DEPARTMENT OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES

Microbial Biofilms in the Bioinformatics Era

Application of High-Throughput DNA Sequencing Technologies in the Metagenomic Study of Marine Biofilms

Kemal Sanli
Department of Biological and Environmental Sciences
Faculty of Science

This thesis will be defended in public on Friday, the 17th of June, 2016, at 10:00, in the lecture hall (Hörsalen) at the Department of Biological and Environmental Sciences, Carl Skottsbergs gata 22B, Gothenburg.

Faculty opponent: Prof. Daniel Huson, Dept. of Algorithms in Bioinformatics, Tübingen University, Germany

Examiner: Prof. Adrian K. Clarke, Dept. of Biological and Environmental Sciences, University of Gothenburg, Sweden

ISBN: 978-91-85529-91-9 (PRINT)
ISBN: 978-91-85529-92-6 (PDF)
http://hdl.handle.net/2077/42424

ABSTRACT

Adverse effects of anthropogenic impact on the environment have become conspicuous in the past century and among others include the gradual increase in the global CO_2 levels, the contamination of air, soil and water by toxic chemicals, and the emergence of antimicrobial resistance among pathogenic microbial species. Microorganisms partake in an extreme diversity of activities in the environment, and hence, constitute the prime candidates to be investigated in understanding of the progression and effects of the aforementioned environmental hazard scenarios. The spectacular rise of massively parallel sequencing (next generation sequencing, NGS) technologies in mid 2000s initiated a renaissance in microbial ecology by allowing the *in situ* investigation of environmental samples at metagenome level, largely eliminating prior laboratory culturing steps. Metagenomics has thereby been established as a new interdisciplinary field and methodology, harmonizing the accumulated knowledge in microbial ecology and genetics with the high-throughput environmental DNA sequence data through the means of bioinformatics analysis resources.

One of the emerging application areas that require a comprehensive microbial investigation is the study of the effects of toxic chemicals on biota in the environment, namely ecotoxicology. In this PhD thesis, bioinformatics software development and microbial ecological data analysis projects are integrated within the field of ecotoxicology. The objective of the thesis is to implement metagenomics as a robust tool in the field of ecotoxicology to gain both community and molecular level insights. Paper I presents FANTOM (Functional and Taxonomic Analysis of Metagenomes), a graphical user interface (GUI)-based metagenomic data analysis tool that provides various statistical analysis and visualization features for biologists with limited bioinformatics experience. PACFM (Pathway Analysis with Circos for Functional Metagenomics), another GUI-based software tool, is presented in Paper II, and it provides researchers in metagenomics with a novel plot and various biochemical pathway analysis features. Paper III is an exploratory study of the marine biofilms (also known as periphython), constituting the first study to sequence the total genomic DNA content of these microbial communities that inhabit the aquatic environment. The metagenomic analysis of the marine biofilms revealed that *Proteobacteria*, *Bacteroidetes* and *Cyanobacteria* are the most abundant organisms in these biofilm communities. In addition, the functional repertoire within the metagenome involved signatures of anaerobic processes including denitrification and methanogenesis, which suggests the presence of low-oxygen zones within the micro-ecosystem formed by the marine biofilms. Paper III also constituted the pilot study for Paper IV, where an experimental design was set up to investigate the toxic effects of the broad spectrum antimicrobial agent, triclosan, on the marine biofilms. High and low levels of triclosan exposure was shown to cause significant changes in the community structure and the functioning of the marine biofilms. A sulfur-based microbial consortium together with several algal groups were hypothesized to partake in the detoxification of triclosan. Hence, metagenomics is shown to be a powerful research tool in the field of ecotoxicology.

This PhD thesis presents novel software tools and applications in the field of metagenomics, combining a wide range of paradigms from several disciplines within a unified solution framework as an attempt to practice and transcend interdisciplinary research.

Keywords: metagenomics, bioinformatics software, microbial biofilms, Next Generation Sequencing, pathway analysis, periphyton, marine biofilms, FANTOM, PACFM

LIST OF PAPERS

This thesis is based on the following papers, which are referred to by their Roman numerals in the text.

I. Sanli, K.*, Karlsson, F.*, Nookaew, I., and Nielsen, J. (2013). FANTOM: Functional and taxonomic analysis of metagenomes. BMC Bioinformatics, 14(1), 38. doi: 0.1186/14711-2105-14-38
Distributed under the Creative Commons Attribution License

II. Sanli, K., Sinclair, L., Nilsson, R. H., Mardinoglu, A., and Eiler, A. (2016). PACFM: Pathway Analysis with Circos in Functional Metagenomics. *Manuscript*.

III. Sanli, K., Bengtsson-Palme, J., Nilsson, R. H., Kristiansson, E., Alm-Rosenblad, M., Blanck, H., and Eriksson, K. M. (2015). Metagenomic sequencing of marine periphyton: taxonomic and functional insights into biofilm communities. Frontiers in Microbiology, 6: 1192. doi: 10.3389/fmicb.2015.01192
Distributed under the Creative Commons Attribution License

IV. Sanli, K., Sinclair, L., Corcoll, N., Nilsson, R. H., Johansson, C. H., Backhaus, T., and Eiler, A. (2016). Triclosan induced community shifts point toward sulfur-based detoxification mechanisms in marine biofilms.
Manuscript.

Papers not included in this PhD thesis are as follows:

V. Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P., Sánchez-García, M., Ebersberger, I., de Sousa, F., Amend, A., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y. J. K., Sanli, K., Eriksson, K. M., Vik, U., Veldre, V., and Nilsson, R. H. (2013), Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods in Ecology and Evolution, 4: 914–919. doi: 10.1111/2041-210X.12073

VI. Eriksson, K. M., Johansson, C. H., Fihlman, V., Grehn, A., Sanli, K., Andersson, M. X., Blanck, H., Arrhenius, Å., Sircar, T., and Backhaus, T. (2015), Long-term effects of the antibacterial agent triclosan on marine periphyton communities. Environmental Toxicology and Chemistry, 34: 2067–2077. doi:10.1002/etc.3030

* Equal contribution

AUTHOR CONTRIBUTIONS

I. Designed, implemented and deployed the software tool FANTOM. Analyzed the metagenomics case study data using the software tool and generated results for the paper. Contributed to drafting of the manuscript and to the revision process with the other co-authors.

II. Designed, implemented and deployed the software tool PACFM. Contributed to the metagenomics case studies with the corresponding authors. Drafted the manuscript and contributed to the revision process with the other co-authors. Coordinated the study.

III. Implemented the metagenomic data processing and analysis pipeline used in the paper. Analyzed the metagenomic data and interpreted the results. Drafted the manuscript and contributed to the revision process with the other co-authors.

IV. Participated in the experiment, sampling and DNA extraction. Analysed the data, interpreted the results, and reported the major findings. Contributed to drafting of the manuscript and to the revision process with the other co-authors.

LIST OF ABBREVIATIONS AND SYMBOLS

AsO_4^{3-}: Arsenate
ATP: Adenosine triphosphate
BLAST: Basic Local Alignment Search Tool
bp: Base pairs
BWT: Burrows-Wheeler Transform
CAZy: Carbohydrate Active Enzymes
CO_2: Carbon dioxide
COG: Clusters of Orthologous Groups
DAG: Directed Acyclic Graph
DGGE: Denaturant Gradient Gel Electrophoresis
eDNA: Extracellular DNA
EMP: Embden-Meyerhof-Parnas
EPS: Extracellular Polymeric Substances
FANTOM: Functional and Taxonomic Analysis of Metagenomes
Fe^{2+}: Ferrous iron
FISH: Fluorescent *in situ* Hybridization
GO: Gene Ontology
GUI: Graphical User Interface
H_2S: Hydrogen sulfide
H_2SO_4: Sulfuric acid
HMM: Hidden Markov Model
KEGG: Kyoto Encyclopedia of Genes and Genomes
Mb: Megabase
MSA: Multiple Sequence Alignment
N_2O: Nitrous oxide
NADH: Nicotinamide adenine dinucleotide
$NADP^+$: (Reduced form of) Nicotinamide adenine dinucleotide phosphate
NADPH: Nicotinamide adenine dinucleotide phosphate
NCBI: National Center for Biotechnology Information
NGS: Next Generation Sequencing
NH_3: Ammonia
NH_4^+: Ammonium
NMDS: Non-metric Multidimensional Scaling
NO: Nitric oxide
NO_2: Nitrite
NO_3^-: Nitrate
ORF: Open Reading Frame
OTU: Operational Taxonomic Unit
PACFM: Pathway Analysis with Circos for Functional Metagenomics
PCR: Polymerase Chain Reaction
PDB: Protein Databank
PPP: Pentose Phosphate Pathway
SO_4^{2-}: Sulfate
SSU rRNA: Small Subunit Ribosomal RNA
T-RFLP: Terminal Restriction Fragment Length Polymorphism
TCA: Tricarboxylic acid
TGGE: Temperature Gradient Gel Electrophoresis
UPGMA: Unweighted-Pair-Group Method with Arithmetic Mean

SAMMANFATTNING

Denna avhandling innefattar arbete kring att ta fram verktyg och programvaror för att möjliggöra och förenkla metagenomiska analyser av organismer och organismsamhällen. Metagenomik är ett relativt nytt forskningsfält som spänner över flera olika vetenskapliga discipliner och har i avhandlingen använts för att bidra till tvärvetenskaplig forskning.

Under det senaste århundradet har människan haft stor negativ inverkan på klimatet och miljön genom bl.a. en gradvis ökning av den globala koldioxidhalten, förorening av luft, mark och vatten och uppkomsten och spridningen av antibiotikaresistens bland patogena mikroorganismer. Mikroorganismer deltar i en lång rad ekologiska processer i miljön och utgör därmed viktiga studieobjekt för att bättre förstå uppkomsten och effekterna av de ovan nämnda miljöproblemen. Utvecklingen av högeffektiv DNA-sekvensering - Next Generation Sequencing (NGS) - under mitten av 2000-talet har revolutionerat våra studier av mikroorganismer. Genom att sekvensera DNA och RNA ur till exempel vattenprover är det numera möjligt att undersöka både vilka mikroorganismer som lever där och vilka ekologiska och funktionella processer de är inblandade i. Innan NGS-metoderna fanns tillgängliga, var man i praktiken hänvisad till att studera de förhållandevis få mikroorganismer som gick att odla i laboratoriet, men många NGS-metoder kräver inte längre odling av mikroorganismerna. Detta gör det möjligt att studera hela organismsamhällen på en gång. Metagenomik har etablerat sig som en relativt ny tvärvetenskaplig metodik som kan harmonisera våra samlade kunskaper inom mikrobiell ekologi och genetik med DNA- och RNA-sekvensdata från miljöprover.

Ett område där metagenomiken har en stor roll att spela är ekotoxikologi - studier av effekterna av kemikalier på flora och fauna i miljön. I avhandlingen har nyutvecklade bioinformatiska programvaror kombinerats med analyser av ekotoxikologiska försök och mikrobiell ekologi. Ett av syftena med avhandlingen har varit att visa att metagenomik är ett kraftfullt verktyg inom ekotoxikologi både på molekylär nivå och på organism- och populationsnivå.

I Paper I presenteras FANTOM (Functional and taxonomic analysis of metagenomes), ett nyutvecklat program som kan analysera metagenom med avseende både på vilka organismer som finns återfinns i metagenomet och vilka ekologiska och funktionella processer som finns representerade däri. FANTOM låter vidare användaren analysera materialet statistiskt och erbjuder flera former av visualisering av resultaten. Programmet är utvecklat för att kunna användas även av biologer med begränsad bioinformatisk erfarenhet. PACFM (Pathway Analysis with Circos for Functional Metagenomics) är ett ytterligare mjukvaruverktyg, även detta med ett grafiskt användargränssnitt, och presenteras i Paper II. PACFM ger forskare ett verktyg för analys och visualisering av biokemiska syntesvägar i metagenom, och gör så på ett mer realistiskt sätt än vad andra vagt liknande program kan erbjuda. Paper III är en studie av marina biofilmer (också kallat perifyton) där det totala genomiska DNA-innehållet i ett mikrobiellt samhälle i marin miljö har sekvenserats. Metagenomikanalysen av dessa marina biofilmer visade att Proteobacteria, Bacteroidetes och Cyanobacteria är de vanligaste organismerna i dessa biofilmer. Dessutom påvisades en funktionell repertoar av anaeroba processer, däribland denitrifikation och metanogenes, vilket tyder på förekomsten av zoner med låga syrehalter inom de mikroekosystem som de marina biofilmerna utgör. Paper III var vidare en pilotstudie inför Paper IV, där en experimentell design upprättades för att undersöka de toxiska effekterna av det antimikrobiella ämnet triklosan på marina biofilmer. Triklosan-exponering visade sig orsaka betydande förändringar i samhällsstrukturen och de funktionella processerna i de marina biofilmerna. Resultaten pekar på att svavelbaserade mikrober samt olika alggrupper kan vara inblandade i detoxifieringen av triklosan. Sammantaget visar avhandlingen att

metagenomik med framgång kan tillämpas inom ekotoxikologi, och att ekotoxikologin har mycket att vinna på att anamma metagenomiska tillvägagångssätt.

Table of Contents

1. INTRODUCTION

Adverse effects of anthropogenic impact on the environment have become conspicuous in the past century and among others include the gradual increase in global CO_2 levels, the contamination of air, soil and water by toxic chemicals and the emergence of antimicrobial resistance among pathogenic microbial species. From global biogeochemical cycles to ecosystem level functions and from primary production in the food web to disease pathogenicity in infections, microorganisms partake in diverse activities in the environment. Hence, the investigation of microorganisms constitutes a high priority research topic for the understanding of progression and consequences of the aforementioned environmental hazard scenarios. However, paradigms and methodologies utilized in the traditional microbiology alone are insufficient to provide solutions to cope up with the complexity of the listed biological phenomena. A comprehensive elaboration about the description of these biological problems, promotion of predictive methodologies for the management of their progression and beyond all, an integrative synthesis of the knowledge and instrumentation at different scales of biological organization is imperative (Clements and Newman, 2003).

Interdisciplinary research fields have emerged to provide the most optimal solutions in such complex cases in the history of science. We know that a large majority of the prominent advances in science has occurred at the intersection of various disciplines (Garner et al., 2013). For example, the foundation of biochemistry emerged from the cooperative achievements of biomedical scientists, biologists and chemists (Chen et al., 2015a). The efforts in this PhD thesis ultimately focus on the utilization of metagenomics in the field of ecotoxicology. In order to initiate the pursuit of this focus, a sound interdisciplinary methodology including the dissection, juxtaposition and synthesis of the constituent disciplines is required. Dissecting the above mentioned research question into its constituent disciplines drags us into the fields of microbiology, microbial ecology, bioinformatics, and ecotoxicology where the first two largely contribute to the fundamental theoretical background and the latter two mainly provide the methodologies applied. Before expanding on these individual disciplines, a thorough description of interdisciplinarity, potential challenges regarding the list of constituent disciplines and a guideline for a definitive reading of the entirety of this PhD thesis are explained below.

1.1. INTERDISCIPLINARITY REVISITED

The ongoing interchangeable usage of the terms crossdisciplinarity, multidisciplinarity, and interdisciplinarity has resulted in the prevalence of a fallacious notion about the ontologies of these approaches to research. Unlike the term disciplinary, which corresponds to the involvement of a single disciplinary approach to a research field, the terms crossdisciplinarity, multidisciplinarity and interdisciplinarity all refer to the involvement of multiple disciplines with subtle differences, although they have been continuously misused interchangeably over time by scholars (Allen et al., 2011). Crossdisciplinarity translates into the examining of an issue, typically relevant for one discipline from the perspective of another (*e.g.* juridical evaluation of embryonic stem cells). Multidisciplinarity refers to the examining of an issue from multiple perspectives without exhibiting any systematic efforts to integrate the investigated disciplines. However in interdisciplinary analysis, an issue is examined from multiple perspectives as a result of the systematic efforts to integrate the paradigms originated from the individual constituent disciplines into a unified solution framework. In contrast to

cross- or multi-disciplinarity, interdisciplinarity requires the harmonization of different paradigms rather than solely depicting a multitude of perspectives in a disintegrated manner. Harmonization of paradigms from multiple disciplines, though, initially requires a solid comprehension of the individual disciplines and furthermore the ability to synthesize new knowledge via the combination of their paradigms that could otherwise not be possible by the utilization of individual disciplines alone (Max-Neef, 2005).

BOX 1. Levels of biological organization are roughly listed as molecules, cells, species, populations, communities and ecosystems in the increasing order of complexity, respectively. Zooming in and out of the biological organization ladder by complexity will naturally introduce more or less levels into the organization. For instance, some textbooks prefer adding organelles to the lower end of the ladder between molecules and cells whereas others introduce the concept of guilds between populations and communities at the higher end of the ladder. The given list covers all relevant biological organizational terminology that is used in this PhD thesis.

A closely related matter to the misuse of interdisciplinarity in biological sciences is an overarching communication gap between the practitioners of the disciplines from the distinct parts of the biological organization ladder which include molecules, cells, species, populations, communities and ecosystems in the increasing order of complexity, respectively (see Box 1). It is not uncommon that the research fields in biology that investigate distinctive levels of biological organization are attributed to be in competition to falsify each other (MacIorowski, 1988). Paradigms of one level may lack the sufficient criteria to be appreciated by the scholars of another. The inferences made in each level are typically justified by three ways of approaching scientific phenomena (Newman and Clements, 2007). First, microexplanation is exhibited by reductionist scientists working with the lower levels of the biological organization ladder such as molecular or cellular levels. Secondly, macroexplanation infers knowledge about the parts of a system by observing the behaviors of the whole. In the last approach, holism bases the inferences of scientific phenomena on consistent patterns or behaviors at higher levels of biological organization without the necessity to report causal links to lower levels. The latter two are frequently applied by ecologists to acquire knowledge unlike the molecular biologists focusing primarily on the microexplanatory aspects of biological systems (Newman and Clements, 2007).

Different approaches to inference-making strategies in biology, all have their own pitfalls that may confine individual researchers or even a research community within a state of scientific oblivion unless practiced in combination with others. For example, biological inferences based solely on holism are prone to prediction errors since the causality of the mechanistic understanding in the lower levels is neglected (Newman and Clements, 2007). On the other side, biological inferences based solely on microexplanation will most likely overlook the emergent properties of the system, constituting perhaps the notorious consequence of reductionism. One of the ambitious goals in this PhD thesis is to provide both microexplanation and macroexplanation to questions addressed at the different levels of biological organization and produce causal links from lower levels to holistic explanations at the ecosystem level. In order to do so, interdisciplinary analysis methodology is utilized to expand each method of acquiring knowledge so as to avoid “naive reductionism” or “pseudo-scientific holism” (Caswell, 1996) through the use of metagenomics, of which greatest

strength stem from providing relevant data for each individual level in the biological organization ladder.

Throughout the *Background* and *Methodology* sections of this PhD thesis, constituent disciplines in metagenomics and ecotoxicology are dissected and elaborately explained. In these sections, brief remarks from the results of individual papers appended to the thesis, are also introduced in order to inform the readers about how the information in the corresponding section is utilized throughout the papers. The section, *Results and Discussion* introduces the major findings of the papers, and presents the applied aspects of the concepts introduced in the *Background* and *Methodology* sections. Finally, *Conclusions and Outlook* section presents a refined summary of the work performed in this PhD thesis as well as taking a critical look at the methodologies applied, findings inferred and prospects pointed out for future studies.

2. AIMS

The aims of this PhD thesis are summarized below.

- Development of bioinformatics-oriented software tools to analyze metagenomics data.
- Description of the functional and taxonomic diversity of marine biofilm communities.
- Utilization of metagenomics as a methodology in the field of ecotoxicology.

In this PhD thesis, bioinformatics software development and microbial ecological data analysis projects are harmonized under the umbrella field of microbial ecology called metagenomics. The ultimate purpose of the PhD project has been the utilization of metagenomics in the field of ecotoxicology as a robust tool to gain both community and molecular level insights on understanding the effects of toxicants on microorganisms in the marine environment. Papers I and II present two software development projects that took place during this PhD period. FANTOM (Functional and Taxonomic Analysis of Metagenomes) was published in Paper I and is a graphical user interface based metagenomic data analysis tool that provides various statistical analysis and visualization features. PACFM (Pathway Analysis with Circos for Functional Metagenomics) provides the researchers in metagenomics with a graphical interface to be utilized for functional metagenomic analyses (Paper II). Paper III is an exploratory study of the marine biofilms, also known as periphyton, constituting the very first study to sequence the microbiota of this phototrophic slime community - as previously referred to - that grows in aquatic environments. Paper III also constituted the pilot study for Paper IV where an experimental design was set up to investigate the toxic effects of the broad spectrum antimicrobial agent, triclosan [5-chloro-2-(2,4-dichloro-phenoxy)-phenol], on the marine biofilm communities.

