Comparative Metagenome Analysis

DISSERTATION

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von:

M. Sc. Suparna Mitra
aus Kolkata (Indien)

Tübingen
2010

Tag der mündlichen Qualifikation:	17.11.2010
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Daniel H. Huson
2. Berichterstatter:	Prof. Dr. Jack A. Gilbert
3. Berichterstatter:	Prof. Dr. Elena Marchiori

Statement of Authorship

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any degree at any educational institution, except where due acknowledgement is made in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged. A detailed definition of my own achievements of that of my cooperation partners and of implementation achievements, in the course of my study and supervised theses have been provided explicitly in Appendix B.

Suparna Mitra
Tübingen, October 2010

This thesis was completed in the period from July 2007 until September 2010 in the *Algorithms in Bioinformatics* group in University of Tübingen under the supervision of Prof. Dr. Daniel Huson.

In accordance with the standard scientific protocol, I will use the personal pronoun "we" to indicate the reader and the writer or (as explained in Appendix B) my scientific collaborators and myself.

Dedicated to

Maa, Baba

&

Sanjit

"I am enough of an artist to draw freely upon my imagination. Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world."

– Albert Einstein

Acknowledgments

It is a pleasure to thank the many people who supported me during tough times in the Ph.D. pursuit and made this thesis possible.

It is difficult to overstate my gratitude to my Ph.D. supervisor Prof. Dr. Daniel H. Huson for introducing me to the astonishing new world of metagenomics and for guiding me through my research. He has taught me, both consciously and unconsciously, how good science and research can be. I appreciate all his contributions of time, good ideas, encouragement, advice and funding to make my Ph.D. experience productive and stimulating. I will never forget the freedom of research that is possible in *Algorithms in Bioinformatics* group; many thanks to him for providing such an excellent working environment. The joy and enthusiasm he has for his every work is contagious and motivational for me. I specially want to thank his will and Tübingen University funding, which made possible for me, to attend several international conferences, to communicate with leading scientists and to visit many places around the globe. Many thanks to Elke also for providing a friendly and homely atmosphere; I always enjoyed nice grill-parties at her place and our refreshing cooking club.

I also owe a huge debt of gratitude to my co-advisor and collaborator Prof. Jack A. Gilbert for his candid support and many fruitful discussions, and for providing several datasets, biological explanations, new ideas for improving my work.

Special thanks goes to Dr. Kay Nieselt for many helpful discussions and for providing expert comments towards the writing and improvement of this thesis. Without her motivation this thesis could never achieve the present quality.

Our department and group members have contributed immensely to my personal and professional time at Tübingen, being a source of friendship as well as good advice, discussion and collaboration. Many thanks to our group secretary Marine, for her endless help in all official and unofficial matters throughout last three years. It is really a pleasure to see her ever-smiling face everyday. Special thanks to our system administrator Jan for his sound help in different computer and cluster related matters. I will remember Celine for helping me with LaTeX problems in writing this thesis and great time we spend together during our coffee and lunchtime discussions. I would

like to acknowledge current workmates: Juliane, Nico; and former group members: Daniel, Alexander, Regula, Johannes and Christian for many fruitful discussions, help, companionship and for providing a friendly atmosphere.

Research presented in this thesis was largely an outcome of different collaborative efforts involving essential contributions from many individuals. First of all I would like to acknowledge Dr. Bernhard Klar for his expert help and support. The statistical methods discussed in this dissertation would not have been possible without his advices. Special thanks goes to Dr. Dawn Field for her contributions of time, new ideas and encouragement. I never can forget a whole night mail-discussion with her. Her efnthusiasm and versatile knowledge in science and research is motivational for me. I also like to thank Prof. Dr. Stephan C. Schuster for many inspiring and stimulating discussions and collaboration in mammoth metagenomics project. Many thanks to Prof. Dr. Peter J. Lockhart for providing their datasets and introducing me to the coastal Fiji water Vibrio project. I also like to thank Dr. Tim Urich for his help towards my biological understanding. Not to forget many other collaborators and students for pleasant and fruitful cooperations and assistance: Max Schubach, Paul Rupek, Wei Wu, Hannelore Clément, Mario Stärk, Dr. Folker Meyer, Prof. Thomas Rattei, Dr. Ida Helene Steen, Simon Domke, Dr. Fangqing Zhao, Dr. Runar Stokke, Anders Lanzén, Andreas Wilke.

My time at Tübingen was made enjoyable in large part due to some friends who have become a part of my life. I am grateful for time spent with Avijit da (Dr. Avijit Pramanik) and Boudi (Mrs. Susmita Pramanik), who always made me feel near home.

I can never forget many people who have not only taught me science but also encouraged me to set my dreams high above and helped me to build my confidence: my high school math teacher (Mrs. Chitra Chakraborty), my MSc. supervisor (Prof. Aditya Chatterjee) among many others.

I wish to thank my family for their love and encouragement; specially my in-laws (Mr. Bhakta Ram Nayak and Mrs. Reba Nayak) for being another parents to me and for their support, care and love, without which I would not have been able to pursue this research. My deep gratitude goes to my brother (Mr. Subhra Mitra) for all the unspoken and joyous moments we shared together and to my near family members for their love, emotional support and care during all these years.

Last of all, but most importantly; I would like to express my deepest gratitude to my parents (Mr. Malay Mitra and Mrs. Indrani Mitra) for being the best parents in the world. In my each and every problem I found them beside me. They believe in me more than myself. They raised me with their immense love and supported me in all my pursuits. And most of all for my loving, supportive and encouraging husband (Dr. Sanjit Nayak) whose faithful support in my life is my strength. 'Thanking' can never express my feelings. You are the sunshine of my life.

I dedicate this thesis to my parents and my husband.

Suparna Mitra
Tübingen, October 2010

Contents

Abstract

Metagenomics is a rapidly growing field of research that aims at studying uncultured organisms to understand the true diversity of microbes, their functions, cooperation and evolution, in environments such as soil, water, ancient remains of animals, or the digestive system of animals and humans.

The recent development of ultra-high throughput sequencing technologies, which do not require cloning or PCR amplification, are fueling a vast increase in the number and scope of metagenome projects. Bioinformatics is faced with the problem of how to handle and analyze these datasets in an efficient and useful way. One goal of these metagenomic studies is to get a basic understanding of the microbial world both surrounding and within us. In a quest for better understanding our microbial community, a main challenge is to compare multiple datasets. Despite the improvements of various techniques, still there is a need for new ways of comparing metagenome datasets, and for fast and user-friendly implementations of such approaches.

This thesis introduces a number of new methods for interactively exploring, analyzing and comparing multiple metagenome datasets:

- The first is a visualization technique for the visual comparison of many large metagenomes in the tree hierarchy;
- the second includes statistical methods for highlighting the significant differences in a pairwise metagenome comparison; and
- the third is a novel approach for visualizing the relationships between multiple metagenome samples combining the use of taxonomic/functional analysis, ecological indices and non-hierarchical clustering to provide a network representation between different datasets. Importantly, the networks provide both the visual definition and metric quantification of non-rooted relationships between metagenome samples.

The methods are illustrated using several published data sets of different types, and are applicable to metagenomes, metatranscriptomes and 16S ribosomal profiles. Not only designed for datasets coming from second generation sequencing platforms, these methods will be applicable to the upcoming third generation sequencing datasets also.

Most metagenome sequencing projects so far have been based on Sanger or Roche-454 sequencing, as only these technologies provide long enough reads, while Illumina sequencing has not been considered suitable for metagenomic studies due to a short read length of only 35 bp. However, now that reads of length 75 bp can be sequenced in pairs, Illumina sequencing has become a viable option for metagenome studies. To this end the performance of two second-generation sequencing technologies are compared by simulating metagenomes. The technical aspects and usefulness of paired reads and different clone length are also portrayed.

All the methods described in this thesis are implemented and freely available with the stand-alone metagenome analysis tool MEGAN.

Chapter 1

Introduction and Background

1.1 Microbes: The Small Wonders

The diversity of species on earth is high. Most of them are microorganisms. They are everywhere. For example, one microliter surface seawater has been estimated to contain thousands of different bacteria, archaea, and ten thousands of different viruses [Azam and Malfatti, 2007]. Bacteria are the main players in the microbial world, performing tasks that include everything from causing disease to fixing nitrogen in the soil. According to en estimate by microbiologist William B. Whitman, there are typically 40 million bacterial cells in one gram of soil and a million bacterial cells in a millilitre of fresh water; alltogether, there are approximately five nonillion (5×10^{30}) bacteria on Earth, forming most of the world's biomass [Whitman et al., 1998]. Microbes play a significant role in global carbon and nutrient cycling. "You can think of microbes as canaries in the coal mine, if you will, because they have an extremely rapid response time. They can double in a day or less. If we can read those responses, and how they might influence cycles, then we would have a very sensitive probe into how environmental changes are occurring" explained Edward F. DeLong [Parson, 2005]. According to Lily Whiteman, "with their mighty collective muscle, microbes control every ecological process, from the decay of dead plants and animals to the production of oxygen" [Whiteman, 2008]. Microbes that colonize the human body during birth or shortly thereafter, and they remain throughout life, are referred to as normal flora [Salyers and Whitt, 2000]. "If all of Earth's microbes died, so would everything else, including us," says Matt Kane. "But if everything else died, microbes

would do just fine". Therefore, Kane concludes, "we need microbes more than they need us". So we can easily say that microbes run the world [Whiteman, 2008].

1.2 Metagenomics

"When we try to pick out anything by itself, we find it is tied to everything else in the universe." – John Muir, 1911

Muir's quote is particularly applicable to the microbial world. It seems to be impossible to identify, classify, and research all microbes. Standard genomics try to enrich pure cultures and study them: for example the taxonomy, the genome, the genes and the pathways. However, only a miniscule fraction – most scientists estimate only about 1% of all microbes can be cultured because of their complex symbiosis with other organisms. Several times scientists have given different estimates of representative genes on Earth, but every new project demonstrated something new, a previously unfathomed repository of biodiversity. "So when you think about biodiversity, and the extent of diversity on the planet, you really get a sense of how little we know about this undiscovered world. We are at the stage of discovery where, everywhere we look, we see new species" says microbiologist Roberto Kolter [Shaw, 2007].

Genomic studies are limited. Because, single organism genome studies involves cloning of its entire genome, which is not often possible due to the interaction of a species within the community with other habitats and sometimes with the host organisms. These studies can not achieve the extend of microbial diversity. The scientific community gained new options with the development of new sequencing techniques and high throughput analysis. The pace of genomic investigations in environmental microbiology and microbial ecology is accelerating. Nowadays a sample can be obtained from a habitat and sequenced directly from the environment. The collective genomes of microbes have been termed the 'metagenome' and these environmental studies are 'metagenomic studies' or shortly 'metagenomics' [Handelsman et al., 1998]. Metagenomics can fill the gap of normal genomics, as sequence data can be obtained directly from the environment where they are with their natural habitats. This is a rapidly growing field of research that aims at studying uncultured organisms to understand the true diversity of microbes, their functions, cooperation and evolution,in environments such as soil, water, ancient remains of animals, or the digestive system of animals and humans. Although it is clear that communities of microbes play a vital role in such systems, a more detailed understanding is only beginning to emerge. In metagenomics, scientists apply the power of genome analysis to entire communities of microbes, without isolating and culturing individual organisms.

A main promise of metagenomics is that it will accelerate drug discovery and biotechnology by providing new genes with novel functions.

1.2.1 The Metagenomics Process

The first step in metagenomic studies begin with obtaining a sample from a particular environment such as marine, soil, or the human biome, extracting genetic material from ideally all of the organisms in the sample and then analyzing the DNA as a collection to gain insights on how members of the community interact, change, and perform complex functions. Typically, laboratory bacteria are then induced to take up and replicate the fragments of the extracted DNA, creating a 'library' containing the genomes of all the bacteria and archaea found in the sampled environment (new technologies facilitate studying a community's DNA directly, bypassing the creation of a library) [Committee on Metagenomics, 2007]. The DNA of the community can then be studied in different approaches. In this thesis we will focus on two main ways: sequence-based metagenomics (studying the community structure, species richness and distribution or taxonomic analysis) and function-based metagenomics (studying the metabolic potential of a community or functional analysis).

1.2.2 Metagenomics Projects

Environmental genomics projects originated from Norm Pace's cultivation independent survey approach for studying natural microbial populations. [Pace et al., 1985, Olsen et al., 1986]. Later advances in DNA sequencing technology, development of improved cloning vectors and streamlined cloning techniques helped the recovery and sequencing of large DNA inserts from naturally occurring microbes. Since 1998, after Jo Handelsman [Handelsman et al., 1998] used the term metagenomics for the first time, many different metagenomics projects have been reported providing remarkable insights into diverse ecological systems. Advances in bioinformatics, refinements of DNA amplification, and the rise of computational power have greatly aided the analysis of DNA sequences recovered from environmental samples. In 2002, Mya Breitbart, Forest Rohwer, and colleagues used environmental shotgun sequencing to show that 200 liters of seawater contains over 5000 different viruses [Breitbart et al., 2002]. After that two remarkable studies "changed the landscape significantly, and showcase the power and potential of shotgun sequencing approaches to characterize natural microbial populations" [DeLong, 2004]. The first focused on an acid mine drainage biofilm [Tyson et al., 2004]. With only $76Mbp$ of sequences Tyson and colleagues showed it was possible to assemble 'near complete' composite genomes of constituent Bac-

teria (*Leptospirillum species*) and Archaea (*Ferroplasma species*). The second spectacular study, the Global Ocean Sampling (GOS) was undertaken by Venter and colleagues. In early 2003, they circumnavigated the globe and collected metagenomic samples throughout. The pilot project, carried out in the Sargasso Sea, discovered DNA from nearly 2000 different species, including 148 types of bacteria never seen before [Venter et al., 2004]. As of 2009, Venter and colleagues thoroughly explored the West Coast of the United States. Currently they are in the midst of a two-year expedition to explore the Baltic, Mediterranian and Black Seas. The Sorcerer II GOS Expedition yielded an extensive dataset consisting of 7.7 million sequencing reads (6.3 billion bp) [Rusch et al., 2007]. "Simply on the basis of size alone, the GOS dataset is a milestone in the endeavour to understand the magnitude and scope of efforts that will be required to make sense of microbial genomic and functional diversity in the sea" [DeLong, 2007]. In addition there are many other projects devoted to marine metagenomics, see for example [Gilbert et al., 2008a, Gilbert et al., 2009, Woyke et al., 2009].

We already know that microbes inhabit the human body. In fact, every person has more than 10 times as many microbes living on and inside his or her body as they have human cells. Although most frequently associated with disease, our microbial community helps us much more than it harms us. As a result, a global initiative was started to characterize the interaction between microbes and the various parts of the human body [Turnbaugh et al., 2007]. Several results have been already reported such as, [Turnbaugh et al., 2006, Gill et al., 2006, Qin et al., 2010]. The National Institutes of Health's Human Microbiome Project (HMP) is one the most important projects designed to take advantage of metagenomic analysis to study human health. The HMP started with the mission of generating resources enabling comprehensive characterization of the human microbiota and analysis of its role in human health and disease. In May 2010, the researchers published a first genomic collection of human microbes [Human Microbiome Jumpstart Reference Strains Consortium et al., 2010].

Sequencing of the soil metagenome will represent the third major microbial sequencing effort after the human microbiome and the marine metagenome. The soil environment has a higher complexity than any other environment on Earth and a concerted international effort is required to obtain the soil metagenome [Handelsman et al., 1998, Tringe et al., 2005, Urich et al., 2008]. After a meeting in Lyon, France in December 2008, an international group of scientists from 23 countries, formed a consortium named 'TerraGenome', solely dedicated to soil metagenomics. The goal is to establish a working public international consortium for the complete sequencing of the metagenome of a reference soil [Vogel et al., 2009]. There are many other projects dedicated to soil metagenomics, for example [Schloss and Handelsman, 2006c, van Elsas et al., 2008].

