A dive into the coral microbiome

Hadrien Gourlé
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics
Uppsala

Doctoral thesis
Swedish University of Agricultural Sciences
Uppsala 2020

Acta Universitatis Agriculturae Sueciae
2020:29

Cover: corals, picture taken off the coast of Mauritius
(photo: Hadrien Gourlé)

ISSN 1652-6880
ISBN (print version) 978-91-7760-576-8
ISBN (electronic version) 978-91-7760-577-5

Print: SLU Service/Repro, Uppsala 20XX

A dive into the coral microbiome

Abstract

Coral reefs are one of the most diverse ecosystems on the planet, harbouring approximately twenty-five percent of the diverse eukaryotic life in the oceans, while also being important economically for millions of people worldwide. Despite their importance, coral reefs are menaced of a very steep decline due to pollution and anthropogenic climate change.

In this thesis, we investigate the microbes that live close-by and inside coral reefs. It is believed that microbiomes, both environmental and coral-associated, play an important role in coral health, both by contributing to nutrient cycling, such as carbon and nitrogen fixation as well as photosynthesis, and by protecting the corals against environmental stressors such as pathogens. These microbiomes can be studied using targeted approaches, such as metabarcoding, or more general and powerful approaches, called metagenomics.

Metagenomics is a relatively new field of study and the first part of this thesis focuses on method development for metagenomics. In paper I, we present InSilicoSeq, a software package to simulate metagenomic Illumina reads. InSilicoSeq is useful for testing new bioinformatics methods as well as benchmarking existing ones.

In paper II and III, we study the composition of the coral microbiome from previously published studies, and the composition and function of the microbiome of the water and upper sediment layer from reefs of the Kenyan coast of the west Indian ocean. We define a putative coral core microbiome at the genus level and take a look at the metabolic pathways that may be active in the surrounding environment of the corals. While the coral core microbiome was largely dominated by one genus, *Endozoicomonas*, the surrounding environment showed great diversity both in taxonomy and in metabolism. We found evidence of antibiotics resistance in the water, which we hypothesise mainly comes from agriculture. We also publish a catalogue of putative expressed pathways and discovered 174 new bacteria in the water and sediment samples.

Keywords: metagenomics, coral reefs, sequencing, bioinformatics, microbiome

Author's address: Hadrien Gourlé, SLU, Department of Animal Breeding and Genetics,
P.O. Box 7023, 750 07 Uppsala, Sweden

Dedication

For a better, kinder, more sustainable world

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I **Gourlé H***, Karlsson O, Hayer J and Bongcam-Rudloff E (2018). Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, vol 35, Issue 3, pp. 521-522.

II **Gourlé H***, Lindsjö O and Hayer J (2020). A Data Driven Review of the Coral Microbiome. (manuscript)

III Wambua S**, **Gourlé H*****, De Villiers E, Karlsson-Lindsjö O, Wambiji N, Macdonald A, Bongcam-Rudloff E and De Villiers S. Molecular Ecology of Coral Reef Microorganisms in the Western Indian Ocean coast of Kenya. (submitted)

* Corresponding author.
** Contributed equally to this publication

The contribution of HG to the papers included in this thesis was as follows:

I HG developed the software and wrote the paper.

II HG planned the study, analysed the data and wrote the paper

III HG contributed to the study planning, analysed the data, and wrote the paper.

Publications not included in this thesis

IV Norling M*, Karlsson-Lindsjö O, **Gourlé H**, Bongcam-Rudloff E, Hayer J (2016). MetLab: an insilico experimental design, simulation and analysis tool for viral metagenomics studies. *Plos One*, vol 11, Issue 8.

V Hernández-de-Diego R*, de Villiers E, Klingström T, **Gourlé H**, Conesa A, Bongcam-Rudloff E (2018). The eBioKit, a stand-alone educational platform for bioinformatics. *Plos Computational Biology*, vol 13, Issue 9.

VI Louyakis A*, **Gourlé H**, Casaburi G, Bonjawo R, Dusher A, Foster J (2018). A year in the life of a thrombolite: comparative metatranscriptomics reveals dynamic changes over diel and seasonal cycles. *Environmental microbiology*, vol 20, Issue 2, pp. 842-861.

* Corresponding author.

List of tables

List of figures

Abbreviations

CO_2	Carbon Dioxide
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNA	Deoxyribonucleic acid
GLM	Generalised Linear Model
KDE	Kernel Density Estimation
MDA	Multiple-Displacement Amplification
ML	Machine Learning
NCBI	National Center for Biotechnology Information
OTU	Operational Taxonomic Unit
RAM	Random-Access Memory
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
WMS	Whole Metagenome Sequencing

1 Introduction

1.1 Coral Reefs

Corals are underwater organisms of the phylum *Cnidaria* that form colonies of stone-like, hard calcified structures, called coral reefs. In their individual forms, corals are polyps. Coral polyps are composed of two layers of cells, covered by a mucus layer and cover a large calcified skeleton (Figure 1). Unlike jellyfish, corals will never enter a medusa stage and will be polyps their whole life. Most corals are stony – or reef-building corals – they will attach to stone or other structures will be slowly building a calcified structure, of the order of a few centimetres per year (Anderson *et al.*, 2017).

Reef-building corals are found in tropical water around the globe. They are thought to be the most diverse ecosystem on the planet, harbouring roughly twenty-five percent of the macroscopic life in the ocean (Spalding and Grenfell, 1997; Plaisance *et al.*, 2011). This diversity makes them invaluable for the survival of many animal species worldwide, such as thousands of species of fish, algae, sponges, crustaceans and more.

Corals have a symbiotic relationship with a unicellular alga: in the vast majority of case, *Symbiodinium,* that provides essential energy to corals via photosynthesis. *Symbiodinium* is an endosymbiont, meaning they live within the tissues of corals.

Coral reefs are extremely important from an economic point of view. They are often valued in billions, sometimes trillions of dollars, both from tourism and the fishing industry (Costanza *et al.*, 1997). Unfortunately, coral reefs are very sensitive to a changing environment and face increased and imminent threats due to anthropogenic climate change. Factors such as the acidification of oceans and mostly the rise in ocean temperature, contribute to their rather quick degradation.

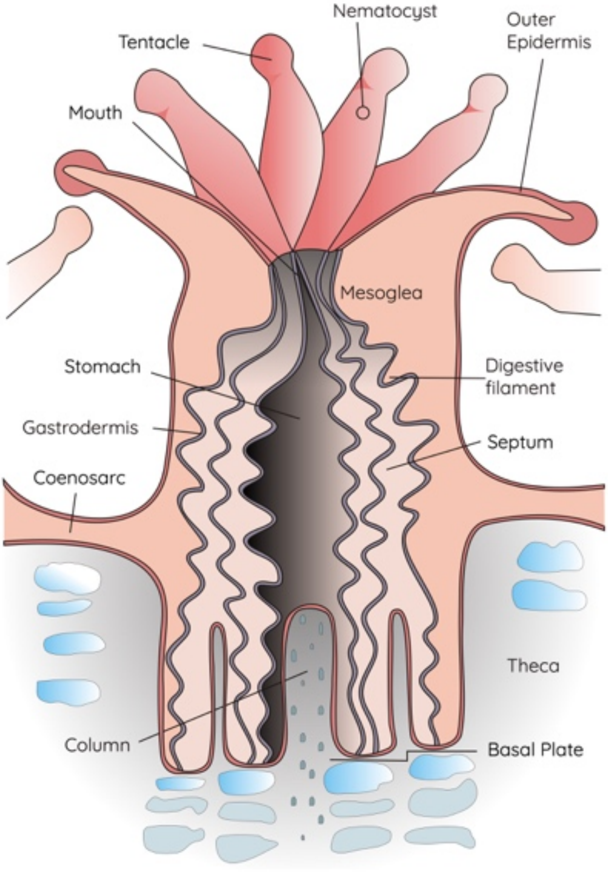


Figure 1: Anatomy of a coral polyp. In pink, the two layers of cells, called the epidermis and gastrodermis. The blue and grey shapes indicate the skeleton produced by the coral polyp.

The most common cause of coral degradation is bleaching. During a bleaching event, corals eject their symbiotic partner resulting in loss of pigmentation, increased susceptibility to disease, and if the bleached state persists, death of the coral (Figure 2) (Graham *et al.*, 2006; Hughes, Kerry, *et al.*, 2017).

Without reducing drastically our CO_2 emissions, the majority of coral reefs will likely not survive much longer, and whole ecosystems may collapse. In the meantime, researchers can do everything they can for helping corals to be more resistant to climate change, but any fix is ultimately temporary (Hughes, Barnes, *et al.*, 2017).

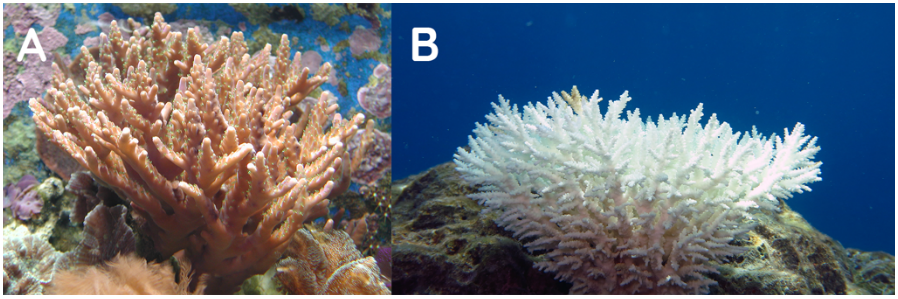


Figure 2: Panel A displays a healthy Acropora sp. coral. Panel B shows a bleached coral. The image to left was taken by Vardhan Patankar and to the right by David Excoffier. Both images are licensed under CC BY-SA 4.0 international.

1.2 The coral microbiome

Microbes can be found in all multicellular organisms. Barring the case of an infection, microbes live a commensal or even symbiotic relationship with their hosts, often providing with important immunological or metabolic functions (Norman, Handley and Virgin, 2014). Microbes are found everywhere and do not need a host to thrive; they are also found in the environment. Those communities of microbes, or microbiomes, outnumber eukaryotic organisms by order of magnitudes (Kallmeyer *et al.*, 2012). Corals most likely have distinct microbiomes in the mucus layer, the tissue and the skeleton (Hernandez-Agreda, Gates and Ainsworth, 2017).

