View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by NORA - Norwegian Open Research Archives

Analysis of sequencing data in environmental genomics

Exploring the diversity of the microbial biosphere

Anders Lanzén

Dissertation for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

2013

Scientific environment

The work presented in this thesis was carried out at the Faculty of Mathematics and Natural Sciences of the University of Bergen (UiB), at the Department of Biology (Marine Microbiology Group) and Centre for Geomicrobiology, as well as the Computational Biology Unit (Jonassen Group) at the UniComputing department of Uni Research; a non-profit research company affiliated with UiB. The project was funded through a PhD grant from the University of Bergen and additional funding for sequencing, laboratory, field and travel expenses was provided by a scholarship from L. Meltzers Høyskolefond. My contributions to the work outlined in papers III and IV (AmpliconNoise) were possible through a long-standing close collaboration between the group of Professor Lise Øvreås in UiB, and Dr. Christopher Quince and Prof. William Sloan at the University of Glasgow. The work outlined in paper V was possible through a research collaboration with the University of Addis Ababa, funded through NUFU (the Norwegian Agency for Development Cooperation).

~ So, really, what you are doing is like trying to understand the ecology of a forest from looking at pulp that you made from a few of its trees, after you chopped them down and processed them. That can't be easy.

Amateur scientist and fellow traveller to Perhentian Kecil

Acknowledgements

There is a whole legion of people to whom I am grateful and without them, this thesis would not be what it is. Lise Øvreås, you have been a fantastic main supervisor. Of course, there were times that I did not see that so clearly, like when being forced to do actual wet labbing with almost no experience (at least for eight years). Then, you decided to leave the continent when I needed you the most to write this thesis. But, "no pain, no gain": without this, I would have gone nine years without touching a pipette and I would not have visited Berkeley. I would also like to thank you and your family for opening your Californian house to me and to Agur.

I was also fortunate enough to have three great co-supervisors, Inge Jonassen, Tim Urich and Pål Puntervoll. Inge, I owe a lot to you, mainly that I managed to keep one foot in bioinformatics. Your supervision and expertise has really provided a unique complement and you have always showed great interest and patience in applying this to exotic problems of microbiology, that I may not always have understood or explained very well. Tim, thank you for the third dimension of supervision, friendship, and incredible patience, and taking time to analyse and discuss the most tiny but important issues in great detail. Also, it is thanks to you that I opened my eyes to microbial ecology, in the first place. Pål, I am grateful for introducing me to a world of new concepts, languages and tools during my time in the Bioinformatics Service Group. This experience was essential for my PhD (incidentally the thesis was written in LyX, for example).

Gratitude also goes to all my colleagues in the Marine Microbiology group and others at the Department of Biology, the Centre for Geobiology and Uni Computing. You have provided a very rich working environment, with diverse knowledge in everything from supercomputers to deep sea vents and microbial metabolism. Thanks especially to Mia Bengtsson and Steffen Jørgensen, for our collaborations and for countless, endless, discussions. Science would have lost some of its magic without either of you. Mia, like you wrote in my copy of your thesis: "tack som fan! <3". And Steffen, I was going to make a joke about a Christmas party but I have to save something for a speech. Special thanks also to Antonio García-Moyano for your expertise and support, to Dominika

Chmolowska for being a dedicated and knowledgable co-worker in the lab, and to all my other co-workers and co-authors: Svenn Helge Grindhaug, Susanne Balzer, Antonio Pagarete, Vigdis Torsvik, Hallgerd Eydal, Addis Simachew, Ingrid Mørkeseth, Baye Sitotaw, Amare Gessesse, Yemisirach Mulugeta, Runar Stokke, Håkon Dahle, Ida Steen, Irene Roalkvam, Christa Schleper, Ramiro Logares, Eva Lindström, Nathalie Reuter, Kjell Petersen, Kidane Tekla, Pawel Stormwasser, Siv Midtun Hollup and all the members in Inge Jonassen's group (especially Animesh and Matus for laughs, support and philosophical insights). Thanks Torbjørn Lium and Særdar Halifu for fantastic 24-7 tech and HPC support (and crazy out-of-work adventures).

Thanks to this thesis project and my supervisors, I had the privilege to visit, work with and get to know some exceptional scientists in Glasgow and Newcastle. I am especially grateful to Chris Quince, Bill Sloan and Tom Curtis for our collaborations and all I have learnt from you. In addition to being a great friend, Chris has arguably acted as an extra, unofficial supervisor. A not-insignificant portion of our work was carried out in various pubs around the world, making it yet more enjoyable.

Another special thanks to all past and present members of the "international lunch table" for fantastic company at work and after: Øystein, Eric, Paolo, Jim, Anne-Laure, Paco, Cecile, Nico, Mari, David, Sofia, Sam, Sara, Cindy, Ana, Fabian, Bea, Becky, Laurent, Mahaut, Valentina. There are so many that I cannot list you all, but I have not forgotten. Your everyday support and friendship has been extremely important, and helped to carry me through (without doubt). So did my Swedish friends, helping me relax and gain perspective during my Stockholm visits and always interested in what it really was I was really working with ("cod DNA?").

Everyone in my family, back in Sweden: You have also meant a lot for this thesis becoming reality, supporting me and showing interest in my work. Thanks to my parents, for taking care of me in Sweden and for telling me to relax when I needed to hear it. And to my beloved grandmother Hillevi, no longer with us, for wise words.

Finally Agur, thank you for everything, for your constant and heartfelt support, and an incredible patience. Also thanks for proof-reading of this thesis, for support with rehearsals of presentations, mathematical problems and R. But, most importantly, thanks for making the last three years the best ones imaginable (actually much better). Although a tough measure, moving from Bergen in advance also provided a final push to finish up quickly, in order to rejoin you in the Basque Country.

Contents

Summary

Most life on this planet is microbial and for the last two decades, environmental genomics has contributed to reveal an impressive biodiversity of this microbial life. This approach applies DNA sequencing to environmental samples, with the significant advantage of not relying on cell cultures, since only a minority of microorganisms are easily cultured in the laboratory. This thesis deals primarily with analysis of microbial diversity based on community profiling. This variant of environmental genomics targets defined marker genes to study the structure of microbial communities. The use of the small subunit ribosomal RNA as a phylogenetic marker is discussed and evaluated, with emphasis on taxonomic classification, estimation of diversity and comparison of community structure between samples. Thanks to improved sequencing technologies, community profiling is an increasingly powerful and cost-efficient technique. Like all methodologies it has limitations and sources of random- and systematic errors, many of which remain poorly understood. In relation to this, a number of recommendations and novel analysis methods are developed and provided. These are subsequently applied to study environmental communities, targeting issues like the "rare biosphere" concept, and variation of community structure across space and environmental gradients.

Taxonomic classification is the process of placing environmental sequences in context of previously studied organisms. Thus, ecologically meaningful information such as putative metabolic functions can be derived. In **Paper I**, a set of resources for taxonomic classification is provided and evaluated. The performance of the resulting framework, CREST (Classification Resources for Environmental Sequence Tags), is shown to compare favourably to existing methods. It also provides a manually curated taxonomy and functionality for comparing composition across datasets. In **Paper II**, a hydrothermal vent-associated microbial mat community is studied, using a set of different environmental genomics methods. Based on this study, several important sources of bias and reproducibility of community profiling are evaluated and discussed. The results highlight the importance of applying complementary methods. They also illustrate the influence of primer choice, PCR bias and whether RNA or DNA is targeted. Random variation, or noise, is another important factor to consider in community profiling

studies. **Papers III** and **IV**, examines the effect of such noise from PCR amplification and pyrosequencing. Currently, this is the most common sequencing method applied to environmental samples. The results of **Paper III** demonstrate that early community profiling studies using pyrosequencing have significantly overestimated the extent of biodiversity, because of noise. To compensate for such noise in amplicon sequence datasets, the program AmpliconNoise was developed. Using "mock communities", a mix of clones with known sequences, the performance of AmpliconNoise is demonstrated and compared to alternative methods. Analyses of diversity in the microbial mat community studied in **Paper II** utilise AmpliconNoise. Resulting estimates are compared to previous findings, from similar environments.

In addition to biodiversity *per se*, the underlying diversity structures of communities and the mechanisms shaping them, remain important but poorly understood issues in microbial ecology. Because of their many useful characteristics, alkaline soda lakes are used as model ecosystem to study several such issues, in **Paper V**. Results reveal that these extreme environments harbour surprisingly high microbial diversity. Interestingly, the most alkaline and saline lakes studied also appear to be the most diverse. Further, it is shown that pH, oxygen level, and sodium- and potassium concentrations can explain 30% of the compositional variance between the lakes studied. The existence of organisms endemic to individual lakes is also indicated. Although soda lakes are relatively uncommon environments, this study provides an example of how fundamental biogeographical questions can be targeted using a careful choice of experimental design and analysis methodology. The results call into question several established notions such as extreme environments generally being less diverse and that few prokaryotic organisms are endemic. Hopefully the findings will inspire future studies, exploring these relationships further.

In summary, the work presented here illustrates the importance of evaluating and optimising the methodology used in environmental genomics, particularly for amplicon sequencing, taxonomic classification, and estimation of phylogenetic diversity. It is likely that methodological limitations have biassed and slowed down data analysis and interpretation of important ecological issues like the rare biosphere and microbial biogeography.

List of publications

Paper I

Lanzén A, Jørgensen SL, Huson DH, Gorfer M, Grindhaug SH, Jonassen I, Øvreås L & Urich T (2012) CREST - Classification Resources for Environmental Sequence Tags, *PLoS ONE* 7: e49334.

Paper II

Lanzén A, Jørgensen SL, Bengtsson MM, Jonassen I, Øvreås L & Urich T (2011) Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA. *FEMS Microbiology Ecology*. 77: 577-589.

Paper III

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

Paper IV

Quince C, Lanzén A, Davenport RJ & Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12: 38.

Paper V

Lanzén A, Simachew A, Gessesse A, Jonassen I & Øvreås L (2012) Surprising prokaryotic and eukaryotic diversity, community structure and biogeography in Ethiopian alkaline lakes. *Submitted to Environmental Microbiology.*

Nomenclature

BLAST	Basic Local Alignment Search Tool
cDNA	Complementary DNA - derived using reverse transcription from RNA
DAPI	4',6-diamidino-2-phenylindole
DNA	Deoxyribonucleic Acid - utilised by living organisms as the primary carrier of information, or the 'genotype'
env. genomics	the application of high-throughput sequencing to nucleic acid samples extracted directly from the environment.
GPL	GNU General Public Licence - a license for freely available source material, such as software
H'	Shannon index (also known as Shannon's diversity index, Shannon-Wiener index or Shannon entropy)
LCA	Lowest Common Ancestor
LSU	the Large Subunit of rRNA
mRNA	messanger RNA - transcripts from protein-coding genes
n-mer	a nucleotide sub-sequence (word) of length n
NCM	Neutral Community Model
OTU	Operational Taxonomic Unit - pragmatic definition used as proxy for species
PCR	Polymerase Chain Reaction
Prokaryote	an organism belonging to the domains Archaea or Bacteria, sharing several methodologically relevant negative characteristics
RDP	the Ribosomal Database Project
Read	A string representing a single sequencing readout from a nucleotide sequence

RNA	Ribonucleic acid
rRNA	ribosomal RNA - here referring either to sequences derived from the ribosome itself or its encoding gene
RT-qPCR	Real-time quantitative PCR
S	Svedberg - a non-SI unit for sedimentation rate
SMRT	Single Molecule Real Time sequencing - a third generation sequencing technology developed by Pacific Biosciences
SSU	the Small Subunit of ribosomal RNA (also known as 16S in prokaryotes and 18S in eukaryotes) - here referring either to the subunit of the ribosome itself or its encoding gene
TAD	Taxon Abundance Distribution - the distribution of relative abundances of taxa in a sequence dataset or biological community

Part I

Synthesis

Chapter 1

Introduction

Microorganisms (*Bacteria*, *Archaea* and unicellular eukaryotes) dominate life on our planet, as well as global biomass and carbon turnover (Whitman et al., 1998). Because of their dominance of global biogeochemical cycles, microorganisms are essential to life and the functioning of the biosphere (Falkowski et al., 2008). Their metabolic versatility and range of habitats is also impressive. Compared to the stricter requirements of larger organisms, microorganisms grow across wide ranges of temperature, pH and salinity, and new findings have repeatedly pushed our perception of the limits of microbial life (Pikuta et al., 2007). A better understanding of microbial ecology may be essential for applications like modelling of large-scale ecological processes, but also for human health and biotechnology. For every human cell of our body, we carry around about ten cells of bacterial symbionts, vital for our well-being (Berg, 1996). Known as the *human microbiome*, this microbial community may even play an important role in determining our mood and mental health (Kinross et al., 2011).

However, the ecological importance of microorganisms has historically been overlooked. Today, this notion is replaced with a growing appreciation for their paramount importance and biodiversity. The rapid scientific progress leading to this paradigm shift would not have been possible if not for molecular techniques like DNA sequencing. These techniques circumvent the need for studying microorganisms through direct observation or culturing. This is particularly important since the majority of organisms cannot easily be cultured. Those that can may be rare and opportunistic organisms, not representing the ecologically important part of the community studied (Handelsman, 2004). Thus, it is only recently that the scale of microbial diversity is starting to be appreciated. Largely responsible is *environmental genomics*, i.e. the application of high-throughput sequencing to nucleic acid samples extracted directly from the environment. It has become an invaluable tool for studying microbial ecology. Like other genomic

techniques, it is quickly gaining ground thanks to the rapid technological development of DNA sequencing, with dropping prices and exponentially increasing throughput.

Environmental genomics has revolutionised the field of microbial ecology, but in many respects it is immature. Limitations and pitfalls remain poorly understood and new techniques develop so fast that benchmarking studies, methodology and recommendations for best practices, lag behind. The situation has even been likened to a "Red Queen" coevolutionary arms race, where "it takes all the running you can do, to keep in the same place" (Carrol, 1872), by Caporaso et al. (2011). In recent years, influential ecological concepts like the *rare biosphere* have been proposed (Sogin et al., 2006), gained popularity, then as quickly been brought into question as a mere consequence of sequencing bias and inappropriate analysis methods (Reeder and Knight 2009). The rare biosphere concept suggests that rare organisms tend to dominate the diversity of natural communities, while they only constitute a small part of the total biomass. Its implications are largely unknown. Either these rare organisms may have important metabolic functions, or be inactive, acting as a "seed bank" (reviewed in Pedrós-Alió, 2012). This is challenging to determine without knowing the structure of the rare biosphere, or indeed that it exists. Another example is the deceptively simple task of estimating the total number of microbial "species". To simply increase our sequencing efforts would be insufficient, until we have gained better knowledge of microbial biogeography (Curtis et al., 2006). However, more strategic use of sequencing combined with increased throughput may contribute to such knowledge.

This cyclic scientific progress is common and illustrates how conceptual understanding is linked to methodology (Kuhn, 1962). Thus, to improve environmental genomics, the assumptions of its methodology must be continuously re-evaluated. It can then be improved, and used to answer essential questions in microbial ecology. This requires interdisciplinary efforts, combining biology, informatics and mathematics. Instead, bioinformatics is sometimes seen as a "magic wand" or "black box" by microbiologists, while fundamental microbial ecology is ignored by bioinformaticians. Such attitudes do not contribute to cross-scientific progress. In this work, I attempt to avoid both, while evaluating, improving and applying *community profiling* methods. Also known as *phylogenetic marker gene profiling*, this is an invaluable technique for studying the diversity and composition of microbial communities. Here, the small subunit ribosomal RNA or its gene (SSU rRNA) is used as a phylogenetic marker and analysed using high throughput sequencing. Hopefully, this thesis can also serve as a primer for using this powerful technique.

The thesis is divided in two major parts. Part I is organised into five chapters. In addition to this introductory chapter, Chapter 2 provides a background to the experimental

and analytical methodology in environmental genomics, focussing on community profiling and sequencing technology. Chapter 3 presents four research questions serving to identify and illustrate gaps in the current knowledge, both of methodological and ecological character. An underlying aim was to close these gaps, and to bridge the gap between bioinformatics and microbial ecology. My contributions in this respect are discussed in Chapter 4, with detailed results available in the five research papers, enclosed in Part II. Finally, Chapter 5, provides concluding remarks and future perspectives.

Chapter 2

Background

2.1 Experimental methodology

2.1.1 Diversity and composition of microbial communities

Except for in microbiology laboratories, microorganisms nearly never exist as monocultures in nature. Understanding the structure of microbial communities and the mechanisms shaping them represent basic but poorly understood questions with significant ecological importance.

