25x25 cm quadrat placed randomly. In each zone 6 replicates were carried out, all of them from the subtidal zone within a range of 1 meter depth. All samples were placed underwater inside plastic containers of 500 ml. Seawater was replaced by 96% ethanol using a 63 µm sieve, and the samples were stored at room temperature until pre-treatment.

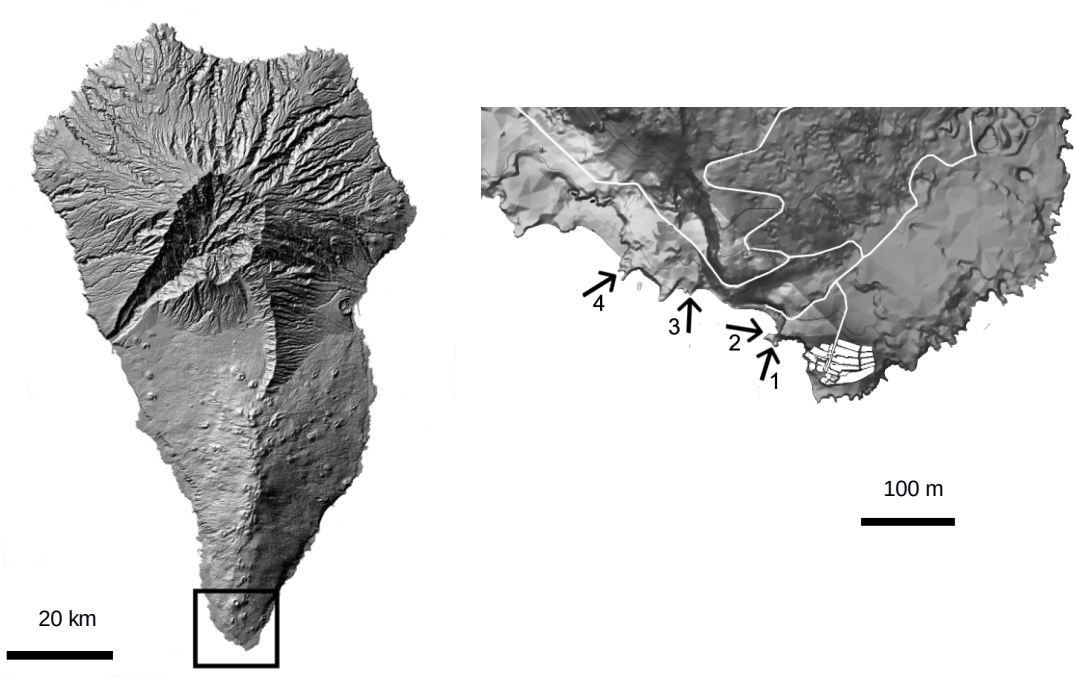

Figure 2. Location of the studied areas in La Palma Island, Canary Islands. Arrows show the sites that were sampled. (1) Vent, (2) Transition, (3) Control and (4) Control Echentive.

2.2. Sample pre-treatment and DNA extraction

Samples were separated into two size fractions (A: > 1 mm; B: 63 µm - 1 mm) using a column of stainless steel sieves (www.cisa.net), washing thoroughly under high-pressure

freshwater. All separated samples were then recovered in 96% ethanol, homogenized using a 600 W hand blender and stored at 5 ºC until DNA extraction. All equipment was thoroughly washed and cleaned with sodium hypochlorite between successive samples.

0,3 g of each homogenized sample fraction were purified using PowerSoil® DNA Isolation Kit (www.mobio.com). DNA concentration of purified extracts was assessed in a Qubit fluorometer (www.lifetechnologies.com). The DNA extracts were stored at -20 ºC.

With some of the samples, two DNA extraction methods were compared: PowerSoil® DNA Isolation Kit, where we only used 0,3 g of the homogenized sample, and PowerMax® Soil DNA Isolation Kit, where we used 10 g of homogenized sample, both as indicated in the protocol. For this comparison, we extracted DNA from 6 samples with the PowerMax® Soil DNA Isolation Kit (V11, V12, V13, V41, V42, V43).

2.3. PCR

We amplified a mitochondrial gene fragment: the mitochondrial cytochrome *c* oxidase I gene (COI), which has been adopted as the standard 'taxon barcode' for most animal groups (Hebert *et al.*, 2003) and is by far the most represented in public reference libraries (Leray *et al.*, 2013).

We used a new highly degenerated primer set called Leray-XT, which includes the reverse primer jgHCO2198 5'-TAIACYTCIGGRTGICCRAARAAYCA-3' (Geller *et al.*, 2013) and a novel forward primer mlCOIintF-XT 5'-GGWACWRGWTGRACWITITAYCCYCC-3', modified from the mlCOIintF primer (Leray *et al.*, 2013) with two more wobble bases incorporated and two inosine nucleotides in the most degenerate positions, for increased universality across eukaryotic groups (Wangensteen *et al.*, 2017). Sample tags (8 bp) were attached to each primer for identifying each sample, in order to prepare a multiplexed mix which was sequenced together in a single run. Amplification of COI used 10 µl of AmpliTaq Gold DNA polymerase, with 1 µl of each 5 µM forward and reverse 8-base tagged primers, 0.16 µg of bovine serum albumin, 5.84 µl of Milli-Q water and 2 µl of purified DNA in a total volume of 20 µl per sample.

The PCR profile included a denaturing step of 10 min at 95 ºC, 35 cycles of 94 ºC 1 min, 45 ºC 1 min and 72 ºC 1 min and a final extension of 5 min at 72 ºC. After PCR, quality

of amplifications was assessed by electrophoresis in agarose gel. All PCR products were purified using MinElute column-based purification kit (QIAgen) (www.qiagen.com).

In order to remove false positives in the future steps and to correct the bias introduced by primer tags (O'Donnell *et al.*, 2016), 3 PCR replicates were made from each fractioned sample, using different sample tags.

2.4. Illumina library preparation and sequencing

Amplified gene fragments were prepared for Illumina sequencing following a one-step PCR-based approach. Illumina adaptor tails and library tags were added using a ligation-based procedure (NEXTflex PCR-Free DNA-Seq Library Prep Kit, Bioo Scientific, USA) prior to sequencing. Four Illumina libraries were built from the DNA amplicon pools (equimolar concentration). All libraries were sequenced together in an Illumina MiSeq platform (Illumina, Inc., San Diego, CA, USA) using a v3 reagent kit 2×250 bp. Sequencing workflow followed manufacturer's protocols and was performed at the genomics laboratoy facility at the University of Salford (Manchester, United Kingdown).

2.5. Bioinformatic analysis

The bioinformatic analyses were carried out in a BioLinux environment, using different alternative versions of a pipeline based on OBITools (Boyer *et al.*, 2016). The analysis pipeline (Figure 3) consisted of three steps: (1) data pre-processing including: paired-end alignment, removal of primers, sequencing adaptors and demultiplexing, and removal of chimeras (artificial amplicons stemming from two or more parent sequences, formed by incomplete template extension), (2) clustering of sequences into MOTUs and (3) taxonomic assignment of the representative sequences of each MOTU using the ecotag algorithm.

The output of the sequencer is a file in FASTQ format, a text file which includes the DNA sequences (reads) and the quality information for each base, so quality control was performed using the software FASTQC (Andrews, 2010). The length of the raw reads was trimmed to a good quality control length between 200 and 240 bp. The next bioinformatic analyses were based on the OBITools metabarcoding software (Boyer *et al.*, 2016). The reads with quality score higher than 40 after the paired-end assembly using

illuminapairedend were kept. The aligned datasets were demultiplexed using ngsfilter. A length filter (obigrep) was applied to the assigned reads (303 – 323 bp) and reads including only A, C, G or T bases were selected. Strictly identical reads were then dereplicated (using obiuniq) and chimeric sequences were detected and removed using the uchime_denovo algorithm implemented in vsearch (http://github.com/torognes/vsearch).

Species are defined operationally as a cluster of similar sequences, and the clusters are known as Molecular Operational Taxonomic Units (MOTUs). Sequences can be clustered into MOTUs even if the sequences have not yet been linked to a taxonomic name. A clustering algorithm is needed to group the related sequences into clusters, so that the resulting MOTUs reflect the real species diversity present in the samples as accurately as possible. The MOTUs were then delimited using two methods: (1) the Bayesian clustering algorithm implemented in CROP (Hao *et al.*, 2011) which generates clusters within user-defined lower (-l) and upper (-u) bound levels of similarity to account for differences in rates of sequence evolution among taxonomic groups. We defined -l = 1.5 and -u = 2.5 (which correspond to an initial clustering level at 95% similarity) because it was shown to create OTUs that closely reflect morphological species grouping by providing the lowest frequency of false positives (splitting of taxa) and false negatives (lumping of taxa) (Leray and Knowlton, 2015). And (2) the SWARM algorithm (Mahe *et al.*, 2015) based in a step-by-step aggregation procedure with a resolution parameter of d = 13 (maximum number of differences allowed to cluster a sequence within a MOTU in each aggregation step). This value was selected from previous works using COI metabarcoding in similar complex marine benthic communities (Wangensteen *et al.*, 2017). For long amplicons and hypervariable markers such as COI, high d values must be used.

The taxonomic assignment of the representative sequences for each MOTU was performed using ecotag (Boyer *et al.*, 2016), which uses a local, customizable, reference database and a tree based approach (using the NCBI taxonomy database) for assigning sequences. Ecotag searches the best hit in the reference database and builds the set of all sequences in the database as similar or more to the best hit as the query sequence is. Then, the taxon that is the most recent common ancestor to all these sequences in the NCBI taxonomy database is assigned to the MOTU. With this procedure, the assigned taxonomic rank varies depending on the similarity of the query sequences and the density of the database (Wangensteen and Turon, 2015). Since we were interested only in Eukaryotic diversity, all MOTUs assigned by ecotag to prokaryotes or to the root of the Tree of Life were removed from the analyses.

The phylogenetic assignment is not exempt of problems. The most obvious being that commonly accepted taxonomic trees (specially those branches relying upon doubtful morphological traits) are often in conflict with molecular phylogenies. Moreover, some molecular phylogenies based on different molecular markers can often be in conflict as well. Even in the presence of phylogenetically coherent taxonomic information in the reference database, the problem remains that the phylogenetic tree obtained from using just a short fragment of marker (unavoidable in metabarcoding) is not expected to match perfectly the real phylogeny. Thus, incongruences and ambiguities are common in any phylogenetic procedure intended to assign metabarcoding data using reference sequences with low similarities, potentially resulting in many query sequences being assigned to higher taxonomic levels or even to wrong taxa. The only way to improve these assignments would be adding as many sequences as possible to the reference databases in order to grow them denser. When any sequence in the reference database matches the query closely (with an identity of 95% or higher in COI), the assignment procedure is reliable, fast and useful. Sadly, the reliability drops quickly as the similarity with the matching references decreases, so that even assignments at the phylum level with identity percentages less than 85% in COI should be considered questionable and treated with caution (Wangensteen and Turon, 2015).