Most of the studies focus on prokaryotic sequences. However, there are also studies on viral metagenomes for example [Breitbart et al., 2002, Culley et al., 2006, Williamson et al., 2008, Li et al., 2010] and many more. Metagenomics has also enriched the recently emerging paleogenomics research. While metagenomics helps us to understand our microbial community, paleogenomics unveils increasingly precise picture of our ancestral vertebrate genomes based on genome sequences and new algorithmic developments. [Hofreiter, 2008]. The method of this research is very similar to typical metagenomic approach. Many recent studies have focused on the ancient bone [Poinar et al., 2006, Green et al., 2006, Ramérez et al., 2009] or hair samples [Gilbert et al., 2007, Miller et al., 2009, Gilbert et al., 2008b] and many others.

The microbial habitats now being examined in metagenomic studies are diverse and expanding. For example typical applications include the study of the microbes in honey bee colony collapse disorder [Cox-Foster et al., 2007], the hindgut microbiata of termites [Warnecke et al., 2007], in a glacier ice metagenome [Simon et al., 2009], the bacterial metagenome of cigarettes [Sapkota et al., 2010], and many others. The Genomes Online Database (GOLD)[1], provides information regarding complete and ongoing microbial projects around the world.

1.3 Role of Statistics in Metagenomics

Metagenomics is a rapidly developing science, promising expansion towards discoveries that can help in the comprehension, cure and prevention of many diseases, in monitoring the impact of pollutants on ecosystems (e.g. for cleaning up contaminated environments) and in mining the rich genetic resource of non-culturable microbes that may lead to the discovery of new genes, enzymes, and natural products.

Statistics has a very important role to play in this because it can provide models and methods to better understand or analyze the data and phenomena in question. Statistics can help to develop innovative and efficient methods that can help to deal with the sparse sampling as well as with the analysis of data from meta-genomics, -transcriptomics, -proteomics. In words of philosopher Ian Hacking, *"The quiet statisticians have changed our world, not by discovering new facts or technical developments, but by changing the ways we reason, experiment, and form our opinions"* (quoted in [Hastie et al., 2009]).

The advances in the throughput and cost-efficiency of sequencing technology

[1] `http://www.genomesonline.org/cgi-bin/GOLD/bin/gold.cgi?page_requested=Microbial`

is fueling a rapid increase in the number and size of metagenomic datasets being generated (see 1.4.6,). Bioinformatics is facing the problem of how to handle and analyze this enormous amount of data in an efficient and useful way. This situation can be described with the words of Rutherford D. Rogers: *"We are drowning in information and starving for knowledge"* (quoted in [Branin and Case, 1998]). In order to make informed decisions, it is important for all the studies, to separate facts from speculations by applying valid statistical analyses.

1.4 DNA Sequencing and Different Technologies

1.4.1 History of DNA

We all know that Crick and Watson, together with Maurice Wilkins, won the 1962 Nobel Prize in Medicine for their discovery of the structure of Deoxyribonucleic acid (DNA). The importance of this breakthrough discovery often overshadows the fact that research into DNA had already begun 75 years before.

DNA was first observed and isolated by the Swiss physician Friedrich Miescher, in the laboratory of Felix Hoppe-Seyler at the University of Tübingen in 1869. He discovered a microscopic substance in the pus of discarded surgical bandages. Since he had isolated it from the cells' "nuclei", he named it "nuclein", a name preserved in today's designation deoxyribonucleic acid [Dahm, 2008]. In 1919, Phoebus Levene identified the four bases: adenine (A), thymine (T), guanine (G), and cytosine (C) and the sugar and phosphate nucleotide unit [Levene, 1919]. In 1937 William Astbury produced the first X-ray diffraction patterns that showed that DNA had a regular structure [Astbury, 1937]. The middle of the twentieth century witnessed some of the most fundamental discoveries in DNA research. In 1944 Oswald T. Avery, Colin MacLeod and Maclyn McCarty suggested in their landmark paper that DNA, not proteins as previously widely believed, was the carrier of genetic information ([Avery et al., 1944]). After that Erwin Chargaff discovered that the base composition of DNA varies between species, but within each species the bases are always present in fixed ratios: the same number of adenine as thymine bases and the same number of cytosine as guanine bases [Chargaff et al., 1949, Chargaff, 1951]. In 1952, Alfred Hershey and Martha Chase confirmed DNA as the genetic material [Hershey and Chase, 1952], and Rosalind Franklin and Maurice Wilkins, decided to try to make a crystal of the DNA molecule. If they could get DNA to crystallize, then they could successfully make an X-ray pattern, thus resulting in understanding how DNA works. Although the first attempt of Astbury (1937) to propose the correct structure of DNA was not successful, as Astbury's insights led directly to the

work of Franklin and Wilkins. Finally, one year later, based on that single X-ray diffraction image, Francis Crick and James Watson proposed a model the structure of DNA, which is now accepted as the correct double-helix model of DNA [Watson and Crick, 1953b, Watson and Crick, 1953a].

1.4.2 Sequencing Process

DNA molecules consist of repeating nucleotides, which are the bases of DNA. The basic principle of DNA sequencing is simple and consists of two main steps. In the first step, labeled nucleotides are inserted into copies of a DNA fragment that involves a technique called DNA amplification. In the second step, the DNA sequence is derived from the locations of the labeled nucleotides which involves separating the DNA fragments according to their lengths. This is often done by electrophoresis in a polyacrylamide gel. The base at the end of each fragment is identified, allowing reconstruction of the DNA sequence. Here we will not describe the extensive details of the different sequencing techniques, but will briefly discuss the fundamental aspects, advantages and disadvantages of them in the following.

1.4.3 1^{st}-Generation Sequencing Techniques

After the discovery of the double-helix structure of DNA, it took almost fifteen years for the first determination of a DNA sequence. When talking about 1^{st} generation sequencing techniques we usually think of the Maxam-Gilbert [Maxam and Gilbert, 1977] and - even more - of the Sanger method published independently in the 1977 [Sanger et al., 1977]. Both inventors shared the Nobel Price in chemistry in 1980 for their works [Gilbert, 1980, Sanger, 1980].

Maxam-Gilbert's technology

In 1976-1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method (also known as "chemical sequencing") based on chemical modification of DNA and subsequent cleavage at specific bases [Maxam and Gilbert, 1977]. At first Maxam-Gilbert's idea became more popular since purified DNA could be used directly, while the initial Sanger method required that each read need to be cloned for production of single-stranded DNA. Due to the use of hazardous chemicals like the extremely toxic hydrazine and because Maxam-Gilbert's technique could not be adopted for large scale automated use, more and more people switched to Sanger sequencing.

Sanger Sequencing

The first of the recent sequencing methodologies was described by Frederick Sanger in 1975 and is widely known as the "Sanger Technique" or "Chain-termination method". Sanger's method is based on the use of dideoxynucleotides (ddNTP's) in addition to the normal nucleotides (NTP's) found in DNA [Sanger et al., 1977]. This method has greatly simplified DNA sequencing.

For many years the Sanger method was the only method used in practice, and over the years it was optimized to make it cheaper and more efficient. It is still used today since it can be used without any special tools in almost every lab. For long reads up to 1000 base pairs, however, more advanced accompanying kits are necessary. The best read length, among the currently available sequencing platforms, makes it still essential for many de-novo sequencing projects. However, Sanger sequencing is still considered to be very expensive and slow. Sequencing one megabase costs $500 dollars - about 1000 times the cost of the cheapest 2_{nd} generation technique.

1.4.4 2^{nd}-Generation Sequencing Techniques

2^{nd} or next-generation sequencing (NGS) techniques apply the idea of "sequencing by synthesis". They are designed for sequencing large amount of DNA substantially faster and cheaper that the early methods by the construction of cyclic-array processes. Three platforms for massively parallel DNA sequencing read production are in reasonably widespread use at present: the Roche/454 FLX (`http://www.454.com`), the Illumina/Solexa Genome Analyzer (`http://www.illumina.com`), and the Applied Biosystems SOLiD^{TM} System (`http://www.appliedbiosystems.com`).

Roche/454 pyrosequencing

The 454 system was the first of the 2^{nd}-generation techniques. It became available commercially in 2005 by 454 Life Sciences. It amplifies DNA inside water bubbles in an oil solution, each bubble containing a single initial DNA molecule and a single primer-coated bead that the DNA can attach to and form a clonal colony (emulsion PCR). The technique is known as 'pyrosequencing' which was developed by Mostafa Ronaghi and Pål Nyrén at the Royal Institute of Technology in Stockholm. The read length is now much longer (400 bases) than those of other 2nd generation sequencing techniques, but it is significantly shorter than those obtained by Sanger sequencing. The overall read accuracy has been constantly improved over the years. [Margulies et al., 2005] reports an accuracy of

96% for the GS20 platform. Later in 2008, [Droege and Hill, 2008] reported a single-read accuracy of $> 99.5\%$ over the first 200 bases. At present (Aug 2010) Roche 454 claims an accuracy of 99% for the 400 bp reads and higher for less bases for the GS FLX Titanium Series (`http://www.454.com/products-solutions/system-features.asp`). However insertion and deletions are the most common errors since it is difficult to determine the correct number of same nucleotides added in sequence by the measured light intensity. Sequencing errors occasionally result from the misinterpretation of homopolymer runs, i.e. stretches of the same base (e.g., TTTTT or AAA). This leads to single-base insertion or deletions ("overcalling", "undercalling"), rather than to substitutions which occur rarely [Droege and Hill, 2008]. In 2007, the preparation of a paired-end library was introduced for 454 sequencing [Korbel et al., 2007].

Illumina (Solexa)

Solexa developed a sequencing technology based on reversible dye-terminators. DNA molecules are first attached to a primer on a slide and amplified so that local clonal colonies are formed (bridge amplification). In 2006, Solexa $1G$ Genetic Analyzer was launched with a claim to re-sequence a human genome for about $\$100,000$ in three months. The Solexa platform has its origins in work by Turcatti and colleagues [Turcatti et al., 2008]. The original company Solexa was bought by Illumina in 2007. Since then the platform is sold under the new name. The current sequencer model is the Genome Analyzer$_{IIe}$ claims to offer a powerful combination of accuracy, read lengths, and paired-end insert sizes at a lower per base ($2 range for a megabase) than 454 (`http://www.illumina.com`). The major disadvantage is the read lengths that are considerably shorter than 454 reads (between 36 and 125 bp). By incorporating chain-terminating nucleotides, complications regarding the homopolymer detection are avoided.

SOLiD

The SOLiD platform is sold by Applied Biosystems (`http://www.appliedbiosystems.com`). It was developed by Kevin McKernan and his colleagues at Agencourt Personal Genomics in 2006. The methodology was first described in 2005 [Shendure et al., 2005]. This technology employs sequencing by ligation. A pool of all possible oligonucleotides of a fixed length are labeled according to the sequenced position. The ligase-based approach helps to avoid polymerase induced errors, but, the length of the reads remains rather short (up to 50 bp). The rate of correctness is about 99.94% over the whole read length according to Applied Biosystems. Since insertions and especially deletions are

technically unlikely, substitutions are the most common error due to hybridized octamers with one mismatch [Mardis, 2008].

Paired-end or mate-pairs sequencing

Next-generation sequencing platforms achieved impressively low costs and high throughputs (see Table 1.1) but most of them are limited by short read lengths. To overcome this limitation, an immediate solution is paired-end tag (PET): to sequence a DNA molecule from the two ends of individual fragments and to keep track of the paired data. The term 'paired-end' refers to the two ends of the same DNA molecule, the two sequences are called 'paired-end reads'. The first description of a paired-end sequencing method can be found by [Hong, 1981]. The first published description of the use of paired ends was in [Edwards and Caskey, 1991]. After that many improvements have been published. Three main 2^{nd}-generation sequencing technologies (454, Illumina and SOLiD) support paired-end modules. Sometimes they are called 'mate pairs'. More accurately, 'paired-end' or 'mate pairs' refers to how the library is made, and then how it is sequenced. The unique 'paired-end' sequencing protocol allows the end user to choose the length of the insert ($200 - 500$ bp) and sequence either end of the insert. Whereas 'mate pair' library sequencing enables the generation of libraries with inserts from 2 to 5 kb in size.

At first Illumina was the only next-generation sequencing platform with the unique combination of short- and long-insert paired-end sequencing libraries (or clones): *short clone* libraries of an average length of 200 bp, say, and *long clone* libraries, of an average length of $2,000$ bp. The paired-end module of Illumina's Genome AnalyzerII_e enables paired-end sequencing up to 2 x 100 bp for fragments ranging from 200 bp to 5 kb (`http://www.illumina.com`). Currently Roche/454's powerful GS FLX Titanium series provide multi-span (3 kb, 8 kb, 20 kb) paired end and long shotgun reads (`http://www.454.com`). The newest SOLiD^{TM}PI System (under development) promises to provide 75 bp fragments with 2 x 75 bp mate-pair and 75 x 35 bp paired-end option (`http://www.appliedbiosystems.com`). Figure 1.1 provides a schematic view of paired-end sequencing methodology.

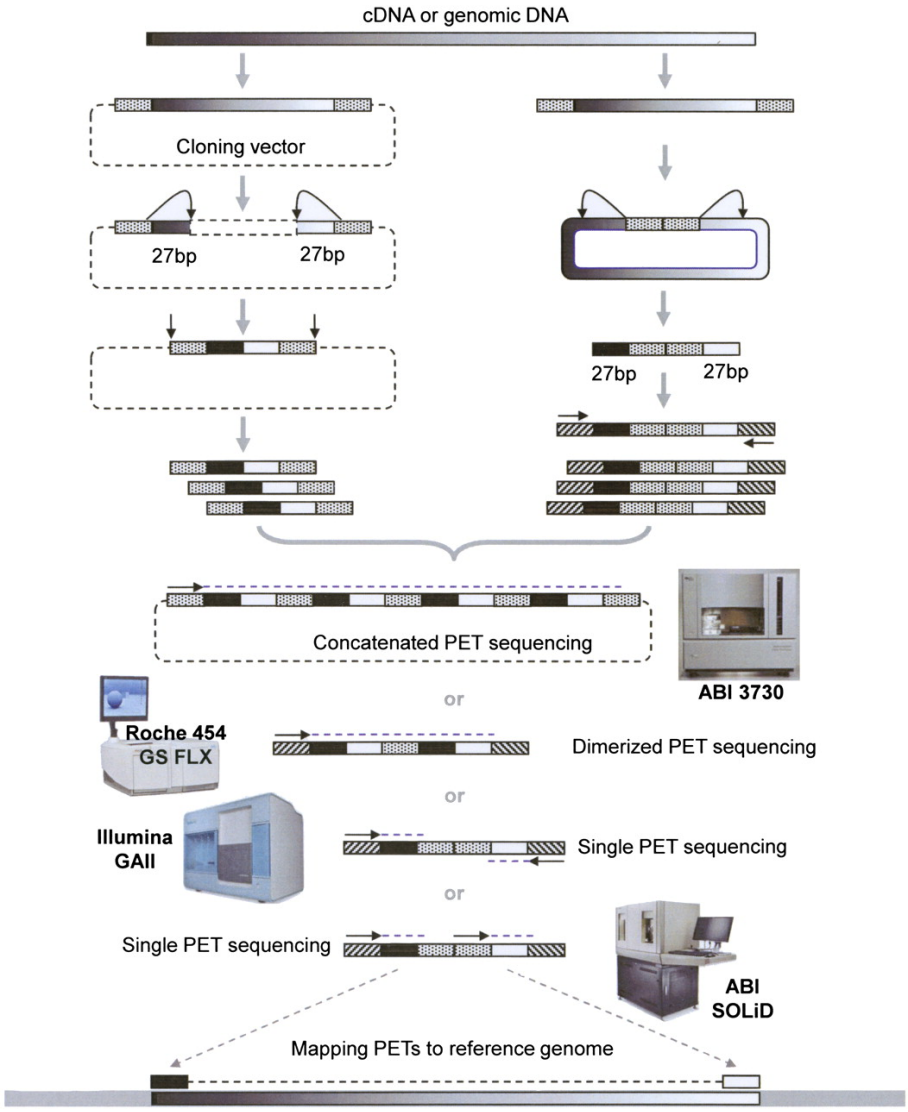

Figure 1.1: A schematic view of paired-end sequencing methodology. Figure taken from [Fullwood et al., 2009].