3. BACKGROUND

As one of the aims of this PhD thesis beside the focus on the bioinformatics software development has been the utilization of metagenomics as a methodology in ecotoxicology, the addressing of this aim from an interdisciplinary analysis perspective may start from the dissection of metagenomics and ecotoxicology into their constituent disciplines. The fundamental paradigms that nourish metagenomics stem from microbiology and microbial ecology as well as the methodological pillars constructed upon bioinformatics. Ecotoxicological paradigms at community level also largely originate from microbial ecology as well as application of methods derived from toxicology into community ecology. Standardized ecotoxicological tests will not be discussed within the scope of this PhD thesis and instead a substantial focus will be given to the establishment of metagenomics to be applied within the field of ecotoxicology.

The following sections will elaborate on the utilized aspects of the disciplines of microbiology and microbial ecology within the extent of this PhD thesis. Microbiology mainly studies the organismal and sub-organismal level biological entities and processes such as the metabolism of nitrogenous compounds or the protein complexes that mediate bacterial motility, whereas microbial ecology is mainly attributed to above population level biological organization as it utilizes ecological paradigms on the microbial scale. Since experimental settings designed within the field of community ecotoxicology immensely employ multi-species microbial biofilms as a test system, biochemical components, functions and emergent properties of biofilm forming microorganisms are elaborated below; thus starting from microbial metabolism to ecological interactions that take place in multi-species biofilms.

3.1. MICROBIOLOGY

3.1.1. MICROBIAL METABOLISM

According to nutritional characteristics, microorganisms are grouped by the carbon sources they utilize, type of reducing equivalents they have, and energy sources they rely on. Bacteria that produce their own carbon sources through the fixation of CO_2 are called autotrophs whereas those that rely on other organisms to obtain organic carbon are called heterotrophs. Energy production in cells requires the transfer of electrons from different nutritional sources. Organotrophs are organisms that drive this electron transfer from one compound to another via organic molecules and if the electrons are utilized from inorganic compounds, the bacterial groups are then dubbed lithotrophic. The microorganisms that utilize sunlight for the source of energy that is required for cellular energy production for biosynthesis and other cellular activities are named phototrophic and the ones that generate ATP solely through the free energy released from chemical reactions are named chemotrophic. Bacteria are frequently attributed by a combination of these nutritional characteristics. For example, chemolithoautotrophs oxidize inorganic compounds to produce electron motive force for ATP generation, produce energy through a sole base of chemical reactions and also fix inorganic carbon. In order to understand the nutritional preferences and biogeochemical functioning of microbial communities, an elaborate description of microbial metabolism is essential and thus explained below.

3.1.1.1. HETEROTROPHIC METABOLISM

RESPIRATION

Heterotrophic bacterial metabolism involves the oxidation of organic compounds as sole energy sources. Carbohydrates, lipids and proteins are the most commonly utilized substrates by heterotrophs. Generation of ATP and reducing equivalents (*e.g.* NADH and NADPH) is achieved by the aerobic and anaerobic oxidation of these substrates through various biochemical pathways and reaction cycles. Aerobic respiration, which provides the maximum yield of energy from one molecule of glucose, involves three distinct steps of processes leading to the generation of 38 ATP molecules in total. The first step is a pathway utilized by both aerobic and anaerobic microbes called the glycolysis (Embden-Meyerhof-Parnas, EMP) pathway. This step results in the generation of net 2 ATP molecules and 2 NADH molecules. Most bacteria, unlike *Cyanobacteria* and all eukaryotes, are unique in the sense that the glucose oxidation may be performed by more than one pathway (Jurtshuk, 1996; Eiler et al., 2016).

In addition to the previously mentioned glycolysis pathway (*i.e.* EMP), different bacterial groups also possess the pentose phosphate pathway (PPP) and the Entner-Doudoroff pathway. The second step of aerobic respiration requires the availability of O_2 in the ambient environment and is called the Krebs (citric acid, tricarboxylic acid, TCA) cycle. In the final step, the transfer of electrons occurs through a series of membrane bound molecules along with oxidative phosphorylation. The utilized organic substrates are completely oxidized to CO_2 and H_2O at the end of the aerobic respiratory pathway (see anaerobic respiratory pathways below). In Paper III and Paper IV, a large majority of bacterial and eukaryotic members of the studied biofilm communities were found to be heterotrophs and the functional metagenomic analyses in Paper III revealed the abundance of DNA sequences matching with the oxidative phosphorylation pathway in these communities.

FERMENTATION

All plants and animals as well as certain microbial groups utilize aerobic respiration as their primary route of energy production in the presence of O_2. In the absence of O_2, bacteria have evolved to utilize alternative pathways to respiration. Fermentation is one of these alternative pathways that certain bacterial groups adopted, in order to grow under anaerobic conditions. Fermentation basically involves the oxidation of reducing equivalents produced during the glycolytic pathway by utilizing organic molecules (or hydrogen) as terminal electron acceptors (Thauer et al., 1977). The incomplete anaerobic dissimilation of glucose results in the formation of simple organic end products such as ethanol, lactic acid, acetic acid, and butyric acid. Bacteria are very commonly named after the fermentation end products they release, albeit the mixed-acid fermentations operated by the members of the family *Enterobacteriaceae* (Clark, 1989).

ANAEROBIC RESPIRATION

Some bacteria utilize alternative electron acceptors such as nitrate (NO_3^-), Mn (VI), Fe (III), arsenate (AsO_4^{3-}), sulfate (SO_4^{2-}), CO_2, or organic compounds including fumarate and methane in order to carry out the energy efficient respiration process in the absence of O_2.

The reduction potential of the listed inorganic compounds are all lower than the reduction potential of O_2. Nitrate is thermodynamically the most favorable terminal electron acceptor for respiration after O_2. Nitrate is utilized in the pathways of denitrification and dissimilatory nitrate reduction as the terminal electron acceptor with energy yields of, 7% and 35% less than that of aerobic respiration, for the respective pathway (Strohm et al., 2007). The high yields of energy derived through these pathways allow bacteria to produce energy levels close to those of oxidative respiration for growth in anaerobic conditions. The majority of the anaerobic respirers are heterotrophic bacteria, although there are autotrophic exceptions (Tichi and Tabita, 2001).

In the denitrification pathway, nitrate is reduced in a stepwise manner to nitrite (NO_2^-), nitric oxide (NO), nitrous oxide (N_2O), and dinitrogen (N_2), respectively. The enzymes required for the individual reduction processes are nitrate reductase, nitrite reductase, nitric oxide reductase and nitrous oxide reductase, respectively. In Paper III (Supplementary Figure S4), we found nearly all steps of the denitrification pathway in the metagenomic dataset of marine biofilms. We detected that the sequence reads matching with nitrous oxide reductase belonged to only flavobacterial orthologues, which hints that the marine biofilms accommodate species that incorporate only partial steps of the denitrification pathway. Nonetheless, we found that, the abundance of *Flavobacteria* in the biofilms secured the full reduction of nitrate to the dinitrogen gas, avoiding the accumulation of intermediary reduction products, especially the greenhouse gas nitrous oxide.

3.1.1.2. AUTOTROPHIC METABOLISM

PHOTOSYNTHESIS

Photosynthesis is the sequence of biochemical processes by which energy emitted by the sun in the form of photons, is stored and utilized by the biota on Earth. Photosynthetic organisms take the primary production role in the energy cycle as opposed to the heterotrophs that rely on autotrophs for survival. Photosynthesis consists of two sets of reaction series, namely light-dependent and light-independent reactions. Light dependent reactions involve the absorption of light, the photolysis of water, reduction of $NADP^+$ and ATP generation. Light-independent reactions are also known as the Calvin-Benson-Besham or simply Calvin cycle and involve the fixation of CO_2 into various carbohydrate forms that are built upon the six-carbon sugars such as glucose and fructose. Apart from the Calvin cycle, bacteria are known to utilize five more pathways to fix inorganic carbon, namely the reductive tricarboxylic acid cycle, the reductive acetyl-CoA or Wood-Ljungdahl pathway, the 3-hydroxypropionate bicycle, the 3-hydroxypropionate/4-hydroxybutyrate, and the dicarboxylate/4-hydroxybutyrate cycles (Fuchs, 2011). The ability to synthesize their own glucose intracellularly, to be used in further anabolic or energy driven reactions, distinguishes autotrophs from the heterotrophic organisms.

In the light dependent phase of photosynthesis, the absorption of light is mediated by light-harvesting complexes involving different pigment molecules that emit light at varying wavelengths. These photosynthetic pigment molecules are classified into three basic groups, namely chlorophylls, carotenoids and phycobilins. Chlorophylls are the predominant pigments in the land plants whereas in the marine phytoplankton, the major light harvesting pigments are carotenoids, usually giving them red, orange or yellow colors (Kirchman, 2008).

Phycobilins are mostly found in *Cyanobacteria* and *Rhodophyta* in the marine environment, allowing these organisms to absorb red, orange, yellow and green light. Moreover, in contrast to the other membrane bound types of pigments, phycobilins form the water-soluble and mobile light-harvesting antenna complex of phycobilisomes (Okafor, 2011). The light harvesting complexes in various microbial groups were among the other indicators of the abundance of photosynthetic organisms in the marine biofilm communities such as the photosystems I and II-related proteins identified through the functional metagenomic analyses in Paper III.

CHEMOSYNTHESIS

Sunlight is not the only energy source in nature that microorganisms use to synthesize their food. There are certain bacterial groups called the chemoautotrophs, or simply chemotrophs, which utilize the energy from the oxidation of inorganic compounds such as ammonia (NH_3), hydrogen sulfide (H_2S), and ferrous iron (Fe^{2+}) to fix CO_2. The chemotrophs take a crucial role in biogeochemical cycles by closing each elemental cycle (Hügler and Sievert, 2011). They also fix carbon by catalyzing redox reactions from a variety of electron donors including S^{2-}, ammonium (NH_4^+) and H_2 as well as electron acceptors including O_2, CO_2, SO_4^{2-}, S^o, and NO_3^-. Thermodynamics of the redox couples and the biochemical features of the utilized metabolic pathways determine the final energy yield in the chemosynthetic pathways (McCollom and Amend, 2005).

AMMONIA OXIDATION AND THE NITROGEN CYCLE

Heterotrophic bacteria catabolize organic nitrogenous compounds to amino acids and inorganic NH_3 through a process called ammonification. When the NH_3 levels in the environment increase, specialized bacteria, accommodating the gene responsible for NH_3 oxidation (*nosZ*), also start growing and producing energy through a reaction series called nitrification. Nitrifiers mostly exist as chemosynthetic autotrophs that convert ammonia to nitrate as the end product (Paerl and Pinckney, 1996; Francis et al., 2007). As previously explained denitrifiers then, convert nitrate to dinitrogen and the nitrogen cycle is closed by a very specialized group of prokaryotes called diazotrophs, converting dinitrogen to ammonia and subsequently to cell proteins through a process called nitrogen fixation.

SULFUR OXIDATION AND THE SULFUR CYCLE

Reduced sulfur compounds, inorganic sulfur and thiosulfate are oxidized by specialized bacteria, producing sulfuric acid (H_2SO_4) throughout the sulfur oxidation process (Friedrich et al., 2005). Certain bacteria including *Thiobacillus denitrificans*, have been found to embody the functional capacity in their genomes to perform sulfur oxidation anaerobically by using nitrate as the terminal electron acceptor (Beller et al., 2006). In Paper IV, we detected the most commonly known sulfur-oxidizing bacterial order, *Thiotrichales* among the 16S ribosomal RNA (rRNA) amplicons of the samples taken from a high level of triclosan exposure concentration, signaling implications for sulfur cycling within the microbial biofilm communities along with the other detected taxa including the sulfate reducing *Desulfobacterales* and the purple sulfur bacteria, *Chromatiales*.

3.1.2. BIOFILMS

The very first bacteria that had ever been observed under the microscope were from a scraped tooth sample that Antonie van Leuvenhoek introduced to microbiology in 1742. During the following centuries scientists did not focus on the habitat or the life form of the initial observation but solely kept their interests in identifying the microbes in various samples due to the urge to describe the microbe-disease relationships. Experimental settings and laboratory tests were developed based on the premise that the pathogenic bacteria may grow freely in liquid cultures. This free-living or "planktonic" form of microbial life is very commonly found in the aquatic environment. However, according to the inferences of marine microbiologists, less than 1% of the microbes observed under the microscope can readily be grown in culture media (Costerton, 2007). Moreover, in the last century, it was discovered that many microorganisms preferentially attach to various surfaces and exhibit a "sessile" life form when possible, as opposed to their free-living counterparts. William J. Costerton defined the concept of "biofilm" as a microbial life form that is found in virtually all environments that encompass a surface substratum, enough nutrients and water for the bacteria to grow (Costerton et al., 1995). There are two opposing views on the motive for the microorganisms to form biofilm structures (Molin, 1999). Firstly, the biofilm communities may be formed by a merely random aggregation of bacterial groups that accommodate the association and interactions to benefit the community structure. According to the second point of view, microbial biofilms are evolved as deterministic structures in response to environmental stimuli and predominate various natural ecosystems as a distinctive life form (Molin, 1999).

Biofilms can be formed by a single species of bacteria as well as the result of communication of a consortium of multiple species invading various biotic and abiotic surfaces. The microconsortium formed by the biofilm species confer distinctive functionalities to the biofilm form of life, including the construction of physiochemical gradients inside a mucilaginous matrix of extracellular polymers. The microbiota is provided with the optimal environment for cell-to-cell communication and horizontal gene transfer to spread genes to resist disturbances such as exposure to antimicrobial agents, temperature and UV irradiation (Decho, 2000). As such, biofilm-forming bacteria in the households and medical settings, have been shown to be highly resistant against chemical disinfection, antibiotics and immunological responses (Costerton et al., 1999; Hall-Stoodley et al., 2004). Biofilm form of life is thus, not surprisingly, found at various environments including the seas and oceans (Cooksey and Wigglesworth-Cooksey, 1995), rivers and streams (Neu and Lawrence, 1997), acid mine drainage sites (Edwards et al., 2000), thermal springs (Ward et al., 1998), wastewater treatment plants (Lazarova and Manem, 1995) as well as in the form of disease causing agents in and on the human body (Singh et al., 2000; Marsh, 2004).

Biofilms are described below according to the three major topics of interest that were also noted in the Supplementary Table I of Paper III, namely biofilm formation, content of extracellular polymeric substances and ecological interactions with regard to their contribution to biogeochemical cycles and energy economy of the community.

3.1.2.1. BIOFILM FORMATION

Biofilm formation is initially triggered by the movement of microorganisms toward a solid substrate surface. Bacterial motility is therefore essential for the initial adhesion of bacterial

colonies to surface substrata. The effective attachment of biofilm species rely both on the surface structures of individual cells and the substratum (O'Toole et al., 2000). Bacteria utilize membrane proteins called adhesins that facilitate the adhesion onto abiotic surface materials (Kachlany et al., 2000; Dunne, 2002) and host organisms (Mittelman, 1996; Amano et al., 1999) with high affinity. For example, *Thiobacillus ferroxidans* uses the membrane bound protein, aporusticyanin, to attach to pyrite (Ohmura and Blake, 1997). In another example, *Staphylococcus aureus* uses fibronectin and collagen-binding proteins to colonize eukaryotic cell surfaces (Foster and Höök, 1998). The initial attachment of biofilm bacteria is also facilitated by the sticky nature of extracellular polysaccharides secreted by certain planktonic bacteria (Mayer et al., 1999). Furthermore, a vast array of functional groups exhibited by the secreted extracellular substances enable the invading bacteria to attach by covalent bonding, hydrogen bonding, hydrophobic interactions, electrostatic and van der Waals forces (Sussman et al., 1993). During the biofilm formation, relying on the changes in the ambient environment, succession of different species takes place. After the succession of primary colonizers, secondary colonizers adhere to the already attached organisms, thereby forming a multi-species community structure (Kolenbrander, 1989).

BACTERIAL MOTILITY

Bacterial dynamics during the biofilm formation phase have previously been shown to involve the crucial role of cellular motility required to reach surfaces (Korber et al., 1994). Bacteria use flagellar, twitching and gliding motility to attach and colonize surfaces (Stewart and Costerton, 2001). The mode of movement is shaped upon the motility proteins that different bacterial groups possess, as described below.

MOTILITY PROTEINS

Microorganisms utilize several protein complexes for motility. Flagella, pili and fimbria are the bacterial motility complexes that take role in biofilm formation (Wimpenny, 1992). A flagellum uses rotary motion, analogous to a propeller with an attached motor protruding from the cytoplasmic membrane and is structurally similar to type III secretion systems in bacteria (Aldridge and Hughes, 2002). The propelling movement is performed by a component called, the filament, which is 20 nm in diameter and is a helical assembly of thousands of copies of the single protein called flagellin. Pili are about ten times thinner surface structures than the flagella and take role in multiple functions including adherence to solid surfaces, twitching motility and conjugation (Bardy et al., 2003). Bacteria carry out a different type of motility by pili than the propeller-like movement provided by flagella. For example, type IV pilus is shot out from the bacterial cell wall at the substratum and the microorganism is then pulled towards the surface with a jerky movement called twitching motility (Pasmore and Costerton, 2003). Type IV pili also perform the recognition role during the uptake of extracellular DNA fragments through a process called transformation (van Schaik et al., 2005). Another filamentous structure that takes role in the initial attachment of biofilm communities is the fimbrium. Fimbria are also known as attachment pili due to their primary role in surface attachment, however, they do not take role in motility. It has been shown in several studies that bacteria lose their adherence ability to solid surfaces when the genes expressing fimbria are knocked out (Prouty et al., 2002). In fact, biofilm formation is shown to be halted in all mutant bacteria that lack motility proteins (Pratt and Kolter, 1998).

Motility proteins were searched for in the metagenome of the marine biofilms as part of the analysis of biofilm-relevant functional content in Paper III.

3.1.2.2. BIOFILM STRUCTURE

"The city of microbes" has previously been used as a metaphor to describe the biofilm structure, due to the selective settlement of community members on different parts of the biofilm, energy storage in the extracellular space in various forms and transfer of genetic material for the collective succession of the community (Watnick and Kolter, 2000). Following the initial attachment, bacterial motility stops and extracellular polymeric substances (EPS) are secreted from individual cells. In fact, the actual living cells make up to a maximum of 10 % of the dry mass of the community while the rest of the extracellular space is covered by the EPS matrix (Flemming and Wingender, 2010).

EXTRACELLULAR POLYMERIC SUBSTANCES

The matrix that the organisms are encased, in a biofilm, is made of a combinatorial aggregation of various biopolymeric materials called EPS. The EPS provides a protective microenvironment for different types of metabolic processes that take place within the biofilms. Further advantages that the EPS matrix provides for the community members and individual biopolymeric components are described below.

EPS FUNCTIONS

In addition to its adhesive support at the initial attachment stage of the primary invader species, the EPS also provides cohesive stability for the community members by immobilizing the cells and positioning them in close proximity (Flemming and Wingender, 2010). Moreover, it has been shown that the EPS contains special chemical cues that marine invertebrates are attracted for settlement (Hadfield and Paul, 2001). Larval development of these invertebrate species is secured by firmer attachment to the EPS than to clean surfaces and the EPS hence, behaves as an environmental placenta for the larva of these species by providing the appropriate conditions prior to metamorphosis. DNA belonging to various invertebrate groups was identified in the taxonomic analysis of Paper III, including the phyla *Arthropoda*, *Mollusca*, and *Cnidaria* in the investigated marine biofilms from the Swedish west coast. It is most likely that the DNA sequences originated from the larva of these invertebrate groups, which utilize the marine biofilms as a temporary settlement habitat.

The EPS matrix includes charged or hydrophobic polysaccharides and proteins to sequester dissolved and particulate nutrients from the water. Biofilm organisms can utilize these nutrients for energy production and also store the excess energy within the extracellular polysaccharides for future use. Furthermore, not only does the EPS adsorb organic compounds that are readily available as nutrients from the ambient environment, but it also sequesters xenobiotics and other organic compounds that are used as biocides (Davey and O'toole, 2000). For example, diclofop methyl, a widely used herbicide, was shown to be degraded and utilized as nutrient source by biofilm organisms (Wolfaardt et al., 1998). Biofilms, therefore, take a purification and detoxification role in the aquatic environment. In Paper IV we hypothesized that certain species in the marine biofilms detoxified triclosan through a sulfurylation reaction between triclosan and sulfate.

In addition to its protective role against the antimicrobials, the EPS also protects the community members from other environmental stressors such as UV radiation, pH shifts, osmotic shock and desiccation by its highly hydrated content (Sutherland, 2001) where the retention of water reaches up to 97% of the total mass (Zhang et al., 1998). Diurnal variation in humidity, thus, is not a lethal problem for the majority of the biofilm community members. Ironically, stability provided by the EPS constitutes a problem in the maritime industry. Microalgae, especially diatoms, secrete large amounts of EPS subsequent to the attachment onto, for example, ship hulls and prepare the conditions for heterotrophic bacteria, protozoans, fungi and invertebrates to settle and cause biofouling, resulting in serious financial damages to the industry (Abbott et al., 2000).