In multiplexed metabarcoding analyses, the sequence data may contain sequencing noise, PCR chimeras (Lenz and Becker, 2008), contaminant sequences, and PCR errors. Eliminating all these error sequences from final data analysis is prerequisite before assigning taxon to the sequences. Some of these errors are eliminated during the pipeline, but the sequences from contaminations due to cross-sampling are not so easy to remove during this step. That is, a small (but usually not negligible) number of reads will be assigned to the wrong sample. The cross-sampling rates will depend on the total abundance of the MOTU. Thus, a small number of reads of the most abundant MOTUs of the dataset will appear randomly spread through most of the sample columns in the raw final matrix resulting from the MOTU clustering procedure. Quantitatively, this would be unimportant, but it represents a big difference for methods based on presence/absence. So, MINT was used to eliminate these sequencing errors (false positives). Multiple intersection of N tags (MINT) consist in an R script (https://www.rstudio.com/) that selects or eliminates replicates that do not meet the rule that has been assigned.

During the sample amplification, 3 PCR replicates were done from each sample to assess the reliability of the sequences that were retained after doing the pipeline. In doing so, we relied on the assumption that sequences that appear in multiple PCR replicates are

more likely to be real rather than randomly generated artefactual sequences. In order to eliminate these false positives, MINT was applied to these replicates in two different ways.

In the MINT-all strategy, MINT was used to eliminate all the MOTUs that only appeared in 1 or 2 of the replicates, so that only the MOTUs that had some reads in all 3 replicates were kept. As an alternative strategy, MINT-two was used to eliminated all MOTUs that only had reads in one replicate, so that MOTUs that appeared in 2 or 3 replicates were kept (Figure 4).

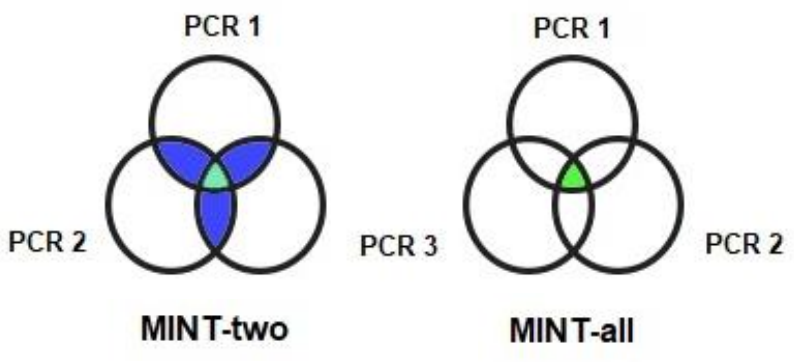

Figure 4. Schematic representation of the 3 PCR replicates showing the MINT strategy. MINT-two: MOTUs that appeared in at least 2 of the 3 replicates. MINT-all: MOTUs that appeared in all 3 replicates.

2.6. Statistical analysis

All analyses were performed in R v 3.3.0 (https://www.R-project.org/). To carry out the community ecology analyses we used the R package vegan (Oksanen *et al.*, 2016).The α-diversity patterns were analyzed applying the rarefaction methods (function rrarefy) (Sanders, 1968) using the number of reads as a proxy for sample size. Pairwise differences in MOTU diversity among sites were evaluated using the nonparametric multiple comparison function (*dunn.test*) implemented in the R package dunn.test. The dunn.test is equivalent to the Kruskall–Wallis and pair-wise Mann–Whitney post hoc tests with Bonferroni correction. MOTU compositional differences among sites were examined using metric multidimensional scaling (MDS).

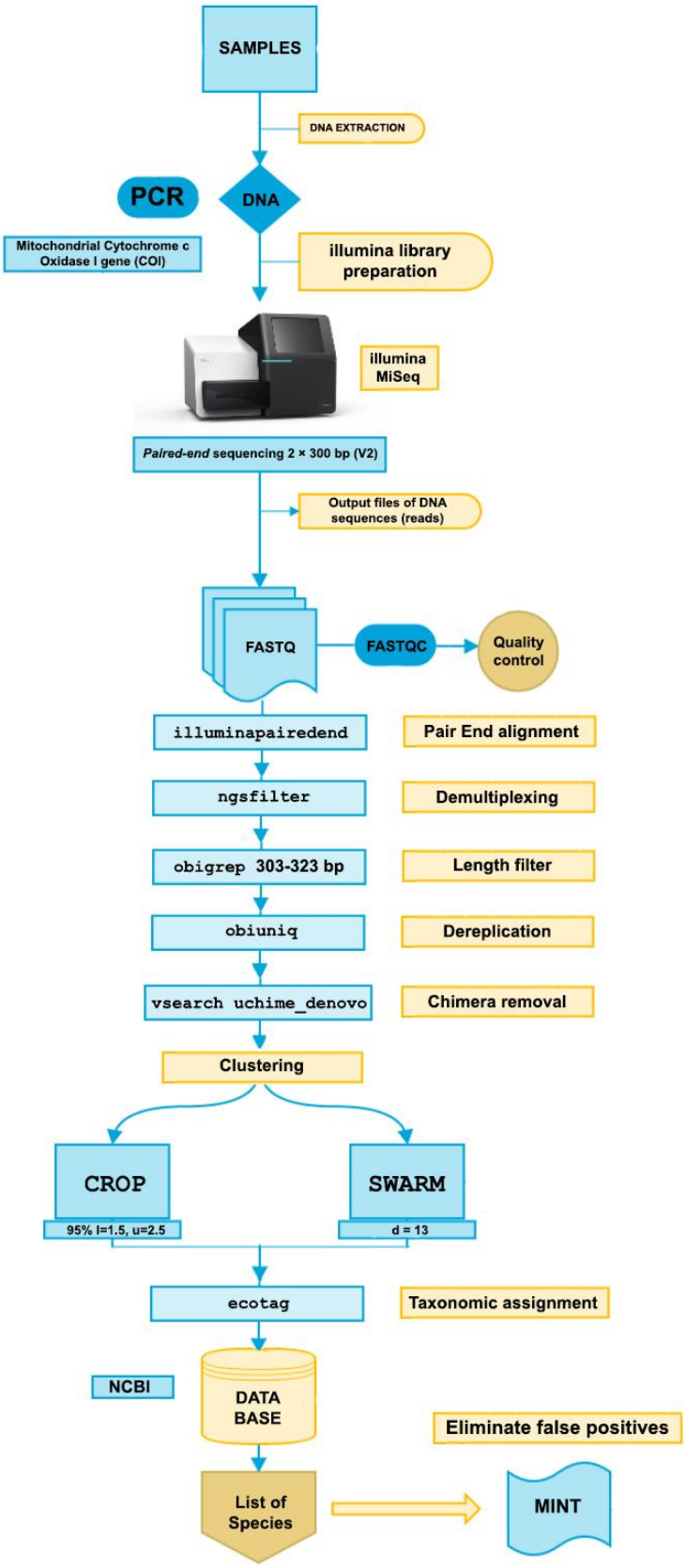

Figure 3. Diagram of the pipeline showing all the steps.

Chapter 3.

Results

3.1. Abundance of MOTUs

We metabarcoded a total of 24 samples (6 from each site: Vent, Transition, Control and Control Echentive). Each sample was separated into 2 size fractions A (> 1 mm) and B (63 µm - 1 mm). Each fraction was analyzed in 3 technical replicates, for a total of 144 technical samples. After the refining procedures, our final dataset from the CROP pipeline comprised of a total of 5,484,370 reads, with an average of 38,086 reads per replicate (range: 1,862 – 95,334). The total reads from the SWARM pipeline were 5,501,299, with an average of 38,203 reads per replicate (range: 1,862 – 95,384).

The number of total MOTUs detected from all samples using CROP clustering before MINT was 6,372. After refining with CROP-MINT-all a total of 2,456 MOTUs remained, from which 1,166 (47.5%) could be assigned to the level of Phylum or lower. Using CROP-MINT-two refining procedure, a total of 3,864 MOTUs remained, from which 1,525 (39.5%) could be assigned to the level of Phylum or lower. Using the SWARM pipeline, the total number of MOTUs before MINT was 9,475. After SWARM-MINT-all we obtained a total of 2,867, from which 1,490 (52.0%) could be assigned to the level of Phylum or lower. Using SWARM-MINT-two we obtained 6,030 MOTUs, from which 2,664 (44.2%) could be assigned to the level of Phylum or lower. Figure 5 shows the abundance of reads for all replicates assigned

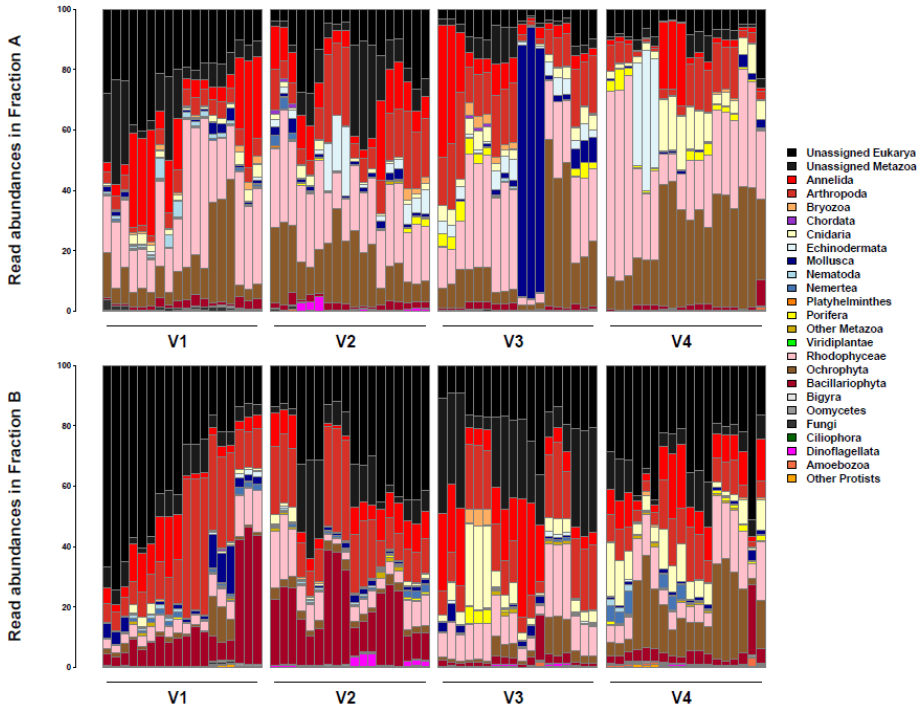

Figure 5.1. Patterns of abundance of metabarcoding reads by phylum for all technical replicates of samples obtained from the fractions A (> 1mm) and fractions B (63µm – 1mm) from the 4 sites Vent (V1), Transition (V2), Control (V3) and Control Echentive (V4). Results for clustering using CROP are shown.

to major eukaryotic groups at a level of Phylum or lower using both clustering methods: CROP represented in Figure 5.1 and SWARM represented in Figure 5.2.

