# CHARACTERISATION OF DIVERSE MICROBIAL COMMUNITIES AND APPLICATION OF NOVEL DETECTION TECHNIQUES

KAISA KOSKINEN

REPORTS FROM THE DEPARTMENT OF ENVIRONMENTAL SCIENCES, LAHTI
FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES
UNIVERSITY OF HELSINKI

# Characterisation of Diverse Microbial Communities and Application of Novel Detection Techniques

**Kaisa Koskinen**

Institute of Biotechnology
and
Department of Environmental Sciences
Faculty of Biological and Environmental Sciences
University of Helsinki
Finland

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki in Auditorium 2 at Viikki Infocenter Korona (Viikinkaari 11), on November 29th 2013, at 12 o'clock noon.

**Supervisors**

Docent Petri Auvinen
Institute of Biotechnology
University of Helsinki
Finland

Professor Martin Romantschuk
Department of Environmental Sciences
Faculty of Biological and Environmental Sciences
University of Helsinki
Finland

**Pre-examiners**

Professor Olivia U. Mason
Earth, Ocean and Atmospheric Science
Florida State University
USA

Dr Alban Ramette
MPI for Marine Microbiology
Germany

**Opponent**

Professor Stefan Bertilsson
Department of Ecology and Genetics
Uppsala University
Sweden

**Custos**

Professor Rauni Strömmer
Department of Environmental Sciences
Faculty of Biological and Environmental Sciences
University of Helsinki
Finland

Layout: Tinde Päivärinta
ISBN 978-952-10-9322-7 (paperback)
ISBN 978-952-10-9323-4 (PDF; http://ethesis.helsinki.fi)
ISSN 1799-0580
Unigrafia
Helsinki 2013

# CONTENTS

# ABSTRACT

Microbes are essential for all life on Earth. They are found in all viable habitats from deep sea sediments and bedrock to high up in the atmosphere with a variety that exceeds by far the eukaryotic diversity. Ecosystem services provided by microorganisms, such as degradation of organic material and mediation of biogeochemical cycles are fundamentally important for the whole biosphere and its inhabitants. Microbes also form symbiotic relationships with multicellular organisms, and play important roles in nutrition and disease. Recent developments in molecular techniques, especially the next generation sequencing technologies and microarray applications, have opened new possibilities in studying diverse microbial communities.

In this thesis, the aim was to determine the diversity and community structure of environmental samples collected from the northern Baltic Sea water column and anaerobic digestion reactor, and to assess how the prevailing abiotic factors affect the microbial community structure. We applied 16S rRNA and ITS gene amplicon sequencing method with 454 sequencing technology to form a detailed taxonomic description of studied communities. The produced sequence data was further utilised in designing probes for a new padlock probe based ligation detection reaction (LDR) microarray that could be employed for specific and sensitive taxonomic identification of microbial groups in diverse communities. The functionality, specificity and sensitivity of the microarray were assessed using artificial and real environmental samples. Additionally, selected amplicon sequencing data analysis methods were compared in order to discover which algorithms work most reliably. In this subproject, we aimed to clarify how significantly the selected analysis methods, specifically denoising and clustering algorithms, affect the results and how comparable the results derived from different analysis pipelines are.

Amplicon sequencing revealed diverse microbial communities in the northern Baltic Sea water column and anaerobic digestion reactor. The pelagic bacterial communities in the northern Baltic Sea were strongly stratified, with aerobic Bacteria such as *Pseudomonas* and *Flavobacterium* dominating in the surface layer and *Oleispira* and sulfate-reducing bacteria in the anoxic deep waters. Based on the sequence data the diversity was assessed one order of magnitude less diverse compared to Atlantic and Pacific ocean bacterial communities. The anaerobic digestion reactor communities were dominated by Bacteria belonging to phyla *Bacteroidetes*, *Firmicutes* and *Thermotogae* and methanogenic Archaea, all essential and typical degraders in anaerobic digestion. The process also supported a diverse fungal community of phyla *Ascomycota* and *Basidiomycota*, including several taxa capable of degrading organic material in anaerobic conditions.

The LDR microarray technology proved sensitive, specific and semiquantitative method for identifying microbes in diverse communities. The proof of principle tests and experiments with real environmental samples showed that if the probes are designed carefully, the detection is comparable to qPCR and amplicon sequencing. The detection limit was 0.01 fmol/µl/template.

Data analysis method comparisons revealed prominent differences in observed operational taxonomic units and relative abundance of identified taxa. The majority of tested methods assessed the species richness too high. Using a functioning denoising method evened out the differences in the number of observed OTUs caused by various clustering algorithms. The ability to filter out the spurious taxa produced by amplification and sequencing, but still retain all the real diversity varied between methods.

This study shows both the potential and the challenges in the use of amplicon sequencing and microarray technologies in studying diverse microbial communities. The results indicate that the padlock based LDR microarray can be designed for very accurate and sensitive identification of microbial groups of interest. The data suggest that amplicon sequencing is a powerful tool in identifying microbes and assessing the diversity but distinguishing between spurious and true community members remain a challenge. There is still work to be done in the development and application of data analysis tools.

## LIST OF ORIGINAL ARTICLES

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

I **Koskinen K**, Hultman J, Paulin L, Auvinen P and Kankaanpää H 2011: Spatially differing bacterial communities in water columns of the northern Baltic Sea. FEMS Microbiology Ecology 75 (1):99-110

II Ritari J*, **Koskinen K***, Hultman J, Kurola JM, Kymäläinen M, Romantschuk M, Paulin L and Auvinen P 2012: Molecular analysis of meso- and thermophilic microbiota associated with anaerobic biowaste degradation. BMC Microbiology 12:121 doi:10.1186/1471-2180-12-121 (*shared first authorship)

III **Koskinen K**, Auvinen P, Björkroth J and Hultman J: Inconsistent denoising and clustering algorithms for amplicon sequence data – a benchmark study. Submitted manuscript.

## AUTHOR'S CONTRIBUTION

I KK prepared the samples for sequencing, analysed the data, interpreted the results and wrote the article with help of the co-authors.

II KK participated in designing the study, prepared the samples for sequencing, analysed the sequence data, contributed in probe design and participated in writing the article.

III KK participated in designing the study, analysed the data, interpreted the results and wrote the article with help of the co-authors.

# ABBREVIATIONS

| | |
|---|---|
| ACE | abundance-based coverage estimator |
| ATP | adenosine triphosphate |
| BLAST | Basic Local Alignment Search Tool |
| CPU | Central Processing Unit |
| DDBJ | DNA Databank of Japan |
| DNA | deoxyribonucleic acid |
| EBI | European Bioinformatics Institute |
| EMBL | European Molecular Biology Laboratory |
| ITS | internal transcribed spacer |
| LDR | ligation detection reaction |
| NGS | next generation sequencing |
| NCBI | National Center for Biotechnology Information |
| OTU | operational taxonomic unit |
| PCR | polymerase chain reaction |
| RDP | Ribosomal Database Project |
| RNA | ribonucleic acid |
| rRNA | ribosomal ribonucleic acid |
| UPGMA | unweighted pair group method with arithmetic mean |

# 1. INTRODUCTION

## 1.1. Microbial diversity

Biodiversity on Earth is composed of three domains of life. All cellular living organisms belong to either Archaea, Bacteria or Eukarya (Madigan et al, 2003). Microbes are a large and diverse group of microscopic organisms that can live as single cells or in cell clusters. All Bacteria and Archaea are microbes, but domain Eukarya contains microbes as well: Fungi, Protozoa and microscopic Algae (Torsvik & Øvreås, 2011). Before the discovery of Archaea in the 1970s (Woese & Fox, 1977), both Bacteria and Archaea were classified as Prokaryotes, indicating these organisms lack a nucleus unlike Eukaryotes. Prokaryotes are thought to be the first organisms on Earth and they represent the majority of life's genetic diversity (Whitman et al, 1998). It has been estimated that biosphere, the place on Earth's surface occupied by living organisms, carries approximately $10^{30}$ to $10^{31}$ microbes, which is two to three orders of magnitude more than the number of animal and plant cells combined (Whitman et al, 1998). Figure 1 illustrates the current knowledge in the three domains of life.

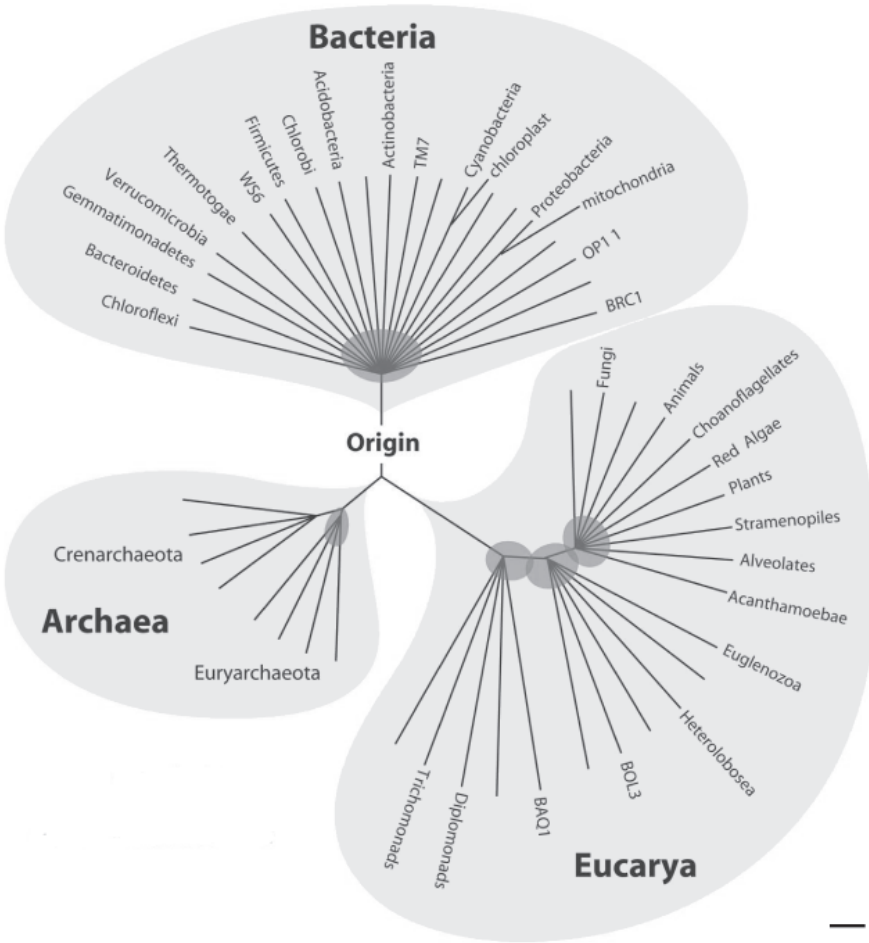


**Figure 1.** A simplified tree of life based on ribosomal RNA sequence comparisons. The grey circles represent the time points of unresolved branching order, and the scale bar indicates rRNA sequence change. Figure modified from Pace, 2009.

The first living organism appeared on Earth about 3.8 billion years ago and since then evolutionary processes have molded the bacterial and archaeal communities (Torsvik & Øvreås, 2011). In comparison, Eukaryotes have been around only for 1.6–2.1 billion years (Knoll, 1992). Bacteria and Archaea are found everywhere: in soil, marine and fresh water, air and deep in sediment of the oceans (Torsvik & Øvreås, 2011; Liu & Jansson, 2010). They provide essential ecosystem services and are crucial for all life forms on Earth. Microbes are critical in central biogeochemical processes, including carbon, nitrogen, sulphur, iron and manganese cycles, nutrient recycling as well as decomposing organic matter, and consequently influence the composition of the atmosphere and affect the climate (Madigan et al, 2003; Nagata, 2008). Their metabolic flexibility, possibility to short generation time and the capability to change genetic material even with distant relatives, make these organisms very adaptable, which has enabled them to occupy nearly every niche on Earth (Torsvik & Øvreås, 2011). Prokaryotes also form symbiotic relationships with multicellular organisms, such as plants and animals, and play important roles in nutrition and disease (Whitman et al, 1998).

### *1.1.1. Classification of organisms and concept of species*

Taxonomy is a field of science that classifies living organisms into groups based on their shared characteristics and provides names to the created groups. The hierarchical system of taxonomic classification and binomial nomenclature was originally developed by Carl von Linné, a Swedish scientist, and the basic idea of his invention is still in use (Linné, 1758-59; Hjelt, 1907). However, the methodology and the determinative characteristics have changed over the centuries, as molecular techniques have gained ground in taxonomy and species identification (Rossello-Mora & Amann, 2001).

#### *1.1.1.1 Taxonomy of Bacteria and Archaea*

Bacterial taxonomy as a discipline is new and dynamic compared to classification of higher organisms. Upon their discovery in the 17th century, Bacteria were described as "infusion animalcules" and no taxonomic classification was attempted (Rossello-Mora & Amann, 2001). Bacteria were seen as a single species that can occur in variety of sizes and shapes. Later scientists developed the classification based on morphology: a Danish scientist, Otto Friedrich Müller (1730-1784), was the first person to classify micro-organisms at the end of the 18th century. He proposed two genera, *Monas* and *Vibrio*, based on round and oblong form. A few decades later Christian Gottfried Ehrenberg (1795–1876) described five helical bacterial genera: *Bacterium*, *Vibrio*, *Spirochaeta*, *Spirillum* and *Spirodiscus*. The ability to isolate micro-organisms and grow them in pure cultures was a major advancement in microbiology and bacterial taxonomy in the late 19th century. Bacteria were easy to study in pure cultures and many tests for distinguishing bacteria by their phenotypic characteristics, such as cytological features, antibiotic susceptibility, chemical analysis, nutrition as well as metabolic, enzymic and reproductive features, were developed (Rossello-Mora & Amann, 2001). In the 1960s, as the awareness of DNA as the genetic material of living organisms increased, the idea of classifying bacteria based on their genomes became a viable option. At first, the base composition of genomes was compared and when mol% G+C values differed significantly, the two bacteria were classified in two different species. The resolution power of base composition comparison was not very effectual and more accurate methods were needed. Brenner and colleagues (1969) developed DNA-DNA hybridisation methodology which soon became a standard in bacterial taxonomy (Rossello-Mora & Amann, 2001; Rossello-Mora, 2006).

A significant breakthrough in taxonomy and molecular microbiology took place in the late 1970s and 1980s when the potential of ribosomal RNA sequences in determining the relationships between organisms was first discovered and applied (Rossello-Mora & Amann, 2001; Ludwig & Schleifer, 1994; Stackebrandt et al, 1985). Woese and Fox published a groundbreaking study in 1977 (Woese & Fox, 1977) describing the three domains of life. They discovered, based on 16S and 18S ribosomal RNA gene data, that all living organisms do not classify in the two domains, as generally accepted before the era of molecular phylogeny, but there is a third group of organisms which doesn't resemble Eukaryotes or prokaryotic Bacteria either in terms of ribosomal RNA sequence similarity or phenotypic features. This group of relatively unknown anaerobes had the expertise of reducing carbon dioxide into methane and was tentatively named *Archaebacteria*, later changed to Archaea to emphasise the distinction between Archaea and Bacteria (Woese & Fox, 1977). Currently Archaea comprises of four phyla: *Euryarchaeota*, *Crenarchaeota*, *Nanoarchaeota* and *Korarchaeota*. *Euryarchaeota* is the most diverse, encompassing all the methanogens and halophiles and *Crenarchaeota* includes Archaea thriving in extreme temperature conditions. *Nanoarchaeota* and *Korarchaeota* are small groups separate from other two phyla, but still little is known about their phylogeny and ecology. It has been even proposed that *Nanoarchaeota* and *Korarchaeota* do not exist as distinct phyla but belong within *Euryarchaeota* and *Crenarchaeota*, respectively (Barns et al, 1996; Huber et al, 2002; Swan & Valentine, 2009).

The species concept and the bases for classification have developed in parallel with advances in laboratory methods that have enabled the acquisition of more exact and reliable taxonomic information on Bacteria and Archaea. In the late 1980s, the Committee on Reconciliation of Approaches to Bacterial Systematics published a report about bacterial species classification (Wayne et al, 1987). They recommended that phylogenetic, descriptive, diagnostic and associative perspectives should be taken into account when proposing a new taxon. It means that a single characteristic can't be used alone for describing a new species, but a set of features that together clearly indicate that a new taxon has been discovered. These different features include information gathered using for example SDS-PAGE method, different chemotaxonomical markers, DNA-DNA similarities and 16S rRNA data (Rossello-Mora & Amann, 2001). The development of sequencing technologies and application of 16 rRNA gene information have not replaced the DNA reassociation that is especially invaluable in differentiating between closely related taxa. Still today, the "polyphasic approach" by combining as many techniques as possible is recommended (Rossello-Mora & Amann, 2001).

#### *1.1.1.2 Taxonomy of Fungi*

Taxonomic classification of Fungi has always been complicated. Very little is known about fungal evolution and even separating an individual fungus from a population may pose a challenge (Carlile et al, 2001). The phylogeny, especially at higher taxonomic levels, is largely unknown. Traditionally Fungi have been classified based on differences in morphological characteristics, and knowledge on ecology, breeding and the ability to infect have been essential in taxonomic classification (Guarro et al, 1999). The developments in DNA techniques have greatly influenced the fungal phylogeny, as ribosomal RNA based studies have seriously questioned the accuracy of conventional biology-based phylogeny. DNA sequence data reveals that Fungi developed as a part of terminal radiation of great eukaryotic groups and it has been suggested that Fungi may actually be more closely related to animals than plants. The main fungal lineages are called *Ascomycota*, *Basidiomycota*, *Zygomycota* and *Chytridiomycota* (Guarro et al, 1999).

#### 1.1.1.3. Concept of species

The concept of microbial species is difficult, if not impossible, to define (Riley, 2011; Cole et al, 2010). Sexually reproducing higher organisms belong to the same species if they are able to produce fertile progeny together (Mayr, 1982, 2001). Due to the asexual reproduction strategy in Bacteria, Archaea and many Fungi this rule does not apply (Madigan et al, 2003). Traditionally a microbial species has been characterised based on extensive set of distinct phenotypic traits (Rossello-Mora & Amann, 2001). Today, we know that using merely phenotypic characteristics for identification can be misleading, not only underestimating the diversity, but also misplacing organisms in the phylogenetic tree. Currently, when a new species is proposed, phenotypic characteristics, 16S rRNA sequence comparisons as well as DNA-DNA hybridisations are required (Rossello-Mora & Amann, 2001). These prerequisites hold the presumption that there is a pure culture of interest that could be subjected to various tests. However, the discovery rate of 16S rRNA sequences from uncultured organisms in natural environments using modern molecular techniques exceeds by far the cultured organisms, which may complicate reliable identification and correct taxonomic placement of sequences in the phylogenetic tree (Prosser et al, 2010). These challenges with species concept, taxonomy and nomenclature make the use of operational taxonomic units (OTUs) for binning sequences into microbial populations very attractive option. OTUs are defined based only on the similarity of sequence reads within a certain dataset (Wooley et al, 2010). An identity threshold is selected, most typically 97%, and all the sequences clustered into a certain OTU are handled as one "species". OTU based methods presume the same rate of evolution in 16S ribosomal RNA gene throughout all living organisms (Kunin et al, 2010; Ward et al, 2011).

### 1.1.2. Measuring microbial diversity

Microbial diversity can be defined by the number of species or different groups (e.g. OTUs) of microbes living in a certain environment, as well as the evenness of the species abundance distribution (Magurran, 2004). In natural environments microbial communities are typically complex and the diversity is difficult to assess and compare. In order to quantify the diversity as objectively as possible, a variety of diversity indices and richness estimates have been developed and applied (Magurran, 2004). These estimators present the diversity data as a single number that takes various aspects, depending of the indices used, of diversity into consideration. The diversity within individual samples or locations can be assessed using alpha diversity measurements whereas comparing the community membership and structure between samples or habitats is accomplished by applying beta diversity calculators (Magurran, 2004; Whittaker, 1960, 1972).

One aspect of alpha diversity is community richness, which describes the number of species present in a certain community (Magurran, 2004). The simplest way of representing the community richness is observed richness, i.e. how many species were observed using a given sampling effort. Regardless of sampled environment and organisms of interest, the observed richness is often far lower than the true richness, indicating that more extensive sampling would yield higher number of observed taxa. Therefore different richness estimate calculators, such as Chao1 (Chao, 1984), ACE (Chao & Lee, 1992), jackknife (Burnham & Overton, 1979; Heltshe & Forrester, 1983) and bootstrap (Smith & van Belle, 1984) estimators are applied in measuring the community richness. These nonparametric estimators use the species abundance and occurrence information and nonparametric model to estimate the total number of species present (Hortal et al, 2006).

In addition to species richness, another important aspect of alpha diversity is community evenness, which is a measure of the evenness in the distribution of species in a sample or environment (Magurran, 2004). The evenness in environmental sample can be represented by Pielou's evenness index which compares the maximum diversity and estimated diversity derived from Shannon index, and Heip's index of evenness which aims to measure evenness independently on species richness, working more reliably with communities carrying very low evenness (Magurran, 2004). The concept of community diversity retains both species richness and evenness. A more evenly distributed community is more diverse than a community with few dominant species but the same species richness. Community diversity can be assessed employing a variety of diversity indices including Shannon index and non-parametric Shannon index, Simpson index and inverse Simpson index (1/D) (Magurran, 2004).