While much is known about the endosymbiont living within the corals, we know relatively little about the coral microbiome and its functions. Similarly, macroscopic life living within the boundaries of reef ecosystems have been the focus of ecologists for years, but the microbes living in the water and upper sediments layers of reef areas remained elusive, despite the fact that they may be one of the most diverse microbial community on the planet, and that they certainly contribute to the ecosystem, being part of the microbial food-web and at the very least enabling nutrition (Mostajir *et al.*, 2015).

That the coral-associated microbial communities play an important role in coral health is not a novel idea (Sorokin, 1973) but it is only recently that microbiome research has started to be conducted in coral reefs environments (Tout *et al.*, 2014; Blackall, Wilson and van Oppen, 2015; Bourne, Morrow and Webster, 2016), and we are starting to discover exactly which microbes living close by – and inside – coral reefs may be extremely important for the reef ecosystem, for processes such as carbon cycling, nitrogen fixation, photosynthesis, and even protection against pathogens via antimicrobial activities (Figure 3).

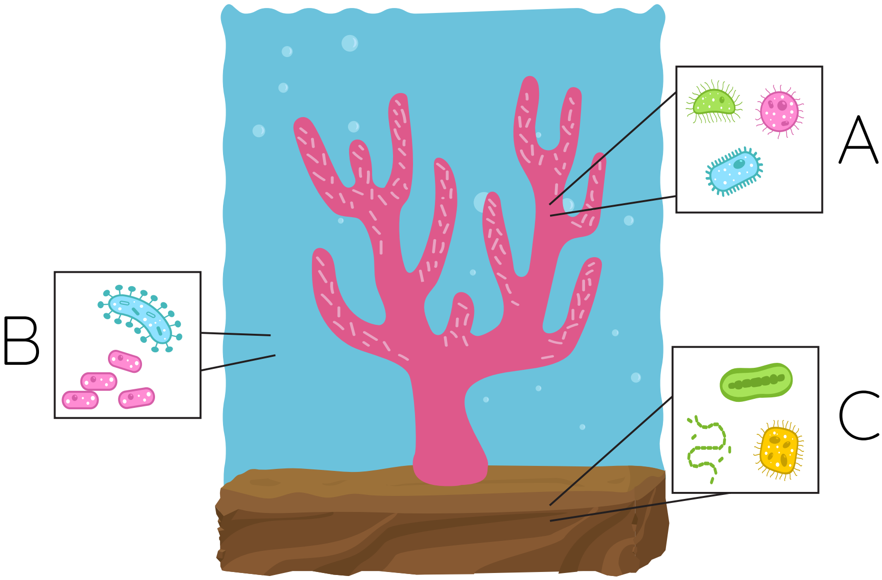


Figure 3: The reef microbiome. Microorganisms are present inside the coral (A), in the water (B) and in the sediment layer of the reef (C). These communities are likely to contribute a great deal into their environment, by means of diverse metabolic pathways, encompassing photosynthesis, carbon and nutrient cycling, nitrogen fixation and more. This image has been designed using resources from freepik.com

1.3 Sequencing

The two following sections will briefly describe the methods that can be used to analyse the genetic material of microbiomes.

Over the last years, high-throughput DNA and RNA sequencing have been claiming the lion's share of the methods used in almost all fields of biological research (Reuter, Spacek and Snyder, 2015; Park and Kim, 2016). From new insights in cancer to bio-prospecting forests for new pharmaceutics, the use of whole-genome or transcriptome sequencing technologies has truly revolutionised how we approach designing modern experiments, especially for microbial ecology, where many organisms that could not be grown in laboratory settings have now been discovered and had their genome sequenced.

In that revolution, one should not forget metabarcoding, which for many years has been an affordable and quick alternative to characterise the taxonomic distribution and difference between microbial communities and compare microbial ecosystems under different conditions. Although we did not collect and sequence samples for metabarcoding in this thesis, we reanalysed metabarcoding data in paper II and the methodology will be further described in the methodology section.

1.3.1 Sequencing Platforms

Several sequencing techniques exist, which we can divide into short- and long-read sequencing. An overview of the current platforms is available in Table 1.

Table 1: Overview of modern sequencing technologies

Platform	Read length	Throughput
Illumina MiSeq	2*150-300	15Gbp
Illunina NovaSeq	2*50-250	up to 3Tbp
Pacbio Sequel	30kb	20Gbp
Oxford Nanopore MinION	up to mb	15Gbp
Oxford Nanopore GridION	up to mb	250Gbp
Oxford Nanopore PromethION	up to mb	up to 10Tbp

The most popular platform for metagenomics is currently Illumina, mainly due to the high throughput necessary for recovering genomes of medium and low abundance species, but the availability and price-tag of the Oxford Nanopore MinION may make it a contender for field sequencing in the very near future.

1.4 Bioinformatics applied to Metagenomics

Once genomes, transcriptomes, or amplicons have been sequenced, different analyses are possible. These analyses are usually platform-independent, meaning that regardless of the sequencing technology used, the tools used might differ but the principles behind the bioinformatics workflows and pipelines used will not differ much.

A typical workflow starts with quality control: the reads produced by the sequencers usually contain errors, as well as technical artefacts such as barcodes and adapter sequences.

1.4.1 Quality Control

Reads as they come out of the sequencer – or, more correctly, the basecalling software – are not free of errors. Errors can come in different flavours: substitutions, insertions, or deletions. Each sequencing platform has their known error profiles, including things such known substitution patterns, deletions on some long homopolymers, or rather large insertions or deletions.

In a read, each base gets assigned a probability of being the correct one by the basecalling software. These probabilities are displayed using PHRED scores, defined as the logarithm of the error probability, as shown in the equation below. PHRED score as reported by the basecalling software usually range from 0 to 40, where their respective probabilities are shown in table 2.

$$Q = -10 \cdot log_{10}P$$

Table 2: PHRED score and their probabilities

Q	P
10	90
20	99
30	99.9
40	99.99

Quality Control is a straightforward step in all bioinformatics analyses, where we (i) remove the adapters and barcodes (the technical sequences used by different technologies to prime flowcells and identify different sample to sequence them in the same run and (ii) trim the bad quality bases (which, in Illumina sequencing, usually happen at the end of the reads), usually Q lower than 5 (Macmanes, 2014).

1.4.2 16S Metabarcoding

16S metabarcoding is a very popular, affordable and quick way to get an overview of the taxonomic composition of a bacterial or archaeal community. In a metabarcoding experiment, only one gene, or part of a gene (The 16S rRNA gene in the case of bacteria and archaea) is sequenced. This gene or part of a gene is called a barcode. Since all bacteria carry that gene and that gene is conserved across the whole kingdom, we can infer the genus of a bacteria only for one or two regions of the 16S gene.

The 16S rRNA gene is about 1600bp long and contains nine hypervariable regions (Figure 4). Many short-read (i.e. Illumina) studies use primer pairs spawning 1 or 2 regions, per example V3-V4 is standard in human microbiome protocols. With long-read sequencing, one can usually sequence the entirety of the 16S gene, allowing for slightly better resolution, and a more complete overview of the sampled population.

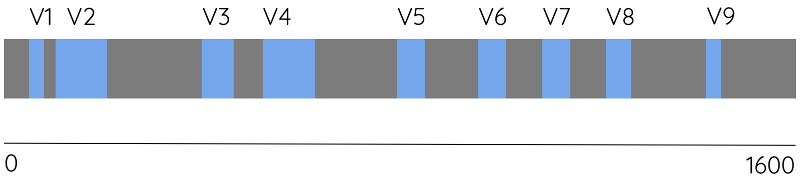


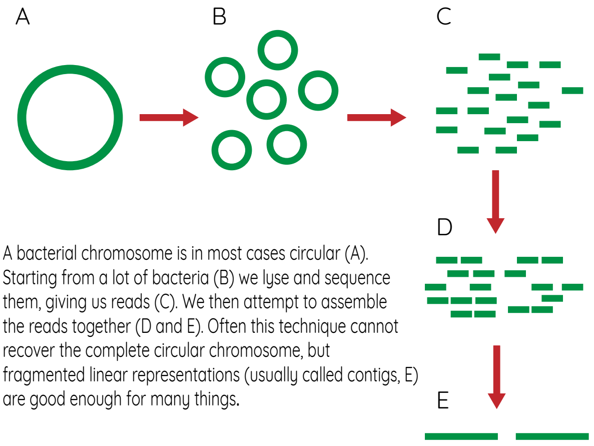
Figure 4: The 16S rRNA hypervariable regions

1.4.3 Metagenome assembly and Binning

As an alternative to metabarcoding, one can sequence the entirety of genomes present in a community.

Once the bad quality bases and technical artefacts have been removed from the reads, we assemble the reads in longer contiguous sequences, called contigs. The current most-used algorithm for short reads makes use of de Bruijn graphs. Briefly, short reads are split into overlapping k-mers (words of length k) and a graph connecting those k-mers together is created. The graph's inaccuracies (dead ends, bubbles, …) are then corrected, and we "simply" have to follow the path to reconstruct the genome or at least the longest possible genomic fragments. (Figure 5)

Current algorithms, combined with the fact that reads from high throughput sequencing are often short (max 300bp for Illumina) do not, most of the time, allow us to reconstruct complete, circularised bacterial chromosomes. Typically for single genomes projects, assemblies will consist of up to 100 contigs for a decent draft bacterial assembly constituted only of short reads. To obtain complete or near-complete genomes, it is possible to complement with long reads.

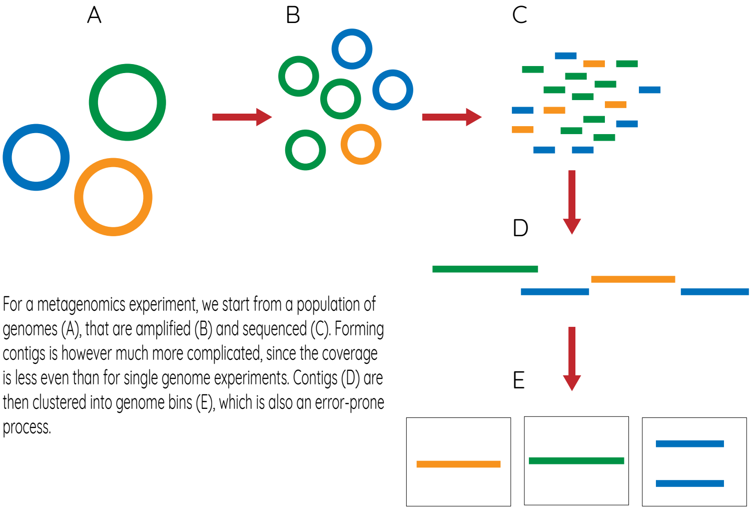


A bacterial chromosome is in most cases circular (A). Starting from a lot of bacteria (B) we lyse and sequence them, giving us reads (C). We then attempt to assemble the reads together (D and E). Often this technique cannot recover the complete circular chromosome, but fragmented linear representations (usually called contigs, E) are good enough for many things.