When describing community structure, the term *diversity* is typically used to describe the degree of variation, e.g. the number of taxa or Operational Taxonomic Units (OTUs; see Section 2.2.1). *Composition* also takes into account abundance metrics and is often discussed at lower taxonomic resolution (e.g. comparing the relative abundance of Archaea in relation to Bacteria). More specific, the term *alpha diversity* was introduced by Whittaker (1972) and refers to local diversity, typically within one sample or site. Whittaker also introduced *beta diversity* referring to the difference between sites of equal size. The term has since been used in several different respects, including measures taking into account differences in abundance of taxa. Because of this ambiguity, this thesis instead refers specifically to comparisons of either composition or alpha diversity.

An important reason for studying community structure is to increase our basal knowledge of microbial biogeography, dispersal and diversity of microorganisms. Another reason is to infer ecological function, or more exactly metabolism, niche, and contribution to biogeochemical cycles of individual community members. Based on community structure, more broad hypotheses can then be formulated about local, regional or global

community function and metabolism. This must always be done with care, since an assumption is made, that genetic or phenotypic similarity also implies functional similarity. The premises and limitations of such assumptions of homology is a matter of much debate and strongly influenced by methodology. In relation to this, the term *functional diversity* is often used, referring to the metabolic and functional repertoire of a community. There may also be functional redundancy to some extent in a community, if two taxa share the same niche or metabolic strategy (Nannipieri et al., 2003). Many techniques exist to study the functioning of communities directly, as opposed to their phylogenetic structure. Molecular techniques such as functional *metagenomics* (termed by Handelsman et al., 1998) and *metatranscriptomics* are very useful for this purpose, but are not discussed in depth in this thesis.

2.1.2 Conventional methods for studying microbial community structure and their limits

Direct observation of microbial communities is particularly challenging due to the microscopic scale and enormous numbers of cells involved. The small scale also adds complexity. In addition, the information that can be gained by direct microscopic observation is typically limited, particularly in prokaryotes. It can even be misleading due to observational bias and morphological plasticity (Justice et al., 2008). In spite of this, microscopy has remained an important tool for identification of microorganisms, since it was first used by Antonie van Leeuwenhoek in 1676. Since then, it has been refined and improved by techniques such as fluorescence microscopy combined with staining, using DAPI or acridine orange (reviewed in Kepner and Pratt, 1994). Using molecular probes, cells belonging to specific taxa can also be stained using fluorescence *in situ* hybridisation (FISH; DeLong et al., 1989).

In addition to analysing microorganisms directly in their natural habitats, early microbiologists like Louis Pasteur (1822-1895) and Robert Koch (1843-1919) developed techniques for isolating and cultivating them in pure cultures. These techniques were later complemented with the use of enrichment cultures by pioneers of microbial ecology like Martinus Beijerinck (1851-1931) and Sergei Winogradsky (1856-1953). Enrichment cultures enabled the selection, isolation and analysis of organisms that did not dominate a particular sample, or in other words, the first studies of microbial community structure.

Doubtless, cultivation-based (or “culture dependent”) techniques remain invaluable tools in microbial ecology. They also represent the only taxonomically valid approach for describing new bacterial species (with the exception of cyanobacteria; Euzéby,

2012). Unfortunately, only a small fraction of viable microorganisms of most environments can be easily cultured. Thus, those retrieved using cultivation are often not representative of the community as a whole. This realisation, although at least partly understood for decades before, was termed the *Great Plate Count Anomaly* by Staley and Konopka (1985). Fortunately, since the mid-1980s, microbiologists have been able to take advantage of and further develop molecular methods to bypass the need for culture dependent studies. Since then, our understanding of microbial communities has expanded and become significantly less biased towards the culturable minority.

2.1.3 Exploration of microbial communities using sequencing

DNA sequencing is the determination of the order of nucleotides in a DNA molecule, resulting in a *sequencing read*. This molecular technique has been of tremendous importance for recent progress in biology. The first forms of nucleotide sequencing instead used RNA as template and depended on laborious restriction digests and two-dimensional gel electrophoresis techniques (Holley et al., 1965). Seven years later, Sogin et al. (1972) argued for using of ribosomal RNA (rRNA) as a phylogenetic marker and use its sequence to determine the evolutionary history of prokaryotic microorganisms, as a means to classify them. Although not considered viable or meaningful by most microbiologists at the time (Sapp, 2005), the usefulness of this approach was later demonstrated by Woese and Fox (1977), who used rRNA sequencing to reveal the three phylogenetic domains of self-replicating life (*Bacteria*, *Archaea* and *Eukaryota*). This work showed that the "Prokaryota" was not a monophyletic group and that humans and all other eukaryotes share a common ancestor with the *Archaea* (except for our microbiome and mitochondria). Most importantly, it pioneered the essential and ongoing work of reconstructing the *Tree of Life*, representing our current understanding of the phylogeny of all living and extinct organisms.

The same year, Sanger et al. (1977) published a new method for DNA sequencing based on polymerase elongation with chain-terminating inhibitors (dideoxy nucleotides). The Sanger method allowed faster, less laborious sequencing and could generate longer sequences than previous methods. It quickly became the established sequencing method ("first generation") and formed the basis for automated Sanger sequencing, which also incorporates fluorescently labelled inhibitors. To this day, it remains the method of choice if both high accuracy and a long read lengths are required (up to about 800 bp).

By using various techniques for selection and separation, rRNA sequences could later be obtained directly from environmental samples (first by Stahl et al., 1984). Such

studies were greatly facilitated by applying a polymerase chain reaction (PCR) for selective DNA amplification of sequencing targets (Saiki et al. 1988; Section 2.1.4). PCR, followed by cloning using plasmid vectors in *Escherichia coli* (resulting in "clone libraries"), was first applied by Giovannoni et al. (1990) to sequence environmental genomic rRNA genes from a water sample from the Saragosso Sea. Ward et al. (1990) utilised a PCR independent variation of this method to sequence cDNA derived using reverse transcription of rRNA (Amann et al., 1995). These culture-independent studies revealed many organisms previously unknown to science, and pioneered an era of microbial exploration, which has continued to this day.

Recently, application of new sequencing methods (see 2.1.5) have demonstrated the vast extent of diversity remaining to be explored. The extent of diversity uncovered by these studies came as a surprise for many, although results agreed reasonably well with estimates based on DNA re-association studies carried out several years earlier (Torsvik et al., 1998). A large portion of the diversity was found to consist of low-abundant organisms and has therefore been termed the *rare biosphere* (Sogin et al., 2006). This can explain why earlier clone library surveys with relatively limited sequencing depths did not reveal this diversity.

There are several reasons for the suitability of rRNA as a phylogenetic marker. Most important, ribosomes are ubiquitous to all self-replicating organisms as they carry out the essential function of protein synthesis. Because of their fundamental importance, they maintain a high degree of conservation in sequence and secondary structure throughout evolution. For the same reason, horizontal gene transfer of the rRNA gene is thought to be very rare. It has been proven possible, however, and appears to have happened several times throughout evolution (Andam and Gogarten, 2011; Kitahara et al., 2012). Another property making rRNA a suitable marker is that it can be obtained in high quantities from most environmental samples, typically constituting approximately 95% of extracted RNA. Further, rRNA genes contain hypervariable regions interspersed with conserved ones, making them ideal for comparative sequence analysis and alignment.

Ribosomal RNA consists of two subunits: one large and one small. In prokaryotes, the large subunit (LSU) consists of two molecules named 5S and 23S after their sedimentation rates (measured in Svedberg; S). The small subunit (SSU) consists of one molecule (16S, here referred to as "SSU"). These three molecules are typically organised as a co-transcribed operon. For practical purposes the very earliest studies targeted the smallest of these, 5S rRNA (e.g. Sogin et al., 1972), but the SSU has since become the *de facto* phylogenetic marker, targeted by a tremendous number of sequencing studies (Tringe and Hugenholtz, 2008). Several studies have also targeted LSU as a important com-

plementary or alternative marker, especially in eukaryotes or to measure intra-species variation. It has also been suggested as a superior prokaryotic marker, but remains less popular, probably due to the relatively low number of LSU sequences in public databases (Yilmaz et al., 2011b).

Nuclear SSU rRNA is also a widely used phylogenetic marker in eukaryotes ("18S", analogous to 16S in prokaryotes). However, common alternatives exist that are more appropriate for particular taxa, e.g. the internal transcribed spacers (ITS) 1 and 2, widely used for Fungi (Santamaria et al., 2012); or the subunit I of cytochrome C oxidase (COI) for Metazoa, plants and other eukaryotes. The later is often referred to as "metabarcoding" when used for community profiling (Taberlet et al., 2012), or simply "barcoding" when used for identification of single species.

As an alternative to sequencing, community profiling can also be carried out using molecular fingerprinting methods, such as denaturing gradient gel electrophoresis (Muyzer et al., 1993), or terminal restriction fragment length polymorphism (Liu et al., 1997). Amplified sequences are then assayed without obtaining the sequences of the community. These techniques allow for relatively rapid comparisons between large numbers of samples at a lower cost than sequencing. However, semi-quantitative comparisons, taxonomic classification and determination of diversity is generally more challenging (Osborn et al., 2000), particularly for complex communities.

2.1.4 Targeted amplification and shotgun sequencing of rRNA

With the advent of new sequencing technologies (Section 2.1.5), a cloning step is no longer necessary, since individual DNA molecules can be used as template for sequencing. However, PCR is required to sequence only a specific gene, such as that encoding SSU rRNA, from extracted genomic DNA. The products of PCR amplification are referred to as *amplicons* and the method as *amplicon sequencing*. Extracted RNA, reverse-transcribed to complementary DNA (cDNA), can also be used as a template for PCR. This results in a profile of the active and abundant part of the community (Amann et al., 1995; Urich et al., 2008), whereas genomic DNA profiles the presence of organisms within the community, including less active, dormant and dead cells (Luna et al., 2002). Relatively small overlap between DNA- and RNA-based clone libraries from the same environment have been demonstrated previously (e.g. Gentile et al., 2006; Moeseneder et al., 2005) indicating that the approaches complement each other in a meaningful way.

To amplify SSU rRNA (or other markers) from a broad group of the community, "universal" oligonucleotide primers are required. Such primers utilise conserved regions

in the rRNA sequence and may also be degenerate, meaning that a cocktail of primers with different nucleotides at one or more degenerate positions, are used. Using standard PCR, it is however not practically possible to achieve true universality, i.e. the possibility to amplify all known microorganisms, should they be present in the sample. Importantly, no primers exist that cover a majority of taxa in each of the three domains of life (*Bacteria*, *Archaea* and *Eukaryota*). Special techniques using shorter, more universal primers have been suggested, but rely on the use of engineered polymerases (Isenbarger et al., 2008).

Depending on the read length of the sequencing technology, it is not necessary to amplify the entire SSU rRNA or its gene. Thus, many primer pairs are designed to amplify hypervariable regions inside the SSU, which can be more informative due to their lower degree of conservation. Several "universal" primer combinations exist and the choice varies depending on preference of individual research groups, organisms targeted and sequencing technology used (Klindworth et al., 2012). Primers with attached *barcodes* can also be used, to facilitate the mixing of several amplicon libraries into a single sequencing reaction (Hamady et al., 2008). The barcodes are then used to identify sequencing reads from individual samples. This technique is also referred to as "multiplexing" (and barcodes as "multiplex identifiers").

An alternative to amplicon sequencing is *shotgun sequencing,* where community DNA or cDNA is used directly as a template for sequencing. The use of shotgun sequencing of genomic DNA is referred to as shotgun *metagenomics* and requires that the DNA is subjected to shearing into smaller, random fragments. This will result in very few sequencing reads from SSU rRNA or other suitable phylogenetic markers. Instead, it is primarily a method for studying the functional structure of communities, rather than their taxonomic or phylogenetic composition and diversity.

Shotgun sequencing of community cDNA, however, can be successfully used as a community profiling method since the rRNA predominates RNA extracted from typical environmental samples (Urich et al., 2008). This method has the added advantages that it can avoid primer bias and other PCR artefacts (see 2.3.2-2.3.3) and that abundant mRNA transcripts are also sequenced. A disadvantage, however, is that it is not straightforward to determine the diversity of a sequenced community, since individual reads will differ in their position within the SSU or LSU rRNA. This can be compensated for to some extent by assembly (Miller et al., 2011; Radax et al., 2012). Known as shotgun metatranscriptomics, this method can also be used as a functional profiling method. This, which is actually the more common version, uses hybridisation or other methods for reducing the amount of rRNA prior to sequencing, thus enriching for mRNA.

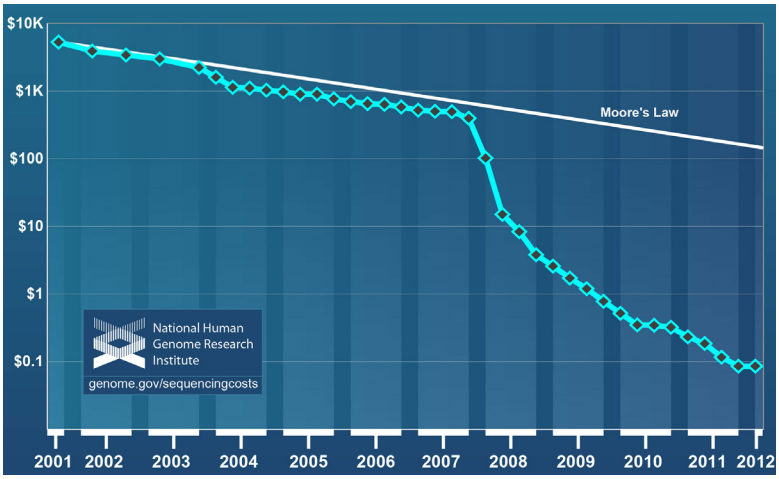


Figure 2.1: Development of sequencing costs per million base-pairs, compared to Moore's law. Source: Wetterstrand (2012) (used with permission).

2.1.5 Pyrosequencing and other "next generation" sequencing platforms

In the 1990s, sequencing was applied to a rapidly increasing range of studies in biology and medicine, most notoriously the completion of the first draft of the human genome in the year 2000. The demand for cheaper and faster sequencing helped drive a development of increased parallelisation of the Sanger method and, later on, of current (high-throughput) sequencing methods. The later are referred to here as *second-* and *third generation sequencing* to separate them from the Sanger method ("first-generation sequencing"). Since then, technical performance and the throughput-to-cost ratio has developed faster than the famous Moore's Law (Wetterstrand, 2012; Figure 2.1), often used to describe long-term performance trends in the computer hardware industry.

The first of the *second-generation* methods include (in chronological order): Massively Parallel Signature Sequencing developed by Lynx Therapeutics (Brenner et al., 2000) and Polony Sequencing (Mitra et al., 2003); in addition to Pyrosequencing**,** Illumina**,** SOLiD and Ion Torrent (see below)**.**

Pyrosequencing ("454") was the first of the second-generation technologies, commercially delivered as sequencing machines. Like Illumina, it is based on *sequencing-by-synthesis*, i.e. reading of each nucleotide base in a sequence during DNA replication. Pyrosequencing is based on stepwise flowing across a pico-titre plate of reagents including one specific deoxynucleoside triphosphate (dNTP). If the dNTP is incorporated, i.e is complementary to the extension position in the template DNA strand, pyrophosphate is released and indirectly detected as light emitted by the enzyme luciferase (Ronaghi et al., 1998). The technology was first commercialised by Pyrosequencing

AB, but the critical miniaturisation and massive parallelisation of the process was developed by 454 Life Sciences (today part of Roche; Margulies et al., 2005). The present version of the platform (GS FLX+ / Titanium) can generate about one million reads per run, with lengths of about 500 bp for amplicons or 750 bp for shotgun sequencing.

Early amplicon sequencing studies using pyrosequencing revealed surprisingly high diversity and novelty in the communities examined (Sogin et al., 2006; Roesch et al., 2007). Since then, it has become the most widely used platform in environmental genomics after Sanger sequencing, mainly thanks to its relatively long read length. This allows for more accurate taxonomic classification and determination of diversity in comparison to the shorter reads generated by other second-generation technologies.

Illumina (previously "Solexa") was developed by Solexa and later acquired by Illumina Inc. The technology is based on a parallelised technology where reversible dye-labelled-terminators are added during replication of a single stranded sequence. The base at that position can then be determined and the dye cleaved off, allowing another round of terminators to be added (Shokralla et al., 2012). Compared to pyrosequencing, Illumina generates shorter reads (150-250 bp), but with higher throughput and for a significantly lower cost. It can also be used with so called "mate-pair" reads, allowing for pairwise assembly of overlapping paired reads above 200 bp long (Rodrigue et al., 2010) Recent protocols for Illumina's MiSeq Personal Sequencer can even produce mate-pairs of length 2x250 bp (Illumina, 2012), theoretically allowing for 500 bp long assembled sequences.

Like pyrosequencing, Illumina has been used successfully for numerous environmental genomics studies (e.g. Qin et al., 2010; Caporaso et al., 2011).