EPS STRUCTURE

Extracellular polymeric substances were previously named as extracellular polysaccharides due to the intensity of sugar molecules in the matrix. However, it was later understood that proteins, enzymes, nucleotides, lipids and other biopolymers such as humic substances were also involved in the structure (Flemming and Wingender, 2010). Chemical analysis of EPS has been cumbersome due to the vast array of those biopolymers in the matrix and therefore EPS has been dubbed "the dark matter of biofilms" (Sutherland, 2001; Flemming et al., 2007). Additionally, composition of the EPS varies between different biofilms, further increasing the complexity of chemical analyses. Diversity of microorganisms, various forms of outside disturbances, temperature and nutrient content in the biofilm communities are the major parameters that affect the EPS composition. Below, components of the EPS, *e.g.* extracellular proteins, enzymes, polysaccharides and DNA that can directly or indirectly be linked to the DNA reads generated by metagenomic sequencing, are described. These EPS components were searched in the public sequence databases in order to explain the biofilm relevant content in the metagenomic sequence reads generated for Paper III.

EXTRACELLULAR PROTEINS

In contrast to early thoughts on the EPS content, we now know that proteins can reach substantial proportions within the matrix structure, outweighing the extracellular polysaccharides (Metzger et al., 2009). More specifically, lectins constitute the majority of the extracellular non-enzymatic protein molecules in the biofilm matrix. They are basically carbohydrate-binding proteins with high affinity, allowing the bacteria to form and stabilize the EPS matrix. They also serve as an authentication gate between the bacterial membrane surface and the EPS with regard to their characteristic specificity to certain sugar molecules. Labeled lectins have previously been used to analyze the carbohydrate composition of EPS matrices produced by biofilms in different environments (Tielker et al., 2005; Diggle et al., 2006; Lynch et al., 2007). Other groups of extracellular proteins detected in the EPS matrix include biofilm associated surface protein (bap), bap-like proteins, amyloids, adhesins and the previously mentioned motility protein complexes such as pili, fimbriae and flagella (Flemming and Wingender, 2010).

EXTRACELLULAR ENZYMES

Enzyme groups found in distinctive biofilm samples include protein-, lipid- and sugar-degrading enzymes as well as oxidoreductases and phosphomonoesterases. These diverse

groups of enzymes provide an external digestive compartment for the biofilm organisms to convert biopolymers into simpler molecules, *e.g.* to be utilized as carbon sources. This enzymatic activity turnover, for instance in the EPS matrix of aquatic biofilms, contribute to global nutrient cycles, although biofilm studies have *hitherto* been restricted mainly to local settings. Enzymatic abundances may also constitute a proxy to estimate the types of different sugar polymers in addition to the specificity of lectins mentioned previously. For example, endocellulases, chitinases, alpha- and beta- glucosidases, beta-xylosidases, N-acetyl-Beta-D-glucosaminidases, chitobiosidases and beta-glucuronidases were previously detected as the polysaccharide degrading enzymes in aquatic biofilms (Flemming and Wingender, 2010). Investigation of the biofilm enzymes is also critical due to their dispersion effect on the biofilms in medical and industrial settings. However, discovery of a single enzyme or any other molecule for the dispersal of biofilms also has the risk to initiate a global scale environmental disaster, due to the drastic roles of biofilms in the aquatic environment such as enacting as self-purification systems (Ahner et al., 1995; Miao et al., 2009). Nevertheless, this risk is potentially bypassed through the variability and complexity of the EPS composition in different environmental biofilms and finding of specific dispersal enzymes targeting biofilm forming pathogenic bacteria may constitute an alternative therapy for infectious diseases related to biofilms. The last types of extracellular enzymes relevant for environmental biofilms are the redox enzymes. The financial damage of biofilms caused by biofouling originates mainly from the presence of redox enzymes found in the EPS matrix and their corrosive activities (Busalmen et al., 2002).

EXTRACELLULAR POLYSACCHARIDES

Polysaccharides are the other group of biopolymers that constitute a large portion of the EPS matrix. The presence of uronic acids and ketal-linked pyruvate in the majority of extracellular polysaccharides, determines their negatively charged molecular structure, although they exist in neutral forms, too. In Paper III, the metagenomic analyses were extended to search for specific sugar-degrading enzymes in the specialized database of Carbohydrate Active Enzymes (CAZy; Cantarel et al., 2009). The CAZy database searches revealed the abundance of carbohydrate-esterase-family 4 and carbohydrate-binding-module-family 50 gene copies in the metagenome of the marine biofilms. These enzyme families are associated with the degradation of chitin-like polymers, which can be explained by the detected presence of mollusks and arthropods within the biofilm communities (Caufrier et al., 2003; Ehrlich, 2010).

EXTRACELLULAR DNA

The final biopolymeric component found in the biofilm matrix, relevant for the functional metagenomic analyses carried out in Paper III is the extracellular DNA (eDNA). The eDNA has previously been detected to be abundant in the biofilm matrix of wastewater biofilms (Frølund et al., 1996). It was observed to take a major structural role in the biofilms of certain species (Wang et al., 2015), whereas no significant link to the EPS structure was detected in the biofilms of others (Izano et al., 2008). Despite its fundamental relevance to metagenomic studies, to my knowledge, there have not been any studies investigating the eDNA to intracellular DNA ratio in multi-species biofilms.

3.1.2.3. ECOLOGICAL INTERACTIONS

The resemblance of biofilm community members to the dwellers of a city is reflected by the intra- and inter-species interactions within the EPS matrix structure. These interactions are driven by specialized bacterial groups that collectively adapt to the microenvironments emerged in the biofilms. Biofilms are therefore typically not homogeneous according both to the spatial distribution of different community members and to the physiochemical properties of the individual microenvironments (De Beer et al., 1994). For instance, it has previously been shown that oxygen concentration and pH drastically drop in the proximity of the surface substratum (Lee and de Beer, 1995). Single species biofilms adapt to these changing micro-environmental conditions by altering their gene expression patterns at different locations within the biofilm. Although not completely analogous to the developmental stages of higher eukaryotes, this phenomenon reminds of the differentiation of multiple organs throughout separate body parts. In multi-species biofilm communities, species distribution at distinctive microenvironments of the biofilm matrix is dependent on the adaptation of both individual species and synergistic relationships between different species (Elias and Banin, 2012). Hence, bacterial evolution in multi-species biofilms is not totally incidental but a result of the progressive interactions and co-evolution within separate microenvironments.

3.2. MICROBIAL ECOLOGY

Microbial ecology is the study of microorganisms throughout a wide range of biological organization levels from individuals to communities and ecosystems. As a discipline incorporating the approaches of traditional ecology into a microbial context, the interactions of microorganisms with the biotic and abiotic components in the environment constitute the essence of microbial ecology. An individual in a microbial ecosystem represent a single living cell or a colony formed by genetically identical cells. A population is defined as a group of individuals belonging to the same species that share the same habitat. A microbial community is formed by two or more populations of organisms that spatially and temporally interact. At the top level of biological organization, an ecosystem exists, comprising the microbial community and rest of the biotic and abiotic factors influencing the functioning of the microbial community. Although experimentation becomes relatively more difficult at complex levels of biological organization in macro ecology due to spatial limitations, microbial ecology is advantageous in the sense that even micro-ecosystem level experiments are operable (Jessup et al., 2004). Since the microbial ecological and the ecotoxicological constituents of this thesis focus mainly on the community level, my focus will be on microbial community ecology in this section and throughout the thesis.

The epicenter of microbial ecology comprises three main questions: Who are the species in the community? What do they do for the community? And how do they accomplish that? The first question essentially addresses the structure of the community. The structure of a community is defined by the species present in the investigated environmental habitat. The community analysis can be expanded to the number of species (richness) and the proportions of species abundances (relative abundance) within the community. Biodiversity refers to the degree of variation in biota at all scales of biological organization and has major implications for the health of an ecosystem (Hughes and Bohannan, 2004). It is investigated at temporal and spatial scales and comparative analyses of biodiversity patterns constitute a major component in microbial ecology research (Gonzalez et al., 2012). The microbial diversity

measures include phylogenetic, species, genotype and gene diversity as well as functional diversity, metabolic diversity and protein diversity (Xu, 2011). Biodiversity has been described by various indices and statistical models in macro-ecology, which have also been adapted to microbial ecology studies (Hughes et al., 2001).

Traditional definition of the term species does not apply to bacteria and archaea due to the distinctive mechanisms of reproduction patterns among the microbial world. Following the attempts of microbiologists that classify bacteria according to their, *e.g.* morphologies, metabolic capabilities, and ecological niches, DNA-based designations of the microbial species concept have been promoted by the microbial ecologists. However, bacterial speciation still remains a debated concept and a consensus on the existence of a "bacterial species" has yet not been established among the scholars (Gevers et al., 2005; Doolittle and Zhaxybayeva, 2009). Instead, a pragmatic approach was taken and clusters of a gene sequence were used to describe microbial diversity in the name of Operational Taxonomic Units (OTUs; Schmidt et al., 2014). According to the OTU-based designation of the microbial diversity, two organisms are accepted to be belonging to the same OTU if their 16S or 18S rRNA gene sequences have at least 97% similarity (Barton and Northup, 2011). 16S and 18S rRNA genes have long been used as the determinants of the OTU concept and constructed the foundation of advances in the microbial diversity research. The ribosomal RNA gene was selected for this purpose due to its universality; hence the occupation of conserved regions throughout all species in the tree of life (see section 4.2.1 in *Methodology*). Through the use of small subunit (SSU) rRNA sequencing, microbial ecologists initiated the discovery of a plethora of microorganisms that might never be achieved solely by culturing. A recent study utilizing Next Generation Sequencing (NGS) technologies by sequencing over 1,000 uncultivated microbial genomes revealed the previously unknown diversification in the bacterial branch and lineages that are overlooked in the current biogeochemical models (Hug et al., 2016).

The second question that microbial ecologists address in their research relates to the ecosystem functions of individual community members and the community itself as an emergent entity. Microbial communities can include autotrophs, heterotrophs, and mixotrophs (Eiler, 2006), thereby constituting a complex food web within the community structure. Moreover, several nutrient cycles may take place within the same community structure, conferring additional complexity on top of the trophic features of community members. In the marine environment, microorganisms, primarily macro algae, diatoms and *Cyanobacteria* cover, up to half of the primary production on Earth (Arrigo, 2005). The food web in the marine waters, which depicts the primary producers as *Cyanobacteria*, micro and macro algae, is shown in Figure 1. Protozoa graze on the primary producers and are eaten by the zooplankton at the higher level. On top of the food chain in the marine environment, the fish consume the zooplankton (Figure 1). Algae are known to absorb toxic chemicals in the marine environment (Ahner et al., 1995), leading to the bioaccumulation of these chemicals throughout the food chain. These chemicals may ultimately reach to the fish and consumers of fish such as humans and other constituent organisms in the ecosystem, thereby setting the basis for the environmental toxicity problem. For example, in Paper IV, we hypothesized that triclosan is immobilized on the cell walls of the red algae (*Rhodophyta*). The red algae is ultimately consumed by the fish or the other intermediary steps in the food chain that have the potential to end up back in human households where the release of triclosan to the environment once started.

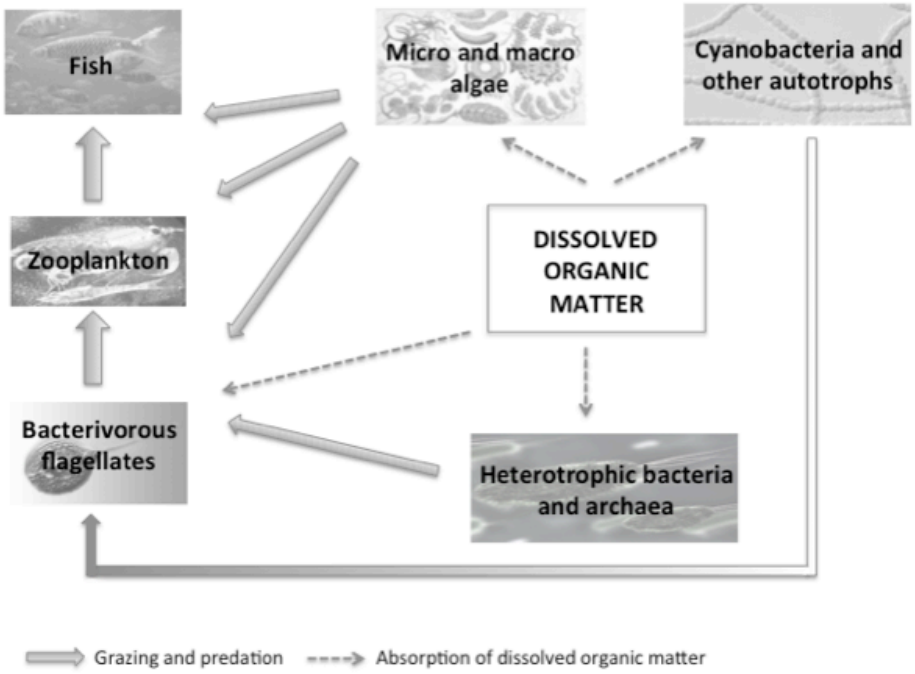

Figure 1. Food web in the marine environment. The food web in the marine waters, comprise primary producers such as *Cyanobacteria* and other phototrophic bacteria, micro and macro algae as well as heterotrophic bacteria, protozoan grazers, zooplankton and fish. Adapted from (Munn, 2004).

Microbial ecologists also address questions regarding the biogeochemical functions of microbial communities (Canfield et al., 2005; Eiler et al., 2014). The trophic classifications of microorganisms based on carbon sources, type of reducing equivalents and utilized energy sources as described in the *Microbiology* section of this thesis as well as emergent properties of communities are investigated as part of microbial community ecology. Finally, the interrelationships between the members of a microbial community with their environment are of interest for microbial ecologists. The interrelationships may involve the positioning of various species within the community structure, cooperation and antagonism among them as well as exchanged signals between them such as the quorum sensing molecules released by the members of a biofilm community (Parsek and Greenberg, 2005).

3.2.1. MARINE POLLUTION

Human societies have regarded the marine environment as a waste-dumping site endowed by nature for centuries. Not only have we overexploited the resources but we have also introduced nonnative organisms to the marine environment, manipulating the dynamics of endemic communities. The disturbances caused by man on the marine environment have peaked since the rise of industrialization and pollution has been added to the aforementioned anthropogenic misconduct. Largely insidious effects of those activities result from the disruptive changes in the ecological dynamics of ambient seawater, ultimately leading towards ecosystem level alterations. It was not so far back in the history when states identified particular anthropogenic interferences in the marine environment as disputable. Marine pollution was defined in 1983 by The United Nations Joint Group of Experts on the Scientific Aspects of Marine Pollution (GESAMP) as "the introduction by man, directly or indirectly, of substances or energy to the marine environment (including estuaries) resulting in deleterious effects such as: harm to living resources; hazards to human health; hindrance of

marine activities including fishing; impairing the quality for use of seawater and reduction of amenities." (GESAMP, 1983).

Pollutants are introduced into the marine environment in the forms of petroleum hydrocarbons, plastics, pesticides and related compounds as well as heavy metals, sewage, radioactive wastes and thermal effluents (Lalli and Parsons, 1997). In the 20th century, a major transformation in the lifestyles of human populations occurred following the discovery and mass production of antibiotics. Diseases and conditions related to bacterial infections have been found to be treatable by antibiotics and antimicrobials. Antibiotics are metabolized in the human body to a large extent prior to excretion, although up to 90% of the parent compound may remain unchanged in certain antibiotic uses including tetracycline and amoxicillin (Hirsch et al., 1999). Furthermore, particular antimicrobial chemicals (*e.g.* triclosan) are administered topically on the human body by the application of personal care products. These antimicrobials are, thus, not metabolized inside the body, but are instead, released directly to the sewage system, thereby increasing the possibility of release into the environment. In addition to fighting against infections in humans, antibiotic and antimicrobial use is largely practiced in livestock farming and aquacultures (Boxall, 2004). Inappropriate disposal both by the consumers and the manufacturing companies is yet another route of anthropogenic discharge of these chemicals into the environment (Boxall, 2004). Hence, antibiotics and antimicrobials have been appended to the above list of chemicals further augmenting marine pollution by directly targeting the microbial marine life, and consequently the entire marine food web. Paper IV in this thesis, investigates the environmental toxicity of triclosan, a widely used antimicrobial compound, to the microbial life in the marine environment.

4. METHODOLOGY

4.1. MODERN METHODS IN MICROBIAL ECOLOGY

The twentieth century witnessed an unprecedented revolution in the history of science by the discovery of the DNA structure (Watson and Crick, 1953). Further findings on DNA sequencing and recombinant DNA technology reinforced the foundation of genetics. The applicability of the paradigm shift that genetics provided for biology had not been feasible in microbial ecology until the discovery of the Polymerase Chain Reaction (PCR; Ho et al., 1989; Muyzer et al., 1993). The PCR allowed researchers, for the first time, to verify the presence of certain genes in organisms by amplifying the targeted gene up to observable amounts. In 1985, Norman Pace and in his colleagues applied PCR and cloning of the 16S ribosomal RNA gene for environmental bacteria (Pace et al., 1985). Consequently, for the first time during the past three centuries of microbial research, a new door was opened to identify the diversity of microbes at genetic level replacing our dependence on growth cultures or mostly microbial morphology-oriented identification through microscopes.

Development of culture independent molecular techniques was invaluable for microbial research since up to 99.9% of the microbial cells sampled from the marine environment were shown to be recalcitrant to culturing in the laboratory (Cho and Giovannoni, 2004). Modern microbial ecology was established upon the fundamental findings on the universal genes and primers, PCR and cloning concepts. However, PCR was used mainly for qualitative purposes in genetics albeit the PCR derived methodologies such as Temperature Gradient Gel Electrophoresis (TGGE; Po et al., 1987), Denaturant Gradient Gel Electrophoresis (DGGE; Muyzer et al., 1993), Fluorescent in situ Hybridization (FISH; Wagner et al., 1993) and later Terminal Restriction Fragment Length Polymorphism (T-RFLP; Liu et al., 1997) were applied to microbial ecology research with a limited resolution to detect the microbial diversity in the environment. PCR, cloning and Sanger method-based amplicon sequencing of marker genes was the most rigorous methodological approach to microbial ecology research during the 1990s and early 2000s. However, Sanger sequencing was too expensive to scale up the sampling efforts and experimentation in microbial ecology. The term metagenomics was coined in 1998 by Jo Handelsman and her colleagues (Handelsman et al., 1998) during the era when Sanger sequencing was still the most widely used sequencing methodology and attained its pervasive use in the field, subsequent to the advances in NGS technologies.

4.1.1. NEXT GENERATION SEQUENCING (NGS) TECHNOLOGIES

Following approximately two decades of domination in the DNA sequencing industry, the first generation of sequencing technologies – the automated Sanger method – started to lose its position in the market to the NGS technologies by the year 2005 (Mardis, 2008). Depending on the applied sequencing chemistry, companies managed to lower the costs of sequencing at large scale, in comparison to the Sanger methodology, by over two orders of magnitude (Shendure and Ji, 2008). Although the accessibility to sequencing facilities has gone through a sensational shift by the democratization of sequencing costs, the contemporary NGS technologies mainly compete in data accuracy, breadth of applications, read length and throughput along with the consumable and instrument prices (Thayer, 2014).

The overall sequencing workflow in the majority of NGS technologies is essentially similar. The processes include template preparation, sequencing and imaging, as well as data analysis, and it is the unique combination of these processes on top of specific chemistries applied that distinguishes them (Metzker, 2010). In Paper III, Roche's 454/Pyrosequencing technology was utilized to sequence the metagenome of the marine biofilms, whereas Illumina sequencing was preferred for the amplicon sequencing of 16S and 18S rRNA genes in Paper IV. 454/Pyrosequencing is based on the quantification of the released pyrophosphate molecules upon nucleotide incorporation by the DNA polymerase enzyme (Ronaghi et al., 1998). Illumina sequencing also relies on the detection of light signals upon the incorporation of fluorescently labeled nucleotides to a template DNA strand by an isothermal DNA polymerase (Bennett, 2004). The sequencing step applied in the two different technologies is called "sequencing by synthesis". The sequences generated from individual runs are called the "reads", and Pyrosequencing (454-GS FLX machine as utilized in Paper III) generates in total a minimum of 35 Mb of 400 bases long reads, whereas Illumina sequencing (MiSeq machine as utilized in Paper IV) generates in total a minimum of 1,500 Mb of 2 x 150 bases long reads in one run (Loman et al., 2012).

4.2. METAGENOMICS

Metagenomics is a young research field that emerged upon a long-term aspiration among the microbial ecologists for studying uncultured microorganisms in order to understand their taxonomic diversity, functional repertoire, cooperation and evolution in environments such as air, soil, water, ancient remains or the digestive systems of animals. In essence, the term has been used synonymously for almost two decades in various contexts with other terms including community genomics, environmental genomics, ecological genomics as well as megagenomics (Riesenfeld et al., 2004; Allen and Banfield, 2005; Handelsman, 2005; Moran et al., 2007). Metagenomics, although widely used to indicate microbial community-genome-sequencing studies, involves several approaches to investigate uncultured microorganisms including SSU rRNA amplicon sequencing, targeted functional gene amplicon sequencing and community- genome shotgun sequencing. In this PhD thesis, community-genome shotgun sequencing and SSU rRNA amplicon sequencing approaches to metagenomic studying of marine biofilm communities were employed in the papers III and IV, respectively. Thus, these approaches are elaborated below.