Given that these richness estimates and diversity indices, as well as the whole ecological theory, were originally developed for studying macrobiota and not microbes using sequence data and OTUs, it is often challenging to assess the suitability of a certain estimate or index for a given dataset and research question, and draw the right conclusions. The relevance of singleton and doubleton OTUs in sequence dataset can easily be overestimated since they may be products of sequencing or PCR errors (Kunin et al, 2010; Huse et al, 2010; Quince et al, 2011). A more reliable approach for comparing alpha diversity between samples is rarefaction analysis combined with community diversity metrics. The rarefaction analysis compares the samples at a certain sequencing depth which eliminates the impact of sampling effort. It has been estimated that the Shannon index applied for rarefied data is a fairly reliable measure (Gihring et al, 2012).

A major goal in ecological research is recognising the processes causing spatial variation between communities. This variation is called beta diversity (Whittaker, 1960), which compares the membership and structure of multiple communities and quantifies differences in both taxon composition and relative abundance. As with alpha diversity, beta diversity can be presented using a range of indices. The main measures of β-diversity include Whittaker's measure $\beta_W$ (Whittaker, 1960, 1972), Cody's measure $\beta_C$ (Cody & Diamond, 1975), Routledge's measures $\beta_R$, $\beta_I$ and $\beta_E$ (Routledge, 1977, 1984) and Wilson & Shmida's measure $\beta_T$ (Wilson & Shmida, 1984). With modern sequencing technologies and large sequence datasets, programs such as LIBHUFF (Singleton et al, 2001), TreeClimber (Schloss & Handelsman, 2006) and UniFrac (Lozupone & Knight, 2005) have been applied in describing beta diversity. Various practical applications are also implemented in popular data analysis pipelines: in Mothur (Schloss et al, 2009) it is possible to compare the membership and structure of multiple communities by creating heatmaps, venn digrams, calculating the share of overlapping community members as well as drawing dendrograms and calculating the statistical significance of the clustering. Qiime (Caporaso et al, 2010) package contains many similar components. These tools are commonly used in visualising beta diversity (Magurran, 2004).

The fundamental goal in microbial ecology is to characterise the structure and function of all microbial communities, and to explain how abiotic and biotic environmental factors and interactions impact these communities (Gentry et al, 2006). This thesis concentrated on microbes in two distinct habitats, marine Bacteria in the northern Baltic Sea water column, and Bacteria, Archaea and Fungi in anaerobic digestion reactor and application and development of modern molecular detection techniques.

### 1.1.3. Microbial communities in marine environment

Microbes are ubiquitously distributed in marine and freshwater habitats on Earth (Liu & Jansson, 2010). It has been estimated that the cellular density for continental shelf and the upper 200 metres in the open ocean is approximately 5 x $10^5$ cells/ml of sea water (Kirchman, 2008). Marine bacteria have traditionally been difficult to culture (Joint et al, 2010), and consequently molecular techniques have been used to characterise the microbial community structure and their functional roles in the marine ecosystem.

Marine bacteria can be divided in eleven major lineages based on their small subunit ribosomal RNA (SSU rRNA) gene: *Cyanobacteria*, various classes of *Proteobacteria* (α,β,γ,δ), *Actinobacteria*, *Lentisphaerae*, *Bacteroidetes*, *Fibrobacter*, *Planctobacteria* and *Chloroflexi* (Miller & Wheeker, 2012). *Cyanobacteria* are probably the most well known marine bacteria, several members of which have a ubiquitous distribution in the marine environment, and are able to fix both nitrogen and carbon. A significant share of marine bacteria is affiliated to *Prochlorococcus* and *Synechococcus*, two major groups of *Cyanobacteria* (Kirchman, 2008). Autotrophic *Cyanobacteria* obtain their energy via photosynthesis: in the presence of carbon dioxide and water they convert the light energy from sun into chemical energy, i.e. sugar and oxygen. In oligotrophic oceans *Cyanobacteria* may account for nearly 90 per cent of the ecosystem's primary production (Kirchman, 2008). Some lineages of *Cyanobacteria* are also capable of producing harmful toxins which hinders the usage and especially the recreation possibilities of affected waters (Chorus et al, 2000; Halinen et al, 2007; Koskenniemi et al, 2007).

Another important group of marine bacteria is so called heterotrophic bacteria, containing taxa from all above mentioned major lineages. Heterotrophic bacteria are the most abundant cellular organisms in the entire biosphere (Kirchman, 2008). They are ubiquitous and abundant in all depths of the ocean and constitute a considerable proportion of microbial genetic diversity in marine habitats (Robinson, 2008). Heterotrophic bacteria need organic carbon-containing compounds as a source of energy and they are responsible for using most of the ocean's dissolved organic matter (Kirchman, 2008).

Ocean as a habitat contains numerous niches of different types for microbes. The water column that often looks uniform environment by naked eye is not static, but modified by various short and long term factors. Short term effects, such as weather, can change the environment rapidly, while longer term effects such as seasonal change affects environmental conditions especially in and near the polar regions, and cause temporal variation (Fuhrman & Hagström, 2008). The response of marine microbial communities to these temporal variations have been studied for decades and the results show that in some environments the community structure can stay very constant for a long time whereas in another habitat rather large daily fluctuations are detected (Acinas et al, 1997; Gilbert et al, 2009, 2012; Hewson et al, 2006a; Lee & Fuhrman, 1991; Riemann & Middelboe, 2002). Another aspect of variation in marine microbial communities at a global scale is spatial variation. It has been established that at the phylum level bacterial communities are rather similar in all localities, but there are only few cosmopolitan taxa at lower taxonomic level. Individual strains of classes *Alphaproteobacteria*, *Gammaproteobacteria*, *Deltaproteobacteria* and phylum *Bacteroidetes* have been found in all studied marine locations but none of the phylum *Firmicutes* (Fuhrman & Hagström, 2008). Probably the best known ubiquitously distributed bacterial species is *Pelagibacter ubique* which belongs to SAR11 clade and class *Alphaproteobacteria*. *Pelagibacter ubique* has been assessed as the most abundant organism on the planet

and was first discovered by Giovannoni and colleagues (1990) in Sargasso Sea by cloning and sequencing the 16S rRNA genes of an unculturable marine community. Using molecular techniques it has been found in almost every seawater sample surveyed since (Joint et al, 2010; Morris et al, 2002). SAR11 cluster is a diverse group of bacteria inhabiting different depths and latitudes (Fuhrman & Hagström, 2008), yet much of what we know about this extensive clade is based on analysis of *P. ubique*'s physiology and genome content (Joint et al, 2010).

Although there are several ubiquitous clades in the marine environment, there is always spatial variation in marine bacterial communities. Marine bacteria are largely affected by environmental parameters, such as temperature and salinity as well as abundance of other microbial groups (Winter et al, 2013). Patchiness may also be caused for example by ocean currents and mixing (Fuhrman & Hagström, 2008), and in coastal areas by rainfall and river discharge (Bouvier & del Giorgio, 2002). Community structure differences according to a certain pattern, for instance decreasing similarity with increasing geographic distance, have been observed in locations were the mixing occurs within geographical constraints, such as in straits, where environmental conditions form a gradient (Riemann & Middelboe, 2002; Hewson et al, 2006b). Differences in water density, strongly influenced by temperature and salinity, form a stratified water column where surface and bottom waters do not mix without vigorous turbulence (Eerola, 1993), and this affect bacterial communities at different depths. Consequently, in locations where vertical and horizontal gradients are present, patchiness and stratification in bacterial communities can be expected (Holmfeldt et al, 2009; Brown et al, 2009; Herlemann et al, 2011; Peura, 2012).

Marine Archaea were discovered in the 1990s, and information on their abundance and function has been received only recently (Ingalls et al, 2006; Karner et al, 2001). Archaea are found everywhere in the water column, but particularly abundantly in deep ocean and extreme environmental conditions, such as exceptionally high temperature and salinity, and low water availability (Kirchman, 2008; Fuhrman & Hagström, 2008). The most abundant planktonic marine Archaea include the first two groups found, *Crenarchaeota* and *Euryarchaeota*. Below 100 m in the ocean water column, *Crenarchaeota* biomass exceeds the total bacterial biomass. Marine Crenarchaeota are known to possess both heterotrophic and autotrophic capabilities and studies suggest that marine Archaea have a significant role in ammonia oxidation. Still, very little is known about these microscopic organisms and the processes they mediate in the deep ocean (Kirchman, 2008; Fuhrman & Hagström, 2008).

Marine microbes are by large responsible for biogeochemical cycles of carbon and nitrogen on Earth and the diversity and community structure of marine microbial communities is of great interest. The International Census of Marine Microbes (ICoMM: http://icomm.mbl.edu/) has initiated a world-wide effort to catalogue all known marine single-cell organisms including Bacteria, Archaea, Protista and associated viruses, study the unknown microbial diversity and explore the collected information in appropriate ecological and evolutionary context.

#### *1.1.3.1. Baltic Sea bacterial communities*

The Baltic Sea is the second largest brackish water basin in the world. Eutrophication, continuous oxygen deficiency and strong stratification of the water column are typical features of the Baltic Sea (Eerola, 1993). Thermocline and halocline prevent the mixing of oxygen-rich surface water and anoxic water from the bottom. Additionally, the presence of toxic hydrogen sulphide in deep waters has adverse effects on the benthic and demersal communities. Phosphate

emissions from land-based sources, internal loading of phosphorus and the abundance of nitrogen intensify the primary production and massive cyanobacterial blooms leading to sedimentation and decomposition, further contributing to anoxia and deterioration of benthic communities (Eerola, 1993). Various anthropogenic pollutants affect the Baltic Sea environment, including industrial chemicals and oil originating from various sources (NAS, 2003). These features make the Baltic Sea a unique environment for a sparse selection of freshwater and marine organisms, migratory species, and glacial relicts (Ojaveer et al, 2010).

Cyanobacterial communities in the area, including toxic and problematic blooms, have been studied for a long time and information on the abundance and dynamics of small-sized picocyanobacteria *Synechococcus* and larger groups such as *Nodularia*, *Aphanizomenon* and *Anabaena* is presented in many publications (Koskenniemi et al, 2007; Andersson et al, 2010; Sivonen et al, 1989a, 1989b; Stal et al, 2003). Some information is also available on the heterotrophic bacterial abundance and activity. It has been detected that the structure of bacterial communities in water column changes according to the seasons. Rather limited data suggest that *Cytophaga-Flexibacter-Bacteroides* group and different classes of *Proteobacteria* dominate the Baltic Sea brackish water bacterial communities, and members of genera *Sphingomonas*, *Pseudomonas* and *Shewanella* have been detected as strikingly common (Riemann et al, 2008; Hagström et al, 2000). Typical freshwater taxa within classes *Actinobacteria*, *Verrucomicrobia*, and *Betaproteobacteria* are also common and detected in the Baltic Sea water column (Holmfeldt et al, 2009; Riemann et al, 2008). However, characteristic marine genera within *Gammaproteobacteria*, such as *Vibrio*, *Pseudoalteromonas*, and *Alteromonas*, and *Roseobacter* within *Alphaproteobacteria* have been conspicuous by their absence (Hagström et al, 2000). Additionally, bacterial diversity in certain special environments such as chemical weapons dumping sites and littoral sediments has also been described in research papers. The methodology in these studies, focused both on water column and sediments, (Holmfeldt et al, 2009; Riemann et al, 2008; Hagström et al, 2000; Edlund et al, 2006, 2008; Edlund & Jansson, 2008; Medvedeva et al, 2009) relies mostly on culturing, fingerprinting, and cloning and sequencing relatively small datasets. These approaches are likely to limit the observations to abundant Bacteria or a certain bacterial groups of special interest leaving unknown, unculturable, rare and seemingly insignificant bacterial groups overlooked. Only since the development of next generation sequencing technologies, has there been a growing interest towards the whole bacterial community, including both rare and abundant community members. Andersson and colleagues (2010) published the first description of the Baltic Sea bacterial communities using pyrosequencing. They found that the samples were dominated, to some extent, by the same lineages as the ocean, for example *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria* and *Proteobacteria*, but also a substantial share of the sequence reads was affiliated to phylogenetic groups characteristic for freshwater ecosystems (Andersson et al, 2010). They also detected annually reoccurring succession patterns within the bacterial communities, the same phenomenon that had been observed in several other marine and freshwater habitats using fingerprinting methods, thus focusing merely on the abundant taxa. Although the short read length of 454 GS-20 technology they used allowed only high taxonomic level identification, it was the first attempt for a highly detailed description of the Baltic Sea bacterial communities (Andersson et al, 2010).

### *1.1.4. Microbial communities in anaerobic digesters*

Anaerobic digestion is a decomposition process where organic material is degraded in the absence of oxygen by numerous different

groups of microorganisms (Hobson & Wheatley, 1993). The process can be divided into four key biological and chemical stages: hydrolysis, acidogenesis, acetogenesis and methanogenesis (Bitton, 2010). At each stage, the process is driven by a specialised group of microbes. In hydrolysis fermentative heterotrophic bacteria break the complex organic molecules down into simple sugars, amino acids and fatty acids. In acidogenesis heterotrophic syntrophic bacteria degrade the simple sugars, amino acids and fatty acids to form carbonic acids, alcohols, hydrogen, carbon dioxide and ammonia. In acetogenesis the simple molecules created in acidogenesis are further digested by acetogenic bacteria to produce acetic acid, carbon dioxide and hydrogen (Bitton, 2010; Okabe & Kamagata, 2010). In methanogenesis Archaea use the products of the preceding stages and convert them into methane, carbon dioxide, and water, which together make up the majority of the biogas emitted from the system (Okabe & Kamagata, 2010; Zehnder, 1978). Consequently, the diversity and community structure of microbes in the process have a major impact on the process efficiency and functionality.

The microbial community composition in anaerobic reactor is greatly influenced by the substrate composition (Hobson & Wheatley, 1993), reactor design as well as operating conditions (McHugh et al, 2003). Temperature seems to affect the microbial community: in mesophilic (temperature about 35 °C) conditions the species richness appears higher and the species composition different compared to thermophilic (temperature about 55 - 60 °C) reactor conditions (Levén et al, 2007). However, the process performance is somewhat equally effective in both temperatures, apart from the more efficient degradation of a few specific compounds and the elimination of pathogens at higher temperatures (Bagge et al, 2005; Levén et al, 2012).

The abundance and distribution of Bacteria and Archaea in anaerobic digestion processes have been studied by several research groups (e.g. McHugh et al, 2003; Levén et al, 2007; Pycke et al, 2011; Riviere et al, 2009; Ros et al, 2013; Sasaki et al, 2011; Schlüter et al, 2008; Shin et al, 2010) and the process is known to carry a diverse community of heterotrophic and syntrophic Bacteria and methanogenic Archaea: the bacterial phyla *Bacteroidetes*, *Chloroflexi*, *Thermotogae*, *Firmicutes*, *Proteobacteria*, *Spirochaetes*, *Actinobacteria*, *Synergistes*, *Planctomycetes*, *Verrucomicrobia*, *Acidobacteria* and *Nitrospira* as well as candidate divisions OD1, PO10, OP9, OP8, OP3, TM6, TM7, EM3 and BA024 have been detected in samples taken from anaerobic digestors (e.g. Okabe & Kamagata, 2010; Levén et al, 2007; Riviere et al, 2009; Levén, 2006). The Archaea present in anaerobic digestion are mainly affiliated to phylum *Euryarchaeota*, but also *Crenarchaeota* have been detected. The often observed archaeal genera in anaerobic reactors include methanogenic *Methanospirillum*, *Methanosarcina*, *Methanoculleus*, *Methanomethylovorans*, *Methanosaeta*, *Methanobrevibacter* and *Methanothermobacter* (e.g. Okabe & Kamagata, 2010; Riviere et al, 2009; Ros et al, 2013; Shin et al, 2010; Levén, 2006).

The analysis of Fungi in anaerobic reactors has typically been limited to pathogenic Fungi and their survival during the digestion (Schnürer & Schnürer, 2006). It has been reported that the quantity of colony forming fungal cells is not reduced during the anaerobic digestion process but the diversity is decreased (Schnürer & Schnürer, 2006). However, a wide variety of fungal lineages can be introduced to the reactor along with the substrate, and since there are fungal groups capable of degrading organic material in anoxic environment, it is surprising that Fungi has attracted so little attention as a functional part of anaerobic digestion (Hobson & Wheatley, 1993; Dumitru et al, 2004; Jennings, 1995; Kinsey et al, 2003).

## 1.2. Molecular methods in studying microbial diversity

Biological study has always been technology driven, and microbial ecology is no exception. Major advances in the field have followed revolutionary instrument innovations, such as the microscope, the discovery of DNA, PCR and most recently the advent of next generation sequencing technology (Margulies et al, 2005; Shendure et al, 2005; Xu, 2011). The current range of molecular methods available is extensive, covering all the techniques relying on extracted community DNA, RNA and proteins, but not requiring cultivation of pure microbial cultures (Prosser et al, 2010). Today, molecular methods are applied to determine the identities and functions of microbes in diverse communities, with a focus on the whole community and its interactions. This is in contrast to traditional methods that typically separated the microbe of interest from its natural physical and biological environment (Prosser et al, 2010). Before the advent of next generation sequencing technologies the term "molecular methods" typically meant either fingerprinting or cloning and sequencing based techniques (Oros-Sichler et al, 2007). Fingerprinting methods are PCR based techniques that in most cases detect differences in the marker gene among diverse community by electrophoretic separation of produced DNA fragments (Oros-Sichler et al, 2007). The resolution power and identification precision of community members are often relatively constricted. The number of microbial groups, OTUs, detected by fingerprinting methods is usually far lower than the actual richness of natural communities with long-tailed rank abundance distribution and the diversity indices calculated based on fingerprints do not provide reliable information on the true diversity (Bent et al, 2007). Consequently, these methods work best as a rapid and comparative analysis of multiple samples, for comparing the diversity and community structure of samples with fairly considerable differences along environmental gradients or experimental factors, or for characterising rather simple and well-studied communities (e.g. Juottonen et al, 2008; Nieminen et al, 2012b). With diverse environments, such as sea water and soil with potentially thousands species per gram of sample and potentially billions of microorganisms (Kirchman, 2008; Daniel, 2011) the resolution power is typically not sufficient. To unravel this great diversity, modern and more powerful estimation methods, such as next generation sequencing and microarray technologies (Margulies et al, 2005; Bentley, 2006; Bentley et al, 2008; Rothberg et al, 2011; Andersen et al, 2010), are applied.

Various molecular methods utilise ribosomal RNA genes or the intergenic transcribed spacer (ITS) regions between ribosomal RNA genes and their special characteristics when studying microbial diversity and phylogeny (Cole et al, 2010; Prosser et al, 2010). Ribosomes are complex molecular devices found in all cells of living organisms, and fundamentally important as primary sites of mRNA translation and thereby, protein synthesis. In Bacteria and Archaea ribosomes consist of a large 50S subunit and a small 30S subunit (Pei et al, 2011). These subunits are complexes with ribosomal RNAs and a set of essential ribosomal proteins. The large subunit contains 23S and 5S rRNA genes and the small subunit includes 16S rRNA which has been widely used in phylogeny (Pei et al, 2011). Eukaryotic ribosomes are composed of 18S (small subunit), 5S, 5.8S and 28S (large subunit) genes (Verschoor et al, 1996) (Figure 2).

The 16S rRNA gene contains nine variable regions and conserved regions in between (Van de Peer et al, 1996). The conserved regions can be used in classifying very distant relatives and the variable regions are used in classifying more closely related groups of organisms, and reliable phylogenetic trees have been constructed based on 16S rRNA gene (Pei et al, 2011; Noller, 1984). Employing rRNA genes in microbial diversity assessment

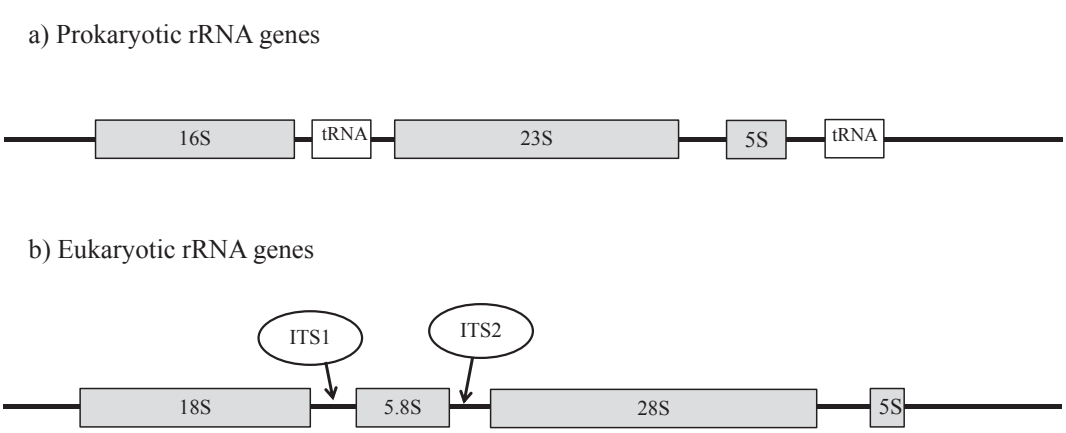


**Figure 2.** A schematic illustration of canonical prokaryotic and eukaryotic rRNA operons.

studies in natural environments was first applied in early 1980s by Norman Pace and colleagues (Stahl et al, 1985), when they used 5S rRNA to characterise the microbial community of a Yellowstone hot spring.

