New DNA-based Approaches for the Study of Microbial Communities

Karst, Søren Michael

DOI (link to publication from Publisher):
10.5278/vbn.phd.engsci.00186

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

NEW DNA-BASED APPROACHES FOR THE STUDY OF MICROBIAL COMMUNITIES

BY
SØREN MICHAEL KARST

DISSERTATION SUBMITTED 2016

AALBORG UNIVERSITY
DENMARK

NEW DNA-BASED APPROACHES FOR THE STUDY OF MICROBIAL COMMUNITIES

by

Søren Michael Karst

AALBORG UNIVERSITY
DENMARK

Dissertation submitted 27th October 2016

Dissertation submitted: 27th October 2016

PhD supervisor: Prof. Per Halkjær Nielsen
Aalborg University

PhD committee: Niels T. Eriksen (chairman)
Institut for Kemi og Biovidenskab
Aalborg University

Søren J. Sørensen
Department of Biology
University of Copenhagen

Thomas Rattei
CUBE – Division Computational Systems Biology
Department of Microbiology and Ecosystem Science
University of Vienna

PhD Series: Faculty of Engineering and Science, Aalborg University

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-830-7

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2016

ENGLISH SUMMARY

Microbes are an essential part of the global biosphere, where they play a role in everything from global nitrogen re-cycling, to treatment of wastewater and human health. The complexity of the microbial world is enormous, partly due to the diversity of microbes, but also due to their tendency to exist in communities with millions of cells, belonging to thousands of different species, which have intricate interactions with each other and the environment. Microbes and their impact have been studied for hundreds of years, using culture-based techniques, but as the extent of microbial complexity has become evident, the need for new techniques to study the microbial world has become pressing. Cutting-edge DNA sequencing techniques have formed the basis for new high-throughput and high resolution methods to study the microbial communities and their activity. With SSU rRNA amplicon sequencing it is now possible to identify the members of complex microbial communities from hundreds of samples in a matter of days. At the same time metagenomics makes it possible to determine genes and reconstruct metabolic pathways present in microbial communities, and thereby give insight into their possible role in the ecosystem. The development of the technologies, since their introduction 10 years ago, has been incredible, but there are still core issues that need to be addressed.

The aim of this PhD project, was to implement and improve DNA-based sequencing methodologies for the study of complex microbial communities. First, we optimized the SSU rRNA amplicon sequencing approach to study the bacterial communities involved in wastewater treatment (Chapter 2). We found that the mechanical lysing of the cells had to be increased, compared to the established standard, and that PCR primers targeting the V1-3 region of the 16S rRNA best captured operational important bacteria. The optimized approach was subsequently used in a range of studies and the experience from these was compiled in a book chapter (Chapter 3). Next, we developed a high-throughput technique to generate high-quality SSU rRNA sequences using reverse transcription of SSU rRNA and synthetic long-read sequencing by molecular tagging (Chapter 4). The method is superior to conventional techniques based on cloning and Sanger sequencing, as it is truly high-throughput and avoids PCR primer bias. With the technique, we generated 30 000 new full-length reference SSU rRNA gene sequences, with a high degree of novel diversity. The method represents a paradigm shift, as it will now be possible to make comprehensive databases for any environment. Finally, we developed the software tool “mmgenome” that enables effortless and reproducible genome extraction from metagenomes (Chapter 5). Compared to other tools, mmgenome can readily integrate information from various binning algorithms and serves as a visualization platform. The tool has been used in numerous studies to extract genomes and document the complex binning process (Chapter 6).

DANSK RESUME

Mikroorganismer er en essentiel del af jordens biosfære, og de spiller en vigtig rolle i alt fra det globale kvælstofkredsløb, til spildevandsrensning og human sundhed. Den globale mikrobielle kompleksitet er enorm, til dels fordi den mikrobielle diversitet er meget høj, men også fordi mikroorganismer foretrækker a gro i samfund med millioner af celler fra flere forskellige arter, som alle interagerer med hinanden og deres omgivelser. Mikroorganismer og deres indflydelse på deres omgivelser har været studeret i flere hundrede år ved hjælp af kultur-baserede teknikker, men efterhånden som man er begyndt at forstå omfanget af den mikrobielle diversitet, er det blevet klart, at der er behov for nye teknikker, der giver et mere detaljeret helhedsbillede. Nye cutting-edge DNA sekvensering teknikker danner grundlag for nye, hurtige analysemetoder med høj opløselighed, som kan anvendes direkte på prøver fra miljøet. Med SSU rRNA amplikon sekvensering er det nu muligt at identificere medlemmer af mikrobielle samfund fra hundrede af prøver på bare få dage. Samtidig er er det blevet muligt, ved hjælp af metagenomics, at bestemme alle generne i et mikrobielt samfund og forudsige mulige metaboliske processer, og derved give indblik i mikroorganismernes aktiviteter og roller. Udviklingen af disse teknologier er gået stærkt siden deres introduktion for 10 år siden, men der er stadig essentielle problemer som der skal tages hånd om.

Formålet med dette PhD projekt, var at implementere og forbedre DNA-baserede sekvenserings metoder brugt til at studere komplekse mikrobielle samfund. I første omgang optimerede vi SSU rRNA amplikon sekvensering til analyse af mikrobielle samfund i aktivt slam fra renseanlæg (Kapitel 2). Vi opdagede at det fysiske element af lyseringsprocessen skulle forøges, sammenlignet med standard metoden, og at PCR primere, der ramte V1-3 regionen af 16S rRNA, var bedst til at ramme bakterier der er vigtige i forhold til rensnings processen. Den optimerede metode blev efterfølgende brugt i en række studier, og erfaringerne fra disse studier blev samlet og udgivet i et bogkapitel (Kapitel 3). Efterfølgende, udviklede vi en hurtig metode til at generere fuld længde SSU rRNA sekvenser af høj kvalitet, ved hjælp af reverse transskription og sekvensering af syntetisk lange reads. Metoden er overlegen sammenlignet med alternative teknikker, på grund af dens hastighed og fordi vi undgår at bruge PCR primere. Med denne nye teknik, fik vi generet 30 000 nye SSU rRNA reference sekvenser, hvor en stor del udgjorde hidtil ukendt diversitet. Metoden repræsenterer et paradigmeskift indenfor feltet, da det nu er muligt at skabe udførlige databaser for hvilket som helst miljø. Derudover udviklede vi en softwarepakke kaldet ”mmgenome”, der gør det nemt at ekstrahere genomer fra metagenomer på en reproducerbar måde (Kapitel 5). Styrken ved mmgenome er visualisering og dokumentation i hvert enkelt trin. Samtidig kan man integrere informationer fra andre typer af analyser, som kan bidrage til en bedre ekstraktion. Værktøjet er blevet brugt i en lang række studier til at ekstrahere genomer (Kapitel 6).

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor Per Halkjær Nielsen for the opportunity to make this PhD thesis. Your advice and support has been invaluable, and I am grateful for the trust you have given me as well as the opportunity and encouragement to pursue my ideas.