Figure 5: Single genome assembly workflow, from a bacterial chromosome to contigs

In the case of metagenomes, special assemblers have seen days and usually try to solve the difficult problem that is metagenome assembly. Depending on the complexity of the community we are studying which may contain from tens to hundreds of organisms in various abundances, we may end up with thousands or millions of contigs. We need then to separate these contigs into genomes, in a process called binning (Figure 6).

The binning process tries to separate the contigs obtained during the assembly into individual organisms. It usually uses a combination of coverage information and tetra-nucleotide frequencies. Organisms in a community will often have different genome coverage, simply because they are found in different abundance. Tetranucleotide frequency is the frequency of each unique 4-mer (a k-mer of length four) in a DNA sequence, which tends to be conserved (Noble, Citek and Ogunseitan, 1998). The current methods are not perfect, and recovered draft genomes usually exhibit various degrees of contamination.



For a metagenomics experiment, we start from a population of genomes (A), that are amplified (B) and sequenced (C). Forming contigs is however much more complicated, since the coverage is less even than for single genome experiments. Contigs (D) are then clustered into genome bins (E), which is also an error-prone process.

Figure 6: Metagenome assembly and binning workflow. From a community of bacteria to putative draft genomes

1.4.4 Genome Annotation and Phylogeny

This is the main strength of whole metagenome studies compared to 16S metabarcoding. Once we have collected draft genomes, we can infer which genes they contain, and get insights about their metabolism, and ultimately their contribution to the community and the environment.

In the case we have assembled and binned genomes that have not been previously discovered, the typical phylogeny workflow is called phylogenetic placement. It is done by taking a known phylogeny (i.e. the current bacterial tree of life) and placing our new genome to an edge of the tree, typically using maximum likelihood methods.

1.4.5 Read-level Taxonomic Classification

While sometimes not included in the "recovery of metagenome-assembled genomes" workflow, a popular method employed for quick community typing of metagenomic dataset is read-level classification, as well as for comparative metagenomics. With this method, the reads are directly matched against databases. Most current methodologies employ either alignment (approximate string matching) against a FM index, or direct k-mer matching against a k-mer database.

1.4.6 Differential Expression and Pathway Analysis

Newer and more challenging, metatranscriptomics leaves us with a much closer picture of the actual metabolism of a bacterial community. While metagenomics will find insights into functions in the genomes, there is little proof that those genes are actually expressed.

Metagenomics allows us to investigate metabolic potential, while metatranscriptomics allows us to see which genes are actually expressed, and therefore more likely to play an actual role in the community's metabolism. The currently best methods for differential expression of metatransciptomes are borrowed from transcriptomics, and fit linear models for each gene, to see if their expression patterns are significantly different across conditions.

2 Aims of the Thesis

The overarching aim of doing research on coral reef is evidently to prevent reefs from going extinct. While we will not achieve this without drastically reduce CO_2 emissions, coral researchers worldwide are trying to delay the decay of reefs, while hopefully policy makers act on CO_2 emissions.

Within a more medium-term timeframe, we hypothesise that the coral microbiome as well as the reef microbial ecosystems play important roles in the health and disease of coral reefs. They might, to go back to that long-term perspective above, be used in coral remediation. Additionally, since the microbiome is usually very sensitive to the environment, we hypothesise that it could be a good indicator of coral health and show signs of imbalance that could be missed while only looking at water properties or the physiology of coral reefs. We may also find biomolecules relevant for human health in those poorly studied systems, that could have applications in human medicine. Lastly, in the context of culturing corals for later reimplantation, it is useful to know about their usual environment.

In a more immediate perspective, the goal of this thesis is firstly to evaluate and build the necessary frameworks to improve environmental metagenomics and metatranscriptomics methodologies. Secondly, we aimed to create a baseline for reef microbial studies in the West Indian Ocean region, which was lacking and can and will be used for many future studies. Lastly, we thought that the mucosal microbiome of the coral may play a crucial role in protection against disease, and we wanted to identify potential microbes, pathways and molecules that could help protect coral against pathogens. This inscribes into the medium-term coral remediation aims, as well as bioprospecting.

Studying and trying to save and preserve coral reefs is extremely important for many people worldwide, and this thesis hopefully contributes even a little bit.

3 Methodology

This section describes the methods that were used for the three studies included in this thesis.

3.1 Simulating Illumina metagenomic data with InSilicoSeq (Paper I)

3.1.1 Illumina Error Types

The error type of Illumina instruments has been subject to a few studies, and we have a decent sense of what are the most common systematic errors that the technology makes (Schirmer *et al.*, 2016; Ma *et al.*, 2019). Most substitution errors are detected by the basecaller and accordingly flagged as poor quality. An increase of errors towards the end of the reads, as well as a higher error rate for the reverse reads, is generally observed. Insertions and deletions (indels) also occur at higher rate than substitutions and are less likely to be flagged as poor quality.

During the early development phase of InSilicoSeq we considered using existing error profiles from the literature, but it appeared that error profiles are machine- and protocol-specific and therefore we decided to model errors based when possible on our metagenomic datasets, or when it was not possible, on the most recent data available from Illumina.

3.1.2 Kernel Density Estimation

The Illumina basecaller transforms the light signal from the instrument into nucleotide calls and assign a probability of that call being incorrect. While the probability distribution is continuous, these probabilities end up being binned into PHRED scores, typically ranging from 1 to 40. For each base of the read, InSilicoSeq estimates the probability distribution using two-dimensional kernel density estimation (Figure 8).

Kernel Density Estimation (KDE) is a non-parametric method used to estimate probability densities (Silverman, 1986). Briefly, for every nucleotide position, the KDE in InSilicoSeq defines a gaussian kernel for each PHRED score at that nucleotide position and then sums the gaussian kernels for producing the density function (Figure 7). The gaussian have each a bandwidth parameter, which has a lot of influence over the final shape of the distributions. A bandwidth too small will lead to undersmoothing, while a bandwidth too large may lead to undersmoothing (Sheather and Jones, 1991).

For the density f of the unknown distribution of basecall probabilities, its Kernel Density Estimator is

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\,(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where K_h is the gaussian function of bandwidth h, and x an observation of the unknown PHRED score distribution.

In InSilicoSeq, we empirically chose the bandwidth parameter h to be $0.2\ /\ s$, s being the sample standard deviation; that bandwidth produced the smoothest "Illumina-like" quality distributions.

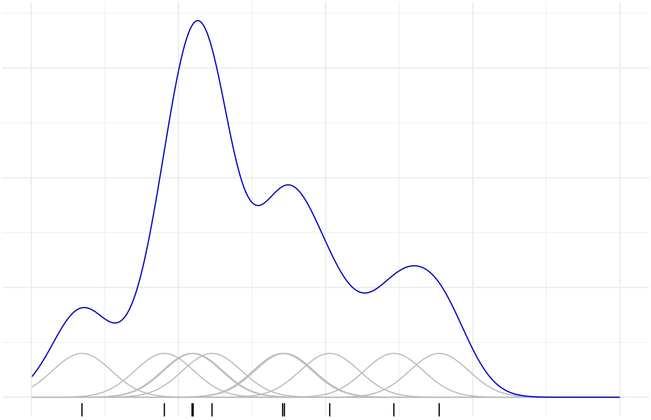


Figure 7: One-dimensional KDE

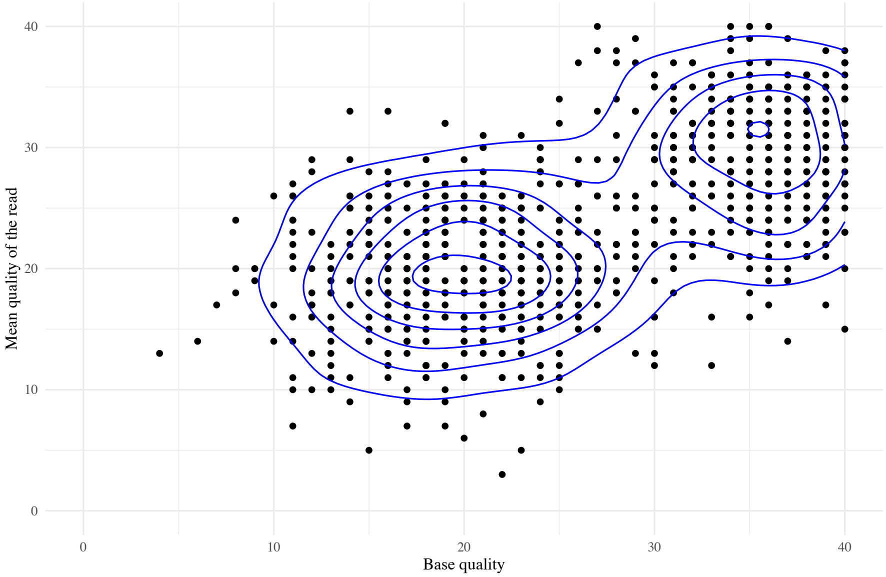


Figure 8: 2D KDE as implemented in InSilicoSeq

3.2 A data-driven Review of the Coral microbiome (Paper II)

A total of nineteen papers were selected, having a reference to a bioproject hosting publicly available data. Of those, six did not have data in the bioproject, or had insufficient metadata and were therefore excluded from the following analysis. The remaining thirteen papers are presented in Table 3.

Table 3: Datasets used for paper II

Authors	Bioproject	Method	Sequencing Technology
Robertson *et al.*, 2016	PRJNA313050	mothur 97%	454
Lawler *et al.*, 2016	PRJNA296835	qiime 1 97%	454
van de Water *et al.*, 2016	PRJNA312472	qiime 1 97%	MiSeq
Till Bayer *et al.*, 2013)	PRJNA189184	mothur 97%	454
Glasl, Herndl and Frade, 2016	PRJNA310360	qiime 1 98%	454
Roder *et al.*, 2015l	PRJNA277291	mothur 97%	454
T. Bayer *et al.*, 2013	SRP010998	mothur 97%	454
Hadaidi *et al.*, 2017	PRJNA352338	mothur 97%	MiSeq
Vezzulli *et al.*, 2013	PRJNA192455	mothut 97%	454
Kellogg, Ross and Brooke, 2016	PRJNA297333	qiime 1 97%	454
van de Water *et al.*, 2017	PRJNA312774	qiime 1 97%	454
Meyer, Paul and Teplitski, 2014	PRJNA231864	qiime 1 97%	454

3.2.1 16S analysis

For many years metabarcoding experiments were done by clustering OTUs together, typically at 97% sequence identity. This was done for two main reasons and has been somewhat controversial. Firstly, clustering into OTUs tends to lump together paralogs (species that have more than one copy of the 16S gene). Secondly, errors due to sequencing will also be merged together, and OTUs should only represent real biological variation, i.e. different species (Nguyen *et al.*, 2016; Edgar, 2018).