1.4.5 3^{rd}-Generation Sequencing Techniques

The 3^{rd} generation or next-next-generation sequencing (NNGS) technologies share a characteristic feature of the "direct" and fast sequencing of single DNA molecules, which distinguish them from the 2^{nd} generation (NGS) technologies. Currently three technology platforms comprise this third generation. Two achieve single-molecule sequencing by incorporating and detecting fluorescently labeled nucleotides: Helicos' Genetic Analysis System `http://www.helicosbio.com`, and Pacific Biosciences' Single Molecule Real Time (SMRT) technology `http://www.pacificbiosciences.com`. The third, Oxford Nanopore's nanopore sequencing uses a fairly different approach to read the DNA molecule `http://www.nanoporetech.com/`.

HeliScope:

The HeliScope is the first NNGS platform and was commercially released in 2008 by Helicos BioSciences Corporation. HeliScope's "true single-molecule sequencing" technology (tSMS) is based on the method described in [Braslavsky et al., 2003]. The read length is rather short ($24 - 32\,bp$). As of Aug 2010, helicos claimes a read length of 25 to 55 base pairs with the accuracy remaining constant regardless of the read length. Main error types are deletions (3%) and insertions (1.5%), whereas substitutions occur rarely (0.2%) (`http://www.helicosbio.com`). [Pushkarev et al., 2009] describes the first single-molecule sequencing of a human genome using the Helicscope technology. This project accomplished the sequencing in $\$48,000$, reasonably lower than previous human sequencing projects [Check Hayden, 2009],

Pacific Biosciences:

Pacific Biosciences was founded in 2004 with the goal of developing "Single Molecule Real Time" (SMRT^{TM}) DNA sequencing technology. SMRT sequencing utilizes the "Zero-mode waveguide" (ZMW), developed in the laboratory of Harold G. Craighead at Cornell University [Levene et al., 2003]. When a nucleotide is incorporated by the DNA polymerase, the fluorescent tag is cleaved off and diffuses out of the observation area where its fluorescence is no longer observable. A detector detects the fluorescent signal of the nucleotide incorporation, and the base call is made according to the corresponding fluorescence of the dye. This technology enables, for the first time, the observation of natural DNA synthesis by a DNA polymerase as it occurs. According to the company the approach is based on eavesdropping on a single DNA polymerase molecule working in a continuous, processive manner `http://www.pacificbiosciences.com`. As reported

in [Metzker, 2010], currently the average read length is $964bp$ (max: $2,805$bp). According to recent news, the CEO Hugh Martin updated that prediction and said "average read length will be 1000-1250 bases, fractionally longer than 454 or Sanger sequencing, with 5% reads between 3-5 kb". This will definitely take (meta)genomic research to the next level very soon. The company is in preparation for the commercial launch of its SMRT DNA sequencing system at the end of this year (2010).

Ion Torrent:

Ion Torrent has developed a DNA sequencing system that directly translates chemical signals (A, C, G, T) into digital information $(0, 1)$ on a semiconductor chip. The result is a sequencing system that is simpler, faster, more cost effective and scalable than any other technology available. This is an entirely new approach of sequencing that enables a direct connection between chemical and digital information. Ion Torrent^{TM} technology doesn't use light – it's the first commercial PostLight^{TM} sequencing technology. This sequencing technology requires no proprietary chemistries or optics because it's based on a well-characterized biochemical process. When a nucleotide is incorporated into a strand of DNA by a polymerase, a hydrogen ion is released as a byproduct. That hydrogen ion carries a charge, which can be detected by the proprietary ion sensor. If a nucleotide, for example a C, is added to a DNA template and a signal is detected, then it is known that the nucleotide was incorporated. Dr. Jonathan Rothberg, a pioneer of high-speed, massively parallel DNA sequencing, founded ion Torrent in August 2007. Ion Torrent is expected to launch the 'Ion Personal Genome Machine sequencer' in late 2010. The instrument will cost under $\$US100,000$ and can complete a run in about an hour.

Oxford Nanopore Technologies:

Nanopore sequencing is not a new idea, however the technology made progress during the last few years. The idea is simple: A single α-hemolysin pore, a $33kD$ protein isolated from *Staphylococcus aureus*, is build into a lipid bilayer separating two compartments. The single stranded DNA is given to one compartment and is forced through the pore following an electrical field put on the chamber filled with physiologic buffer into the other compartment. While DNA passes through, the pore is partly blocked and the ionic current is modulated in a specific way. This technique was first described by John Kasianowicz in 1996 to measure the length of a DNA fragment [Kasianowicz et al., 1996]. Similar to the SMRT technology of Pacific Biosciences, the nanopore technology has the potential to generate

long reads (up to several thousand base pairs). According to [Rusk, 2009] read accuracy is quite high (99.8%), and the error correction is expected to be straightforward. As of Aug 2010, Oxford Nanopore Technologies, has not yet disclosed the timelines for introduction of their first sequencing system.

1.4.6 Advances in Technologies

By the beginning of 1990s, only a limited number of groups were able to sequence DNA up to $100,000$ bases at extremely high costs. The race between Celera Genomes and the Human Genome Project to sequence the human genome inspired scientists and engineers to come up with automated techniques that not only speed up the process of DNA sequencing but also substantially lower its cost. DNA sequencing is now done routinely all round the world. There are now many laboratories that can sequence 100 million bases or more every year.

Next-generation DNA sequencing has started a revolution in genomics and created the opportunity for large-scale sequencing projects. To further motivate the research in sequencing technologies, the X-Prize Foundation announced the Archon Genomics X-Prize on October 4, 2006 as a joint effort of the X-Prize Foundation and the J. Craig Venter Science Foundation. The team that first sequences 100 diploid human genomes with 6 billion bp each, in 10 days for less than $\$10,000$ per genome with a coverage of 98% and with not more than 1 error per $100,000$ bases; will be awarded \$10 million prize donated by diamond prospector Stewart Blusson (`http://genomics.xprize.org`). According to J. Craig Venter, the overall probability that this goal will be ever achieved is "close to 100%" [Pennisi, 2006]. A comparison of the performance and cost of different sequencing platforms is provided in Table 1.1. At the moment Sanger sequencing and the four major NGS contenders are being used for metagenomics. Advantage of Sanger sequencing is based on the long read length and the high read accuracy (up to 99,999%). However, NNGS technologies are knocking on the door and will hit the market very soon with the promise to produce even longer reads than Sanger.

Rapid Growth in Genomic data

New sequencing technologies are accelerating the generation of biological sequencing data. The phenomenal growth of sequence data (Figure 1.2) in GenBank is unabated and challenging to manage.

Platform	Sequencing chemistry	Read length (nt)	Quality	Run time	Max. data (per run)	Machine cost (US$)	Sequencing cost (per Mb)
Sanger	chain terminator	$800 - 1000$	$10^{-4} - 10^{-5}$	1 hour	256 Kb	$350,000$	$\sim 500$$
Roche/454's GS FLX Titanium	pyro-sequencing	400	$10^{-3} - 10^{-4}$	8.4 hours	460 Mb	$500,000$	$\sim 20$$
Illumina/ Solexa's GAII	reversible terminator	$75 - 100$	$10^{-2} - 10^{-3}$	$4_{\S}, 9_{\dagger}$ days	$18_{\S}, 35_{\dagger}$ Gb	$540,000$	$\sim 0.50$$
Life/APG's SOLiD 3	cleavable probe SBL	50	$10^{-2} - 10^{-3}$	$7_{\S}, 14_{\dagger}$ days	$30_{\S}, 50_{\dagger}$ Gb	$595,000$	$\sim 0.50$$
Helicos BioScience's HeliScope	reversible terminator	$32_{\ddagger}$	10^{-2}	$8_{\S}$ days	$37_{\S}$ Gb	999,000	$< 0.50$$
Pacific Bioscience's SMRT	real-time	$1000_{\ddagger}$ -$1250_{\ddagger}$ (exp.)	NA	15 min (exp.)	$25_{\S}$ Gb	695,000	NA
SBL: sequencing by ligation; §: Fragment run; †: Mate-pair run; ‡: Average read-lengths; NA: Not available							

Table 1.1: Comparison of sequencing technologies. Information derived from [Metzker, 2010], [Kircher and Kelso, 2010] and the company websites and news reports as of Sept 2010.

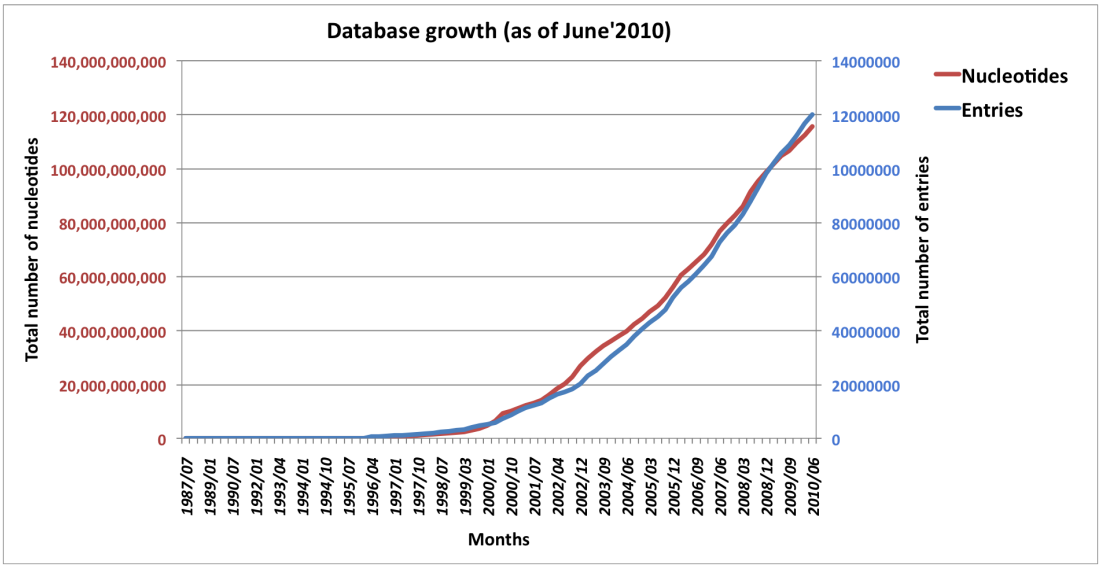

Figure 1.2: Exponential growth of genetic data due to next-generation sequencing technologies. Data taken from: `http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html` and `ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt`.

The GenBank release notes of October 2007 state that "from 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months". As of 15 June 2010, GenBank release-178.0, genetic sequence-data has reached 115, 624, 497, 715 bases, from 120, 604, 423 reported sequences (`ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt`). Upcoming sequencing technologies will accelerate this growth by several magnitudes.

Chapter 2

Main Computational Challenges

The goal of metagenomics is to understand the extent, diversity, ecology and evolution of microbial ecosystems. Characterizing the complexity of microbial communities will help us to understand ecological processes. As mentioned in Chapter 1, the developments of DNA sequencing technologies have had an enormous impact on genomic research. With metagenomics, microbial biology is experiencing a remarkable period of rapid expansion. Different sampling methods and DNA extraction techniques represent challenge for standardizing the metagenomics methods. The sheer mass of DNA fragments of a heterogenous environmental sample presents a major computational challenge.

The analysis of such sequence datasets is aimed at determining and comparing the biological diversity and the functional activity of different microbial communities. The initial aim of the computational analysis of a metagenomic dataset is to answer the following two questions:

- *Who is out there?*
  Determine the taxonomic content of a dataset by estimating which taxa are present in which relative proportions. Additionally, determine the presence or absence of key species of interest.

- *What are they doing?*
  Determine the functional content of a dataset by estimating what types of genes are present in which relative proportions. Additionally, determine which metabolic pathways of interest are supported.

Improvements in novel methods, culturing techniques, and physical separation methods, along with the generation of complete genome sequences for model microorganisms, and in some cases the assembly of whole genomes, are necessary

to interpret metagenomic sequence data. Computationally, species identification relies on the use of reference databases or reference phylogenies that contain sequences of known origin and gene function. The most prominent databases are the NR and NT databases [Benson et al., 2005]. Unfortunately, substantial database biases toward model organisms present a major hurdle for metagenomic analysis, and in a typical metagenome dataset as much as 90% of the reads may exhibit no similarity to any known sequence.

There is a need for new computational tools to solve these problems. As an initiative, in 2007, our group published and released the first stand-alone analysis tool for metagenomic data, called MEGAN (MEta Genome ANalyzer) [Huson et al., 2007]. Initially, the aim was to provide a tool for studying the taxonomic content of a single dataset. In Chapter 3, details about taxonomic and functional analysis using MEGAN are presented.

As already described in Section 1.4, the recent development of new, less expensive, ultra-high throughput sequencing technologies [Margulies et al., 2005, Bentley, 2006] that can produce huge numbers of DNA reads at an affordable cost, has boosted the number and scope of metagenomic sequencing projects. It has resulted into a dramatic increase in the volume of sequence data that must be analyzed. The analysis of metagenomic datasets is an immense conceptual and computational challenge. First two basic computational problems in metagenomics are to estimate the taxonomic content and the functional content of a given dataset. A further task is to compare the contents of different metagenomic datasets. The basic question that comes in mind is:

- *How does content of different metagenomes differ?*
  Compare the taxonomic and functional content of multiple datasets in an attempt to correlate significant differences of the genomic content to environmental differences.

Taxonomic and functional differences between metagenomic samples can help us to understand the influence of different biological factors on interaction of microbial life in a wide range of habitats. There are a number of different systems and resources for metagenome or similar analysis. These are offered in the form of databases, web portals, web services, small packages and basic stand-alone programs [Overbeek et al., 2005, Krause et al., 2008, Markowitz et al., 2006, Markowitz et al., 2008, von Mering et al., 2007, McHardy et al., 2006, Dutilh et al., 2008, Seshadri et al., 2007, Teeling et al., 2004, Meyer et al., 2008].

These resources are mainly focused on the analysis of individual metagenomes and currently do not have the capacity for rapid and highly-interactive comparison of multiple datasets. Furthermore, many of these resources allow taxonomic analyses. In our experience currently only the MG-RAST web

server [Overbeek et al., 2005, Meyer et al., 2008] provides a readily useable service for the analysis of a new metagenomic dataset. Moreover, while web portals are attractive because they offer large computational resources for data analysis, some scientists have concerns about uploading their unpublished data to a website.

The initial motivation for this work was to develop improved methods for metagenome comparison, while providing a freely available software tool. Interest for pursuing this study also grew from a desire to apply statistical methods to the field of metagenomics. Several tools or packages have been developed in recent years for comparing communities considering statistical aspects [Singleton et al., 2001, Schloss et al., 2004, Schloss and Handelsman, 2005, Schloss and Handelsman, 2006a, Lozupone et al., 2006, Schloss and Handelsman, 2006b]. However, as described in [Schloss et al., 2009], "a number of limitations will affect their use as sequencing capacity increases and studies become more complex". An implementation of these algorithms while focusing to overcome their limitations for handling large data can be found in [Schloss et al., 2009].

Still all of "these methods aim to assess **whether**, rather than **how** two communities differ" [White et al., 2009]. 'How' is important to understand the contribution of microbes to the community. Until the beginning of our research in 2007, there was only one published work that applied statistical method to addressed this question [Rodriguez-Brito et al., 2006].