I would also like to thank my many colleagues in the environmental biotechnology group for contributing to my PhD study being a fun and enjoyable experience. Special thanks to Mads Albertsen and Rasmus Kirkegaard for the great collaboration as well as serious and not so serious discussions.

Lastly, I would like to thank my family and friends for always supporting me during my PhD study.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

Microbes are everywhere and in unfathomable numbers. Estimates predict there are up towards 3×10^{30} microbial cells on earth, which is roughly equivalent to 50% of the earth's total plant biomass[1,2]. Most microbes live in soil, oceans and sediments, where they are the main drivers of the biochemical re-cycling of the building blocks of life, e.g. carbon, nitrogen and phosphorus. Re-cycling is essential for the continued existence of life and the scale of recycling is staggering. Looking at nitrogen alone, the flux in and out of the atmosphere facilitated by microbes, is in the range of 500 million tons of nitrogen per year, which is equivalent of 5% of the nitrogen in all living biomass on earth[3,4].

Since the discovery of the first microbes by Antonie van Leeuwenhoek in the 17th century[5], we have realized that microbes play an important part in our daily lives. Initially microbes were mostly associated with diseases, but in the 21st century, we started to realize the magnitude of their involvement in all human activities. For example, studies of the human microbiome have shown that our natural existing microbial community is involved in digestion, nutrient uptake and stimulation of the immune system[6]. Furthermore, microorganisms are involved in critical parts of our infrastructure, such as biological wastewater treatment[7], biogas production[8] and agriculture[9].

The last hundred years of human activity have had a significant impact on our global nutrient cycles and hence also disturbed the natural balance of our ecosystem[10]. For example, due to the demand for fertilizer used in food production, industrial fixation of nitrogen from the atmosphere is currently at the same scale as biotic fixation (see **Figure 1**). As nitrogen is often the growth limiting factor in terrestrial and oceanic ecosystems, this has led to increased microbial activity with eutrophication, loss of biodiversity and emissions of the greenhouse gas N_2O[4]. Hence, understanding the microbes involved in nitrogen transformation is important in order to understand how human activity impact the complex global nutrient cycles. Interestingly, this field is one of the most extensively studied in microbiology, starting with Sergei Winogradsky describing the first nitrifying and nitrogen-fixing bacteria more than a century ago[11]. Despite the long history of research, discoveries are still being made, which completely alter our fundamental understanding of the organisms involved in the global nitrogen cycle. For example, it was recently shown that complete nitrification, oxidation of ammonia to nitrate, could be performed by a single bacteria, instead of two, which had been a dogma for 100 years[12,13]. This underlines that even

for key microbial processes, which have been researched extensively, our understanding is most likely still incomplete.

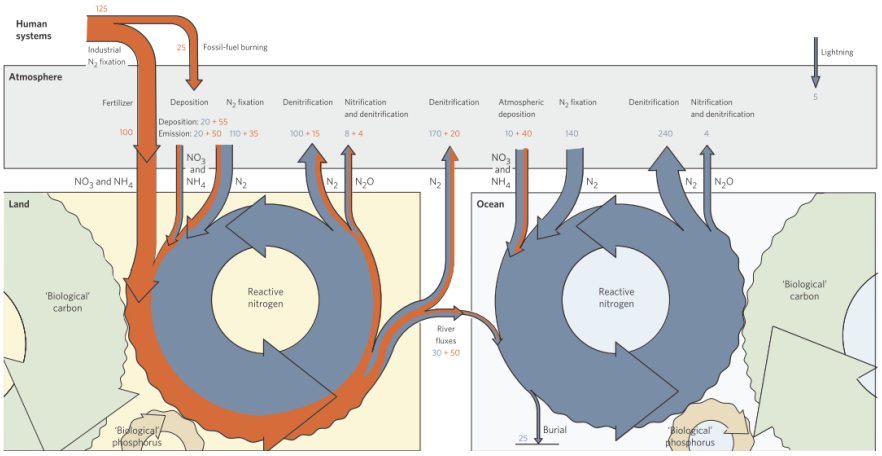

***Figure 1**: Overview of the global nitrogen cycle*[3]*. Arrows illustrate flux (Tg N per year) between two major reservoirs. Blue signifies natural fluxes and orange signifies contribution from human activities. Microbes are the main drivers of the natural N_2-fixation, nitrification and denitrification. Flux estimates are known to be between +/-20-50%.* *Reprinted by permission from Macmillan Publishers Ltd: Nature; Gruber, N. & Galloway 2008, copyright 2008.*

There are many other examples of large microbial driven global issues, where our understanding is severely lacking, such as antibiotic resistance[14,15], waste treatment[7,8] and our surroundings impact on the healthy human microbiome[16]. There is a growing acknowledgement of the need to better understand the microbes, which is reflected by the increase in large microbiome research programs being funded as e.g. the Earth Microbiome Project which aims to construct a global catalogue of uncultured microorganisms based on more than 200 000 samples[17]. However, our ability to study the microbes and their impact efficiently is dependent on the development of new technologies.

1.1. THE STUDY OF COMPLEX MICROBIAL SYSTEMS

In most natural systems microbes exist together in microbial communities, often consisting of thousands of different species[18]. Hence, the ability of the microbial communities to perform complex functions, such as methane production from waste, is often an emerging function of the community. Key organisms perform the core conversions, but depend on several secondary partners to be able to perform their task[19]. Hence, to systematically deconstruct this complexity it is important to know the identity and functional capabilities of the organisms. Therefore, the study of complex microbial communities requires methods with high sensitivity, resolution and throughput[20].

Traditionally, the identity and function of microorganisms have been studied by isolating and culturing them in the laboratory, followed by various studies under the microscope, growth studies and biochemical profiling[21]. Most of our current knowledge about microbes have been obtained this way. However, due to the unknown, diverse growth preferences of most microbes, culturing is an iterative and time consuming process, which is often performed manually. This makes it very impractical to scale, although new high-throughput approaches are starting to emerge[22,23]. Furthermore, microorganisms behave differently *in vivo* compared to *in situ*, due to phenotypic and genotypic acclimatization to the artificial growth conditions. Therefore, to efficiently obtain the identity and function of environmental microorganisms, culture independent methods are required that can be applied to samples taken directly from the ecosystem and scaled to capture community dynamics and interaction in time and space.

1.2. CULTURE INDEPENDENT TECHNIQUES

Many modern culture-independent techniques for determining the identity of microorganisms depends on the small subunit ribosomal RNA (SSU rRNA) gene. The SSU rRNA gene is a good phylogenetic marker gene, as it contains alternating conserved and variable regions, and is present in all forms of life (see Figure 2). The SSU rRNA gene codes for a piece of RNA that is an integral part of the ribosome, and its alternating sequence of conserved and variable regions arises from the fact that some regions of the rRNA are more mechanistically important for the function of the ribosome than others.