Published in 2016, dada2 offers a new approach to metabarcoding data analysis, correcting systematic errors present in sequencing datasets and skipping clustering into OTUs, instead assigning taxonomy to every sequence variant (Callahan *et al.*, 2016). Dada2 can also detect organisms that have multiple copies of the 16S gene, and deal with them appropriately. It outputs less spurious sequences than previously described methods.

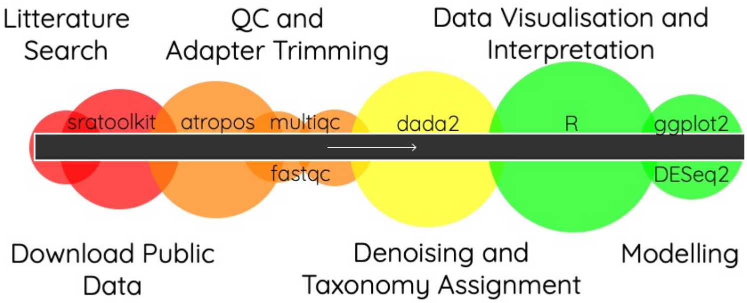


Figure 9: Pipeline used for paper II

3.2.2 Modelling of the core microbiome

To identify the microbes that were present in all datasets, we employed the DESeq2 software package (Love, Huber and Anders, 2014). Initially developed for transcriptomics and differential RNA-seq analysis, DESeq2 has proven robust, and is known to perform well with count matrices for other data types than RNA-seq.

Briefly, DESeq2 fits a generalised linear model (GLM) for each observation (row) from a count matrix of the form

$$K_{ij} \sim NB(\mu_{ij}, \alpha_{ij})$$

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = k_j \beta_i$$

where α_{ij} is the dispersion parameter, also called the within-group variance. DESeq2 estimates this within-group variance via maximum likelihood (Figure 10). The mean μ_{ij} is composed of a size factor s and q_{ij}, which is the expected true count, proportional to the expected true fragment concentration x_j and a coefficient β_i. In our study, we used contrasts to make the specific comparison of groups. DESeq2 calculates contrasts after fitting the GLM and calculate the Wald statistic by multiplying the coefficient β_i by a contrast vector on the numerator, and by taking the square root of the product of the covariance matrix by the contrast vector on the denominator.

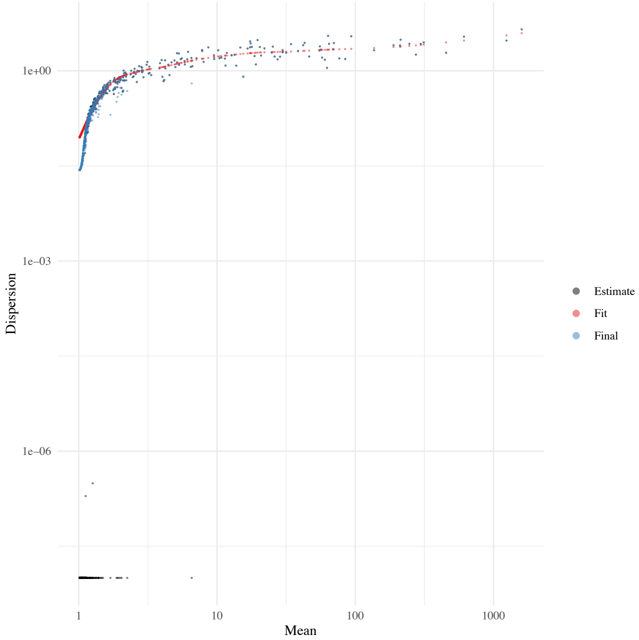


Figure 10: Dispersion estimates for the core microbiome modelling

3.3 Molecular Ecology of Coral Reef Microorganisms in the Western Indian Ocean coast of Kenya (Paper III)

3.3.1 Sample collection and preparation

The three samples sites were chosen for their distinct human activities. Mombasa and Malindi were arguably the most polluted, due to their numerous touristic activities, and their proximity to large industrial hubs. Kilifi lies a bit farther from the coast, and suffers less from pollution, besides boat traffic at large of the island.

For the water samples, we collected four litres per sample that were then filtered through a 0.2 µm membrane to retain only the microbial cells. The sediment samples were collected using a 10 ml syringe barrel. DNA was extracted using Mobio PowerWater and PowerSoil isolation kits, respectively. Samples were then amplified by multiple displacement amplification (MDA) using Qiagen's REPLI-G kit.

3.3.2 Choice of sequencing platform and depth

Illumina sequencing was chosen due to the diversity of the samples. We performed a pilot study using Illumina MiSeq and extrapolated the needed sequencing depth to recover ninety-five percent of the microbial population. Given the needed sequencing depth, Illumina was the only acceptable choice, and the NovaSeq instrument would give us the best price per base. The pilot study was also used to verify our protocol and check that contamination (human and phytoplankton) was minimal. We used nonpareil (Rodriguez-R and Konstantinidis, 2014) for extrapolating the coverage needs (Figure 11). It was estimated that to recover 95% of the organisms living in the community, we would need between 172 and 343 Gigabases of data.

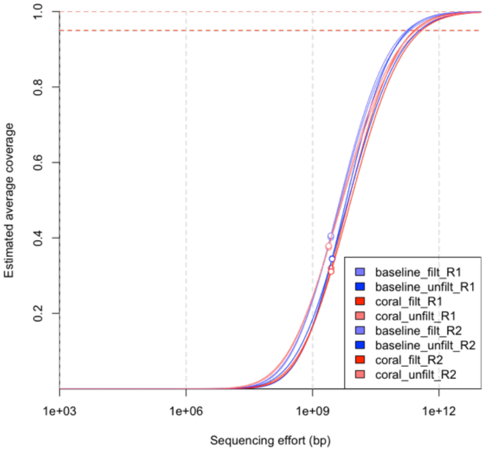


Figure 11: Nonpareil extrapolation curves

3.3.3 Bioinformatics Analyses

The bioinformatics pipeline used in Paper III can be divided in 3 parts (Figure 12):

- Taxonomic classification
- Metagenome assembly and binning
- Protein assembly and annotation

In this paper we used kraken2 (Wood, Lu and Langmead, 2019), which uses exact k-mer matches to classify reads against a nucleotide database (nt in this case). Each k-mer originating from a sequence is assigned to the lowest common ancestor of all the sequences in the database containing that k-mer. The hits from the forward and reverse read are then compared to each other, and their lowest ancestor is selected if they are not in agreement.

Assembly was performed with megahit (Li *et al.*, 2015). Megahit is a de-Bruijn graph-based assembler that is optimised for metagenomes. It uses comparatively less-memory than other assemblers, making it possible to assemble larger metagenomes with a reasonable RAM footprint. This is made possible by the use of a succinct de Bruijn graph, which is a compressed representation of a de Bruijn graph. The assemblies were then binned (or clustered) with the objective of recovering draft genomes from individual organisms. Metabat2 was used for the binning (Kang *et al.*, 2019). Metabat2 uses coverage information, obtained by mapping the reads back to the assembly with bowtie2 (Langmead and Salzberg, 2012), and tetranucleotide frequencies to cluster similar contigs together, using graph partitioning with the contigs as nodes, and their similarity scores as edges. The binning quality is then checked with checkm (Parks *et al.*, 2015) and the bins are eventually refined with refinem (Parks *et al.*, 2017).

For the protein assemblies, we used the Plass software (Steinegger, Mirdita and Söding, 2018). Plass is a greedy assembler, that does not use a de Bruijn graph, but rather relies on all vs all overlap calculation of reads, which plass achieves in linear time. The plass assemblies, as well as the genome bins passing the quality threshold were then annotated using gtdb-tk (Chaumeil *et al.*, 2019) and eggnog-mapper (Huerta-Cepas *et al.*, 2017).

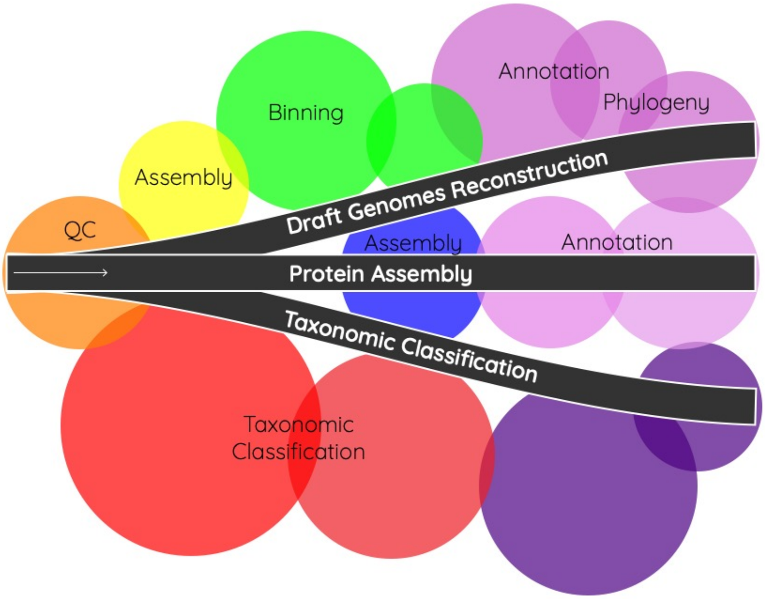


Figure 12: Pipeline used for paper III

4 Results and Discussion

As mentioned earlier the objectives of the thesis were two-fold. Firstly, it was important to investigate and develop new methods for metagenomics. Secondly, we wanted to employ stat-of-the-art methods to investigate the microbial communities living within reefs.