SOLiD was developed by Applied Biosystems (now Life Technologies) and is based on *sequencing-by-ligation*. This technology relies on the the differential sensitivity of the enzyme DNA ligase for base-pairing mismatches (Shokralla et al., 2012). SOLiD is comparable to Illumina in terms of cost and throughput, yields slightly shorter read lengths (75 bp), but outperforms other existing methods in terms of accuracy (Glenn, 2011). Although no published studies to date used SOLiD for community profiling, it has been used for functional profiling (e.g. Liu et al., 2011b).

Ion Torrent (now owned by Life Technologies) is the newest of the second-generation platforms. It is based on a similar principle as pyrosequencing, but instead of indirect detection of released pyrophosphate by light, the positively charged hydrogen ion released along with it is detected. This is carried out by a proprietary ion sensor technology (Rothberg et al., 2011). During its two years of commercial availability, a number

of upgraded protocols and reagents have been released (including disposable sequencing "chips"). The present version (318) allows for a read length of 200 bp, at less than one tenth of the cost per base of pyrosequencing, but with lower accuracy (Glenn, 2011; LifeTechnologies, 2012). Ion Torrent has already been successfully used for amplicon sequencing studies by e.g. Whiteley et al. (2012) and Jünemann et al. (2012).

All second-generation sequencing technologies require thousands of copies of each unique DNA molecule to be sequenced. Because of this, they are depending on a PCR amplification step ("*in vitro* cloning"), carried out as part of the sequencing protocol. Pyrosequencing, Ion Torrent and SOLiD utilise a method called emulsion PCR, while Illumina utilise another known as bridge PCR. The template for this amplification is a single DNA molecule, replacing the need for cloning prior to sequencing. However, errors are introduced during any PCR reaction, which contributes to lower sequencing accuracy (see 2.3.2). In the case of amplicon sequencing, these errors are added to those already present from amplicon library construction.

Third generation sequencing technologies, however are PCR-independent, able to sequence individual DNA molecules (reviewed in Schadt et al., 2010). In theory, the approach allows for virtually unlimited read lengths with consistently high accuracy, as opposed to second-generation methods, where increased length has to be balanced vs. accuracy, partly because of their dependence on PCR. Commercialised third generation technologies include **Helicos** (Thompson and Steinmann, 2010; now discontinued) and **Single Molecule Real Time sequencing** (**SMRT**), developed by Pacific Biosciences (Eid et al., 2009). Neither platform is optimal for environmental genomics studies, SMRT because of its higher error rate and Helicos because of short read lengths. However, techniques like circular consensus sequencing (Travers et al., 2010) show great potential for amplicon sequencing using SMRT. Hybrid approaches combining SMRT and Illumina for genome sequencing (Koren et al., 2012) could also have potential for shotgun metatranscriptomics. Yet more promising are the many technologies in development, including Oxford Nanopores (Stoddart et al., 2009), yet to release proof-of-principle data at the time of writing.

2.2 Sequence analysis of community profiling data

2.2.1 Taxonomic classification

Taxonomic classification is the process of predicting the taxonomical positions or "memberships" of organisms from a studied community, based on similarity to previously studied taxa or environmental clades. Either all unique sequences obtained are

classified, or representatives from OTUs, alternatively contigs from assembly of shotgun sequencing reads. As discussed in Section 2.1.3, SSU rRNA is the most common marker used in prokaryotes and also useful for classification of eukaryotes. Several large projects also organise and align available SSU rRNA sequences into databases such as SILVA (Pruesse et al., 2007), RDP (Cole et al., 2009) and Greengenes (DeSantis et al., 2006). These represent invaluable resources for taxonomic classification and phylogenetic studies.

For each sequence to be classified, most classification methods utilise one of the following three strategies to identify a subset of similar sequences. Either each query sequence is:

1. pairwise aligned (using e.g. BLAST) to a set of reference sequences with known taxonomic affiliation;
2. fitted into an existing multiple alignment of reference sequences; or
3. divided into words of length *n* bp ("n-mers"), and the word composition compared to a reference dataset.

Classification is then based on similarity to the sequences in this subset. In the most trivial strategy, only the most similar reference sequence is selected and the query sequence is classified as belonging to the same taxon, possibly using a minimum similarity or scoring function. Alternatives to this nearest neighbour strategy is to find the lowest common ancestor (LCA) out of a set of nearest neighbours, a strategy first used for metagenomic sequences in the program MEGAN (Huson et al., 2007). This is trivial as long as the phylogenetic tree of the reference sequences is known, which unfortunately is typically not the case. Instead such classifiers must approximate it using a reference-to-taxonomy mapping, in the case of MEGAN the "NCBI Taxonomy" (Federhen, 2012). Another strategy is to apply Bayes' theorem, resulting in a Naive Bayes Classifier, so called because of its strong (naive) assumptions of independence among underlying features, i.e. word frequencies. Table 2.1 lists some of the most common classification tools, their strategies (according to the list above) and classification algorithms. In addition, tree-based hybrid methods exist that first identify a subset of related sequences, then predict a phylogenetic tree including the query sequence and related sequences.

Liu et al. (2008) compared the performance of a number of different classification strategies. For this, they used the RDP reference database (and cropped subsequences thereof), as well as three environmental datasets. In this comparison, methods based on multiple alignment (strategy 2) or tree-based methods showed higher accuracy for

Table 2.1: Common taxonomic classifiers for SSU rRNA datasets

Name	Reference	Strategy	Classification algorithm
PANGEA	Giongo et al. (2010)	1	Nearest neighbour
MARTA	Horton et al. (2010)	1	LCA-like
CLOTU	Kumar et al. (2011)	1	Nearest neighbour
GAST (VAMPS)	Huse et al. (2008)	1+3 with tree	LCA
SINA	Pruesse et al. (2012)	2	LCA
Greengenes Classifier	DeSantis et al. (2006)	2+3	Nearest neighbour
RDP Classifier	Wang et al. (2007)	3	Naive Bayes
MOTHUR	Schloss and Westcott (2011)	3 (default) or 1	Naive Bayes or LCA
SSuMMo	Leach et al. (2012)	Hidden Markov Models	Nearest neighbour
CREST	Paper I	1	LCA

"leave-one-out" testing with reference sequences. Environmental datasets, however, are often dominated by taxa that have not yet been taxonomically described. For these, Strategies 1 (pairwise alignments) and 2 (nucleotide composition), provided better results. Further, tree-based methods are generally more computationally demanding, which presents another challenge for large scale community profiling. The results of Liu et al. also indicated that accuracy of assignments vary depending on SSU rRNA regions, especially for short reads with lengths around 100 bp.

Besides those tailored for SSU rRNA, several specific methods also exist for other phylogenetic markers. In addition to community profiling data, many classification tools are also available for shotgun metagenomics. However, a fundamental problem associated with such analysis is that large numbers of genes in most genomes have undergone horizontal transfer at some point during their evolutionary history (Andam and Gogarten, 2011). Reads from such genes break the underlying assumption that sequence similarity corresponds to phylogenetic similarity, which can be compensated for by classifying only reads from suitable phylogenetic markers (Liu et al., 2011a).

2.2.2 Using Operational Taxonomic Units (OTUs) as proxies for microbial species

As mentioned in Section 2.1.2, isolation of microorganisms in pure culture remains the only accepted route to describing new bacterial or archaeal species, in spite of the problems associated with it. Apart from this approach, there is no consensus among microbial biologists concerning a species concept and it remains a debated topic (Achtman and Wagner, 2008; Caro-Quintero and Konstantinidis, 2011). Several studies suggest that genetic information alone may be insufficient to define a meaningful species concept, at any rate if only one marker gene is sequenced. Combining genetic and

ecological information may be a more successful approach (Reviewed in Fraser et al., 2009).

In environmental genomics, Operational Taxonomic Units (OTUs) based on genetic similarity are typically used as a proxy for species. Several definitions of OTUs exist, depending on method and preference. In SSU rRNA-based amplicon sequencing, a clustering approach is commonly used. This approach takes advantage of the fact that all sequences are obtained from a homologous region and thus can be directly compared. Result from such comparisons are collected in a global similarity matrix and are normally based on a global multiple alignment, or exhaustive pairwise alignments between all unique sequences. Agglomerative hierarchical clustering can then be carried out based on this matrix and OTUs are defined as all such clusters above a chosen similarity cutoff (commonly 97%, for SSU rRNA).

A problem with alignment-based clustering methods is that processing time often scales with the square of the number of unique sequences ($O(n^2)$). Several hybrid approaches have been developed that optimise this process for large datasets using dynamic programming (Cai and Sun, 2011), n-mer composition (Edgar, 2010; Ghodsi et al., 2011) or heuristic algorithms (Li and Godzik, 2006; Seguritan and Rohwer, 2001).

The 97% similarity cutoff commonly used to define OTUs in SSU rRNA amplicon studies was first suggested for full-length sequences of the gene, by Stackebrandt and Goebel (1994). It has later been suggested that this does not correspond to taxonomically described species or natural genetic clusters of variation, and a cutoff of 99% has instead been proposed (Acinas et al., 2004; Stackebrandt and Ebers, 2006). However, as the degree of variation differs between hypervariable regions of the SSU rRNA, the same cutoff value will give rise to different numbers of OTUs depending on which region that is sequenced (Kim et al., 2011). Further, lower cutoffs may increase accuracy by compensating for errors introduced by PCR or sequencing (see Section 2.3).

Several benchmarking studies (e.g. Sipos et al., 2010; White et al., 2010; Sun et al., 2011) have compared the consistency and quality of different OTU clustering approaches. Consensus results show that maximum-linkage clustering outperformed heuristic approaches and alternative manners of merging hierarchical clusters, such as average-linkage or UPMGA (Unweighted Pair Group Method with Arithmetic Mean).

An alternative to OTUs is to analyse clusters of sequences classified taxonomically to the same genus or higher ranks. A promising hybrid approach was developed by White et al. (2010) that utilise taxonomic annotations for semi-supervised OTU clustering.

2.2.3 Diversity estimates, comparison and extrapolation of richness

The most widely used estimate of alpha diversity in community profiling studies is *richness*, which is simply a count of the number of OTUs or taxa present in a community, habitat or sample. Richness and other diversity measures also depend on the size, heterogeneity and population of the environmental sample investigated. Because of the requirements of most PCR and second-generation sequencing protocols, relatively large samples compared to their microbial inhabitants are typically needed, in order to extract sufficient amounts of nucleic acid. This is sometimes desirable in order to compensate for fine-grained environmental heterogeneity. Regardless, a consequence is that few sequencing studies can provide a complete census of organisms present, except in the most simple of communities. To compensate, a number of methods exist that estimate *total richness* based on the captured diversity structure. An underlying assumption of such methods is that the sequence dataset analysed represent a random sample of the underlying diversity. This may not necessarily be true after PCR amplification, however.

One of the most simple richness estimators is Chao-1 (Chao, 1984). This non-parametric estimate is based only on the shares of observed OTUs (or taxa) represented by exactly one and two reads, respectively. While useful for estimating a minimum level of total richness, the Chao-1 and other commonly used non-parametric estimates (e.g. ACE; Chao and Lee, 1992) have been shown not to converge with increasing sequencing depth when applied to complex communities, to a large extent caused by sequencing artefacts (Gihring et al., 2011).

Parametric estimation can also be used to predict total richness (Hong et al., 2006). A disadvantage with this approach is that a specific shape must be assumed for the underlying taxon-abundance distribution (TAD). Quince et al. (2008) have developed a Bayesian method utilising a Markov chain Monte Carlo algorithm, to sample and optimise a range of TAD parametrisations, along with associated probabilities of fitting the underlying TAD. Based on this, estimates of total richness can be calculated. An advantage of this method is that estimates are provided as Bayesian confidence intervals rather than point estimates, allowing for significance assessments when comparing the richness between samples. These estimates are also less sensitive to sequencing artefacts than non-parametric estimates (Øvreås et al., unpublished).

A number of indices have also been proposed that take into account more aspects of community diversity than simply richness. One such diversity index is the Shannon index (*H'*), originally proposed to quantify the entropy in strings of text (Shannon,

1948). *H'* is defined as :

$$H' = -\sum_{i=1}^{R} p_i \log(p_i)$$

where total richness is R and relative abundance p. Another widely used diversity index is the Simpson index (D), which equals the probability that two entities (sequence reads) randomly taken from a dataset will represent the same class (i.e. taxon or OTU; Simpson, 1949). It equals:

$$D = \sum_{i=1}^{R} p_i^2$$

Both of the mentioned indices take into account the *evenness* of the community, i.e. how equal the different taxa or OTUs are numerically. Evenness can be described as a quotient between measured *H'* and its theoretical maximum, but this is problematic since such a calculation requires total richness to be known. In either case, an evenness of 1 indicates that all OTUs (taxa) are present at exactly the same abundance.

To allow unbiassed comparisons of diversity, datasets need to be derived using the same methods. Differences in e.g. extraction method, primer choice or PCR conditions can otherwise lead to biases that are not easily compensated (see 2.3.2-2.3.3). To compare datasets derived with the same methodology, but with significantly different size, random sub-sampling can be used. This is especially important when comparing total richness or estimates like Chao-1, whereas parametric estimates, evenness and diversity indices are less sensitive to such bias (Gihring et al., 2011). Another approach is rarefaction, which uses repeated random sub-sampling to calculate how observed richness depends on the sequencing effort in number of reads (Gotelli and Colwell, 2001). Results can be illustrated as a rarefaction curve and allows for an intuitive manner to compare richness, evenness and total sequencing efforts. Richness can also be compared between datasets at specific sequencing depth, but the technique cannot be used to estimate total richness by extrapolation (Gotelli and Colwell, 2001).

2.2.4 Comparison of community composition across datasets

One of the strengths of sequencing-based community profiling is that it allows for a large number of biological samples to be processed and sequenced at a relatively low cost. Lundin et al. (2012) have demonstrated that relatively few sequence reads per dataset (~1,000) may suffice to reveal 90% of the trends in compositional difference. This would allow for hundreds of barcoded datasets to be sequenced in one single pyrosequencing run, for example. Compared to molecular fingerprinting methods, it also allows for more direct compositional comparisons, that are not limited to only

predominating taxa. It is preferable that datasets to be compared are derived using the same set of methods, particularly for amplicon datasets where primer choice otherwise might bias composition (see 2.3.2).

Following taxonomic classification, relative abundances can be compared directly between datasets in order to find taxa present at significantly different relative abundance. For pairwise comparisons, a suitable statistical method for this is Fisher's exact test (Parks and Beiko, 2010). It is also important to adjust calculated p-values for multiple hypothesis testing, using e.g. Bonferroni correction. For comparisons between multiple datasets, it is often more useful to calculate a dissimilarity measure between each pair of datasets. This can be calculated using normal Euclidean distance. However, a range of dissimilaritiy measures more suitable for ecological data also exist. One of the most widely used for community profiling is the Bray-Curtis dissimilarity (d_{BC}; Bray and Curtis, 1957). This is analogous to the rectilinear Manhattan or "taxicab" distance, standardised by the sum of all taxon (or OTU) abundances, such that it is bound between 0 and 1. It is given by the formula:

$$d_{BC} = \frac{\sum_{i=1}^{R} |p_i - q_i|}{\sum_{i=1}^{R} (p_i + q_i)}$$

where $\boldsymbol{p}$ is the abundance in the first and $\boldsymbol{q}$ in the second dataset, and R is the combined richness of the two samples.

Based on the resulting dissimilarity matrix, multivariate statistical techniques can be used for explorative data analysis. An example is hierarchical clustering, which can handle non-metric dissimilarities like Bray-Curtis. In addition to clustering, ordination methods like non-metric multidimensional scaling (NMDS) are very useful for analysing relationships between datasets. Some ordination methods like principal components analysis (PCA) and clustering methods like k-means, require Euclidean distances. To compensate for problems this may cause when applied to ecological data, Hellinger transformation can be applied (normalisation to relative abundance and square root transformation; Legendre and Gallagher, 2001).

When comparing the composition between datasets of unequal size it is important to compensate by removing rare taxa (or OTUs) below the detection limit in the smaller dataset, especially when analysing presence or absence rather than using a dissimilarity index (Gobet et al., 2012). For amplicon sequence datasets, it is also common to remove all OTUs represented by only one read (singletons) before comparisons.

2.3 Sources of random and systematic errors, and methods for compensation

2.3.1 Sample handling, nucleic acid extraction and reverse transcription

The first step in preparation of an environmental sequencing library is the extraction of nucleic acid (DNA or RNA) from collected samples. However, such extractions are not always possible to carry out in the field, making it necessary to preserve samples temporarily. The time between collection and preservation of a sample has potential to influence the community, since it can involve severe stress factors, e.g. subjecting anaerobic organisms to oxygen, or filtering of a water sample. It is thus important to minimise this time, especially when analysing mRNA, whose half-life can be as short as a few minutes (Selinger et al., 2003).

The choice of preservation method may also have an influence on nucleic acid yield and quality. Simister et al. (2011) studied this influence for sponge endosymbiont samples, comparing preservation in liquid nitrogen to RNAlater (a buffered saturated solution of ammonium sulphate). The former was found to be favourable, but due to the small differences and the complications of handling liquid nitrogen in the field, RNAlater was nonetheless recommended.