There are several reasons why the small subunit ribosomal RNA gene is ideal for studying evolutionary relationships: ribosomal RNA genes are found in all organisms, major changes in ribosomal RNA gene due to horizontal gene transfer are unlikely since it would disrupt the gene function, and consequently the evolutionary history of organisms can be reconstructed fairly reliably from rRNA gene sequence (Pei et al, 2011). Additionally, a vast amount of rRNA gene sequence data is publicly available in sequence databases, such as Ribosomal Database Project (RDP, rdp.cme.msu.edu) (Cole et al, 2009), Max Planck Institute for Marine Microbiology's Silva database (www.arb-silva.de) (Quast et al, 2013), Lawrence Berkeley National Laboratory's Greengenes (greengenes.lbl.gov) (DeSantis et al, 2006a) and National Institute of Health's NCBI (www.ncbi.nlm.nih.gov) (Benson et al, 2009). There are also established broad range primers targeting the conserved regions in 16S rRNA gene that amplify most of the known bacterial and archaeal community members, as well as primers targeting specific groups or rare taxa (Prosser et al, 2010). However, there are also pitfalls that need to be taken into account when applying 16S rRNA gene based approaches: First, the number of ribosomal operons may vary between 1 and 15 in a single genome (Klappenbach et al, 2001) and the intra-genomic variability between the copies can be up to several percentages. This may complicate the identification of microbes of interest and lead to the spurious classification of one organism into more than one taxon (von Wintzingerode et al, 1997). Second, even if major horizontal gene transfer is categorised as unlikely (Pei et al, 2011), it has been suggested that the functioning of the ribosome is largely dependent on its secondary structure and not on the 16S rRNA gene nucleotide sequence itself (Kitahara et al, 2012), indicating that a functioning ribosome may have received genomic material through horizontal gene transfer which could lead to misleading phylogenetic signal. Additionally, the rather conserved nature of the gene hinders the reliable discrimination between closely related species or strains, and thus fine ecological adaptations and speciation generally cannot be detected (Cole et al, 2010).

### *1.2.1. DNA sequencing*

DNA sequencing is a process of determining the order of nucleotides in a stretch of DNA. The first sequencing methods were published

in the 1970s when first Sanger and colleagues and then Maxam and Gilbert published their sequencing methods (Sanger & Coulson, 1975; Maxam & Gilbert, 1977). At first upon its publication in 1977, the Maxam-Gilbert sequencing system became more popular because of the ease to use compared to the first Sanger method. However, later the same year Sanger and colleagues published a new sequencing strategy relying on sequencing by synthesis and natural 2'-deoxynucleotides (dNTP) and chain terminating 2',3'-dideoxynucleotides (ddNTPs) (Sanger et al, 1977). This new Sanger sequencing method soon became the most popular way of determining the nucleotide sequence and later improvements with automation in laboratory work and electrophoresis allowed parallelisation and more throughput (Metzker, 2005). The Sanger method dominated the field for several decades.

Sanger sequencing method was the first sequencing technique applied in metagenomic studies of natural environments (Hugenholtz & Tyson, 2008). Despite the automation in methodology, the list of required steps in protocol is long and the stages laborious: the gene of interest needs to be amplified or the environmental DNA fragmented. Then the clone libraries are constructed of the resulting amplicons or DNA fragments and these fragments are sequenced individually (Metzker, 2005). However, Sanger sequencing method has been widely applied with functional genes and community structure and diversity studies using 16S rRNA gene (Rastogi & Sani, 2011), and it is still in use in many laboratories. Yet, it is not feasible to sequence tens of thousands or millions sequences per project using Sanger sequencing and therefore the rare microbial groups in studied environment are easily overlooked (Rastogi & Sani, 2011). Consequently, the limiting factors such as cost and laboriousness and the significant PCR and cloning bias related to the method led to a need for new and more powerful technologies for sequencing.

Over the last few years the next generation sequencing (NGS) technologies have revolutionised the field of genome and community sequencing, and consequently microbial ecology (Metzker, 2010). NGS technologies typically pursue high throughput by parallelising the sequencing process, producing thousands to millions of sequences concurrently. The major advance in NGS is indeed the huge amount of data produced at considerably low cost compared to Sanger method (Rastogi & Sani, 2011; Metzker, 2010). Projects that took years or months to finish with Sanger sequencing can now be carried out in days, at a fraction of the cost. Divergent features of different NGS technologies, such as read length, throughput and differing error types (homopolymer, insertion/deletion, substitution, random) facilitates the coexisting of multiple platforms in the market (Ståhl & Lundeberg, 2012). Decisions as to which NGS platform to use depends on various factors, and beyond cost considerations, the most important of which is study setup and objectives, along with what kind of error schema is the least harmful for the research problem in question. The drawback of most of these new technologies is their short read length (Table 1) (Ståhl & Lundeberg, 2012; Mardis, 2008).

Currently DNA sequencing is one of the most rapidly developing technology fields of biology. Future advances promise single-molecule sensitivity, uninterrupted real-time sequencing and low cost, but these innovations depend upon the development of advanced micro- and nanostructures (Ståhl & Lundeberg, 2012).

#### *1.2.1.1. 454 pyrosequencing*

The first next generation sequencing platform and commercial alternative to Sanger sequencing was Genome Sequencer GS20 by the company 454 Life Sciences, later bought by Roche. The method bases on pyrosequencing that was invented and developed by Nyrén and colleagues (Nyrén, 2001; Ronaghi et al, 1996, 1998). The first automated pyrosequencing system was published and sold already in

**Table 1.** Overview of next generation sequencing technologies

| DNA sequencing platforms | read length | M sequences /run | throughput Gb/day | throughput Gb/run | typical error type | references |
|---|---|---|---|---|---|---|
| GS FLX Titanium, 454/Roche | 450 | 0,7 1 | 0,45 | 0,45 | homopolymer | Margulies et al, 2005 |
| GS FLX+, 454/Roche | 700 | 1 | 0,7 | 0,7 | homopolymer | |
| HiSeq 2000, Illumina | 2 x 100 | 6000 | 54 | 600 | substitutions | Bentley 2006, 2008 |
| HiSeq 2500, Illumina | 2 x 150 | 1200 | 45 | 600 | substitutions | |
| MiSeq, Illumina | 2 x 250 | 16 | 4,5 5 | 7,5 8,5 | substitutions | |
| SOLiD 5500xl, Life Technologies | 75 + 35 / 2 x 60 | 2800 | 20 30 | 180 | substitutions | Shendure et al, 2005 |
| PGM, Ion Torrent/ Life Technologies | 400 | 4 5,5 | 3 6 | 1 2 | homopolymer | Rothberg et al, 2011 |
| Ion Proton | 200 | 60 80 | 60 120 | 10 | homopolymer | |
| PacBio *RS*, Pacific Biosciences | 5000 | 0,5 | 2,5 | 1,5 | indels | Eid et al, 2009 |
| Oxford Nanopore Technologies | 10 000* | ND | ND | ND | indels | Clarke et al, 2009 |

* estimated values for the sake of comparison (Ståhl & Lundeberg 2012)
ND=not determined

1999 but this first instrument was restricted to sequence only one sample at a time. Parallelisation was improved and later version was capable of handling 96 samples in one run (Nyrén, 2007). The significant step was taken in year 2005 when Rothberg and colleagues (Margulies et al, 2005) published a new technique of pyrosequencing, a highly parallel sequencing system with substantially larger throughput compared to previous pyrosequencing or capillary electrophoresis sequencing instruments (Margulies et al, 2005).

454-pyrosequencing is based on sequencing by synthesis and the process is monitored by charge-coupled device (CCD) camera (Metzker, 2010). Adapters are attached to fragments of correct size which are captured on micron-scale beads and amplified in water in oil emulsion, a process called emulsion PCR (Dressman et al, 2003). Next, the beads containing the copied fragments are deposited in picoliter-scale wells on a picotiter plate. The reaction mixture in wells also contains enzymes DNA polymerase, ATP sulfurylase, luciferase, apyrase and adenosine5'-phosphosulfate and luciferin as substrates. At each of 1779 cycles (current FLX+ pattern B) one of the four nucleotides is introduced and resulting incorporation releases pyrophosphate inducing a burst of light via ATP sulfurylase and luciferin. Apyrase degrades the unincorporated nucleotides and ATP before the next nucleotide is added (Margulies et al, 2005). The pattern of detected incorporation events reveals the nucleotide sequence of individual templates. The raw data is a series of images which are quantised and normalised, and converted into flowgrams. Flowgram data is the starting point for sequence analysis (Shendure & Ji, 2008).

#### *1.2.1.2. 454 in environmental microbiology*

454-pyrosequencing is highly advantageous in many fields of science: de novo sequencing and assembly of genomes and metagenomes, targeted resequencing, transcriptome sequencing of cells, tissues and entire organisms (RNA-seq), gene discovery and studying diverse microbial communities by metagenomics and metatranscriptomics (http://454.com/). All these approaches

are applicable in medical, biological and environmental research as well, but especially in environmental microbiology metagenomics has become a very popular tool in studying diverse microbial communities (Nelson et al, 2010). 454-pyrosequencing has been applied in studying the community structure and function in countless habitats on land and sea, including natural grassland soil metagenome (Delmont et al, 2012), both freshwater and marine communities (e.g. Dinsdale et al, 2008; Biddle et al, 2008; Nakai et al, 2011), as well as assessing the feasibility of metatranscriptome sequencing in studying the active marine microbe community and gene expression by combining metatranscriptomic and metagenomic approaches (Gilbert et al, 2008; Frias-Lopez et al, 2008). The term metagenomics was first used by Handelsman and colleagues (1998) and it stands for studying the whole microbial community structure and function based on sequence data derived from genetic material extracted directly from samples. Consequently, the sample may contain a variety of organisms such as bacteria, archaea, viruses, fungi, plants and animals, and the produced sequence data is a fragmented puzzle of all the genes and intergenic regions from all organisms in the sample (Xu, 2011). Therefore, the data includes large amount of information on the whole community structure and function. Currently however, the most widely applied next generation sequencing method in environmental microbiology is called amplicon sequencing.

#### *1.2.1.3. 454 amplicon sequencing*

In amplicon sequencing, a specific genetic region shared by the members of the microbial community, is amplified using universal primers to produce fragments of similar length suitable for sequencing. Amplification success is expected to be unbiased across taxa, and thus each sequence read represents a random sample of the genetic diversity in the sample DNA. Often the amplified fragment is a region in 16S ribosomal RNA gene or other marker gene (Cole et al, 2010; Huber et al, 2007; Sogin et al, 2006). With current 454 sequencing technology, usually two or three variable regions in 16S rRNA gene are selected for amplification.

The amplicon sequencing method with 454 for studying diverse microbial communities was first applied by Sogin and colleagues (2006) when they published a study of bacterial diversity in North Atlantic Deep Water and Axial Seamount in the northeast Pacific Ocean. They designed universal primers flanking the V6 region of bacterial 16S rRNA gene and included 454 Life Science's A and B sequencing adapters to the 5' end of the primers. They amplified the sample DNA, prepared the libraries and ran the eight samples in 16 separate lanes recovering almost 120 000 sequence reads representing the forward and reverse reads of the samples (Sogin et al, 2006). The rarefaction analysis and the abundance-based coverage estimator ACE and the Chao1 estimator of species diversity indicated at least an order of magnitude greater bacterial diversity in studied samples than in any published article about microbial community diversity at the time. The results suggested that additional sampling would result in significant increase in observed operational taxonomic units (Sogin et al, 2006).

One year later, the same group of scientists published a new study about the bacterial and archaeal diversity of deep sea hydrothermal vents in northeast Pacific Ocean - with improved methodology (Huber et al, 2007). They designed a set of unique 5 nucleotides long sequence tags between the 454 Life Sciences A sequence adapter and the forward primer (Figure 3). Every sample was amplified with a forward primer carrying a different tag sequence and the resulting sequence data could be sorted to samples using bioinformatics tools and the tag sequence information. This way, they were able to run 16 samples in one sequencing run without reducing the area at disposal on the picotiter

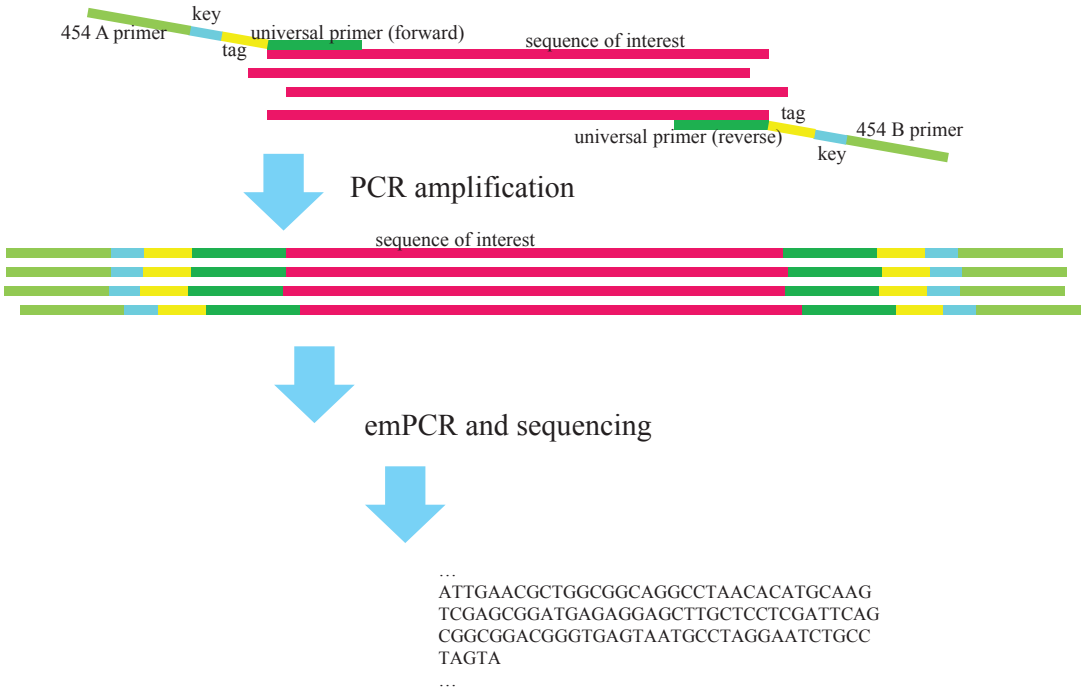


**Figure 3.** A schematic picture of amplicon sequencing method.

plate by partitioning gaskets, thus increasing the number of wells available for sequencing (Huber et al, 2007).

Since the revolutionary publication by Huber and colleagues (2007) 454 amplicon sequencing has been applied to almost all imaginable environments from deep sea and deep biosphere (Nyyssönen et al, *in press*) to upper troposphere (DeLeon-Rodriguez et al, 2013). There are publications about microbial diversity in domestic showerheads (Vornhagen et al, 2013), human body parts such as belly buttons (Hulcr et al, 2012), skin (Hanski et al, 2012) and mouth (Huang et al, 2011), cow rumen (Fouts et al, 2012; Mao et al, 2012), intestines of various animals (Le Roy et al, 2012; Su et al, 2013), food products (Nam et al, 2012; Nieminen et al, 2012a) as well as soil, water and sediment (e.g. Comeau et al, 2012; Deng et al, 2012; Siam et al, 2012; Zhang et al, 2012).

### *1.2.2. DNA microarrays*

DNA microarrays are a parallel detection tool that consists of multiple microscopic spots containing oligonucleotide probes that hybridise to their specific complementary targets (Andersen et al, 2010), and a solid platform on which the probes are immobilised (Blalock, 2003). The presence of sequence of interest in studied sample is indicated by fluorescent signal detected by a special microarray scanner (Andersen et al, 2010; Blalock, 2003). DNA microarrays were originally developed for whole genome gene expression study purposes (Schena et al, 1995) but the method has become a popular tool with a lot of potential in microbiological research as well. Microbial diagnostic microarrays are of great value in environmental microbiology especially in characterising microbial community structure and function of diverse samples. As a high throughput method, microarrays allow parallel identification of microbes and genes from several samples simultaneously which makes them an efficient

and rapid tool for monitoring changes in diverse microbial communities (Andersen et al, 2010; Bodrossy & Sessitsch, 2004).

There are several commercial microarrays readily available, such as PhyloChip (Brodie et al, 2006) for phylogenetic identification of most known Bacteria and Archaea based on their 16S rRNA genes, and GeoChip (He et al, 2007) targeting various functional genes responsible for instance carbon, nitrogen, sulfur, and phosphorus cycling, metal resistance and reduction, as well as organic contaminant degradation (Andersen et al, 2010). The advantages of microarray technology include the rapidness and detection of rare taxa. In theory, microarrays are not as affected by abundant microbial groups as standard amplicon sequencing because every hybridisation between the probe and target is an independent process (Brodie, 2011). However, as PhyloChip contains hundreds of thousands of ~25mer oligonucleotide probes and GeoChip more than twenty four thousand 50mers, potential cross-hybridisation between nearly identical sequences and probes may skew the results, even though this problem has been considered carefully in probe design (Andersen et al, 2010). A possible disadvantage is that only targeted microbial groups can be detected and no new groups are found (Andersen et al, 2010).

#### *1.2.2.1. Phylogenetic microarrays*

Phylogenetic microarrays are used as a high throughput tool for identifying microbes in diverse samples (Andersen et al, 2010). There are several different types of phylogenetic microarrays to meet the differing demands (Brodie, 2011). The probes can be in situ hybridised or spotted, and depending on the probe density one can study the presence and absence of tens to thousands of microbial taxa in a sample with one single experiment. The probes can be designed to target a phylogenetically broad group of microbes, such as class, or very strictly target one single strain of interest (Brodie, 2011). If the commercially available microarrays are not applicable in the project in question it is possible design a phylogenetic microarray for a specific environment, as long as certain matters have been taken into account: The oligonucleotide probes must be designed with great care. The melting temperature must be nearly the same, for example by designing probes with similar length, using directed modification of spacer lengths and adding tertiary amine salts to the hybridisation buffer. Still, often there are considerable differences in maximal hybridisation capacity between the probes (Blalock, 2003; Loy et al, 2002; Peplies et al, 2003).

#### *1.2.2.2. Sensitivity and specificity*

Albeit the phylogenetic microarrays hold a lot of promise there are certain limitations and problems, such as sensitivity and specificity that may restrict their use (Wagner et al, 2007). Sensitivity refers to the quantities of target DNA that can be detected, and various factors may affect: long probes allow better affinity to the target molecule and consequently higher sensitivity, but unspecific hybridising may occur (Lomakin & Frank-Kamenetskii, 1998; Öhrmalm et al, 2010). Several studies have shown that a microbial group or species can only be detected by a microarray method if its copy number in studied environment is large enough and its relative abundance more than 0.04-5% of the whole community (Loy et al, 2002; Wagner et al, 2007; Franke-Whittle et al, 2005; Hultman et al, 2008; Palmer et al, 2006). Consequently, choosing a phylogenetic microarray for a research method is reasonable when the species/populations of interest are not among the rarest in studied environment. Fortunately, the sensitivity can be improved, within limits, by taking a good notice of the quality and sensitivity of probes and detection appliances and optimising the laboratory protocols (Wagner et al, 2007). The other possibility is to amplify selectively the target gene fragments prior to microarray

analysis using primers specific for species or population of interest (Loy et al, 2005).

Specificity denotes how well the probe hybridises to its fully matching target sequence. Short probes result in better target specificity and even one nucleotide mismatch can be discriminated, but lower affinity may cause false negative observations (Lomakin & Frank-Kamenetskii, 1998; Öhrmalm et al, 2010). Especially, in natural microbial communities that often harbour closely related taxa that differ only by one nucleotide within the probe's target region the specificity is of primary importance (Andersen et al, 2010). Since the optimal hybridisation conditions (hybridisation and washing buffer compositions and temperature) typically vary between oligonucleotide probes, the employed conditions are not working equally and some probes will not produce a signal in the presence of a perfect match target (Wagner et al, 2007). However, there are ways to minimise the effect of differing hybridisation capacities and to discriminate between two targets with only one nucleotide difference. The imperfect specificity can be circumvented for example by applying multiple probes per one target: The set of probes may possess hierarchical or parallel specificity, or a probe containing a single-base mismatch, providing a negative control for each probe on the array. The presence of target organism is then confirmed by positive signal in sufficiently many or all matching probes or with control probes (Brodie et al, 2006; Loy et al, 2002; DeSantis et al, 2005; Liu et al, 2001). Another approach to increase the specificity with multiple probes is so called ligation detection reaction (Baner et al, 2003; Busti et al, 2002; Castiglioni et al, 2004). Furthermore, using distinct control probes have been proven to reduce the problems caused by noise, and true positive signals are easier to distinguish with improved statistical significance (Ritari et al, 2009).