All the above mentioned tools including [White et al., 2009] can only be used to the data obtained from 16S rRNA surveys and are not suitable for random high-throughput metagenome projects. In a similar time of pursuit of our study other applications have been developed for the statistical analysis [Kristiansson et al., 2009] and for assessing differences between groups of metagenomes [Gianoulis et al., 2009]. Still there was a lack of readily available tool for complete analysis and interactive comparison of metagenome samples.

Our intention was to provide a software tool that can compare metagenome datasets visually and statistically to detect significant differences in occurrence of taxa between two metagenomes. The methods are described in Chapter 4. In contrast to existing software our approach is able to provide p-values information for each node in a pairwise comparison of metagenomes [Mitra et al., 2009]. Following our path later in 2010, [Parks and Beiko, 2010] wanted to look at this question more closely considering biological relevance.

Later on, we moved forward with a more complex and obvious question:

- *How to compare multiple datasets?*
  Compare multiple metagenome datasets simultaneously.

This question of multiple comparison is first addressed [White et al., 2009] for a clinical setting. However, their method is limited to the comparison of two treatment populations each comprising multiple samples. At the beginning of our study there was no method or tool available to compare different multiple datasets simultaneously. Our method brings together established approaches from three different domains (metagenomics, ecology and phylogenetics) in a novel way. Chapter 5 outlines this algorithmic approach. This method provides a network representation of the relationships between different datasets and is applicable to metagenomes, metatranscriptomes as well as 16S ribosomal profiles [Mitra et al., 2010a].

As described in Section 1.4, many different sequencing techniques are available, with different pros and cons. That leads to another question:

- *Which technology is most suitable for a given metagenomic project?*
  Comparison of available technologies based on their performance and cost.

To answer this question, in Chapter 6 the performance of two second-generation sequencing technologies is compared by simulating metagenomes. The technical aspects and usefulness of paired reads and different clone length are also portrayed.

In Chapter 7 we apply the introduced analysis methods in four diverse metagenome projects.

Chapter 8 concludes the topics of this thesis and reviews the achievements of this work in the context of current research and developments in the field of metagenomics.

Chapter 3

Metagenome Analysis using MEGAN

In this chapter we describe the taxonomic and functional analysis of metagenome sample using MEGAN with an objective to answer the previously mentioned questions:

- *Who is out there?*
- *What are they doing?*

3.1 Getting Started with MEGAN

MEGAN is a software tool that was initially developed for taxonomic analysis, using a homology-based method to bin sequence reads. The analysis infers taxon assignments by comparing sequence reads with known sequences contained in databases. The ideal setup for performing metagenome analyses using MEGAN is a powerful desktop work-station with at least 8 GB of main memory and a large and fast local disk. The program is written in Java and requires a JRE version 1.5 or newer. Installers for all major operating systems are available from: `www-ab.informatik.uni-tuebingen.de/software/megan`.

Pre-processing of sequence reads: Given a file of DNA sequence reads obtained by sequencing an environmental sample using environmental shotgun sequencing [Mardis, 2008, Shendure and Ji, 2008], the first computational step is to compare the reads against one or more reference databases using a pairwise alignment tool such as BLAST (a sequence similarity search

tool) [Altschul et al., 1990]. This is usually the computationally most demanding step of any computational analysis of metagenomic data. For example, one giga-base of sequence requires on the order of 10,000 CPU hours for a BLASTX comparison against the NCBI-NR database. Typical analyses are a comparison against the NCBI-NR [Benson et al., 2005] database using BLASTX [Altschul et al., 1990], against the NCBI-NT [Benson et al., 2005] database using BLASTN [Altschul et al., 1990], or against one or more whole-genome sequences using BLASTZ [Schwartz et al., 2003]. However, MEGAN is not tied to any particular comparison method or database.

MEGAN needs to parse the header lines associated with entries in the reference database to extract taxonomic and functional information. Thus the program requires that input is provided either in the BLAST 'standard format' (plain text, `-m 0`) or the 'XML format' (`-m 7`), but not the 'tab delimited format' (`-m 8`).

Upon launch, MEGAN first loads its own version of the NCBI-taxonomy and then displays the first three levels of the taxonomy (Figure 3.1). Once this step

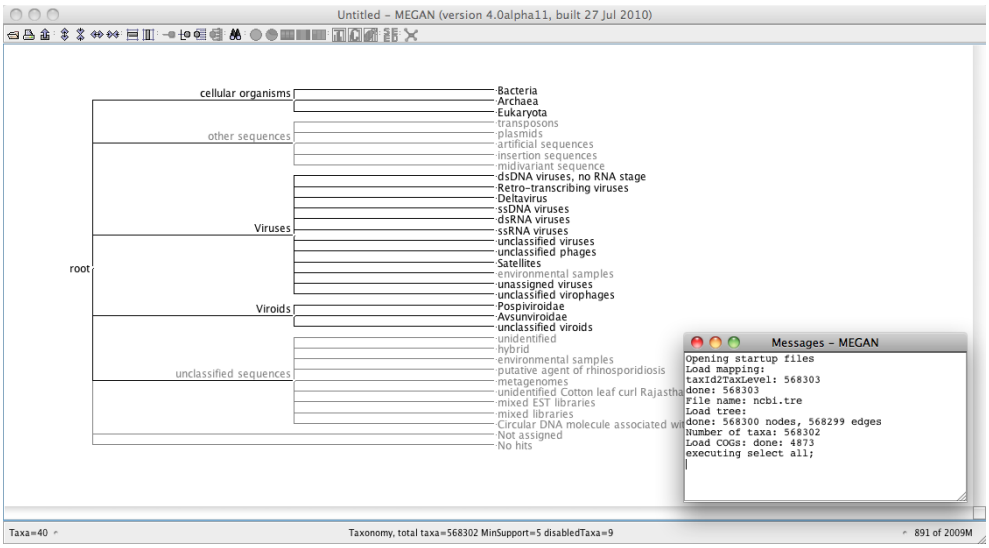

Figure 3.1: MEGAN window after launch with its own version of NCBI taxonomy.

is completed, the user can start a new analysis by importing a BLAST file using the 'Import from BLAST' option (Figure 3.2). MEGAN will parse the BLAST file (and the reads file, if present) and will then perform an initial taxonomic and functional analysis of the data. All reads, matches and results of the analysis are saved in an "RMA" file. (RMA stands for Read-Match-Archive. This is a compressed binary format especially designed for metagenomic data. The size of an RMA file is typically about 20–30% of the size of the original reads and

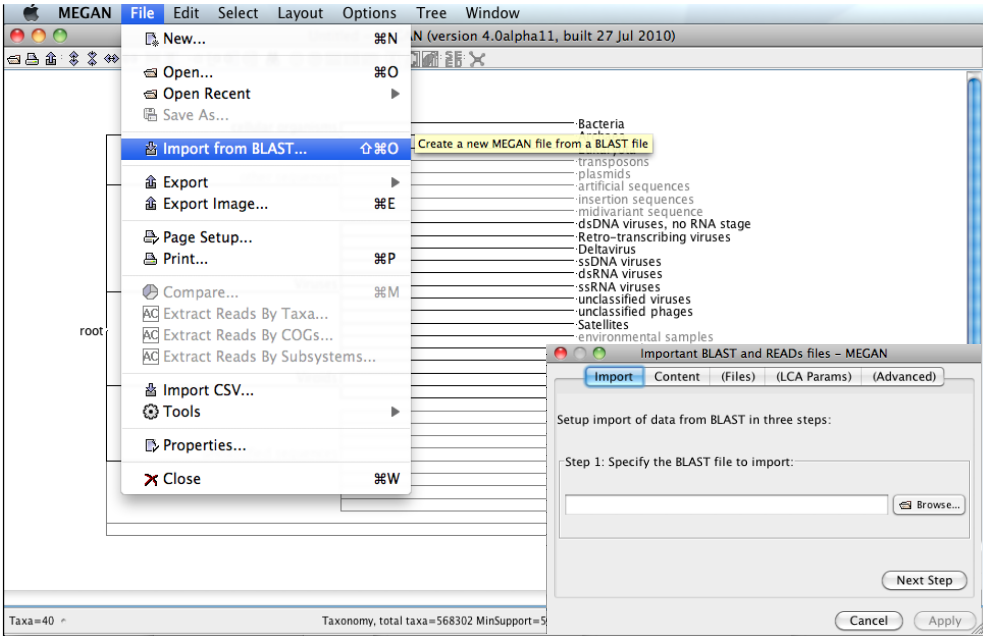

Figure 3.2: Importing BLAST file in MEGAN.

BLAST file.) The initial analysis of a dataset by MEGAN can take a number of hours and may require a lot of computer memory, depending on the size of the dataset. However, once the initial analysis has been completed, opening and working with multiple RMA files is very fast and memory efficient.

Example dataset under consideration: In this chapter, we will illustrate the metagenome analysis process using a published mouse gut dataset obtained using random shotgun sequencing, as a running example. In more detail, this is the "obese mouse" dataset (ob1: 677,384 reads) from a study reported in [Turnbaugh et al., 2006].

3.2 Taxonomic Analysis

The diversity of the microbial world is believed to be huge. However, only about 6,000 microbial species have been named [Kuever et al., 2005] and many of these are represented by only a few genes, at most, in public sequence databases. Moreover, current databases are biased toward organisms of specific interest and were not explicitly populated to represent a wide sampling of biodiversity. For this reason, taxonomic analysis currently cannot be based on high similarity sequence matching, but rather depends on the detection of homologies using quite sensitive methods.

One approach is to use *phylogenetic markers* to distinguish between different species in a sample. The most widely used marker gene is the SSU rRNA gene; others include RecA, EF-Tu, EF-G, HSP70 and RNA polymerase B (RpoB) [Venter et al., 2004]. Advantages of this approach are that such genes have been studied in detail and for some there are high quality phylogenies available that can be used as a reference to place reads from a metagenomic dataset. However, this approach is not unproblematic: On the one hand, the use of "universal" primers to target specific genes suffers from the problem that such primers are not truly universal and so only a portion of the true diversity is captured [Wu and Eisen, 2008]. On the other hand, while the use of a random shotgun approach can overcome this problem, less than 1% of the reads in a random shotgun dataset will correspond to commonly used phylogenetic marker genes [von Mering et al., 2007], which seems very wasteful, as 99% percent of the reads will remain unused (and unclassified).

Moreover, the goal of taxonomic analysis is not only to provide an estimation of the types of organisms present in a sample, but also to corral the sequence reads by taxonomic identity to facilitate further analysis, for example to study the GC content or to attempt the assembly of particular genomes.

Our approach is to compare reads against the NCBI-NR database (or some other appropriate database) to find homologous sequences, thus making use of the fact that homologies are easier to detect on the protein level. For the abovementioned reason that current databases provide only a poor coverage of the true diversity of organisms, we treat all sequence matches of high significance as equally valid indications that the given read represents a gene that is present in the corresponding organism. In more detail, we place each read on the lowest common ancestor (in the NCBI taxonomy) of all the organisms that are known to contain the gene present in the read. So, in essence, the placement of a read is governed by the gene content of the available reference genomes and thus we will refer to our method as the *LCA-gene content* approach.

An attractive feature of this "LCA-gene content" approach is that it is inherently conservative and is more prone to error toward non-informative assignments of reads (to high-level nodes in the taxonomy) than toward false-positive assignments (placing reads from one species onto the node of another species). In particular, genes that are susceptible to horizontal gene transfer will not be assigned to any of the participating species, as long as more than one is hit in the reference database. MEGAN uses the NCBI taxonomy to bin all reads of a given metagenome dataset. The NCBI taxonomy provides names and IDs for over 568,000 taxa, including approximately 287,000 eukaryota, 28,000 bacteria and 62,000 viruses [1]. The species are hierarchically classified at the levels of:

[1] Visit `http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi` for updated numbers.

superkingdom, kingdom, phylum, class, order, family, genus, and species (and some unofficial clades in between like groups, subspecies).

To perform a taxonomic analysis, MEGAN places each read of a given dataset onto one of the taxa (or "nodes") of the NCBI taxonomy, based on the BLAST matches provided for the read. If a read has significant matches to sequences in more than one species, MEGAN will assign the read to the "lowest common ancestor" (LCA) node of the species in the taxonomy, thus using a simple LCA algorithm [Huson et al., 2007].

We will now describe how to perform a taxonomic analysis with MEGAN using the obese mouse data from the above mentioned mouse gut datasets [Turnbaugh et al., 2006]. The first step is to compare the 677,384 reads against

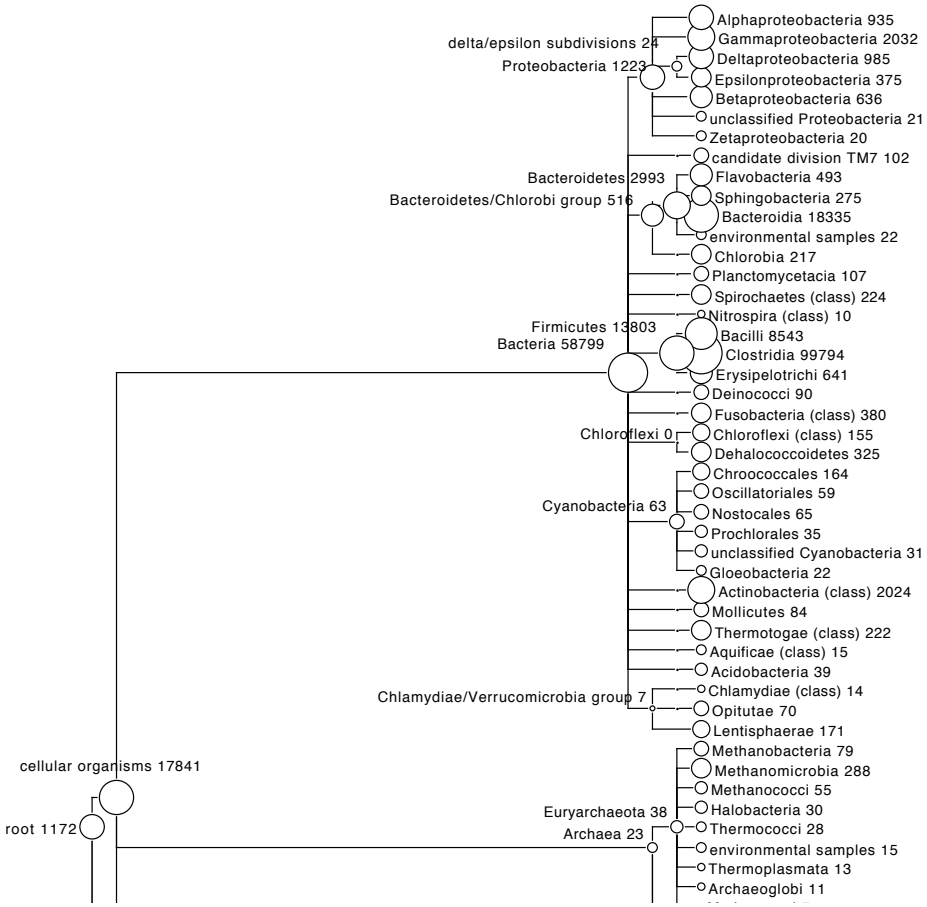

Figure 3.3: Figure shows a part of taxonomic analysis of 677,384 reads of the obese mouse gut dataset [Turnbaugh et al., 2006] by MEGAN. Each circle represents a node of the NCBI taxonomy and is labeled by the number of reads assigned to the node. The sizes of the circles are logarithmically scaled to visually represent the numbers.

the NCBI-NR database using BLASTX. The result of this step is a file of size

10.57 GB consisting of 311, 755, 132 BLAST alignments of reads to sequences in the database. The second step is then to process the BLAST file and reads using MEGAN to obtain an RMA file `Obese.rma` of size 3.44 GB, in which all reads have been taxonomically placed using the LCA algorithm. MEGAN can then be used to interactively explore the dataset. In Figure 3.3 we show the assignment of reads to the NCBI taxonomy, down to the taxonomic rank of class (a part of the image is displayed here). In this figure, each node is labeled by the number of reads assigned to it and the size of the nodes are scaled logarithmically to represent the number of reads. The program provides the exact numbers of reads assigned to any given node, and the number of hits in to any nodes in the subtree rooted at the node. Moreover, with different chart tools one can visualize the distribution of the assigned reads. The program also allows one to interactively inspect the assignment of reads to a specific node, to drill down to the individual BLAST hits that support the assignment of a read to a node, and to export all reads (and their matches, if desired) that were assigned to a specific part of the NCBI taxonomy.