Several studies have evaluated the influence of nucleic acid extraction methods on community profiling (e.g. Cuív et al., 2011; Simister et al., 2011; Terrat et al., 2012). A range of protocols exist, differing in whether RNA or DNA is extracted, or both simultaneously. Protocols also differ in the method for cell lysis. Physical lysis methods use e.g. bead-beating or freeze-thawing, while chemical methods use e.g. lysozyme and a mixture of other substances. The most appropriate extraction method depends on a range of factors such as the type of environment and organisms targeted (especially the type of cell walls); preferred nucleic acid; and analytical constraints. Regardless, the choice of extraction procedure can have a severe influence on the resulting community profile and fail to recover certain taxa, especially in complex environments like soil (Terrat et al., 2012).

In addition to the potential systematic errors from sample handling and extraction, reverse transcription of extracted RNA into cDNA is another source of systematic and random errors. Therefore, it has been recommended to always use technical duplication and never compare RNA-derived datasets using different primers or reverse transcription conditions (Ståhlberg et al., 2004).

2.3.2 PCR amplification bias and random drift

As already mentioned, many primer pairs exist that target different taxonomic groups and regions of the SSU rRNA. Primarily, differences in taxonomic coverage can bias results, if organisms present in the community have SSU rRNA sequences that do not match the primers used. The extent of such bias depends on the number of mismatches, their positions in the primer and the annealing temperature used (Sipos et al., 2007; Wu et al., 2009). Diversity estimates obtained will also depend on which SSU rRNA region that is targeted. In addition, shorter amplicon lengths may also skew the community profile and increase apparent diversity (Engelbrektson et al., 2010). Use of primers with degenerate positions can also bias results by preferential amplification of templates with the nucleobases G or C (Polz and Cavanaugh, 1998).

In addition to the systematic errors discussed, PCR may also introduce significant random error, skewing the community profile and leading to high variance in relative abundances between technical replicates, particularly for rare taxa. This effect, termed *PCR drift* by Polz and Cavanaugh (1998), is caused by the exponential nature of PCR and can be decreased by minimising the number of amplification cycles and using technical replication. Replicates may be pooled after PCR.

2.3.3 Chimeras, misincorporations and other PCR artefacts

In addition to amplification bias, several artefacts can arise during PCR. These include chimeric sequences formed by two different DNA molecules, point mutations and partial sequence deletions. Such artefacts can lead to several analytical problems. In addition to increasing diversity estimates, they can also suggest the existence of organisms that do not exist.

Taq polymerase, the high-temperature adapted DNA polymerase typically used in PCR, lacks exonuclease proofreading activity and therefore causes a relatively high rate of misincorporations during strand synthesis. This misincorporation rate has been estimated between $3x10^{-3}$ to $3x10^{-5}$ per nucleotide and cycle (von Wintzingerode et al., 1997). In either case, PCR can lead to a significant share of sequences having one or more point mutations, increasing with the number of amplification cycles used. The use of proof-reading DNA polymerases for PCR can decrease the rate of misincorporations, but may at the same time worsen other PCR artefacts (Gury et al., 2008).

Partial sequence deletions are caused by the formation of secondary structures such as hairpins. In addition to PCR, such artefacts can form with high frequency during

reverse transcription, making them particularly problematic in RNA-derived sequence datasets (von Wintzingerode et al., 1997).

The formation of chimeric sequences is a widespread problem, already identified in the early days of PCR in the late 1980s (von Wintzingerode et al., 1997). Most chimeras are generated from incomplete extension during PCR with the resulting fragment acting as a primer in the next amplification cycle. Like PCR drift, the effect be reduced by minimising the number of cycles used. Chimeras form more frequently when DNA molecules of shorter sequence lengths are used as template, which particularly is the case for cDNA (von Wintzingerode et al., 1997). Other factors also influence chimera formation, choice of DNA polymerase, annealing temperature and other PCR conditions, as well as the diversity of the community studied (Fonseca et al., 2012). All of these factors may also influence the extent of other PCR artefacts. In a study by Osborn et al. (2000), the type of *Taq* polymerase used had a larger influence on community fingerprinting results than any other conditions tested, including template concentration and number of cycles used.

2.3.4 Detection and removal of chimeric sequences

Detection of chimeric sequences is essentially a binary classification problem. All chimeras exist as recombinations of two or more parent sequences. Thus, their detection would be trivial if all chimera-free parent sequences were known, which unfortunately is not the case. In typical environmental datasets, many novel sequences are instead encountered. Most existing algorithms used for chimera detection utilise reference datasets of, ideally, chimera-free full-length sequences, to which the investigated community is compared using e.g. pairwise alignments. Investigated sequences may also be subdivided and their parts aligned separately to the reference sequences. A heuristic threshold or classification algorithm is then used to identify chimeras, which exhibit significantly differential similarity between their partial sequences.

Several methods that use variations on the approach described above include CheckChimera (Robison-Cox et al., 1995), Bellerophon (Huber et al., 2004), CCode (Gonzalez et al., 2005), Mallard (Ashelford et al., 2006) and ChiSeqI (Arigon et al., 2008). However, these were developed for the longer reads of first-generation sequencing and typically show poor performance for shorter reads from e.g. pyrosequencing. To compensate for this, a new generation of chimera classification tools were developed, including ChimeraSlayer (Haas et al., 2011), UCHIME (Edgar et al., 2011), Perseus (**Paper IV**) and DECIPHER (Wright et al., 2012). The former three methods utilise an alignment

strategy, whereas DECIPHER performs taxonomic classification and identifies subsequences (30-mers) that are not expected in the taxon to which the sequence has been classified. Perseus and UCHIME can also identify chimeras without using a reference dataset, instead analysing the relative abundances of sequences. A detailed discussion of these methods and their performance is found in Section 4.3.

2.3.5 Noise, artefacts and compensation in pyrosequencing and Ion Torrent data

By definition, all sequencing methods incorporate a step where the underlying sequence of nucleobases in a DNA molecule is measured (or multiple copies thereof, for first- and second generation methods). Thus, they are influenced by systematic- and random measurement errors, or in other words: *bias* and *noise*. Additional noise stems from the PCR amplification inherent to second-generation sequencing methods and. Furthermore, the template molecule may already contain errors before sequencing, stemming from the PCR carried out during amplicon library preparation, or from reverse transcription (see 2.3.2-2.3.3). First-generation sequencing is further biased by the cloning step necessary prior to sequencing.

Apart from PCR errors, the major source of noise in pyrosequencing stems from the measurement of light intensities. Ideally, a specific measured light intensity represents the number of nucleobases incorporated during each measurement step. In reality, however, noise contributes to a gaussian-like distribution of measured intensities for each ideal intensity. This is particularly problematic when more than one base is incorporated, representing a *homopolymer* stretch, and the variation of measured intensities increases with homopolymer length (Margulies et al., 2005). The variation also increases with base position in the sequence, giving rise to more noise in the 3' end of sequences (Balzer et al., 2010). The situation is essentially the same for Ion Torrent, although pH is measured instead of light intensities and available data indicate a higher noise level compared to pyrosequencing (Loman et al., 2012).

Adjusted measures of light intensity in pyrosequencing, or pH in Ion Torrent, are termed *flow values*. The name is derived from one type of nucleobase being introduced at a time (*flowed* across the pico-titre plate), during the sequencing-by-synthesis reaction. The default software delivered with these platforms can compensate to some extent for systematic errors, but cannot compensate for random noise. During base-calling, i.e. translation from measured flow values to sequence strings, each value is instead rounded to the closest integer, representing the homopolymer length. The result is a number of sequences with homopolymer stretches of incorrect lengths. A quality

score is also assigned to each base representing its reliability, consequentially being lower in longer homopolymer stretches.

A number of methods have been suggested in order to increase the accuracy of pyrosequenced datasets. Huse et al. (2007) studied the accuracy of the first pyrosequencing version (GS20) and found errors to be concentrated to reads significantly shorter than average or with ambiguous base calls ("Ns"). Removing these decreased the error rate from 0.5 to 0.2 % in the dataset evaluated. However, later studies using newer versions of the pyrosequencing were unable to reproduce a similar decrease in error rate, using such filtering (Schloss et al., 2011). Other filtering methods also incorporate trimming of reads based on quality (Kunin et al., 2009) and alignments based on predicted secondary structure of SSU rRNA (Cole et al., 2009).

For amplicon sequence datasets, a number of methods exist that utilise greedy agglomerative clustering algorithms, followed by a selection of unique representative sequence for each resulting OTU (Huse et al., 2010; Kunin and Hugenholtz, 2010). Others use iterative probabilistic clustering algorithms that incorporate flow value distributions (**Papers III, IV;** Reeder and Knight, 2010). These methods are discussed further in Section 4.3. Another recently developed method, DADA, use a similar clustering algorithm, but does not take into account flow values (Rosen et al., 2012). This method might also be a viable alternative for noise-reduction of other sequencing platforms such as Illumina. Yet another method, HPCall, improves base-calling to better predict homopolymer lengths in sequences (Beuf et al., 2012).

In addition to random noise and PCR errors, a number of systematic artefacts exist, specific to pyrosequencing (Balzer et al., 2010). One such artefact is duplicated reads, the removal of which is recommended for shotgun metagenome data (Gomez-Alvarez et al., 2009). In amplicon- and shotgun SSU rRNA data, identical reads are expected to appear naturally, however. Removing them would bias obtained taxon-abundance distributions rather than correct them.

Chapter 3

Research questions

As mentioned in Chapter 1, environmental genomics is a relatively young field, even more so at the onset of my PhD project. The limitations and preconditions of its rapidly developing methodology remain poorly understood, with several knowledge gaps. A "culture gap" between bioinformatics and microbial ecology risk to widen them. The major aim of this PhD thesis was to identify and helping to close such gaps, both of methodological and ecological characters. Specifically, the use of SSU rRNA-targeted pyrosequencing was investigated. A number of knowledge gaps were identified and four specific research questions were devised, to target a selection of these:

Q1: How to determine taxonomic composition and novelty of microbial communities?

The purpose of taxonomic classification is to map sequences derived from environmental samples to described taxa. Ultimately, it can also complement the underlying taxonomy, leading to improved systematics and evolutionary understanding. By tradition, only organisms cultured and studied in the laboratory have valid taxonomical standing. However, a growing part of the known microbial diversity is derived only from sequence data of uncultured organisms. How to complement microbial systematics with this type of data remains an unresolved issue. Regardless, an underlying goal of taxonomic classification should always be to derive as meaningful and accurate information as possible, from the community studied.

When I started to work on this thesis, none of the available classification methods were deemed sufficient for the community profiling data considered here (from SSU rRNA amplicon sequencing or shotgun metatranscriptomics). To address this, new resources for classification were developed, described in detail in **Paper I**. The sub-questions

below summarise the challenge of taxonomic classification and were formulated to aid the design of new methods:

Q1.1: Does a suitable set of reference sequences exist, with sufficient coverage of SSU rRNA sequences from all three domains of life?

Q1.2: How accurate is the taxonomic classification of the reference dataset? Is it updated in relation to current phylogenetic studies?

Q1.3: Which method and set of parameters offers the best classification accuracy using the chosen reference dataset?

Q1.4: Can novel sequences with unusually low similarity to reference sequences be identified and distinguished from sequencing noise?

Q1.5: How can the predicted diversity and taxonomic composition best be illustrated?

Q2: What is the reproducibility, extents and sources of bias of SSU rRNA-targeted pyrosequencing?

Prior to the work presented here, few studies, if any, had appropriately controlled the reproducibility of SSU rRNA-targeted pyrosequencing. For amplicon sequencing, many error sources are known to exist (see 2.3), including nucleic acid extraction, PCR, reverse transcription and choice of sequencing platform. All of these can potentially bias the estimates of community composition and diversity. However, the relative influences of individual sources of bias are largely unknown, as well as their relations to different experimental protocols, including primer choice and reverse transcription. As demonstrated using clone libraries by e.g. Moeseneder et al. (2005), the choice of nucleic acid analysed also has strong influence on results (either rRNA, or its gene from DNA). It is likely, however, that the lower sequencing depth as well as random errors inherent to cloning increased these differences. The use of second-generation sequencing has removed the need for a cloning step prior to sequencing, which has certainly helped to reduce such errors. In conclusion, the difference between the active (RNA) and present (DNA) organisms in environmental communities also remains an open question.

In **Paper II**, the extent of primer bias and other sources of variation were evaluated, using a hydrothermal vent associated microbial mat community as a model system. Several datasets were derived from two biological replicates of such mats, using both shotgun- and amplicon sequencing, from DNA and RNA. In addition to systematic error sources, the reproducibility of the amplicon sequencing protocol used was also estimated.

Q3: How diverse are microbial communities and to what extent can this be determined?

The diversity of microbial communities and the mechanisms shaping it remains a subject of much debate. Several explanations for the extent of microbial diversity and the rare biosphere (see Section 2.1.3) have been suggested, including sequencing artefacts (Reeder and Knight, 2009), host-virus interactions (Thingstad, 2000) and other mechanisms related to dispersal and grazing (reviewed in Pedrós-Alió, 2012). In order to test such ecological explanations, it is critical to understand the contribution of methodological artefacts to sequence diversity. Without it, no definition of the rare biosphere or estimate of its extent can be complete. Further, exhaustive studies remain unfeasible in most ecosystems, in spite of exponentially increasing sequencing capacity. Consequentially, extrapolation is necessary to infer total OTU richness. This would be challenging even with a perfect and unbiassed sequencing technology, but without understanding the consequence of noise introduced by PCR and sequencing, it is impossible. In other words, before the first part of the question proposed here can be answered ("how diverse are microbial communities?"), its second part must be addressed ("to what extent can this be determined"). This second part can be divided into the following sub-questions:

Q3.1: What is the extent of systematic and random errors introduced during library preparation and sequencing?

Q3.2: How can such errors influence current diversity estimates?

Q3.3: Can these errors be removed or compensated for, using e.g. filtering or clustering methods?

Q3.4: Given a successful compensation for errors (Q3.3), what is the extent of remaining errors and how do these influence diversity estimates?

These questions were addressed in **Papers III** and **IV**, by using "mock communities" with known real diversity and by providing a set of compensation methods (AmpliconNoise). My contributions to these papers consisted mainly in evaluation and implementation of new functionality.

The first part of Q3 ("how diverse are microbial communities?") was then addressed in **Papers II** and **V**. Even after the application of AmpliconNoise, significant errors will still remain and influence diversity estimates. Unfortunately, their extent and consequences (**Q3.4**) are expected to differ between natural ecosystems and the mock community samples used in **Papers III** & **IV**. Thus, important work remains in order to answer this question in a satisfactory manner (see Chapter 5).

Q4: How does diversity and composition of microbial communities vary across space and environmental gradients?

This fundamental question is difficult to answer, because of the multitude of microbial habitats, their complexity, and the many relevant physicochemical and biological parameters affecting indigenous microorganisms. The question is intentionally formulated very generally, bordering on the naive. Yet, it illustrates our incomplete knowledge of spatial heterogeneity across microbial habitats, both in terms of physichochemical conditions and community structure. To disentangle these patterns of microbial biogeography and the mechanisms that shape them is crucial for understanding the influence of environmental gradients on microbial communities. **Paper V** evaluates the influence of three such gradients, namely salinity, pH and dissolved oxygen. More specific, their correlation to diversity and composition of communities in alkaline soda lakes were studied. Soda lakes represent excellent model ecosystems, thought to harbour relatively unique communities of limited diversity. The question of spatial variability was also targeted using replication. Further, sequence datasets derived from both DNA and RNA were analysed. These were compared with respect to variation within and between lakes, asking the question of whether diversity and compositional patterns behave the same for genomic (potential) and transcribed (active) rRNA.

Chapter 4

Discussion

4.1 Taxonomic classification of SSU rRNA sequence data

Taxonomic classification (**Q1**) enables comparisons of environmental communities from different studies, and linking of taxonomic identity with function. The latter serves the purpose of predicting ecological roles of members in the studied community. It is particularly relevant in community profiling studies where, as opposed to functional metagenomic profiling, this cannot be directly inferred from similarity to protein-coding genes.

Preferably, a taxonomy should be coherent with its underlying evolutionary history, as inferred using sequencing data (or must be, according to phylogenetic nomenclature; Cantino and de Queiroz, 2010). Ecological roles can then be inferred from phylogenetic marker sequences mapped to the taxonomy, based on knowledge about the taxa to which they are classified. An important factor is the similarity between query and reference sequences. Those nearly identical to reference sequences from well-studied microbial species can be placed at higher resolution, whereas sequences with low identity must be classified more conservatively. It can also be of interest to study such sequences more closely, due to their novelty. Environmental sequences can be useful to include as references, either to improve classification accuracy or to infer possible ecological roles of sequences with low similarity to cultured, well-described species. They can also improve the accuracy of phylogenetic trees and thus the taxonomy itself (Nilsson et al., 2011).