#### *1.2.2.3. Universal microarrays*

The above mentioned problems with oligonucleotide probes displaying different maximal hybridisation capacities can be avoided by using universal microarrays. Universal microarrays do not have oligonucleotide probes targeting the microbial groups of interest in samples but a selection of artificial sequences called Zip-codes (Gerry et al, 1999). These Zip-code sequences are designed to have similar thermodynamic properties and the hybridisation can be performed at a certain temperature leading to a more stringent and rapid hybridisation (Gerry et al, 1999). The Zip-codes target molecules containing a complement piece of sequence in studied sample, and consequently the method needs to be linked with e.g. ligation detection reaction approach. A universal microarray can be used for any microbe community, given the selected probe set is designed to target the microbial groups of interest in studied samples (Gerry et al, 1999; Favis et al, 2000).

#### *1.2.2.4. Ligation detection reaction (LDR)*

Ligation detection reaction microarray approach was first applied for detecting low abundance point mutations and small insertions and deletions in cancer cells (Gerry et al, 1999; Favis et al, 2000). The method is based on two probes, one discriminative and one common probe. The discriminating probe carries a fluorescent label on its 5' end and the 5' phosphorylated common probe contains a Zip-code on its 3' end (Busti et al, 2002). These two probes are designed to hybridise adjacently to their target sequence in the sample. The target sequence is typically an amplified fragment from a marker gene: in environmental microbiology 16S rRNA or ITS gene is often amplified with universal primers in order to characterise the microbial community structure of samples (Wagner et al, 2007; Hultman et al, 2008). If the probes' target is present and hybridisation occurs to the perfectly matching template, the 3' end of the discriminating probe and 5' end

of the common probe are ligated together by DNA ligase (Busti et al, 2002). Even one nucleotide mismatch at 3' terminal position of the discriminating probe can prevent the ligation (Khanna et al, 1999). The ligated products, now carrying fluorescent label on their 5' end and the Zip-code on their 3' end, are hybridised on a universal microarray containing the complementary Zip-code sequences (Gerry et al, 1999). The positions on microarray containing the hybridised ligation product can be detected by laser scanning and identifying the locations (Gerry et al, 1999; Favis et al, 2000). The basic idea of the ligation detection reaction microarray is presented in Figure 4.

As with any microarray technology, probe design is crucial with LDR. The probes need to target the sequence or microbial group of interest, but no other groups (Andersen et al, 2010). The difficulty with environmental samples is the huge amount of unknown organisms in the samples for analysis, particularly the rare taxa, and the great amount of organisms that have not been sequenced, and thus can't be found in sequence databases. These challenges may result in selection of wrong microbes of interest in terms of community structure or the probes may cross-hybridise to similar sequences from related or unrelated organisms (Gentry et al, 2006).

### *1.2.2.5. Microarrays in environmental microbiology*

Microarrays have been employed to study microbial diversity in several research projects. The microbial community structure of natural habitats including water column (Castiglioni et al, 2004; Rudi et al, 2000), sediment (El Fantroussi et al, 2003) and soil as well as man-made environments such as anaerobic digester (Franke-Whittle et al, 2009) and compost (Hultman et al, 2008) have been characterised using phylogenetic microarrays. Microarrays provide a practical analysis method for complex microbial communities with relatively low cost and easiness (Andersen et al, 2010). However, it's been speculated that new next generation sequencing methods with lower costs and improved sequence length and quality may reduce the interest towards microarray technologies which may evolve into preparative tools (Brodie, 2011).

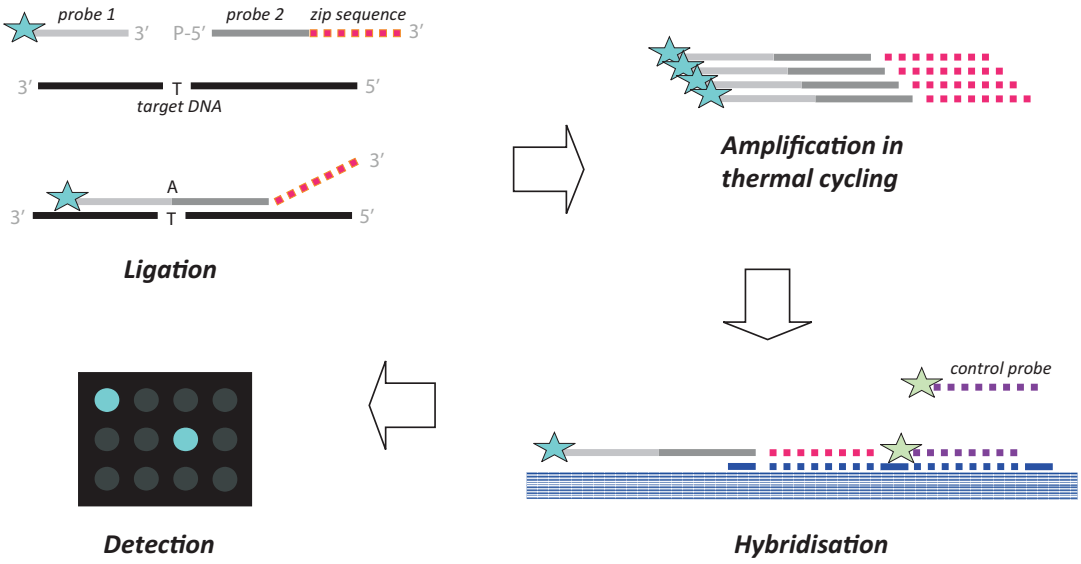


**Figure 4.** The principle of ligation detection reaction (LDR) microarray: a schematic picture presenting the ligation detection reaction and hybridisation on microarray. Figure modified from Hultman et al, 2008.

## 1.3. Microbial community data analysis

Bioinformatics is a multidisciplinary field of science that develops computational data processing methods for applications in biology (Lesk, 2002). The development and application of NGS sequencing and microarray technologies in microbial ecology has shifted the bottleneck from producing data to analysing vast amounts of complicated information. This shift has put pressure on the development of new data analysis tools, and the increasing need for computational power and methodology has become the new challenge (Teeling & Glöckner, 2012). The whole field of microbial ecology, especially when working with sequence data, has become very computer assisted and today's microbial ecologist must learn to use several data analysis tools, and even write their own scripts and programs if suitable applications are not available. Fortunately, there are a growing number of publicly available programs that can be utilised in data analysis, and the recently published algorithms are designed to handle large datasets.

### *1.3.1. Amplicon sequence data analysis*

The purpose of sequence data analysis is to extract all the relevant information hidden in the nucleotide sequence data and translate it into meaningful knowledge. To identify and eliminate the errors in the data originating from laboratory procedures and sequencing technologies is vitally important (Quince et al, 2011). The selection of methods, from sampling and DNA extraction to PCR conditions and sequencing method may all bias the resulting data. These biases and errors influence the downstream analyses and possibly results and conclusions if not tackled by proper experimental design and sequence data analysis (Schloss et al, 2011).

For the last few years several research groups around the world have developed computer programs specifically designed for amplicon sequence data analysis. A restricted list of the methods is presented in Table 2.

Mothur (Schloss et al, 2009), RDP Pyrosequencing Pipeline and other Ribosomal Database Project tools (Cole et al, 2009), as well as Qiime (Caporaso et al, 2010) represent the most applied complete amplicon sequence data analysis pipelines, at least in the sense of being most cited (www.scopus.com). The data analysis pipeline programs can typically be utilised from initial processing of the raw data to hypothesis testing and visualising the results. There are also a countless number of programs designed to compute a certain step in data analysis (Table 2). As a rule, the data analysis workflow, regardless of the programs applied, follows the same routine, possibly with small differences in the order of certain tasks (Figure 5).

Generally, the data analysis initiates with preprocessing. The first step is denoising the flowgram (or fasta formatted) data. Denoising refers to the correcting of the raw data and translating it to human readable DNA sequence. Denoising is accomplished by clustering the flowgrams using frequency based heuristics or approximate likelihood with empirically derived error distributions (Quince et al, 2011; Reeder & Knight, 2010). Another option is to use the fasta files in alignment based error correction (Bragg et al, 2012). AmpliconNoise (PyroNoise) (Quince et al, 2011) and DeNoiser (Reeder & Knight, 2010), as well as Mothur shhh.flows (Schloss et al, 2011), Patrick Schloss's translation of PyroNoise algorithm, use the flowgram data. AmpliconNoise is highly effective in noise removal but it requires a lot of memory and CPU time (Bragg et al, 2012) and therefore it is not operable for most researchers with large datasets. DeNoiser algorithm is much faster and the maximum memory requirement is only a small fraction of AmpliconNoise's but the performance is rather poor (Quince et al, 2011). Acacia (Bragg et al, 2012) uses fasta files and requires even less memory and CPU time than DeNoiser. According to

**Table 2.** Sequence data analysis methods

| method | function | references |
|---|---|---|
| CANGS DB | pipeline | Pandey et al, 2011 |
| CLOTU | pipeline | Kumar et al, 2011 |
| CloVR-ITS | pipeline | White et al, 2013 |
| Mothur | pipeline | Schloss et al, 2009 |
| Pangea | pipeline | Giongo et al, 2010 |
| PlutoF | pipeline | Abarenkov et al, 2010 |
| PyroTagger | pipeline | Kunin & Hugenholtz, 2010 |
| Qiime | pipeline | Caporaso et al, 2010 |
| RDP Pyrosequencing Pipeline | pipeline | Cole et al, 2009 |
| SCATA | pipeline | Brandström Durling et al, submitted |
| SEED | pipeline | Větrovský & Baldrian, 2013 |
| W.A.T.E.R.S | pipeline | Hartman et al, 2010 |
| ITSx | pre-processing | Bengtsson-Palme et al, 2013 |
| ITSxtractor | pre-processing | Nilsson et al, 2010 |
| PyroTrimmer | pre-processing | Oh et al, 2012 |
| SEQTRIM | pre-processing | Falgueras et al, 2010 |
| V-Xtractor | pre-processing | Hartmann et al, 2010 |
| Acacia | denoising | Bragg et al, 2012 |
| AmpliconNoise | denoising | Quince et al, 2011 |
| DADA | denoising | Rosen et al, 2012 |
| DeNoiser | denoising | Reeder & Knight, 2010 |
| Mothur shhh.flows | denoising | Schloss et al, 2011 |
| PyroNoise | denoising | Quince et al, 2009 |
| Single-linkage preclustering (SLP) | denoising | Huse et al, 2010 |
| ESPRIT | aligning | Sun et al, 2009 |
| Greengenes NAST aligner | aligning | DeSantis et al, 2006b |
| Infernal | aligning | Nawrocki et al, 2009 |
| MAFFT | aligning | Katoh & Standley, 2013 |
| Muscle | aligning | Edgar, 2004 |
| SILVA Incremental Aligner (SINA) | aligning | Quast et al, 2013 |
| BEBaC | clustering | Cheng et al, 2012 |
| CD-HIT | clustering | Li et al, 2001; Niu et al, 2010 |
| CLUSTOM | clustering | Hwang et al, 2013 |
| CrunchClust | clustering | Hartmann et al, 2012 |
| ESPRITTree | clustering | Cai & Sun, 2011 |
| Mothur average neighbor | clustering | Schloss et al, 2009 |
| Two-Stage Clustering (TSC) | clustering | Jiang et al, 2012 |

**Table 2 continuing**

| method | function | references |
|---|---|---|
| USEARCH algorithms: UBlast, USearch, UClust | database search and clustering | Edgar, 2010 |
| Bellerophon | chimera removal | Huber et al, 2004 |
| CCODE | chimera removal | Gonzalez et al, 2005 |
| Chimera_check | chimera removal | Maidak et al, 2001 |
| ChimeraSlayer | chimera removal | Haas et al, 2011 |
| Mallard | chimera removal | Ashelford et al, 2006 |
| Perseus | chimera removal | Quince et al, 2011 |
| Pintail | chimera removal | Ashelford et al, 2005 |
| Uchime | chimera removal | Edgar et al, 2011 |
| BLAST | taxonomy | Altschul et al, 1997 |
| GAST | taxonomy | Huse et al, 2008 |
| Megan | taxonomy | Huson et al, 2007 |
| pplacer | taxonomy | Matsen et al, 2010 |
| RDP Classifier | taxonomy | Wang et al, 2007 |
| SeqMatch | taxonomy | Wang et al, 2007 |
| SimRank | taxonomy | DeSantis et al, 2011 |
| VITCOMIC | taxonomy and visualisation | Mori et al, 2010 |
| Libshuff | statistical method | Singleton et al, 2001 |
| Metastats | statistical method | White et al, 2009 |
| STAMP | statistical method | Parks & Beiko, 2010 |
| Unifrac | statistical method | Lozupone & Knight, 2005 |
| Flowsim | data simulation | Balzer et al, 2010 |
| Grinder | data simulation | Angly et al, 2012 |
| MetaSim | data simulation | Richter et al, 2008 |

Bragg and colleagues (2012), the developers of the program, Acacia's error correction is comparable to AmpliconNoise.

Preprocessing of sequence data may also include extracting the targeted 16S rRNA and ITS sequences, because the raw data sometimes contain contaminants and sequences derived from other genes. Programs, such as V-Xtractor (Hartmann et al, 2010) and ITSxtractor (Nilsson et al, 2010), are developed for extracting 16S rRNA and ITS sequences, respectively, from the raw data. This function is important when working with fungal ITS sequences, since fungal ITS primers often amplify plant derived material as well.

After preprocessing and denoising, the sequences are in most cases aligned. Aligning methods can be divided into two categories: unsupervised alignment methods and alignment to a reference database (Schloss, 2010): Multiple sequence alignment algorithms such as muscle (Edgar, 2004) and ESPRIT (Sun et al, 2009) work without a reference and the idea is to align "all against all". With large next generation

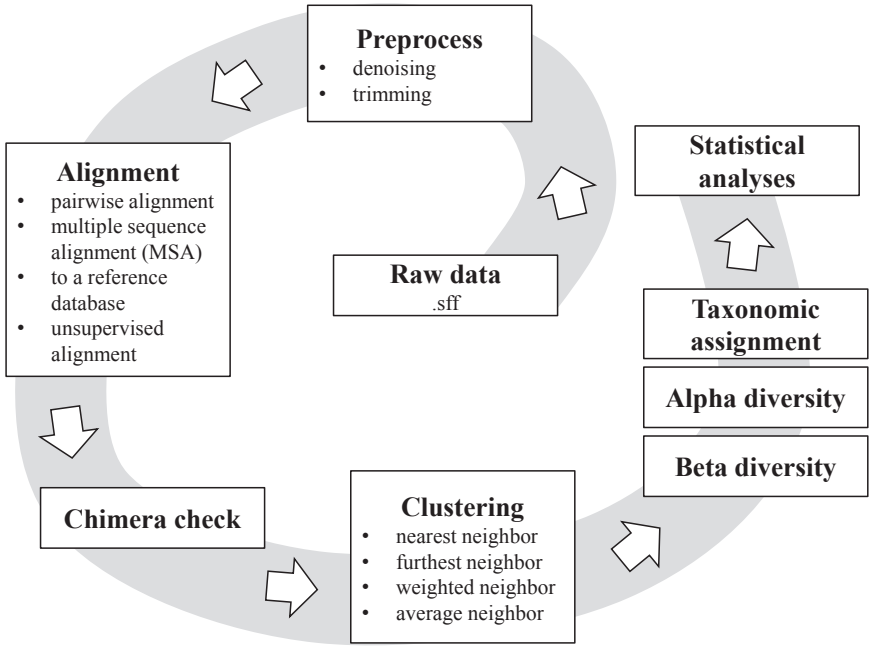


**Figure 5.** A schematic picture of a typical data analysis workflow.

sequencing datasets this approach is often unfeasible and reference based alignment algorithms have been developed (Schloss et al, 2009; Caporaso et al, 2010; Cole et al, 2009; DeSantis et al, 2006b). These methods employ a reference database of thousands of manually-curated good quality sequences and the same regions are aligned to the same positions every time allowing unchanging annotation of position-dependent features such as primer annealing locations and secondary structures (DeSantis et al, 2006b). Additionally, it is not necessary to align the sequences against each other, but every sequence against the database, diminishing the memory and CPU requirements (Schloss, 2010). The distances can be calculated using different criteria: ignoring gaps in alignment, treating a gap of any length as one mutation, or interpreting a gap of "n" nucleotides long as "n" mutations (Lesk, 2002). The choices made with alignment parameters and alignment quality considerably affects the distances between sequences and consequently the all downstream analyses (Schloss, 2010).

Yet another source of bias can, and should be identified and removed in the aligned sequence data. Chimeric sequences are hybrid products of multiple parent sequences that are formed during PCR amplification (Kopczynski et al, 1994; Liesack et al, 1991). The abundance of chimeric sequences in a dataset may vary greatly from only a few sequence reads to almost half of the entire dataset (Quince et al, 2009; Huber et al, 2004; Ashelford et al, 2006). Several factors may influence the chimera formation, including pairwise sequence identity between the target genes, relative abundance of PCR templates in the sample and the number of PCR cycles (Acinas et al, 2005; Lahr & Katz, 2009; Thompson et al, 2002; Wang & Wang, 1996, 1997). Consequently, the frequency of chimera formation can be influenced to some extent by checking the laboratory procedures, for example using low number of cycles in PCR amplification (Lahr & Katz, 2009), but there are always factors that are beyond control. There are two different approaches that can be applied: the sequence data can be checked by comparing

it against a curated reference database with good quality sequences. If the two ends of the query sequence form best alignments with unrelated reference sequences, the query sequence is potentially chimeric (e.g. Haas et al, 2011). Possible chimeras can also be checked by comparing against the abundant sequences in the same dataset or sample under scrutiny. The basic idea is that the chimeric sequences should be more infrequent than the parent sequences (e.g. Quince et al, 2011). There are various chimera checking programs available, such as Perseus (Quince et al, 2011) that uses sequence abundance information to identify the chimeras, Chimera Slayer (Haas et al, 2011) that searches against chimera-free database and Uchime (Edgar et al, 2011) that can be run both with a reference database and in *de novo* mode.

Clustering sequences into OTUs is the first step in assessing community diversity. Whether two sequences are clustered into same OTU depends both on their similarity and the other sequences in the same dataset, as well as the clustering algorithms used (Schloss & Westcott, 2011). There are four classic hierarchical clustering algorithms: the nearest (single-linkage), furthest (complete linkage), weighted and average (Unweighted Pair Group Method with Arithmetic Mean) neighbour algorithms (Legendre & Legendre, 1998). These classic algorithms were written before the era of massively parallel sequencing and colossal datasets, and consequently they are computationally intensive. Therefore, scientists have developed and employed heuristics to cluster the sequence reads into OTUs, for example ESPRITTree (Cai & Sun, 2011), CD-HIT (Niu et al, 2010), Uclust (Edgar, 2010), to allow analysis of large datasets without the access to efficient computer facilities. Unfortunately, these algorithms do not work as accurately as the best of the classic hierarchical clustering methods and the average neighbour algorithm is still recommended method based on empirical observations (Schloss & Westcott, 2011). The clustered OTUs can be subjected to various alpha and beta diversity measurements, such as diversity index and richness estimate assessments and similarity calculations between communities (Table 2.).

As the fundamental goal in microbial ecology is to characterise the structure and function of all microbial communities (Gentry et al, 2006), the identification of microbes present in studied environment is of great interest. Taxonomic assignment is performed by querying the sequence reads or a representative sequence of each OTU against a reference sequence database using a search algorithm, such as BLAST (Altschul et al, 1997), naive bayesian classifier (Wang et al, 2007) or USEARCH (Edgar, 2010). There are three major sequence data repositories, National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) at the European Molecular Biology Laboratory (EMBL) and DNA Databank of Japan (DDBJ), and data submitted to one of the repositories will be shared daily across all three, and data releases are made frequently (Benson et al, 2009). These repositories include all the submitted genome and gene sequences of all organisms sequenced. There are also databases concentrating only on 16S rRNA sequences. These include SILVA (Quast et al, 2013), a quality checked and aligned ribosomal RNA sequence database, Ribosomal Database Project database (Cole et al, 2009) with good quality bacterial and archaeal sequences, and Greengenes (DeSantis et al, 2006a) with a good quality and chimera checked 16S rRNA sequences.

When the organisms are identified and the diversity assessed in the studied environment, the other major interest in ecological research is to study the variation between membership and structure of multiple communities by quantifying the differences in both taxon composition and relative abundance, i.e. beta diversity (Whittaker, 1960). Differences in taxonomic assignment and OTU-based diversity can be examined using statistical analyses. Typically

the aim of the study is to discover if there is a statistically significant difference in the overall diversity or in the abundance distribution or diversity of a certain microbial group between treatment and control. Metastats (White et al, 2009) can be employed in search of differentially abundant features, i.e. taxonomic groups or OTUs, in treatment and control samples. Unifrac (Lozupone & Knight, 2005) and Libshuff (Singleton et al, 2001) methods describe whether two or more studied communities have the same community structure. The diversity between samples or treatments can also be visualised using ordination plots (Magurran, 2004).

Amplicon sequence data analysis, including all the steps described above, has developed substantially since the first publication by Sogin and colleagues in 2006 (Sogin et al, 2006). Lots have been learned about the possibilities of the method and the problems concerning data inaccuracy (Quince et al, 2011; Schloss et al, 2011). Research groups are developing even better algorithms to fix the biases and sequencing errors, but at the same time a new problem has emerged: using different methods/parameters for the various steps of the data analysis may lead to amazingly dissimilar error rates. The choices in methods will affect which sequences reads are accepted in analysis and which are discarded (Schloss et al, 2011), and the differences may show in the final results. Especially the number of observed OTUs and various diversity estimates are sensitive to the number of sequence reads (sampling effort) (Gihring et al, 2012), but also the abundance of different taxa in samples may be skewed. This causes problems particularly if the results of different studies are compared but the data analysis is not redone with one single data analysis protocol.