4.2.1. SSU rRNA AMPLICON SEQUENCING

As a breakthrough in microbial research, the evolutionary significance of the SSU rRNA gene was discovered, which facilitated the screening of microbial diversity in environmental samples without necessitating a culturing protocol (Pace et al., 1985; Woese, 1987). PCR amplified sequences (amplicons) of the 16S (prokaryotes) or 18S (eukaryotes) components of the SSU rRNA gene were chosen as the preferred marker for the microbial diversity for several reasons, including the straightforward procedure of its amplification in most situations (Xu, 2011).

First of all, the SSU rRNA gene is universally found in all organisms allowing the identification of all cellular organisms at any environment. The ribosomal RNA transcribed by this gene is functionally homologous in the archaeal, bacterial and eukaryotic domains of life, which ensures the validity of comparisons in subsequent taxonomic analyses. Third, the

homogeneity of its function throughout the tree of life has resulted in conserved and variable regions across the gene sequence, which enables us to perform high-resolution phylogenetic analyses (Hartmann et al., 2010). Relatively constant evolutionary rate of the gene allows deducing the divergence times of broad taxonomic groups such as archaea, bacteria, plants and animals. Lastly, the aforementioned conserved and variable regions allow us to design PCR primers both universally for individual domains in the tree of life and also for specific target groups (Vandenkoornhuyse et al., 2002; Baker et al., 2003).

The downsides of the utilization of 16S or 18S rRNA gene as a marker-gene in amplicon sequencing studies include high variability of the copy numbers of this gene in different organisms (Klappenbach et al., 2001). For example, in Paper III, the extracted 18S rRNA genes were found to be approximately ten times more abundant than the 16S rRNA genes in the metagenome of the marine biofilms. However, whole genome similarity searches revealed in fact, that bacterial sequences dominated the annotated portion of the biofilm metagenome. Another disadvantage of the SSU rRNA gene amplicon sequencing is the within species sequence variability at extreme cases, where operons of the 16S rRNA gene within the same genome were identified to cover variable regions that qualify these genes as belonging to different species (Pei et al., 2010). Lastly, unless supported with a hypothesis-driven approach, direct metabolic evidence is not warranted by amplicon sequencing of the rRNA gene. However, the use of SSU rRNA gene amplicon sequencing in Paper IV, revealed tangible clues about a sulfur-based metabolic consortium among the marine biofilms established in a flow-through microcosm system.

Once the amplicon sequencing of the SSU rRNA gene is performed, processing of the sequence reads is applied through bioinformatics pipelines tailored for the utilized NGS technology. The sequence-processing pipeline utilized in Paper IV is described in Sinclair et al. (Sinclair et al., 2015), hence the interested readers are recommended to see the *Methods* section of this publication for an elaborate description. Several alternative workflows and software tools have been published, describing the initial processing of amplicon sequence reads (Schloss et al., 2009; Caporaso et al., 2010; Edgar, 2013), and the standard output of these tools is a list of OTUs with abundance data for individual samples. The downstream analysis steps typically involve methodologies largely borrowed from macro-ecological analyses including the calculation of within-community diversity indices such as the Chao1 index, richness, Pielous's evenness and Shannon-Weaver index and between-community similarity measures such as the Bray-Curtis similarity or Euclidian distance measures. Moreover, statistical hypothesis testing and multivariate analyses are further employed to identify significant differences between the community compositions of differentially treated samples. In Paper IV, comparative analyses were performed with the DeSeq2 method (Love et al., 2014). In addition, the Non-metric Multidimensional Scaling (NMDS) method was used to reduce the multidimensional OTU abundance data to be represented by a two-dimensional plot. In this PhD thesis, all microbial ecology analyses except for the DeSeq2 analysis were performed in the R package *vegan* (Dixon, 2003).

4.2.2. COMMUNITY-GENOME SHOTGUN SEQUENCING

In metagenomics research, environmental samples are investigated through a procedure involving extraction and isolation of DNA, sequencing, and data processing and analysis. During the initial attempts to sequence amplicons of the SSU rRNA gene in the late 1990s through the early 2000s, the DNA sequences were required to be prepared through a cloning and library preparation step prior to sequencing, dubbed clonal amplification (Pace et al., 1985). This step and its associated biases were eliminated by the default amplification step of NGS technologies, without the necessity of a cloning procedure (Metzker, 2010). Whole genome shotgun sequencing was already utilized in 1990s during the accomplishments of the first bacterial (Fleischmann et al., 1995) and human genome sequencing projects (Venter et al., 2001). However, sequencing the collective genome of a microbial community by the previously used cloning based shotgun sequencing approaches was practically impossible. The elimination of the cloning step along with the drastically reduced sequencing costs allowed for the application of shotgun sequencing at microbial community scale, which is referred to as community-genome shotgun sequencing or simply shotgun metagenomics. Cloning associated steps prior to the NGS application of shotgun sequencing such as the genome mapping to construct scaffolds of genomic sequences (Dear, 2001) were thereby shifted to the bioinformatics analysis step following DNA sequencing.

Direct sequencing of environmental samples has provided further valuable insights into the life styles and metabolic capabilities of uncultured organisms occupying various environmental niches, allowing us to investigate not only which microorganisms are there but also what they are capable of doing in the environment. Over the last two decades, metagenomics projects such as the exploration of the microbiome of the Sargasso Sea (Venter et al., 2004), the Human Microbiome Project (Turnbaugh et al., 2007), and the Earth Microbiome Project (Gilbert et al., 2014) have expanded our knowledge about the microorganisms living in the environment as well as in and on humans. Discovery of the prevalence of proteorhodopsins in the oceans (Béja et al., 2001; DeLong, 2005) and the enterotypes of the human gut microbiota (Qin et al., 2010; Arumugam et al., 2011) were among the major findings that metagenomics introduced to the biology literature.

Software tools have been made available to run the metagenomic analysis pipelines and the ones utilized in Paper III are described in the *Bioinformatics* section below. The individual analysis steps required for the bioinformatics sequence processing and analysis pipeline in a community-genome shotgun-sequencing project, are shown in Figure 2. There has been software development endeavors targeting specific steps in the whole analysis pipeline such as FastQC (Andrews, 2010), PRINSEQ (Schmieder and Edwards, 2011), and Fastx-toolkit (Gordon and Hannon, 2010) for the sequence quality check and initial filtering; MetaVelvet (Namiki et al., 2012), SoapDenovo (Luo et al., 2012), and Ray Meta (Boisvert et al., 2012) for the sequence assembly; FragGeneScan (Rho et al., 2010), MetaGene (Noguchi et al., 2006), and Prodigal (Hyatt et al., 2010) for the gene prediction; NCBI BLAST (Camacho et al., 2009) , Blat (Kent, 2002), BWA (Li, 2013), and Bowtie 2 (Langmead and Salzberg, 2012) for the sequence alignment and read mapping, Metaxa (Bengtsson et al., 2011), rRNASelector (Lee et al., 2011) for the extraction of SSU rRNA gene sequences from the metagenome, MEGAN (Huson et al., 2007), STAMP (Parks et al., 2014), ShotgunFunctionalizeR (Kristiansson et al., 2009) and FANTOM (Paper I) for the functional and taxonomic analyses (Figure 2). There have also been large scale efforts to provide the

whole metagenomics analysis pipeline as a web-service by projects such as mg-RAST (Meyer et al., 2008), CAMERA (Seshadri et al., 2007), IMG/M (Markowitz et al., 2008) and EBI-Metagenomics (Mitchell et al., 2016).

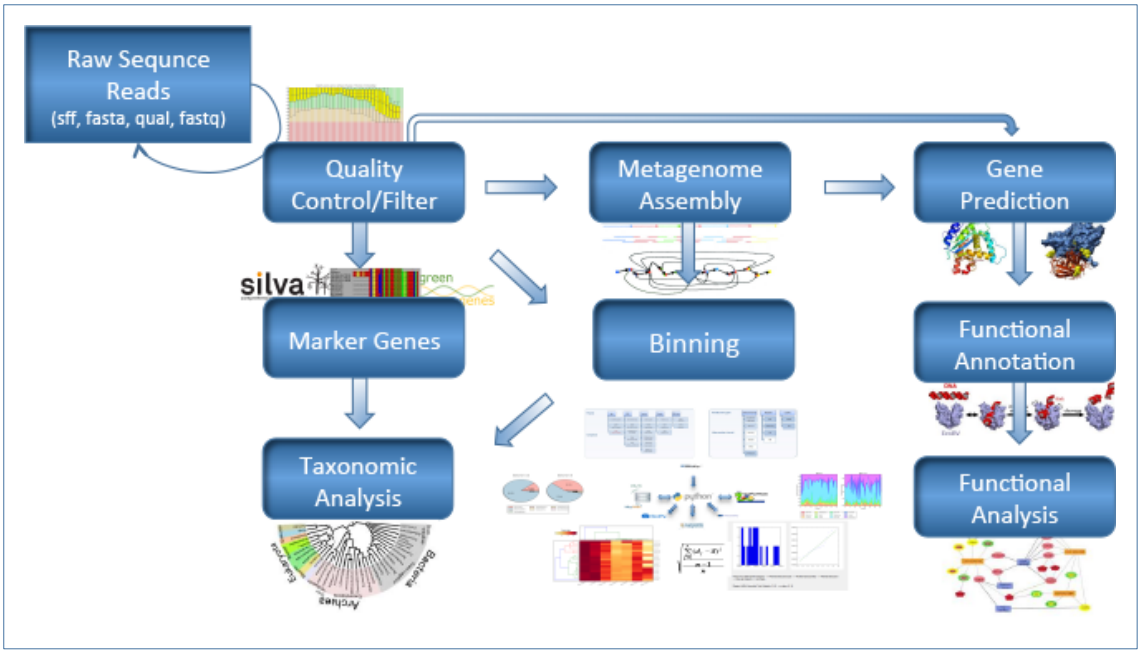

Figure 2: Bioinformatics analysis pipeline of metagenomics data. The sequence sequence reads generated by NGS technologies are initially processed through several quality control and filtering steps. The quality filtered reads are then annotated by binning the reads through mapping to reference genomes or by nucleotide composition-based algorithms. Functional annotation is performed by gene prediction and searching similarities to sequence databases. In cases when the sequence data represents a low complexity microbiota and sufficient sequencing depth is reached, contigs are formed by assembling reads after the quality control and filtering step. The annotated reads are finally subjected to taxonomic or functional metagenomic analyses.

4.3. BIOINFORMATICS

The term bioinformatics was coined in the early 1970s by the Dutch scientists Paulien Hogeweg and Ben Hasper on the seeds that the epochal discoveries of DNA and protein structures had sawn. The initial aim for the usage of the term was "to study informatic processes in biotic systems" (Hogeweg, 2011). Along with the discovery of the DNA structure by Watson and Crick, the major credit for the foundations of the contemporary bioinformatics advances should probably go to the double Nobel Prize winning researcher Frederick Sanger who first published the peptide sequence of insulin in 1955 and established a methodology to scale DNA sequencing projects in 1978 with the method entitled to his name (Sanger sequencing; Sanger and Coulson, 1978). As the protein sequence generation precedes the DNA, the initial efforts to make sense out of genetic information, was focused on protein sequences (Attwood et al., 2011). Margaret Dayhoff was the first researcher to collect protein sequence information and analyze it, which led her to the following conclusion: "There is a tremendous amount of information regarding the evolutionary history and biochemical function implicit in each sequence and the number of known sequences is growing explosively. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it" (Dayhoff, 1967). The latter sentence basically defines the major aim in the research field of bioinformatics.

The initial efforts in DNA and protein sequence data generation brought along the

immediate need for an organized storage and access to the accumulated sequence data. The Protein Data Bank (Bernstein et al., 1977) and Genbank (Benson et al., 2008) databases were curated to suffice this prerequisite of the nascent discipline. The discipline itself ultimately grew organically from the needs of researchers to access and analyze (primarily biomedical) data, which appeared to be accumulating at alarming rates simultaneously in different parts of the world. Advances in computer science and more specifically sequence alignment algorithms allowed the scientists to process, analyze, and store the growing numbers of sequences in the initial databases. From an information technology perspective, therefore, bioinformatics can be defined as a scientific discipline encompassing acquisition, storage, processing, analysis, interpretation and visualization of biological information (Singh, 2015c).

The core aims of bioinformatics include 1) the providing of access to existing biological information and submitting new entries as new data is generated, and 2) the development of software tools and other such resources that aid in the exploration of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterized sequences. 3) The third aim is to use these tools to analyze generated data and interpret the results in a biologically meaningful manner. The bioinformatics methodologies below are approached from a metagenomics analysis perspective and will cover all analysis steps taken in the metagenomic analysis pipeline of Paper III as illustrated in Figure 2.

4.3.1. DATA GENERATION AND ANALYSIS

Once the initial bioinformatics steps including data acquisition, organization and storage are completed, the data-driven phase of bioinformatics begins. Although there are multiple other sources of molecular biological data to be acquired from biological systems including gene expression, proteomics, and metabolomics, the extent of this PhD thesis will cover the information derived solely from DNA sequences and the inferences made from them. Consequently, the following step of identifying a fragment of DNA of unknown origin is to scan the sequences for open reading frames and predict genes. In this thesis, assembly of short reads to contigs was not considered within the metagenomics analysis pipeline due to 1) the relatively long read lengths generated by the 454 technology (avg. 350 bp), which allowed us to perform taxonomic and functional annotation of a substantial portion of the sequence reads, and 2) as expected from the previously asserted large species diversity of the marine biofilms, depth of sequencing was insufficient, and thus will not be discussed here.

4.3.1.1. GENE PREDICTION

Identification (or prediction) of open reading frames (ORFs) and genes is the initial step for the annotation of raw DNA sequences in a metagenomic analysis pipeline. ORF or gene prediction can be performed by at least three different ways. First, the start and stop codons can be scanned for throughout the whole set of DNA sequences, and any coding regions can be extracted by the algorithms based on decision trees. The second approach utilizes hidden Markov models (HMMs) in gene prediction and was implemented in several software tools including Glimmer (Delcher et al., 2007), Metagene (Noguchi et al., 2006) and FragGeneScan (Rho et al., 2010). FragGeneScan was used in the gene prediction step of the analysis pipeline in Paper III due to its accuracy in metagenomics applications. Finally, gene prediction can be performed by searching public sequence databases for genes. For example, the SSU rRNA identification tool used in Paper III, namely Metaxa (Bengtsson et al., 2011),

uses a combination of HMM and sequence database searches to extract SSU rRNA sequences from metagenomic datasets.

4.3.1.2. SEQUENCE ALIGNMENT

As mentioned among the approaches to gene prediction in the previous section, searching DNA sequences in databases requires the alignment of sequences to the individual sequence entries in the corresponding databases. An alignment is simply the pairwise comparison of individual characters of the query and target sequences. Evolutionary processes including insertions, deletions and substitutions are taken into account by introducing bonuses and penalties for nucleotide matches, gaps and mismatches in pairwise sequence alignments. In addition, scoring matrices are built to evaluate the probability of the alignment of two nucleotides. Alignments can be performed to cover the entirety of the query and the target sequences by global alignment algorithms. Alternatively shorter but more accurate similarities can be searched within the subsets of the two sequences by local alignment algorithms.

The sequence alignment problem becomes computationally expensive for the character-by-character scoring of nucleotide pairs as the length of sequences increase. Dynamic programming was initially introduced to solve this problem for the global and local alignment algorithms, named after the researchers Needleman & Wunsch and Smith & Waterman, for the individual algorithm, respectively (Needleman and Wunsch, 1970; Smith and Waterman, 1981). Although the pairwise alignment problem was overcome within a reasonable time through the use of dynamic programming based algorithms, searching for the ever-increasing number of sequences in the archives one by one requires certain heuristics and indexing schemes to reduce the time spent for a database search. Similarity searches in databases are performed largely by algorithms that do not guarantee that the best match is found, but rather report the most probable alignment of the query sequence to the individual database entries (Bansal, 2005). Indices are the auxiliary data structures that these algorithms produce from either of the query sequences or database entries or in some cases both. Alignment algorithms are grouped into three categories according to the utilized indexing scheme, namely hash-table- (also know as word), suffix tree- or merge sorting-based algorithms (Li and Homer, 2010). The most commonly used database search algorithm, BLAST (Altschul et al., 1997), is essentially a hash-table based method, relying on the scanning of k-mer sized subsequences of the query in the database by using a hash table whose keys are the k-mer sequences. The initial contiguous match in the target sequence, called the seed, is extended and joined without gaps in BLAST and afterwards refined by a Smith-Waterman algorithm in the following step. It finally reports the statistically significant local alignments in the output. A suffix tree is a data structure that stores all the suffixes of a string, enabling fast string matching. The fast read mapping algorithms utilized in the alignment steps of most NGS sequence alignment software are based on Burrows-Wheeler Transform (BWT)-based algorithms that utilize suffix trees (Burrows and Wheeler, 1994). In Paper I, the taxonomic count data was retrieved by mapping human gut metagenome reads to a database of whole microbial genomes by using the tool Bowtie (Langmead et al., 2009), which uses a BWT-based algorithm to map the reads to individual genomes.

4.3.1.3. PHYLOGENETIC ANALYSIS

In addition to searching databases for unknown sequences, alignments are also useful to learn about phylogenetic relationships between sequences. Multiple sequence alignments are constructed for this purpose by stacking orthologous sequences from different species. Thereby, reasonably conserved regions from orthologous sequences are extracted, and finally an evolutionary tree is inferred based on the multiple sequence alignments. Phylogenetics is the study of the evolutionary relationships between genes, individuals, populations or species, and constitutes the foundational methodology of a larger study field called systematics. The term phylogeny refers to the relationship among the aforementioned biological organization units that depicts a common ancestry between the units at a particular time point in evolution (Krane and Raymer, 2003). A phylogenetic tree is the typical graphical representation of these evolutionary relationships and is composed of various arrangements of nodes and branches. Taxonomy refers to the naming and classification of the nodes from the tip of the phylogenetic tree to the root node. Phylogenetic trees usually convey three distinct sources of information about the evolutionary history of the biological units including relatedness, degree of divergence and the taxonomic affiliation of their common ancestor (Krane and Raymer, 2003). Constructing multiple sequence alignments is the preliminary step for phylogenetic analyses. The following step of inferring phylogenetic trees can be performed by several alternative methodologies including Unweighted-Pair-Group Method with Arithmetic Mean (UPGMA)-, neighbor joining-, maximum likelihood (ML)- and maximum parsimony-based methods (Day and Edelsbrunner, 1984; Hyde et al., 2013). A ML-based tree inference tool, MLTreeMap (Stark et al., 2010), was used for the reconstruction of the phylogenetic tree shown in Figure S1, in Paper III.

4.3.2. FUNCTIONAL ANNOTATION AND BIOLOGICAL DATABASES

Biological databases are essential tools for the researchers in a myriad of fields in biology from protein structural studies to the investigation of the types of chemicals released to the environment. As pointed out at the beginning of the *Bioinformatics* section, the initial concerted efforts to organize accumulated biological information were the establishment of PDB for the protein structural data and Genbank database for nucleic acid sequences. These databases are attributed to be archival or primary databases, meaning that the experimental results are initially stored in these databases with limited interpretation or annotation provided by the submitting researcher (Singh, 2015a). There are also secondary databases, which mostly use the data in the archival ones but refine the information through an established curation strategy relying on the paradigm that the database attempts to convey to the researchers of particular, specialized fields of biology. Examples of these specialized or secondary databases include databases storing information of genomics and sequence data of whole genomes, protein mutations and polymorphism, 2D gel- and mass spectrometry-based proteomics data, metabolic reaction and pathway data or more specialized information related to genes, proteins, enzymes, protein complexes and biochemical pathways such as antibiotic resistance genes, membrane transport proteins, and carbohydrate-degrading enzymes. There are currently 1,685 different biological databases listed in the Molecular Biology Database Collection (Rigden et al., 2016).

4.3.2.1. REFSEQ

The often limited annotations provided for the archival database entries are in fact prone to ambiguity during the time of the actual database usage by the sequence alignment tools or manual entry retrieval processes by the users of the database. Due to the reasons including duplicate entries for the same biological entity, lack of information to reinforce reliability and consistency to clarify the multiple entries and uncertainties regarding the source of the database entry (*e.g.* experimental or *in silico*), the archival databases indeed attenuate the sequence read annotation process and may risk obstructing the downstream bioinformatics analyses (Singh, 2015a). The RefSeq database project, ran by the National Center for Biotechnology Information (NCBI), was introduced to address these issues and to constitute a reliable and consistent biological data resource for individual level of molecular information in the central dogma, namely DNA, RNA and protein (Pruitt et al., 2007). The non-redundant data representation for biological entities provided by the NCBI databases was utilized in the sequence annotation step of Paper III by the alignment of metagenomic sequences to the non-redundant protein and the non-redundant nucleotide databases.