SILVA (Pruesse et al., 2007) is a database that address the issues above, i.e. using aligned SSU rRNA sequence data to construct a phylogenetically consistent taxonomy including environmental sequences. However, no method prior to the work presented

here was capable of using SILVA or similar resources for classification of large sequence datasets. Neither did any method allow for direct detection of taxa with high novelty (low similarity to reference sequences). To this end, a classification scheme named "CREST" (Classification Resources for Environmental Sequence Tags; **Paper I**) was developed.

CREST was implemented both as a standalone program (GPL licensed), a web server and as part of the existing program MEGAN (Huson et al. 2007; Figure 4.1). This modular approach was intended to make it accessible for as many scientists as possible, regardless of background and computer skills. CREST uses an alignment-based classification strategy and the user can classify sequences using two different reference databases: (1) SilvaMod, resulting from a manual curation of the SILVA database and taxonomy, or (2) Greengenes (see below). It also allows for construction and use of custom databases and taxonomies using the program ARB (Ludwig et al., 2004). In addition, CREST reports total composition of one or more datasets, as well as taxon-specific diversity. Novel sequences are also identified and classified more conservatively.

The accuracy of CREST was compared to other classification tools intended for SSU rRNA sequence datasets (see 2.2.1), using two different cross-validation techniques, as well as environmental sequence data. Except for one test case, CREST outperformed the RDP Classifier (Wang et al., 2007), likely the most widely used classification tool for SSU rRNA sequence data. Classification tools like SINA aligner, SSuMO (Leach et al., 2012) and GAST (Huse et al., 2008) could not be tested using cross-validation due to the nature of their reference data. However, tests using environmental datasets were carried out for these tools, in all cases with discouraging results.

Greengenes is an SSU rRNA reference alignment and database similar to SILVA. During the development of CREST, Greengenes announced their own custom taxonomy, much like SILVA, taking into account the tree resulting from hierarchical clustering of reference sequence alignments (McDonald et al., 2012). A difference between the two is that Greengenes uses a custom-developed algorithm to taxonomically annotate sequences in its database, whereas SILVA was annotated manually. Further, Greengenes does not yet include nuclear eukaryotic sequences (18S rRNA). Neither does it provide a tool for classification of large datasets. However, training data derived from a subset of Greengenes is available for the RDP Classifier. When compared to the default training data of the RDP Classifier, this improved classification accuracy significantly, especially for archaeal taxa. Thanks to encouraging results, Greengenes was adopted into CREST as an alternative to SilvaMod. Comparisons between the two (**Paper I**) were inconclusive with the best accuracy depending on testing scheme. SilvaMod

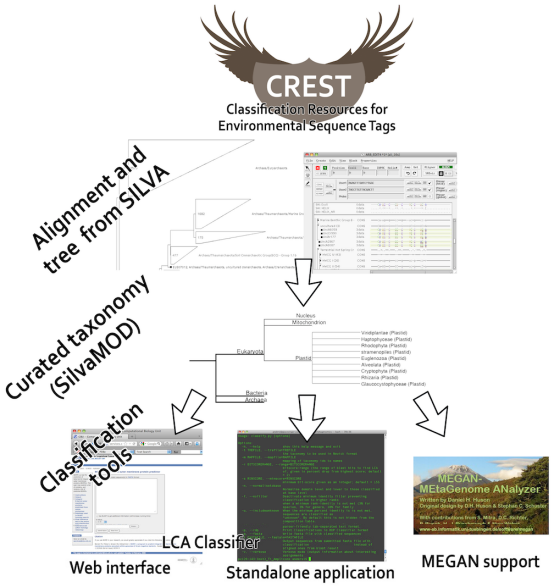


Figure 4.1: Overview of CREST: The flow of information during the construction of a new reference database such as SilvaMod is represented by arrows. The classification tools MEGAN or LCAClassifier use CREST taxonomy files for classification of environmental sequences aligned to the reference database with Megablast.

generally performed better when applied to environmental data, whereas Greengenes showed better cross-validation results. The taxonomic annotation approach utilised in Greengenes shows promise, since it can avoid time-consuming manual updates when incorporating new sequences, or findings from phylogenetic studies.

4.2 Bias and reproducibility of SSU rRNA-targeted pyrosequencing

There are many methodological issues associated with community profiling , such as reproducibility and sources of bias (see Section 2.3). These should always be carefully considered in studies utilising community profiling. In **Paper II**, the influence and causes of PCR bias were investigated, as well as differences between RNA- and DNA derived datasets (**Q2**). Reproducibility, and the influence of DNA extraction, could also be estimated to some extent. This analysis was carried out in connection to an ecological study of a microbial community from a hydrothermal vent associated biofilm. The degenerate, prokaryotic "universal" primer pair used for amplicon sequencing in the study, targets the V5-V6 hypervariable region of the SSU rRNA. V6 is a commonly targeted region in amplicon sequencing and known to be the second most suitable for determination of diversity and taxonomic affiliation, after V3 (Jeraldo et al., 2011). The pair was chosen because it has been shown to provide the highest possible

coverage among bacteria and archaea without mismatch, by Jørgensen et al. (2012).

The community investigated in **Paper II** was diverse, but relatively uneven and dominated by mesophilic *Epsilonproteobacteria*. Two biological replicates taken from adjacent mats were sampled and used to construct various sequencing libraries, yielding DNA- and RNA- (cDNA) derived amplicon- as well as shotgun datasets. Amplicon libraries were also prepared twice from one cDNA pool, allowing for estimation of reproducibility of the sequencing step. These two replicates were more similar than any other pair of datasets. One of the three DNA amplicon datasets was derived with an alternative extraction method, but did not deviate more from the other two.

Comparisons based on taxonomic composition (**Paper I**) showed that sequence datasets (**Paper II**) clustered mainly according to library type. When comparing compositional differences between sequences derived from the same sample, the biggest differences were found between DNA- vs. RNA-derived datasets, followed by shotgun- vs. amplicon datasets. These consistently differed more than the two biological replicates did from each other, when analysed with the same experimental strategy. Further analysis revealed that both DNA vs. RNA and shotgun vs. amplicon differences were mostly systematic, with the same taxa showing consistent and significant differences in all pairwise comparisons made. The taxa with the highest overrepresentation in DNA- compared to RNA-derived datasets were both archaeal.

The differences between amplicon- and shotgun datasets were also investigated further, using shotgun sequence data. Using linear regression, it was shown that about half of this variation could be explained by two factors: primer mismatch, and nucleobase composition at degenerate positions (**Paper II**). As expected, only two "nucleobase types" were statistically significant, namely (1) G or C, and (2) A or T. The first are responsible for stronger base pairing, requiring a higher melting temperature. It is possible that reverse transcription or shotgun sequencing also biased results against certain taxa, related to e.g. G/C-content, not possible to control using this simple methodology. However, the indicated causes of primer bias agree well with earlier studies of multi-template PCR (e.g. Polz and Cavanaugh, 1998; Wu et al., 2009). It also demonstrates that using universal primers with degenerate positions other than G/C or A/T may be problematic and compromise the semi-quantitative rigour of sequencing results.

Although not examined in detail, **Paper II** also indicates an impressive technical reproducibility of the methodology employed. Between the two technical replicates sequenced with different platforms, the maximum class level difference in relative abundance was very similar to that reported in a benchmarking study by Pilloni et al. (2012) (about 10% of relative abundance). Further, all OTUs represented by over 20 reads were found in both replicates (compared to 96% in the study by Pilloni et al.). This

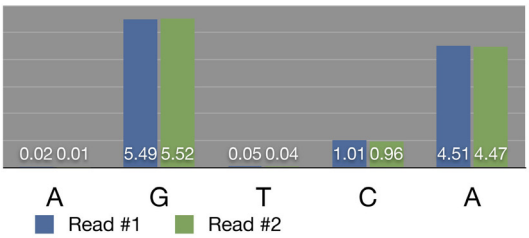
strongly contrasts the discouraging results of Zhou et al. (2011), possibly because of differences in the experimental procedure. For example, Zhou et al. did not use the two-step (reconditioning) PCR protocol shown to minimise the PCR bias caused by barcoded primers and lead to better reproducibility (Berry et al., 2011).

Paper II also indicates that bias caused by differing extraction methods was limited. This is supported in **Paper V** where the choice of extraction protocol had no discernible effect on clustering patterns of datasets. The choice of protocol used for harvesting cells from their aquatic habitat, however, had significant influence on the abundance of several taxa.

Because of its high resolution compared to functional profiling, community profiling is a technique especially suited for exploring the diversity of rare taxa and communities with low biomass. However, technical reproducibility may decrease with lower abundance (as noted by Legge, 2012). This is important to take into account when studying the rare biosphere, whose members often show higher spatial or temporal variation in abundance (**Paper V**; Youssef et al., 2010; Peura et al., 2012). In addition to inactive taxa, expected to show such distributions, part of this variation could be artifactual. Other studies have shown more conserved relative abundances for rare taxa (Bowen et al., 2012; Kirchman et al., 2010). As a semi-quantitative method, only *relative* abundances can be measured. Complementary methods, such as quantitative real-time PCR (RT-qPCR) or microscopy-based cell counting (**Paper V**), can be used to achieve quantitative results. However, RT-qPCR targeting DNA-derived SSU rRNA will not provide an estimate of cell numbers due to variations in the number of rRNA gene copies between taxa (Lee et al., 2009).

4.3 Dealing with sequence noise and determination of microbial diversity

The microbial biodiversity and omnipresence revealed by community profiling surveys is one of the most important findings in modern ecology. However, the extent and implications of this diversity remain poorly understood. The *rare biosphere* is one of its fascinating aspects, but many technical questions must be addressed to better understand it (**Q3.1-Q3.4**). This calls for the development of algorithms for noise-filtering and accurate diversity estimation, preferably with implementations easily used by most of the scientific community. For this purpose, *PyroNoise* was developed (**Paper III**). Later it was succeeded by *AmpliconNoise* (**Paper IV**). My contributions to these software packages were development (mainly of auxiliary workflow components), testing, benchmarking and documentation.




```
Read #1  GGGGGCAAAAA
         |||||  ||||
Read #2  GGGGGGCAAAA

Similarity = 9/11 = 82%
```

Figure 4.2: Example of homopolymer noise and its influence on base-calling in pyrosequencing. Although the two flowgrams in the example were generated from identical underlying sequences they yield reads with a pairwise similarity of only 83%.

PyroNoise (**Paper III**) is an algorithm for correction of noise-induced pyrosequencing artefacts in amplicon datasets, implemented in C and allowing for parallel execution. The idea behind PyroNoise was to work directly with the measured intensities, or *flowgrams*, for each sequence, because these contain information that is lost after base-calling (see Section 2.1.5). For example, two very similar flowgrams can lead to base-called sequences that do not appear similar when aligned (see Figure 4.2). What PyroNoise does can be described as merging such flowgrams into clusters, in cases where abundance data supports that they were generated from identical underlying sequences.

To parametrise PyroNoise, clones with known sequences were mixed into a *mock community* and pyrosequenced. The probability densities of measured intensities surrounding homopolymer lengths were then used, to derive the likelihood of a given flow value distribution being derived from a given nucleotide sequence. Expanding on this, a Bayesian model was derived, able to predict the total probability of the dataset, given assumptions of which underlying sequences that generated the flowgrams obtained. The PyroNoise algorithm uses expectation-maximisation to find a local maximum of this total probability, by re-assigning flowgrams to nucleotide sequences iteratively. The starting point is the mapping of each flowgram to the nucleotide sequence that would arise from normal base-calling.

Performance of PyroNoise was evaluated and compared to: (1) the "standard method" of OTU clustering as applied by Sogin et al. (2006) when defining the *rare biosphere*; (2) the RDP pipeline that incorporates quality score filtering (Cole et al., 2009); and (3) the assembly method CAP3 (Huang and Madan, 1999). Results showed that PyroNoise, followed by chimera removal and maximum-linkage clustering with 97% similarity cutoff, estimated OTU richness precisely when tested on data from the mock community. Other methods overestimated this richness by at least six times. For unique sequences rather than OTUs, PyroNoise overestimated richness by about 50%, whereas the standard method predicted a richness two orders of magnitude higher. Applied to

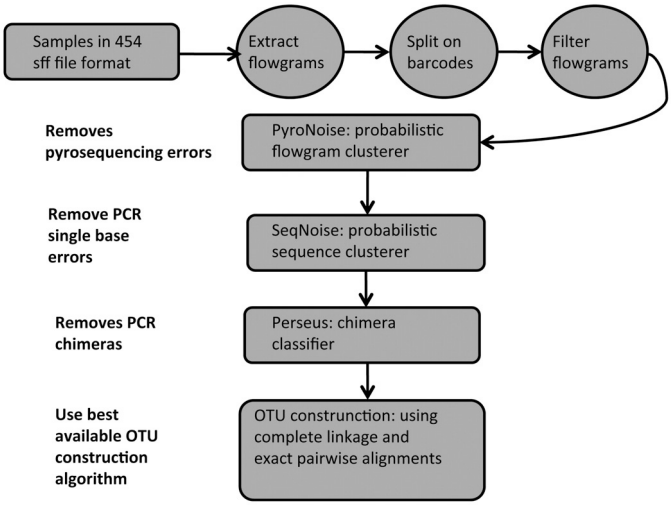


Figure 4.3: Overview of the AmpliconNoise workflow.

environmental data, PyroNoise predicted a 40% lower OTU richness compared to other methods.

In **Paper IV**, PyroNoise was re-parametrised to handle data from the newest pyrosequencing version (GS FLX Titanium) and re-implemented to decrease processing time. It was also complemented by two separate algorithms: *SeqNoise*, for removal of PCR noise; and *Perseus* for removal of chimeras. Together these are incorporated into a workflow called *AmpliconNoise* (see Figure 4.3). SeqNoise uses a similar Bayesian model as PyroNoise, although it clusters noise-corrected, base-called sequences to compensate for PCR misincorporations, rather than flowgrams.

As opposed to other chimera removal methods, Perseus (**Paper IV**) uses an approach that does not require a reference dataset. Instead, it utilises abundance data. For each sequence in the environmental dataset, an assumption is made that it could be chimeric and formed from two or more parents found in the same dataset. The probability of this is evaluated and compared to the probability that the sequence and its two parents evolved naturally from a common ancestor. Only parent sequences of equal or higher abundance are considered, since each chimera present will have undergone at least one amplification cycle less than its parents.

During the development of AmpliconNoise, various attempts were made to include a trimming step based on quality scores, followed by more rigorous filtering, prior to the PyroNoise algorithm. However, preliminary results indicated that this did not increase the accuracy of the obtained sequences. However, a filtering step similar to that suggested by Huse et al. (2007) is applied, as well as trimming.

Performance of AmpliconNoise was evaluated with existing and new mock communities, and compared to the methods DeNoiser (Reeder and Knight, 2010) and single-linkage pre-clustering (SLP; Huse et al., 2010). From the original mock community, the predicted number of unique sequences or OTUs never deviated more than 2% from the true number. SLP and DeNoiser over-predicted OTU richness with about 25%, at 97% similarity. For lower cutoffs, SLP underestimated richness. Further, SLP consistently increased per-base error rates in clustered sequences, whereas AmpliconNoise decreased it. The later finding is especially relevant for analysis of taxonomic composition, since the accuracy of classification is severely decreased by sequencing errors (particularly for nucleotide-composition based methods; Lanzén et al. 2011). Using SLP with e.g. the RDP Classifier should thus be considered unsuitable, and the use of raw reads as preferable. This issue arguably deserves more attention, rather than only focussing on diversity estimation in combination with sequencing noise.

Lee et al. (2012) have confirmed the superior accuracy of AmpliconNoise, also taking into account a novel aspect, namely predictions of relative abundances. The most important finding of **Papers III** and **IV,** however, is how sequencing and PCR noise can lead to inflated diversity estimates, if not properly compensated for. The problem of overestimation was independently confirmed by Kunin et al. (2009), contemporary with **Paper III**. This led to the realisation that earlier studies had overestimated the extent of the rare biosphere. Most of the scientific community adapted quickly and a clear majority of community profiling studies now incorporate noise compensation methods. AmpliconNoise, and the simplified DeNoiser implementation of PyroNoise, were also incorporated in widely used amplicon sequence analysis packages like QIIME (Reeder and Knight, 2010) and MOTHUR (Schloss et al., 2011).

Perseus, the method for chimera removal included in AmpliconNoise, was also compared independently to other methods. Results indicate that Perseus performs better than ChimeraSlayer (**Paper IV**) and comparably to UCHIME (Schloss et al., 2011), when applied to pyrosequenced amplicon data. DECIPHER (Wright et al., 2012) has not yet been compared to Perseus. According to its authors it performs better than UCHIME with longer sequence reads. UCHIME was used to complement AmpliconNoise in **Paper V**, because Perseus failed to identify a number of chimeras whose parents were not part of the sequence dataset. Instead, these chimeras appeared to be formed by recombinations between PCR products with cDNA-templates not covered by the primers used. This illustrates a shortcoming with the reference-free strategy of Perseus, and how the two methods can complement each other.