The use of the SSU rRNA gene for identification (16S in bacteria/archaea and 18S in eukaryotes) was pioneered in 1977 by Woese and Fox[24] and revolutionized our ability to identify members of the microbial world. For 30 years the 16S rRNA gene has been isolated and amplified using polymerase chain reaction (PCR), cloned and then Sanger sequenced to determine the DNA sequence[25]. The approach works well for generating full-length 16S rRNA sequences from environmental samples, but one of the biggest limitations is low throughput. Still, analysis of SSU rRNA gene sequences provided insight into a microbial world, which was much more complex than anticipated from culturing and microscopy. Approximately 10 years ago, the SSU rRNA sequencing framework was adapted to high-throughput DNA sequencing platforms (first Roche 454[26] and later Illumina[27]). Today, microbial community analysis using SSU rRNA amplicon sequencing can be performed on hundreds of samples, producing millions of sequences from each, in a matter of days.

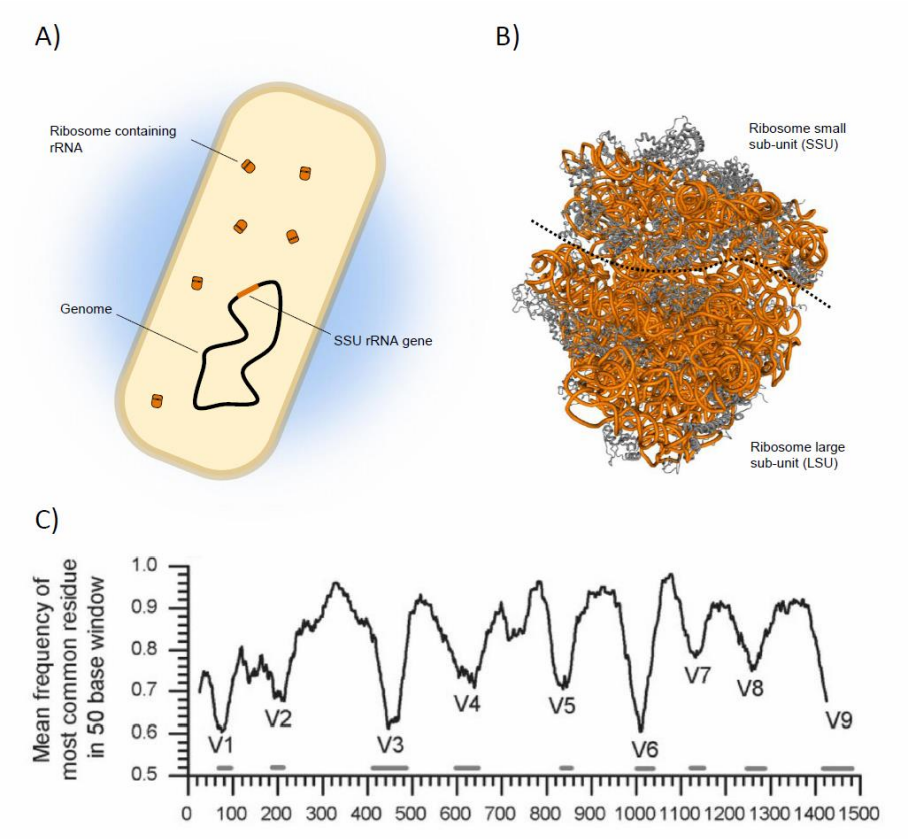

Figure 2: A) Simple illustration of a bacterial cell showing the genome, which contains the SSU rRNA gene, as well as the ribosomes containing the transcribed and processed SSU rRNA. B) 3-D structure depiction of a ribosome (PDB structure ID 4V5D[28]) consisting of a small and large subunit. rRNA highlighted with orange and ribosomal proteins are highlighted with grey. C) Nucleotide conservation across all known bacterial 16S rRNA sequences (adopted from Ashelford et al. 2005[29]). The x-axis shows basepair position and y-axis shows the frequency of the most observed bases at a given position across all sequences (running average).

The availability of a large number of SSU rRNA gene sequences has also been used as the basis for developing single cell microscopy based methods, which have enabled *in situ* identification, as well as the study of ecophysiology and activity. One of the core methods is Fluorescence *in situ* hybridization (FISH), which uses DNA oligo probes complementary to a phylogenetic distinct region of the target bacterial rRNA to tag the ribosomes of cells with fluorescent probes, so they can be identified and counted under a microscope[30]. Several methods derived from FISH were also developed to explore *in situ* activity and ecophysiology such as microautoradiography FISH (substrate utilization and activity)[31], enzyme-labelled fluorescence FISH (Exoenzyme expression)[32], microsphere adhesion to cells FISH (Surface properties)[32], Raman-FISH microspectroscopy (detection of cellular components[33] and

incorporation of labelled substrates[34]). The microscopy based techniques lack the throughput of the SSU rRNA amplicon sequencing, but allows researchers to link identity and activity. There are also efforts being made to combine some of the microscopy based methods with flow cytometry[35] and microfluidics[36] to increase the throughput.

1.2.1. META-OMIC APPROACHES

Modern high-throughput instruments for DNA sequencing and mass-spectrometry have paved the way for meta-omic approaches that can study the metabolic potential (metagenomics), gene expression (metatranscriptomics), protein expression (metaproteomics) and metabolite content (metabolomics) to shed light on the functions of complex microbial communities[37]. Metagenomics and metatranscriptomics can indicate function and activity in complex samples. Due to the complex regulation of mRNA transcription, turnover and translation, gene presence and gene expression does not necessarily result in protein synthesis and activity[38], and therefore trends found by metagenomics and transcriptomics are often supported with other techniques, such as single cell *in situ* techniques[39] or proteomics[40]. Metaproteomics has been applied to microbial communities with some success, but complicated sample preparation, high dynamic range between low abundant and high abundant proteins as well as lack of reference genomes still limits the widespread application in complex samples[41]. Metabolomics is considered to be in its infancy in application to complex microbial communities[42] as microbial cells can contain hundreds of different metabolites with very diverse chemical compositions, and many of these chemicals are often uncharacterized, and therefore difficult to identify and put into context[42].

For the high throughput study of complex microbial communities in environmental ecosystems, the core tools used today are SSU rRNA amplicon sequencing and metagenomics. There are, however, despite their widespread use, there is still critical shortcomings of the methods that needs to be addressed, and this has been the focus of the present PhD thesis.

1.3. AIM

The overall aim of this PhD study was to implement and improve next generation sequencing methodologies for the study of complex microbial communities. Specifically, the main objectives were:

1. Optimize 16S rRNA amplicon sequencing for analysis of the bacterial communities in wastewater treatment.
2. Develop a method for obtaining full-length SSU rRNA sequences to improve reference database and expand our knowledge of the tree of life.
3. Develop metagenomic tools for genome binning and apply them to study important bacteria in complex wastewater treatment samples.