3.3 Functional Analysis

Initially, MEGAN was designed only to provide a taxonomic analysis of a dataset. For the functional analysis the first approach was based on NCBI's Clusters of Orthologous Groups (COG) classification [Tatusov et al., 1997, Tatusov et al., 2003], which was developed to cluster annotated genes into functionally related groups. As an example, in Figure 3.4 we display a chart of abundances for several COG categories for the obese mouse gut dataset as computed by MEGAN. This chart corresponds closely to a similar analysis reported in [Turnbaugh et al., 2006]. A COG analysis is easy for MEGAN to perform when given the result of a BLAST comparison against the NCBI-NR database because the representatives of the COG families are present in that database. However, while COGs are still used in publications, the COG classification in no longer curated and has thus become stale.

As a next step towards a sophisticated functional analysis, MEGAN used the Gene Ontology (GO) [Ashburner et al., 2000] as a classification structure for binning environmental sequences. The Gene Ontology provides three sets of structured vocabularies (ontologies): *Molecular function*, *biological process* and *cellular component*. GO is regularly updated and widely used in many biological databases, gene expression and annotation studies, and has been referred to as "the most successful example of systematic description of biology" [Rhee et al., 2008]. The GO ontology currently contains around 28, 000 terms. As the large number of terms and relationships can be unwieldy for some ap-

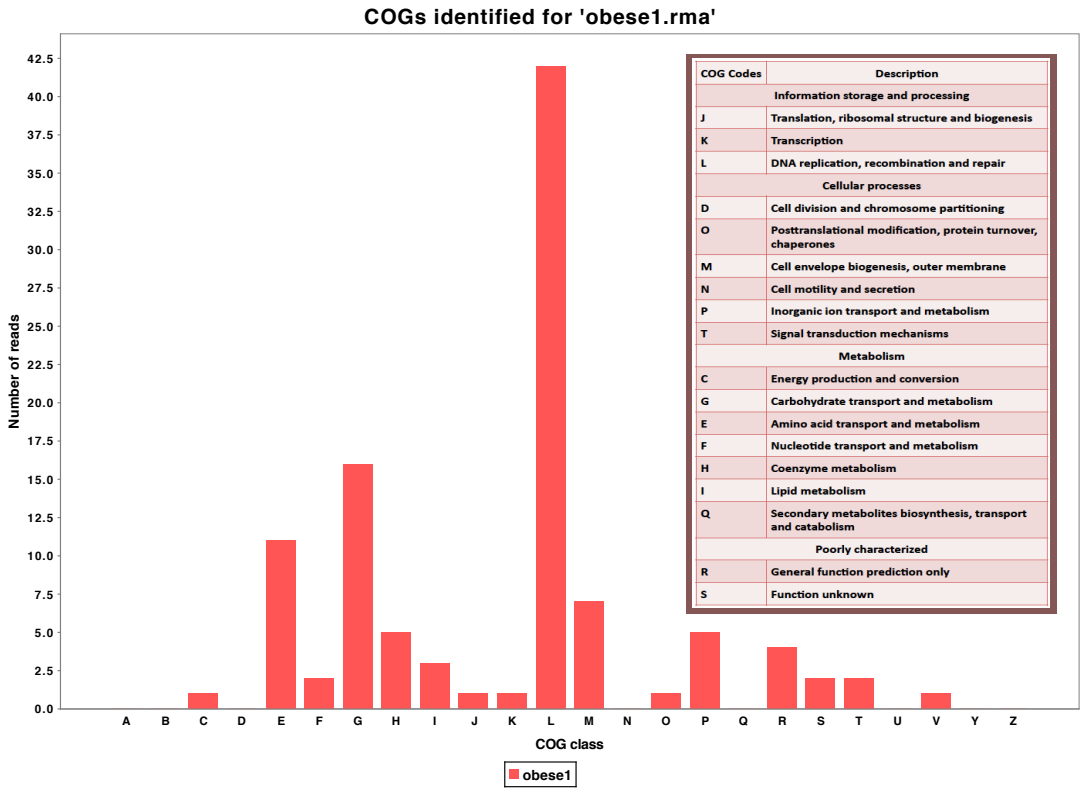

Figure 3.4: Abundances of several COG categories for the obese mouse gut dataset with description of COG categories. The result complies with the fact that presented in [Turnbaugh et al., 2006].

plications, the Gene Ontology Consortium also provides reduced, summarizing ontologies called 'GO-slims'. At present, four slims are provided on the GO website: the *generic GO slim*, *GOA and whole proteome analysis*, *plant GO slim* and *yeast GO slim*. Additionally, a prokaryotic subset of all GO terms is also provided. MEGAN has the ability to summarize results based on GO slims. MEGAN uses an LCA-like approach to assign each read to at most one node in each of the three GO onotologies. Since GO identifiers are not reported directly in a BLAST result file, MEGAN employs ref-seq identifiers [Pruitt et al., 2009] and a lookup-table to assign GO terms to read matches. In Figure 3.5 we show a part of a MEGAN analysis focused on the *metabolic process* ontology and *generic GO slim*, for the obese mouse dataset. As in the case of a taxonomic analysis, in a functional analysis the user can search for specific nodes of interest, view all reads (and their matches) assigned to a given node, or save all reads (and their matches, if desired) assigned to a given node.

We now believe that the SEED classification [Overbeek et al., 2005] is more

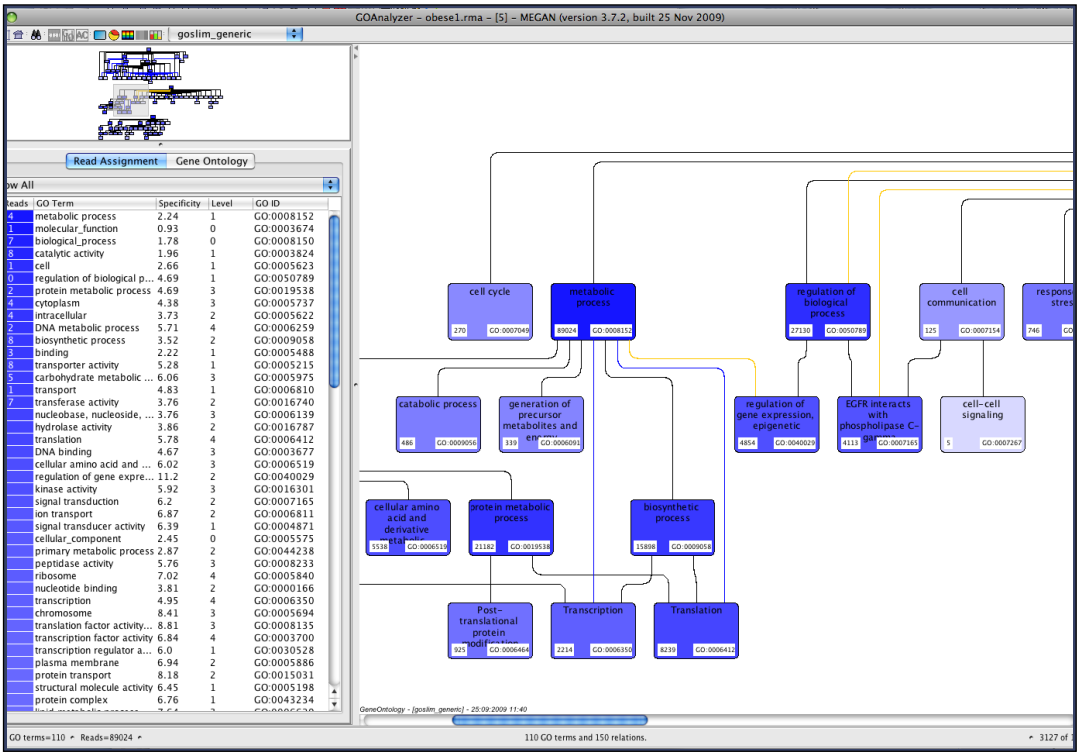

Figure 3.5: Example of a functional analysis, showing a part of the *metabolic process* ontology of an obese mouse gut dataset. The GOAnalyzer window is divided into three parts: at the top-left an overview panel shows the whole GO graph, on the left all the GO terms are listed and on the right the three GO ontologies are displayed in separate subgraphs.

powerful and more popular for functional analysis and our SEED Analyzer will replace the GOAnalyzer in version 4 of MEGAN.

3.3.1 Functional Assignment Based on MEGAN-SEED

The next major release of MEGAN, version 4, uses the SEED classification [Overbeek et al., 2005] for functional analysis. In this classification, genes are assigned to functional roles and different functional roles are grouped into subsystems. The SEED classification can be represented by a rooted tree whose internal nodes represent the different subsystems and whose leaves represent the functional roles. Note that the tree is "multi-labeled" in the sense that different leaves may represent the same functional role, if it occurs in different subsystems. The current tree has about 10,000 nodes.

To perform a functional analysis, MEGAN assigns each read to the func-

tional role of the highest scoring gene in a BLAST comparison against a protein database. MEGAN provides a hierarchical representation of SEED, where reads are mapped to SEED subsystems using the 'seed2ncbi.gz' file from the SEED server[2]. Figure 3.6 shows a part of the functional analysis of the obese mouse gut sample. The program reports the numbers of reads assigned to each functional role.

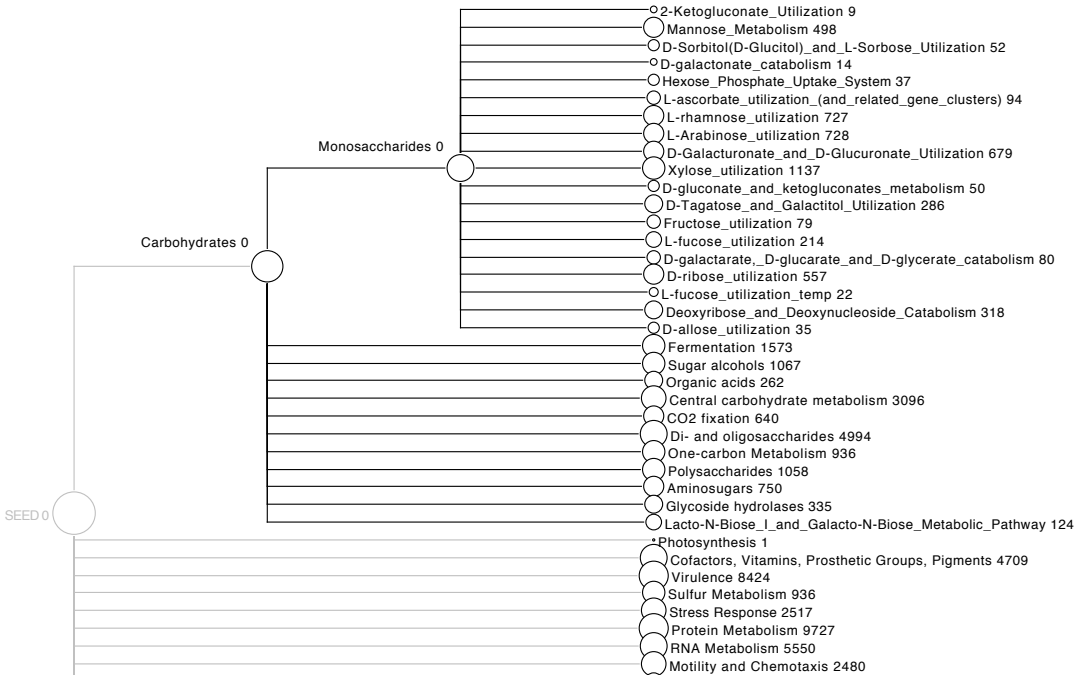

Figure 3.6: Figure shows a part of functional analysis of an obese mouse gut dataset based on the functional content using SEED subsystems, where the subtree of 'Carbohydrates' is partly shown. Each circle represents a node of the SEED subsystem and is labeled by the number of reads assigned to the node. The sizes of the circles are logarithmically scaled to visually represent the numbers.

3.3.2 KEGG analysis with MEGAN

To obtain a metabolic pathway (KEGG map), MEGAN attempts to match each read to a KO-accession number, using the best hit to a reference sequence for which a KO-accession number is known. MEGAN then calculates the number of hits to each KEGG pathway and reports these numbers to the user. The user can request to see the hits to a given pathway and an appropriate image of the

[2] `ftp://ftp.theseed.org/misc/Data/idmapping/seed2ncbi.gz`

pathway is generated by coloring the pathways based on the KEGG mapping. MEGAN's KEGG-Viewer is able to "scale" the color of the enzymes according to their read abundances (scaled in yellow to red in Figure 3.7. This color gradient

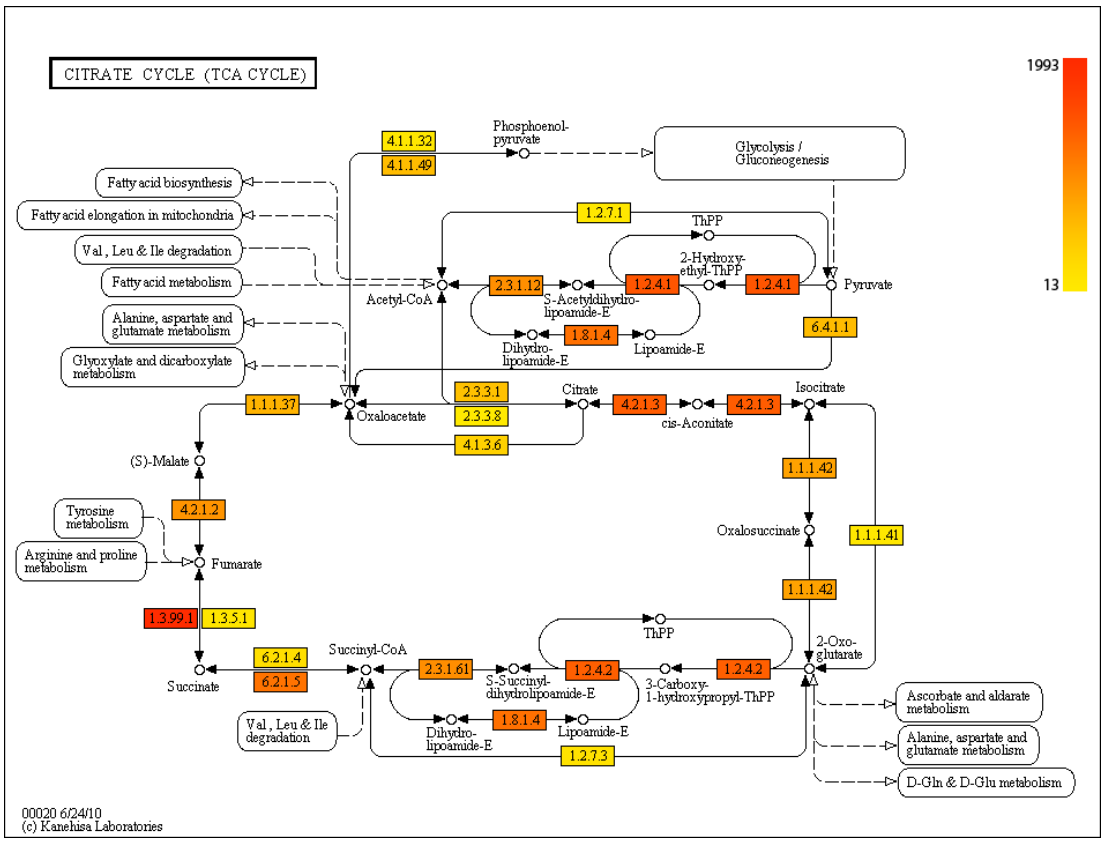

Figure 3.7: Figure shows 'citrate cycle KEGG map' as an example of a KEGG analysis with MEGAN for the obese mouse sample.

can help to visualize the enzyme kinetics as the abundance of reads assigned to an enzyme can be proportional to the turnover frequency (TOF) associated with that enzyme. Furthermore MEGAN allows one to analyze several datasets together, using different colors to show which parts of a pathway are present in which datasets.