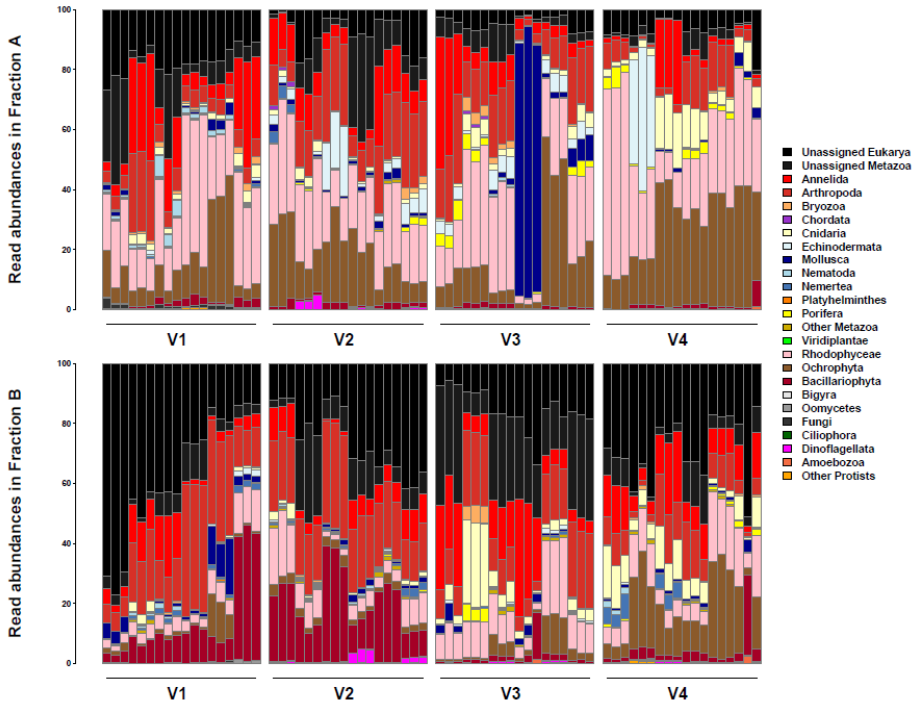

Figure 5.2. Patterns of abundance of metabarcoding reads by phylum for all technical replicates of samples obtained from the fractions A (> 1mm) and fractions B (63µm – 1mm) from the 4 sites Vent (V1), Transition (V2), Control (V3) and Control Echentive (V4). Results for clustering using SWARM are show n.

3.2. α-Diversity patterns

The rarefaction analysis of the different fraction sizes of the sampled communities resulted in different patterns for fraction A (> 1mm) and fraction B (63µ – 1mm). MOTU Richness Plots showing the patterns of α-diversity for CROP and SWARM can be seen in Figure 6 and 7, being represented in Figure 6 before doing MINT and in Figure 7 after doing MINT.

In Figure 6.1. for the fraction A, after performing the Dunn Test ($p < 0.05$), we obtained significant differences between V1 and V4. And the same for fraction B, significant differences between V1 and V4. In Figure 6.2. for the fraction A we didn't obtain any significant difference after doing the Dunn Test but in fraction B we could appreciate significant differences between V1 – V2 and V1 – V4.

In Figure 7 we obtained the same results as in Figure 5 after performing the Dunn Test. In Figure 7.1 we obtained significant differences between V1 and V4 for both, fraction A and fraction B. In Figure 7.2., for the fraction A, there weren't any significant difference after performing the Dunn Test but in fraction B we could see significant differences between V1 – V2 and V1 – V4.

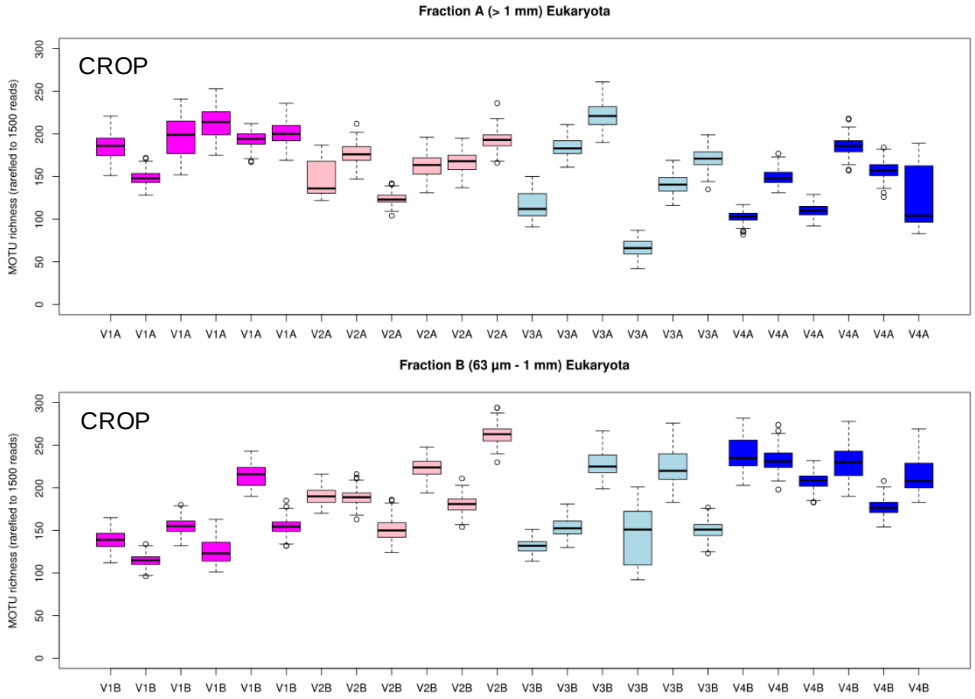

Fig 6.1. MOTU Richness Plots showing patterns of α-diversity before doing MINT of the fractions A (> 1mm) and fractions B (63µm – 1mm) from the 4 sites Vent, Transition, Control and Control Echentive. CROP clustering. Results obtained by rarefaction analysis to 1,500 reads per sample.

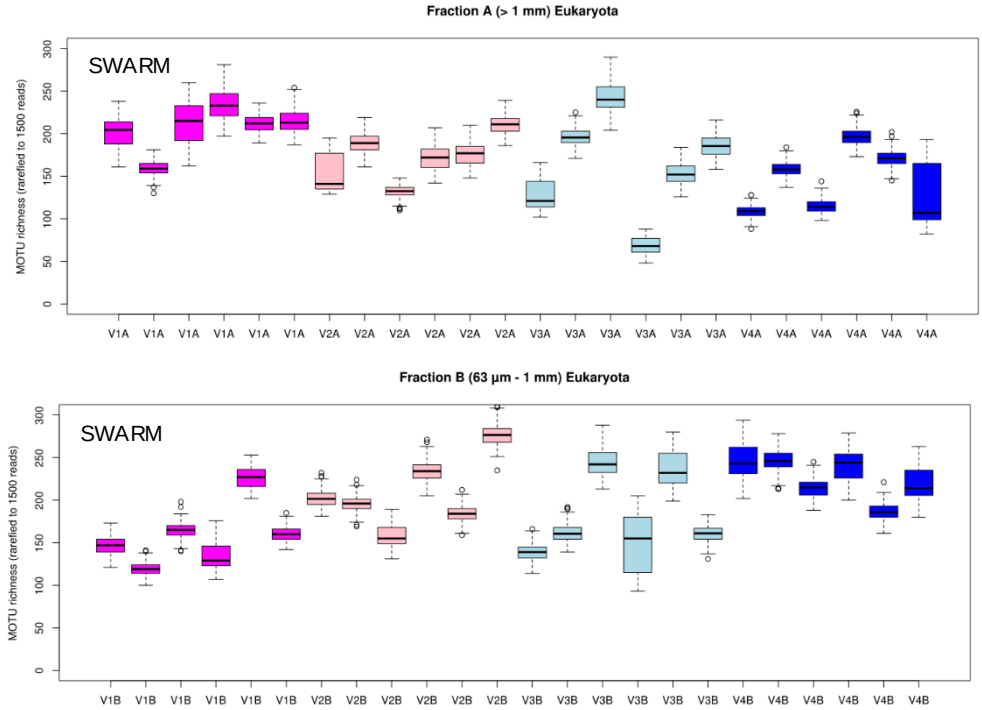

Fig 6.2. MOTU Richness Plots showing patterns of α-diversity before doing MINT of the fractions A (> 1mm) and fractions B (63µm – 1mm) from the 4 sites Vent, Transition, Control and Control Echentive. SWARM clustering. Results obtained by rarefaction analysis to 1,500 reads per sample.

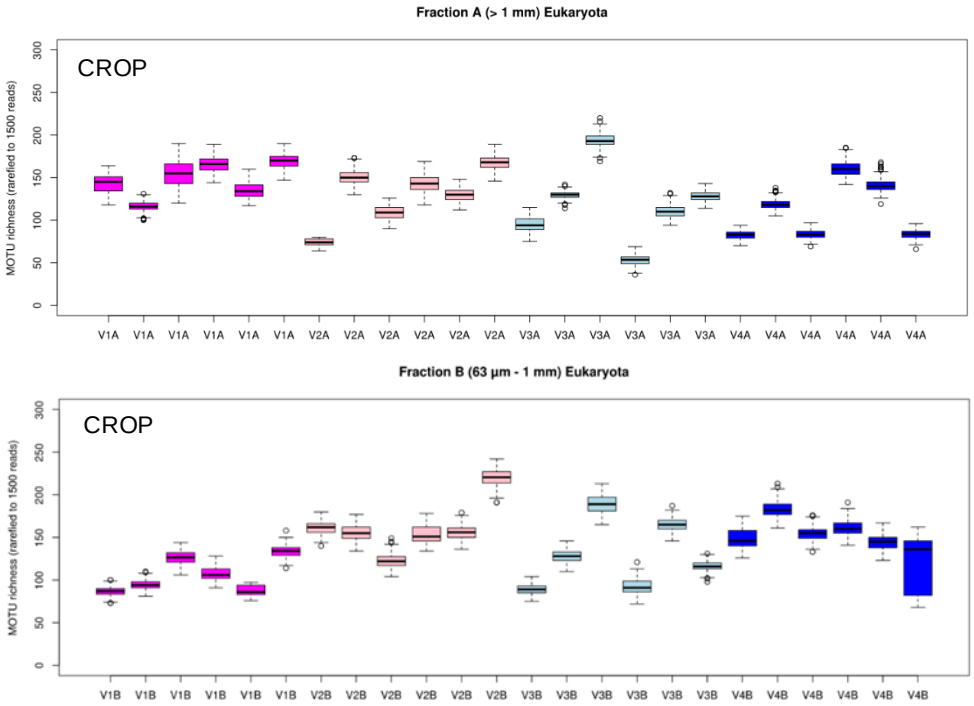

Fig 7.1. MOTU Richness Plots showing patterns of α-diversity after doing MINT of the fractions A (>1mm) and fractions B (63µm – 1mm) from the 4 sites Vent, Transition, Control and Control Echentive. CROP clustering. Results obtained by rarefaction analysis to 1,500 reads per sample.