### *1.3.2. Microarray data analysis*

Microarray raw data is composed of scanned microarray images that need to be extracted and normalised prior to the actual analysis (Quackenbush, 2002). Data normalisation strives for elimination or standardisation of variation caused by all other factors except the studied variability, and still maintaining the real biological variation. These factors include variation between and within microarrays caused by technical issues, such as nucleic acid quantity and quality in samples, differences in labelling and probe concentrations as well as hybridisation efficiency and washing (Blalock, 2003; Quackenbush, 2002). The variance of the feature of interest may be of similar size as the variation caused by experimental design, and consequently without successful normalisation and analysis the result would not be seen (Blalock, 2003). There are number of data normalisation methods available for microarray data, including global intensity normalisation, intensity-dependent linear normalisation, variance stabilising normalisation and spiked control normalisation that can be applied singly or in conjunction (Quackenbush, 2002; Huber et al, 2002; Polanski & Kimmel, 2007). Unfortunately, these normalisation methods are not optimised for phylogenetic microarrays that usually employ mismatch probes as an indicator for differentiating between background hybridisation and real positive signals (DeSantis et al, 2005). Even though the microarray technology has been widely used in medical and biological research, there are still unknown or poorly understood mechanisms causing bias and variation within and between microarrays causing difficulties in data analysis and interpretation (Steger et al, 2011). A promising method for phylogenetic microarray data normalisation bases on the application of per-spot hybridisation control oligonucleotide probes (Ritari et al, 2009). After eliminating all removable bias from the data, the analysis can be run using for instance R statistical environment and marray package (Yang et al, 2007) from the Bioconductor project (Gentleman et al, 2004).

# 2. AIMS OF THE STUDY

We aimed to characterise diverse microbial communities in environmental samples using amplicon sequencing. The goal was to determine the microbial community structure and diversity in the northern Baltic Sea water column and anaerobic digestion reactor operating at different temperatures and organic loads, and to assess how the prevailing abiotic factors affect the studied communities. We also strived to develop and assess ligation detection reaction microarray method in identifying microbes in environmental samples. The objective was to design a fast, specific and sensitive microarray method that could be applied in monitoring changes in microbial community structure during anaerobic digestion process. Additionally, we assessed existing sequence data analysis methods. The main goal was to evaluate magnitude and direction of bias of widely applied amplicon sequence data denoising and clustering tools, and identify the algorithms that work most reliably.

# 3. MATERIALS AND METHODS

The materials and key methods are summarised in Table 3 and described in detail in indicated articles.

**Table 3.** Materials and key methods used in the study.

| materials | article |
|---|---|
| **biological samples** | |
| sea water bacteria samples | I |
| anaerobic reactor microbe samples | II |
| **data sets** | |
| Quince Titanium | III |
| modified atmosphere packaked poultry | III |
| **methods** | |
| **DNA extractions** | |
| FastDNA Spin Kit for Soil | I, II |
| **padlock probes** | |
| protocol for circular padlock probe ligation reaction | II |
| probe design | II |
| primers and protocol for amplification of padlock ligation reactions | II |
| **microarray** | |
| fabrication and testing | II |
| hybridisation protocols | II |
| **pyrosequencing** | |
| description of amplicon sequencing laboratory protocols | I, II |
| **data analysis** | |
| Acacia | III |
| BEBaC | III |
| CD-HIT | III |
| DeNoiser | III |
| ESPRITTree | III |
| Mothur | I, II, III |
| Qiime | III |
| UCLUST | III |

# 4. RESULTS AND DISCUSSION

## 4.1. Microbial diversity in environmental samples (I, II)

We studied diverse microbial communities in environmental samples using 454 amplicon sequencing method.

### *4.1.1. Bacterial diversity in the northern Baltic Sea (I)*

Bacterial communities in the northern Baltic Sea were characterised. The goal was to determine the community structure and assess the diversity of the bacterial communities in different locations and depths, as well as to study how the abiotic factors, such as oxygen concentration, pH, temperature, salinity and nutrient concentrations affect the bacterial communities. The samples were taken in 2008 from three HELCOM-COMBINE monitoring stations (LL7, LL12 and TPDEEP) from three depths in the water column and Kruunuvuorenselkä near Helsinki from two depths. Based on the special features of the Baltic Sea region and previous publications we hypothesised to find limited bacterial species diversity in our samples (Hällfors et al, 1981) compared to oceanic habitats and highest diversity in the surface water samples (Höfle & Brettar, 1995). We also expected to observe the measured environmental factors affecting the community structure.

Taxonomic affiliations were defined using Ribosomal Database Project Classifier (Wang et al, 2007). The results show that the northern Baltic Sea carries a diverse and patchy bacterial community, a mixture of salt- and freshwater bacterial taxa, with many potentially novel phylotypes: A total of 23 bacterial classes were identified, and as previously found in many locations in the world's oceans and seas (Gilbert et al, 2009, 2012; DeLong et al, 2006; Ghiglione & Murray, 2012; Krause et al, 2012; Venter et al, 2004), phylum *Proteobacteria* constituted the majority of the sequences in studied dataset. Classes *Gammaproteobacteria*, *Flavobacteria*, *Betaproteobacteria*, *Alphaproteobacteria* and *Actinobacteria* were abundant and detected in all samples. Especially *Gammaproteobacteria* and *Alphaproteobacteria* are very abundant and frequently encountered in marine habitats, particularly free-living marine bacterioplankton communities, but also distributed throughout sediments and water columns in freshwater environments (Gilbert et al, 2012; DeLong et al, 1993; Nold & Zwart, 1998; Fuhrman et al, 2006; Giovannoni et al, 1995; Hiorns et al, 1997). However, relatively commonly occurring class in our surface samples, *Betaproteobacteria*, is a typical freshwater taxon that is, according to literature, largely absent from open ocean environments (Nold & Zwart, 1998; Methé et al, 1998). The most abundant bacterial genera in the whole dataset were *Pseudomonas*, *Oleispira*, *Flavobacterium*, *Oceanospirillum*, *Rubritalea*, *Rhodobacter*, *Fluviicola* and *Sulfurimonas*. Most of these genera are typical marine bacteria found in various marine habitats in different locations worldwide (Alonso et al, 2007; Dang & Lovell, 2002; Grote et al, 2012; Holt et al, 1994; Kube et al, 2013; Moore et al, 2006; Satomi et al, 2002; Scheuermayer et al, 2006; Yakimov et al, 2003). A total of 5% of the sequence reads couldn't be reliably classified beyond the domain level indicating the presence of novel taxa. Figure 6 shows the bacterial groups present in samples at class level.

The results suggest the northern Baltic Sea carries wide variety of bacterial taxa – however, according to richness estimates the community is considerably less diverse than in the Atlantic and Pacific oceans as well as Western English Channel where the amount of identified bacterial genera is also substantially higher (Gilbert et al, 2009; Brown et al, 2009; Sogin et al, 2006). At least part of these

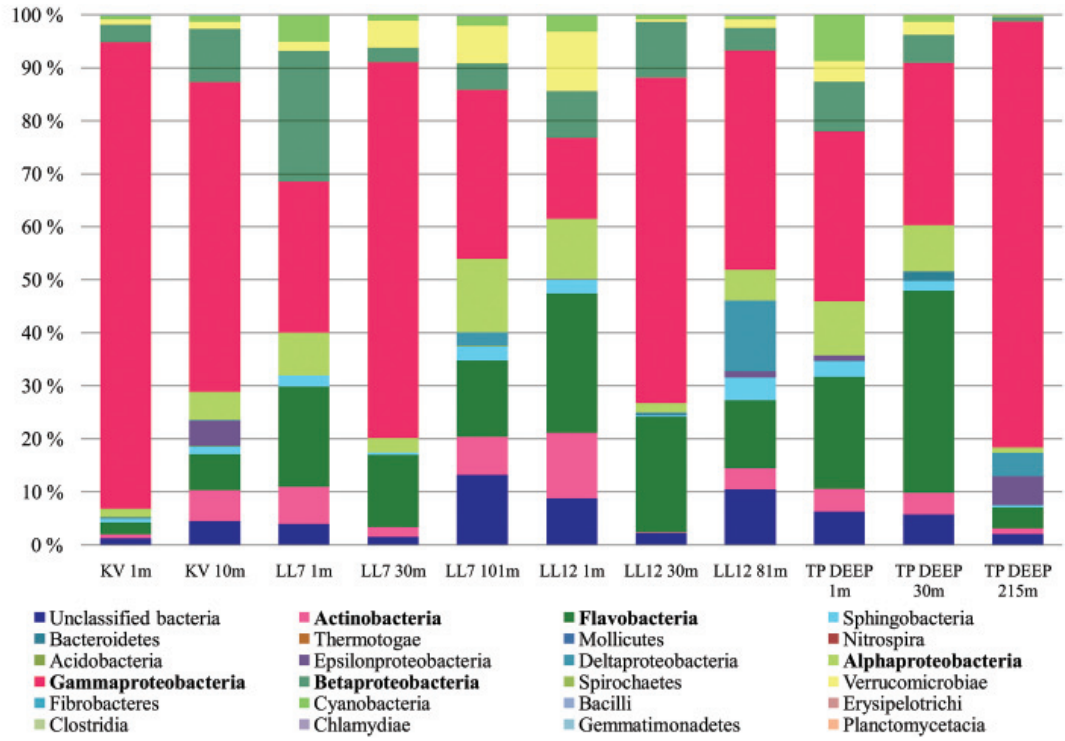


**Figure 6.** Baltic Sea bacterial communities in each sample determined by RDP Classifier (Wang et al, 2007). KV=Kruunuvuori; LL7, LL12 & TPDEEP=HELCOM-COMBINE monitoring stations. The most abundant bacterial classes are marked in bold.

differences can be explained by the application of short fragments (~100 bp) in all of these earlier publications and 454 GS-20 technology in Sogin's and Brown's publications which may inflate the number of operational taxonomic units significantly. Additionally, the number of sequence reads per sample was very high in all these three studies (Gilbert et al, 2009; Brown et al, 2009; Sogin et al, 2006) which is easily reflected in richness estimates that are sensitive to the sampling effort (Gihring et al, 2012). It has also been shown that datasets analysed using dissimilar methodology cannot be compared reliably (III).

Changes in physical and chemical environment, like shifts in nutrient concentrations, temperature, salinity, oxygen and depth may cause notable changes in bacterial community structure and diversity. The structural changes often affect the functional potential of the community which can be reflected to the whole ecosystem (Herlemann et al, 2011; Ojaveer et al, 2010; Lozupone & Knight, 2007; Wu et al, 2006). The bacterial groups present in studied samples were unevenly distributed, especially vertically. Redundancy analysis confirmed that the abundance of *Oleispira*, *Sulfurimonas*, *Oceanospirillum* and *Desulfobacterium* correlated with depth, lower pH and salinity, lack of oxygen, nitrite and nitrate as well as the presence of silicate, ammonium, and phosphate. Sulfate-reducing *Desulfobacterium* (Brysch et al, 1987) belongs to class *δ-Proteobacteria* which members are typically found in benthic microbial communities, but very rarely in oxygen rich surface water (Nold & Zwart, 1998), and analogous distribution was detected in our dataset. Herlemann and colleagues (2011) observed similar composition change at the oxygen-sulfide transition zone, below which

the sulfate-reducing bacteria dominated the community. The preference of bacterial genera *Pseudomonas*, *Flavobacterium*, *Rhodobacter*, *Fluviicola*, *Rubritalea* and *Polaribacter* for surface water and the related environmental conditions such as higher pH, oxygen concentration and nitrite was affirmed. Furthermore, the OTU based sequence data analysis showed a clear stratification of bacterial communities at pelagic sampling sites (I: Figure 2): the samples taken from the same depth bore more resemblance than samples taken at same location from different depths. This phenomenon has been observed before (e.g. Brown et al, 2009) and it is characteristic for stratified water bodies. In the Baltic Sea area the stratification is largely caused by temperature and salinity gradients, and due to the lack of mixing also other environmental factors that affect the bacterial communities such as oxygen and nutrient concentrations often differ dramatically between surface and near bottom waters. However, the oxygen depleted deep water communities were not less diverse than the surface water samples, refuting our hypothesis on the highest diversity in the surface water. The estuarine samples from Kruunuvuori clustered separately from pelagic samples in UPGMA analysis and formed their own branch. This can be explained by differing hydrochemical conditions between pelagic and estuarine samples: estuarine conditions include higher temperature, lower salinity, even oxygen concentrations due to vertical mixing and higher nutrient concentrations. Additionally, transport of minerals and riverine bacterial populations from Vantaa River and different sampling season may have impacted the bacterial community structure.

Salinity has been proposed to be the most determining factor for bacterial communities in aquatic environments (Wu et al, 2006). According to Herlemann and colleagues (2011) the most determining factors for bacterial community composition in the Baltic Sea brackish water are indeed salinity and depth, in this particular order. In our study, we were unable to differentiate between these two variables, most probably due to small number of samples and the fact that measured environmental factors correlated with depth.

Our bacterial community composition results did not fully resemble the previous studies conducted in the area. At least part of the differences between publications is caused by technical issues. Firstly, the choice of methods may have crucial impact on the results, for instance by selecting primers that amplify slightly different community. Furthermore, different practices in sample collection and storage, DNA extractions, PCR conditions, and sequencing may affect the data quality and shift the community structure (Lahr & Katz, 2009; Carrigg et al, 2007; Claesson et al, 2010; Engelbrektson et al, 2010; Hazen et al, 2013; Klindworth et al, 2012; Molbak et al, 2006; Pinto & Raskin, 2012; Plassart et al, 2012; Sergeant et al, 2012; Wallenius et al, 2010; Wu et al, 2010; Yuan et al, 2012). However, as the first publication in such depth about northern Baltic Sea bacterial communities, our study is a good beginning towards a better understanding of brackish water Bacteria and factors affecting the community structure and diversity.

### *4.1.2. Microbial diversity in anaerobic reactor (II)*

Microbial communities in anaerobic digestion reactor operating at different temperatures and organic loads were characterised. The aim was to study the diversity of Bacteria, Archaea and Fungi present in meso- and thermophilic anaerobic conditions, and use the resulting information in designing a new microarray for monitoring the community structure changes in response to physical and chemical factors during anaerobic digestion process. Based on literature (e.g. Hobson & Wheatley, 1993; Bitton, 2010; Okabe & Kamagata, 2010), we expected to find syntrophic Bacteria and methanogenic Archaea that are able to degrade organic material and produce methane in

anaerobic conditions. As the community structure and diversity of Fungi have not been characterised before, we were excited to see the fungal community that would be revealed. After all, organic waste contains fungal cells that are carried to the anaerobic reactor along with the waste and many of those Fungi are able to survive and grow in anaerobic conditions (Dumitru et al, 2004; Jennings, 1995; Kinsey et al, 2003). We anticipated that meso- and thermophilic temperatures would carry somewhat divergent communities and that increasing organic load would affect the microbial community structure and diversity (Levén et al, 2007).

Bacterial and archaeal communities in study reactor resembled the results in earlier publications (e.g. Levén et al, 2007; Riviere et al, 2009; Ros et al, 2013; Sasaki et al, 2011; Godon et al, 1997; Sekiguchi, 2006; von Wintzingerode et al, 1999; Wu et al, 2001). Bacterial phyla *Bacteroidetes*, *Firmicutes* and *Thermotogae* are frequently encountered in anaerobic digesters and were the most abundant bacterial phyla in our study reactor. Class *Flavobacteria* was present only in mesophilic reactor whereas *Thermotogae*, as the name implies, was detected exclusively in thermophilic reactor. The relatively less abundant phyla, *Actinobacteria* and *Proteobacteria* preferred mesophilic conditions, and *Spirochaetes*, *Synergistes* and *Verrucomicrobia* were present only there. Several candidate phyla consisting merely of environmental clones were also detected. All the identified Archaea belonged to phylum *Euryarchaeota* which is known to thrive in anaerobic digestion process. At class and genus level our findings were also supported by ealier publications as we identified numerous methanogens including *Methanosarcina*, *Methanobrevibacter*, *Methanosphaera*, *Methanospirillum*, *Methanosphaerula* and *Methanobacterium*. These Archaea are fundamentally important to the digestion process as they convert the byproducts of bacterial degradation into methane (Bitton, 2010).

When the cooperation of syntrophic Bacteria and metanogenic Archaea in anaerobic digestion is a well-known phenomenon and their functional roles in this multiphase process determined (Okabe & Kamagata, 2010) there has been little interest in the analysis of Fungi. Consequently, the diversity and community structure of Fungi in anaerobic digestion process is largely unknown. Our results suggest that there is a very diverse fungal community in anaerobic reactors. Two fungal phyla, *Ascomycota* and *Basidiomycota* were identified, *Ascomycota* by far the most abundant. *Saccharomycetes* and *Eurotiomycetes* were dominant at class level and several classes including but not limited to *Leotiomycetes*, *Pezizomycetes*, *Tremellomycetes*, *Agaricomycetes*, *Microbotryomycetes* were detected more infrequently. We identified as many as 33 fungal genera in the study reactors. The most abundant fungal genus was *Candida*, detected in both temperatures and organic loads. The second and third most commonly found Fungi were *Penicillium* and *Mucor*, respectively, *Penicillium* more prevalent in mesophilic and *Mucor* thermophilic temperature. These Fungi include groups capable of fermenting organic waste, and they are commonly found in soil, plant surfaces, and degrading organic matter (Madigan et al, 2003). Additionally, these Fungi are known to be able to degrade organic material in anaerobic conditions and their role in anaerobic digestion should be studied further (Dumitru et al, 2004; Jennings, 1995; Kinsey et al, 2003). Figure 2 (II) summarises the most abundant microbial groups present in the reactor samples.

## 4.2. Ligation detection reaction microarray with padlock probes (II)

The amplicon sequencing data were employed in designing padlock probe based ligation detection reaction microarray for monitoring

the microbial communities in anaerobic digestion process.

### 4.2.1. Specificity and sensitivity (II)

Microarrays have great potential in routine monitoring of diverse environmental samples (Andersen et al, 2010; Bodrossy & Sessitsch, 2004). However, the specificity of microarrays with oligomeric probes targeting the marker gene may pose problems when analysing diverse communities carrying closely related strains with (almost) identical nucleotide sequences in the marker gene, because oligonucleotides lack the high specificity in recognition of nucleic acids and hybridisation to mismatching targets occurs (Lomakin & Frank-Kamenetskii, 1998; Öhrmalm et al, 2010; Wu et al, 2005). Another challenge is the sensitivity, as the abundant phylotypes in a community are easily detected and the rare biosphere, the phylotypes with low relative abundance, falls below the detection limit (Wagner et al, 2007). Our ligation detection reaction microarray tackles these challenges by using a linear single stranded DNA probe with target recognition sequences situated at both termini of the probe, applying the similar idea as the discriminative probe and common probe in previous LDR microarray (e.g. Hultman et al, 2008; Busti et al, 2002; Castiglioni et al, 2004; Gerry et al, 1999; Favis et al, 2000), which in theory should prevent any unspecific hybridisation. The linear probe forms a circular molecule in ligation reaction if both target recognition sites are hybridised to their target with a perfect match (II: Figure 3). The circular probe is then PCR amplified with 5' phosphorylated forward primer and 5' Cy3 labelled reverse primer after which the phosphorylated strand is degraded. The single stranded Cy3 labelled products are hybridised on microarray and after scanning it is possible to detect which microbial groups were present. This padlock probe method allows the detection of unamplified microbial targets in environmental samples and it has been previously applied successfully in pathogen detection and gene variant analysis (van Doorn et al, 2007; Hardenbol et al, 2005; Jarvius et al, 2006; Li et al, 2009; Szemes et al, 2005).

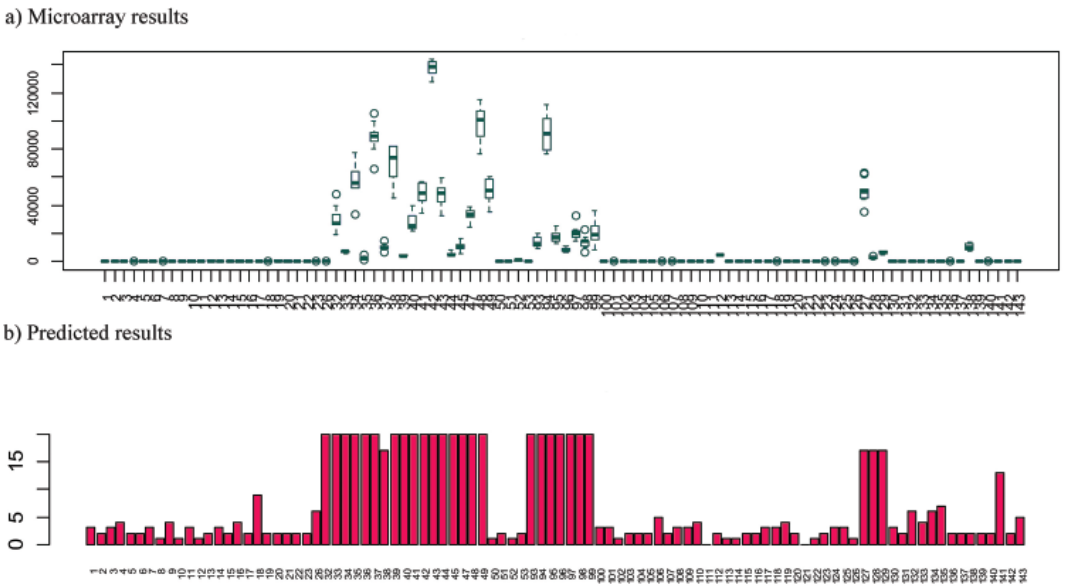


**Figure 7.** Proof of principle experiment with artificial templates. a) Microarray results and b) predicted results of the same artificial community. Figure modified from Ritari et al, 2012.