As different genes present in different organisms in a consortium of microbes often will not operate together in a single pathway, MEGAN allows one to restrict the pathway analysis to any taxon or set of taxa in the NCBI taxonomy [Kanehisa and Goto, 2000].

3.4 Discussion

In this chapter, we have addressed the basic computational questions in a metagenome analysis and introduced the taxonomic and functional assignment techniques implemented in MEGAN that help to answer these questions. Our next goal is to develop different methods to compare the content of these metagenomes (or metatranscriptomes).

Chapter 4

Visual and Statistical Comparison of Metagenomes

In this chapter we describe a visual comparison method for multiple metagenome samples, and two statistical comparison techniques to find significant differences in a pairwise comparison of metagenomes. This chapter is committed in addressing the question:

- *How does the content of different metagenomes differ?*

Data processing: Throughout this chapter, we illustrate the techniques for comparing metagenome datasets using six published mouse gut samples, obtained using random shotgun sequencing [Turnbaugh et al., 2006, Turnbaugh et al., 2008], a human gut sample [Gill et al., 2006] and two highly different metagenome datasets: a soil sample [Tringe et al., 2005] and a marine sample [Rusch et al., 2007] obtained using Sanger sequencing technology, as a running example. Details of the samples are given in Table 4.1.

As discussed in Section 3.1, all the reads were blasted against the NCBI-NR [Benson et al., 2005] database using BLASTX [Altschul et al., 1990] and then processed by MEGAN (default settings) to obtain a taxonomic profile of each metagenome (as illustrated in 3.2). First we describe how to visually compare multiple datasets (in Section 4.1). Then, we demonstrate two methods to detect taxa for which the number of assigned reads differs in a statistically significant way in the comparison of exactly two datasets (in Section 4.2).

Datasets	Names	Number of reads	Associated studies
Obese mouse gut sample	obese1	677,384	[Turnbaugh et al., 2006]
Lean mouse gut sample	lean1	1,046,611	
Diet induced obesity in mouse gut sample	western1	9,072	[Turnbaugh et al., 2008]
	western3	10,997	
	CARB-R1	10,773	
	FAT-R1	10,681	
Human gut sample	human_gut	approx. 145,000	[Gill et al., 2006]
A subset of soil sample	soil	approx. 140,000	[Tringe et al., 2005]
A subset of marine sample	sea	approx. 150,000	[Rusch et al., 2007]

Table 4.1: Example datasets under consideration for comparison.

4.1 Visual Comparison of Metagenomes

Visualization of biological data can significantly help to reveal structures and patterns in an easy way before a complex and costly analysis. In a comparative analysis, different datasets are brought together and compared for taxonomic and functional content. To compare multiple datasets visually, we defined a *multiple-comparison tree view* which is displayed as a hierarchical tree, where each node shows the taxonomic assignments obtained for the different datasets under consideration. An important feature is the ability to interactively collapse or expand the presented tree at different levels of the taxonomy, so as to be able to start at a high-level view and then to drill down to a low-level comparison.

Visual comparison using taxonomic content:

In order to visually compare the taxonomic content of multiple datasets, we define a new *multiple-comparison tree view* in which an arbitrary number of different datasets are displayed together on a subtree of the NCBI taxonomy. To perform such a visual comparison using MEGAN 2 (or later version), first all datasets should be opened in MEGAN. Then the user should select the Compare menu item to generate a new document that contains a comparison of all datasets. The comparison can be done using either absolute counts or normalizing over all reads, the latter choice being of interest when the compared datasets are very different in size. The comparison document opens in a new window and the user can then interactively explore the comparison.

As a multiple comparison example, shown in Figure 4.1 we have taken six mouse gut datasets for comparison. In such a view, each node in the NCBI taxonomy is shown as a set of "meters" (pie charts or heat maps are also possible) indicating the number of reads (normalized, if desired) from each dataset that have been assigned to that node.

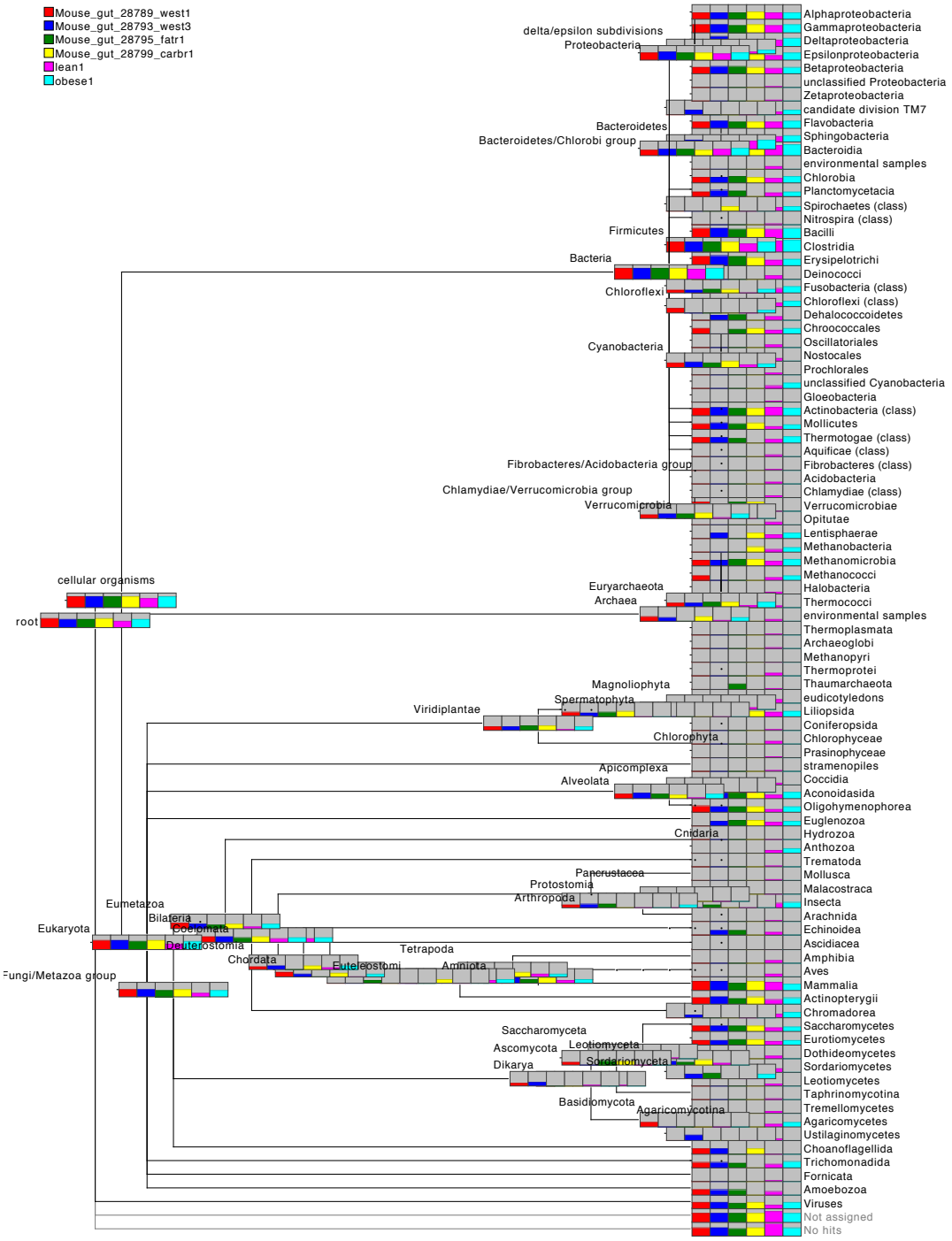

Figure 4.1: Comparative visualization of six mouse gut datasets with the NCBI taxonomy collapsed at the class level of the NCBI taxonomy.

Visual comparison using functional content:

In a similar fashion, MEGAN supports the simultaneous analysis and comparison of the functional content of multiple metagenomes using a new SEED-based tree view (see Figure 4.2). After comparing the taxonomic content in MEGAN we already got the comparison in a new window. Then choosing the 'SEED' menu directly from the taxonomic comparison window allowed us to get the functional comparison of the samples in a new window (Figure 4.2) based on their SEED content (as described in 3.3.1).

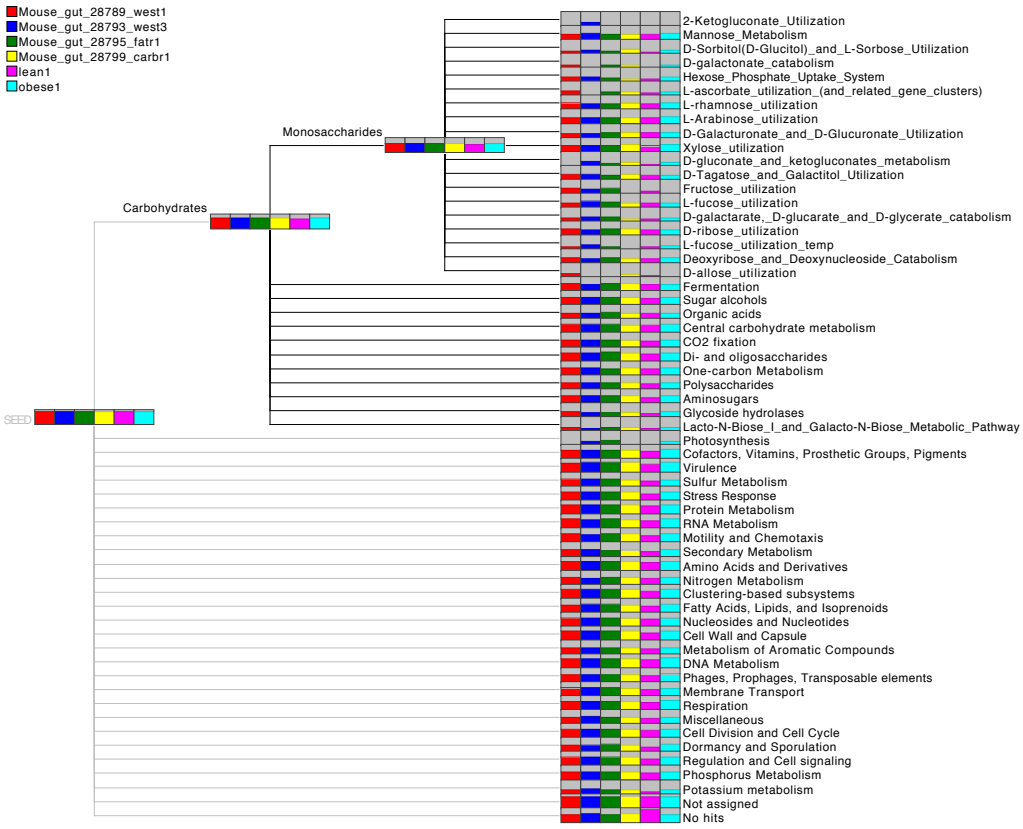

Figure 4.2: Comparative visualization of six mouse gut datasets based on their functional content using SEED subsystems. The subtree is shown for only 'Carbohydrates'.

4.2 Statistical Comparison of Two Metagenomes

Comparative visualizations are useful to obtain an impression of how two datasets differ. For a more detailed analysis, one requires information on the statistical

significance of observed differences. Our approach is based on comparing the gene content of two metagenomes.

4.2.1 Finding Significant Difference with Support Value

To address the question of metagenome comparison, initially we adapted a test developed for comparing curated subsystems in metagenomic data [Rodriguez-Brito et al., 2006], which will be described in the following.

Resampling procedures:

For a specific subsystem, the subsequent steps are performed to obtain tests with nominal level $\alpha = 10\%$.

Step 1: Repeat the following steps *1a* to *1c* exactly $n = 1000$ times.

1a: Choose randomly $N = 10000$ proteins (with replacement) from metagenome 1 which yields sample 1. Count the number of times the subsystem is seen in sample 1.

1b: Repeat step 1*a* with metagenome 2.

1c: Calculate the difference in counts between sample 1 and sample 2.

Step 2: Compute the median of the M differences obtained in *step 1*.

Step 3: Repeat the following steps *3a* to *3c* $n = 1000$ times.

3a: Choose randomly $N = 10000$ proteins (with replacement) from either metagenome 1 or metagenome 2 and count the number of times the subsystem is seen in the sample.

3b: Repeat step *3a*.

3c: Calculate the difference in counts between the two samples in *3a* and *3b*.

Step 4: Compute the lower and upper $\alpha/2$-quantile of the differences obtained in *step 3*. To this end, order the M differences from lowest to highest and determine the appropriate order statistics; if $M = 1000$ and $\alpha/2 = 5\%$, this would be the 50^{th} and 950^{th} element.

Step 5: If the median from *step 2* is between the quantiles computed in *step 4* there is no statistically significant difference between the two samples. If the median is smaller than the lower or larger than the upper quantile there is a statistically significant difference between the two samples.

This test uses a resampling technique to determine for which subsystems a difference in counts is significant. However, as mentioned in Chapter 2, this method only provide information 'whether' the two communities differ and doesn't even say anything about 'how'. As an basic initiative towards answering this question, this method can be extended by defining a *support value* as the proportion of deviation given by,

$$SV = \pm \frac{2|M - P_{50}|}{P_{95} - P_5} \tag{4.1}$$

based on the average difference M of pairs of values sampled from the two different datasets and the percentile values P_x obtained by resampling from the *same* dataset. A positive support (proportion of deviation) indicates that the difference is in favour of the first metagenome, whereas a negative sign indicates the opposite. With MEGAN 2.0, it is possible to apply this test to any level of the NCBI taxonomy [Huson et al., 2009].

We have applied this test in two case studies; the first using human gut and obese mouse gut sample and the second using marine and soil sample. Figure 4.3 and 4.4 are examples of finding significant differences, computed using the support value test.

Significant differences in the comparison of human gut and mouse gut metagenomes

As a first example, Figure 4.3 shows the statistically significant different nodes in a comparison of human and mouse gut metagenomes, five of them listed in Table 4.2 at two levels ('phylum' and 'class') of the NCBI taxonomy.

Rank	1	2	3	4	5
Phylum	Actinobacteria	Firmicutes	Euryarchaeota	Chordata	Ascomycota
Support	+282.88	+115.30	+30.93	−10.12	−6.96
Class	Actinobacteria	Clostridia	Methanobacteria	Mollicutes	Bacilli
Support	+282.70	+110.21	+87.0	+46.66	+25.01

Table 4.2: The five most different nodes with respect to the support values in a comparison of a human gut metagenome [Gill et al., 2006], and obese mouse gut metagenomes [Turnbaugh et al., 2006]. A positive support (proportion of deviation) indicates that the difference is in favour of the human gut dataset, whereas a negative sign indicates the opposite.