As mentioned in section 2.2.2, methods using heuristic speed-up of agglomerative OTU clustering were recently developed to handle large sequence datasets. However, for

pyrosequencing, AmpliconNoise reduces the number of unique sequences in most environmental datasets to such an extent that hierarchical clustering is no longer problematic. Further, heuristic clustering methods should be used with caution, given their reduced accuracy. However, AmpliconNoise and similar methods are not available for sequencing data from other platforms than pyrosequencing, for example Illumina. For these platforms, agglomerative clustering methods instead have an important role to play in reducing data size. It is theoretically possible to use AmpliconNoise with Ion Torrent data, which shares the same file format for flowgrams. However, specific parametrisation has not yet been carried out.

AmpliconNoise is more computationally demanding than alternative methods. This can be a serious limitation, especially at high sequencing depths. If computer power is limited, and accurate diversity estimates not important, DeNoiser may be sufficient for removing a majority of the noise and allowing OTU clustering. A method using graphics processing units to speed up AmpliconNoise analysis is also available (Gao and Bakos, 2012). Further, a promising noise-filtering method called DADA was recently developed (Rosen et al., 2012). According to the authors, it outperformed AmpliconNoise on the datasets analysed in **Paper IV**, in terms of both accuracy and processing time.

Another shortcoming of AmpliconNoise is its limitation to amplicon sequence data. Recently, Miller et al. (2011) developed the method EMIRGE for noise reduction and diversity estimation in shotgun sequence data. However, it requires Illumina data.

4.4 Community structure in environmental datasets

To investigate the diversity of environmental microbial communities (**Q3**), as well as the heterogeneity and variation of community structure across gradients (**Q4**), soda lakes were chosen as a model ecosystem (**Paper V**). Several factors make soda lakes particularly suitable for this purpose. Firstly, many relevant physicochemical and biological gradients can easily be targeted in these environments, both inside individual lakes and by comparing several lakes to each other. Salt concentrations and pH are examples of the later category, while depth-related gradients such as oxygen and light are examples of the former. These are typically practical to measure due to the limited lake depths. They are also steep due to high productivity and biomass (Zinabu and Taylor, 1997). The biomass also facilitates isolation of high-quality RNA and DNA from limited amounts of lake water.

Secondly, soda lakes are considered extreme environments because of their high pH

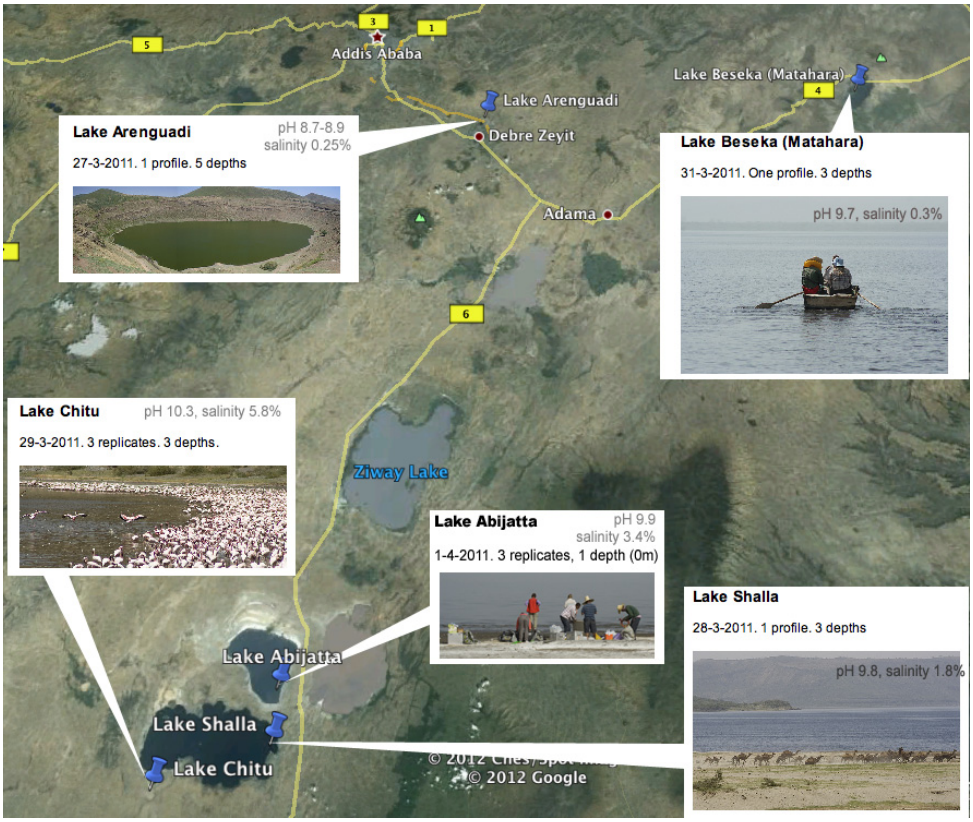


Figure 4.4: Geographical position of the soda lakes studied in Paper V, annotated with sampling regime, pH, salinity and photographs taken during sampling. Map source: Google Inc.

and salinity. Consequentially, they are expected to harbour limited diversity, more easily covered with community profiling. The extreme conditions also make soda lakes uninhabitable to all except adapted alkaliphiles. Thus, successful immigration is expected to be low. With large distances between them, soda lakes can be thus be considered biogeographical islands.

Soda lakes in the Ethiopian Rift Valley were studied in **Paper V**, selected for differing pH and salinities (see Figure 4.4). Three types of environmental and physicochemical variations were considered:

1. environmental heterogeneity and random community structure variation (between biological replicates);
2. internal depth-related environmental gradients (limited dispersal barriers); and
3. environmental gradients between lakes (significant dispersal barriers).

The studied communities (**Paper V**) appeared more diverse than expected, considering their high salinities and pH. Interestingly, the two most "extreme" lakes, showed the highest richness estimates. In previously studied neutral freshwater lakes (Logue et al., 2012; Peura et al., 2012; both using AmpliconNoise), OTU richness was approximately twice as high as in the studied soda lakes, at corresponding sequencing depths. Another SSU rRNA region was targeted by these studied, however, and water samples were not pre-filtered. Both factors may have biased the comparison of diversity to the favour of the neutral lakes. Taken together, this calls into question the notion that more extreme aquatic habitats are generally less diverse.

It is possible that limited dispersal may act to increase diversity and evolution of endemic organisms in these habitats, as previously suggested in euxinic freshwater lakes (Barberán and Casamayor, 2011). If this is the case, it calls into question another established principle, namely the first half of Baas-Becking's famous hypothesis: "everything is everywhere" (O'Malley, 2008). It is possible that the bias of culture-based studies influenced Baas-Becking to make this conclusion, where many cosmopolitan opportunistic organisms are often found in cell cultures regardless of sample origin (Souza et al., 2012).

One factor making many organisms unculturable is a dependence on chelators of insoluble Fe(III), called *siderophores*, from other organisms. These act as growth factors for many organisms that do not produce their own siderophores, even though the ability may be retained and utilised occasionally (D'Onofrio et al., 2010). It was suggested by Lewis (2010) that this dependence prevents sporulating cells from colonising environments with suboptimal conditions. It can be argued that this mechanism leads to a distribution pattern that disagrees with the Baas Becking hypothesis, at least as far as the active part of communities is concerned. Other observations of endemicity (e.g. Barberán and Casamayor, 2011; Martiny et al., 2006) also disagree with this hypothesis, although it arguably is rather inexactly formulated. As noted by Zinger et al. (2011), the topic is being debated actively, fuelled by contrasting results between studies. This inconsistency may have more to do with the taxonomic resolution considered, than underlying ecological principles (Souza et al., 2012; Zinger et al., 2011). This demonstrates the importance of accurate taxonomic classification (**Q1**) and careful methodological choices.

For all samples studied (**Papers II** and **V**) parametric estimates and rarefaction indicated that sequencing depth was far from exhaustive, with rare OTUs contributing considerably to diversity. Similar trends were observed in at least one of the mentioned neutral lake studies (Logue et al., 2012), as well as numerous studies of other environments (e.g Kirchman et al., 2010; Agogué et al., 2011; Gobet et al., 2012), including the

use of alternative marker genes (Vos et al., 2012). Such findings support the existence of a rare biosphere, although its contribution to global biodiversity and cell abundances may be significantly smaller than first thought.

A comparison of parametric richness estimates from the soda lake datasets (**Paper V**) demonstrates that richness did not vary significantly between spatial replicates. Compositional analysis also showed that replicates shared a substantial core community. The same applied between depths in certain lakes. However, Shannon- and and Simpson diversity estimates varied to a larger extent, indicating that taxon-abundance distributions were less conserved than richness, or absence- and presence of OTUs. This demonstrates the importance of spatial replication, even in habitats considered relatively homogenous. Unfortunately, replication has been relatively uncommon in community profiling studies (Prosser, 2010). In this case (**Paper V**), practical constraints limited the degree of replication possible, doubtless a common situation.

It was challenging to relate compositional variations to differences in environmental and physical parameters, because compositional variation between lakes did not show clear patterns of co-variance with such parameters. Sodium, potassium, oxygen and pH (in decreasing order) did however appear significantly correlated with community composition. Together these explained 30% of between-habitat variation. This was expected and agrees well with previous studies of aquatic environments (e.g. (Barberán and Casamayor, 2011; Herlemann et al., 2011; Lozupone and Knight, 2007). In soda lake sediments, pH has also showed a strong compositional influence, but with the opposite trend between richness and pH (Xiong et al., 2012). This could either indicate that sediment diversity is shaped by different mechanisms than aquatic diversity or, more likely, be an artefact of the insufficient noise compensation used by Xiong et al. (2012).

In addition to the soda lakes, diversity estimates from the hydrothermal microbial mats (**Paper II**) were of particular interest. Firstly, the study was one of the first to use AmpliconNoise on environmental data. Secondly, the community showed many similarities to “Marker 52”, the most diverse of the communities studied in Sogin et al. (2006), and later Huber et al. (2007). Both derived their energy from hydrothermal vents, had similar pH and taxonomic composition. Merged datasets from the microbial mat communities yielded 982 OTUs in total (97% similarity), similar to the richness of the lake community used for testing of PyroNoise (**Paper III**). According to rarefaction analysis, the “Marker 52” dataset is about five times as OTU-rich at corresponding sequencing depth (Huber et al., 2007), but inadequate noise reduction likely explains this discrepancy. A more recent study of similar environments using better noise compensation, similar to SLP, estimated richness values more similar to those of the microbial

mats studied in **Paper II** (Huber et al., 2010).

These findings support those of **Paper III** and indicate that the study introducing the term *"rare biosphere"* (Sogin et al., 2006), significantly overestimated its richness. Although a rare biosphere still appears to exist, its extent and the suggested mechanisms shaping it, would also benefit from re-examination. For example, Sogin et al. (2006) noted that the majority of rare OTUs showed lower similarity to reference strains and suggested their antiquity as an explanation, i.e. that these OTUs have persisted over geological time scales, largely thanks to the competitive benefits of being rare (Pedrós-Alió, 2012; Thingstad, 2000). Yet, such rare taxa are suggested to periodically take over environmental niches from dominating ones, enabling the microbial community to adapt more quickly to environmental changes (Sogin et al., 2006). If this happens frequently enough, it arguably would contradict the higher degree of novelty observed in the rare biosphere. Further, studies using other experimental conditions and methods for noise removal (including **Paper II**) often fail to reproduce such a high degree of novelty. Improved reference databases for taxonomic classification (e.g. **Paper I**) may also explain this discrepancy. Another explanation is that the majority of rare organisms in one site, typically are dominant elsewhere. The communities examined in **Paper V** contain several examples of such OTUs.

4.5 Complementarity of environmental genomics approaches

An important aspect of taxonomic classification (**Q1**) is the ability to compare composition across datasets and different experimental approaches (e.g, sequencing strategy, primer differences or even marker genes). This approach was important to the analyses carried out in **Papers II** and **V**, where it was used to compare composition between amplicon- and shotgun sequence datasets. With amplicon datasets, a taxonomy-independent approach can be employed instead, comparing the distribution of OTUs defined by clustering. Especially for novel or poorly categorised taxa, this can provide a better resolution. A disadvantage, however, is that many OTU-clustering algorithms are not deterministic and sensitive to minor differences in query sequences. Further, to add new datasets requires the entire clustering to be repeated, which can be computationally costly.

There has been some debate regarding which of these two practices that is more appropriate (taxonomy dependent vs. independent). Hybrid approaches have also been suggested (e.g. Lozupone and Knight 2005; Schloss and Westcott 2011; Sul et al. 2011). It is interesting to note that analyses based on taxonomic clusters and OTU dis-

tributions resulted in nearly identical clustering patterns in **Paper V.** However, clusters dissolved when deeper taxonomic levels were used (order or class rank).

It is possible that several studies reporting insignificant results using "taxonomy-dependent" approaches may have suffered from overly conservative classifications, for example labelling a sequence *Gammaproteobacteria* (class rank), rather than *Methylococcaceae* (family rank). The latter allows inference of more useful information, i.e. that we are dealing with a methane oxidiser. This potential pitfall of taxonomic classification can be worsened by poor coverage or classification of reference sequences. Taxonomy-independent approaches, based on OTU composition alone, can cause equally detrimental problems when noise is not compensated for (see Section 4.3).

To counteract these problems, 'functional biogeography' has been suggested as an alternative to traditional community profiling, to "allow for the possibility that the traits themselves disperse irrespective of their original hosts" (Raes et al., 2011). As demonstrated here (**Paper V**) and in countless other studies (e.g. Lozupone and Knight, 2007; King et al. 2010; Finkel et al., 2012; reviewed in Martiny et al., 2006), clear biogeographical patterns do exist and can be revealed also by community profiling, whereas functional patterns in cases lack similar resolution (Raes et al., 2011). Thus, complementary approaches using both functional and phylogenetic profiling clearly have a role to play in biogeographical studies, as in the definition of meaningful species concepts for *Archaea* and *Bacteria.* The comparative analysis carried out in **Paper II** demonstrates this complementarity of shotgun- and amplicon sequencing, even when considering only taxonomic composition. Using this combination, shotgun sequencing can be used to quantify the bias in amplicon sequence datasets, providing a valuable quality control. Regression analysis (as carried out in **Paper II)** could even be used to compensate for such primer bias.

The results of **Paper II** also illustrate the complementarity of RNA- and DNA-based sequencing. This is supported by other studies showing that the abundance of rRNA and its gene can differ substantially between taxa, particularly in environments with high overall activity (Rodríguez-Blanco et al., 2009). Presence of dead and dormant organisms are also expected to contribute to these differences (Luna et al., 2002; Jones and Lennon, 2010). However, in **Paper II**, the majority of taxa were encountered both in DNA and RNA, albeit at different relative abundances. This was not the case in several earlier studies based on clone-libraries (e.g. Moeseneder et al., 2005), probably caused by their lower sequencing depth and cloning bias.

Chapter 5

Conclusions and future perspectives

The results described in this thesis have contributed to increase the understanding of sequencing-based community profiling methods, their limitations, and sources of errors and bias. The work has also provided freely available software for the purpose of taxonomic classification and compensation of error sources. Their applicability to problems in microbial ecology was demonstrated, and in doing so, novel ecological insights were obtained. However, many of these insights, both of technical and ecological nature, warrant further investigation.

Early use of high throughput sequencing for community profiling overestimated community diversities significantly, due to unresolved methodological issues. However, the existence of a rare biosphere remains plausible and is supported by several theoretical and empirical studies (including **Papers II** and **V**). Several questions remain, critical to better understanding the ecological and evolutionary consequences of the rare biosphere:

- Whether most rare organisms are consistently rare, or experience habitats or periods with high abundance;
- How accurately diversity can be estimated, after the application of AmpliconNoise (**Q3.4**); and
- The extent of microbial diversity in different environmental communities, and what determines it.

These questions are intimately related to microbial biogeography and heterogeneity (**Q4**). Compensation of sequencing noise and other artefacts is fundamental for finding meaningful biogeographical patterns, as is accurate taxonomic classification or OTU clustering. Thus, results from earlier studies lacking proper noise compensation should

always be questioned, especially if lack of correlation was taken to indicate lack of a connection (as in e.g. Daghino et al., 2012). Re-analysis of existing datasets could instead have a meaningful role to play in re-evaluating biogeographical patterns targeted by previous studies, highlighting the usefulness of data depositories like the NCBI Sequence Read Archive (NCBI, 2013) and standardised metadata formats (Yilmaz et al., 2011a).