1.1. IMPLEMENTING A STANDARD METHOD FOR 16S RRNA AMPLICON SEQUENCING IN ACTIVATED SLUDGE

During the past five years 16S rRNA amplicon sequencing have become the backbone of most microbial ecology studies. However, despite its widespread use, there are still numerous methodological problems that needs to be addressed. In this PhD thesis one of the main focus areas has been to implement and optimize bacterial 16S rRNA gene community analysis in activated sludge, which is the topic of Albertsen *et al*. 2015 (see Chapter 2). Our accumulated experience from working with the approach in activated sludge for the last four years has been compiled into a book chapter in "Experimental Methods in Wastewater Treatment"[43] (see Chapter 3). The optimized methods and workflows are now the accepted standard within the field of microbial community analysis in activated sludge, and are used in our effort to analyze the thousands of samples within the framework of the Microbial Database for Activated Sludge (MiDAS[44]; www.midasfieldguide.org).

A major issue in relation to taxonomic classification is incomplete reference databases[45,46], and to help solve this problem we developed a novel high-throughput method for generating full-length SSU rRNA gene sequences, which is described in Karst et al. 2016 (see Chapter 4Chapter 4). The main results from the publications will be highlighted and put into context in the next sections.

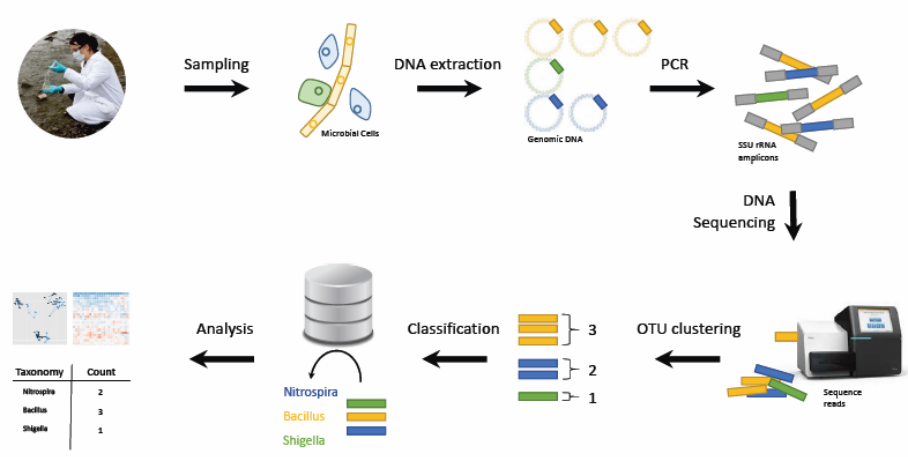

Figure 3: Conceptual overview of SSU rRNA amplicon sequencing (adapted from Karst et al. 2016, see Chapter 3).

1.1.1. DNA EXTRACTION – SMALL CHANGES WITH LARGE IMPACTS

Bacterial cells vary greatly in their resistance to lysis, mainly due to the composition of their cell envelope and the strength of their biofilm matrix[47]. A general example is that some Gram-positive bacteria are more difficult to lyse compared to Gram-negative bacteria due to their peptidoglycan layer[47]. Different extraction efficiencies mean that the DNA from some species might be overrepresented compared to others, thus skewing their relative abundance. Many different parameters can be changed, when optimizing DNA extraction, such as chemicals, enzymes, temperature and physical force. However, even though this has been the topic of numerous articles[48–50], the large difference in sample type (i.e. both matrix and microbial composition) makes it essential to optimize the methods for each individual environment.

Several studies have also tried to optimize DNA extraction for activated sludge, and the general observation was that it is important to include bead beating (mechanical lysing)[51,52]. However, few studies sequenced the samples, and none investigated the impact of the bead beating intensity or compared the results to DNA extraction independent methods to evaluate the results. In Albertsen *et al*., 2015 (see Chapter 2), we conducted the most comprehensive method evaluation and optimization of DNA extraction in activated sludge to date. By analyzing a wide range of bead beating settings, we found that increasing the bead beating, compared to recommendations in the standard protocol, resulted in dramatic increase in relative abundance of notoriously difficult to lyse phyla, such as Actinobacteria and Alphaproteobacteria. In addition, we could show that the impact of different bead beating settings, on the change in the apparent microbial community structure, was of the same magnitude, as the change which occurred naturally over five months in the wastewater treatment plant. This is a prime example of how "small" protocol changes easily can overshadow the biological changes, and the importance of protocol standardization for cross study comparison. To facilitate community wide adoption of standard methods we have started the www.midasfieldguide.org initiative that also includes online versions of all optimized protocols for microbial community analysis of activated sludge.

1.1.2. THE EFFECT OF USING DIFFERENT PCR PRIMERS ON THE OBSERVED MICROBIAL COMMUNITY STRUCTURE

Another crucial step in any 16S rRNA microbial community analysis is the choice of primers, as the primers determine which variable region will be amplified and hence has a significant impact on the observed microbial community[53]. The 16S rRNA primers are usually designed based on reference databases containing full-length sequences of the 16S rRNA gene (e.g. the SILVA database[54]). During primer design, the 16S rRNA gene sequences are aligned and regions with high similarity are considered conserved and further analysis as potential priming sites[55]. The 16S rRNA gene is approximately 1500 bp in most bacteria and consists of 9 variable regions (V1-V9) with 10 relative conserved flanking regions (see Figure 2C). To obtain most

phylogenetic resolution from a 16S rRNA analysis, sequencing of the whole gene would be preferred. However, due to limitations of the current high-throughput next generation sequencing technologies[56], usually only fragments below 550 bp are possible to sequence.

Depending on the ecosystem there are different primer preferences, e.g. human gut studies frequently use primers targeting V1-3 or V3-5 regions[57], while the Earth Microbiome Project uses V4[27]. The choice of primers often relies on historical use or *in silico* evaluation[27,58], and for the primer evaluations that has been made, primers are often assessed based on performance in mock communities[57], change in relative abundances, diversity measures[59] and very rarely compared to PCR-free approaches[60]. To find the best suited primer set for activated sludge, we screened three different broadly used 16S rRNA primer sets (V1-3, V4, V3-4), and compared their performance with PCR-free data from metagenomics, transcriptomics and quantitative FISH of selected phylogenetic groups. The observed microbial community structure for the different primers was very different, and none of them similar to the PCR-free results, illustrating that an objective best primer set is difficult to define. Of the three primer-sets, we chosed the V1-3, as this primer set covered the same range of phyla as the metagenome and was best at capturing the phyla Chloroflexi and Actinobacteria. These phyla are of special interest as they contain groups that are important in the activated sludge process, such as *Tetrasphaera* involved in biological P-removal, filamentous Chloroflexi important for floc structure, and *Ca.* Microthrix known for causing bulking[61]. Hence, there is no perfect choice of primers, and the primers should always be evaluated in respect to the goals of the analysis. In addition, using different primers between studies makes comparisons impossible.