4.1 Simulating Illumina metagenomic data with InSilicoSeq (Paper I)

Software for simulating genomics and metagenomics existed previously, but most were outdated, undocumented or unmaintained (Escalona, Rocha and Posada, 2016). InSilicoSeq makes it easier to simulate metagenomes from a user-defined community or random genomes from the NCBI. As such, InSilicoSeq has already been used in the development of several new metagenomics software (Georgiou *et al.*, 2019; Valdes, Stebliankin and Narasimhan, 2019; Kalantar *et al.*, 2020; Mallawaarachchi, Wickramarachchi and Lin, 2020)

4.1.1 Speed

InSilicoSeq is multi-threaded and can generate half a million reads in less than 10 minutes (Figure 13).

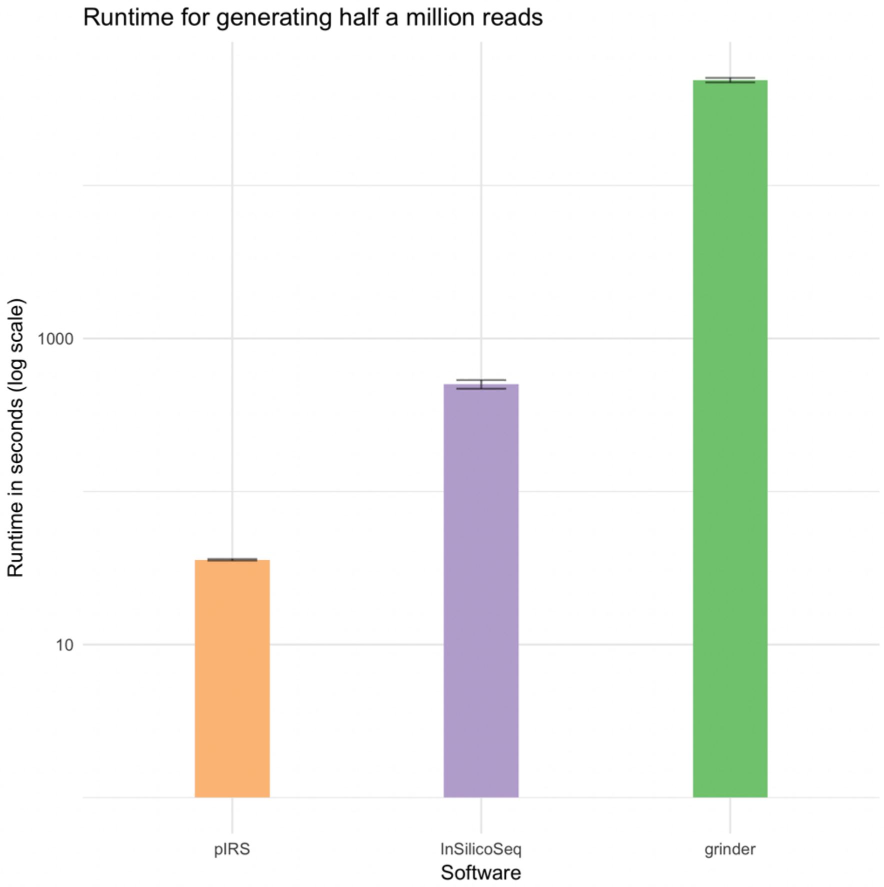


Figure 13: Speed of read simulation software

4.1.2 Accuracy

InSilicoSeq produces realistic Illumina data. Of the other software tested, only ART (Huang *et al.*, 2012) models as closely the per-base quality (Figure 14). However, even ART fails to properly model the mean sequence quality distribution, while InSilicoSeq at least produces low-quality sequences occasionally (Figure 15).

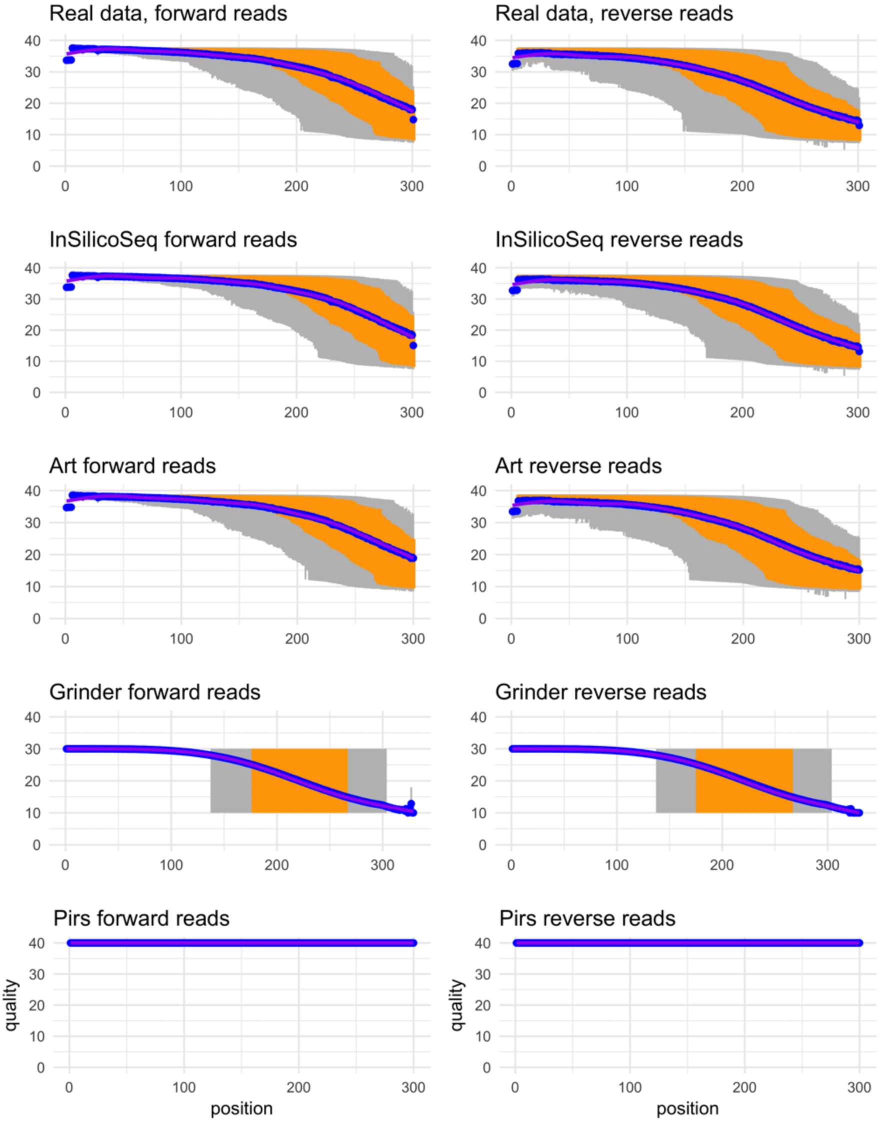


Figure 14: Base-level accuracy for real data and selected simulation software

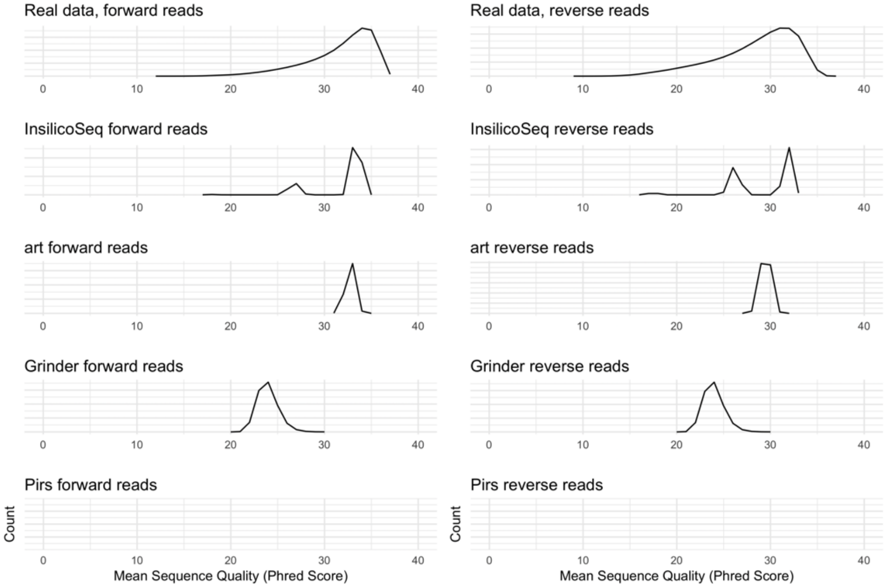


Figure 15: Read mean qualities for real data and selected simulation software

4.1.3 Current and future benchmarks

One of the primary goals when developing InSilicoSeq was to benchmark taxonomic classifier and binning algorithms against each other, in order to choose the most suited tools for analysing our data. Performing this benchmark has been problematic, mainly due to lack of computing power and storage. Our benchmark framework for taxonomy classification tools, available at https://github.com/HadrienG/2019_classifiers_benchmark, was not able to run on our hardware in the desired timeframe. The software architecture is nevertheless there, and it would be desirable to complete that benchmark in the future. Concerning the binning software used in paper III, we chose metabat2 in accordance with the initial CAMI results, the most recent metabat2 paper (Sczyrba *et al.*, 2017; Kang *et al.*, 2019), as well as preliminary internal results on smaller datasets (https://github.com/SGBC/metagenome-binning).

4.2 A data-driven Review of the Coral microbiome (Paper II)

In this paper, we reanalysed 405 samples from 13 different 16S metabarcoding studies (Table 3) and attempted to define a minimal core microbiome for corals. This putative core microbiome is composed of 24 bacteria and largely dominated by *Endozoicomonas* (Table 4). *Endozoicomonas* has previously been hypothesized to be the dominating member of the coral microbiome (Till Bayer *et al.*, 2013; Pogoreutz *et al.*, 2018), although their function is currently unclear.