4.3.2.2. KEGG

Another intensely utilized biological data resource throughout this PhD thesis is KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Kanehisa and Goto, 2000). The KEGG database is a collection of several databases providing information at varying levels of biological functional organization including molecular-level, pathway and broader-level functions. Molecular level information is provided within the KEGG Orthology (KO) database in which individual molecular data entries are associated with orthologous proteins and enzymes annotated from gene catalogs of complete genomes. The KO database is a secondary database constructed by restructuring the database entries provided in RefSeq, Genbank and other publicly available data resources. Molecular level information is further organized within the broader level functional categories of the KEGG Pathway database and the KEGG Brite hierarchies (Kanehisa et al., 2008). The KEGG databases enable bioinformatics researchers to systematically conceptualize gene functions and to link gene-level information with a higher order functional categorization represented by interconnected entries at different hierarchy levels. The interconnections between different categories and the functional hierarchy information provided in the KEGG databases are utilized in Paper II. Paper II presents a software tool developed as part of this PhD thesis, namely PACFM, providing a comprehensive functional visualization tool with additional utilities to explore and manipulate pathway data annotated by the KEGG databases.

4.3.2.3. GENE ONTOLOGY

Another systematic effort to organize functional biological data under the "controlled vocabularies (ontologies)" of molecular functions, biological processes and cellular components is the Gene Ontology (GO) database (Ashburner et al., 2000). The GO database involves biological entities called the GO terms, structured as a directed acyclic graph (DAG). In this data structure, specialized database entities may have multiple broader categories that they partake in, which means a child node in the database may have multiple parent nodes unlike a taxonomy tree, in which each child node has only one parent. For example, the biological process term hexose biosynthesis has two parents, hexose metabolism

and monosaccharide biosynthesis. The DAG structure of GO (which is also utilized in the curation of the KEGG Brite database) constitutes a problem when assigning broad level functions to metagenomic sequence reads. The molecular level annotations are associated with multiple pathways (broader level functions) and several of the associated pathways end up among the significant analysis results even though they are absent in the investigated environmental microbiome. This problem is addressed in Paper II, and the bioinformatics software PACFM was designed to provide several options for tackling these pitfalls.

4.3.2.4. PFAM

Multiple sequence alignments are explained in the previous sections where DNA or amino acid sequences are stacked within a single alignment. Visual inspection of MSAs often reveals conserved regions throughout the entirety of an MSA. There are more sophisticated methods developed to discover and extract these conserved regions from MSAs. One such methodology is to model the conserved regions of protein sequences by Hidden Markov Models (HMMs). Pfam or Protein families database is a collection of around 16,000 conserved regions of proteins modeled and stored as distinct HMMs (Finn, 2012). Pfam models can be considered as short amino acid sequence patterns, which exhibit conserved functional or structural units of proteins. Hence, while the previously detailed databases such as KEGG and GO organize biochemical functions into broader level groups than individual enzymatic functions and protein features, Pfams model sub-protein level biological information such as protein families. The software tool HMMER (v.2; Eddy, 1998) was used in Paper III to assign Pfam annotations to individual sequence reads from the metagenome of the marine biofilms.

Apart from the databases listed above, several others including the taxonomy databases NCBI Taxonomy (Eddy, 1998) and SILVA (Quast et al., 2013) as well as protein functional databases such as UniProt Knowledgebase (Apweiler et al., 2004), COG (Tatusov et al., 2000), TIGRFAMs (Haft et al., 2003) and CAZy (Haft et al., 2003) were used in the software development or data analysis steps of the Papers I and III.

4.4. COMMUNITY ECOTOXICOLOGY

Ecotoxicology, by definition, is the science of contaminants in the biosphere and their effects on the constituents of the biosphere, including humans (Newman and Unger, 2003). The emergence of the discipline originates from the investigation of the effects of specific abiotic factors, namely toxic chemicals released by humans to the environment. Therefore, ecotoxicology is largely influenced by the paradigms stemming from ecology. As the ecology covers a wide range of levels in the biological organization from individual species to the ecosystem level interactions, so does the ecotoxicology research by differentiating into subfields for certain biological organization levels. The use of economically important or charismatic species to investigate toxicant responses in organisms has been a trend among researchers at the level of organismal physiology or biomolecular level (Handy and Depledge, 1999; Gunnarsson et al., 2009; Celander, 2011) Researchers trained in ecology have implemented their paradigms to introduce plausible explanations for the effects of toxicants on the organisms at the community and ecosystem levels of biological organization (Blanck, 2002; Backhaus et al., 2008; Blanck et al., 2009). Although researchers working with the two different aspects of ecotoxicology with reasonably distinctive paradigms tend to approach to

the environmental toxicity problem in a heterogeneous manner, the ultimate aim of the field of community ecotoxicology is to integrate the mechanistic explanation of the lower biological hierarchy levels with the responses relying on the emergent characteristics of populations and communities (Newman and Clements, 2007).

4.4.1. FIELD SAMPLING

The microbial biofilm communities sampled in Paper III were allowed to colonize and grow on rectangular glass slides (150 mm × 20 mm) at 1.5 m depth at five sampling sites at the mouth area of the Gullmar fjord on the Swedish west coast. The samples represent a relatively small coastal territory within the fjord, the most distant sites being 11 km apart, from the inner to the outer archipelago of the Swedish west coast. A combined sample of four distinct time points in a year from the outer archipelago was also sent for sequencing as part of the study. For each site, one sample, corresponding to a surface area of approximately 100 cm^2, was taken. Individual sites were different in terms of water depth, bottom characteristics and distance to land. Water movement from currents, tide and weather-dependent high and low waters were expected to connect each site and allow them to share drifted microbiota. Five of the samples to be sequenced were from the sampling day of 23rd of July 2004. The sixth sample was obtained by pooling equal amounts of DNA from the sampling occasions on 28th of April, 23rd of July, 30th of August and 21st of September, 2004 respectively.

4.4.2. FLOW-THROUGH MICROCOSM (AQUARIA) EXPERIMENTS

The long-term exposure to triclosan in Paper IV was performed using a flow-through microcosm system located at the Sven-Loven Center for Marine Sciences at Kristineberg, Sweden. The flow through microcosm system allows the settlement of new microbiota in a continuous fashion with the constant inflow of seawater in addition to the predetermined concentrations of triclosan exposure. Seawater is pumped from the Gullmar fjord into the laboratory through a nylon mesh of 1 mm filter size before being distributed into 25 different glass aquaria including the controls and replicate treatments through the triclosan exposure gradient. The incoming water had an approximate flow rate of 220 mL/minute in the individual microcosms. The toxicant solutions were refreshed every third day by preparing in deionized water and were injected to the microcosms with respect to the experimental design procedure according to which the exposure gradient was set between 0.316 nM and 1000 nM concentration of triclosan. The microbial biofilms were let to be established on glass discs mounted inside polyethylene racks. The flow-through microcosms were let to run for 18 days, considered as a long-term exposure period, and the established microbial biofilms in each aquarium were collected and sampled subsequently.

5. RESULTS AND DISCUSSION

5.1. PAPER I

One of the outstanding strengths of metagenomics as a molecular gauge, is the researchers' ability to address taxonomic or functional analysis aspects of microbial ecology directly from a single data source. Researchers in metagenomics have the flexibility to approach DNA sequence data analysis by separating the microbiota from the biochemical capacity of the sampled environment or merge the two information sources and interpret them jointly. Since its inception, the traditional methodology to present results of metagenomic studies has been performed by dividing the results into individual distinct sections of taxonomy and functional analysis interpretations. This tradition may boil down to the inadequacy of interdisciplinary formulation portrayed at the beginning of this thesis, as such the field of metagenomics largely originates from microbial ecology, and microbial ecologists often adopt ecological paradigms in their research, without considering the necessity to link community level inferences to the lower levels of biological organization such as the biochemical functions inside cells. This dichotomy of treating metagenomics data has largely been the prevalent approach to the analysis of DNA sequence reads extracted from environmental samples. Paper I presents FANTOM, an open-source software tool that stems from the need for an easy-to-use tool to explore the often complex metagenomics data, which exhibit the aforementioned dichotomy of the taxonomic and functional aspects of microbial ecology research.

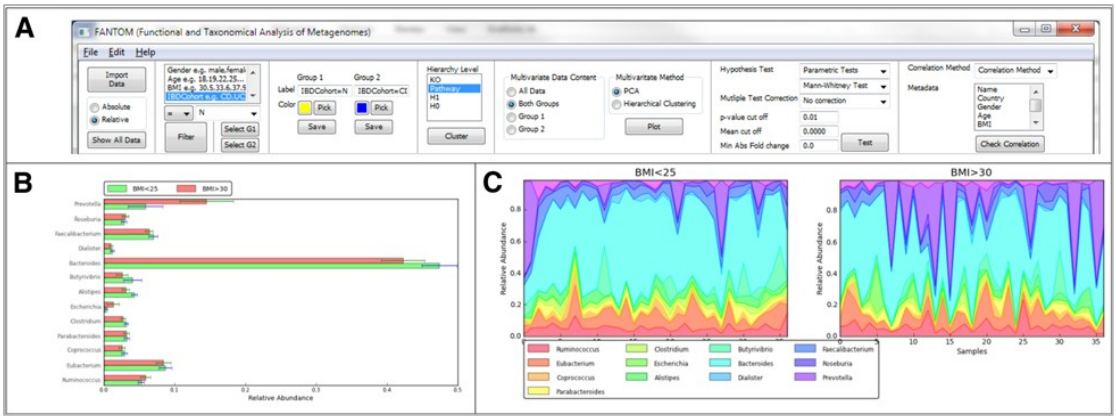

Figure 3. Graphical user interface (GUI) of FANTOM and comparative analysis plots generated through the GUI. A. Data selection/filtering and statistical analysis panel B. Bar graph comparing two groups according to the pathway abundances. C. Area plots showing the fluctuating levels of functional categories in the samples of two different groups.

FANTOM was developed as a stand-alone, open source, and graphical user interface (GUI)-based (see Figure 3) software tool. The limitations of the existing statistical analysis and visualization tools for metagenomics data analysis and the reluctance of users to upload their data to online services set the major motivations for FANTOM to be developed. FANTOM allows users to analyze metagenomics data in connection with NCBI taxonomy, KEGG, COG, PFAM and TIGRFAM databases. It features exploratory and comparative analyses of metagenomics data integrated with individual sample metadata for sophisticated statistical analyses such as principal component analysis or several options for hypothesis testing. The software tool was used to reveal significant analysis results of a comprehensive

human gut metagenome data covering healthy, obese and individuals associated with several diseases of the human gastrointestinal tract including inflammatory bowel disease, Crohn's disease and ulcerative colitis (Qin et al., 2010). One of the most striking results found by using FANTOM on this human gut metagenome data was the detection of a significant deficiency of the archaeal species *Methanobrevibacter* sp. in the gastrointestinal tracts of Crohn's disease patients. FANTOM was also used in Paper III where the metagenomic sequence reads from the marine biofilm samples were assigned to various taxonomic levels and *Proteobacteria, Bacteroidetes*, and *Cyanobacteria* were found to dominate the marine biofilms. In conclusion, an open source, standalone and user-friendly software tool, for data analysis and data mining of shotgun metagenomics studies has been introduced to the research community.

5.2. PAPER II

Through the door that FANTOM opened in the analysis of metagenomics data, a second software tool called PACFM was developed and presented in Paper II. PACFM utilizes the hierarchical data structure of the KEGG Brite database for providing a novel graph in pathway visualization. It also addresses a commonly observed problem in the analysis of functional metagenomics data where a gene is counted multiple times as belonging to different pathways, and several of the associated pathways may end up among the abundant functional categories in the analysis results even though they may not exist in the investigated environment. Naïve pathway abundance calculation as stated in previous literature sources (Sharon et al., 2011), may obstruct the biological inferences made about the microbiota in complex metagenomics samples where neither the actual microbial diversity nor the full capacity of biochemical functions are known. PACFM introduces the concept of pathway associations of enzymes in order to distinguish generically used enzymes of a biochemical pathway from the unique ones that determine the rate-limiting steps of the associated pathway. In order to do so, the tool takes full advantage of the flexible visualization features of another widely used tool, Circos (Krzywinski et al., 2009), which was originally developed to illustrate chromosomal maps. PACFM draws the conceptualization of functional metagenomics data up to a further level by providing seven distinct sources of information regarding the biochemical potential of a metagenomic sample as follows: KEGG Brite categories at 1) the top level, 2) second level, 3) third (pathway) level, 4) database hierarchy information, 5) a manually curated database subset, 6) pathway associations of individual enzymes, and 7) the key/rate-limiting enzyme information (see Figure 4).

In addition to providing the metagenomics researchers with an improved way of visualizing pathway abundance data, PACFM also presents a wide array of methods for filtering and normalizing the abundance counts annotated by the individual KEGG Orthology identifier. The software tool was shown in Paper II to uncover novel results in previously published studies including an open ocean depth profiling (DeLong et al., 2006) and a human gut metagenomics study of the obese and lean twins (Turnbaugh et al., 2009) that may shed new light on the biological interpretations of the microbial traits in these environments. As an example, although there are several manual interference options to amend functional annotations in PACFM, an automatized run eliminated all of the human related disease categories in KEGG from the results of significant differences between lean and obese individuals. By setting the pathway association cutoff for individual enzymes of pathways to 1, 2 and 3 we observed that the otherwise abundantly reported methane metabolism pathway

in a sample from the study of an open ocean depth profile, might indeed not be found in reality. The pathway visualization feature of PACFM was also utilized in Paper III, illustrating the full biochemical capacity of marine biofilm samples in one plot. PACFM serves researchers in metagenomics with an easy to use, standalone and integrated pathway visualization software tool and also allows them to easily elaborate on the pathway analysis results through automated and manual manipulation means, with the overall goal to help users to avoid misleading biological inferences.

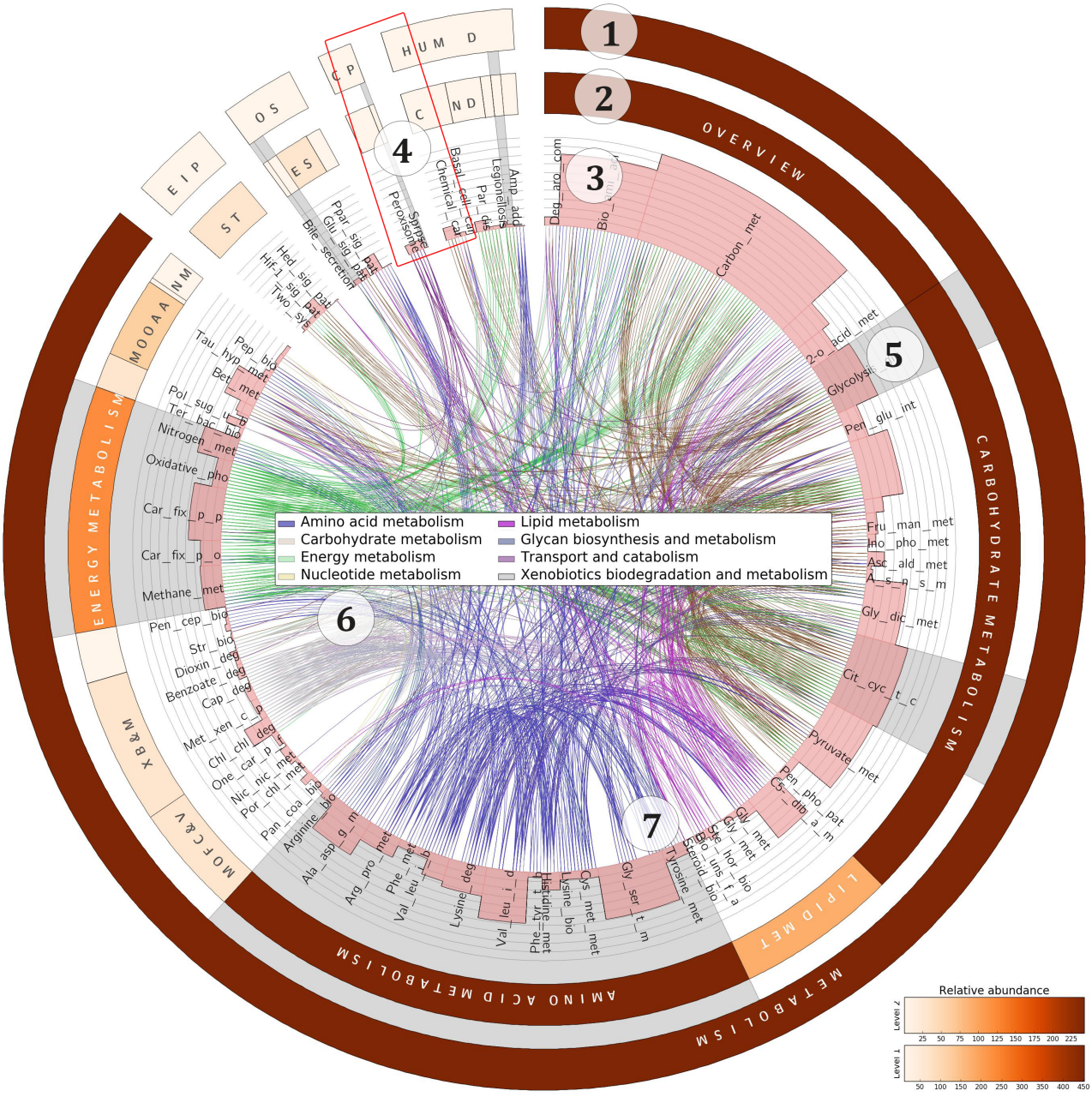

Figure 4. Final output plot of PACFM. Abundance data represented at 1) the top level, 2) second level, 3) third (pathway) level KEGG hierarchies as well as 4) database hierarchy information, 5) a manually curated database subset, 6) pathway associations of individual enzymes, and 7) the key/rate-limiting enzyme information are shown.

5.3. PAPER III

The aquatic biofilms, also known as periphyton, constitute an unprecedented tool for microbial ecologists, environmental scientists and ecotoxicologists for investigating the aquatic microbial life due to several advantages. First, they represent a micro-ecosystem that is attached on a surface and open to all sorts of ecological interactions with the ambient environment. The multispecies nature of these biofilm communities allows the researchers to adopt traditional ecological paradigms to apply on this micro-ecosystem. The communities harbor a large diversity of organisms comprising viruses, bacteria, algae, fungi, protozoans, and metazoans. By accommodating phototrophic (from which the name periphyton originates) and heterotrophic members from different levels of the food web (Figure 1), the biofilms represent an ecologically relevant model system, which enables the possibility of controlled experimentation at the microbial community level. This is a giant leap on top of the traditional microbiological experimentation based on culturing of single microbial species. However, until the community-genome shotgun sequencing based exploratory study was carried out in Paper III, the biodiversity and functional capacity residing in the marine biofilms was poorly described. In Paper III, we used metagenomics to characterize marine biofilm communities from the sampling sites described in Figure 5 at the Swedish west coast.

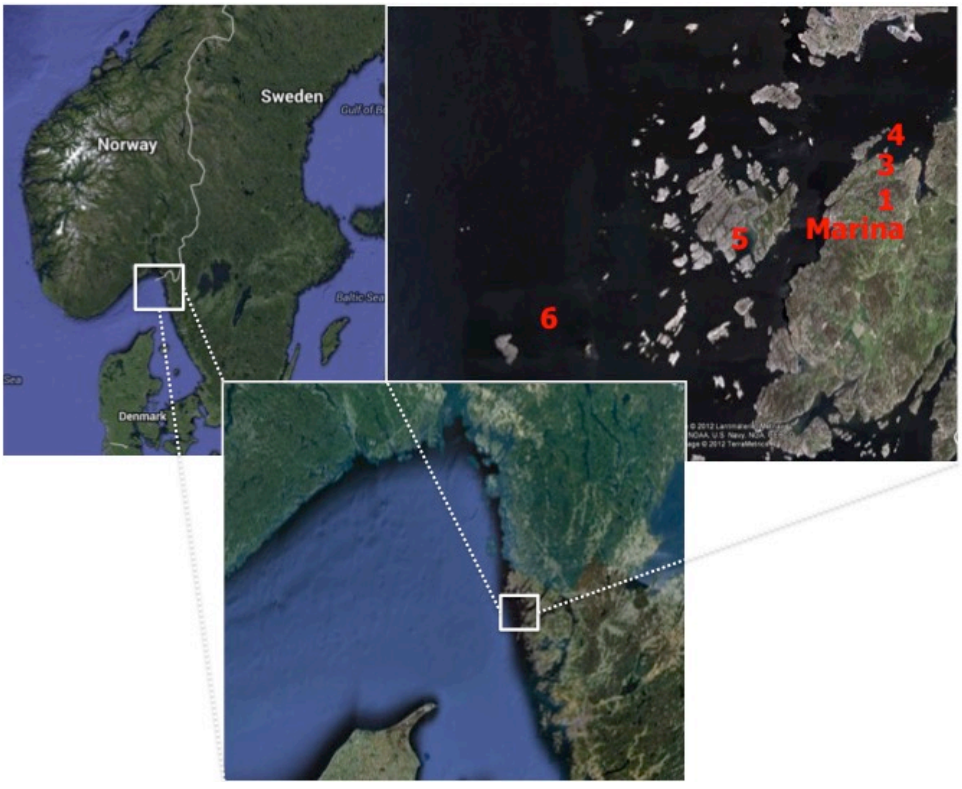

Figure 5. Sampling sites of the marine biofilms in Paper III. The samples represent a relatively small coastal territory within Gullmar fjord on the Swedish west coast. The most distant sites are 11 km apart, from the inner to the outer archipelago of the Swedish west coast.