In conclusion, continuous and critical re-evaluation of methodology should be prioritised, especially in recent techniques such as environmental genomics. More established ecological concepts, derived from the study of larger organisms, also deserve critical re-evaluation when applied to the microbial world. The overwhelming complexity of microbial life provides a challenge to the development of general models, able to predict community structures. Still, important fundamental questions of microbial biogeography can be successfully targeted, given a careful choice of model ecosystem and experimental design. The soda lake community study (**Paper V**) provides a good example and may contribute to bridging the gap between bioinformatics and microbial ecology (see Chapter 1).

The diversity of the studied communities showed a counter-intuitive trend, with regard to salinity and pH. Although such factors limit the range of possible indigenous life, the predicted OTU richness was higher in the two most saline and alkaline lakes studied. A more ambitious study covering more lakes would increase the potential to reveal these trends, i.e. the underlying ecological and physicochemical relations to diversity and composition. It is possible that trophic interactions play a role in this, e.g. by limiting the diversity of bacteriovorous grazers or opportunistic organisms. Another possible explanation is that pH increases available dissolved carbon dioxide for photosynthetic organisms, leading to higher productivity and biomass (Grant, 2006). Indeed, a correlation between cell density and richness was established in one of the lakes studied. Therefore, future studies should seek to include measurements of production and cell density with phylogenetic diversity.

Limited effective immigration and endemism may also play a part in shaping the indigenous communities of soda lakes. As demonstrated by e.g. Sloan et al. (2006), neutral community models (NCMs) may help to reveal such connections. Observations in favour of this approach include that taxon-abundance distributions closely fitted log-series distributions, coherent with random colonisation from a large metacommunity (Hubbell, 2001). The applicability of an NCM does not imply that only random factors shape the community structure, however. In the studied communities, 30% of compositional variation could be explained by physicochemical factors (**Paper V**).

A large fraction of the OTUs were only encountered in one of the five lakes studied

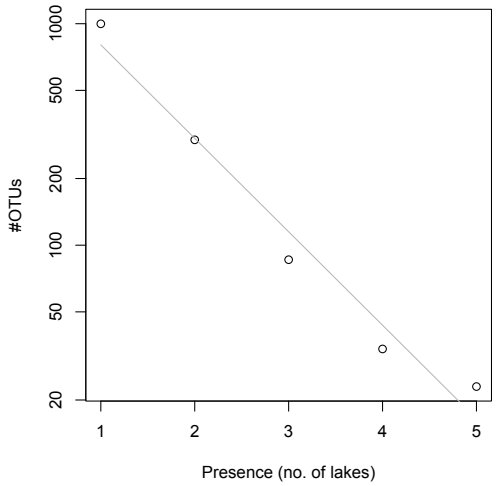


Figure 5.1: The number of OTUs (y-axis) shared between exactly *n* of the studied soda lakes (x-axis). To compensate for environmental heterogeneity and differing sequences depths, only OTUs from the surface of lakes at abundance above the minimum detection limit were included. The grey line indicates a fitted log-linear correlation.

(**Paper V**). Further, the number of OTUs shared between lakes showed an inverse log-linear correlation to the number of lakes considered (Figure 5.1). Since a limited number of soda lakes exist on Earth, this substantiates the existence of microorganisms endemic to individual lakes, as suggested earlier for stratified freshwater lakes (Barberán and Casamayor, 2011). This presents a problem to the NCM approach suggested above, because speciation is not considered in Hubbel's original framework. However, a number of recent models have incorporated speciation into NCMs and could prove useful for this purpose (reviewed in Kopp, 2010).

Although limited sequencing data exists from soda lakes, the data presented here would benefit from a meta-analysis, incorporating such datasets as they become available. Inclusion of existing data from neutral lakes, or clone libraries from soda lakes, can also be considered. A problem with this approach is differences between methodology, particularly primer choice and SSU rRNA region targeted. As demonstrated in **Paper II**, these factors must always be taken into account, or compensated for by establishing the influence of primer bias. Further, global similarity-based OTU clustering is not possible when combining datasets from different SSU rRNA regions. However, as demonstrated in **Paper V**, taxonomic classification using CREST (**Paper I**) provides a viable alternative to this approach.

The classification scheme and taxonomies available in CREST could also benefit from improvements, to increase the accuracy of this approach further. For example, nucle-

otide composition could be integrated with the current alignment-based method into a multiple classifier system. Another issue is the classification of novel sequences. Currently, those are bundled into a separate cluster, when several of them share a common parent taxon. CREST could be extended to avoid such clusters forming polyphyletic groups and increase classification accuracy. For example, a taxonomy-independent clustering approach could be employed for sequences with low similarity to reference sequences.

In spite of rigorous compensation, the influence of remaining sequencing noise and other artefacts cannot be excluded when comparing patterns of diversity across habitats. However, it is interesting to note that over half of the OTUs in the soda lakes studied (**Paper V**) were shared between at least two datasets. Several OTUs shared by many lakes were also consistently rare. This indicates that remaining noise is not solely responsible for the observed patterns and that increased sequencing efforts and better replication could help to improve our understanding of these patterns in diversity.

An important remaining task is to establish the degree of remaining sequencing artefacts after compensation with AmpliconNoise or other methods (**Q3.3**). The mock communities used for this purpose (**Papers III** and **IV**) are not expected to provide a realistic picture of remaining noise in complex environmental communities. A conceptually simple approach would be the construction and sequencing of more complex mock communities, incorporating several hundreds of clones with known sequences, mixed in predetermined proportions according to a distribution common in nature, e.g. log-normally. Although practically demanding to construct, these would provide useful standardised resources for benchmarking of sequencing technologies and downstream analysis tools.

Bibliography

Achtman M. and Wagner M. (2008). Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol*, 6(6):431–440.

Acinas S. G., Klepac-Ceraj V., Hunt D. E., Pharino C., Ceraj I. *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, 430(6999):551–554.

Agogué H., Lamy D., Neal P. R., Sogin M. L., and Herndl G. J. (2011). Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Mol Ecol*, 20(2):258–274.

Amann R. I., Ludwig W., and Schleifer K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*, 59(1):143–169.

Andam C. P. and Gogarten J. P. (2011). Biased gene transfer in microbial evolution. *Nat Rev Microbiol*, 9(7):543–555.

Arigon A.-M., Perrière G., and Gouy M. (2008). Automatic identification of large collections of protein-coding or rRNA sequences. *Biochimie*, 90(4):609–614.

Ashelford K. E., Chuzhanova N. A., Fry J. C., Jones A. J., and Weightmanva A. J. (2006). New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl Environ Microbiol*, 72(9):5734–5741.

Balzer S., Malde K., Lanzén A., Sharma A., and Jonassen I. (2010). Characteristics of 454 pyrosequencing data–enabling realistic simulation with flowsim. *Bioinformatics*, 26(18):i420–i425.

Barberán A. and Casamayor E. O. (2011). Euxinic freshwater hypolimnia promote bacterial endemicity in continental areas. *Microb Ecol*, 61(2):465–472.

Berg R. D. (1996). The indigenous gastrointestinal microflora. *Trends Microbiol*, 4(11):430–435.

Berry D., Mahfoudh K. B., Wagner M., and Loy A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl Environ Microbiol*, 77(21):7845–7849.

Beuf K. D., Schrijver J. D., Thas O., Criekinge W. V., Irizarry R. A., and Clement L. (2012). Improved base-calling and quality scores for 454 sequencing based on a Hurdle Poisson model. *BMC Bioinformatics*, 13(1):303.

Bowen J. L., Morrison H. G., Hobbie J. E., and Sogin M. L. (2012). Salt marsh sediment diversity: a test of the variability of the rare biosphere among environmental replicates. *ISME J*, 6:2014–2023.

Bray J. and Curtis J. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27:326–349.

Brenner S., Johnson M., Bridgham J., Golda G., Lloyd D. H. *et al.* (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 18(6):630–634.

Cai Y. and Sun Y. (2011). ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res*, 39(14):e95.

Cantino P. and de Queiroz K. (2010). PhyloCode - International Code of Phylogenetic Nomenclature, v4c. `http://www.ohio.edu/phylocode/`.

Caporaso J. G., Lauber C. L., Walters W. A., Berg-Lyons D., Lozupone C. A. *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A*, 108 Suppl 1:4516–4522.

Caro-Quintero A. and Konstantinidis K. T. (2011). Bacterial species may exist, metagenomics reveal. *Environ Microbiol*, 14(2):347–355.

Carrol L. (1872). *Through the looking glass and what Alice found there.* Macmillan, London.

Chao A. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, 11(4):265–270.

Chao A. and Lee S.-M. (1992). Estimating the Number of Classes via Sample Coverage. *Journal of the American Statistical Association*, 87(417):210–217.

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

Curtis T. P., Head I. M., Lunn M., Woodcock S., Schloss P. D., and Sloan W. T. (2006). What is the extent of prokaryotic diversity? *Philos Trans R Soc Lond B Biol Sci*, 361(1475):2023–2037.

Cuív O. P., Aguirre de Cárcer D., Jones M., Klaassens E., Worthley D. *et al.* (2011). The Effects from DNA Extraction Methods on the Evaluation of Microbial Diversity Associated with Human Colonic Tissue. *Microbial Ecology*, 61:353–362.

Daghino S., Murat C., Sizzano E., Girlanda M., and Perotto S. (2012). Fungal Diversity Is Not Determined by Mineral and Chemical Differences in Serpentine Substrates. *PLoS ONE*, 7(9):e44233.

DeLong E. F., Wickham G. S., and Pace N. R. (1989). Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science*, 243(4896):1360–1363.

DeSantis T. Z., Hugenholtz P., Larsen N., Rojas M., Brodie E. L. *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, 72(7):5069–5072.

D'Onofrio A., Crawford J., Stewart E., Witt K., Gavrish E. *et al.* (2010). Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chemistry & biology*, 17(3):254–64.

Edgar R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

Edgar R. C., Haas B. J., Clemente J. C., Quince C., and Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200.

Eid J., Fehr A., Gray J., Luong K., Lyle J. *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

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

Euzéby J. (2012). List of new names and new combinations previously effectively, but not validly, published. *Int J Syst Evol Microbiol*, 62(Pt 11):2549–2554.

Falkowski P. G., Fenchel T., and Delong E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science*, 320(5879):1034–1039.

Federhen S. (2012). The NCBI Taxonomy database. *Nucleic Acids Res*, 40(Database issue):D136–D143.

Finkel O. M., Burch A. Y., Elad T., Huse S. M., Lindow S. E. *et al.* (2012). Distance-Decay Relationships Partially Determine Diversity Patterns of Phyllosphere Bacteria on Tamrix Trees across the Sonoran Desert. *Appl Environ Microbiol*, 78(17):6187–6193.

Fonseca V. G., Nichols B., Lallias D., Quince C., Carvalho G. R. *et al.* (2012). Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic Acids Res*, 40(9):e66.

Fraser C., Alm E. J., Polz M. F., Spratt B. G., and Hanage W. P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science*, 323(5915):741–746.

Gao Y. and Bakos J. (2012). GPU Acceleration of Pyrosequencing Noise Removal. In *Application Accelerators in High Performance Computing (SAAHPC), 2012 Symposium on*, pages 94 –101.

Gentile G., Giuliano L., D'Auria G., Smedile F., Azzaro M. *et al.* (2006). Study of bacterial communities in Antarctic coastal waters by a combination of 16S rRNA and 16S rDNA sequencing. *Environ Microbiol*, 8(12):2150–2161.

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

Gihring T. M., Green S. J., and Schadt C. W. (2011). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environ Microbiol*, 14(2):285—290.

Giongo A., Crabb D. B., Davis-Richardson A. G., Chauliac D., Mobberley J. M. *et al.* (2010). PANGEA: pipeline for analysis of next generation amplicons. *ISME J*, 4(7):852–861.

Giovannoni S. J., Britschgi T. B., Moyer C. L., and Field K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, 345(6270):60–63.

Glenn T. C. (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, 11(5):759–769.

Gobet A., Böer S. I., Huse S. M., van Beusekom J. E. E., Quince C. *et al.* (2012). Diversity and dynamics of rare and of resident bacterial populations in coastal sands. *ISME J*, 6(3):542–553.

Gomez-Alvarez V., Teal T. K., and Schmidt T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J*, 3:1314–1317.

Gonzalez J. M., Zimmermann J., and Saiz-Jimenez C. (2005). Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics*, 21(3):333–337.

Gotelli N. J. and Colwell R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4(4):379–391.

Grant W. D. (2006). *Extremophiles*, chapter Alkaline environments and biodiversity. Encyclopedia of Life Support Systems (EOLSS). UNESCO, Eolss Publishers, Oxford, UK.

Gury J., Zinger L., Gielly L., Taberlet P., and Geremia R. A. (2008). Exonuclease activity of proofreading DNA polymerases is at the origin of artifacts in molecular profiling studies. *Electrophoresis*, 29(11):2437–2444.

Haas B. J., Gevers D., Earl A. M., Feldgarden M., Ward D. V. *et al.* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*, 21(3):494–504.

Hamady M., Walker J. J., Harris J. K., Gold N. J., and Knight R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods*, 5:235–237.

Handelsman J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, 68(4):669–85.

Handelsman J., Rondon M. R., Brady S. F., Clardy J., and Goodman R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*, 5(10):R245–R249.

Herlemann D. P., Labrenz M., Jürgens K., Bertilsson S., Waniek J. J., and Andersson A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J*, 5(10):1571–1579.

Holley R., Apgar J., Everett G., Madison J., Marquisee M. *et al.* (1965). Structure of a ribonucleic acid. *Science*, 147:1462–1465.

Hong S.-H., Bunge J., Jeon S.-O., and Epstein S. S. (2006). Predicting microbial species richness. *Proc Natl Acad Sci U S A*, 103(1):117–122.

Horton M., Bodenhausen N., and Bergelson J. (2010). MARTA: a suite of Java-based tools for assigning taxonomic status to DNA sequences. *Bioinformatics*, 26(4):568–569.

Huang X. and Madan A. (1999). CAP3: A DNA sequence assembly program. *Genome Res*, 9(9):868–77.

Hubbell S. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Monographs in Population Biology. Princeton University Press.

Huber J. A., Cantin H. V., Huse S. M., Mark Welch D. B., Sogin M. L., and Butterfield D. A. (2010). Isolated communities of Epsilonproteobacteria in hydrothermal vent fluids of the Mariana Arc seamounts. *FEMS Microbiology Ecology*, 73(3):538–549.

Huber J. A., Welch D. B. M., Morrison H. G., Huse S. M., Neal P. R. *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science*, 318(5847):97–100.

Huber T., Faulkner G., and Hugenholtz P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, 20(14):2317–2319.

Huse S., Huber J., Morrison H., Sogin M., and Welch D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*, 8(7):R143.

Huse S. M., Dethlefsen L., Huber J. A., Welch D. M., Relman D. A., and Sogin M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet*, 4(11):e1000255.

Huse S. M., Welch D. M., Morrison H. G., and Sogin M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol*, 12(7):1889–1898.

Huson D. H., Auch A. F., Qi J., and Schuster S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res*, 17(3):377–386.

Illumina (2012). MiSeq scientific data. `http://www.illumina.com/systems/miseq/scientific_data.ilmn`. Company website.

Isenbarger T. A., Finney M., Ríos-Velázquez C., Handelsman J., and Ruvkun G. (2008). Miniprimer PCR, a new lens for viewing the microbial world. *Appl Environ Microbiol*, 74(3):840–849.

Jeraldo P., Chia N., and Goldenfeld N. (2011). On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys. *Environ Microbiol*, 13(11):3000–3009.

Jones S. E. and Lennon J. T. (2010). Dormancy contributes to the maintenance of microbial diversity. *Proc Natl Acad Sci U S A*, 107(13):5881–5886.

Justice S. S., Hunstad D. A., Cegelski L., and Hultgren S. J. (2008). Morphological plasticity as a bacterial survival strategy. *Nat Rev Microbiol*, 6(2):162–168.

Jørgensen S. L., Hannisdal B., Lanzén A., Baumberger T., Flesland K. *et al.* (2012). Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proc Natl Acad Sci U S A*.

Jünemann S., Prior K., Szczepanowski R., Harks I., Ehmke B. *et al.* (2012). Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing. *PLoS One*, 7(8):e41606.

Kepner R. L. and Pratt J. R. (1994). Use of fluorochromes for direct enumeration of total bacteria in environmental samples: past and present. *Microbiol Rev*, 58(4):603–615.

Kim M., Morrison M., and Yu Z. (2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods*, 84(1):81–87.

King A. J., Freeman K. R., McCormick K. F., Lynch R. C., Lozupone C. *et al.* (2010). Biogeography and habitat modelling of high-alpine bacteria. *Nat Commun*, 1:53.

Kinross J. M., Darzi A. W., and Nicholson J. K. (2011). Gut microbiome-host interactions in health and disease. *Genome Med*, 3(3):14.

Kirchman D. L., Cottrell M. T., and Lovejoy C. (2010). The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ Microbiol*, 12(5):1132–1143.