Table 4: The putative coral core microbiome

Genus	Mean	Log Fold Change	p value
Endozoicomonas	1,731.70	1.1586	0.0731
Acinetobacter	34.51	-1.3349	0.0136
Rubritalea	33.03	-1.1755	0.0327
Acholeplasma	28.18	-1.3279	0.0111
SUP05_cluster	20.60	-0.8288	0.0842
Parahaliea	17.96	-1.2033	0.0355
Ascidiaceihabitans	16.68	-0.7208	0.0924
Delftia	14.83	-1.3969	0.0032
Undibacterium	12.39	-1.4143	0.0026
Sphingobium	11.73	-1.1520	0.0166
Methylobacterium	10.83	-1.4906	0.0018
Fulvivirga	10.21	-1.0302	0.0520
Streptococcus	8.95	-1.2378	0.0090
Corynebacterium_1	8.58	-1.2939	0.0086
Coxiella	7.67	-1.4753	0.0074
Turneriella	7.62	-1.1979	0.0210
Aquabacterium	7.10	-1.3663	0.0037
Coraliomargarita	6.99	0.7788	0.0866
Micrococcus	6.92	-1.0060	0.0315
OM43_clade	6.51	-1.2169	0.0163
Lawsonella	6.26	-1.1695	0.0146
Chryseobacterium	6.22	-0.9954	0.0451
Aureispira	5.87	-1.0024	0.0483
Luminiphilus	5.33	-0.7981	0.0874

Due to the limitation of 16S analyses, we of course only have resolution to the genus level. A logical next step is to conduct whole metagenome sequencing experiments on the coral reef microbiome, in order to identify the bacterial species being part of the coral core microbiome, and eventually decipher their function. 16S studies can be valuable for comparing microbiome composition, but insights on the functionality of the microbiome is very limited. Still, knowing the composition of a healthy coral microbiome is a great interest; some studies have notably tried to transplant a healthy microbiome cocktail into diseases corals (Damjanovic *et al.*, 2019).

It is noteworthy to point out that the results of our analysis consistently found fewer OTUs than were previously reported. We accredit these differences to the dada2 algorithm as explained in the methods. Systematic errors present in the sequencing data were mostly corrected, leading to less spurious variants than with classical OTU clustering.

Lastly, the diversity in bleached or lesioned samples was much less important than in the healthy mucus or tissue samples. It is consistent with previous findings in other species, where a diseased state causes a microbiome imbalance (Lozupone *et al.*, 2012).

4.3 Molecular Ecology of Coral Reef Microorganisms in the Western Indian Ocean coast of Kenya (Paper III)

This study provided with the first metagenomes from the coast on the West Indian Ocean. The goal of this paper was to (i) get a baseline community from the microbes surrounding the reef and (ii) hypothesise what metabolism these communities eventually bring to their environment.

4.3.1 Rationale

In paper II we discussed the limitations of 16S sequencing and discussed how important it was to pivot to WMS for the coral microbiome. We chose however to focus here on the surrounding environment of the corals, instead of on the coral themselves. We will come back to that in the conclusion of this thesis but designing and carrying a WMS experiment on the coral microbiome has and will be proven challenging. While we developed and refined our coral metagenomics and metatranscriptomics protocols, we decided to sequence the coral environment: water and upper sediment layer, in parallel. The surrounding environment of coral reef remains interesting, as the different communities are expected to interact with each other and exchange nutrient, as well as being affected by the same environmental stressors; this makes the surrounding microbiomes of coral reefs promising markers against pollution and other environmental challenges.

4.3.2 Genome binning

A total of 782 genome bins were recovered from the assemblies, of which 197 had > 50% completeness and < 25% contamination (Figure 16). Of those, 94 bins were > 75% completeness and 26 > 90%.

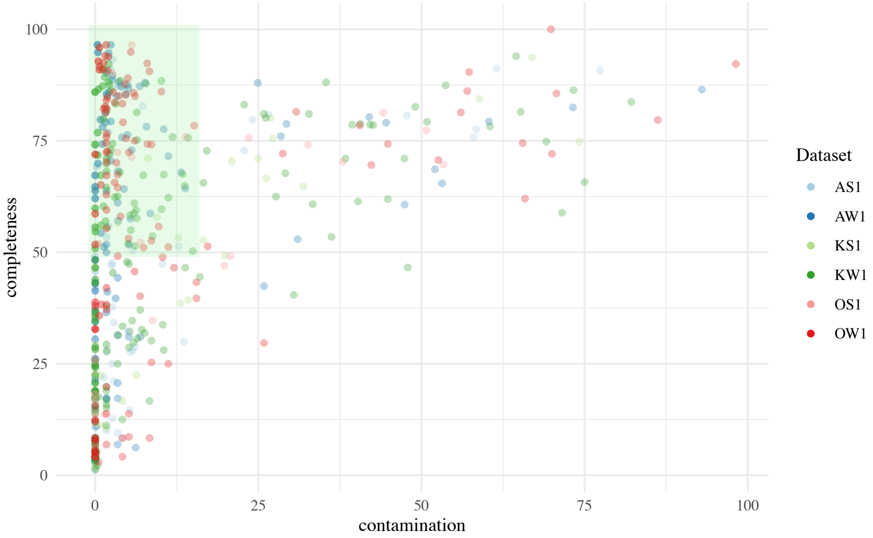


Figure 16: completeness and contamination of the genome bins

Of the 197 bins with > 50% completeness, 166 came from the water samples. We hypothesised that the sediment samples were harder to assemble, possibly due to abnormal GC content or more repeat-rich genome, or uneven coverage.

Additionally, the majority of assembled genomes were *Proteobacteria*, followed by *Bacteroidetes* (Figure 17). The majority of *Bacteroidetes* exhibited potential vancomycin resistance, which may have come from the use of avoparcin in agriculture (Bager *et al.*, 1997; Nilsson, 2012). Avoparcin is an antibiotic very similar to vancomycin, which is not permitted to use in the European Union for fear of antibiotic resistance but is in use in Kenya.

4.3.3 Protein Assembly

Protein assembly is a new and promising method to get more functional insights from a metagenome than from classical metagenome assembly. Protein assembly on average recovers more putative proteins than "classical" metagenome binning (Steinegger, Mirdita and Söding, 2018). As with metagenome assembly, it is however challenging to know if those proteins are expressed by the microbial cells: we more often than not talk about metabolic potential in such experimental setup.

As with the rest of the dataset, the metabolic potential of the water and sediment communities was very diverse. We found 424 different pathways in the datasets. Amongst the most abundant pathways were notably biosynthesis of antibiotics, biosynthesis of secondary metabolism, carbon fixation and metabolism, and quorum sensing. These pathways indicate that surrounding communities may provide with important nutrient exchange within the reef ecosystem, as well as being an important first line of defence against pathogens.

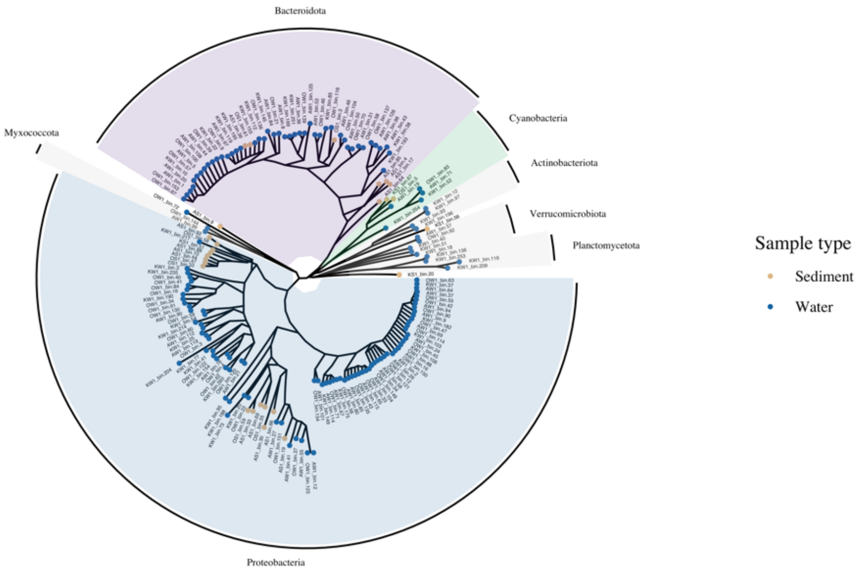


Figure 17: phylogenetic tree of the 176 bacterial genome bins

5 Conclusion and Future Perspectives

The objectives of this thesis were two-fold. Firstly, we wanted to improve methods used for metagenomics. With Paper I, we provided the bioinformatics community with a simple, yet powerful way to simulate metagenomic sequencing data. As stated previously, our simulator was already used for developing several new bioinformatics tools.

Several methodological improvements can still be made, which would benefit the whole field. Metagenome binning is not a solved problem yet, and some Machine Learning (ML) algorithms, such as DBSCAN or affinity propagation have shown promise. At the moment of this writing, there are still benchmarking needs to be filled. Only one large scale independent benchmarking of metagenomics methods has been performed so far, the CAMI challenge. For reproducibility and replicability's sake, it would be valuable to present and publish one more benchmarking framework.

Secondly, we wanted to investigate the bacterial communities living in close proximity and within coral reefs. In Paper II we re-analysed public datasets to try to decipher the members of the coral microbiome. In paper III, we take a more modern approach, WMS, to sequences the microbial communities living in the water and upper sediment layer of coral reefs in Kenya.

In paper II, we propose a putative core microbiome for tropical coral reefs. Largely dominated by *Endozoicomonas*, its function remains to be understood. In paper III, we presented 176 novel bacterial genomes, as well as the first metagenomes from the coast of the West Indian Ocean. Additionally, the pathway analysis revealed potential exchanges between microbial communities.

Our work on The Kenyan metagenomes provides with an important baseline for future reef studies and was the first metagenomics study of the regions' reefs. In the future it would be desirable to study the coral microbiome itself, both from the surface mucus layer and the coral tissue. We have already collected DNA and RNA samples from the surface mucus layers of twenty-five *Acropora spp.* coral colonies, at the same three samples sites described in paper III. The samples were sequenced at SciLifeLab, and the data analysis is ongoing. Given the large amount of data, getting preliminary results may well take twelve months or so.

We foresee that this analysis will provide the valuable functional insights into the coral microbiome that were missing from previous 16S metabarcoding studies, and that in these methods lie the future of coral reef microbiology. We now have a pretty broad idea of the taxonomy of the coral microbiome and have putative functions for bacteria living alongside in the reef. The logical next step is to look at the function of the coral microbiome.

There are a lot of considerations and hurdles to overcome when planning and conducting a metatranscriptomics experiment. Firstly, sampling sites tend to be far from laboratories, and, RNA degrading quickly, it is important to be able to transport it quickly and safely to cold storage. Dry ice and RNA later are a must but are expensive and hard to get in some countries. Secondly, and one of the reasons coral microbiome have almost extensively been studied using metabarcoding, is that removal of the host genetic material is difficult. Not only one must be sure that samples are not dominated by the studied coral but neither by the genetic material of Symbiodinium. Thirdly, metatranscriptomics has poorly established workflows and is a very recent methodology. Few studies have seen light, and while they are thought to perform decently, the RNA-Seq methods have not been adapted to the hierarchical nature of the variables in metatranscriptomes. Indeed, the gene expression levels in metatranscriptomics experiments can be explained by (i) high gene expression and (ii) high abundance of the bacteria whose gene it is. There is a need for adapting the linear modelling used in RNA-Seq to analyse both metagenomics and metatranscriptomics data. Promising leads are joint modelling or hierarchical testing.