1.1.3. IMPROVING THE SSU rRNA REFERENCE DATABASES BY ORDERS OF MAGNITUDE THROUGH NEW METHODS

An important part of the data processing in SSU rRNA amplicon sequencing is taxonomic classification of the sequences. Proper classification links the sequence with a name and thereby the biological information available in the literature. In taxonomic classification, each sequence is compared to the sequence of references in a database (e.g. SILVA[54], RDP[62] or Greengenes[63]), and inherits the taxonomy of the reference sequence(s)[64]. The database sequences used for classification are primarily near full-length, high quality sequences of SSU rRNA genes, which are used to create a reference alignment and a phylogenetic reference tree, which is used to manually curate the taxonomy of the database[54]. The quality of the used databases influences the classification, and has a large impact on the observed diversity.

Using different databases produce different results, as they are based on different reference sequences and curated by different experts[46]. In addition, some reference sequences simply lack taxonomic information[65]. This can be due to lack of curation

or that no developed taxonomy nomenclature is available, which is the case for many sequences from a majority of uncultured organisms[54]. Addressing the flaws of the SSU rRNA databases will greatly improve the value provided by the SSU rRNA based microbial community analysis. However, curating everything in the databases are a monumental task. A solution is to create ecosystem specialized databases, where effort has been made to add reference sequences and develop the taxonomy of ecosystem critical organisms. An example of such an effort is the MiDAS taxonomy that we maintain for activated sludge, where the fraction of abundant OTUs classified to genus level was improved from 53% to 91% by manually curating the existing taxonomy within the SILVA database[44].

Another and more fundamental problem is the diversity missing from the databases. The global bacterial species richness on earth is estimated[66] to be as high as 10^{12}. This is in stark contrast to the fact that the total number of non-redundant high quality SSU rRNA sequences in the SILVA database is merely close to 700 000 (SILVA SSU Ref NR 128). Missing reference sequences can result in poor classification or completely missing classification, which can have a large impact on the observed community structure and dynamics[46]. For less studied ecosystems, such as soil and plants, the effect is more pronounced, compared to extensively studied ecosystems, such as the human gut, where many of the SSU rRNA sequences in the reference databases originate from[45].

Historically, the vast majority of SSU rRNA reference sequences in the database are obtained by PCR and Sanger sequencing, and are therefore subject to primer bias, meaning some taxonomic groups are underrepresented or completely missing as indicated by metagenomic studies[67]. The bacterial reference databases are severely underpopulated, and it is even worse for the archaea and eukaryote databases, as no good full-length primers exist for these domains[58].

A major task is to populate the SSU rRNA databases with reference sequences to cover the missing diversity. The sequences need to be high quality, near full-length and free of primer bias. The conventional Sanger sequencing and cloning approach depends on primers and scales poorly. Some PCR-free metagenome approaches exist (e.g. EMIRGE[68]), but these have problems with strain diversity and their performance have been questioned[45]. Therefore, we developed a new method for generating reference sequences (Karst *et al*. 2016, see Chapter 4). The method combines reverse transcription of full-length SSU rRNA[69] and synthetic long read sequencing using molecular tagging[70]. This method is free of conventional primer bias, as the SSU rRNA sequences are generated directly from SSU rRNA. The molecular tagging approach allows to recreate full-length cDNA sequences from the short reads, which is not possible with commercially available short-read library protocols[56]. With our new method we were able to generate more than 30 000 full-length, high quality SSU rRNA reference sequences (>1200 bp, 0.17% error rate, 0.19% chimera rate), from

all domains of life, based on samples from five different environments (activated sludge, anaerobic digestion, fresh water, human gut and soil).

The method is scalable, and now makes it possible for a single researcher to generate 500 000 SSU rRNA sequences in less than a week using a single run on an Illumina HiSeq 4000 system – this is equivalent to 25 % of the high-quality reference sequences ever submitted to the SILVA database. With this technique, we predict millions of SSU rRNA sequence will be generated in the coming years, which will allow us to conduct microbial community analysis more effectively. It will be possible to make specialized reference databases for ecosystems and use these to design more effective primers and FISH probes. Databases with broad coverage of the microbial diversity will also make it feasible to determination community structure and relative abundance based on phylotype methods instead of OTU based methods, which is computational easier[65] and avoids artifacts from OTU clustering[71]. Furthermore, the new method will increase our knowledge of the tree of life and evolution in general. It will especially increase our insight into the Archaea and Eukaryotic SSU rRNA phylogeny and diversity, which has been missed due to lack of universal primer sets to generate full-length SSU rRNA sequences[58,72,73].

1.2. GENOME-CENTRIC METAGENOMICS: FROM IDENTIFICATION TO FUNCTIONAL POTENTIAL

Metagenomics was first introduced in 1998 by Handelsman *et al*.[74], who defined it as *"functional analysis of the collective genomes of soil microflora, which we term the metagenome of the soil"*. However, as with 16S rRNA amplicon sequencing, the past five years, the application of metagenomics has seen a dramatic increase due to the exponential decrease in cost of DNA sequencing. This has, for example, enabled establishment of comprehensive gene catalogues from the human gut[75] and numerous studies of species-function relationships that shape the ecological properties of the human microbiome[76]. However, a perquisite for functional insights based on metagenomic data, is the existence of good reference databases with genomes closely related to the microbes in the specific ecosystem. Unfortunately, the genome databases are far from complete[77] except in the case of the human microbiome, where large projects have been undertaken to systematically improve the databases[78].

1.2.1. GENOME BINNING FROM METAGENOMES

The conventional approach to obtain genomes has been through culturing and isolation, but as the majority of species are difficult to culture, other approaches have been explored. As metagenomics involves sequencing of the complete genetic information of a given sample, it should, in theory, allow reconstruction of the genomes of the individual community members. In this ideal scenario, it would be possible to generate the complete metabolic models for these organisms from the

genetic material, and based on these models, predict activity and roles of the organisms, which can be validated by *in situ* methods.

Presently, reconstruction of individual genomes from metagenomes requires binning of the individual scaffolds from metagenomic assemblies into separate groups or "bins"[79], which constitute genomes individual community members. Historically, sequence composition based methods[80–82] have been used (e.g. tetra nucleotide frequencies), but since 2013 methods based on differential coverage approaches have become the standard within the field, although often supplemented by composition based approaches[83–85].

In Chapter 5 we describe our tool "mmgenome" that can be used to extract genomes from metagenomes. The tool extends on the framework defined in original multi-metagenome approach from Albertsen *et al*., 2013. Contrary to many other recent binning tools, mmgenome is not an automated one-stop tool, but rather an interactive-visualization suite, that can be used to integrate meta-information from a large number of different binning algorithms and other analysis. In this way, multiple sources of evidence can be used in a supervised binning approach. Visualizing the metagenome data and binning process, allows the user to better detect data artifacts, and makes the whole binning process more transparent and reproducible.