Kitahara K., Yasutake Y., and 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*.

Klindworth A., Pruesse E., Schweer T., Peplies J., Quast C. *et al.* (2012). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*, 41(1):e1.

Kopp M. (2010). Speciation and the neutral theory of biodiversity: Modes of speciation affect patterns of biodiversity in neutral communities. *Bioessays*, 32(7):564–570.

Koren S., Schatz M. C., Walenz B. P., Martin J., Howard J. T. *et al.* (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*, 30:693–700.

Kuhn T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press, Chicago.

Kumar S., Carlsen T., Mevik B.-H., Enger P., Blaalid R. *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.

Kunin V., Engelbrektson A., Ochman H., and Hugenholtz P. (2009). Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environ Microbiol*, 12(1):118–123.

Kunin V. and Hugenholtz P. (2010). PyroTagger: A fast, accurate pipeline for analysis of rRNA amplicon pyrosequence datas. *The Open Journal*.

Lanzén A., Huson D. H., Jørgensen S. L., Øvreas L., and Urich T. (2011). Fast and accurate classification of environmental rRNA sequences using MEGAN and the Silva SSURef databases. Poster at the 12th Symposium on Bacterial Genetics and Ecology, Corfu, Greece.

Leach A. L. B., Chong J. P. J., and Redeker K. R. (2012). SSuMMo: rapid analysis, comparison and visualization of microbial communities. *Bioinformatics*, 28(5):679–686.

Lee C. K., Herbold C. W., Polson S. W., Wommack K. E., Williamson S. J. *et al.* (2012). Groundtruthing Next-Gen Sequencing for Microbial Ecology-Biases and Errors in Community Structure Estimates from PCR Amplicon Pyrosequencing. *PLoS ONE*, 7(9):e44224.

Lee Z. M.-P., Bussema C., and Schmidt T. M. (2009). rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Res*, 37(Database issue):D489–D493.

Legendre P. and Gallagher E. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2):271–280.

Legge R. (2012). *Analysis of microbial diversity by amplicon pyrosequencing*. PhD thesis, University of Nebraska-Lincoln.

Lewis K. (2010). The Uncultured Bacteria. `http://schaechter.asmblog.org/schaechter/2010/07/the-uncultured-bacteria.html`. Blog entry in Small Things Considered (ed. Moselio Schaechter).

Li W. and Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

LifeTechnologies (2012). Ion Torrent. `http://www.invitrogen.com/site/us/en/home/brands/Ion-Torrent.html`. Company Website.

Liu B., Gibbons T., Ghodsi M., Treangen T., and Pop M. (2011a). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, 12 Suppl 2:S4.

Liu W. T., Marsh T. L., Cheng H., and Forney L. J. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol*, 63(11):4516–4522.

Liu Z., DeSantis T. Z., Andersen G. L., and Knight R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res*, 36(18):e120.

Liu Z., Klatt C. G., Wood J. M., Rusch D. B., Ludwig M. *et al.* (2011b). Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. *ISME J*, 5(8):1279–1290.

Logue J. B., Langenheder S., Andersson A. F., Bertilsson S., Drakare S. *et al.* (2012). Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species-area relationships. *ISME J*, 6(6):1127–1136.

Loman N. J., Misra R. V., Dallman T. J., Constantinidou C., Gharbia S. E. *et al.* (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*, 30(5):434–439.

Lozupone C. and Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, 71(12):8228–8235.

Lozupone C. A. and Knight R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A*, 104(27):11436–11440.

Ludwig W., Strunk O., Westram R., Richter L., Meier H. *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res*, 32(4):1363–1371.

Luna G. M., Manini E., and Danovaro R. (2002). Large fraction of dead and inactive bacteria in coastal marine sediments: comparison of protocols for determination and ecological significance. *Appl Environ Microbiol*, 68(7):3509–3513.

Lundin D., Severin I., Logue J. B., Ostman O., Andersson A. F., and Lindström E. S. (2012). Which sequencing depth is sufficient to describe patterns in bacterial alfa- and beta-diversity? *Environmental Microbiology Reports*, 4(3):367–372.

Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S. *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

Martiny J. B. H., Bohannan B. J. M., Brown J. H., Colwell R. K., Fuhrman J. A. *et al.* (2006). Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol*, 4(2):102–112.

McDonald D., Price M. N., Goodrich J., Nawrocki E. P., DeSantis T. Z. *et al.* (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*, 6:610–618.

Miller C. S., Baker B. J., Thomas B. C., Singer S. W., and Banfield J. F. (2011). EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol*, 12(5):R44.

Mitra R. D., Shendure J., Olejnik J., Edyta-Krzymanska-Olejnik, and Church G. M. (2003). Fluorescent in situ sequencing on polymerase colonies. *Anal Biochem*, 320(1):55–65.

Moeseneder M. M., Arrieta J. M., and Herndl G. J. (2005). A comparison of DNA- and RNA-based clone libraries from the same marine bacterioplankton community. *FEMS Microbiol Ecol*, 51(3):341–352.

Muyzer G., de Waal E. C., and Uitterlinden A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol*, 59(3):695–700.

Nannipieri P., Ascher J., Ceccherini M. T., Landi L., Pietramellara G., and Renella G. (2003). Microbial diversity and soil functions. *European Journal of Soil Science*, 54(4):655–670.

NCBI (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 41(D1):D8–D20.

Nilsson R. H., Ryberg M., Sjökvist E., and Abarenkov K. (2011). Rethinking taxon sampling in the light of environmental sequencing. *Cladistics*, 27(2):197–203.

O'Malley M. A. (2008). 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Stud Hist Philos Biol Biomed Sci*, 39(3):314–325.

Osborn A. M., Moore E. R. B., and Timmis K. N. (2000). An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environ Microbiol*, 2(1):39–50.

Parks D. H. and Beiko R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26(6):715–721.

Pedrós-Alió C. (2012). The rare bacterial biosphere. *Ann Rev Mar Sci*, 4:449–466.

Peura S., Eiler A., Bertilsson S., Nykänen H., Tiirola M., and Jonesbend R. I. (2012). Distinct and diverse anaerobic bacterial communities in boreal lakes dominated by candidate division OD1. *ISME J*, 6(9):1640–1652.

Pikuta E. V., Hoover R. B., and Tang J. (2007). Microbial extremophiles at the limits of life. *Crit Rev Microbiol*, 33(3):183–209.

Pilloni G., Granitsiotis M. S., Engel M., and Lueders T. (2012). Testing the limits of 454 pyrotag sequencing: reproducibility, quantitative assessment and comparison to T-RFLP fingerprinting of aquifer microbes. *PLoS One*, 7(7):e40467.

Polz M. F. and Cavanaugh C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol*, 64(10):3724–3730.

Prosser J. I. (2010). Replicate or lie. *Environ Microbiol*, 12(7):1806–1810.

Pruesse E., Peplies J., and Glöckner F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14):1823–1829.

Pruesse E., Quast C., Knittle K., Fuchs B. M., Ludwig W. *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, 35:7188–7196.

Qin J., Li R., Raes J., Arumugam M., Burgdorf K. S. *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

Quince C., Curtis T. P., and Sloan W. T. (2008). The rational exploration of microbial diversity. *ISME J*, 2(10):997–1006.

Radax R., Rattei T., Lanzen A., Bayer C., Rapp H. T. *et al.* (2012). Metatranscriptomics of the marine sponge Geodia barretti: tackling phylogeny and function of its microbial community. *Environ Microbiol*, 14(5):1308–1324.

Raes J., Letunic I., Yamada T., Jensen L. J., and Bork P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol*, 7:473.

Reeder J. and Knight R. (2009). The 'rare biosphere': a reality check. *Nat Methods*, 6(9):636–637.

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

Robison-Cox J. F., Bateson M. M., and Ward D. M. (1995). Evaluation of nearest-neighbor methods for detection of chimeric small-subunit rRNA sequences. *Appl Environ Microbiol*, 61(4):1240–1245.

Rodrigue S., Materna A. C., Timberlake S. C., Blackburn M. C., Malmstrom R. R. *et al.* (2010). Unlocking short read sequencing for metagenomics. *PLoS One*, 5(7):e11840.

Rodríguez-Blanco A., Ghiglione J.-F., Catala P., Casamayor E. O., and Lebaron P. (2009). Spatial comparison of total vs. active bacterial populations by coupling genetic fingerprinting and clone library analyses in the NW Mediterranean Sea. *FEMS Microbiol Ecol*, 67(1):30–42.

Roesch L. F. W., Fulthorpe R. R., Riva A., Casella G., Hadwin A. K. M. *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*, 1(4):283–290.

Ronaghi M., Uhlén M., and Nyrén P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363, 365.

Rosen M. J., Callahan B. J., Fisher D. S., and Holmes S. P. (2012). Denoising PCR-amplified metagenome data. *BMC Bioinformatics*, 13(1):283.

Rothberg J. M., Hinz W., Rearick T. M., Schultz J., Mileski W. *et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352.

Saiki R. K., Gelfand D. H., Stoffel S., Scharf S. J., Higuchi R. *et al.* (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491.

Sanger F., Nicklen S., and Coulson A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467.

Santamaria M., Fosso B., Consiglio A., Caro G. D., Grillo G. *et al.* (2012). Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform*, 13(6):682–695.

Sapp J. (2005). The prokaryote-eukaryote dichotomy: meanings and mythology. *Microbiol Mol Biol Rev*, 69(2):292–305.

Schadt E. E., Turner S., and Kasarskis A. (2010). A window into third-generation sequencing. *Hum Mol Genet*, 19(R2):R227–R240.

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

Schloss P. D. and Westcott S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl Environ Microbiol*, 77(10):3219–3226.

Seguritan V. and Rohwer F. (2001). FastGroup: a program to dereplicate libraries of 16S rDNA sequences. *BMC Bioinformatics*, 2:9.

Selinger D. W., Saxena R. M., Cheung K. J., Church G. M., and Rosenow C. (2003). Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome Res*, 13(2):216–223.

Shannon C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27:379–423, 623–656.

Shokralla S., Spall J. L., Gibson J. F., and HajibabaehI M. (2012). Next-generation sequencing technologies for environmental DNA research. *Mol Ecol*, 21(8):1794–1805.

Simister R. L., Schmitt S., and Taylor M. W. (2011). Evaluating methods for the preservation and extraction of DNA and RNA for analysis of microbial communities in marine sponges. *J Exp Mar Bio Ecol*, 397(1):38–43.

Simpson E. (1949). Measurement of diversity. *Nature*, 163(4148):688.

Sipos M., Jeraldo P., Chia N., Qu A., Dhillon A. S. *et al.* (2010). Robust computational analysis of rRNA hypervariable tag datasets. *PLoS One*, 5(12):e15220.

Sipos R., enAnna J Székely, Palatinszky M., Révész S., Márialigeti K., and Nikolausz M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol*, 60(2):341–350.

Sloan W. T., Lunn M., Woodcock S., Head I. M., Nee S., and Curtis T. P. (2006). Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol*, 8(4):732–740.

Sogin M. L., Morrison H. G., Huber J. A., Welch D. M., Huse S. M. *et al.* (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, 103(32):12115–12120.

Sogin S., Sogin M., and Woese C. (1972). Phylogenetic measurement in procaryotes by primary structural characterization. *J Mol Evol*, 1:173–184.

Souza V., Eguiarte L. E., Travisano M., Elser J. J., Rooks C., and Siefert J. L. (2012). Travel, Sex, and Food: What's Speciation Got to Do with It? *Astrobiology*, 12(7):634–640.

Stackebrandt E. and Ebers J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today*, 33:152–155.

Stackebrandt E. and Goebel B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J Syst Bacteriol*, 44(4):846–849.

Stahl D. A., Lane D. J., Olsen G. J., and Pace N. R. (1984). Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science*, 224(4647):409–411.

Staley J. T. and Konopka A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol*, 39:321–346.

Stoddart D., Heron A. J., Mikhailova E., Maglia G., and Bayley H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc Natl Acad Sci U S A*, 106(19):7702–7707.

Ståhlberg A., Håkansson J., Xian X., Semb H., and Kubista M. (2004). Properties of the reverse transcription reaction in mRNA quantification. *Clin Chem*, 50(3):509–515.

Sul W. J., Cole J. R., da C Jesus E., Wang Q., Farris R. J. *et al.* (2011). Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. *Proc Natl Acad Sci U S A*, 108(35):14637–14642.

Sun Y., Cai Y., Huse S. M., Knight R., Farmerie W. G. *et al.* (2011). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform*, 13(1):107–121.

Taberlet P., Coissac E., Pompanon F., Brochman C., and Willerslev E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol*, 21(8):2045–2050.

Terrat S., Christen R., Dequiedt S., Lelièvre M., Nowak V. *et al.* (2012). Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. *Microb Biotechnol*, 5(1):135–141.

Thingstad T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.*, 45(6):1320–1328.

Thompson J. F. and Steinmann K. E. (2010). Single molecule sequencing with a HeliScope genetic analysis system. *Curr Protoc Mol Biol*, Chapter 7:Unit7.10.

Torsvik V., Daae F. L., Sandaa R. A., and Ovreås L. (1998). Novel techniques for analysing microbial diversity in natural and perturbed environments. *J Biotechnol*, 64(1):53–62.

Travers K. J., Chin C.-S., Rank D. R., Eid J. S., and Turner S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res*, 38(15):e159.

Tringe S. G. and Hugenholtz P. (2008). A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol*, 11(5):442–446.

Urich T., Lanzén A., Qi J., Huson D. H., Schleper C., and Schuster S. C. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE*, 3(6):e2527.

von Wintzingerode F., Göbel U. B., and Stackebrandt E. (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev*, 21(3):213–229.

Vos M., Quince C., Pijl A. S., de Hollander M., and Kowalchuk G. A. (2012). A Comparison of rpoB and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity. *PLoS One*, 7(2):e30600.

Wang Q., Garrity G. M., Tiedje J. M., and Cole J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*, 73(16):5261–5267.

Ward D. M., Weller R., and Bateson M. M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, 345(6270):63–65.

Wetterstrand K. (2012). DNA Sequencing Costs. `http://www.genome.gov/sequencingcosts/`. Data from the NHGRI Genome Sequencing Program (GSP).

White J. R., Navlakha S., Nagarajan N., Ghodsi M.-R., Kingsford C., and Pop M. (2010). Alignment and clustering of phylogenetic markers–implications for microbial diversity studies. *BMC Bioinformatics*, 11:152.

Whiteley A. S., Jenkins S., Waite I., Kresoje N., Payne H. *et al.* (2012). Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *J Microbiol Methods*, 91(1):80–88.

Whitman W. B., Coleman D. C., and Wiebe W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A*, 95(12):6578–6583.

Whittaker R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon*, 21(2):213–251.

Woese C. R. and Fox G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74(11):5088–5090.

Wright E. S., Yilmaz L. S., and Noguera D. R. (2012). DECIPHER, a Search-Based Approach to Chimera Identification for 16S rRNA Sequences. *Appl Environ Microbiol*, 78(3):717–725.

Wu J.-H., Hong P.-Y., and Liu W.-T. (2009). Quantitative Effects of Position and Type of Single Mismatch on Single Base Primer Extension. *J Microbiol Methods*, 77:267–275.

Xiong J., Liu Y., Lin X., Zhang H., Zeng J. *et al.* (2012). Geographic distance and pH drive bacterial distribution in alkaline lake sediments across Tibetan Plateau. *Environ Microbiol*, 14(9):2457–2466.

Yilmaz P., Kottmann R., Field D., Knight R., Cole J. R. *et al.* (2011a). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol*, 29(5):415–420.

Yilmaz P., Kottmann R., Pruesse E., Quast C., and Glöckner F. O. (2011b). Analysis of 23S rRNA genes in metagenomes - A case study from the Global Ocean Sampling Expedition. *Syst Appl Microbiol*, 34(6):462–469.

Youssef N. H., Couger M. B., and Elshahed M. S. (2010). Fine-scale bacterial beta diversity within a complex ecosystem (Zodletone Spring, OK, USA): the role of the rare biosphere. *PLoS One*, 5(8):e12414.

Zhou J., Wu L., Deng Y., Zhi X., Jiang Y.-H. *et al.* (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J*, 5(8):1303–1313.

Zinabu G. and Taylor W. (1997). Bacteria-chlorophyll relationships in Ethiopian lakes of varying salinity: are soda lakes different? *J Plankton Res*, 19(5):647–654.

Zinger L., Lejon D. P. H., Baptist F., Bouasria A., Aubert S. *et al.* (2011). Contrasting diversity patterns of crenarchaeal, bacterial and fungal soil communities in an alpine landscape. *PLoS One*, 6(5):e19950.

Øvreås L., Quince C., Sloan W. T., Lanzén A., Davenport R. *et al.* (unpublished). Determining the Microbial Diversity and Species Abundance Patterns in Arctic Soils using Rational Methods.