The specificity and sensitivity of the probes were tested with 80 nucleotides long synthetic dsDNA oligos as templates which were exact matches to the probes. A set of synthetic templates, 10 fmol of each, were pooled together to produce artificial microbial communities. We compared the microarray results of these test communities and the predicted results from the corresponding template pool to decipher the specificity of the probes (Figure 7). The results show that the templates present in the pool hybridised with their probes and the microarray signal intensities were clearly distinguishable from the nearly non-existent signals from the rest of the probes suggesting good specificity. Nevertheless, the signal intensities differed notably even though the templates were added in the pool at equal concentrations. About ten per cent of the probes did not hybridise to their target and six probes gave false positive signals presenting opposite results on microarray compared to *in silico* examination of the probes and templates.

The sensitivity was tested using the same probe-specific synthetic oligos as templates. We tested four template pools, each containing 24 synthetic oligos at different concentrations (Figure 8). With the highest concentration (1 fmol/µl/template) the signals were strongest, albeit the variation between probes was noted. The template pool with lowest concentration (0.001 fmol/µl/template) gave practically no signal on microarray. Almost all the probes gave stronger signal with higher template concentration and most probes were able to detect the exactly matching templates with 0.01 fmol/µl/template concentrations suggesting that this microarray could be applied in monitoring pursuing semiquantitative detection of microbial taxa over at least three orders of magnitude.

### *4.2.2. Application of the microarray to anaerobic reactor samples (II)*

We tested the padlock probe based ligation detection reaction microarray technology with real environmental samples from anaerobic reactor operating at different temperatures and organic loads. The probes were designed to target bacterial, archaeal and fungal phylotypes present in reactor, and the aim was to monitor the changes in the microbial communities and to identify the taxa present at a certain time point. The microbial communities grouped with measured factors rather similarly when using amplicon sequencing and microarray datasets in redundancy analysis, suggesting the microarray method could be used for

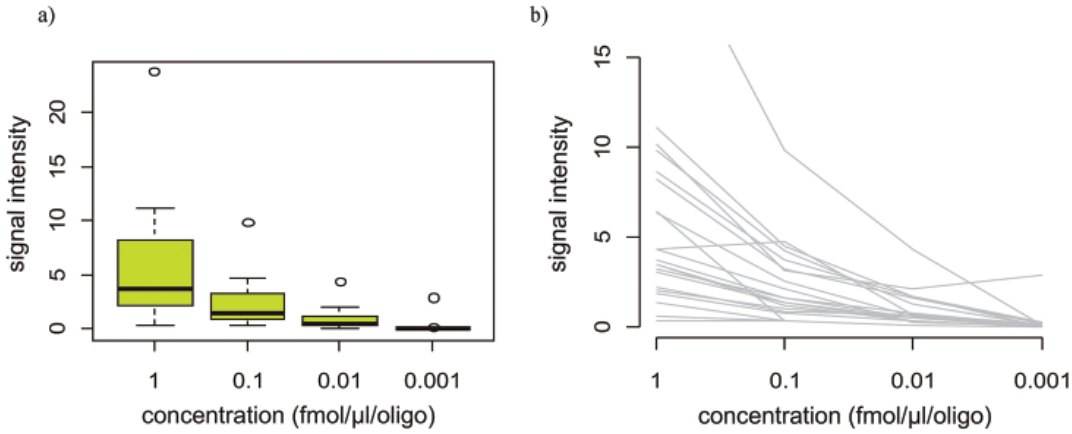


**Figure 8.** Microarray signals for different concentrations of synthetic template oligos.
(a) Boxplots showing the distribution of signals in each concentration and (b) line plots showing mean signals of individual probes. Figure modified from Ritari et al, 2012.

monitoring the changes during the digestion process (II: Figure 5).

As the microbial community structure in anaerobic reactor resembles the community structure in many other habitats with relatively long tail in rank abundance curve (Andersen et al, 2010), the detection limit is often too high for the less abundant phylotypes. The limited read length of the amplicon sequence data forced us to design part of the probe set using the best matches in the nucleotide database and their whole 16S rRNA sequence, leading to uncertainty whether these probes really target the microbial groups of interest. The results show that the probes having 100% match in the amplicon data gave the most consistent signals on microarray compared to the probes designed using partial matches or the database sequences. Figure 4 (II) shows the conformity of ligation detection reaction microarray with padlock probes, amplicon sequencing and qPCR methods, indicating the capability to detect the phylotypes of interest in real diverse samples using microarray assuming that the probes are designed carefully. The proof of principle data with synthetic templates as well as the experiment with anaerobic reactor samples suggest that the microarray can be applied in semiquantitative detection of microbes of interest, providing the detection limit (0.01 fmol/µl/template) is taken into account. The sensitivity of this new padlock probe based ligation detection reaction microarray is better compared to the traditional LDR microarray (Hultman et al, 2008; Busti et al, 2002; Castiglioni et al, 2004) and does not require PCR amplification of the target molecules, eliminating one source of bias in community structure characterisation. This concept is applicable to any environment as long as the probes are designed with care, and particularly suitable for monitoring purposes when samples are collected repeatedly and the analysis results are needed rapidly.

## 4.3. Data analysis bias (III)

The bias caused by different stages in laboratory work when studying diverse microbial communities using sequencing based methods are widely known and appreciated (e.g. Lahr & Katz, 2009; Carrigg et al, 2007; Sergeant et al, 2012; Yuan et al, 2012). However, after sequencing the data analysis stage can skew the results as well. We studied how the selection of analysis methods affects amplicon sequencing results. We hypothesised that selected methods would affect the number of observed operational taxonomic units and the detection of particularly rare taxa, but the magnitude of the impact was unclear.

We compared Denoiser (Reeder & Knight, 2010), Mothur shhh.flows (Schloss et al, 2011) and Acacia (Bragg et al, 2012) for denoising, and Uclust (Edgar, 2010), CD-HIT (Niu et al, 2010), Mothur average neighbour (Schloss et al, 2009), ESPRITTree (Cai & Sun, 2011) and BEBaC (Cheng et al, 2012) programs for clustering. The comparisons between data analysis pipelines were conducted using published data: an artificial soil microbial community with relatively even diversity profile, made up by amplifying selected clones and consequently fully known community structure and diversity, served as a benchmark (Quince et al, 2011). The same analyses were also applied to well-studied modified atmosphere packaked poultry microbe communities with a few very dominant community members and some extremely rare ones (Nieminen et al, 2012a) to test how consistently the algorithms work with distinct datasets with dissimilar diversity profiles.

We found drastic differences in the number of OTUs and inconsistencies in taxonomic composition. The results show that the choice of analysis methods may inflate the number of OTUs by an order of magnitude and difference between methods is dependent on data quality as well as bacterial diversity and abundance distribution in

studied samples (III: Figure 1 a). Comparisons using benchmark dataset revealed that most of the analysis pipelines overestimated the diversity in two ways: the number of OTUs was assessed too high with all methods except BEBaC (Cheng et al, 2012), and in many cases the number of identified taxa was higher than in the original clone library (III: Figure 1. b). Especially Acacia (Bragg et al, 2012) denoising algorithm with all tested clustering methods, quality score based trimming with Uclust, CD-HIT (Niu et al, 2010) and ESPRITTree (Cai & Sun, 2011), and untrimmed data overrated the number of taxa present in the community. The extra taxa originated primarily from PCR and sequencing based errors and there were considerable differences between data analysis methods in how well these errors were recognised and eliminated. The problem of the other extreme, i.e. filtering out the real diversity, was most prominent among methods producing OTU numbers smaller than the median.

The differences in relative abundance of identified taxa were studied using STAMP bioinformatics software (Parks & Beiko, 2010). Distance matrices (III: Figure 2) depict the average of phylum, class, order, family and genus level differences in studied datasets. The amount and focus of differences between methods were dependent on analysed dataset. Quince Titanium and natural poultry product microbe community (III: Figure 2 a and b) show very significant and great difference between Acacia (Bragg et al, 2012) and Mothur shhh.flows denoising (Schloss et al, 2011) with all tested clustering methods but the phenomenon was not present with marinated poultry product microbe community (III: Figure 2 c) representing the least diverse of studied datasets. Additionally, in poultry product microbe communities, both natural and marinated (III: Figure 2 b, c), BEBaC clustering (Cheng et al, 2012) method results differed most from other analysis pipelines. According to these observations, the analysis methods substantially affect the results: clustering method seems to have a great impact on the number of OTUs and denoising algorithm influences more on taxonomic affiliations.

These results clearly show the importance of the choice of data analysis methods, and demonstrate how biased comparisons can be made if two datasets are analysed using different algorithms. Moreover, the significance of functioning denoising and clustering methods is evident, because amplification and sequencing produce a considerable amount of errors and extra taxa that have to be filtered out from the data before further analysis. Additionally, considering all the bias causing stages in the workflow before data analysis, starting from experimental design and sampling to laboratory protocols sensitive to bias, such as nucleic acid extraction and PCR, and recognising that each one of these steps may skew or change the resulting community structure potentially by multifold (Hazen et al, 2013; Hong et al, 2009; Hurt et al, 2001; Leff et al, 1995; Sipos et al, 2010), it is challenging to be certain that the acquired results really represent the studied natural microbial community and the conclusions are correct, especially if there are no other data supporting the findings. These problems highlight the importance of careful research planning and taking actions towards minimising the bias at every step.

# 5. CONCLUSIONS

New molecular techniques have revolutionised the field of microbial ecology. The development of highly parallel sequencing and microarray technologies has enabled the study of diverse microbial communities and linking them to biotic and abiotic factors of interest by extensive experimental design.

In this thesis, the microbial community structure and diversity in the northern Baltic Sea water column and anaerobic digestion reactor were characterised using amplicon sequencing method with 454 sequencing technology. The aim was to assess how the prevailing abiotic factors affect the studied communities. We also developed and assessed a new padlock probe based ligation detection reaction microarray method for monitoring the changes in microbial community structure during anaerobic digestion process. Additionally, the performance of selected sequence data analysis methods was evaluated.

The results described in this thesis show the potential of amplicon sequencing method in characterising complex environmental samples in detailed fashion. The bacterial communities in the northern Baltic Sea water column were strongly stratified and inhabited by a vast selection of different bacterial groups. Aerobic Bacteria, such as *Pseudomonas* and *Flavobacterium* dominated in the surface layer, and *Oleispira* and sulfate-reducing bacteria in the anoxic deep waters. However, the diversity was assessed one order of magnitude less diverse compared to oceanic habitats.

The anaerobic digestion reactor communities were dominated by Bacteria belonging to phyla *Bacteroidetes*, *Firmicutes* and *Thermotogae* and metanogenic Archaea. These groups are all typical degraders in anaerobic digestion. The process also supported a diverse fungal community of phyla *Ascomycota* and *Basidiomycota*, including several taxa capable of degrading organic material in anaerobic conditions. The most abundant genera were *Candida*, *Penicillium* and *Mucor*.

The results also suggest that 454 amplicon sequence data can be applied in designing probes for microarrays. The new padlock probe based ligation detection reaction (LDR) microarray proved to be fast, specific and relatively sensitive without the need of amplifying the target molecule. The detection limit was 0.01 fmol/µl/template. Amplicon sequencing and LDR microarray method produced concordant community structure results and both showed the capability of semiquantitative identification of microbial community members.

As much promise as amplicon sequencing method holds in characterising diverse microbial communities, there are factors causing bias and errors. Amplicon sequencing raw data contain large amounts of false diversity that has to be filtered out during the course of data analysis and different analysis methods perform the task with varying success. We found prominent differences in observed operational taxonomic units and relative abundance of identified taxa. There were also notable differences between algorithms in the ability to filter out the spurious taxa produced by amplification and sequencing, but still retain all the real diversity.

As the progress of highly parallel sequencing technologies and microarray methods continues, and producing large amounts of data becomes even more available throughout the research community, the future challenges will include the development and application of functioning analysis tools which would reliably discriminate between the real and false diversity in studied samples. Additionally, less bias producing practices in laboratory will be of great importance since the most of the bias is still produced in different stages of laboratory protocols.

# 6. ACKNOWLEDGEMENTS

This thesis work was carried out in DNA sequencing and genomics laboratory at Institute of Biotechnology, University of Helsinki and was financially supported by Maj and Tor Nessling Foundation and Tekes (the Finnish Funding Agency for Technology and Innovation). The present and former heads of the Institute of Biotechnology Professors Tomi Mäkelä and Mart Saarma are thanked.

I wish to express my gratitude to everyone involved in this process. I thank my supervisors Doc. Petri Auvinen, the laboratory director in DNA sequencing and genomics laboratory, and Professor Martin Romantschuk at the Department of Environmental Sciences. I am grateful to Petri for accepting me as a part of his research group and providing me excellent working facilities and support. I am thankful to Martin for introducing me to the world of science. During my master's thesis project under Martin's supervision I realised how fascinating science can be. I am deeply grateful to both of you for your guidance. Furthermore, I would like to thank my thesis advisory committee members Professor Kaarina Sivonen and Dr Jarno Tuimala. I am also grateful to Professor Olivia Mason and Dr Alban Ramette for carefully reviewing my thesis and for valuable and constructive comments. Dr Christopher Wheat is thanked for assisting me with the language and for fearlessly giving me the most honest feedback: "Your introduction is incredibly dry". ☺

My co-authors Jenni Hultman, Lars Paulin, Harri Kankaanpää, Jarmo Ritari, Jukka Kurola, Maritta Kymäläinen, Martin Romantschuk, Johanna Björkroth and Petri Auvinen are warmly thanked for professional cooperation and contribution to the study. I thank all the lab members, current and former, for creating an inspiring working environment. Specially, I would like to thank Jenni, Pia and Miia, for sharing the same and exceptionally high-quality sense of humour. Dr Jenni is specifically acknowledged for her patient and long term contribution in introducing me to sequencing and sequence data analysis, and writing supportive statements. I thank you for being an excellent conference travelling supervisor, training partner and dear friend. I'm grateful to Pia and PKS Bioinformatics for the persistent help with unix and bioinformatics. I'm indebted to you for writing me scripts and improving them with relevant toy supplements. Dr Miia is thanked for sharing all kinds of adventures with me and being a dear friend. In addition, Dr C is acknowledged for the highly educational hands-on practicals on 454, and sometimes working overtime in lab and office.

My colleagues Teija, Juhana, Olli-Pekka, Velma and Tuuli are thanked for efficiently disconnecting me from my thesis over lunch in recent months. Your unscientific input was invaluable. Velma is also thanked for agreeing to be my first student, Tuuli for modelling in lab photos that I used in my presentations and Teija for continuously brainstorming about the theme at my karonkka dinner. Lasse is thanked for providing amusement by persistently attempting and failing to beat me at running. Additionally, Eeva-Marja, Fitsum, Hanna, Harri, Kirsi, Kui, Margarita, Matias, Olli, Panu, Pedro, Päivi, and Ursula (the list is not by ranking, but in alphabetical order) are acknowledged for the everyday support.

I wish to acknowledge Professor Timo Kairesalo, Dr Olli-Pekka Penttinen and Taru Nordman for all the help with administrative issues during my studies and finalising this thesis. I also would like to thank my graduate school EnSTe (Finnish Doctoral Programme in Environmental

Science and Technology) for the annual meetings which provided me both useful information and peer support from other students. I am also grateful for travelling grants that allowed me to travel to interesting conferences and workshops every year.

I also would like to thank my friends and family for all the support. My dear sister Laura and friends Tea, Pia and Heli are thanked for high-quality discussions on *everything*, being so exceptionally lovely and always there for me. Sisäkumi is acknowledged for organising refreshing holidays both domestic and abroad. Finally, I would like to thank my family for the support throughout my studies.

Helsinki, November 2013

Kaisa

# 7. REFERENCES

Abarenkov K, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, *et al.* (2010). PlutoF—a Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences. Evolutionary Bioinformatics 6:189-196.

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol 71:8966-8969.

Acinas SG, Rodriguez-Valera F, Pedrós-Alió C (1997). Spatial and temporal variation in marine bacterioplankton diversity as shown by RFLP fingerprinting of PCR amplified 16S rDNA. FEMS Microbiol Ecol 24:27-40.

Alonso C, Warnecke F, Amann R, Pernthaler J (2007). High local and global diversity of Flavobacteria in marine plankton. Environ Microbiol 9:1253-1266.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. - Nucleic Acids Res.1997 Sep 1;25(17):3389-402.

Andersen GL, He Z, DeSantis TZ, Brodie EL, Zhou J (2010). The use of Microarrays in Microbial Ecology. In: Environmental Molecular Microbiology. Caister Academic Press, pp 87-109.

Andersson AF, Riemann L, Bertilsson S (2010). Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. ISME J 4:171-181.

Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW (2012). Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res 40:e94.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2006). New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Appl Environ Microbiol 72:5734-5741.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol 71:7724-7736.

Bagge E, Sahlström L, Albihn A (2005). The effect of hygienic treatment on the microbial flora of biowaste at biogas plants. Water Res 39:4879-4886.

Balzer S, Malde K, Lanzen A, Sharma A, Jonassen I (2010). Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim. Bioinformatics 26:i420-5.

Baner J, Isaksson A, Waldenstrom E, Jarvius J, Landegren U, Nilsson M (2003). Parallel gene analysis with allele-specific padlock probes and tag microarrays. Nucleic Acids Res 31:e103.

Barns SM, Delwiche CF, Palmer JD, Pace NR (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci U S A 93:9188-9193.

Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, *et al.* (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 :n/a-n/a.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009). GenBank. Nucleic Acids Res 37:D26-31.

Bent SJ, Pierson JD, Forney LJ, Danovaro R, Luna GM, Dell'anno A, *et al.* (2007). Measuring species richness based on microbial community fingerprints: the emperor has no clothes. Appl Environ Microbiol 73:2399-401; author reply 2399-401.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.

Bentley DR (2006). Whole-genome re-sequencing. Curr Opin Genet Dev 16:545-552.

Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH (2008). Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc Natl Acad Sci U S A 105:10583-10588.

Bitton G (2010). Anaerobic Digestion of Wastewater and Biosolids. In: Wastewater Microbiology. John Wiley & Sons, Inc., pp 409-435.

Blalock EM (2003). A Beginner's Guide to Microarrays. Kluwer Academic Publishers: Boston, MA.

Bodrossy L, Sessitsch A (2004). Oligonucleotide microarrays in microbial diagnostics. Curr Opin Microbiol 7:245-254.

Bouvier TC, del Giorgio PA (2002). Compositional changes in free-living bacterial communities along a salinity gradient in two temperate estuaries. Limnol Oceanogr 47(2):453-470.

Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW (2012). Fast, accurate error-correction of amplicon pyrosequences using Acacia. Nat Meth 9:425-426.

Brandström Durling M, Clemmensen KE, Stenlid J, Lindahl B. SCATA - An efficient bioinformatic pipeline for species identification and quantification after high-throughput sequencing of tagged amplicons. (*submitted*)

Brenner DJ, Fanning GR, Rake AV, Johnson KE (1969). Batch procedure for thermal elution of DNA from hydroxyapatite. Anal Biochem 28:447-459.

Brodie EL, Desantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL, *et al.* (2006). Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl Environ Microbiol 72:6288-6298.

Brodie EL (2011). Phylogenetic Microarrays (PhyloChips) for Analysis of Complex Microbial Communities. In: Handbook of Molecular Microbial Ecology I. John Wiley & Sons, Inc., pp 521-532.

Brown MV, Philip GK, Bunge JA, Smith MC, Bissett A, Lauro FM, *et al.* (2009). Microbial community structure in the North Pacific ocean. ISME J 3:1374-1386.

Brysch K, Schneider C, Fuchs G, Widdel F (1987). Lithoautotrophic growth of sulfate-reducing bacteria, and description of Desulfobacterium autotrophicum gen. nov., sp. nov. Arch Microbiol 148:264-274.

Burnham KP, Overton WS (1979). Robust Estimation of Population Size When Capture Probabilities Vary Among Animals. Ecology 60:927-936.

Busti E, Bordoni R, Castiglioni B, Monciardini P, Sosio M, Donadio S, *et al.* (2002). Bacterial discrimination by means of a universal array approach mediated by LDR (ligase detection reaction). BMC Microbiol 2:27.

Cai Y, Sun Y (2011). ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Res 39:e95.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. Nat Meth 7:335-336.

Carlile MJ, Watkinson SC, Gooday GW (2001). The Fungi. Academic Press: London.

Carrigg C, Rice O, Kavanagh S, Collins G, O'Flaherty V (2007). DNA extraction method affects microbial community profiles from soils and sediment. Appl Microbiol Biotechnol 77:955-964.