As expected, Actinobacteria are more dominant in the human gut, manifested through a high abundance of *Bifidobacterium longum*, *Bifidobacterium adolescentis* and *Collinsella aerofaciens ATCC 25986*. All three species are known

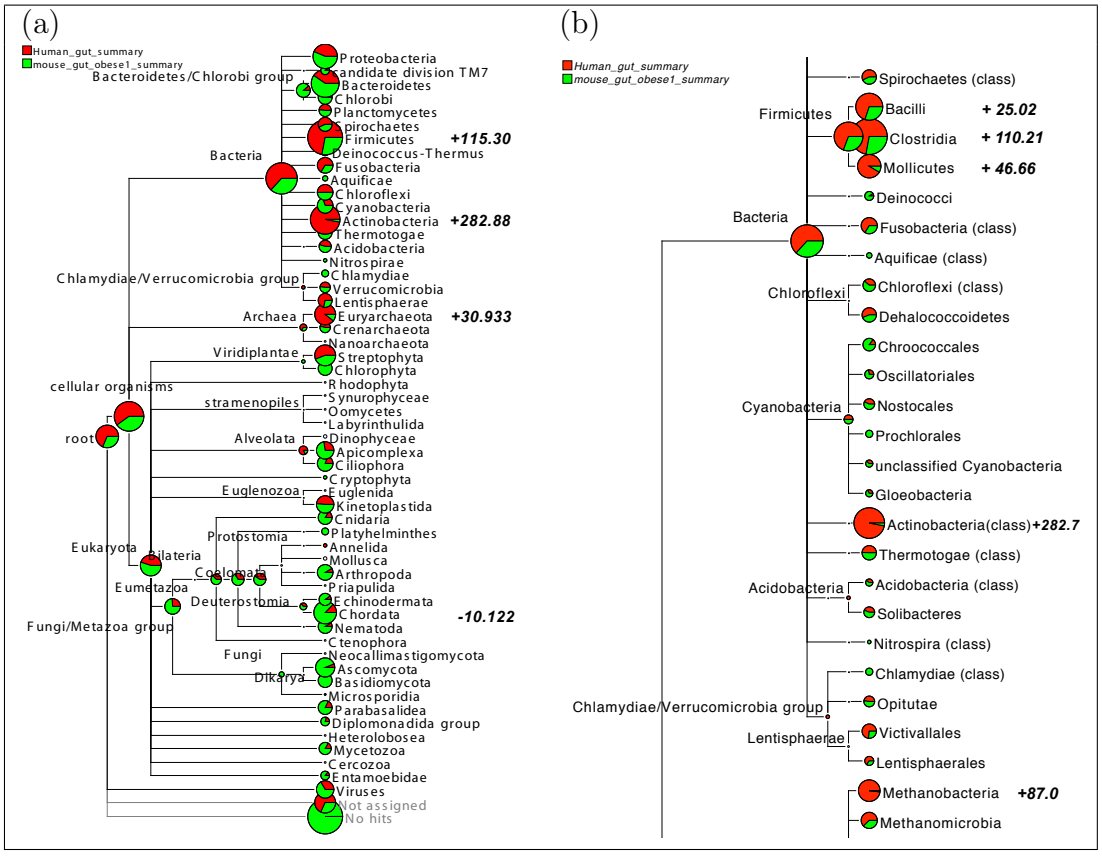

Figure 4.3: Two comparative tree views of a human gut metagenome [Gill et al., 2006] shown in red and a mouse gut metagenome [Turnbaugh et al., 2006] shown in green, as computed by MEGAN 2.0, using normalized counts. In (a), an overview of the taxonomy down to the phylum level is shown, whereas in (b) a part of a class-level analysis is displayed. In bold the support values as listed in Table 4.2 are shown. Furthermore pie charts are used to indicate the abundance of assigned reads for two datasets at each node.

to be normal inhabitants of the human intestine. Also, Firmicutes are more dominant in the human gut, mostly in the form of Clostridia, Lactobacillales and Mollicutes. Clostridia and Lactobacillales can live in intestinal tracts of animals and humans, however it is not clear why the levels of abundance differ in the two datasets. The human dataset also contains *Eubacterium dolichum DSM 3991* whose presence has previously been established by its isolation from the human gut flora. *Mesoplasma florum* is considered a commensal strain in humans and an animal parasite.

A striking contrast between the two datasets also seems to be the high abundance of Euryarchaeota/Methanobacteria. As previously reported, the main representative of this group is *Methanobrevibacter smithii*, a well-known archaeal

inhabitant of the human gut, see [Gill et al., 2006, Eckburg et al., 2005].

The fourth most significant phylum is 'Chordata'. However, in our experience, the class of Chordata is always problematic in this type of metagenomic analysis. This is most likely due to the high complexity and large sequence space covered by higher eukaryote and especially vertebrate genomes. This is further aggravated by database biases toward model organisms and the problem of false annotation of vertebrate genetic elements.

The amount of hits mapped to Ascomycota was significantly higher in the mouse gut probe, mostly reads assigned to yeast species like *Saccharomyces* and *Candida*. It is well known that these yeast species can be found in caeca of mouse [Wells et al., 2007] and rat [Lambert et al., 1967]. As stated in [Turnbaugh et al., 2006], the mouse gut probe was extracted from its caecum, whereas the human probe was taken from distal gut.

Interestingly, the proportion of mouse gut reads that exhibit no hits to the NR database is much higher than for the other dataset. This probably reflects the different read lengths produced by the employed sequencing technologies (Sanger for the human gut sample, 454 for the mouse one). An additional potential explanation may be that there is a bias in NR database that favors human endosymbionts and parasites.

Significant differences in the comparison of marine and soil metagenomes

As a second example, Figure 4.4 shows the statistically significant different nodes in a comparison of marine [Rusch et al., 2007] and soil [Tringe et al., 2005] metagenomes. The five most statistically significant differences in numbers of reads assigned to taxon classes in the comparison of marine and soil metagenomes are listed in Table 4.3.

	1	2	3	4	5
Phylum	Proteobac.	Cyanobac.	Acidobacteria	Chlorophyta	Chloroflexi
Support	+37.95	+33.54	−29.31	+22.67	−18.83
Class	Prochlorales	Thermoprotei	Oligohymen.	Aconoidasida	Prasinophyceae
Support	+267.33	+82.36	+52.36	+50.36	+52.33
Proteobac.: Proteobacteria; Cyanobac.: Cyanobacteria; Oligohymen.: Oligohymenophorea					

Table 4.3: The five most statistically significant different nodes with respective support values in a comparison of marine [Rusch et al., 2007] and soil [Tringe et al., 2005] datasets. A positive support (proportion of deviation) indicates that the difference is in favour of the soil dataset, whereas a negative sign indicates the opposite.

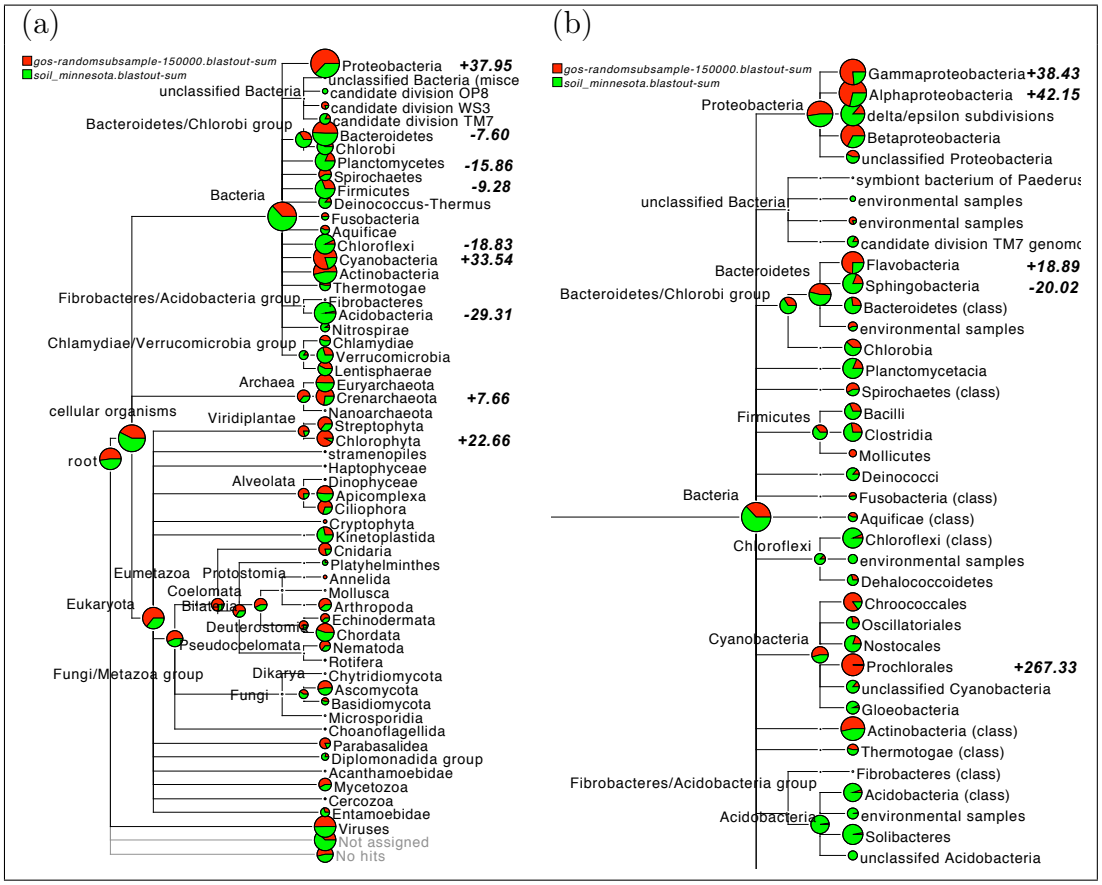

Figure 4.4: Two multiple-comparative tree views of a soil metagenome [Tringe et al., 2005] shown in green and a marine metagenome [Rusch et al., 2007] shown in red, as computed by MEGAN 2.0. In (a), we show an overview of the taxonomy down to the phylum level, whereas in (b) we display a part of a class-level analysis. In bold we show the support values as listed in Table 4.3.

We now briefly discuss some of the main differences summarized in Figure 4.4 and Table 4.3. Our analysis reiterates the well-known fact that soil metagenomes are significantly more complex than marine ones. However, this diversity is underrepresented in current reference databases. Therefore, more reads are assigned to the proteobacterial phylum in the marine dataset than in the soil one, in particular *Pseudomonas mendocina ymp*, *Shewanella* (aquatic bacteria), and some unclassified gamma proteobacteria, such as *marine gamma proteobacteria HTCC2080*, *HTCC2143* and *EBAC20E09*. Differences in the number of reads assigned to Cyanobacteria can be attributed to *Synechococcus* and *Prochlorococcus marinus* which both belong to the most abundant bacterial species in marine surface water [Venter et al., 2004].

Significantly more reads are assigned to Acidobacteria in the soil dataset, most mapping to *Solibacter usitatus Ellin6076*, a soil bacterium. However, since the Acidobacteria are a very divergent class of taxa, this discrepancy could be due to the low amount of currently sequenced species within this group.

The fact that reads hitting Chlorophyta are more present in the marine dataset is due to the number of hits to Prasinophyceae, which are marine algae. The existence of fresh water variants may explain the small number of hits in soil. Reads that match Chloroflexi are found more often in the soil than in the marine dataset, in particular *Herpetosiphon aurantiacus ATCC 23779*, which was originally isolated from a lake in Minnesota, the same state from which the soil sample was taken. The fact that Thermoprotei are favored by the marine sample is due to reads assigned to *Nitrosopumilus maritimus SCM1*, which is a mesophilic (not thermophilic) salt-water bacterium. The groups Oligohymenophorea and Aconoidasida both belong to the phylum Alveolata comprising a very divergent group of unicellular eukaryotes, some of them are capable of photosynthesis. Accordingly, the marine dataset contains significantly more reads of these eukaryotic clades than the soil dataset. Interestingly, most hits within Aconoidasida belong to the taxon *Plasmodium falciparum*, the pathogen of malaria. Since it is known that *P. falciparum* possesses a chloroplast-like organelle which presumably was derived in a common ancestor of Apicomplexa [Lang-Unnasch et al., 1998], a possible explanation may be that these reads come from a marine species that is closely related to the Aconoidasida, which itself is not well represented in the NR database.

Some remarks and discussions related to finding significant difference with resampling procedure

Though we have successfully adapted, extended and applied the resampling test of [Rodriguez-Brito et al., 2006], there are some drawbacks in the approach. According to the resampling process (described earlier in the beginning of this subsection 4.2.1) we reconsider the facts:

1. The use of $\alpha = 10\%$, $n = 1000$ and $N = 1000$ is merely an example.
2. Let p_1 denotes the proportion of proteins in sample 1 belonging to the specific subsystem. In step *1a*, a sequence of N Bernoulli trials is performed with probability of success p_1. Hence, the random number of counts in step *1a* has a binomial distribution with parameters N and p_1.
   Replacing p_1 by p_2, the corresponding fact holds for sample 2.
3. In *step 3*, we draw randomly from a mixture of two binomial distributions with mixture proportion 0.5.

4. Assume that we replace the median in *step 2* by the mean as an alternative measure of central location. Then computing the mean of M differences of counts in N trials is equivalent to the computation of the difference of the counts in $N \cdot M$ trials.
   This shows that the precision of the estimate of the difference in proportions in *step 2* is essentially based on sample size $N \cdot M$.

5. Due to slightly varying definitions of empirical quantiles, we could choose the 51^{th} and 950^{th} or the 50^{th} and 951^{th} element as well in *step 4*.

6. In *step 5*, we have to specify what is meant by "there is no statistically significant difference between the two samples". A more precise statement would be "there is no statistically significant difference of the proportions of proteins belonging to the specific subsystem between the two samples". That means we aim at testing the hypothesis $H_0 : p_1 = p_2$ against the alternative $H_0 : p_1 \neq p_2$.

Why the procedure of resampling test might not work properly

There exists a large amount of literature on bootstrap and resampling based tests, e.g., [Davison and Hinkley, 1997] and [Good, 2004]. Despite the variety of methods, they usually obey two basic requirements:

1. The test statistic in the resampling step is the same as that used for the original data.

2. Resampling is done from a distribution which satisfies the relevant null hypothesis.

Since steps *3a* and *3b* described above are identical, the two samples obtained in these steps come from the same mixture distribution; hence, the second point is satisfied in the procedure of [Rodriguez-Brito et al., 2006]. However, the first point is clearly violated since the test statistic in the resampling step *3c* is solely the difference of N counts, whereas the original test statistic consists of the median of M such differences. As a consequence, the variability of the original test statistic is much smaller than the variability of the resampled statistics. This leads to a conservative test, i.e. a test which nearly never rejects the hypothesis if it is true. Hence, the attained level of the test is much smaller than the nominal level. Such behavior has also a serious effect on the test under the alternative hypothesis: the power of the test is very low compared with tests which attain their nominal level.

4.2.2 Directed Homogeneity Test

In order to find a more sophisticated and statistically valid approach we introduced the *Directed Homogeneity test* for comparing metagenome samples. The test is based on basic statistical ideas and helps to investigate significant difference in a particular taxa in two datasets. The test provides answers to two questions:

- *Is there a significant difference in the proportions of occurrences of reads at the child node and at the parent node in two datasets?* ⇒ up test

- *Is there a significant difference in the distribution of reads among the children of a particular node in two datasets?* ⇒ down test

To answer these questions we combined two tests, the up and the down test, in our *Directed Homogeneity test*. Both test proportions. For this test we assume, in a taxonomy tree 'up' means towards the root or parent node (left side in the tree) and 'down' means towards the children or lower level nodes (right side in the tree).

For better understanding the situation is depicted in Figure 4.5 as a basic example. In the case of the up test, for each intermediate node we take the proportion of occurrence of reads at that particular node and at the parent node for two datasets and perform a two-sample test for equality of proportions with continuity correction. In Figure 4.5, *a* and *b* display this situation. We consider the proportion of occurrence of reads at the child node and at the parent node for both the datasets (shown in red and blue). Now if there is a significant difference in the proportion, then the left side of the child will be highlighted (as in Figure 4.5.*a*) and no highlight indicates the opposite (as in Figure 4.5.*b*).