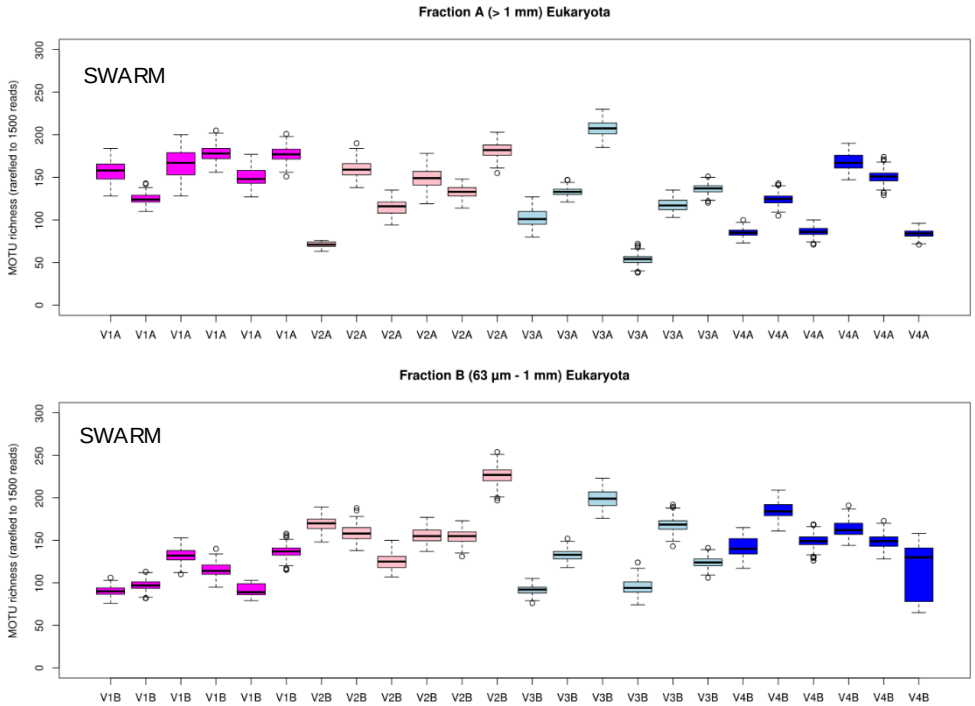

Fig 7.2. MOTU Richness Plots showing patterns of α-diversity after doing MINT of the fractions A (>1mm) and fractions B (63µm – 1mm) from the 4 sites Vent, Transition, Control and Control Echentive. Results obtained by rarefaction analysis to 1,500 reads per sample.

3.3. Ordination patterns after CROP or SWARM with MINT

Figure 8 shows the multidimensional scaling (MDS) of the results from the clustering algorithm (CROP or SWARM) after using MINT correction. The representation of the MDS shows a similar pattern in the four situations with small variations. There is barely any remarkable difference between CROP and SWARM. After using MINT we can observe that there is no remarkable differences in the obtained ordination patterns between using MOTUs appearing in 2 or 3 replicates (MINT-two) or MOTUs appearing in all replicates (MINT-all). The total number of remaining MOTUs is the only important variable which varies in each situation.

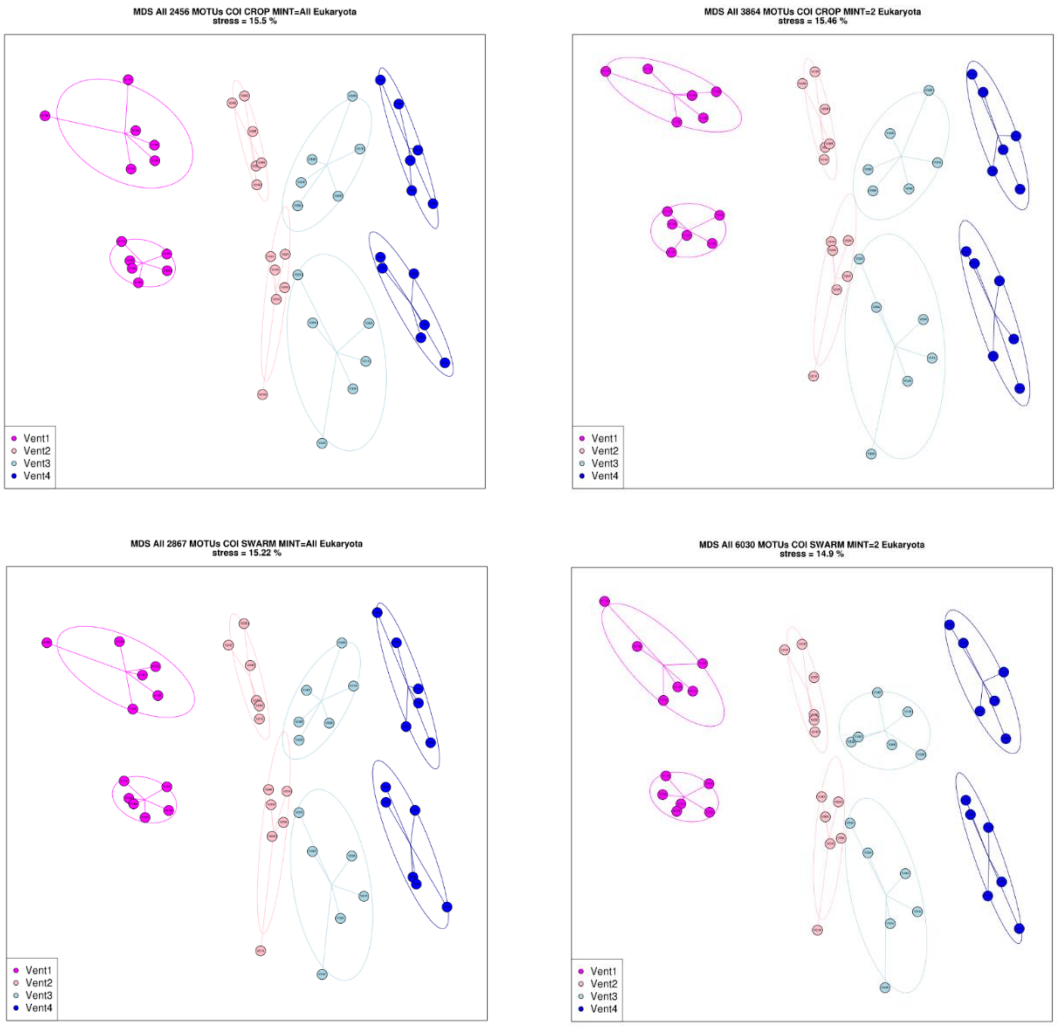

Figure 8. nMDS plots showing the robustness of ordination patterns using CROP or SWARM with both types of MINT correction (MINT-all and MINT-two).

3.4. Comparing DNA extraction methods

In Figure 9 we can observe the patterns of abundance of reads per sample of 6 samples (3 samples from V1 and 3 samples from V4) comparing different extraction methods which are based in two different sample weights. PowerMax® Soil DNA Isolation Kit allowed to purify DNA from 10 g of sample and PowerSoil® DNA Isolation Kit purified DNA from 0.3 g, both from the same company *MoBio Laboratories, Inc (now QIAGEN).*

3 PCR replicates of each sample were done for the samples treated with PowerSoil® DNA Isolation Kit and 2 PCR replicates were done for PowerMax® Soil DNA Isolation Kit.

We can observe similar MOTUs and read abundances in most of the replicates. In the case of V41 and V42 we found differences between both kits for fractions A. In the replicates of V41 using PowerSoil® DNA Isolation Kit we can appreciate more reads of Rhodophyceae and using PowerMax® Soil DNA Isolation Kit there are less reads of Rhodophyceae but more of Annelida, Arthropoda, Cnidaria and Mollusca. In the replicates of V42 using PowerSoil® DNA Isolation Kit we can appreciate more reads of Echinodermata and using PowerMax® Soil DNA Isolation Kit there are less reads of Echinodermata but more of unassigned reads, Annelida, Arthropoda, Cnidaria and Mollusca.

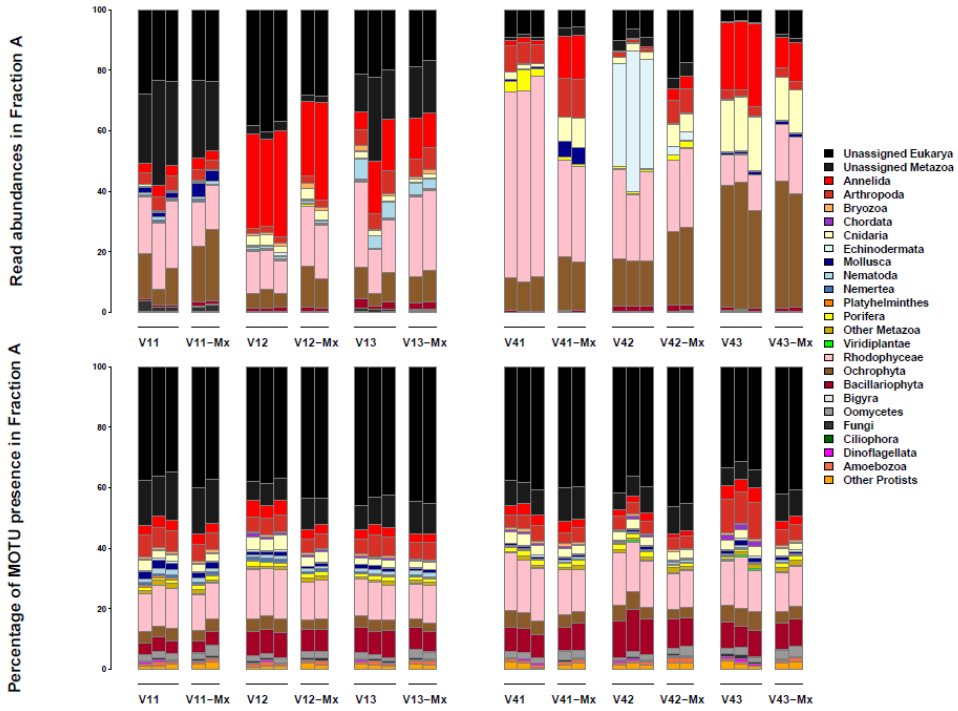

Figure 9. Patterns of abundance of metabarcoding reads per sample obtained using two DNA extraction kits, PowerSoil® DNA Isolation Kit and PowerMax® Soil DNA Isolation Kit.

Chapter 4.

Discussion

The field of eukaryotic biodiversity assessment from marine benthic communities has benefited from a growth in the number of tools available to analyze the growing number of COI DNA gene sequences. As we have shown in this study, both OTU- and replicate-based methods have unique challenges that affect one's ability to implement the method and interpret the results. The results presented in this study enable researchers to better interpret and overcome these challenges. There are a several extensions of this research that deserve further consideration.

4.1. Sample pre-treatment, the benefits of sieving and size fractionation

Following the protocol done in Wangensteen and Turon (2015) we can see advantages in the choice of size fractionation and filtering through a column of sieves. The partitioned metabarcoding of size fractions allows characterization of structurally complex communities at different levels, which could be impossible if each sample was homogenized altogether, due to the high number of DNA copies from organisms of bigger biomass outnumbering the smaller ones (Wangensteen and Turon, 2015). An additional advantage of this process is the removal of a significant fraction of microorganisms (prokaryotes and smallest microeukaryotes), jointly with probably most of the extra-organismal DNA in the form of small remains, cell debris, or extracellular DNA (Creer *et al.*, 2016), which are not retained in the last sieve (63 µm). These microorganisms are known to be genetically diverse and under-represented in genetic databases. Their presence introduces some additional problems during bioinformatic analyses, specifically for clustering and taxonomical assignation algorithms, most notably when using COI as marker, and they are better removed from the samples whenever they are not the main study target (Wangensteen and Turon, 2015).

4.2. Choosing the correct DNA extraction kit

DNA extraction is a fundamental step in metabarcoding and a number of commercial kits are available in the market. Selecting the correct kit can save crucial time on kit optimization and extraction repeatability. Factors to be considered for selecting a kit include:

- Sample origin and humic content: humic substances need to be removed with a proper kit, as they can inhibit downstream applications like PCR.
- Preparation method: depending if the samples are fresh or previously frozen.
- Intended use: considering the quality and purity of the DNA you want to obtain.
- Sample quantity: depends on how much sample is available and how many replicates we want to analyze.
- Price.