1.2.2. FUNDEMENTAL PROBLEMS IN GENOME BINNING

In practice, it can be very difficult to retrieve complete genomes from metagenomes as the metagenome *de novo* assembly often is very fragmented in complex samples. The main culprit is genetically similar regions in the population[86,87]. These similarities can be within a genome, in the form of genes with multiple copy numbers, e.g. rRNA or transposons that cannot be readily bridged by the short read lengths. The similarities can also be between genomes, in the form of conserved genes or micro-diversity (presence of almost sequence-identical strains). To illustrate this, we simulated the impact of micro-diversity on *de novo* assembly of a metagenome containing two very similar strains, with sequence identity between the two, varying from 94-100% (see Figure 4). At 94% the two strains assembled separately into complete genomes, at 100%, the two strains assembled into one genome. In between we observed varying degrees of fragmentation with up to thousands of small fragments in one of the assemblies. If this is extrapolated to a complex microbial community sample with hundreds of species, and many of them are represented by strain diversity, it becomes evident that it can be complicated to recreate genomes. This also means that it is often impossible to recreate even the dominant species in the metagenomes[88], and without visualizing the data using tools such as, mmgenome, this will often be missed.

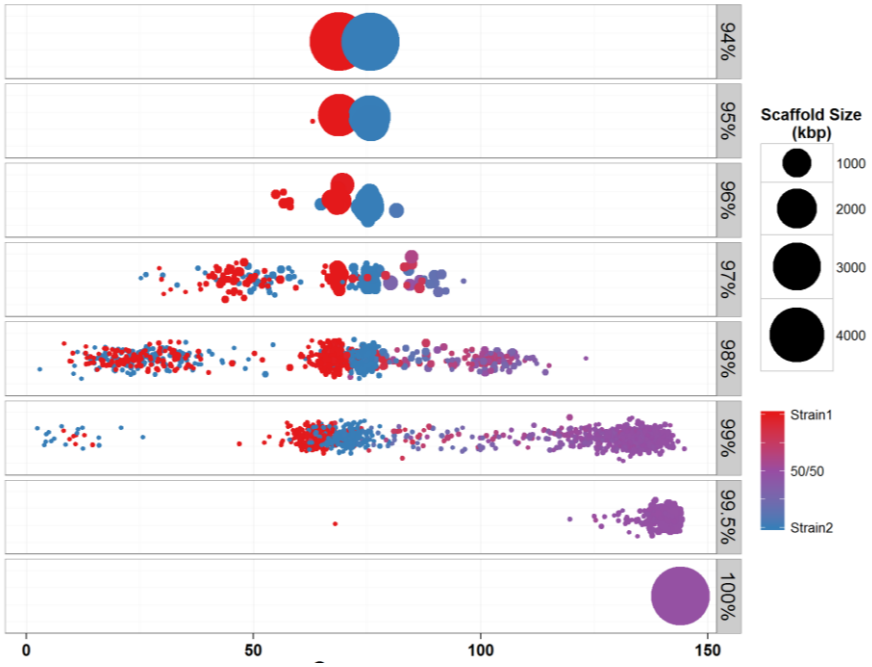

Figure 4: Simulation of the impact of metagenome micro-diversity on de novo assembly. x-axis shows scaffold read coverage and y-axis shows different levels of sequence identity between the two strains. The circles represent scaffolds, and the circle size represents scaffold size. The color illustrates which strain mapped reads originate from.

1.2.3. ESTIMATING COMPLETENESS IN METAGENOME DERIVED BINS

Before using the bins to determine metabolic potential of a given organism, it is important to estimate the completeness of the genome in the bin. A prevalent strategy is to detect the number of lineage-specific essential single copy genes, as implemented by the software CheckM[89]. The rationale for the approach is, that if all the lineage-specific essential genes are represented, the genome is probably complete. In addition, if the genes are only there in single copies, the bin is probably not contaminated by other organisms. However, the method should be used with care. As shown above, co-assembly of strains will occur in metagenome datasets, and there will be a higher risk that core-genes co-assemble, as these are often highly conserved between strains. Hence, it can be difficult to properly detect completeness and contamination in metagenome derived population genome bins and the numbers should always be supplemented by visual inspection of the binning process, where it is relatively easy to visually identify potential strain related problems.

1.2.4. APPLYING GENOME-CENTRIC METAGENOMICS TO RETRIEVE FUNCTIONAL IMPORTANT BACTERIA IN ACTIVATED SLUDGE

When used carefully metagenomics can be a very powerful tool and can, in some circumstances, enable recovery of circular genomes from complex communities. In the paper "Genomic and in situ investigations of the novel uncultured Chloroflexi associated with 0092 morphotype filamentous bulking in activated sludge" (Chapter 6), we extracted the genome from the Chloroflexi genus B45 using the mmgenome workflow. We had previously shown, through the MiDAS survey, that the bacteria was abundant and possible involved in bulking in Danish wastewater treatment plants. From the genome a metabolic model was created, which was used as basis for MAR-FISH studies to determine the *in situ* activities of the genera.

1.3. CONCLUSION

- The 16S rRNA amplicon sequencing approach was optimized for study of bacterial communities in activated sludge. More rigorous physical lysis was implemented and PCR primers targeting the V1-3 region of the SSU rRNA, were found to best target key bacteria in activated sludge.
- The 16S rRNA amplicon sequencing was used in a wide range of studies on activated sludge from wastewater treatments plants, and used as standard protocols in the MiDAS fieldguide initiative.
- A new method for generating full-length, high quality SSU rRNA sequences was developed, by combining reverse transcription of SSU rRNA and synthetic long-read sequencing by molecular tagging. In a single study, we were able to generate 30 000 full-length, high quality SSU rRNA sequences from all domains of life, coming from 5 different ecosystems. The mean error rate was 0.17% and the chimera rate was 0.19%. A major fraction of the generated SSU rRNA sequences represented novel diversity, especially for Eukaryotes, where 63% of the sequences were less than 97% identical to sequences in the reference databases. We predict the method will change the way microbial community analysis studies are conducted, with more focus on ecosystem specific reference databases.
- The tool ´mmgenome´ was developed for manual genome binning from metagenomes. The tool has emphasis on interactive data visualization, transparency and reproducibility and integration with data from other types of analysis. ´mmgenome´ is open source and has been used by many researchers around the world.
- The ´mmgenome´ tool was used to extract and recreate a complete circular genome from the important activated sludge bacterial phylotype B45 from the phylum Chloroflexi. The genome was used to generate a metabolic model for the organism of which key features were validated through *in situ* studies.

1.4. PERSPECTIVES

The present PhD work has taken advantage of the exponential decrease in sequencing cost of high-throughput short-read sequencing using the Illumina platform. However, now we stand on the verge of a new paradigm shift; from the high-throughput short-read era to the high-throughput long-read era. The advances in long-read DNA sequencing technologies will fundamentally change the microbial community studies. There are currently true long-read technologies such SMRT sequencing (Pacific Biosciences) and nanopore sequencing (Oxford Nanopore) as well as the synthetic long-read technologies (Illumina, 10X Genomics and molecular tagging), which can produce read lengths from 10-100 kbp, at a throughput that makes them available for metagenome sequencing. This is a dramatic increase compared short-read technologies with read lengths of < 1 kbp. Longer reads radically simplifies *de novo* assembly of sequence data[90], and long-read technologies are today routinely used to sequence and assemble genomes from pure cultures[91,92], and the first examples of the techniques being applied to low complexity microbial communities have been published[12].