The study design of Paper III in terms of metagenomic sequencing differs from the traditional methodology applied to the planktonic microorganisms by filtrating water samples through micrometer size filters. Despite the numerous advantages of biofilms for microbial ecological experimentation, the extracellular polymeric matrix where the organisms are encased, obstructs the separation of eukaryotes from the bacterial community members during

the sampling of microbiota. Therefore, taxonomic affiliation of the sequence reads was done entirely through bioinformatics methodologies. By extracting the ribosomal small subunit gene sequences from the metagenomic dataset, we found approximately ten times more eukaryotic rRNA sequences compared to prokaryotic genes. However the whole metagenome-based similarity searches showed that bacterial phyla including *Proteobacteria*, *Bacteroidetes* and *Cyanobacteria* are the most abundant organisms in these biofilms (Figure 6). Intriguingly, the DNA sequences belonging to marine invertebrates were highly abundant in the biofilms, confirming the literature regarding the inhabiting of the biofilm matrix by invertebrate larvae until their developmental stage of metamorphosis (Hadfield and Paul, 2001).

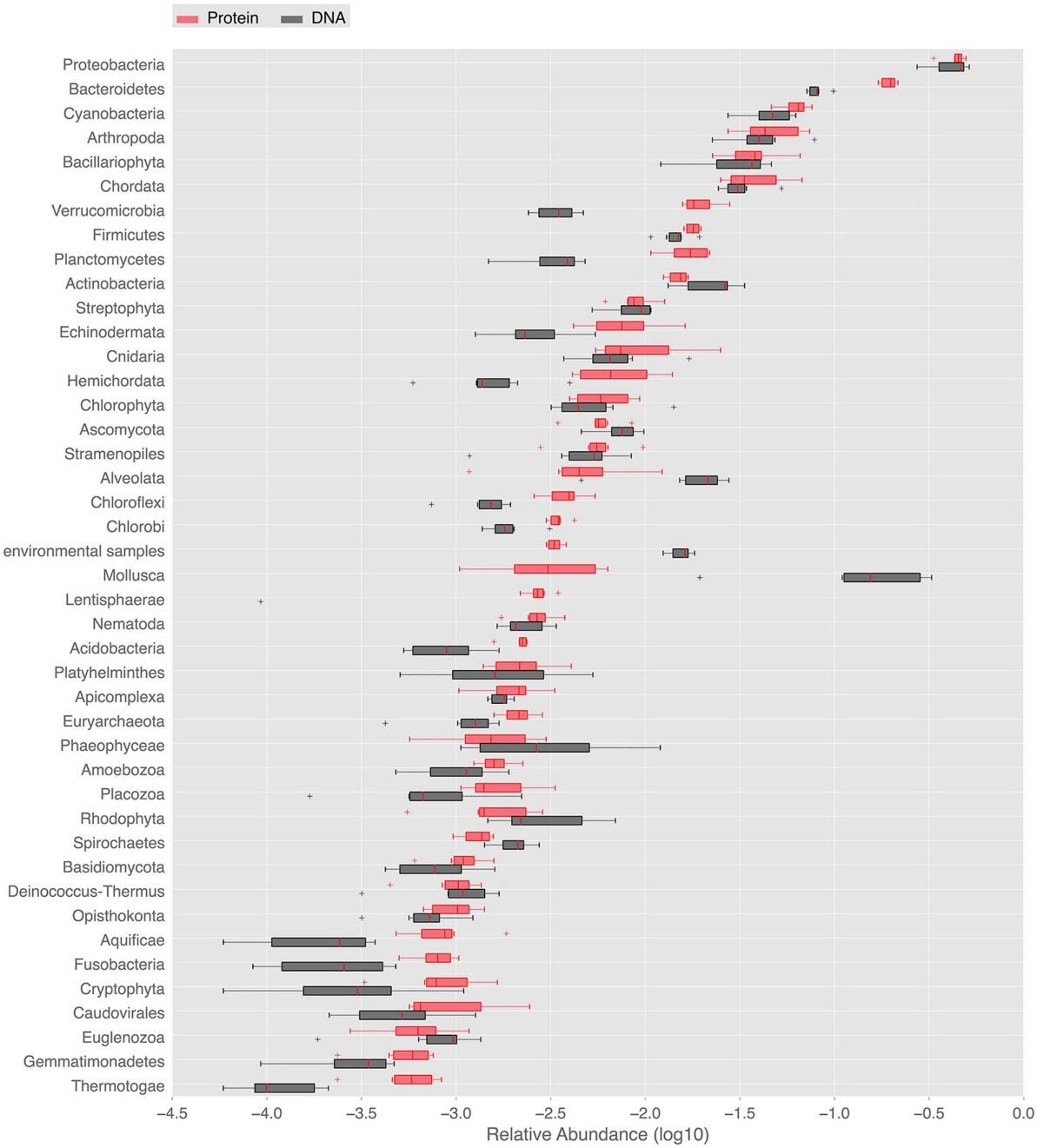

Figure 6. Ranked abundances of phyla identified in the periphyton metagenome. Relative abundances of phyla found in periphyton according to whole metagenome similarity searches to public databases.

In the biofilm metagenome, we detected phototrophic members such as *Cyanobacteria*, the alpha-proteobacterial genus *Roseobacter,* micro- and macro-algae as highly abundant. We also assessed the metabolic pathways that predispose these communities to an attached lifestyle. Functional indicators of the biofilm form of life in periphyton involved genes coding for enzymes that catalyze the production and degradation of extracellular polymeric substances, mainly in the form of complex sugars such as starch and glycogen-like meshes together with chitin. Genes coding for 278 different transporter proteins were detected in the metagenome, constituting the most abundant protein complexes. Our finding of genes encoding for enzymes that participate in anaerobic respiratory pathways, such as denitrification and methanogenesis hints at the presence of anaerobic or low-oxygen micro-zones within the micro-ecosystem formed by the marine biofilms.

5.4. PAPER IV

With the toolset provided by metagenomic sequencing and the established analysis software, the implementation of controlled experimental settings into microbial ecology and ecotoxicology is aimed in Paper IV. We used amplicon sequencing of taxonomic markers, such as 16S and 18S rRNA genes to improve our ability to study the effects of an antimicrobial compound, namely triclosan on the composition of environmental biofilm communities. Triclosan has become a ubiquitous contaminant in many environmental compartments due to the large-scale antimicrobial consumption (Halden and Paull, 2005; SCCS, 2010), and environmental risks from triclosan have been identified for a wide range of non-target organisms (Johansson et al., 2014). A flow-through microcosm experiment was performed over 18 days with marine biofilms exposed to a concentration gradient of triclosan ranging from 0.316 nM to 1000 nM. We showed that triclosan exposure causes shifts in bacterial and eukaryotic community composition, not only expressed by an overall decrease in microbial diversity, but also validating previous findings on the algal sensitivity and metazoan tolerance to the antimicrobial. We further found an increased relative abundance of a sulfur-based syntrophic consortium at high triclosan exposure, including the taxa *Desulfobacterales, Thiotrichales*, *Campylobacterales* and *Chromatiales*.

In contrast to the eradication of other algal groups from the biofilms at high-level triclosan concentrations, the increase in the relative abundance of red algae inspired the investigation of the differences between the distinct groups of algae. In alignment with the increased relative abundance of the sulfur-based microbial consortium, one of the distinguishing structural components of red algae was found to be the density of sulfated-polysaccharides within their cell wall structure (Hernández-Sebastià et al., 2008). Triclosan has previously been shown to form a sulfated ether (Triclosan-O-Sulfate; Chen et al., 2015b), which hinted us at a potential detoxification mechanism provided by red algae against triclosan, through the sulfurylation reaction of triclosan and sulfate. Moreover, additional algal taxa were identified also in significantly increased abundances at the high triclosan exposure level, including *Phaeophyceae*, and *Ulvophyceae*. These algal groups also accommodate sulfated polysaccharides in their cell walls, which were previously shown to exhibit plant elicitor defence roles against pests, diseases or other organisms (Benhamou, 1996; Mercier et al., 2001; Hernández-Sebastià et al., 2008; Domozych et al., 2012). We hypothesized that the large supply of sulfate provided by the algal groups as well as echinoderms stimulates the growth of sulfate-reducing groups including the order of *Desulfobacterales* in the biofilm samples from the high triclosan exposure level. The sulfide produced by the sulfate-reducers are oxidized by the purple sulfur bacteria (*Chromatiales*)

and lithoheterotrophic members of the orders *Thiotrichales* and *Campylobacterales* forming a sulfur-based unique consortium not identified by the metagenomic sequencing analyses in Paper III.

6. CONCLUSIONS AND OUTLOOK

The foundation of a scientific discipline is consolidated by the pillars of controlled experimentation and the accuracy of the employed gauges, which allows reproducibility. Metagenomics is a young research field, which has yet not achieved the switch between its "hypothesis-generating" role into a "hypothesis-driven" one in the field of microbial ecology, partly due to the tunnel vision imposed and pigeonholed as the *de facto* route by the global scale exploratory endeavors. The phrases "stamp collecting" and "fishing expeditions" have previously been used to criticize the contribution of large-scale data analysis-based methodologies and research fields to the accumulation of biological knowledge (Hunter, 2006; Ning and Lo, 2010; Mouritsen, 2011). The consolidation of metagenomics as a robust instrument in biology and as a research field highly relies on amendments in two aspects: 1) adoption of the appropriate interdisciplinary methodology and 2) increasing the number of controlled experimental settings. The *Introduction* section emphasizes the significance of interdisciplinarity in metagenomics and proposes a methodology by dissecting and juxtaposing its constituent disciplines, which ultimately leads to the synthesis of novel biological knowledge.

From a methodological perspective, among the constituent disciplines of metagenomics, bioinformatics has established the strongest pillars since the field has emerged as virtually a natural result of NGS, much like the case with all other 'omics data-based research fields. The software tool FANTOM in Paper I filled a gap in the statistical analysis and visualization of comparative metagenomics data. There certainly are plenty of alternative methods elsewhere, that utilize similar analysis features as implemented in FANTOM. However, at the least, a beginner level programming experience and statistical analysis knowledge is required in most extant applications. For example, there is a large number of packages developed for the statistical analysis language, R that can be utilized for different aspects of metagenomics analyses (Oksanen et al., 2007; White et al., 2009; Kristiansson et al., 2009; Zhang et al., 2013; Love et al., 2014), and the Python programming language community has continuously provided software packages and libraries to fill in certain gaps in the 'omics analyses (Cock et al., 2009; Caporaso et al., 2010; Pedregosa et al., 2011). FANTOM is based on several statistical analysis libraries including numpy, scipy and matplotlib as well as the relational database management features of sqlite and the graphical user interface programming library wxPython. There are also alternative GUI-based software tools introduced prior to FANTOM. The MG-RAST server (Meyer et al., 2008), MEGAN (Huson et al., 2007) and STAMP (Parks et al., 2014) are only a few, which currently include the majority of features implemented in FANTOM. One of the major strengths of FANTOM is the ability to analyze data through various biological database resources. I would like to acknowledge the database curation efforts for unifying and conceptualizing functional and taxonomic data sources from various databases by cross-referencing them in the development teams of the M5NR database (Wilke et al., 2012) and the UniProt Knowledgebase (Magrane and Consortium, 2011). Ultimately, FANTOM and the other listed software tools and databases serve to conceptualize the large-scale biological data and provide sound biological inferences based on relevant statistics.

Conceptualization of molecular biology data relies on the added value that stem from expert knowledge refined into specialized databases. These specialized databases are of paramount importance for metagenomic research both for the identification and

characterization of organisms as in the example of taxonomic analyses and in order to group detected genes into broader biochemical organization levels as in functional metagenomics analyses. The KEGG databases (Kanehisa et al., 2002) are excellent resources to group detected genes into the so-called metabolic pathways or protein complexes. The curation strategy of the KEGG databases involves the retrieval of sequences from individual genomes stored in the RefSeq database (Pruitt et al., 2007) and organization of the functional information from those genomes into orthologous groups (KEGG Orthology database), pathway modules (KEGG Modules database), pathways (KEGG Pathways database), or broader categories grouped into distinct hierarchies (KEGG Brite database). The hierarchical annotation of metagenomic data is an indispensible way of approaching functional information coded in the microbiome of distinct environments in order to perform top-down and bottom-up biochemical examinations. However, since the KEGG databases do not distinguish between different genomes according to the environmental biomes that individual organisms originate from, during their curation processes, the annotation of metagenomic reads by the KEGG databases result in grouping of the detected genes in the microbiome of a specific environment into broader functional categories that may or may not be present in the investigated environmental context. The latter case leads to incorrect interpretation and inferences about the biochemical capacity of the inspected microbiota. The software tool presented in Paper II, PACFM, both provides a novel graphical representation reflecting the hierarchical structure of the KEGG Brite database and allows researchers to manually inspect or automatically eliminate pathways, which do likely not exist in the investigated environment. This in turn helps the researchers improve their biological data conceptualization capability during functional metagenomics analyses.

Once the toolset for the bioinformatics analyses in metagenomics was ready to be applied, the exploratory metagenomics analyses in Paper III paid back with valuable insights on the research of marine biofilms. Shotgun metagenomic sequencing of the marine biofilms revealed the potential for low-oxygen micro-zones within the biofilm matrix through the inferences made upon the pathways belonging to energy metabolism. The abundance of ABC transporters in the metagenome indicates the intensity of the ongoing metabolite transport within the biofilm interior. The previously attributed phototrophic nature of the marine biofilms was shown to be inaccurate through the detection of abundant heterotrophic groups. The previously asserted name to represent the multi-species marine biofilms, periphyton (Sand-Jensen and Borum, 1991; Rosemond et al., 1993), should therefore be used with caution in future studies. The presence of viral genes was slightly touched upon in our efforts; thus, investigation of the diversity and function of viruses in the marine biofilms is a topic to be dug further into. The results of Pfam database annotations hinted towards the high abundance of reverse transcriptase genes (Pfam description: *Reverse transcriptase: RNA-dependent DNA polymerase*) in the biofilm metagenome, which can be of viral origin. Furthermore, the proposed abundance of extracellular DNA in the biofilm structure according to previous literature on biofilms, could not be examined in our study due to the difficulties associated with the separation of DNA found in the EPS matrix from intracellular DNA.

The first three papers in this PhD thesis mainly involve the traditional way of applying metagenomics paradigms into research, be it software development or exploratory investigation of environmental samples (albeit the metagenomics analysis of multi-species biofilms in Paper III is among the rare attempts to investigate the microorganisms encased in a biopolymeric matrix). Paper IV is an attempt of long-term perturbation of a microbial

ecosystem through a controlled experimental setting via an antimicrobial agent. In contrast to the traditional metagenomics methodology of sampling and sequencing pre-established communities in their natural environments, the biofilms of this study are let to grow on clean glass discs for three weeks under exposure to the antimicrobial agent triclosan. Such controlled experimentation of microbial communities has been performed by ecotoxicologists since the 1980s in order to assess the risks of chemical compounds released to the environment. However, previous studies have been based on methodologies yielding endpoints of low-resolution snapshots of the actual microbial community *via* microscopy (Dahl and Blanck, 1996; Devilla et al., 2005) or based on qualitative measures such as mortality (McPeek and Peckarsky, 1998; Oberdörster et al., 2006). Although there have been quantitative measures applied for the risk assessment of natural microbial communities through microbial activities including carbon utilization or photosynthesis (Blanck et al., 1988; Bonilla et al., 1998; Eriksson et al., 2009), these methods represent the activities exhibited by a subset of the entire community. In Paper IV, a laborious experiment was performed by the involvement of a large group of researchers, and the effects of triclosan on the biofilm community members was investigated through the DNA amplicon sequencing of 16S and 18S rRNA genes, theoretically covering the entirety of the microbial biofilm community. The results not only showed the anticipated change in the community composition, but also hinted towards the mechanisms that the biofilm community members employed to cope with the exposed antimicrobial agent. Although, further investigation of the biofilm microbiome at different exposure concentrations through the utilization of shotgun metagenomic sequencing may validate the findings in Paper IV and reveal even a broader array of detoxification and antimicrobial resistance mechanisms, the results of the amplicon sequencing study proved the methodology to be a *de facto* step in future employment of DNA sequencing for ecotoxicological experimentation.

Individual steps of a typical metagenomics sampling and analysis pipeline, from metadata recording to sequence assembly and binning as well challenges associated with individual steps were previously covered in several excellent reviews (Hamady and Knight, 2009; Wooley et al., 2010; Gilbert and Dupont, 2011; Ju and Zhang, 2015). The previously described exploratory or hypothesis-generating nature of the majority of metagenomics studies is prone to disorientation and lack of conceptualization during the bioinformatics analysis steps. In order to avoid haphazardly applied bioinformatics analyses, a sound experimental design with a clear aim is of utmost importance. The clearer the aim, the more efficient the downstream methodological approaches will be. For example, expensive and time-consuming sequencing efforts are not required if the aim is to solely identify the dominant microbial groups in a community when low-resolution but cheaper and faster community profiling approaches are applicable (Hamady and Knight, 2009). Researchers should also make feasible choices with respect to the study goals and a consolidated decision should be made between SSU rRNA amplicon sequencing and shotgun sequencing where the latter is still orders of magnitude more expensive and computationally more time consuming albeit the depth and breadth of sequence reads it generates. As a rule of thumb, pilot studies are suggested to be performed via amplicon sequencing in order to get an idea about the species composition and diversity of the microbiota in the studied environment. Shotgun metagenomic sequencing can then be applied to only those samples, which suggest leads of novel discoveries regarding the biochemical capacity of the investigated microbiome.

The crucial steps of sampling, such as filtering and metadata recording, should be thoroughly examined. As such, depending on the study goals microbial communities can be stratified to varying size fractions covering large eukaryotes, small protists and bacteria as well as viruses. Minimum information standards regarding the recording and reporting of metadata for the investigated environmental microbiome (Yilmaz et al., 2011) should be followed in order both to increase the number of parameters that can explain data variability and also to eliminate the risk of future inquiries for the credibility of the performed metagenomics study. Scientific reproducibility of the results follows in the wake of adherence to such scientific community standards (Jasny et al., 2011). In order to extract the maximal amount of intracellular DNA, either custom protocols should be optimized or ready-made kits should be purchased relevant for the environments such as soil, water and gastrointestinal tracts or for the exceptional cases of multi-species life forms such as the biofilm samples. Since DNA sequencing constitutes only a proxy for the detection of actual processes going on *in vivo*, identified leads of novel mechanisms should be validated at gene expression and/or protein levels through complementary methodologies including quantitative/real-time PCR (qPCR), mass spectrometry and enzyme activity assays.

The computational aspects of metagenomics analyses mainly involve DNA sequence data processing pipelines, mathematical models and software tools tailored primarily to handle the ever-increasing volume of data generated by NGS technologies. Furthermore, established paradigms in the fields of microbial ecology, comparative genomics and gene expression analyses are continuously implemented into metagenomics pipelines by improved strength of relevant statistics. A common fallacy among the bioinformatics researchers is the approach to metagenomics data analysis pipelines with the certainty that a major finding is warranted. A comprehensive understanding of strengths and weaknesses of "big data" analysis schemes and reference biological databases through the interdisciplinary methodology portrayed in this thesis and exhorted by many other scholars, should characterize any research effort in metagenomics. It should not be forgotten that the glory of large-scale sequence data generation methodologies will not subside a potential failure to explain the ongoing biological processes in the investigated environment. Therefore, biological inference-oriented software development projects should be promoted, which will drastically reduce the time and effort spend on analyzing and interpreting metagenomics data as in the examples of the mg-RAST server, MEGAN, STAMP, FANTOM and PACFM. Open source software tools that are easy to use and apply will eventually replace the autocracy of bioinformatics analysts at the final step of long-term metagenomics endeavors. Finally modeling efforts of environmental microbial communities should be promoted to increase the predictive power of the existing methods upon environmental disturbances such as in ecotoxicological experiments. Modeling approaches reinforced by metabolic engineering paradigms such as the application of flux balance analysis in (meta)genome scale metabolic models (Price et al., 2004; Österlund et al., 2012; Hanemaaijer et al., 2015; Mardinoglu and Nielsen, 2015) will leverage our capability of conceptualizing nucleotide sequence data. Robust community-scale models applied by the integrative utilization of metagenomics, metatranscriptomics, metaproteomics and metabolomics will complement the final pieces in the puzzle of the central dogma of molecular biology and lead towards an era of maturity in high-throughput data 'omics biology.