PowerMax® Soil DNA Isolation Kit and PowerSoil® DNA Isolation Kit are based in the same protocol but they work with different amounts of sample. The first one purifies DNA from 10 g of sample and the second one 0.3 g. There are also differences in the price, PowerMax® Soil DNA Isolation Kit costs $24.6 per sample and PowerSoil® DNA Isolation Kit costs $5.14 per sample.

PowerMax® Soil DNA Isolation Kit has the advantage that we can use more quantity of sample but at a higher cost. 48 samples were analyzed in our study, so this means buying 5 PowerMax® Soil DNA Isolation Kits (a total of $1,230), which would involve high spending just for DNA extraction.

As we are working with 48 samples, the PowerSoil® DNA Isolation Kit for 100 samples would be a good option for our DNA extraction. But then we come to the question: are 0.3 g sufficient to analyze the biodiversity of the community? We start from an original sample of 150 g of where we only take 0.3 g. As we can see in Figure 8 there are barely differences between kits. In order to answer our question, we can affirm that 0.3 g are enough to obtain repeatable results for presence / absence of MOTUs and even for the relative abundance of reads per phylum. The comparison between the kits is perfect for fractions B, whereas in the case of fractions A, only two samples showed different relative abundances of reads (due to increased abundance of one rhodophycean or one echinoderm MOTU, respectively). This result suggests that the repeatability could be improved if the samples had been more thoroughly homogenized before the DNA extraction. However, the main conclusion is that the PowerSoil® DNA Isolation Kit allows to obtain representative enough DNA samples at a lower cost.

4.2. Importance of the choice of eDNA metabarcoding markers

The importance of marker choice in eDNA metabarcoding has recently been emphasized (Wangensteen *et al.*, 2017). Because there is no ideal universal metabarcode, marker choice should be specific to the target taxonomic group, and validation is required before application of the metabarcoding analysis in situ. In this case, we are studying a whole community so we need a universal primer capable of amplifying the wide array of taxonomic groups. We have used the Leray fragment of the mitochondrial cytochrome *c* oxidase I gene (COI) as a metabarcoding marker although it has been criticized arguing that it does not contain suitably conserved regions for short amplicon-based eDNA applications (Deagle *et al.*, 2014). COI presents two major advantages compared to other possible markers. First, the steadily growing international effort, led by the Consortium for the Barcode of Life (CBOL), to develop a public DNA barcoding database with curated taxonomy enormously facilitates taxonomic assignment. The Barcode of Life Data System (BOLD) database is based mainly in COI barcoding and currently includes over 4 million sequences belonging to more than 500,000 different species, curated and identified by expert taxonomists. These data have been gathered by thousands of researchers working worldwide across decades and it is highly unlikely that any comparable effort might be undertaken for any other marker in the next future. Metabarcoding studies may take full advantage of this invaluable resource only if they choose COI as a marker. Secondly, the high mutation rate of COI practically ensures the unequivocal identification at the species level, whereas the highly conserved sequence of other markers is often impossible to distinguish at the genus or family levels, or even at higher ranks (Wangensteen and Turon, 2015). Species-level identification is important because we can see which species have been affected by the acidification of the sea water with the increase of CO_2 dissolved in water.

4.3. Quantification of abundance

A controversial issue in metabarcoding is whether this technique has a quantitative value. There is some evidence of a positive relationship between species abundance and DNA sequence counts for several species (Shaw *et al.*, 2016; Klymus *et al.*, 2017). Many factors can affect sequence count (Deagle *et al.*, 2013), including DNA extraction method, sample storage, PCR bias, preferential amplification, spatial biases, cell integrity, gene copy

number, and sequencing error. The abundance of species is not an important factor to be taken into account when studying the species richness of the community, since this depends just on the presence/absence of species. Our results for the relative abundance of reads belonging to each phylum showed patterns that broadly matched the expected abundances of the different phyla in the studied ecosystems. Thus, apparently, using highly-degenerated universal primers, such as the Leray-XT primer set, on community DNA from thoroughly homogenized samples, yields metabarcoding data with at least some quantitative value. However, more research is needed before species abundance can be reliably estimated from eDNA metabarcoding surveys.

4.4. Comparing clustering algorithm

High-throughput sequencing technologies can generate millions of amplicons (or barcode sequences) in a single run, and these sequences need to be clustered into molecular operational taxonomic units (MOTUs) before being used for diversity estimates or other statistical analyses. Sequences are clustered into MOTUs based on a similarity threshold that mirrors natural intraspecific divergence. This approach generates “species equivalents” that can be used to approximate species numbers in metabarcoding studies irrespective of DNA reference database coverage (Clare *et al.*, 2016).

A number of analytical programs are used to define MOTU and most rely on some sort of clustering or threshold approach. Because of the increasing sizes of today’s amplicon datasets, fast and greedy *de novo* clustering heuristics are usually the preferred practical approach to produce MOTUs (Ghodsi *et al.*, 2011). Shared steps in these current algorithms are: an amplicon is drawn out of the amplicon pool and becomes the center of a new OTU (centroid selection), this centroid is then compared to all other amplicons remaining in the pool. Amplicons for which the distance is within a global clustering threshold, *t*, to the centroid are moved from the pool to the OTU. The OTU is then closed. These steps are repeated as long as amplicons remain in the pool (Mahé *et al.*, 2014) (Figure 10).

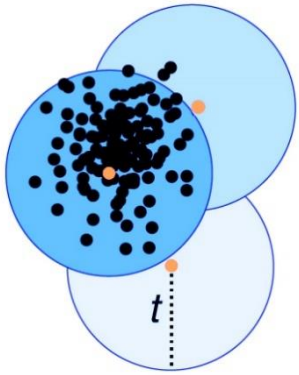

Figure 10. Schematic view of the clustering approach based on centroid selection and a global clustering threshold, *t*, where closely related amplicons can be placed into different OTUs.

In defining OTUs, some studies showed that it is difficult to use a consistent threshold because there is considerable overlap in the maximum intra-taxon distance between taxonomic levels (Schloss and Westcott, 2011). It may be useful when targeting a closely related group of organisms, but it is not the best strategy when dealing with complex, taxonomically-wide samples. Setting a constant arbitrary similarity threshold for defining MOTUs across the whole dataset will necessarily miss some real diversity in those groups having experienced recent evolutionary radiations, whereas it will overestimate the number of species of slowly evolving groups with high intra-specific genetic variability. Algorithms based in a shifting similarity threshold for defining the MOTUs are then likely more appropriate for describing the real diversity in these widely targeted studies (Wangensteen and Turon, 2015).

To avoid using a hard threshold value in clustering as implemented in hierarchical and heuristic methods, Hao *et al.* (2011) proposed a Gaussian mixture model-based clustering algorithm termed Clustering 16S rRNA for OTU Prediction (CROP). It adopts an unsupervised probabilistic Bayesian clustering algorithm and uses a soft threshold for defining OTUs. The CROP algorithm bypasses setting an often-subjective hard cut-off threshold thus may effectively reduce the effects of PCR and sequencing errors in inferring OTUs.

In addition, amplicons from one species can be subsumed into the OTU of a genetically closely related species with a very dissimilar ecology if that second species had a higher abundance, leading to erroneous ecological interpretations. To solve these issues, (Mahé *et al.*, 2014) developed SWARM — a novel method that avoids both fixed global clustering thresholds, and input-order dependency due to centroid selection. They

implemented an exact, yet fast, de novo clustering method that produces meaningful OTUs and reduces the influence of clustering parameters.

In our study, we compare both clustering algorithm (Fig. 11). CROP was set with the following parameters: l=1.5, u=2.5 (producing an initial threshold of similarity of around 95%). And SWARM was set with a distance of d=13 mismatches, from a total fragment length of 313 bp of the Leray fragment.

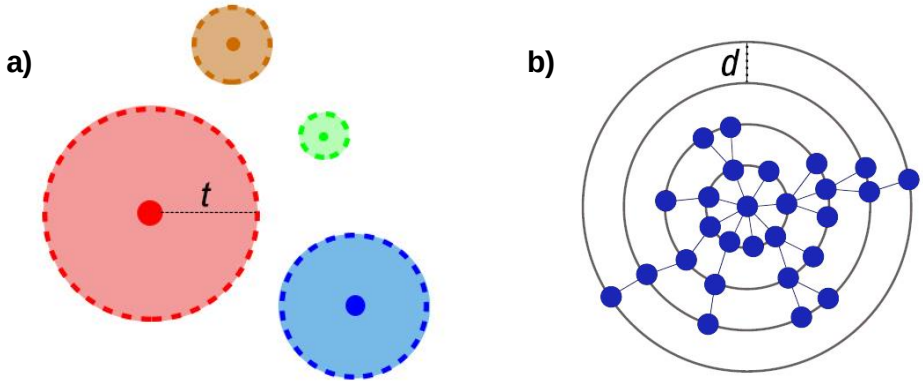

Figure 11. Schematic view of both clustering algorithm, CROP (a) and SWARM (b).

Figure 8 shows the multidimensional scaling (MDS) of both clustering algorithm where similar ordination patterns were recovered from CROP and SWARM. Samples from each site are grouped together with a certain overlap of the ellipses between V2 and V3 after using the MINT correction. However, the ordination patterns obtained from both clustering methods are almost exactly the same, despite the completely different principles underlying both methods (one being a heuristic Bayesian method, and the other a deterministic step-by-step aggregation algorithm). Comparing the 4 sites, V2, V3 and V4 are closer to each other, while V1 (the CO_2-vent site) is further away, demonstrating that V1 has a higher difference in the composition of the MOTUs, probably due to the differences of acidity in the water.

In terms of differences between fractions, there is a large difference between fraction A and fraction B which can be expected since we are separating populations according to size (63 µm - 1 mm).

In conclusion, both clustering methods yielded similar results regarding the ecological differences among the communities present in all sites, but they yielded different number of MOTUs. With CROP, we obtained 6,372 MOTUs and with SWARM we obtained 9,475

MOTUs. The higher number of MOTUs is mainly due to low-abundance MOTUs, typically including less than 20 reads, whereas the most abundant MOTUs including hundreds or thousands of reads are, in general, equally retrieved by both clustering methods.

The amount of time needed for the calculations is a critical factor when we choose an appropriate method for analyzing large scale datasets. With our initial dataset, clustering with CROP took 51 hours to reach a valid output solution, whereas clustering with SWARM took only 5 minutes to reach to the deterministic solution.

Other disadvantages of CROP are that, although using a soft threshold for defining OTUs, lineages evolve at variable rates so no single cut-off value can accommodate the entire tree of life. A single global clustering threshold will inevitably be too relaxed for slow-evolving lineages and too stringent for rapidly evolving one (Koeppel and Wu, 2013). Also, the input order of amplicons may influence the clustering results. Previous centroid selections are not re-evaluated as clustering progresses, which can generate inaccurately formed OTUs, where closely related amplicons can be separated and unrelated amplicons can be grouped (Westcott and Schloss, 2015).

For all these reasons, we conclude that clustering with SWARM is the best solution for treating COI metabarcoding datasets obtained from complex samples.

4.5. Estimates of α-diversity

Alpha diversity refers to the diversity within a particular area or ecosystem, and is usually expressed by the number of species (i.e., species richness) in that ecosystem. One of the aims of metabarcoding is to objectively determine which species are present in a given environmental sample. However, it is presently impossible to establish a one-to-one correspondence between morphological species and MOTUs obtained by any metabarcoding marker (Wangensteen and Turon, 2015). Neither read abundance can be used to estimate the number of individuals per OTU. Therefore, rarefaction must be used to normalize the number of reads per sample before comparing values for alpha diversity (Yu *et al.*, 2012). In our dataset, MOTU richness was rarefied to 1,500 reads per sample, where we obtained values of 50-300 MOTUs per sample.