In paper III, we published a large bacterial catalogue, as well as a protein catalogue, which may be used to design more targeted monitoring experiments. While physicochemical properties of coastal waters have remained stable even in polluted reefs, we believe that bacteria may be good indicators of coral reef health, and our findings may help designing primer-based approaches for continuous, cheap coral reef monitoring.

Overall, this thesis helped to lay foundations for coral reef work in West Africa and elsewhere. Deciphering the functions and metabolism of the coral microbiome is within reach, and the microbiome may play an important role in coral health, and coral sustainability if we can, in a medium-term perspective, monitor more efficiently coral health by monitoring surrounding bacteria, and perhaps even reimplant healthy microbiomes in damaged ecosystems. Then, if we reduce drastically our CO_2 emissions within the next decades, coral may have a shot at survival.

References

Anderson, K. D. *et al.* (2017) “Variation in growth rates of branching corals along Australia’s Great Barrier Reef,” *Scientific reports*, 7(1), p. 2920. doi: 10.1038/s41598-017-03085-1.

Bager, F. *et al.* (1997) “Avoparcin used as a growth promoter is associated with the occurrence of vancomycin-resistant Enterococcus faecium on Danish poultry and pig farms,” *Preventive veterinary medicine*, 31(1–2), pp. 95–112. doi: 10.1016/s0167-5877(96)01119-1.

Bayer, T. *et al.* (2013) “Bacteria of the genus Endozoicomonas dominate the microbiome of the Mediterranean gorgonian coral Eunicella cavolini,” *Marine ecology progress series*, 479, pp. 75–84. doi: 10.3354/meps10197.

Bayer, Till *et al.* (2013) “The microbiome of the Red Sea coral Stylophora pistillata is dominated by tissue-associated Endozoicomonas bacteria,” *Applied and environmental microbiology*, 79(15), pp. 4759–4762. doi: 10.1128/AEM.00695-13.

Blackall, L. L., Wilson, B. and van Oppen, M. J. H. (2015) “Coral-the world’s most diverse symbiotic ecosystem,” *Molecular ecology*, 24(21), pp. 5330–5347. doi: 10.1111/mec.13400.

Bourne, D. G., Morrow, K. M. and Webster, N. S. (2016) “Insights into the Coral Microbiome: Underpinning the Health and Resilience of Reef Ecosystems,” *Annual review of microbiology*, 70, pp. 317–340. doi: 10.1146/annurev-micro-102215-095440.

Callahan, B. J. *et al.* (2016) “DADA2: High-resolution sample inference from Illumina amplicon data,” *Nature methods*, 13(7), pp. 581–583. doi: 10.1038/nmeth.3869.

Case, R. J. *et al.* (2007) “Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies,” *Applied and environmental microbiology*, 73(1), pp. 278–288. doi: 10.1128/AEM.01177-06.

Chaumeil, P.-A. *et al.* (2019) “GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database,” *Bioinformatics* . doi: 10.1093/bioinformatics/btz848.

Costanza, R. *et al.* (1997) “The value of the world’s ecosystem services and natural capital,” *Nature*, 387(6630), pp. 253–260. doi: 10.1038/387253a0.

Damjanovic, K. *et al.* (2019) “Experimental Inoculation of Coral Recruits With Marine Bacteria Indicates Scope for Microbiome Manipulation in Acropora tenuis and Platygyra daedalea,” *Frontiers in microbiology*, 10, p. 1702. doi: 10.3389/fmicb.2019.01702.

Edgar, R. C. (2018) “Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences,” *PeerJ*, 6, p. e4652. doi: 10.7717/peerj.4652.

Escalona, M., Rocha, S. and Posada, D. (2016) “A comparison of tools for the simulation of genomic next-generation sequencing data,” *Nature reviews. Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 17, p. 459. doi: 10.1038/nrg.2016.57.

Georgiou, A. *et al.* (2019) “META$^\mathbf2$: Memory-efficient taxonomic classification and abundance estimation for metagenomics with deep learning,” *arXiv [q-bio.GN]*. Available at: http://arxiv.org/abs/1909.13146.

Glasl, B., Herndl, G. J. and Frade, P. R. (2016) “The microbiome of coral surface mucus has a key role in mediating holobiont health and survival upon disturbance,” *The ISME journal*, 10(9), pp. 2280–2292. doi: 10.1038/ismej.2016.9.

Graham, N. A. J. *et al.* (2006) “Dynamic fragility of oceanic coral reef ecosystems,” *Proceedings of the National Academy of Sciences of the United States of America*, 103(22), pp. 8425–8429. doi: 10.1073/pnas.0600693103.

Hadaidi, G. *et al.* (2017) “Stable mucus-associated bacterial communities in bleached and healthy corals of Porites lobata from the Arabian Seas,” *Scientific reports*, 7, p. 45362. doi: 10.1038/srep45362.

Hernandez-Agreda, A., Gates, R. D. and Ainsworth, T. D. (2017) “Defining the Core Microbiome in Corals’ Microbial Soup,” *Trends in microbiology*. Elsevier, 25(2), pp. 125–140. doi: 10.1016/j.tim.2016.11.003.

Huang, W. *et al.* (2012) “ART: a next-generation sequencing read simulator,” *Bioinformatics* , 28(4), pp. 593–594. doi: 10.1093/bioinformatics/btr708.

Huerta-Cepas, J. *et al.* (2017) “Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper,” *Molecular biology and evolution*, 34(8), pp. 2115–2122. doi: 10.1093/molbev/msx148.

Hughes, T. P., Barnes, M. L., *et al.* (2017) “Coral reefs in the Anthropocene,” *Nature*, 546(7656), pp. 82–90. doi: 10.1038/nature22901.

Hughes, T. P., Kerry, J. T., *et al.* (2017) “Global warming and recurrent mass bleaching of corals,” *Nature*, 543(7645), pp. 373–377. doi: 10.1038/nature21707.

Kalantar, K. L. *et al.* (2020) “IDseq - An Open Source Cloud-based Pipeline and Analysis Service for Metagenomic Pathogen Detection and Monitoring,” *bioRxiv*. doi: 10.1101/2020.04.07.030551.

Kallmeyer, J. *et al.* (2012) “Global distribution of microbial abundance and biomass in subseafloor sediment,” *Proceedings of the National Academy of Sciences of the United States of America*, 109(40), pp. 16213–16216. doi: 10.1073/pnas.1203849109.

Kang, D. D. *et al.* (2019) “MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies,” *PeerJ*, 7, p. e7359. doi: 10.7717/peerj.7359.

Kellogg, C. A., Ross, S. W. and Brooke, S. D. (2016) “Bacterial community diversity of the deep-sea octocoral Paramuricea placomus,” *PeerJ*, 4, p. e2529. doi: 10.7717/peerj.2529.

Langmead, B. and Salzberg, S. L. (2012) “Fast gapped-read alignment with Bowtie 2,” *Nature methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Lawler, S. N. *et al.* (2016) “Coral-Associated Bacterial Diversity Is Conserved across Two Deep-Sea Anthothela Species,” *Frontiers in microbiology*, 7, p. 458. doi: 10.3389/fmicb.2016.00458.

Li, D. *et al.* (2015) “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph,” *Bioinformatics* , 31(10), pp. 1674–1676. doi: 10.1093/bioinformatics/btv033.

Love, M. I., Huber, W. and Anders, S. (2014) “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lozupone, C. A. *et al.* (2012) “Diversity, stability and resilience of the human gut microbiota,” *Nature*, 489(7415), pp. 220–230. doi: 10.1038/nature11550.

Ma, X. *et al.* (2019) “Analysis of error profiles in deep next-generation sequencing data,” *Genome biology*, 20(1), p. 50. doi: 10.1186/s13059-019-1659-6.

Macmanes, M. D. (2014) “On the optimal trimming of high-throughput mRNA sequence data,” *Frontiers in genetics*, 5, p. 13. doi: 10.3389/fgene.2014.00013.

Mallawaarachchi, V., Wickramarachchi, A. and Lin, Y. (2020) “GraphBin: Refined binning of metagenomic contigs using assembly graphs,” *Bioinformatics* . doi: 10.1093/bioinformatics/btaa180.

Meyer, J. L., Paul, V. J. and Teplitski, M. (2014) “Community shifts in the surface microbiomes of the coral Porites astreoides with unusual lesions,” *PloS one*, 9(6), p. e100316. doi: 10.1371/journal.pone.0100316.

Mostajir, B. *et al.* (2015) “Microbial Food Webs in Aquatic and Terrestrial Ecosystems,” in Bertrand, J.-C. et al. (eds.) *Environmental Microbiology: Fundamentals and Applications: Microbial Ecology*. Dordrecht: Springer Netherlands, pp. 485–509. doi: 10.1007/978-94-017-9118-2_13.

Nguyen, N.-P. *et al.* (2016) “A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity,” *NPJ biofilms and microbiomes*, 2, p. 16004. doi: 10.1038/npjbiofilms.2016.4.

Nilsson, O. (2012) “Vancomycin resistant enterococci in farm animals - occurrence and importance,” *Infection ecology & epidemiology*, 2. doi: 10.3402/iee.v2i0.16959.

Noble, P. A., Citek, R. W. and Ogunseitan, O. A. (1998) “Tetranucleotide frequencies in microbial genomes,” *Electrophoresis*, 19(4), pp. 528–535. doi: 10.1002/elps.1150190412.

Norman, J. M., Handley, S. A. and Virgin, H. W. (2014) “Kingdom-agnostic metagenomics and the importance of complete characterization of enteric microbial communities,” *Gastroenterology*, 146(6), pp. 1459–1469. doi: 10.1053/j.gastro.2014.02.001.

Park, S. T. and Kim, J. (2016) “Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing,” *International neurourology journal*, 20(Suppl 2), pp. S76-83. doi: 10.5213/inj.1632742.371.

Parks, D. H. *et al.* (2015) “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes,” *Genome research*, 25(7), pp. 1043–1055. doi: 10.1101/gr.186072.114.