7. ACKNOWLEDGEMENTS

I would like to thank Hans Blanck and Martin Eriksson for giving me the opportunity to start this PhD project. As an inexperienced PhD student with background training largely in molecular biology, Hans brought me into a scientific culture-shock through his attempts to indoctrinate how any biological phenomenon has to be ecologically relevant. Hans, I truly internalized this approach through the end of my PhD. I would like to especially thank you for inspiring me on the interdisciplinarity aspects and for your contributions to the kappa and the Papers III and IV.

The substantial contribution of Martin Eriksson in both scientific and administrative aspects of my PhD thesis has been crucial. I would like to thank Martin for his generous support at certain times. This PhD thesis would not be complete without his help. Thank you Martin also for your contributions to the Papers III and IV.

Henrik Nilsson kindly accepted to be my co-supervisor at the beginning of my PhD and his contribution, especially during the manuscript revision periods were decisive for the fate of my papers. “This baby is ready to fly!” What a motivating expression for a juvenile PhD student. I also owe my strong ambition for using the archaic text editors to you Henrik. How would I become the bioinformatics practitioner that I am now if I did not see you type “vi” on the command line? I would like to thank you very much, for being my unofficial mentor during this PhD period and for your contributions to the kappa, and the Papers II, III, and IV.

I would like to thank Henrik Aronsson, who apart from his achievements in vesicular transport in plant molecular biology, has worked in the Department of Biological and Environmental Sciences as virtually the life insurance of a PhD student. I also would like to acknowledge his benevolent efforts for me to accomplish this PhD project. I am sure it would not be possible without his final touches. Many thanks Henrik. Thank you also for contributing to the revision of the kappa.

I would like to thank Alexander Eiler whom I met by chance as an office-mate in Chalmers. Alex, you have taught me how to follow the actions that I believe were right in academia. Your utmost humility in both scientific discussions and personal communications has set an ideal model to pursue throughout my future life. Thank you very much. I have been honored. I hereby also would like to acknowledge your contributions to the kappa and the Papers II, III, and IV.

I thank Erik Kristiansson for arranging an office for me in Chalmers and kindly inviting me to his group discussions. I also would like to thank you Erik for being by my side throughout the down periods of my PhD time. I truly appreciate and will not forget what you have done. Thank you also for your contributions to the Papers II, III, and IV.

I also would like to thank Ingela Dahllöf whose timely involvements gave me additional motivation for the PhD project.

I would like to acknowledge my MSc thesis supervisors from the Systems Biology group at Chalmers, who also co-authored Paper I. Special thanks to Jens Nielsen, Intawat Nookaew, and Fredrik Karlsson. I also would like to thank Adil Mardinoglu for his constructive scientific discussions and kind support during my PhD. Adil’s contribution was also invaluable in Paper II.

I would like to acknowledge Magnus-Alm Rosenblad who started as my co-supervisor at the beginning of my PhD and suggested me to attend a workshop in genomics in Czech Republic. Although, we could not see the end of this PhD marathon together, I am grateful that he informed me about this workshop and indirectly helped me shape my ideas about bioinformatics and metagenomics. I also would like to thank Mats Töpel for his administrative role for the bioinformatics cluster where I spent a huge amount of time for my analyses. Thanks for your patience for my reckless exploitation of the cluster resources at times.

I would like to convey my warmest gratefulness to Lucas Sinclair and his immense knowledge about computers. Thank you very much Lucas also for your friendship. I will not forget our scientific and ethics-related discussions as alternative recreational activities.

I thank Thomas Backhaus and Adrian K. Clarke for being my examiners.

I would like to thank the former and current members of the community ecotoxicology group in the department of Biological and Environmental Sciences. Triranta, Henrik, Mikael, Åsa, Natalia, Ida, Viktor F., Maria and Marianne: thank you all!

The dream team of innebandy players: Somnath, Filipe, Thomas, Alex and Ivana, Ruud, Jonna, Fabian, Yann, Bernard, Jason, Bengt, and Azeez. Thank you all for introducing me to this weird game. I will probably keep missing the Friday afterwork as long as I stay in Gothenburg. Bernard, Filipe, and Jonna: special thanks for organizing the weekly events.

I would like to thank the PhD students in the department of Biological and Environmental Sciences: Andrei, Karine, Lisa, Victor C, Victor G, Ivan, Josué, Fernanda, Romina, Priscilla, and all others. Thank you for reminding me of my PhD despair and of not being alone in this experience. ☺

I would like to thank Johanna B, Jenny, Nurun, Zeynep, Anna P, and Daniela, just for being great female biologists around me, as if one needs any more reason to be thankful while working with computers. ☺

I would like to thank the members of the GoBiG group: Johan, Fredrik, Viktor J, Anna J, Chandan, Amir, Tobias, Francesco, and Mariana. Thank you Johan for organizing this alternative PhD group outside the Botanical Garden, for me to brainstorm and socialize with other PhD students working with bioinformatics in Gothenburg.

Sven, Niclas, Ylva, and Ingela L: You have been the secret heroes and heroines of this and many other PhD projects. Thank you very much for your endless help in administrative, practical and technical issues.

Berkay, Melisa, Barış T, Can P, Efe, Selçuk Ç, Ender ve diğer tüm FTAS yönetici ve üyelerine doktora sürecimdeki sosyal yaşantıma katkılarından dolayı teşekkür ederim.

Göreceli olarak sonsöz: Bu doktora tezinin tamamlanmasındaki katkılarından dolayı aileme sonsuz teşekkürlerimi sunarım. İyi ki varsınız!

8. REFERENCES

Abbott, A., Abel, P., Arnold, D., and Milne, A. (2000). Cost–benefit analysis of the use of TBT: the case for a treatment approach. *Science of the Total Environment* 258**,** 5-19.

Ahner, B.A., Kong, S., and Morel, F.M. (1995). Phytochelatin production in marine algae. 1. An interspecies comparison. *Limnology and Oceanography* 40**,** 649-657.

Aldridge, P., and Hughes, K.T. (2002). Regulation of flagellar assembly. *Current Opinion in Microbiology* 5**,** 160-165.

Allen, E.E., and Banfield, J.F. (2005). Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology* 3**,** 489-498.

Allen, R., Rick, S., and William, N. (2011). Interdisciplinary Research: Case Studies of Integrative Understandings of Complex Problems.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25**,** 3389-3402.

Amano, A., Nakagawa, I., and Hamada, S. (1999). Studying initial phase of biofilm formation: molecular interaction of host proteins and bacterial surface components. *Methods in Enzymology* 310**,** 501.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. *Reference Source*.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., and Magrane, M. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 32**,** D115-D119.

Arrigo, K.R. (2005). Marine microorganisms and global nutrient cycles. *Nature* 437**,** 349-355.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., and Batto, J.-M. (2011). Enterotypes of the human gut microbiome. *Nature* 473**,** 174-180.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25**,** 25-29.

Attwood, T., Gisel, A., Bongcam-Rudloff, E., and Eriksson, N. (2011). *Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective.* INTECH Open Access Publisher.

Backhaus, T., Sumpter, J., and Blanck, H. (2008). "On the ecotoxicology of pharmaceutical mixtures," in *Pharmaceuticals in the Environment*. Springer), 257-276.

Baker, G., Smith, J.J., and Cowan, D.A. (2003). Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods* 55**,** 541-555.

Bansal, A.K. (2005). Bioinformatics in microbial biotechnology–a mini review. *Microbial Cell Factories* 4**,** 19.

Bardy, S.L., Ng, S.Y., and Jarrell, K.F. (2003). Prokaryotic motility structures. *Microbiology* 149**,** 295-304.

Barton, L.L., and Northup, D.E. (2011). *Microbial Ecology.* Wiley-Blackwell.

Béja, O., Spudich, E.N., Spudich, J.L., Leclerc, M., and Delong, E.F. (2001). Proteorhodopsin phototrophy in the ocean. *Nature* 411**,** 786-789.

Beller, H.R., Chain, P.S., Letain, T.E., Chakicherla, A., Larimer, F.W., Richardson, P.M., Coleman, M.A., Wood, A.P., and Kelly, D.P. (2006). The genome sequence of the obligately chemolithoautotrophic, facultatively anaerobic bacterium Thiobacillus denitrificans. *Journal of Bacteriology* 188**,** 1473-1488.

Bengtsson, J., Eriksson, K.M., Hartmann, M., Wang, Z., Shenoy, B.D., Grelet, G.-A., Abarenkov, K., Petri, A., Rosenblad, M.A., and Nilsson, R.H. (2011). Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie Van Leeuwenhoek* 100**,** 471-475.

Benhamou, N. (1996). Elicitor-induced plant defence pathways. *Trends in Plant Science* 1**,** 233-240.

Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics* 5**,** 433-438. doi: 10.1517/14622416.5.4.433.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2008). GenBank. *Nucleic Acids Research* 36**,** D25-D30.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank. *European Journal of Biochemistry* 80**,** 319-324.

Blanck, H. (2002). A critical review of procedures and approaches used for assessing pollution-induced community tolerance (PICT) in biotic communities. *Human and Ecological Risk Assessment* 8**,** 1003-1034.

Blanck, H., Eriksson, K.M., Grönvall, F., Dahl, B., Guijarro, K.M., Birgersson, G., and Kylin, H. (2009). A retrospective analysis of contamination and periphyton PICT patterns for the antifoulant irgarol 1051, around a small marina on the Swedish west coast. *Marine Pollution Bulletin* 58**,** 230-237.

Blanck, H., Wängberg, S.-Å., and Molander, S. (1988). "Pollution-induced community tolerance—a new ecotoxicological tool," in *Functional testing of aquatic biota for estimating hazards of chemicals*. ASTM International).

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol* 13**,** R122.

Bonilla, S., Conde, D., and Blanck, H. (1998). The photosynthetic responses of marine phytoplankton, periphyton and epipsammon to the herbicides paraquat and simazine. *Ecotoxicology* 7**,** 99-105.

Boxall, A.B. (2004). The environmental side effects of medication. *EMBO Reports* 5**,** 1110-1116.

Burrows, M., and Wheeler, D. (Year). "A block-sorting lossless data compression algorithm", in: *DIGITAL SRC RESEARCH REPORT*: Citeseer).

Busalmen, J., Vazquez, M., and De Sanchez, S. (2002). New evidences on the catalase mechanism of microbial corrosion. *Electrochimica Acta* 47**,** 1857-1865.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10**,** 1.

Canfield, D.E., Kristensen, E., and Thamdrup, B. (2005). *Aquatic geomicrobiology.* Gulf Professional Publishing.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research* 37**,** D233-D238.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., and Gordon, J.I. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7**,** 335-336.

Caswell, H. (1996). Demography meets ecotoxicology: untangling the population level effects of toxic substances. *Ecotoxicology: A hierarchical treatment***,** 255-292.

Caufrier, F., Martinou, A., Dupont, C., and Bouriotis, V. (2003). Carbohydrate esterase family 4 enzymes: substrate specificity. *Carbohydrate Research* 338**,** 687-692.

Celander, M.C. (2011). Cocktail effects on biomarker responses in fish. *Aquatic Toxicology* 105**,** 72-77.

Chen, S., Arsenault, C., Gingras, Y., and Larivière, V. (2015a). Exploring the interdisciplinary evolution of a discipline: the case of Biochemistry and Molecular Biology. *Scientometrics* 102**,** 1307-1323.

Chen, X., Casas, M.E., Nielsen, J.L., Wimmer, R., and Bester, K. (2015b). Identification of Triclosan-O-Sulfate and other transformation products of Triclosan formed by activated sludge. *Science of the Total Environment* 505**,** 39-46.

Cho, J.-C., and Giovannoni, S.J. (2004). Cultivation and growth characteristics of a diverse group of oligotrophic marine Gammaproteobacteria. *Applied and Environmental Microbiology* 70**,** 432-440.

Clark, D.P. (1989). The fermentation pathways of Escherichia coli. *FEMS Microbiology Reviews* 5**,** 223-234.

Clements, W.H., and Newman, M.C. (2003). *Community ecotoxicology.* John Wiley & Sons.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., and Wilczynski, B. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25**,** 1422-1423.

Cooksey, K., and Wigglesworth-Cooksey, B. (1995). Adhesion of bacteria and diatoms to surfaces in the sea: a review. *Aquatic Microbial Ecology* 9**,** 87-96.

Costerton, J.W. (2007). *The biofilm primer.* Springer Science & Business Media.

Costerton, J.W., Lewandowski, Z., Caldwell, D.E., Korber, D.R., and Lappin-Scott, H.M. (1995). Microbial biofilms. *Annual Reviews in Microbiology* 49**,** 711-745.

Costerton, J.W., Stewart, P.S., and Greenberg, E. (1999). Bacterial biofilms: a common cause of persistent infections. *Science* 284**,** 1318-1322.

Dahl, B., and Blanck, H. (1996). Toxic effects of the antifouling agent Irgarol 1051 on periphyton communities in coastal water microcosms. *Marine Pollution Bulletin* 32**,** 342-350.

Davey, M.E., and O'toole, G.A. (2000). Microbial biofilms: from ecology to molecular genetics. *Microbiology and Molecular Biology Reviews* 64**,** 847-867.

Day, W.H., and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification* 1**,** 7-24.

Dayhoff, M.O.T.B., C (1967). Margaret O.Dayhoff Papers,. *Archives of the National Biomedical Research Foundation, Washington, D.C., USA.*

De Beer, D., Stoodley, P., Roe, F., and Lewandowski, Z. (1994). Effects of biofilm structures on oxygen distribution and mass transport. *Biotechnology and Bioengineering* 43**,** 1131-1138.

Dear, P.H. (2001). "Genome Map," in *eLS*. John Wiley & Sons, Ltd).

Decho, A.W. (2000). Microbial biofilms in intertidal systems: an overview. *Continental Shelf Research* 20**,** 1257-1273.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23**,** 673-679.

Delong, E.F. (2005). Microbial community genomics in the ocean. *Nature Reviews Microbiology* 3**,** 459-469.

Delong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.-U., Martinez, A., Sullivan, M.B., Edwards, R., and Brito, B.R. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311**,** 496-503.

Devilla, R.A., Brown, M.T., Donkin, M., and Readman, J.W. (2005). The effects of a PSII inhibitor on phytoplankton community structure as assessed by HPLC pigment analyses, microscopy and flow cytometry. *Aquatic Toxicology* 71**,** 25-38.

Diggle, S.P., Stacey, R.E., Dodd, C., Cámara, M., Williams, P., and Winzer, K. (2006). The galactophilic lectin, LecA, contributes to biofilm development in Pseudomonas aeruginosa. *Environmental Microbiology* 8**,** 1095-1104.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* 14**,** 927-930.

Domozych, D., Ciancia, M., Fangel, J.U., Mikkelsen, M.D., Ulvskov, P., and Willats, W.G.T. (2012). The cell walls of green algae: a journey through evolution and diversity. *Frontiers in Plant Science* 3. doi: 10.3389/fpls.2012.00082.

Doolittle, W.F., and Zhaxybayeva, O. (2009). On the origin of prokaryotic species. *Genome Research* 19**,** 744-756.

Dunne, W.M. (2002). Bacterial adhesion: seen any good biofilms lately? *Clinical Microbiology Reviews* 15**,** 155-166.

Eddy, S. (1998). "HMMER2 Profile hidden Markov models for biological sequence analysis".).

Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 10**,** 996-+. doi: 10.1038/nmeth.2604.

Edwards, K.J., Bond, P.L., Gihring, T.M., and Banfield, J.F. (2000). An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* 287**,** 1796-1799.

Ehrlich, H. (2010). Chitin and collagen as universal and alternative templates in biomineralization. *International Geology Review* 52**,** 661-699.

Eiler, A. (2006). Evidence for the ubiquity of mixotrophic bacteria in the upper ocean: implications and consequences. *Applied and Environmental Microbiology* 72**,** 7431-7437.

Eiler, A., Mondav, R., Sinclair, L., Fernandez-Vidal, L., Scofield, D.G., Schwientek, P., Martinez-Garcia, M., Torrents, D., Mcmahon, K.D., and Andersson, S.G. (2016). Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria. *The ISME Journal*.

Eiler, A., Zaremba‐Niedzwiedzka, K., Martínez‐García, M., Mcmahon, K.D., Stepanauskas, R., Andersson, S.G., and Bertilsson, S. (2014). Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics. *Environmental Microbiology* 16**,** 2682-2698.

Elias, S., and Banin, E. (2012). Multi‐species biofilms: living with friendly neighbors. *FEMS Microbiology Reviews* 36**,** 990-1004.

Eriksson, K., Clarke, A., Franzen, L.-G., Kuylenstierna, M., Martinez, K., and Blanck, H. (2009). Community-level analysis of psbA gene sequences and irgarol tolerance in marine periphyton. *Applied and environmental microbiology* 75**,** 897-906.

Finn, R.D. (2012). Pfam: the protein families database. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269**,** 496-512.

Flemming, H.-C., Neu, T.R., and Wozniak, D.J. (2007). The EPS matrix: the "house of biofilm cells". *Journal of Bacteriology* 189**,** 7945-7947.

Flemming, H.-C., and Wingender, J. (2010). The biofilm matrix. *Nature Reviews Microbiology* 8**,** 623-633.

Foster, T.J., and Höök, M. (1998). Surface protein adhesins of Staphylococcus aureus. *Trends in Microbiology* 6**,** 484-488.

Francis, C.A., Beman, J.M., and Kuypers, M.M. (2007). New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *The ISME Journal* 1**,** 19-27.

Friedrich, C.G., Bardischewsky, F., Rother, D., Quentmeier, A., and Fischer, J. (2005). Prokaryotic sulfur oxidation. *Current Opinion in Microbiology* 8**,** 253-259.

Frølund, B., Palmgren, R., Keiding, K., and Nielsen, P.H. (1996). Extraction of extracellular polymers from activated sludge using a cation exchange resin. *Water Research* 30**,** 1749-1758.

Fuchs, G. (2011). Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annual Review of Microbiology* 65**,** 631-658.

Garner, J., Porter, A.L., Borrego, M., Tran, E., and Teutonico, R. (2013). Facilitating social and natural science cross-disciplinarity: Assessing the human and social dynamics program. *Research Evaluation* 22**,** 134-144.

GESAMP. (1983). Report of the Thirteenth Session, Geneva, Switzerland, 28. February-4 March 1983. IMO/FAO/UNESCO/WMO/WHO/IAEA/UN/UNEP Joint Group, of Experts on Scientific Aspects of Marine Pollution. *Reports and Studies (18):50p. WMO, Geneva, Switzerland.*

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van De Peer, Y., Vandamme, P., and Thompson, F.L. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology* 3**,** 733-739.

Gilbert, J.A., and Dupont, C.L. (2011). Microbial metagenomics: beyond the genome. *Annual Review of Marine Science* 3**,** 347-371.

Gilbert, J.A., Jansson, J.K., and Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biology* 12**,** 69.

Gonzalez, A., King, A., Robeson Ii, M.S., Song, S., Shade, A., Metcalf, J.L., and Knight, R. (2012). Characterizing microbial communities through space and time. *Current Opinion in Biotechnology* 23**,** 431-436.

Gordon, A., and Hannon, G. (2010). Fastx-toolkit. *Computer program distributed by the author, website* http://hannonlab. *cshl. edu/fastx_toolkit/index. html [accessed 2014–2015]*.

Gunnarsson, L., Kristiansson, E., Rutgersson, C., Sturve, J., Fick, J., Förlin, L., and Larsson, D. (2009). Pharmaceutical industry effluent diluted 1: 500 affects global gene expression, cytochrome P450 1A activity, and plasma phosphate in fish. *Environmental Toxicology and Chemistry* 28**,** 2639-2647.

Hadfield, M.G., and Paul, V.J. (2001). Natural chemical cues for settlement and metamorphosis of marine invertebrate larvae. *Marine Chemical Ecology***,** 431-461.

Haft, D.H., Selengut, J.D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research* 31**,** 371-373.

Halden, R.U., and Paull, D.H. (2005). Co-Occurrence of Triclocarban and Triclosan in U.S. Water Resources. *Environmental Science & Technology* 39**,** 1420-1426. doi: 10.1021/es049071e.

Hall-Stoodley, L., Costerton, J.W., and Stoodley, P. (2004). Bacterial biofilms: from the natural environment to infectious diseases. *Nature Reviews Microbiology* 2**,** 95-108.

Hamady, M., and Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research* 19**,** 1141-1152.

Handelsman, J. (2005). Metagenomics or megagenomics? *Nature Reviews Microbiology* 3**,** 457-458.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* 5**,** R245-R249.

Handy, R., and Depledge, M. (1999). Physiological responses: their measurement and use as environmental biomarkers in ecotoxicology. *Ecotoxicology* 8**,** 329-349.

Hanemaaijer, M., Röling, W.F., Olivier, B.G., Khandelwal, R.A., Teusink, B., and Bruggeman, F.J. (2015). Systems modeling approaches for microbial community studies: from metagenomics to inference of the community structure. *Frontiers in Microbiology* 6.