Our data showed differences between the diversity from the vent site (V1) and the Echentive control site (V4), the one that is farthest away from the vent site. In addition,

fractions A and B follow an opposite pattern, that is, alpha diversity is higher in fraction A from V1 but lower in fraction B. Thus, as we move from V4 (the control area) towards V1 (area where the CO_2 concentration is higher), a decreasing biodiversity gradient is detected in organisms of smaller body size (63 µm to 1 mm), while the detected molecular biodiversity of organisms of bigger sizes (> 1 mm) seems to increase. Among the reported effects of a higher concentration of carbon dioxide from other studies, there is a clear reduction in diversity, biomass and trophic complexity of benthic marine assemblages, major declines in the number of many calcifying organisms and increased abundances of erect macroalgae, seagrass and soft-corals (Hall-Spencer *et al.*, 2008) (Inoue *et al.*, 2013) (Linares *et al.*, 2015; Enochs *et al.*, 2015). Other authors have also demonstrated, using natural vents, the capacity of adaptation of calcifying organisms (Johnson *et al.*, 2012) and the important role that non-calcifying ones would play in future conditions (Russell *et al.*, 2013). However, more vent systems are necessary to represent more oceans and different environments, in order to assess how biodiversity will adapt to future ocean acidification.

4.6. Contamination

The most serious pitfall of eDNA is probably the risk of contamination and hence the possibility of false positive results. Contamination of samples can occur from taking the samples in the field to every step of analyses in the laboratory. If several localities are sampled after one another in the field, there is a risk of cross-contamination: target DNA carried unintentionally from one locality to another. Laboratory contamination is especially serious because of the frequent use of PCR in eDNA studies, generating billions of DNA copies, which can readily spread throughout the laboratory. The use of HTS technologies has further complicated the contamination issues, as they produce a very high throughput of DNA sequences likely to reveal tiny amounts of lab-source PCR products (Thomsen and Willerslev, 2015). Cross-contamination in the laboratory seems almost unavoidable, and it is essential to apply conservative cut-offs for minimum percentages of sequences obtained in a sample, and/or the amplification success in independent PCR reactions, before including a recovered taxon as authentic. A strict clean-lab protocol using decontamination procedures and physical separation of laboratories for pre- and post-PCR work will significantly limit the contamination risks (Champlot *et al.*, 2010). Inclusion of DNA extraction blanks and PCR blanks, as well as field blanks, to monitor contamination is essential.

4.7. Dealing with false positives

One of the main challenges associated with metabarcoding is the risk of erroneous DNA sequences. Errors can occur either during sampling, during PCR or during sequencing. PCR-generated errors include point mutations and formation of chimeric molecules (Lenz and Becker, 2008). However, most errors are probably generated during sequencing. In order to eliminate these errors, we performed 3 PCR replicates from each sample.

PCR replicates allow researchers to optimize diversity detection by counteracting effects of PCR stochasticity (Leray and Knowlton, 2015), which might be especially high in low template complex DNA extracts (Murray *et al.*, 2015). Analytically, if PCR replicates are made, they are often used additively, that is, through pooling the sequences of a single sample's PCR replicates to maximize diversity detection (e.g. (Burgar *et al.*, 2014; Leray and Knowlton, 2015)). While this certainly reduces the risk of missing taxa, false positives are a potential consequence due to accumulation of artefactual sequences. To counteract this, a few studies have used PCR replicates in a restrictive context by only retaining sequences that are shared by a number of a sample's PCR replicates (Hope *et al.*, 2014). However, the most widely used strategy for eliminating erroneous sequences is to set a minimum sequence copy number below which sequences are discarded. While many authors opt for removing singletons (e.g. (Burgar *et al.*, 2014), others set higher copy number thresholds (e.g. (Giguet-Covex *et al.*, 2014). In the MINT strategy we used here, we kept singletons in order to not reduce the detected diversity, since they will be removed after using this script if they only appear in one replicate. One specific group of PCR artefacts are chimeric sequences, and if they are not detected and removed they might erroneously inflate the richness and diversity measurements of the results (Bjørnsgaard Aas *et al.*, 2017). In our study, we eliminated the chimeras in the pipeline as it is a recommended step before the clustering procedures, although it is unlikely to change the results considerably if other measures to counteract the effect of PCR and sequencing errors are taken.

The number of total MOTUs detected from all samples using CROP clustering before MINT was 6,372. After refining with CROP-MINT-all a total of 2,456 MOTUs remained. Using CROP-MINT-two refining procedure, a total of 3,864 MOTUs remained. Using the SWARM pipeline, the total number of MOTUs before MINT was 9,475. After SWARM-MINT-

all we obtained a total of 2,867. Using SWARM-MINT-two we obtained 6,030 MOTUs. Although the number of detected MOTUs was decreased, the MOTUs removed by our stringent cleaning procedures (MINT) likely corresponded to artefactual MOTUs originated by different kinds of random errors. These removed MOTUs never showed high abundance values, and the resulting cleaned datasets are robust and more representative of the real diversity present in the sampled communities.

4.8. Taxonomic assignment

Many approaches can be used to assign taxonomy to OTUs detected in a metabarcoding study. Some authors argue that for COI markers, identifications below 98% identity might be error prone (Clare *et al.*, 2011), thus taxonomic assignments should be limited to matches above that threshold. However, such fixed approaches, might result in a low taxonomic assignment success, more so if the target taxa are not well characterized in reference databases, as is commonly the case for small-sized marine benthic communities. While species-level identification is essential in some studies, in most cases, incorporating higher taxonomic assignments in addition to species-level identifications allows increasing the ecological inference of the study. In those cases, multi-level taxonomic assignments that assign higher taxonomic levels to lower identities might be useful (Alberdi *et al.*, 2017).

Multiple algorithms and approaches exist to perform taxonomic assignments but for our analysis with reference sequences from public databases we used sequences obtained from two sources: in silico ecoPCR against the release 117 of the EMBL nucleotide database and a second set of sequences for our metabarcoding fragment obtained from the Barcode of Life Data systems (BOLD) (Ratnasingham and Hebert, 2007) and a custom R script. This newly generated database included 190,101 reference sequences (August 2017) from a wide taxonomic range.

The rates of unassigned sequences in our results suggest that important gaps still exist for the COI marker in the genetic repositories, which would prevent the detailed identification of many (probably most) marine organisms to a level below order or family, in agreement with the concerns expressed by other authors (Leray and Knowlton, 2015).

Chapter 5.

Conclusion and future remarks

Vent systems are not perfect predictors of future ocean ecology owing to temporal variability in pH, spatial proximity of populations unaffected by acidification and the unknown effects of other global changes in parameters such as temperature, currents and sea level. However, such vents acidify sea water on sufficiently large spatial and temporal scales to integrate ecosystem processes such as production, competition and predation (Hall-Spencer *et al.*, 2008). Some experimental and modelling predictions confirm that differential responses of benthic species to decreased pH can lead to substantial changes in community structure (Raven *et al.*, 2005; Riebesell *et al.*, 2007; Feely *et al.*, 2004; Orr *et al.*, 2005; Hoegh-Guldberg *et al.*, 2007; Davies *et al.*, 2007; Fine and Tchernov, 2007; Kuffner *et al.*, 2008). It is unknown whether these groups of species will adapt to survive the rapid rate of ocean acidification predicted to occur due to anthropogenic CO_2.

Although a number of studies exist that have studied marine biodiversity in communities affected by ocean acidification, to the best of our knowledge, no previous study has been performed in the Atlantic Ocean. So, this CO_2 shallow vent discovered in the Canary Islands can be used as a proxy for ocean acidification studies on Atlantic communities. This opportunity to observe the tipping points at which principal groups of marine organisms are affected by lowered pH proves that, even without considering the global warming, the projected rise in atmospheric CO_2 concentration is hazardous, as ocean acidification will probably bring about reductions in biodiversity and radically alter ecosystems (Hall-Spencer *et al.*, 2008).

Over the past 10 years, advances in sequencing technology and accompanying methodological breakthroughs have revolutionized our ability to study community biodiversity. Using this approach, thousands of species present in any environmental sample can be detected by high-throughput DNA sequencing and identified using public databases. This next generation sequencing can provide us with a lot of information so there is a great need to develop effective methods to analyze these data.

Yet, during the next few years, the improvement of the current PCR-based metabarcoding methods is still the most promising development. There is still much room for improvement in the design of universal primers (by incorporating deoxyinosine nucleotides in crucial positions), as well as in the design of specific primers for selected groups of organisms of special interest, which would allow the selective amplification of just the target group of organisms, with the resulting increase in the sequencing depth of target sequences. No doubt new metabarcoding markers will be discovered, which will lead to enhanced quantitative and qualitative biodiversity assessment.

Another crucial point where improvement is expected in the near future is the increasing breadth and coverage of existing reference databases. Although for particular studies it is possible to generate the database of interest (e.g. by sequencing all species of a given group known to occur in very local studies), for most studies concerned with assessment of general biodiversity, the correct assignment of MOTUs is directly dependent on dense and correctly curated databases. Marine barcoding databases are currently much less populated than their terrestrial counterparts, and significant portions of the tree of life are still under-represented or missing altogether. We urge marine biodiversity researchers to contribute to the growth of denser and more useful reference databases by sequencing at least the most prominent taxa found in marine metabarcoding studies, thus obtaining a validation of the assignments and contributing to improve the databases. Of course, ecological inferences are possible even in the absence of a precise taxonomic assignment, but the lower the number of "unassigned" MOTUs (or assigned just at high taxonomic levels), the better the knowledge we will obtain of the structural properties and the functioning of the ecosystems.

We demonstrate that the diversity detected in metabarcoding studies can drastically change according to the laboratory set-up and the different parameters and thresholds employed during the bioinformatic workflow. While it is likely that none of the approaches employed perfectly reflects reality, it is clear that certain choices critically increase the reliability of the results. Thus, we encourage researchers to acknowledge the benefits as well as potential biases and limitations of the different set-ups when designing a metabarcoding study, and adjust the interpretation of the data to the level of uncertainty. We also highlight the importance of adjusting parameters to the marker region and taxonomic range based on empirical data rather than relying on general rules of thumb or standard settings of available software.

Finally, if metabarcoding is to reliably answer biological questions, we believe that continuous technical refinement and illumination of strengths, weaknesses, and biases are needed in order to ensure unbiased, standardized and optimal detection of true diversity.

Chapter 6.

References

Abbasi, T. and Abbasi, S. A. (2011) 'Ocean Acidification: The Newest Threat to the Global Environment', *Critical Reviews in Environmental Science and Technology,* 41(18), pp. 1601-1663.

Alberdi, A., Aizpurua, O., Gilbert, M. T. P. and Bohmann, K. (2017) 'Scrutinizing key steps for reliable metabarcoding of environmental samples', *Methods in Ecology and Evolution*, pp. 1-14.

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data.

Baird, D. J. and Hajibabaei, M. (2012) 'Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing', *Molecular Ecology,* 21(8), pp. 2039-2044.

Baldwin, C. C., Collette, B. B., Parenti, L. R., Smith, D. G. and Springer, V. G. (1996) 'Collecting fishes', in Lang, M.A. & Baldwin, C.C. (eds.) *Methods and Techniques of Underwater Research*: American Academy of Underwater Sciences, pp. 11-33.