Castiglioni B, Rizzi E, Frosini A, Sivonen K, Rajaniemi P, Rantala A, *et al.* (2004). Development of a universal microarray based on the ligation detection reaction and 16S rrna gene polymorphism to target diversity of cyanobacteria. Appl Environ Microbiol 70:7161-7172.

Chao A, Lee S (1992). Estimating the number of classes via sample coverage. J Am Stat Assoc 87:210-217.

Chao A (1984). Nonparametric estimation of the number of classes in a population. Scand J Statist 11:265-270.

Cheng L, Walker AW, Corander J (2012). Bayesian estimation of bacterial community composition from 454 sequencing data. - Nucleic Acids Res.2012 Jul;40(12):5240-9.Epub 2012 Mar 9.

Chorus I, Falconer IR, Salas HJ, Bartram J (2000). Health risks caused by freshwater cyanobacteria in recreational waters. J Toxicol Environ Health B Crit Rev 3:323-347.

Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, *et al.* (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res 38:e200.

Clarke J, Wu H, Jayasinghe L, Patel A, Reid S, Bayley H (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nano 4:265-270.

Cody ML, Diamond JM (1975). Ecology and Evolution of Communities. Belknap Press of Harvard University Press: Cambridge (Mass.) .

Cole JR, Konstantinidis KT, Farris RJ, Tiedje JM (2010). Microbial Diversity and Phylogeny: Extending from rRNAs to Genomes. In: Environmental Molecular Microbiology. Caister Academic Press, pp 1-19.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141-D145.

Comeau AM, Harding T, Galand PE, Vincent WF, Lovejoy C (2012). Vertical distribution of microbial communities in a perennially stratified Arctic lake with saline, anoxic bottom waters. Sci Rep 2:604.

Dang H, Lovell CR (2002). Numerical dominance and phylotype diversity of marine Rhodobacter species during early colonization of submerged surfaces in coastal marine waters as determined by 16S ribosomal DNA sequence analysis and fluorescence in situ hybridization. Appl Environ Microbiol 68:496-504.

Daniel R (2011). Soil-Based Metagenomics. In: Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats. Wiley-Blackwell, pp 83-92.

DeLeon-Rodriguez N, Lathem TL, Rodriguez-R LM, Barazesh JM, Anderson BE, Beyersdorf AJ, *et al.* (2013). Microbiome of the upper troposphere: species composition and prevalence, effects of tropical storms, and atmospheric implications. Proc Natl Acad Sci U S A 110:2575-2580.

Delmont TO, Prestat E, Keegan KP, Faubladier M, Robe P, Clark IM, *et al.* (2012). Structure, fluctuation and magnitude of a natural grassland soil metagenome. ISME J 6:1677-1687.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N, *et al.* (2006). Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science 311:496-503.

DeLong EF, Franks DG, Alldredge AL (1993). Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. Limnology Oceanography 38:924-934.

Deng Y, He Z, Xu M, Qin Y, Van Nostrand JD, Wu L, *et al.* (2012). Elevated carbon dioxide alters the structure of soil microbial communities. Appl Environ Microbiol 78:2991-2995.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* (2006a). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72:5069-5072.

DeSantis TZ, Stone CE, Murray SR, Moberg JP, Andersen GL (2005). Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. FEMS Microbiol Lett 245:271-278.

DeSantis TZ,Jr, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, *et al.* (2006b). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res 34:W394-9.

DeSantis T, Keller K, Karaoz U, Alekseyenko A, Singh N, Brodie E, *et al.* (2011). Simrank: Rapid and sensitive general-purpose k-mer search tool. BMC Ecology 11:11.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, *et al.* (2008). Functional metagenomic profiling of nine biomes. Nature 452:629-632.

van Doorn R, Szemes M, Bonants P, Kowalchuk GA, Salles JF, Ortenberg E, *et al.* (2007). Quantitative multiplex detection of plant pathogens using a novel ligation probe-based system coupled with universal, high-throughput real-time PCR on OpenArrays. BMC Genomics 8:276.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A 100:8817-8822.

Dumitru R, Hornby JM, Nickerson KW (2004). Defined anaerobic growth medium for studying Candida albicans basic biology and resistance to eight antifungal drugs. Antimicrob Agents Chemother 48:2350-2354.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and throughput. Nucleic Acids Res 32:1972-1997.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194-2200.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-2461.

Edlund A, Jansson JK (2008). Use of bromodeoxyuridine immunocapture to identify psychrotolerant phenanthrene-degrading bacteria in phenanthrene-enriched polluted Baltic Sea sediments. FEMS Microbiol Ecol 65:513-525.

Edlund A, Hårdeman F, Jansson JK, Sjöling S (2008). Active bacterial community structure along vertical redox gradients in Baltic Sea sediment. Environ Microbiol 10:2051-2063.

Edlund A, Soule T, Sjöling S, Jansson JK (2006). Microbial community structure in polluted Baltic Sea sediments. Environ Microbiol 8:223-232.

Eerola P (1993). The Baltic Sea - our common environment: background paper of the state of the Baltic Sea.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, *et al.* (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323:133-138.

El Fantroussi S, Urakawa H, Bernhard AE, Kelly JJ, Noble PA, Smidt H, *et al.* (2003). Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. Appl Environ Microbiol 69:2377-2382.

Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H, *et al.* (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J 4:642-647.

Falgueras J, Lara AJ, Fernandez-Pozo N, Canton FR, Perez-Trabado G, Claros MG (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics 11:38-2105-11-38.

Favis R, Day JP, Gerry NP, Phelan C, Narod S, Barany F (2000). Universal DNA array detection of small insertions and deletions in BRCA1 and BRCA2. Nat Biotechnol 18:561-564.

Fouts DE, Szpakowski S, Purushe J, Torralba M, Waterman RC, MacNeil MD, *et al.* (2012). Next generation sequencing to define prokaryotic and fungal diversity in the bovine rumen. PLoS One 7:e48289.

Franke-Whittle IH, Goberna M, Pfister V, Insam H (2009). Design and development of the ANAEROCHIP microarray for investigation of methanogenic communities. J Microbiol Methods 79:279-288.

Franke-Whittle IH, Klammer SH, Insam H (2005). Design and application of an oligonucleotide microarray for the investigation of compost microbial communities. J Microbiol Methods 62:37-56.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, *et al.* (2008). Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A 105:3805-3810.

Fuhrman JA, Hagström Å (2008). Bacterial and Archaeal Community Structure and its Patterns. In: Microbial Ecology of the Oceans. Wiley-Blackwell, pp 45-90.

Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. Proceedings of the National Academy of Sciences 103:13104-13109.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80.

Gentry TJ, Wickham GS, Schadt CW, He Z, Zhou J (2006). Microarray applications in microbial ecology research. Microb Ecol 52:159-175.

Gerry NP, Witowski NE, Day J, Hammer RP, Barany G, Barany F (1999). Universal DNA microarray method for multiplex detection of low abundance point mutations. J Mol Biol 292:251-262.

Ghiglione JF, Murray AE (2012). Pronounced summer to winter differences and higher wintertime richness in coastal Antarctic marine bacterioplankton. Environ Microbiol 14:617-629.

Gihring TM, Green SJ, Schadt CW (2012). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. Environ Microbiol 14:285-290.

Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, *et al.* (2009). The seasonal structure of microbial communities in the Western English Channel. Environ Microbiol 11:3132-3139.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, *et al.* (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One 3:e3042.

Gilbert JA, Steele JA, Caporaso JG, Steinbruck L, Reeder J, Temperton B, *et al.* (2012). Defining seasonal marine microbial community dynamics. ISME J 6:298-308.

Giongo A, Crabb DB, Davis-Richardson AG, Chauliac D, Mobberley JM, Gano KA, *et al.* (2010). PANGEA: pipeline for analysis of next generation amplicons. ISME J 4:852-861.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990). Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60-63.

Giovannoni S, Mullins T, Field K (1995). Microbial Diversity in Oceanic Systems: rRNA Approaches to the Study of Unculturable Microbes. In: Springer Berlin Heidelberg, pp 217-248.

Godon JJ, Zumstein E, Dabert P, Habouzit F, Moletta R (1997). Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. Appl Environ Microbiol 63:2802-2813.

Gonzalez JM, Zimmermann J, Saiz-Jimenez C (2005). Evaluating putative chimeric sequences from PCR-amplified products. Bioinformatics 21:333-337.

Grote J, Schott T, Bruckner CG, Glöckner FO, Jost G, Teeling H, *et al.* (2012). Genome and physiology of a model Epsilonproteobacterium responsible for sulfide detoxification in marine oxygen depletion zones. Proceedings of the National Academy of Sciences 109:506-510.

Guarro J, GeneJ, Stchigel AM (1999). Developments in fungal taxonomy. Clin Microbiol Rev 12:454-500.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, *et al.* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21:494-504.

Hagström Å, Pinhassi J, Zweifel UL (2000). Biogeographical diversity among marine bacterioplankton. Aquat Microb Ecol 21:231-244.

Halinen K, Jokela J, Fewer DP, Wahlsten M, Sivonen K (2007). Direct Evidence for Production of Microcystins by *Anabaena* Strains from the Baltic Sea. Appl Environ Microbiol 73:6543-6550.

Hällfors G, Niemi Å, Ackefors H, Lassig J, Leppäkoski E (1981). Biological Oceanography. In: The Baltic Sea. Elsevier Oceanography Series: Amsterdam, The Netherlands, pp 219-274.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5:R245-9.

Hanski I, von Hertzen L, Fyhrquist N, Koskinen K, Torppa K, Laatikainen T, *et al.* (2012). Environmental biodiversity, human microbiota, and allergy are interrelated. Proceedings of the National Academy of Sciences .

Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, Brundage T, *et al.* (2005). Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. Genome Res 15:269-275.

Hartman AL, Riddle S, McPhillips T, Ludascher B, Eisen JA (2010). Introducing W.A.T.E.R.S.: a workflow for the alignment, taxonomy, and ecology of ribosomal sequences. BMC Bioinformatics 11:317-2105-11-317.

Hartmann M, Howes CG, VanInsberghe D, Yu H, Bachar D, Christen R, *et al.* (2012). Significant and persistent impact of timber harvesting on soil microbial communities in Northern coniferous forests. ISME J 6:2199-2218.

Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH (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. J Microbiol Methods 83:250-253.

Hazen TC, Rocha AM, Techtmann SM (2013). Advances in monitoring environmental microbes. Curr Opin Biotechnol 24:526-533.

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, *et al.* (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J 1:67-77.

Heltshe JF, Forrester NE (1983). Estimating Species Richness Using the Jackknife Procedure. Biometrics 39:1-11.

Herlemann DP, Labrenz M, Jurgens K, Bertilsson S, Waniek JJ, Andersson AF (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. ISME J 5:1571-1579.

Hewson I, Steele JA, Capone DG, Fuhrman JA (2006a). Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. Mar Ecol Prog Ser 311:67-77.

Hewson I, Steele JA, Capone DG, Fuhrman JA (2006b). Remarkable Heterogeneity in Meso- and Bathypelagic Bacterioplankton Assemblage Composition. Limnol Oceanogr 51:1274-1283.

Hiorns WD, Methe BA, Nierzwicki-Bauer SA, Zehr JP (1997). Bacterial diversity in Adirondack mountain lakes as revealed by 16S rRNA gene sequences. Appl Environ Microbiol 63:2957-2960.

Hjelt OEA (1907). Carl Von Linne´s Betydelse Såsom Naturforskare Och Läkare. Almqvist & Wiksell: Uppsala.

Hobson P, Wheatley A (1993). Anaerobic Digestion: Modern Theory and Practice. Elsevier Applied Science.

Höfle MG, Brettar I (1995). Taxonomic diversity and metabolic activity of microbial communities in the water column of the central Baltic Sea. Limnol Oceanogr 40:868-874.

Holmfeldt K, Dziallas C, Titelman J, Pohlmann K, Grossart HP, Riemann L (2009). Diversity and abundance of freshwater Actinobacteria along environmental gradients in the brackish northern Baltic Sea. Environ Microbiol 11:2042-2054.

Holt JG, Krieg NR, Sneath PHA, Staley JT, Williams ST (1994). Bergey's Manual of Determinative Bacteriology. Williams & Wilkins: Baltimore, Maryland, USA.

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J 3:1365-1373.

Hortal J, Borges PA, Gaspar C (2006). Evaluating the performance of species richness estimators: sensitivity to sample grain size. J Anim Ecol 75:274-287.

Huang S, Yang F, Zeng X, Chen J, Li R, Wen T, *et al.* (2011). Preliminary characterization of the oral microbiota of Chinese adults with and without gingivitis. BMC Oral Health 11:33-6831-11-33.

Huber JA, Welch DBM, Morrison HG, Huse SM, Neal PR, Butterfield DA, *et al.* (2007). Microbial Population Structures in the Deep Marine Biosphere. Science 318:97-100.

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417:63-67.

Huber T, Faulkner G, Hugenholtz P (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20:2317-2319.

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1:S96-104.

Hugenholtz P, Tyson GW (2008). Microbiology: Metagenomics. Nature 455:481-483.

Hulcr J, Latimer AM, Henley JB, Rountree NR, Fierer N, Lucky A, *et al.* (2012). A Jungle in There: Bacteria in Belly Buttons are Highly Diverse, but Predictable. PLoS ONE 7:e47712.

Hultman J, Ritari J, Romantschuk M, Paulin L, Auvinen P (2008). Universal ligation-detection-reaction microarray applied for compost microbes. BMC Microbiol 8:237-2180-8-237.

Hurt RA, Qiu X, Wu L, Roh Y, Palumbo AV, Tiedje JM, *et al.* (2001). Simultaneous recovery of RNA and DNA from soils and sediments. Appl Environ Microbiol 67:4495-4503.

Huse SM, Welch DM, Morrison HG, Sogin ML (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol 12:1889-1898.

Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML (2008). Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing. PLoS Genet 4:e1000255.

Huson DH, Auch AF, Qi J, Schuster SC (2007). MEGAN analysis of metagenomic data. Genome Res 17:377-386.

Hwang K, Oh J, Kim TK, Kim BK, Yu DS, Hou BK, *et al.* (2013). CLUSTOM: a novel method for clustering 16S rRNA next generation sequences by overlap minimization. PLoS One 8:e62623.

Ingalls AE, Shah SR, Hansman RL, Aluwihare LI, Santos GM, Druffel ER, *et al.* (2006). Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. Proc Natl Acad Sci U S A 103:6442-6447.

Jarvius J, Melin J, Goransson J, Stenberg J, Fredriksson S, Gonzalez-Rey C, *et al.* (2006). Digital quantification using amplified single-molecule detection. Nat Methods 3:725-727.

Jennings DH (1995). The Physiology of Fungal Nutrition. Cambridge University Press: Cambridge, United Kingdom.

Jiang X, Zhang H, Sheng H, Wang Y, He Y, Zou F, *et al.* (2012). Two-Stage Clustering (TSC): A Pipeline for Selecting Operational Taxonomic Units for the High-Throughput Sequencing of PCR Amplicons. PLoS ONE 7:e30230.

Joint I, Mühling M, Querellou J (2010). Culturing marine bacteria – an essential prerequisite for biodiscovery. Microbial Biotechnology 3:564-575.

Juottonen H, Tuittila ES, Juutinen S, Fritze H, Yrjala K (2008). Seasonality of rDNA- and rRNA-derived archaeal communities and methanogenic potential in a boreal mire. ISME J 2:1157-1168.

Karner MB, DeLong EF, Karl DM (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature 409:507-510.

Katoh K, Standley DM (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Khanna M, Park P, Zirvi M, Cao W, Picon A, Day J, *et al.* (1999). Multiplex PCR/LDR for detection of K-ras mutations in primary colon tumors. Oncogene 18:27-38.

Kinsey G, Paterson R, Kelley J (2003). Filamentous fungi in water systems. In: The Handbook of Water and Wastewater Microbiology. Academic Press, pp 77-819.

Kirchman DL (2008). Introduction and overview. In: Microbial Ecology of the Oceans. Wiley-Blackwell: New Jersey, pp 1-26.

Kitahara K, Yasutake Y, Miyazaki K (2012). Mutational robustness of 16S ribosomal RNA, shown by experimental horizontal gene transfer in Escherichia coli. Proc Natl Acad Sci U S A 109:19220-19225.

Klappenbach JA, Saxman PR, Cole JR, Schmidt TM (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. Nucleic Acids Research 29:181-184.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, *et al.* (2012). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research .

Knoll A (1992). The early evolution of eukaryotes: a geological perspective. Science 256:622-627.

Kopczynski ED, Bateson MM, Ward DM (1994). Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultivated microorganisms. Appl Environ Microbiol 60:746-748.

Koskenniemi K, Lyra C, Rajaniemi-Wacklin P, Jokela J, Sivonen K (2007). Quantitative Real-Time PCR Detection of Toxic *Nodularia* Cyanobacteria in the Baltic Sea. Appl Environ Microbiol 73:2173-2179.

Krause E, Wichels A, Gimenez L, Lunau M, Schilhabel MB, Gerdts G (2012). Small changes in pH have direct effects on marine bacterial community composition: a microcosm approach. PLoS One 7:e47035.

Kube M, Chernikova TN, Al-Ramahi Y, Beloqui A, Lopez-Cortez N, Guazzaroni M, *et al.* (2013). Genome sequence and functional genomic analysis of the oil-degrading bacterium Oleispira antarctica. Nat Commun 4.

Kumar S, Carlsen T, Mevik BH, Enger P, Blaalid R, Shalchian-Tabrizi K, *et al.* (2011). CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. BMC Bioinformatics 12:182-2105-12-182.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol 12:118-123.

Kunin V, Hugenholtz P (2010). PyroTagger: A fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. The Open Journal:1-8.

Lahr DJG, Katz LA (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. BioTechniques 47:857-866.

Le Roy T, Llopis M, Lepage P, Bruneau A, Rabot S, Bevilacqua C, *et al.* (2012). Intestinal microbiota determines development of non-alcoholic fatty liver disease in mice. Gut.

Lee S, Fuhrman JA (1991). Spatial and temporal variation of natural bacterioplankton assemblages studied by total genomic DNA cross-hybridisation. Limnol Oceanogr 36:1277-1287.

Leff LG, Dana JR, McArthur JV, Shimkets LJ (1995). Comparison of methods of DNA extraction from stream sediments. Appl Environ Microbiol 61:1141-1143.

Legendre P, Legendre L (1998). Numerical Ecology. Elsevier: Amsterdam.

Lesk AM (2002). Introduction to Bioinformatics. Oxford University Press: Oxford.

Levén L (2006). Anaerobic Digestion at Mesophilic and Thermophilic Temperature With Emphasis on Degradation of Phenols and Structures of Microbial Communities.

Levén L, Nyberg K, Schnürer A (2012). Conversion of phenols during anaerobic digestion of organic solid waste--a review of important microorganisms and impact of temperature. J Environ Manage 95 Suppl:S99-103.

Levén L, Eriksson ARB, Schnürer A (2007). Effect of process temperature on bacterial and archaeal communities in two methanogenic bioreactors treating organic household waste. FEMS Microbiol Ecol 59:683-693.

Li JB, Gao Y, Aach J, Zhang K, Kryukov GV, Xie B*, et al.* (2009). Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. Genome Res 19:1606-1615.

Li W, Jaroszewski L, Godzik A (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17:282-283.

Liesack W, Weyland H, Stackebrandt E (1991). Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria. Microb Ecol 21:191-198.

Linné Cv (1758-59). Systema Naturae: Per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis. : Holmiae.

Liu W-T, Jansson JK (2010). Environmental molecular microbiology.

Liu WT, Mirzabekov AD, Stahl DA (2001). Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. Environ Microbiol 3:619-629.

Lomakin A, Frank-Kamenetskii MD (1998). A theoretical analysis of specificity of nucleic acid interactions with oligonucleotides and peptide nucleic acids (PNAs). J Mol Biol 276:57-70.

Loy A, Schulz C, Lucker S, Schopfer-Wendels A, Stoecker K, Baranyi C*, et al.* (2005). 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order "Rhodocyclales". Appl Environ Microbiol 71:1373-1386.

Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J*, et al.* (2002). Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl Environ Microbiol 68:5064-5081.

Lozupone C, Knight R (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228-8235.

Lozupone CA, Knight R (2007). Global patterns in bacterial diversity. Proc Natl Acad Sci U S A 104:11436-11440.

Ludwig W, Schleifer KH (1994). Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol Rev 15:155-173.

Madigan MT, Martinko JM, Parker J (2003). Brock Biology of Microorganisms. Prentice Hall, Pearson Education, Inc.: New York, the United States of America.

Magurran AE (2004). Measuring Biological Diversity. Blackwell: Malden (Ma.).

Maidak BL, Cole JR, Lilburn TG, Parker CT,Jr, Saxman PR, Farris RJ*, et al.* (2001). The RDP-II (Ribosomal Database Project). Nucleic Acids Res 29:173-174.

Mao S, Zhang R, Wang D, Zhu W (2012). The diversity of the fecal bacterial community and its relationship with the concentration of volatile fatty acids in the feces during subacute rumen acidosis in dairy cows. BMC Vet Res 8:237-6148-8-237.

Mardis ER (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387-402.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA*, et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Matsen F, Kodner R, Armbrust EV (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11:538.

Maxam AM, Gilbert W (1977). A new method for sequencing DNA. Proceedings of the National Academy of Sciences 74:560-564.