In the present PhD work we leveraged synthetic long reads to drastically improve the existing rRNA databases (see Chapter 4) and enable comprehensive ecosystem specific rRNA databases. The increase in high-throughput long-read area will enable ecosystem specific genome databases within the next five years. This will radically change our ability to study microbial ecosystems. However, the genomes are only the starting point and the challenge will be to develop high-throughput methods to understand the function and interaction of the bacteria in the environment.

LITERATURE

1. Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C. & D'Hondt, S. From the Cover: Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc. Natl. Acad. Sci.* **109,** 16213–16216 (2012).

2. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95,** 6578–6583 (1998).

3. Gruber, N. & Galloway, J. N. An Earth-system perspective of the global nitrogen cycle. *Nature* **451,** 293–6 (2008).

4. Knoll, A. H., Canfield, D. E. & Konhauser, K. O. Fundamentals of Geobiology. *Fundam. Geobiol.* 36–48 (2012). doi:10.1002/9781118280874

5. Corliss, J. O. The Leeuwenhoek Legucy. *J. Protozool.* **39,** 649–649 (1992).

6. Blaser, M., Bork, P., Fraser, C., Knight, R. & Wang, J. The microbiome explored: recent insights and future challenges. *Nat. Rev. Microbiol.* **11,** 213–7 (2013).

7. Nielsen, P. H., Saunders, A. M., Hansen, A. A., Larsen, P. & Nielsen, J. L. Microbial communities involved in enhanced biological phosphorus removal from wastewater—a model system in environmental biotechnology. *Curr. Opin. Biotechnol.* **23,** 452–459 (2012).

8. Vanwonterghem, I., Jensen, P. D., Ho, D. P., Batstone, D. J. & Tyson, G. W. Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Curr. Opin. Biotechnol.* **27,** 55–64 (2014).

9. Chaparro, J. M., Sheflin, A. M., Manter, D. K. & Vivanco, J. M. Manipulating the soil microbiome to increase soil health and plant fertility. *Biol. Fertil. Soils* **48,** 489–499 (2012).

10. Gilbert, J. A. & Neufeld, J. D. Life in a World without Microbes. *PLoS Biol.* **12,** 1–3 (2014).

11. Winogradsky, S. Contributions a la morphologie des organismes de la nitrification. *Arch. Sci. Biol.* **1,** 87–137 (1892).

12. Daims, H. *et al.* Complete nitrification by Nitrospira bacteria. *Nature* **528,**

504–509 (2015).

13. van Kessel, M. A. H. J. *et al.* Complete nitrification by a single microorganism. *Nature* **528,** 555–559 (2015).

14. Sommer, M. O. A. & Dantas, G. Antibiotics and the resistant microbiome. *Curr. Opin. Microbiol.* **14,** 556–563 (2011).

15. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74,** 417–433 (2010).

16. Kembel, S. W. *et al.* Architectural design influences the diversity and structure of the built environment microbiome. *ISME J.* **6,** 1469–79 (2012).

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

18. Flemming, H.-C. *et al.* Biofilms: an emergent form of bacterial life. *Nat. Rev. Microbiol.* **14,** 563–575 (2016).

19. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10,** 538–550 (2012).

20. Zhou, J. *et al.* High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *MBio* **6,** 33–36 (2015).

21. Madigan, M. T. *Brock Biology of Microorganisms*. (Pearson/Benjamin Cummings, 2009).

22. Stewart, E. J. Growing unculturable bacteria. *J. Bacteriol.* **194,** 4151–4160 (2012).

23. Lagier, J. C. *et al.* The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. *Clin. Microbiol. Rev.* **28,** 237–264 (2015).

24. Pace, N. R., Sapp, J. & Goldenfeld, N. Classic Perspective: Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proc. Natl. Acad. Sci.* **109,** 1011–1018 (2012).

25. Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation . Phylogenetic Identification and In Situ Detection of Individual Microbial

Cells without Cultivation. *Microbiol. Rev.* **59,** 143–169 (1995).

26. Welch, D. B. M. & Huse, S. M. Microbial Diversity in the Deep Sea and the Underexplored 'Rare Biosphere'. *Handb. Mol. Microb. Ecol. II Metagenomics Differ. Habitats* 243–252 (2011). doi:10.1002/9781118010549.ch24

27. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl,** 4516–22 (2011).

28. Voorhees, R. M., Weixlbaumer, A., Loakes, D., Kelley, A. C. & Ramakrishnan, V. Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70S ribosome. *Nat. Struct. Mol. Biol.* **16,** 528–533 (2009).

29. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. 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 (2005).

30. Amann, R. I., Krumholz, L. & Stahl, D. A. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *J. Bacteriol.* **172,** 762–770 (1990).

31. Lee, N. *et al.* Combination of fluorescent in situ hybridization and microautoradiography-a new tool for structure-function analyses in microbial ecology. *Appl. Environ. Microbiol.* **65,** 1289–1297 (1999).

32. Nielsen, P. H., Kragelund, C., Seviour, R. J. & Nielsen, J. L. Identity and ecophysiology of filamentous bacteria in activated sludge. *FEMS Microbiol. Rev.* **33,** 969–98 (2009).

33. Brehm-stecher, B. F. & Johnson, E. A. Single-Cell Microbiology: Tools, Technologies, and Applications Single-. *Microbiol. Mol. Biol. Rev.* **68,** 538–559 (2004).

34. Huang, W. E. *et al.* Raman-FISH: Combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function. *Environ. Microbiol.* **9,** 1878–1889 (2007).

35. Liu, P. *et al.* Microfluidic fluorescence in situ hybridization and flow cytometry (μFlowFISH). *Lab Chip* **11,** 2673–9 (2011).

36. Chrimes, A. F., Khoshmanesh, K., Stoddart, P. R., Mitchell, A. & Kalantar-zadeh, K. Microfluidics and Raman microscopy: current applications and future challenges. *Chem. Soc. Rev.* **42,** 5880–5906 (2013).

37. Zhang, W., Li, F. & Nie, L. Integrating multiple 'omics' analysis for microbial biology: Application and methodologies. *Microbiology* **156,** 287–301 (2010).

38. Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329,** 533–8 (2010).

39. McIlroy, S. J. *et al.* 'Candidatus Competibacter'-lineage genomes retrieved from metagenomes reveal functional metabolic diversity. *ISME J.* **8,** 613–24 (2014).

40. Herbst, F. A. *et al.* Enhancing metaproteomics-The value of models and defined environmental microbial systems. *Proteomics* **16,** 783–798 (2016).