Hartmann, M., Howes, C.G., Abarenkov, K., Mohn, W.W., and Nilsson, R.H. (2010). V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods* 83**,** 250-253.

Hernández-Sebastià, C., Varin, L., and Marsolais, F. (2008). "Sulfotransferases from plants, algae and phototrophic bacteria," in *Sulfur Metabolism in Phototrophic Organisms*. Springer), 111-130.

Hirsch, R., Ternes, T., Haberer, K., and Kratz, K.-L. (1999). Occurrence of antibiotics in the aquatic environment. *Science of the Total Environment* 225**,** 109-118.

Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K., and Pease, L.R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* 77**,** 51-59.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Comput Biol* 7**,** e1002021.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., and Ise, K. (2016). A new view of the tree of life. *Nature Microbiology***,** 16048.

Hughes, J.B., and Bohannan, B.J.M. (2004). "Section 7 update: Application of ecological diversity statistics in microbial ecology," in *Molecular Microbial Ecology Manual,* eds. G.A. Kowalchuk, F.J. Bruijn, I.M. Head, A.D. Akkermans & J.D. Elsas. (Dordrecht: Springer Netherlands), 3223-3246.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology* 67**,** 4399-4406.

Hügler, M., and Sievert, S.M. (2011). Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. *Marine Science* 3.

Hunter, D.J. (2006). Genomics and proteomics in epidemiology: treasure trove or "high-tech stamp collecting"? *Epidemiology* 17**,** 487-489.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. *Genome Research* 17**,** 377-386.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11**,** 1.

Hyde, K.D., Udayanga, D., Manamgoda, D.S., Tedersoo, L., Larsson, E., Abarenkov, K., Bertrand, Y.J., Oxelman, B., Hartmann, M., and Kauserud, H. (2013). Incorporating molecular data in fungal systematics: a guide for aspiring researchers. *arXiv preprint arXiv:1302.3244*.

Izano, E.A., Amarante, M.A., Kher, W.B., and Kaplan, J.B. (2008). Differential roles of poly-N-acetylglucosamine surface polysaccharide and extracellular DNA in

Staphylococcus aureus and Staphylococcus epidermidis biofilms. *Applied and Environmental Microbiology* 74**,** 470-476.

Jasny, B.R., Chin, G., Chong, L., and Vignieri, S. (2011). Again, and again, and again…. *Science* 334**,** 1225-1225.

Jessup, C.M., Kassen, R., Forde, S.E., Kerr, B., Buckling, A., Rainey, P.B., and Bohannan, B.J. (2004). Big questions, small worlds: microbial model systems in ecology. *Trends in Ecology & Evolution* 19**,** 189-197.

Johansson, C.H., Janmar, L., and Backhaus, T. (2014). Triclosan causes toxic effects to algae in marine biofilms, but does not inhibit the metabolic activity of marine biofilm bacteria. *Marine Pollution Bulletin* 84**,** 208-212. doi: http://dx.doi.org/10.1016/j.marpolbul.2014.05.010.

Ju, F., and Zhang, T. (2015). Experimental Design and Bioinformatics Analysis for the Application of Metagenomics in Environmental Sciences and Biotechnology. *Environmental Science & Technology* 49**,** 12628-12640.

Jurtshuk, P. (1996). "Bacterial Metabolism," in *Medical Microbiology,* ed. S. Baron. 4 ed (University of Texas Medical Branch at Galveston, Galveston, Texas).

Kachlany, S.C., Planet, P.J., Bhattacharjee, M.K., Kollia, E., Desalle, R., Fine, D.H., and Figurski, D.H. (2000). Nonspecific Adherence by Actinobacillus actinomycetemcomitans Requires Genes Widespread inBacteria and Archaea. *Journal of Bacteriology* 182**,** 6169-6176.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., and Tokimatsu, T. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research* 36**,** D480-D484.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28**,** 27-30.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Research* 30**,** 42-46.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. *Genome research* 12**,** 656-664.

Kirchman, D.L. (2008). "Introduction and Overview," in *Microbial Ecology of the Oceans,* ed. D.L. Kirchman. 2 ed (College of Marine and Earth Studies, University of Delaware: John Wiley & Sons, Inc.), 3.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. (2001). rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Research* 29**,** 181-184.

Kolenbrander, P.E. (1989). Surface recognition among oral bacteria: multigeneric coaggregations and their mediators. *Critical Reviews in Microbiology* 17**,** 137-159.

Korber, D.R., Lawrence, J.R., and Caldwell, D.E. (1994). Effect of motility on surface colonization and reproductive success of Pseudomonas fluorescens in dual-dilution continuous culture and batch culture systems. *Applied and Environmental Microbiology* 60**,** 1421-1429.

Krane, D.E., and Raymer, M.L. (2003). "Distance Based Methods of Phylogenetics," in *Fundamental Concepts of Bioinformatics,* eds. D.E. Kranei & M.L. Raymer. (Benjamin Cummings, 1301 Sansome Street, San Francisco, CA: Pearson Education Inc.), 77-97.

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* 25**,** 2737-2738.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research* 19**,** 1639-1645.

Lalli, C., and Parsons, T.R. (1997). "Marine Pollutants," in *Biological Oceanography: An Introduction*. 2 ed (Linacre House, Jordan Hill, Oxford OX2 8DP: Elsevier Butterworth-Heinemann), 251-156.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9**,** 357-359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10**,** R25.

Lazarova, V., and Manem, J. (1995). Biofilm characterization and activity analysis in water and wastewater treatment. *Water Research* 29**,** 2227-2245.

Lee, J.-H., Yi, H., and Chun, J. (2011). rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *The Journal of Microbiology* 49**,** 689-691.

Lee, W., and De Beer, D. (1995). Oxygen and pH microprofiles above corroding mild steel covered with a biofilm. *Biofouling* 8**,** 273-280.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* 11**,** 473-483.

Liu, W.-T., Marsh, T.L., Cheng, H., and Forney, L.J. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology* 63**,** 4516-4522.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., and Pallen, M.J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* 30**,** 434-439.

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**,** 1-21.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., and Liu, Y. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1**,** 1-6.

Lynch, D.J., Fountain, T.L., Mazurkiewicz, J.E., and Banas, J.A. (2007). Glucan-binding proteins are essential for shaping Streptococcus mutans biofilm architecture. *FEMS Microbiology Letters* 268**,** 158-165.

Maciorowski, A.F. (1988). Populations and communities: linking toxicology and ecology in a new synthesis. *Environmental Toxicology and Chemistry* 7**,** 677-678.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database* 2011**,** bar009.

Mardinoglu, A., and Nielsen, J. (2015). New paradigms for metabolic modeling of human cells. *Current Opinion in Biotechnology* 34**,** 91-97.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 24**,** 133-141.

Markowitz, V.M., Ivanova, N.N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.-M.A., Grechkin, Y., Dubchak, I., and Anderson, I. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research* 36**,** D534-D538.

Marsh, P. (2004). Dental plaque as a microbial biofilm. *Caries Research* 38**,** 204-211.

Max-Neef, M.A. (2005). Foundations of transdisciplinarity. *Ecological Economics* 53**,** 5-16.

Mayer, C., Moritz, R., Kirschner, C., Borchard, W., Maibaum, R., Wingender, J., and Flemming, H.-C. (1999). The role of intermolecular interactions: studies on model systems for bacterial biofilms. *International Journal of Biological Macromolecules* 26**,** 3-16.

Mccollom, T., and Amend, J. (2005). A thermodynamic assessment of energy requirements for biomass synthesis by chemolithoautotrophic micro‐organisms in oxic and anoxic environments. *Geobiology* 3**,** 135-144.

Mcpeek, M.A., and Peckarsky, B.L. (1998). Life histories and the strengths of species interactions: combining mortality, growth, and fecundity effects. *Ecology* 79**,** 867-879.

Mercier, L., Lafitte, C., Borderies, G., Briand, X., Esquerré-Tugayé, M.T., and Fournier, J. (2001). The algal polysaccharide carrageenans can act as an elicitor of plant defence. *New Phytologist* 149**,** 43-51.

Metzger, U., Lankes, U., Fischpera, K., and Frimmel, F.H. (2009). The concentration of polysaccharides and proteins in EPS of Pseudomonas putida and Aureobasidum pullulans as revealed by 13C CPMAS NMR spectroscopy. *Applied Microbiology and Biotechnology* 85**,** 197-206.

Metzker, M.L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics* 11**,** 31-46.

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., and Wilke, A. (2008). The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9**,** 386.

Miao, A.-J., Schwehr, K.A., Xu, C., Zhang, S.-J., Luo, Z., Quigg, A., and Santschi, P.H. (2009). The algal toxicity of silver engineered nanoparticles and detoxification by exopolymeric substances. *Environmental Pollution* 157**,** 3034-3041.

Mitchell, A., Bucchini, F., Cochrane, G., Denise, H., Ten Hoopen, P., Fraser, M., Pesseat, S., Potter, S., Scheremetjew, M., and Sterk, P. (2016). EBI metagenomics in 2016-an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research* 44**,** D595-D603.

Mittelman, M.W. (1996). Adhesion to biomaterials. *Bacterial adhesion: molecular and ecological diversity. New York: Wiley-Liss, Inc***,** 89-127.

Molin, S. (1999). Microbial activity in biofilm communities. *Dental plaque revisited: oral biofilms in health and disease***,** 73-78.

Moran, M., Belas, R., Schell, M., González, J., Sun, F., Sun, S., Binder, B., Edmonds, J., Ye, W., and Orcutt, B. (2007). Ecological genomics of marine Roseobacters. *Applied and Environmental Microbiology* 73**,** 4559-4569.

Mouritsen, O.G. (2011). Lipidology and lipidomics—quo vadis? A new era for the physical chemistry of lipids. *Physical Chemistry Chemical Physics* 13**,** 19195-19205.

Munn, C.B. (2004). *Marine microbiology.* London: Bios Scientific.

Muyzer, G., De Waal, E.C., and Uitterlinden, A.G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* 59**,** 695-700.

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

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48**,** 443-453.

Neu, T.R., and Lawrence, J.R. (1997). Development and structure of microbial biofilms in river water studied by confocal laser scanning microscopy. *FEMS Microbiology Ecology* 24**,** 11-25.

Newman, M., and Unger, M. (2003). Fundamentals of ecotoxicology Lewis Publishers. *Boca Raton, Florida* 458.

Newman, M.C., and Clements, W.H. (2007). *Ecotoxicology: a comprehensive treatment.* cRc Press.

Ning, M., and Lo, E.H. (2010). Opportunities and challenges in Omics. *Translational Stroke Research* 1**,** 233-237.

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research* 34**,** 5623-5630.

O'toole, G., Kaplan, H.B., and Kolter, R. (2000). Biofilm formation as microbial development. *Annual Reviews in Microbiology* 54**,** 49-79.

Oberdörster, E., Zhu, S., Blickley, T.M., Mcclellan-Green, P., and Haasch, M.L. (2006). Ecotoxicology of carbon-based engineered nanoparticles: effects of fullerene (C 60) on aquatic organisms. *Carbon* 44**,** 1112-1120.

Ohmura, N., and Blake, R. (1997). Aporusticyanin mediates the adhesion of Thiobacillus ferrooxidans to pyrite. *IBS Biomine* 97**,** 1-10.

Okafor, N. (2011). "Taxonomy, Physiology, and Ecology of Aquatic Microorganisms," in *Environmental Microbiology of Aquatic and Waste Systems,* ed. N. Okafor. (Springer Dordrecht Heidelberg London New York: Springer Science+Business Media), 82.

Oksanen, J., Kindt, R., Legendre, P., O'hara, B., Stevens, M.H.H., Oksanen, M.J., and Suggests, M. (2007). The vegan package. *Community Ecology Package* 10.

Österlund, T., Nookaew, I., and Nielsen, J. (2012). Fifteen years of large scale metabolic modeling of yeast: developments and impacts. *Biotechnology Advances* 30**,** 979-988.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. (1985). Analyzing natural microbial populations by rRNA sequences. *ASM American Society for Microbiology News* 51**,** 4-12.

Paerl, H., and Pinckney, J. (1996). A mini-review of microbial consortia: their roles in aquatic production and biogeochemical cycling. *Microbial Ecology* 31**,** 225-247.

Parks, D.H., Tyson, G.W., Hugenholtz, P., and Beiko, R.G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30**,** 3123-3124.

Parsek, M.R., and Greenberg, E. (2005). Sociomicrobiology: the connections between quorum sensing and biofilms. *Trends in Microbiology* 13**,** 27-33.

Pasmore, M., and Costerton, J.W. (2003). Biofilms, bacterial signaling, and their ties to marine biology. *Journal of Industrial Microbiology and Biotechnology* 30**,** 407-413.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12**,** 2825-2830.

Pei, A.Y., Oberdorf, W.E., Nossa, C.W., Agarwal, A., Chokshi, P., Gerz, E.A., Jin, Z., Lee, P., Yang, L., and Poles, M. (2010). Diversity of 16S rRNA genes within individual prokaryotic genomes. *Applied and Environmental Microbiology* 76**,** 3886-3897.

Po, T., Steger, G., Rosenbaum, V., Kaper, J., and Riesner, D. (1987). Double-stranded cucumovirus associated RNA 5: experimental analysis of necrogenic and non-necrogenic variants by temperature-gradient gel electrophoresis. *Nucleic Acids Research* 15**,** 5069-5083.

Pratt, L.A., and Kolter, R. (1998). Genetic analysis of Escherichia coli biofilm formation: roles of flagella, motility, chemotaxis and type I pili. *Molecular Microbiology* 30**,** 285-293.

Price, N.D., Reed, J.L., and Palsson, B.Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology* 2**,** 886-897.

Prouty, A., Schwesinger, W., and Gunn, J. (2002). Biofilm Formation and Interaction with the Surfaces of Gallstones by Salmonella spp. *Infection and Immunity* 70**,** 2640-2649.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35**,** D61-D65.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., and Yamada, T. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464**,** 59-65.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41**,** D590-D596.

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research* 38**,** e191-e191.

Riesenfeld, C.S., Schloss, P.D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* 38**,** 525-552.

Rigden, D.J., Fernández-Suárez, X.M., and Galperin, M.Y. (2016). The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. *Nucleic Acids Research* 44**,** D1-D6.

Ronaghi, M., Uhlén, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science* 281**,** 363.

Rosemond, A.D., Mulholland, P.J., and Elwood, J.W. (1993). Top-down and bottom-up control of stream periphyton: effects of nutrients and herbivores. *Ecology* 74**,** 1264-1280.

Sand-Jensen, K., and Borum, J. (1991). Interactions among phytoplankton, periphyton, and macrophytes in temperate freshwaters and estuaries. *Aquatic Botany* 41**,** 137-175.

Sanger, F., and Coulson, A. (1978). The use of thin acrylamide gels for DNA sequencing. *FEBS Lett* 87**,** 107-110.

Sccs (Scientific Committee on Consumer Safety) (2010). "Opinion on triclosan (antimicrobial resistance)". (Brussels: Scientific Committe on Consumer Safety).

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., and Robinson, C.J. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75**,** 7537-7541.

Schmidt, T.S., Rodrigues, J.F.M., and Von Mering, C. (2014). Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput Biol* 10**,** e1003594.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27**,** 863-864.

Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biology* 5.

Sharon, I., Bercovici, S., Pinter, R.Y., and Shlomi, T. (2011). Pathway-based functional analysis of metagenomes. *Journal of Computational Biology* 18**,** 495-505.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology* 26**,** 1135-1145.

Sinclair, L., Osman, O.A., Bertilsson, S., and Eiler, A. (2015). Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the Illumina platform. *PloS One* 10**,** e0116955.

Singh, G.B. (2015a). "Biological Databases," in *Fundamentals of Bioinformatics and Computational Biology,* ed. G.B. Singh. (Springer Cham Heidelberg New York Dordrecht London: Springer International Publishing), 37-73.

Singh, G.B. (2015c). "Introduction to Bioinformatics," in *Fundamentals of Bioinformatics and Computational Biology,* ed. G.B. Singh. (Springer Cham Heidelberg New York Dordrecht London: Springer International Publishing), 3-4.

Singh, P.K., Schaefer, A.L., Parsek, M.R., Moninger, T.O., Welsh, M.J., and Greenberg, E. (2000). Quorum-sensing signals indicate that cystic fibrosis lungs are infected with bacterial biofilms. *Nature* 407**,** 762-764.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* 147**,** 195-197.

Stark, M., Berger, S.A., Stamatakis, A., and Von Mering, C. (2010). MLTreeMap-accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* 11**,** 461.

Stewart, P.S., and Costerton, J.W. (2001). Antibiotic resistance of bacteria in biofilms. *The Lancet* 358**,** 135-138.

Strohm, T.O., Griffin, B., Zumft, W.G., and Schink, B. (2007). Growth yields in bacterial denitrification and nitrate ammonification. *Applied and Environmental Microbiology* 73**,** 1420-1424.

Sussman, M., Denyer, S., and Gorman, S. (1993). *Microbial biofilms: formation and control.* Blackwell Scientific Publication.

Sutherland, I.W. (2001). Biofilm exopolysaccharides: a strong and sticky framework. *Microbiology* 147**,** 3-9.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28**,** 33-36.

Thauer, R.K., Jungermann, K., and Decker, K. (1977). Energy conservation in chemotrophic anaerobic bacteria. *Bacteriological Reviews* 41**,** 100.

Thayer, A.M. (2014). "Next-Gen Sequencing Is A Numbers Game", in: *Chemical & Engineering News.* American Chemical Society).

Tichi, M.A., and Tabita, F.R. (2001). Interactive control of Rhodobacter capsulatus redox-balancing systems during phototrophic metabolism. *Journal of bacteriology* 183**,** 6344-6354.

Tielker, D., Hacker, S., Loris, R., Strathmann, M., Wingender, J., Wilhelm, S., Rosenau, F., and Jaeger, K.-E. (2005). Pseudomonas aeruginosa lectin LecB is located in the outer membrane and is involved in biofilm formation. *Microbiology* 151**,** 1313-1323.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., and Affourtit, J.P. (2009). A core gut microbiome in obese and lean twins. *Nature* 457**,** 480-484.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J.I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* 449**,** 804.

Van Schaik, E.J., Giltner, C.L., Audette, G.F., Keizer, D.W., Bautista, D.L., Slupsky, C.M., Sykes, B.D., and Irvin, R.T. (2005). DNA binding: a novel function of Pseudomonas aeruginosa type IV pili. *Journal of Bacteriology* 187**,** 1455-1464.

Vandenkoornhuyse, P., Baldauf, S.L., Leyval, C., Straczek, J., and Young, J.P.W. (2002). Extensive fungal diversity in plant roots. *Science* 295**,** 2051-2051.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., and Holt, R.A. (2001). The sequence of the human genome. *Science* 291**,** 1304-1351.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., and Nelson, W. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304**,** 66-74.

Wagner, M., Amann, R., Lemmer, H., and Schleifer, K.-H. (1993). Probing activated sludge with oligonucleotides specific for proteobacteria: inadequacy of culture-dependent methods for describing microbial community structure. *Applied and Environmental Microbiology* 59**,** 1520-1525.

Wang, S., Liu, X., Liu, H., Zhang, L., Guo, Y., Yu, S., Wozniak, D.J., and Ma, L.Z. (2015). The exopolysaccharide Psl–eDNA interaction enables the formation of a biofilm skeleton in Pseudomonas aeruginosa. *Environmental Microbiology Reports* 7**,** 330-340.

Ward, D.M., Ferris, M.J., Nold, S.C., and Bateson, M.M. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiology and Molecular Biology Reviews* 62**,** 1353-1370.

Watnick, P., and Kolter, R. (2000). Biofilm, city of microbes. *Journal of Bacteriology* 182**,** 2675-2679.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids. *Nature* 171**,** 737-738.

White, J.R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 5**,** e1000352.

Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E.M., Kyrpides, N., Mavrommatis, K., and Meyer, F. (2012). The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* 13**,** 141.

Wimpenny, J.W. (1992). "Microbial systems," in *Advances in Microbial Ecology*. Springer), 469-522.

Woese, C.R. (1987). Bacterial evolution. *Microbiological Reviews* 51**,** 221.

Wolfaardt, G., Lawrence, J., Robarts, R., and Caldwell, D. (1998). In situ characterization of biofilm exopolymers involved in the accumulation of chlorinated organics. *Microbial Ecology* 35**,** 213-223.

Wooley, J.C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput Biol* 6**,** e1000667.

Xu, J. (2011). "Microbial Ecology in the Age of Metagenomics," in *Handbook of Molecular Microbial Ecology, Volume I: Metagenomics and Complementary Approaches,* ed. F.J.D. Bruijn. First ed (John Wiley & Sons, Inc: Wiley-Blackwell).

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., Gilbert, J.A., Karsch-Mizrachi, I., Johnston, A., and Cochrane, G. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* 29**,** 415-420.

Zhang, H., Meltzer, P., and Davis, S. (2013). RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics* 14**,** 244.

Zhang, X., Bishop, P.L., and Kupferle, M.J. (1998). Measurement of polysaccharides and proteins in biofilm extracellular polymers. *Water Science and Technology* 37**,** 345-348.