Parks, D. H. *et al.* (2017) “Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life,” *Nature microbiology*, 2(11), pp. 1533–1542. doi: 10.1038/s41564-017-0012-7.

Plaisance, L. *et al.* (2011) “The diversity of coral reefs: what are we missing?,” *PloS one*, 6(10), p. e25026. doi: 10.1371/journal.pone.0025026.

Pogoreutz, C. *et al.* (2018) “Dominance of Endozoicomonas bacteria throughout coral bleaching and mortality suggests structural inflexibility of the Pocillopora verrucosa microbiome,” *Ecology and evolution*, 8(4), pp. 2240–2252. doi: 10.1002/ece3.3830.

Reuter, J. A., Spacek, D. V. and Snyder, M. P. (2015) “High-throughput sequencing technologies,” *Molecular cell*, 58(4), pp. 586–597. doi: 10.1016/j.molcel.2015.05.004.

Robertson, V. *et al.* (2016) “Highly Variable Bacterial Communities Associated with the Octocoral Antillogorgia elisabethae,” *Microorganisms*, 4(3). doi: 10.3390/microorganisms4030023.

Roder, C. *et al.* (2015) “Microbiome structure of the fungid coral Ctenactis echinata aligns with environmental differences,” *Molecular ecology*, 24(13), pp. 3501–3511. doi: 10.1111/mec.13251.

Rodriguez-R, L. M. and Konstantinidis, K. T. (2014) “Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets,” *Bioinformatics* , 30(5), pp. 629–635. doi: 10.1093/bioinformatics/btt584.

Schirmer, M. *et al.* (2016) “Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data,” *BMC bioinformatics*, 17, p. 125. doi: 10.1186/s12859-016-0976-y.

Sczyrba, A. *et al.* (2017) "Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software," *Nature methods*, 14(11), pp. 1063–1071. doi: 10.1038/nmeth.4458.

Sheather, S. J. and Jones, M. C. (1991) "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society. Series B, Statistical methodology*. [Royal Statistical Society, Wiley], 53(3), pp. 683–690. Available at: http://www.jstor.org/stable/2345597.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. CRC Press. Available at: https://market.android.com/details?id=book-e-xsrjsL7WkC.

Sorokin, Y. I. (1973) "Trophical Role of Bacteria in the Ecosystem of the Coral Reef," *Nature*, 242(5397), pp. 415–417. doi: 10.1038/242415a0.

Spalding, M. D. and Grenfell, A. M. (1997) "New estimates of global and regional coral reef areas," *Coral reefs* , 16(4), pp. 225–230. doi: 10.1007/s003380050078.

Steinegger, M., Mirdita, M. and Söding, J. (2018) "Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold," *bioRxiv*. doi: 10.1101/386110.

Tout, J. *et al.* (2014) "Variability in microbial community composition and function between different niches within a coral reef," *Microbial ecology*, 67(3), pp. 540–552. doi: 10.1007/s00248-013-0362-5.

Valdes, C., Stebliankin, V. and Narasimhan, G. (2019) "Large scale microbiome profiling in the cloud," *Bioinformatics* , 35(14), pp. i13–i22. doi: 10.1093/bioinformatics/btz356.

Vezzulli, L. *et al.* (2013) "16SrDNA Pyrosequencing of the Mediterranean Gorgonian Paramuricea clavata Reveals a Link among Alterations in Bacterial Holobiont Members, Anthropogenic Influence and Disease Outbreaks," *PloS one*, 8(6), p. e67745. doi: 10.1371/journal.pone.0067745.

van de Water, J. A. J. M. *et al.* (2016) "Spirochaetes dominate the microbial community associated with the red coral Corallium rubrum on a broad geographic scale," *Scientific reports*, 6, p. 27277. doi: 10.1038/srep27277.

van de Water, J. A. J. M. *et al.* (2017) "Comparative Assessment of Mediterranean Gorgonian-Associated Microbial Communities Reveals Conserved Core and Locally Variant Bacteria," *Microbial ecology*, 73(2), pp. 466–478. doi: 10.1007/s00248-016-0858-x.

Wood, D. E., Lu, J. and Langmead, B. (2019) "Improved metagenomic analysis with Kraken 2," *Genome biology*. Springer, 20(1), p. 257. doi: 10.1186/s13059-019-1891-0.

Popular science summary

Imagine a sandy beach, turquoise water and a myriad of fish swimming around colourful stony structures. These structures – called reefs – are formed very slowly by small organisms called corals. Corals harbour many different species of fish, sea urchins, algae, and more. They are very important economically, bringing tourism revenue, protecting coast from extreme weather events and many people worldwide depend on them for food. Corals live a symbiotic relationship with an alga, *Symbiodinium*, that provides energy to them.

Corals are surprisingly fragile and sensitive to environmental changes. Their leading cause of death is called bleaching. During a bleaching event, corals eject *Symbiodinium* and lose their colour. Bleaching mostly happens with sudden rise of temperature but can also be caused by more salinity or acidity in the water. If the coral is in a prolonged state of bleaching, it dies. Due to global warming, bleaching, as well as other diseases, happens more frequently, and corals are menaced of extinction if humanity does not reduce their CO_2 emissions.

Corals, like every other organism on earth, is surrounded by microorganisms, such as bacteria and viruses. Microorganisms are found inside all living things, such as humans, and corals! In this thesis, we hypothesise that bacteria play an important role in the health of coral reefs. Firstly we focus on methodological improvements, since the study of microbes' genomes from the environment is a relatively new science. Secondly, we try to identify which microbes are found in all coral reefs around the world. Lastly, we focus on function – what the microbes do, not just what they are – of the bacteria surrounding reefs on the Kenyan coast of the west Indian ocean.

We find that the bacterial communities living on the reef are very diverse, and that they probably participate by exchanging nutrients (food) with their surroundings. Additionally, we find evidence of antibiotic resistance in the water communities, which may partly be caused by the use of antibiotics in agriculture. We also find evidence that some bacteria on the reef also produce antibiotics, which could be a sign that bacterial communities play a role in protecting coral reefs from microorganisms that cause diseases.

To conclude, this work provides with a first good idea of which bacteria can be found within corals, and what bacteria surrounding the reef are doing. In the future, we would want to monitor the bacterial communities we have studied. We think they are more sensitive than corals to environmental changes, and the monitoring could predict health issues in the reef. Additionally, we would like to investigate further the function of the bacteria that live inside corals.

Acknowledgements

I would like to start acknowledging the funding agencies, without which this work would not have been possible. This work has been funded by the Swedish Research Council, grant number 2015-03443_VR, as well as by a SciLifeLab biodiversity grant. I'd like to further thank SciLifeLab, without which sequencing would not have been possible. Similarly, large computational parts of this projects would not have been possible without the support of the Google Cloud Platform that granted us research credits. I would also like to thank the DEANN project, for awarding me a travel grant for a study visit to the United States.

Erik, without you I simply would not have pursued a doctoral education and would not have the life that I have now. You welcomed me in the group as an undergraduate student, and never let me go. You gave me a world of opportunities, and the independence I needed to grow as a scientist and a person. **Mikael**, I'm sorry that I did not focus more on the viruses, but I always felt that you were there, ready to help if I needed it, thank you. This work would not be the same without the incredible, kind, overly enthusiastic, **Oskar Lindsjö.** You kept my preposterous ideas in check. Or added your own, nobody is really sure. You supported my febrile hands in the lab. You came to me with the most Oskar bioinformatics problems imaginable. You helped me when I needed the most.

My Swedish journey so far would not be my Swedish journey without a bit of French in it. From my undergraduate up 'til now, you became my friend, then my friend and supervisor. **Juliette**, thanks for everything! Thank you for the barbecues, the laughs, the sing-along Spotify sessions, all the conversations we had in a mangled, awful language mix of French and English and Swedish. We both can't spell in our native language anymore, but it'll be fine.

Santie and **Sammy**, I don't even know how to thank you. Your efforts with the sampling were superhuman. I really hope our paths will cross again, and that you started something good with marine research there in Kilifi. I'm glad I've gotten to be a part of it.

Doing research in bioinformatics can be lonely. Not in Sweden! **NBIS** and their Uppsala beer club, the amazing **SciLifeLab** nf-core team in Stockholm, **SLUBI**, as well as the **MedBioinfo** research school, I've been spoilt with networking opportunities, bioinformatics research discussion, and a whole regiment of like-minded people, to share my research interests with. Thanks to all the people that came and went into our research section, and I wish you the best wherever you are. **The carpentries** community, you hold a special place in my heart. You are composed of so many incredible scientists and educators, and made a better teacher and a better person. From ANGUS 2017 in Davis, you have not stopped to amaze me!

Jamie, **Dyanna** and the rest of the UF astrobiology lab in Cape Canaveral, thank you from welcoming me for my short trip to Florida. I immediately felt like a part of the group, however how short I was meant to stay. I'm looking forward to finishing our unfinished collaboration now that this thesis is almost finished.

I'd like to thank the current and former PhD students at HGEN, VH and SLU in general. I'll miss you folks! **Agnese**, **Chrissie** and **Suvi**, thanks for the skiing (or not, **Suvi)**, the laughs, and the sharing of our occasional desperation in front of the adversity of research. A big thank you to the virology gang, notably **Emeli, Hedvig, Pontus** and **Hari** amongst others.

There was a time where I wasn't the most studious person, but teachers stood by me and gently or less gently pushed me in the right direction. **David, Aline** and **Françoise**, thank you for giving me a well-rounded education and believing in me when I didn't believe in myself.

Mathieu, and the whole 4 rue des Échelles, thank you for all the fun we had, I miss you folks! **Mattias,** thank you for being the night butterfly that you were when we met, and introduced me to so many friends. You've shaped how my study years were and my life in Sweden currently is. **Patrick**, thanks for all the beers.

I'd like to thank my parents for supporting me when I left home, and for still being supportive of all the choices I make in life. A big thank you as well to **Marie** and **Tore**, for becoming my second family here in Sweden, you've been nothing but welcoming.

Marielle, thanks you for being there for me, for waking up with me, for laughing with me every day, for not closing the cabinet doors, for arguing with me, for loving me and for everything else. Thank you for being you.

Lastly, I'd like to finish by thanking my closest collaborator. You have been supporting with me during the writing of this thesis, forcing me to take breaks, and to play with you so I would not overwork myself, forcing me to go outside and pick up your poop and trying to prevent you to chase birds and rabbits and hedgehogs. **Echo**, you're a mighty hunter, a ferocious beast and you know it.