Barry, J., Hall-Spencer, J. and Tyrrell, T. (2010) 'In situ perturbation experiments: natural venting sites, spatial/temporal gradients in ocean pH, manipulative in situ pCO_2 perturbations.', in Riebesell, U., Fabry, V., Hansson, L. & Gattuso, J.-P. (eds.) *Guide to best practices for ocean acidification research and data reporting.* Luxembourg: Publications Office of the European Union, pp. 123-126.

Bjørnsgaard Aas, A., Davey, M. L. and Kauserud, H. (2017) 'ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities', *Molecular Ecology Resources,* 17(4), pp. 730-741.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. and Coissac, E. (2016) 'obitools: a unix-inspired software package for DNA metabarcoding', *Molecular Ecology Resources,* 16(1), pp. 176-182.

Burgar, J. M., Murray, D. C., Craig, M. D., Haile, J., Houston, J., Stokes, V. and Bunce, M. (2014) 'Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed', *Molecular Ecology,* 23(15), pp. 3605-3617.

Bálint, M., Schmidt, P.-A., Sharma, R., Thines, M. and Schmitt, I. (2014) 'An Illumina metabarcoding pipeline for fungi', *Ecology and Evolution,* 4(13), pp. 2642-2653.

Caldeira, K. and Wickett, M. E. (2003) 'Oceanography: Anthropogenic carbon and ocean pH', *Nature,* 425(6956), pp. 365-365.

Champlot, S., Berthelot, C., Pruvost, M., Bennett, E. A., Grange, T. and Geigl, E.-M. (2010) 'An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications', *PLOS ONE,* 5(9), pp. e13042.

Chariton, A. A., Court, L. N., Hartley, D. M., Colloff, M. J. and Hardy, C. M. (2010) 'Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA', *Frontiers in Ecology and the Environment,* 8(5), pp. 233-238.

Clare, E. L., Barber, B. R., Sweeney, B. W., Hebert, P. D. N. and Fenton, M. B. (2011) 'Eating local: influences of habitat on the diet of little brown bats (Myotis lucifugus)', *Molecular Ecology,* 20(8), pp. 1772-1780.

Clare, E. L., Chain, F. J. J., Littlefair, J. E. and Cristescu, M. E. (2016) 'The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data', *Genome,* 59(11), pp. 981-990.

Cornwall, C. E. and Hurd, C. L. (2016) 'Experimental design in ocean acidification research: problems and solutions', *ICES Journal of Marine Science,* 73(3), pp. 572-581.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C. and Bik, H. M. (2016) 'The ecologist's field guide to sequence-based identification of biodiversity', *Methods in Ecology and Evolution,* 7(9), pp. 1008-1018.

Cressey, D. (2015) 'Seawater studies come up short', *Nature,* 524(7563), pp. 18-19.

Davies, A. J., Roberts, J. M. and Hall-Spencer, J. (2007) 'Preserving deep-sea natural heritage: Emerging issues in offshore conservation and management', *Biological Conservation,* 138(3-4), pp. 299-312.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P. and Karsenti, E. (2015) 'Eukaryotic plankton diversity in the sunlit ocean', *Science,* 348(6237).

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F. and Taberlet, P. (2014) 'DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match', *Biology Letters,* 10(9).

Deagle, B. E., Thomas, A. C., Shaffer, A. K., Trites, A. W. and Jarman, S. N. (2013) 'Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count?', *Molecular Ecology Resources,* 13(4), pp. 620-633.

Enochs, I. C., Manzello, D. P., Donham, E. M., Kolodziej, G., Okano, R., Johnston, L., Young, C., Iguel, J., Edwards, C. B., Fox, M. D., Valentino, L., Johnson, S., Benavente, D., Clark, S. J., Carlton, R., Burton, T., Eynaud, Y. and Price, N. N. (2015) 'Shift from coral to macroalgae dominance on a volcanically acidified reef', *Nature Clim. Change,* 5(12), pp. 1083-1088.

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., ErsÉUs, C., Gusarov, V. I., Edwards, M. E., Johnsen, A., StenØlen, H. K., Hassel, K., Kauserud, H., Yoccoz, N. G., BrÅThen, K. A., Willerslev, E., Taberlet, P., Coissac, E. and Brochmann, C. (2012) 'New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems', *Molecular Ecology,* 21(8), pp. 1821-1833.

Feely, R. A., Doney, S. C. and Cooley, S. R. (2009) 'Ocean Acidification: Present Conditions and Future Changes in a High-CO_2 World', *Oceanography,* 22(4), pp. 36-47.

Feely, R. A., Sabine, C. L., Lee, K., Berelson, W., Kleypas, J., Fabry, V. J. and Mille ro, F. J. (2004) 'Impact of anthropogenic CO_2 on the $CaCO_3$ system in the oceans', *Science,* 305(5682), pp. 362-366.

Fine, M. and Tchernov, D. (2007) 'Scleractinian coral species survive and recover from decalcification', *Science,* 315(5820), pp. 1811-1811.

Fonseca, V. G., Carvalho, G. R., Nichols, B., Quince, C., Johnson, H. F., Neill, S. P., Lambshead, J. D., Thomas, W. K., Power, D. M. and Creer, S. (2014) 'Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes', *Global Ecology and Biogeography,* 23(11), pp. 1293-1302.

Friedrich, T., Timmermann, A., Abe-Ouchi, A., Bates, N. R., Chikamoto, M. O., Church, M. J., Dore, J. E., Gledhill, D. K., Gonzalez-Davila, M., Heinemann, M., Ilyina, T., Jungclaus, J. H., McLeod, E., Mouchet, A. and Santana-Casiano, J. M. (2012) 'Detecting regional anthropogenic trends in ocean acidification against natural variability', *Nature Clim. Change,* 2(3), pp. 167-171.

Gattuso, J.-P. and Buddemeier, R. W. (2000) 'Ocean biogeochemistry: Calcification and CO_2', *Nature,* 407(6802), pp. 311-313.

Gazeau, F., Quiblier, C., Jansen, J. M., Gattuso, J.-P., Middelburg, J. J. and Heip, C. H. R. (2007) 'Impact of elevated CO_2 on shellfish calcification', *Geophysical Research Letters,* 34(7), pp. 1-5.

Geller, J., Meyer, C., Parker, M. and Hawk, H. (2013) 'Redesign of PCR primers for mitochondrial cytochrome *c* oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys', *Molecular Ecology Resources,* 13(5), pp. 851-861.

Ghodsi, M., Liu, B. and Pop, M. (2011) 'DNACLUST: accurate and efficient clustering of phylogenetic marker genes', *BMC Bioinformatics,* 12(1), pp. 271.

Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.-J., Griggo, C., Gielly, L., Domaizon, I., Coissac, E., David, F., Choler, P., Poulenard, J. and Taberlet, P. (2014) 'Long livestock farming history and human landscape shaping revealed by lake sediment DNA', 5, pp. 3211.

Guardiola, M., Uriz, M. J., Taberlet, P., Coissac, E., Wangensteen, O. S. and Turon, X. (2015) 'Deep-Sea, Deep-Sequencing: Metabarcoding Extracellular DNA from Sediments of Marine Canyons', *PLOS ONE,* 10(10), pp. e0139633.

Guinotte, J. M. and Fabry, V. J. (2008) 'Ocean Acidification and Its Potential Effects on Marine Ecosystems', *Annals of the New York Academy of Sciences,* 1134(1), pp. 320-342.

Hall-Spencer, J. M., Rodolfo-Metalpa, R., Martin, S., Ransome, E., Fine, M., Turner, S. M., Rowley, S. J., Tedesco, D. and Buia, M. C. (2008) 'Volcanic carbon dioxide vents show ecosystem effects of ocean acidification', *Nature,* 454(7200), pp. 96-99.

Hao, X., Jiang, R. and Chen, T. (2011) 'Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering', *Bioinformatics,* 27(5), pp. 611-618.

Hebert, P. D. N., Cywinska, A., Ball, S. L. and DeWaard, J. R. (2003) 'Biological identifications through DNA barcodes', *Proceedings of the Royal Society B-Biological Sciences,* 270(1512), pp. 313-321.

Hernández, C. A., Clemente, S., Sangil, C. and Hernández, J. C. (2015) 'High-resolution ocean pH dynamics in four subtropical Atlantic benthic habitats', *Biogeosciences Discuss.,* 12, pp. 19481-19498.

Hernández, C. A., Sangil, C. and Hernández, J. C. (2016) 'A new CO_2 vent for the study of ocean acidification in the Atlantic', *Marine Pollution Bulletin,* 109(1), pp. 419-426.

Hoegh-Guldberg, O., Mumby, P. J., Hooten, A. J., Steneck, R. S., Greenfield, P., Gomez, E., Harvell, C. D., Sale, P. F., Edwards, A. J., Caldeira, K., Knowlton, N., Eakin, C. M., Iglesias-Prieto, R., Muthiga, N., Bradbury, R. H., Dubi, A. and Hatziolos, M. E. (2007) 'Coral reefs under rapid climate change and ocean acidification', *Science,* 318(5857), pp. 1737-1742.

Hope, P. R., Bohmann, K., Gilbert, M. T. P., Zepeda-Mendoza, M. L., Razgour, O. and Jones, G. (2014) 'Second generation sequencing and morphological faecal analysis reveal unexpected foraging behaviour by *Myotis nattereri* (Chiroptera, Vespertilionidae) in winter', *Frontiers in Zoology,* 11(1), pp. 39.

Hopkins, G. W. and Freckleton, R. P. (2002) 'Declines in the numbers of amateur and professional taxonomists: implications for conservation', *Animal Conservation,* 5, pp. 245-249.

Hönisch, B., Ridgwell, A., Schmidt, D. N., Thomas, E., Gibbs, S. J., Sluijs, A., Zeebe, R., Kump, L., Martindale, R. C., Greene, S. E., Kiessling, W., Ries, J., Zachos, J. C., Royer, D. L., Barker, S., Marchitto, T. M., Moyer, R., Pelejero, C., Ziveri, P., Foster, G. L. and Williams, B. (2012) 'The Geological Record of Ocean Acidification', *Science,* 335(6072), pp. 1058-1063.

Inoue, S., Kayanne, H., Yamamoto, S. and Kurihara, H. (2013) 'Spatial community shift from hard to soft corals in acidified water', *Nature Clim. Change,* 3(7), pp. 683-687.

Johnson, V. R., Russell, B. D., Fabricius, K. E., Brownlee, C. and Hall-Spencer, J. M. (2012) 'Temperate and tropical brown macroalgae thrive, despite decalcification, along natural CO_2 gradients', *Global Change Biology,* 18(9), pp. 2792-2803.

Jones, J. B. (1992) 'Environmental-Impact of Trawling on the Seabed - A Review', *New Zealand Journal of Marine and Freshwater Research,* 26(1), pp. 59-67.

Kelly, R. P., Port, J. A., Yamahara, K. M., Martone, R. G., Lowell, N., Thomsen, P. F., Mach, M. E., Bennett, M., Prahler, E., Caldwell, M. R. and Crowder, L. B. (2014) 'Harnessing DNA to improve environmental management', *Science,* 344(6191), pp. 1455.

Klymus, K. E., Marshall, N. T. and Stepien, C. A. (2017) 'Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes', *PLOS ONE,* 12(5), pp. e0177643.

Knowlton, N. (1993) 'Sibling Species in the Sea', *Annual Review of Ecology and Systematics,* 24(1), pp. 189-216.