Mayr E (2001). What Evolution is. Basic Books: New York, NY.

Mayr E (1982). The Growth of Biological Thought: Diversity, Evolution, and Inheritance. The Belknap Press of Harvard University Press: Cambridge, Mass.

McHugh S, Carton M, Mahony T, O'Flaherty V (2003). Methanogenic population structure in a variety of anaerobic bioreactors. FEMS Microbiol Lett 219:297-304.

Medvedeva N, Polyak Y, Kankaanpää H, Zaytseva T (2009). Microbial responses to mustard gas dumped in the Baltic Sea. Mar Environ Res 68:71-81.

Methé BA, Hiorns WD, Zehr JP (1998). Contrasts between marine and freshwater bacterial community composition: Analyses of communities in Lake George and six other Adirondack lakes. Limnol Oceanogr 43(2):368-374.

Metzker ML (2005). Emerging technologies in DNA sequencing. Genome Res 15:1767-1776.

Metzker ML (2010). Sequencing technologies mdash] the next generation. Nat Rev Genet 11:31-46.

Miller CB, Wheeker PA (2012). Biological Oceanography. Wiley-Blackwell.

Molbak L, Sommer HM, Johnsen K, Boye M, Johansen M, Moller K, *et al.* (2006). Freezing at -800 degrees C distorts the DNA composition of bacterial communities in intestinal samples. Curr Issues Intest Microbiol 7:29-34.

Moore EB, Tindall B, Santos VP, Pieper D, Ramos J, Palleroni N (2006). Nonmedical: Pseudomonas. In: Springer New York, pp 646-703.

Mori H, Maruyama F, Kurokawa K (2010). VITCOMIC: visualization tool for taxonomic compositions of microbial communities based on 16S rRNA gene sequences. BMC Bioinformatics 11:332.

Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420:806-810.

Nagata T (2008). Organic Matter?Bacteria Interactions in Seawater. In: Microbial Ecology of the Oceans. John Wiley & Sons, Inc., pp 207-241.

Nakai R, Abe T, Takeyama H, Naganuma T (2011). Metagenomic analysis of 0.2-mum-passable microorganisms in deep-sea hydrothermal fluid. Mar Biotechnol (NY) 13:900-908.

Nam YD, Park SL, Lim SI (2012). Microbial composition of the Korean traditional food "kochujang" analyzed by a massive sequencing technique. J Food Sci 77:M250-6.

NAS (2003). Oil in the Sea III: Inputs, Fates and Effects. National Academy of Sciences, National Academy Press: Washington, D.C.

Nawrocki EP, Kolbe DL, Eddy SR (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335-1337.

Nelson KE, Bryan PA, White BA (2010). Genomics and Metagenomics: History and Progress. In: Environmental Molecular Microbiology. Caister Academic Press, pp 21-36.

Nieminen TT, Koskinen K, Laine P, Hultman J, Sade E, Paulin L, *et al.* (2012a). Comparison of microbial communities in marinated and unmarinated broiler meat by metagenomics. Int J Food Microbiol 157:142-149.

Nieminen TT, Valitalo H, Sade E, Paloranta A, Koskinen K, Bjorkroth J (2012b). The effect of marination on lactic acid bacteria communities in raw broiler fillet strips. Front Microbiol 3:376.

Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, Bergsten J, *et al.* (2010). An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. Fungal Ecology 3:284-287.

Niu B, Fu L, Sun S, Li W (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. - BMC Bioinformatics.2010 Apr 13;11:187.doi: 10.1186/1471-2105-11-187.

Nold SC, Zwart G (1998). Patterns and governing forces in aquatic microbial communities. Aquat Ecol 32:17-35.

Noller HF (1984). Structure of ribosomal RNA. Annu Rev Biochem 53:119-162.

Nyrén P (2007). The history of pyrosequencing. Methods Mol Biol 373:1-14.

Nyrén P (2001). Method for sequencing DNA based on the detection of the relese of pyrophosphate and enzymatic nucleotide degradation. Patents: US 6 258 568BI and WO98/28440.

Nyyssönen M, Hultman J, Ahonen L, Kukkonen I, Paulin L, Laine P, *et al.* Taxonomically and functionally diverse microbial communities in deep crystalline rocks of the Fennoscandian shield. The ISME Journal. (*in press*)

Oh J, Kim BK, Cho WS, Hong SG, Kim KM (2012). PyroTrimmer: a software with GUI for pre-processing 454 amplicon sequences. J Microbiol 50:766-769.

Öhrmalm C, Jobs M, Eriksson R, Golbob S, Elfaitouri A, Benachenhou F, *et al.* (2010). Hybridization properties of long nucleic acid probes for detection of variable target sequences, and development of a hybridization prediction algorithm. Nucleic Acids Res 38:e195.

Ojaveer H, Jaanus A, MacKenzie BR, Martin G, Olenin S, Radziejewska T, *et al.* (2010). Status of Biodiversity in the Baltic Sea. PLoS ONE 5:e12467.

Okabe S, Kamagata Y (2010). Wastewater treatment. In: Environmental Molecular Microbiology. caister Academic Press: Norfolk, UK, pp 191-210.

Oros-Sichler M, Costa R, Heuer H, Smalla K (2007). Molecular Fingerprinting Techniques to Analyze Soil Microbail Communities. In: Modern Soil Microbiology. CRC Press, pp 355-386.

Palmer C, Bik EM, Eisen MB, Eckburg PB, Sana TR, Wolber PK, *et al.* (2006). Rapid quantitative profiling of complex microbial populations. Nucleic Acids Res 34:e5.

Pandey RV, Nolte V, Boenigk J, Schlotterer C (2011). CANGS DB: a stand-alone web-based database tool for processing, managing and analyzing 454 data in biodiversity studies. BMC Res Notes 4:227-0500-4-227.

Parks DH, Beiko RG (2010). Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26:715-721.

Pei A, Oberdorf WE, Nossa CW, Chokshi P, Blaser MJ, Yang L, *et al.* (2011). Diversity of 23S rRNA Genes within Individual Prokaryotic Genomes. In: Handbook of Molecular Microbial Ecology I. John Wiley & Sons, Inc., pp 17-27.

Peplies J, Glöckner FO, Amann R (2003). Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. Appl Environ Microbiol 69:1397-1407.

Peura S (2012). Bacterial communities in stratified humic lakes.

Pinto AJ, Raskin L (2012). PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. PLoS ONE 7:e43093.

Plassart P, Terrat S, Thomson B, Griffiths R, Dequiedt S, Lelievre M, *et al.* (2012). Evaluation of the ISO Standard 11063 DNA Extraction Procedure for Assessing Soil Microbial Abundance and Community Structure. PLoS ONE 7:e44279.

Polanski A, Kimmel M (2007). Bioinformatics. Springer: Berlin.

Prosser J, Jansson JK, Liu W- (2010). Nucleic-acid-based Characterization of Community Structure and Function. In: Environmental Molecular Microbiology. Caister Academic Press, pp 63-86.

Pycke BFG, Etchebehere C, van dC, Negroni A, Verstraete W, Boon N (2011). A time-course analysis of four full-scale anaerobic digesters in relation to the dynamics of change of their microbial communities. Water Science & Technology 63:769-775, doi:10.2166/wst.2011.307.

Quackenbush J (2002). Microarray data normalization and transformation. Nat Genet 32 Suppl:496-501.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590-6.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, *et al.* (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Methods 6:639-641.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011). Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12:38-2105-12-38.

Rastogi G, Sani R (2011). Molecular Techniques to Assess Microbial Community Structure, Function, and Dynamics in the Environment. In: Springer New York, pp 29-57.

Reeder J, Knight R (2010). Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Methods 7:668-669.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008). MetaSim—A Sequencing Simulator for Genomics and Metagenomics. PLoS ONE 3:e3373.

Riemann L, Leitet C, Pommier T, Simu K, Holmfeldt K, Larsson U, *et al.* (2008). The Native Bacterioplankton Community in the Central Baltic Sea Is Influenced By Freshwater Bacterial Species. Appl Environ Microbiol 74:503-515.

Riemann L, Middelboe M (2002). Stability of bacterial and viral community compositions in Danish coastal waters as depicted by DNA fingerprinting techniques. Aquat Microb Ecol 27:219-232.

Riley MA (2011). Population Genomics Informs Our Understanding of the Bacterial Species Concept. In: Handbook of Molecular Microbial Ecology I. John Wiley & Sons, Inc., pp 75-82.

Ritari J, Koskinen K, Hultman J, Kurola JM, Kymalainen M, Romantschuk M, *et al.* (2012). Molecular analysis of meso- and thermophilic microbiota associated with anaerobic biowaste degradation. BMC Microbiol 12:121-2180-12-121.

Ritari J, Paulin L, Hultman J, Auvinen P (2009). Application of hybridization control probe to increase accuracy on ligation detection or minisequencing diagnostic microarrays. BMC Research Notes 2:249.

Riviere D, Desvignes V, Pelletier E, Chaussonnerie S, Guermazi S, Weissenbach J, *et al.* (2009). Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. ISME J 3:700-714.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyrén P (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242:84-89.

Ronaghi M, Uhlén M, Nyrén P (1998). A Sequencing Method Based on Real-Time Pyrophosphate. Science 281:363-365.

Ros M, Franke-Whittle IH, Morales AB, Insam H, Ayuso M, Pascual JA (2013). Archaeal community dynamics and abiotic characteristics in a mesophilic anaerobic co-digestion process treating fruit and vegetable processing waste sludge with chopped fresh artichoke waste. Bioresour Technol 136:1-7.

Rosen M, Callahan B, Fisher D, Holmes S (2012). Denoising PCR-amplified metagenome data. BMC Bioinformatics 13:283.

Rossello-Mora R (2006). DNA-DNA Reassociation Methods Applied to Microbial Taxonomy and Their Critical Evaluation. In: Springer Berlin Heidelberg, pp 23-50.

Rossello-Mora R, Amann R (2001). The species concept for prokaryotes. FEMS Microbiol Rev 25:39-67.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, *et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348-352.

Routledge RD (1977). On Whittaker's Components of Diversity. Ecology 58:1120-1127.

Routledge RD (1984). Estimating Ecological Components of Diversity. Oikos 42:23-29.

Rudi K, Skulberg OM, Skulberg R, Jakobsen KS (2000). Application of sequence-specific labeled 16S rRNA gene oligonucleotide probes for genetic profiling of cyanobacterial abundance and diversity by array hybridization. Appl Environ Microbiol 66:4004-4011.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74:5463-5467.

Sanger F, Coulson AR (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441-448.

Sasaki D, Hori T, Haruta S, Ueno Y, Ishii M, Igarashi Y (2011). Methanogenic pathway and community structure in a thermophilic anaerobic digestion process of organic solid waste. Journal of Bioscience and Bioengineering 111:41-46.

Satomi M, Kimura B, Hamada T, Shigeaki Harayama S, Fujii T (2002). Phylogenetic study of the genus *Oceanospirillum* based on 16S rRNA and *gyrB* genes: emended description of the genus *Oceanospirillum*, description of *Pseudospirillum* gen. nov., *Oceanobacter* gen. nov. and *Terasakiella* gen. nov. and transfer of *Oceanospirillum jannaschii* and *Pseudomonas stanieri* to *Marinobacterium* as *Marinobacterium jannaschii* comb. nov. and *Marinobacterium stanieri* comb. nov. Int J Syst Evol Microbiol 52:739-747.

Schena M, Shalon D, Davis RW, Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467-470.

Scheuermayer M, Gulder TA, Bringmann G, Hentschel U (2006). *Rubritalea marina* gen. nov., sp. nov., a marine representative of the phylum 'Verrucomicrobia', isolated from a sponge (Porifera). Int J Syst Evol Microbiol 56:2119-2124.

Schloss PD, Westcott SL (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol 77:3219-3226.

Schloss PD (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput Biol 6:e1000844.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75:7537-7541.

Schloss PD, Handelsman J (2006). Introducing TreeClimber, a test to compare microbial community structures. Appl Environ Microbiol 72:2379-2384.

Schloss PD, Gevers D, Westcott SL, (2011). Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. PLoS ONE 6:e27310.

Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, Gartemann K, *et al.* (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J Biotechnol 136:77-90.

Schnürer A, Schnürer J (2006). Fungal survival during anaerobic digestion of organic household waste. Waste Manag 26:1205-1211.

Sekiguchi Y (2006). Yet-to-be cultured microorganisms relevant to methane fermentation processes. Microbes Environ 21:1-15.

Sergeant MJ, Constantinidou C, Cogan T, Penn CW, Pallen MJ (2012). High-Throughput Sequencing of 16S rRNA Gene Amplicons: Effects of Extraction Procedure, Primer Length and Annealing Temperature. PLoS ONE 7:e38094.

Shendure J, Ji H (2008). Next-generation DNA sequencing. Nat Biotechnol 26:1135-1145.

Shendure J, Porreca G, Reppas N, Lin X, McCutcheon J, Rosenbaum A, *et al.* (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science (New York, N.Y.) 309:1728-1732.

Shin SG, Lee S, Lee C, Hwang K, Hwang S (2010). Qualitative and quantitative assessment of microbial community in batch anaerobic digestion of secondary sludge. Bioresour Technol 101:9461-9470.

Siam R, Mustafa GA, Sharaf H, Moustafa A, Ramadan AR, Antunes A, *et al.* (2012). Unique prokaryotic consortia in geochemically distinct sediments from Red Sea Atlantis II and discovery deep brine pools. PLoS One 7:e42872.

Singleton DR, Furlong MA, Rathbun SL, Whitman WB (2001). Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. Appl Environ Microbiol 67:4374-4376.

Sipos R, Székely A, Révész S, Márialigeti K (2010). Addressing PCR Biases in Environmental Microbiology Studies. In: Humana Press, pp 37-58.

Sivonen K, Kononen K, Esala A, Niemelä SI (1989a). Toxicity and isolation of the cyanobacterium *Nodularia spumigena* from the southern Baltic Sea in 1986. Hydrobiologia 185:3-8.

Sivonen K, Kononen K, Carmichael WW, Dahlem AM, Rinehart KL, Kiviranta J, *et al.* (1989b). Occurence of the Hepatotoxic Cyanobacterium *Nodularia spumigena* in the Baltic Sea and Structure of the Toxin. Appl Environ Microbiol 55:1990-1995.

Smith EP, van Belle G (1984). Nonparametric Estimation of Species Richness. Biometrics 40:119-129.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, *et al.* (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103:12115-12120.

Stackebrandt E, Ludwig W, Fox GE (1985). 16S ribosomal RNA oligonucleotide cataloguing. In: Methods in Microbiology. Academic Press, pp 75-107.

Stahl DA, Lane DJ, Olsen GJ, Pace NR (1985). Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. Appl Environ Microbiol 49:1379-1384.

Ståhl PL, Lundeberg J (2012). Toward the single-hour high-quality genome. Annu Rev Biochem 81:359-378.

Stal LJ, Albertano P, Bergman B, von Bröckel K, Gallon JR, Hayes PK, *et al.* (2003). BASIC: Baltic Sea cyanobacteria. An investigation of the structure and dynamics of water blooms of cyanobacteria in the Baltic Sea—responses to a changing environment. Cont Shelf Res 23:1695-1714.

Steger D, Berry D, Haider S, Horn M, Wagner M, Stocker R, *et al.* (2011). Systematic Spatial Bias in DNA Microarray Hybridization Is Caused by Probe Spot Position-Dependent Variability in Lateral Diffusion. PLoS ONE 6:e23727.

Su Y, Li B, Zhu WY (2013). Fecal microbiota of piglets prefer utilizing DL-lactate mixture as compared to D-lactate and L-lactate in vitro. Anaerobe 19:27-33.

Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, *et al.* (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. Nucleic Acids Res 37:e76.

Swan BK, Valentine DL (2009). Diversity of Archaea. In: Els. John Wiley & Sons, Ltd.

Szemes M, Bonants P, de Weerdt M, Baner J, Landegren U, Schoen CD (2005). Diagnostic application of padlock probes--multiplex detection of plant pathogens using universal microarrays. Nucleic Acids Res 33:e70.

Teeling H, Glöckner FO (2012). Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. Brief Bioinform 13:728-742.

Thompson JR, Marcelino LA, Polz MF (2002). Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. Nucleic Acids Res 30:2083-2088.

Torsvik VL, Øvreås L (2011). DNA Reassociation Yields Broad-Scale Information on Metagenome Complexity and Microbial Diversity. In: Handbook of Molecular Microbial Ecology I. John Wiley & Sons, Inc., pp 3-16.

Van de Peer Y, Chapelle S, De Wachter R (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. Nucleic Acids Res 24:3381-3391.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, *et al.* (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science 304:66-74.

Verschoor A, Srivastava S, Grassucci R, Frank J (1996). Native 3D structure of eukaryotic 80s ribosome: morphological homology with E. coli 70S ribosome. J Cell Biol 133:495-505.

Větrovský T, Baldrian P (2013). Analysis of soil fungal communities by amplicon pyrosequencing: current approaches to data analysis and the introduction of the pipeline SEED. Biol Fertility Soils :1-11.

Vornhagen J, Stevens M, McCormick DW, Dowd SE, Eisenberg JN, Boles BR, *et al.* (2013). Coaggregation occurs amongst bacteria within and between biofilms in domestic showerheads. Biofouling 29:53-68.

Wagner M, Smidt H, Loy A, Zhou J (2007). Unravelling microbial communities with DNA-microarrays: challenges and future directions. Microb Ecol 53:498-506.

Wallenius K, Rita H, Simpanen S, Mikkonen A, Niemi RM (2010). Sample storage for soil enzyme activity and bacterial community profiles. J Microbiol Methods 81:48-55.

Wang GC, Wang Y (1997). Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. Appl Environ Microbiol 63:4645-4650.

Wang GC, Wang Y (1996). The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. Microbiology 142 ( Pt 5):1107-1114.

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

Ward DM, Melendrez MC, Becraft ED, Klatt CG, Wood JM, Cohan FM (2011). Metagenomic Approaches for the Identification of Microbial Species. In: Handbook of Molecular Microbial Ecology I. John Wiley & Sons, Inc., pp 105-109.

Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, *et al.* (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic Bacteriology 37:463-464.

White JR, Maddox C, White O, Angiuoli S, Fricke WF (2013). CloVR-ITS: Automated internal transcribed spacer amplicon sequence analysis pipeline for the characterization of fungal microbiota. Microbiome 1:6.

White JR, Nagarajan N, Pop M (2009). Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS Comput Biol 5:e1000352.

Whitman WB, Coleman DC, Wiebe WJ (1998). Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences 95:6578-6583.

Whittaker RH (1972). Evolution and Measurement of Species Diversity. Taxon 21:213-251.

Whittaker RH (1960). Vegetation of the Siskiyou Mountains, Oregon and California. Ecol Monogr 30:279-338.

Wilson MV, Shmida A (1984). Measuring Beta Diversity with Presence-Absence Data. J Ecol 72:1055-1064.

Winter C, Matthews B, Suttle CA (2013). Effects of environmental variation and spatial distance on Bacteria, Archaea and viruses in sub-polar and arctic waters. ISME J 7:1507-1518.

von Wintzingerode F, Selent B, Hegemann W, Gobel UB (1999). Phylogenetic analysis of an anaerobic, trichlorobenzene-transforming microbial consortium. Appl Environ Microbiol 65:283-286.

von Wintzingerode F, Gobel UB, Stackebrandt E (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21:213-229.

Woese CR, Fox GE (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088-5090.

Wooley JC, Godzik A, Friedberg I (2010). A Primer on Metagenomics. PLoS Comput Biol 6:e1000667.

Wu C, Carta R, Zhang L (2005). Sequence dependence of cross-hybridization on short oligo microarrays. Nucleic Acids Research 33:e84-e84.

Wu JH, Liu WT, Tseng IC, Cheng SS (2001). Characterization of microbial consortia in a terephthalate-degrading anaerobic granular sludge system. Microbiology 147:373-382.

Wu JY, Jiang XT, Jiang YX, Lu SY, Zou F, Zhou HW (2010). Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. BMC Microbiol 10:255-2180-10-255.

Wu QL, Zwart G, Schauer M, Kamst-van Agterveld MP, Hahn MW (2006). Bacterioplankton community composition along a salinity gradient of sixteen high-mountain lakes located on the Tibetan Plateau, China. Appl Environ Microbiol 72:5478-5485.

Xu J (2011). Microbial Ecology in the Age of Metagenomics. In: Handbook of Molecular Microbial Ecology I. John Wiley & Sons, Inc., pp 111-122.

Yakimov MM, Giuliano L, Gentile G, Crisafi E, Chernikova EN, Abraham W-, *et al.* (2003). *Oleispira antarctica* gen. nov., sp. nov., a novel hydrocarbonoclastic marine bacterium isolated from Antarctic coastal sea water. Int J Syst Evol Microbiol 53:779-785.

Yang YH, Paquet A, Dudoit S (2007). Marray: Exploratory analysis for two-color spotted microarray data.

Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ (2012). Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome. PLoS ONE 7:e33865.

Zehnder AJB (1978). Ecology of methane formation. In: Water Pollution Microbiology. John Wiley & Sons: London, United Kingdom, pp 349-376.

Zhang M, Liu W, Nie X, Li C, Gu J, Zhang C (2012). Molecular analysis of bacterial communities in biofilms of a drinking water clearwell. Microbes Environ 27:443-448.