41. Hettich, R. L., Pan, C., Chourey, K. & Giannone, R. J. Metaproteomics : Harnessing the Power of High Performance Mass Spectrometry to Identify the Suite of Proteins That Control Metabolic Activities in Microbial Communities. *Anal. Chem.* **85,** 4203–4214 (2013).

42. Tang, J. Microbial metabolomics. *Curr. Genomics* **12,** 391–403 (2011).

43. van Loosdrecht, M. C., Nielsen, P. H., Lopez-Vazquez, C. M., & Brdjanovic, D. *Experimental Methods in Wastewater Treatment | IWA Publishing*. *Water Intelligence Online* (2016).

44. McIlroy, S. J. *et al.* MiDAS: The field guide to the microbes of activated sludge. *Database* **2015,** 1–8 (2015).

45. Schloss, P. D., Girard, R. A., Martin, T., Edwards, J. & Thrash, J. C. Status of the Archaeal and Bacterial Census: an Update. *MBio* **7,** e00201-16 (2016).

46. Werner, J. J. *et al.* Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J.* **6,** 94–103 (2012).

47. Frostegård, Å. *et al.* Quantification of bias related to the extraction of DNA directly from soils. *Appl. Environ. Microbiol.* **65,** 5409–5420 (1999).

48. Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7,** e33865 (2012).

49. Smith, B., Li, N., Andersen, A. S., Slotved, H. C. & Krogfelt, K. A. Optimising bacterial DNA extraction from faecal samples: comparison of three methods. *Open Microbiol. J.* **5,** 14–7 (2011).

50. Kennedy, N. A. *et al.* The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* **9,** 1–9 (2014).

51. Vanysacker, L. *et al.* Bacterial community analysis of activated sludge: an evaluation of four commonly used DNA extraction methods. *Appl. Microbiol. Biotechnol.* **88,** 299–307 (2010).

52. Guo, F. & Zhang, T. Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Appl. Microbiol. Biotechnol.* (2012). doi:10.1007/s00253-012-4244-4

53. Debelius, J. *et al.* Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biol.* **17,** 217 (2016).

54. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590-6 (2013).

55. Wang, Y. & Qian, P. Y. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS One* **4,** (2009).

56. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17,** 333–351 (2016).

57. Ward, D. V. *et al.* Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. *PLoS One* **7,** e39315 (2012).

58. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41,** 1–11 (2013).

59. Cai, L., Ye, L., Tong, A. H. Y., Lok, S. & Zhang, T. Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets. *PLoS One* **8,** 1–11 (2013).

60. Tremblay, J. *et al.* Primer and platform effects on 16S rRNA tag sequencing. *Front. Microbiol.* **6,** 1–15 (2015).

61. Mielczarek, A. T., Nguyen, H. T. T., Nielsen, J. L. & Nielsen, P. H. Population dynamics of bacteria involved in enhanced biological phosphorus removal in Danish wastewater treatment plants. *Water Res.* **47,** 1529–1544 (2013).

62. Cole, J. R. *et al.* Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42,** 633–642 (2014).

63. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).

64. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–7 (2007).

65. Schloss, P. D. & Westcott, S. L. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* **77,** 3219–3226 (2011).

66. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci.* **113,** 5970–5975 (2016).

67. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523,** 208–211 (2015).

68. Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W. & Banfield, J. F. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol.* **12,** R44 (2011).

69. Botero, L. M. *et al.* Poly(A) polymerase modification and reverse transcriptase PCR amplification of environmental RNA. *Appl. Environ. Microbiol.* **71,** 1267–75 (2005).

70. Burke, C. & Darling, A. E. Resolving microbial microdiversity with high accuracy full length 16S rRNA Illumina sequencing. *bioRxiv* 10967 (2014). doi:10.1101/010967

71. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* **12,** 1889–1898 (2010).

72. Epstein, S. & López-García, P. 'Missing' protists: A molecular prospective. *Biodivers. Conserv.* **17,** 261–276 (2008).

73. Tsagkogeorga, G. *et al.* An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models. *Bmc Evol. Biol.* **9,** (2009).

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

75. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).

76. Vieira-Silva, S. *et al.* Species–function relationships shape ecological properties of the human gut microbiome. *Nat. Microbiol.* **124,** 16088 (2016).

77. Albertsen, M., Hansen, L. B. S., Saunders, A. M., Nielsen, P. H. & Nielsen, K. L. A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. *ISME J.* 1094–1106 (2011). doi:10.1038/ismej.2011.176

78. Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328,** 994–9 (2010).

79. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* **72,** 557–78, Table of Contents (2008).

80. Strous, M. *et al.* Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440,** 790–794 (2006).

81. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37–43 (2004).

82. Wrighton, K. C. *et al.* Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science (80-.).* **337,** 1661–1665 (2012).

83. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11,** 1144–1146 (2014).

84. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* (2013). doi:10.1038/nbt.2579

85. Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts

in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* **23,** 111–120 (2013).

86. Temperton, B. & Giovannoni, S. J. Metagenomics: microbial diversity through a scratched lens. *Curr. Opin. Microbiol.* 1–8 (2012). doi:10.1016/j.mib.2012.07.001

87. Turaev, D. & Rattei, T. High definition for systems biology of microbial communities: Metagenomics gets genome-centric and strain-resolved. *Curr. Opin. Biotechnol.* **39,** 174–181 (2016).

88. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331,** 463–7 (2011).

89. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–55 (2015).

90. Miller, J. R., Koren, S. & Sutton, G. Genomics Assembly algorithms for next-generation sequencing data. *Genomics* **95,** 315–327 (2010).

91. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12,** 733–735 (2015).

92. Remus-Emsermann, M. N. P. *et al.* Complete genome sequence of Pseudomonas citronellolis P3B5, a candidate for microbial phyllo-remediation of hydrocarbon-contaminated sites. *Stand. Genomic Sci.* **11,** 1–12 (2016).

CHAPTER 2. BACK TO BASICS - THE INFLUENCE OF DNA EXTRACTION AND PRIMER CHOICE ON PHYLOGENETIC ANALYSIS OF ACTIVATED SLUDGE COMMUNITIES

CHAPTER 3. EXPERIMENTAL METHODS IN WASTEWATER TREATMENT

CHAPTER 4. THOUSANDS OF PRIMER-FREE, HIGH-QUALITY, FULL-LENGTH SSU RRNA SEQUENCES FROM ALL DOMAINS OF LIFE

CHAPTER 5. MMGENOME: A TOOLBOX FOR REPRODUCIBLE GENOME EXTRACTION FROM METAGENOMES

CHAPTER 6. GENOMIC AND IN SITU INVESTIGATIONS OF THE NOVEL UNCULTURED CHLOROFLEXI ASSOCIATED WITH 0092 MORPHOTYPE FILAMENTOUS BULKING IN ACTIVATED SLUDGE

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-830-7

AALBORG UNIVERSITY PRESS