Koeppel, A. F. and Wu, M. (2013) 'Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units', *Nucleic Acids Research,* 41(10), pp. 5175-5188.

Kuffner, I. B., Andersson, A. J., Jokiel, P. L., Rodgers, K. S. and Mackenzie, F. T. (2008) 'Decreased abundance of crustose coralline algae due to ocean acidification', *Nature Geoscience,* 1(2), pp. 114-117.

Leclercq, N. I. c., Gattuso, J. E. A. N. P. and Jaubert, J. E. A. N. (2000) 'CO_2 partial pressure controls the calcification rate of a coral community', *Global Change Biology,* 6(3), pp. 329-334.

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T. A., Black, K. D. and Pawlowski, J. (2015) 'High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems', *Scientific Reports,* 5, pp. 13932.

Lenz, T. L. and Becker, S. (2008) 'Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci — Implications for evolutionary analysis', *Gene,* 427(1–2), pp. 117-123.

Leray, M. and Knowlton, N. (2015) 'DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity', *Proceedings of the National Academy of Sciences of the United States of America,* 112(7), pp. 2076-2081.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T. and Machida, R. J. (2013) 'A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents', *Frontiers in Zoology,* 10, pp. 14.

Linares, C., Vidal, M., Canals, M., Kersting, D. K., Amblas, D., Aspillaga, E., Cebrián, E., Delgado-Huertas, A., Díaz, D., Garrabou, J., Hereu, B., Navarro, L., Teixidó, N. and Ballesteros, E. (2015) 'Persistent natural acidification drives major distribution shifts in marine benthic ecosystems', *Proceedings of the Royal Society B: Biological Sciences,* 282(1818).

Lucey, N. M., Lombardi, C., Florio, M., DeMarchi, L., Nannini, M., Rundle, S., Gambi, M. C. and Calosi, P. (2016) 'An in situ assessment of local adaptation in a calcifying polychaete froma shallow CO_2 vent system', *Evolutionary Applications,* 9(9), pp. 1054-1071.

Mahe, F., Rognes, T., Quince, C., de Vargas, C. and Dunthorn, M. (2015) 'Swarm v2: highly-scalable and high-resolution amplicon clustering', *Peerj,* 3, pp. 12.

Mahé, F., Rognes, T., Quince, C., de Vargas, C. and Dunthorn, M. (2014) 'Swarm: robust and fast clustering method for amplicon-based studies', *PeerJ,* 2, pp. e593.

McElhany, P. and Shallin Busch, D. (2013) 'Appropriate pCO_2 treatments in ocean acidification experiments', *Marine Biology,* 160(8), pp. 1807-1812.

Meadows, A. S., Ingels, J., Widdicombe, S., Hale, R. and Rundle, S. D. (2015) 'Effects of elevated CO_2 and temperature on an intertidal meiobenthic community', *Journal of Experimental Marine Biology and Ecology,* 469, pp. 44-56.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., Kitoh, A., Knutti, R., Murphy, J. M., Noda, A., Raper, S. C. B., Watterson, I. G., Weaver, A. J. and Zhao, Z.-C. (2007) 'Global Climate Projections', in Solomon, S., Qin, D., Manning, M., CheN, Z., Marquis, M., Averyt, K.B., Tignor, M. & Miller, H.L. (eds.) *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change: Vol. 1*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 748-845.

Morse, J. W., Arvidson, R. S. and Lüttge, A. (2007) 'Calcium Carbonate Formation and Dissolution', *Chemical Reviews,* 107(2), pp. 342-381.

Murray, D. C., Coghlan, M. L. and Bunce, M. (2015) 'From Benchtop to Desktop: Important Considerations when Designing Amplicon Sequencing Workflows', *PLOS ONE,* 10(4), pp. e0124671.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Henry, H. and Stevens, M. 2016. Vegan: Community Ecology Package. R package version 2.3-3.a.

Oppermann, B. I., Michaelis, W., Blumenberg, M., Frerichs, J., Schulz, H. M., Schippers, A., Beaubien, S. E. and Kruger, M. (2010) 'Soil microbial community changes as a result of long-term exposure to a natural CO_2 vent', *Geochimica Et Cosmochimica Acta,* 74(9), pp. 2697-2716.

Orr, J. C., Fabry, V. J., Aumont, O., Bopp, L., Doney, S. C., Feely, R. A., Gnanadesikan, A., Gruber, N., Ishida, A., Joos, F., Key, R. M., Lindsay, K., Maier-Reimer, E., Matear, R., Monfray, P., Mouchet, A., Najjar, R. G., Plattner, G.-K., Rodgers, K. B., Sabine, C. L., Sarmiento, J. L., Schlitzer, R., Slater, R. D., Totterdell, I. J., Weirig, M.-F., Yamanaka, Y. and Yool, A. (2005a) 'Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms', *Nature,* 437(7059), pp. 681-686.

O'Donnell, J. L., Kelly, R. P., Lowell, N. C. and Port, J. A. (2016) 'Indexed PCR Primers Induce Template-Specific Bias in Large-Scale DNA Sequencing Studies', *PLOS ONE,* 11(3), pp. e0148698.

Pavan-Kumar, A., Gireesh-Babu, P. and Lakra, W. (2015) 'DNA Metabarcoding: A New Approach for Rapid Biodiversity Assessment', *Cell Science and Molecular Biology,* 2(1), pp. 1-9.

Pearman, J. K. and Irigoien, X. (2015) 'Assessment of Zooplankton Community Composition along a Depth Profile in the Central Red Sea', *PLOS ONE,* 10(7), pp. e0133487.

Porzio, L., Garrard, S. L. and Buia, M. C. (2013) 'The effect of ocean acidification on early algal colonization stages at natural CO_2 vents', *Marine Biology,* 160(8), pp. 2247-2259.

Ratnasingham, S. and Hebert, P. D. N. (2007) 'bold: The Barcode of Life Data System (http://www.barcodinglife.org)', *Molecular Ecology Notes,* 7(3), pp. 355-364.

Raulf, F. F., Fabricius, K., Uthicke, S., de Beer, D., Abed, R. M. M. and Ramette, A. (2015) 'Changes in microbial communities in coastal sediments along natural CO_2 gradients at a volcanic vent in Papua New Guinea', *Environmental Microbiology,* 17(10), pp. 3678-3691.

Raven, J., Caldeira, K., Elderfield, H., Hoegh-Guldberg, O., Liss, P., Riebesell, U., Shepherd, J., Turley, C. and Watson, A. (2005) *Ocean acidification due to increasing atmospheric carbon dioxide*: Royal Society Special Report.

Reeder, J. and Knight, R. (2010) 'Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions', *Nat Meth,* 7(9), pp. 668-669.

Riebesell, U., Schulz, K. G., Bellerby, R. G. J., Botros, M., Fritsche, P., Meyerhofer, M., Neill, C., Nondal, G., Oschlies, A., Wohlers, J. and Zollner, E. (2007) 'Enhanced biological carbon consumption in a high CO_2 ocean', *Nature,* 450(7169), pp. 545-548.

Riebesell, U., Zondervan, I., Rost, B., Tortell, P. D., Zeebe, R. E. and Morel, F. M. M. (2000) 'Reduced calcification of marine plankton in response to increased atmospheric CO_2', *Nature,* 407(6802), pp. 364-367.

Riesenfeld, C. S., Schloss, P. D. and Handelsman, J. (2004) 'Metagenomics: Genomic Analysis of Microbial Communities', *Annual Review of Genetics,* 38(1), pp. 525-552.

Rodolfo-Metalpa, R., Lombardi, C., Cocito, S., Hall-Spencer, J. M. and Gambi, M. C. (2010) 'Effects of ocean acidification and high temperatures on the bryozoan Myriapora truncata at natural CO_2 vents', *Marine Ecology-an Evolutionary Perspective,* 31(3), pp. 447-456.

Russell, B. D., Connell, S. D., Uthicke, S., Muehllehner, N., Fabricius, K. E. and Hall-Spencer, J. M. (2013) 'Future seagrass beds: Can increased productivity lead to increased carbon storage?', *Marine Pollution Bulletin,* 73(2), pp. 463-469.

Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., Wanninkhof, R., Wong, C. S., Wallace, D. W. R., Tilbrook, B., Millero, F. J., Peng, T.-H., Kozyr, A., Ono, T. and Rios, A. F. (2004) 'The Oceanic Sink for Anthropogenic CO_2', *Science,* 305(5682), pp. 367-371.

Sanders, H. L. (1968) 'Marine Benthic Diversity: A Comparative Study', *The American Naturalist,* 102(925), pp. 243-282.

Schloss, P. D., Gevers, D. and Westcott, S. L. (2011) 'Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies', *PLOS ONE,* 6(12), pp. e27310.

Schloss, P. D. and Westcott, S. L. (2011) 'Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis', *Applied and Environmental Microbiology,* 77(10), pp. 3219-3226.

Schmidt, P.-A., Bálint, M., Greshake, B., Bandow, C., Römbke, J. and Schmitt, I. (2013) 'Illumina metabarcoding of a soil fungal community', *Soil Biology and Biochemistry,* 65, pp. 128-132.

Shaw, J. L. A., Clarke, L. J., Wedderburn, S. D., Barnes, T. C., Weyrich, L. S. and Cooper, A. (2016) 'Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system', *Biological Conservation,* 197, pp. 131-138.

Shokralla, S., Spall, J. L., Gibson, J. F. and Hajibabaei, M. (2012) 'Next-generation sequencing technologies for environmental DNA research', *Molecular Ecology,* 21(8), pp. 1794-1805.

Taberlet, P., Coissac, E., Hajibabaei, M. and Rieseberg, L. H. (2012a) 'Environmental DNA', *Molecular Ecology,* 21(8), pp. 1789-1793.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. and Willerslev, E. (2012b) 'Towards next-generation biodiversity assessment using DNA metabarcoding', *Molecular Ecology,* 21(8), pp. 2045-2050.

Thomsen, P. F. and Willerslev, E. (2015) 'Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity', *Biological Conservation,* 183, pp. 4-18.

Wangensteen, O. S., Guardiola, M., Palacín, C. and Turon, X. (2017) 'Metabarcoding shallow marine hard-bottom communities: unexpected diversity and database gaps revealed by two molecular markers', *Submitted for publication.*

Wangensteen, O. S. and Turon, X. (2015) 'Metabarcoding Techniques for Assessing Biodiversity of Marine Animal Forests', in Rossi, S., Bramanti, L., Gori, A. & Orejas Saco del Valle, C. (eds.) *Marine Animal Forests: The Ecology of Benthic Biodiversity Hotspots*. Cham: Springer International Publishing, pp. 1-29.

Wei, Z.-G., Zhang, S.-W. and Jing, F. (2016) 'Exploring the interaction patterns among taxa and environments from marine metagenomic data', *Quantitative Biology,* 4(2), pp. 84-91.

Westcott, S. L. and Schloss, P. D. (2015) 'De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units', *Peerj,* 3, pp. 23.

Wheeler, Q. D., Raven, P. H. and Wilson, E. O. (2004) 'Taxonomy: Impediment or expedient?', *Science,* 303(5656), pp. 285-285.

Yu, D. W., Ji, Y. Q., Emerson, B. C., Wang, X. Y., Ye, C. X., Yang, C. Y. and Ding, Z. L. (2012) 'Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring', *Methods in Ecology and Evolution,* 3(4), pp. 613-623.