The down test incorporates Person's Chi-squared test to compare the distribution of the two datasets on the children of a particular node. Figure 4.5, *c* and *d* display this situation. If there is a significant difference in the distribution of reads among the children of the parent node, then the right (or down in the tree) side of the parent will be highlighted (as shown in Figure 4.5.*c*). No significant difference in the test causes no highlighting, as displayed in Figure 4.5.*d*.

The program uses the two tests to highlight all nodes in the whole tree for which either test asserts a statistically significant difference. To be precise, if the *p*-value of the up test is below a critical level (e.g. 0.01), then the part of the node that faces the parent will be highlighted, whereas a significant *p*-value for the down test will result in the part of the node that faces the children being highlighted (see [Mitra et al., 2009] for details of the test).

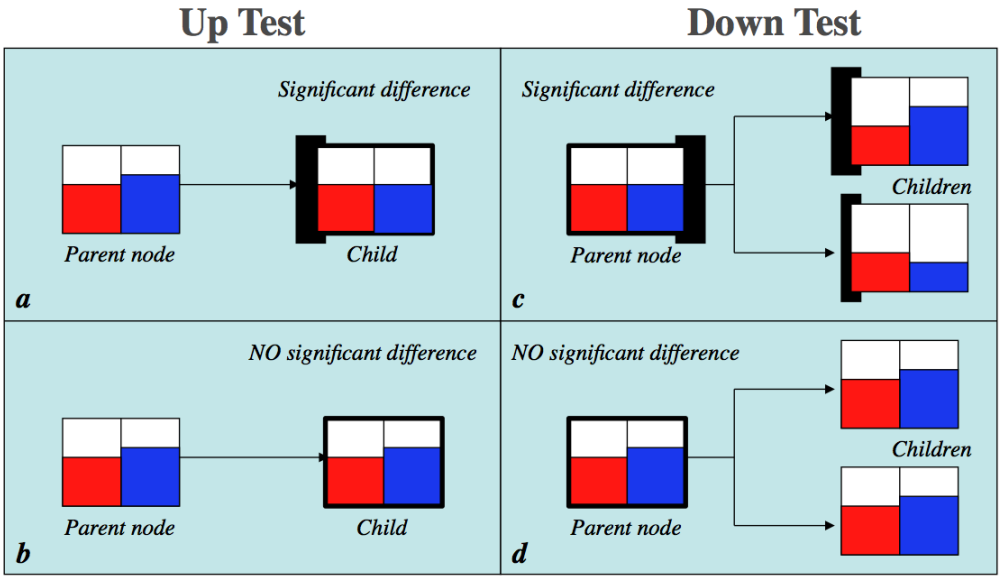

Figure 4.5: Directed Homogeneity test, including up and down test in a comparison of two datasets (shown in red and blue). Up test: A significant difference in the proportion of occurrence of reads at the child node and at the parent node, is causing the left side of the child to be highlighted (black) in *a*; whereas in *b* the difference is not significant. Down test: A significant difference in the distribution of reads among the children of the parent node, is causing the right side of the parent to be highlighted (black) in *c*; whereas in *d* the difference is not significant.

Note that such a difference may occur due to a number of causes, such as different sampling or sequencing technologies etc, and so will not necessarily reflect a biologically interesting difference. If the two datasets were obtained using different sequencing technologies, some adjustments to the analysis can be made to account for the different read-length distributions of multiple data sets.

Since a large number of tests are being performed during the comparison of two datasets, we have to address the problem of multiple testing: in a large number of tests we will see some false significant results by chance. To address this, we have implemented two well-known correction methods, namely the *Bonferroni* and the *Holm-Bonferroni* corrections [Shaffer, 1995, Holm, 1979]. It should be emphasized that controlling the family-wise error rate is not always needed, e.g. in more exploratory screening experiments. In other cases, the main aim is to decide whether the two samples come from different distributions. The overall conclusion that this is indeed the case need not be erroneous even if some of the (sub) null hypotheses are falsely rejected.

Statistical Background

Here we describe some fundamental knowledge about necessary statistical tests in the context of our method.

Two-sample test for equality of proportions. By testing equality of proportions between two samples, we can determine whether two probabilities are the same. We are interested to test whether proportions of occurrences of reads on a particular node in two datasets are similar or not. Suppose we consider the proportion of occurrence of reads on the *Gammaproteobacteria* node within *Bacteria* in two datasets A and B as p_1 and p_2. In terms of the abundance parameter, we write the null and alternative hypotheses as, $H_0 : p_1 = p_2$ vs. $H_0 : p_1 \neq p_2$, where p_1 and p_2 are the proportions of occurrences of a particular node within a higher level node for each datasets. We can perform significance tests based on the χ^2 statistic, as

$$z = \frac{(\hat{p_1} - \hat{p_2}) - 0}{\sqrt{(\frac{\overline{p}(1-\overline{p})}{n_1}) + (\frac{\overline{p}(1-\overline{p})}{n_2})}} \quad and \quad \chi^2 = z^2 \tag{4.2}$$

where n_1 and n_2 are the total number of reads in *Bacteria* for two datasets A and B; and $\overline{p}$ = proportion of total reads in *Gammaproteobacteria* and in *Bacteria*. When the p-value of the test is above a critical level (e.g. 0.01), we can say that the data are consistent with the null hypothesis.

Multiple testing problem. In statistics, the multiple comparisons, or multiple testing, problem occurs when one considers a set, or family, of statistical inferences simultaneously [Miller, 1981].

Family-wise error rate (FWER). This is the probability that we will get at least one false positive result, when multiple tests are performed. In order to retain a prescribed familywise error rate α in an analysis involving more than one comparison, the error rate for each comparison must be more stringent than α. Many statistical methods have been developed for this. We will use the *Bonferroni* correction [Shaffer, 1995] and *Holm-Bonferroni* correction [Shaffer, 1995, Holm, 1979].

Bonferroni correction. In this test the target α (typically $\alpha = 0.05$ or 0.01) is divided by the number of tests being performed. If the unadjusted p-value is less than the *Bonferroni*-corrected target α, then the null hypothesis is rejected. If the unadjusted p-value is greater than the $Bonferroni$-corrected target α, then the null hypothesis is not rejected.

Holm-Bonferroni correction. The *Bonferroni* test is called a "single-step" method, whereas Holm's test is a stepwise method. It is also called a sequential rejection method, because it examines each hypothesis in an ordered

sequence, and the decision to accept or reject the null hypothesis depends on the results of the previous hypothesis tests. This test is less conservative than the Bonferroni correction, and is therefore more powerful. The Holm's test uses a stepwise procedure to examine the ordered set of null hypotheses, beginning with the smallest p-value, and continuing until it fails to reject a null hypothesis.

Finding Significant Differences with Directed Homogeneity Test

In the pairwise comparison of taxonomic content of two datasets, the user can turn on the Directed Homogeneity test by selecting the Highlight Differences menu item. The user has the option to choose *no correction*, *Bonferroni* or *Holm-Bonferroni* for the adjusted p-value correction.

In Figure 4.6 we show the result of a pairwise comparison between the obese and lean mouse gut datasets [Turnbaugh et al., 2006]. From the black highlighting and the up p-values we can easily see that the significant differences between the two datasets mainly lie in the Bacteroidetes/Chlorobi group and Firmicutes classes. From the down p-values we can say that the difference in read numbers for the Bacteroidetes node is mostly due to the difference in read numbers for the Favobacteria class and the difference for the Firmicutes is mostly due to the difference for the Bacilli class. We didn't choose any multiple testing correction for this figure because using no correction will result in a maximum number of significant different nodes. The user can further investigate all possible nodes, having significant difference. To inspect any interesting node the user can refine this view to a lower-level comparison. These differences are similar to the differences reported in [Turnbaugh et al., 2006].

Moreover, our statistical method provides one p-value for the up-test and one for the down-test. From the up p-value, we can easily see that the proportional difference in number of reads assigned to the Bacteroidetes/Chlorobi group (UPv= 0.0) and the Firmicutes (UPv= 0.0) is highly significant between the two datasets, whereas the down p-value gives us additional comparative information about the children of these two nodes.

From the down p-value, we can say that the difference in read numbers for the Bacteroidetes/ Chlorobi group node (DPv= $2.77E-9$) is mostly caused by the difference in read numbers for Bacteroidetes phyla. The difference for Firmicutes (DPv= 0.0) is mostly caused by the difference for the Bacilli and Clostridia classes (see Figure 4.7). Figure 4.8 shows the same part of the tree, only the p-values are computed using the Bonferroni correction, which augments the p-values for each particular test based on the number of tests being performed. This correction is used to reduce problems associated with multiple comparisons, but it can significantly increases the risk of committing type II errors. In Figure 4.9

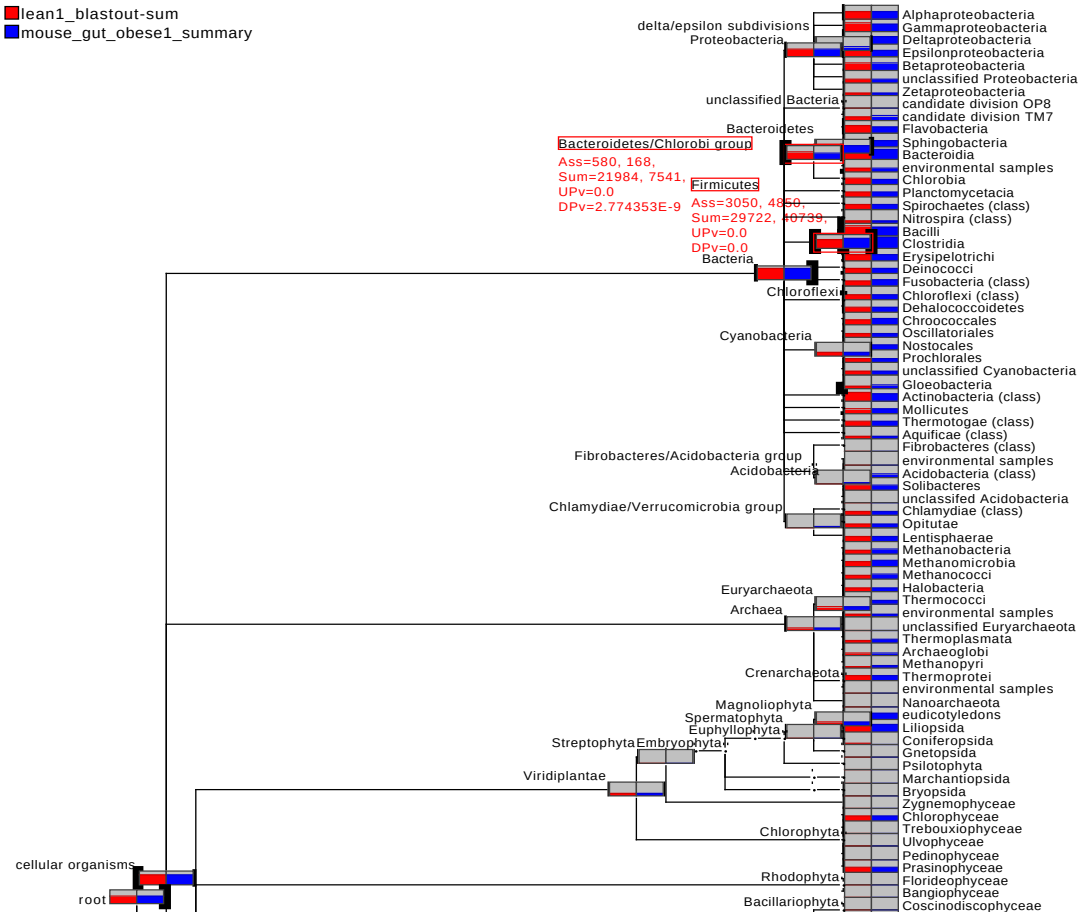

Figure 4.6: Pairwise comparison of two metagenome datasets, one from the gut of a lean mouse (red) and and one from an obese mouse (blue) collapsed at 'Class' level. Black highlighting on the left side of a node indicates that the up-test of the Directed Homogeneity test indicates a significant difference, whereas black highlighting on the right side indicates a significant difference detected by the down test. The thickness of the highlighting is logarithmically proportional with the significance.

the p-values are computed using the Holm-Bonferroni correction. Using either of the corrections, the results for the nodes of interest are still significant.

Testing the Performance of the Directed Homogeneity Test

Finally we tested the performance of our method with highly different environmental metagenome datasets. We used the above mentioned soil [Tringe et al., 2005] and marine sample [Rusch et al., 2007] for this study. We used these to see which differences and/or similarities between these datasets can

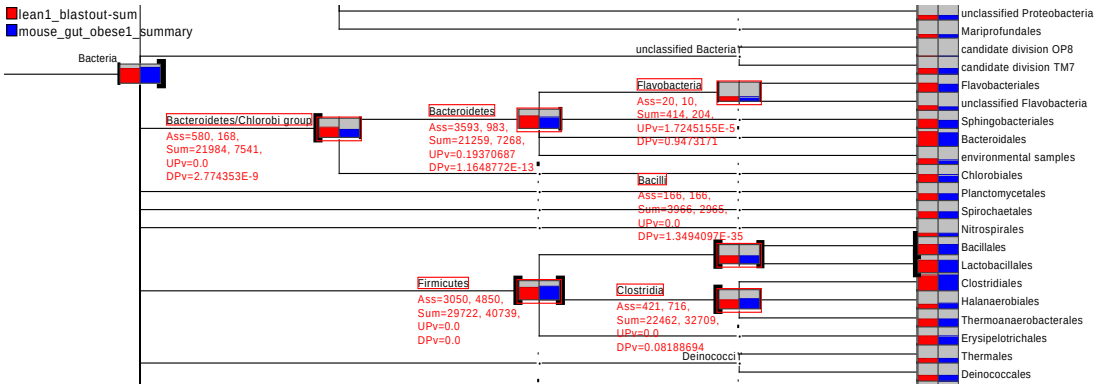

Figure 4.7: A part of the lean and obese mouse datasets comparison view (tree collapsed at 'Order' level). The labels UPv and DPv indicate the p-values associated with the up and down parts of the Directed Homogeneity test (Uncorrected).

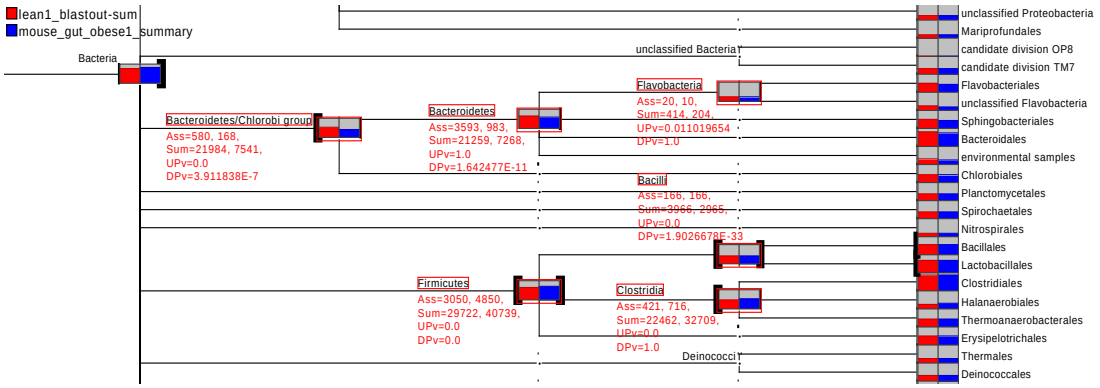

Figure 4.8: Same tree as Figure 4.7, here the p-values are computed using the Bonferroni correction.

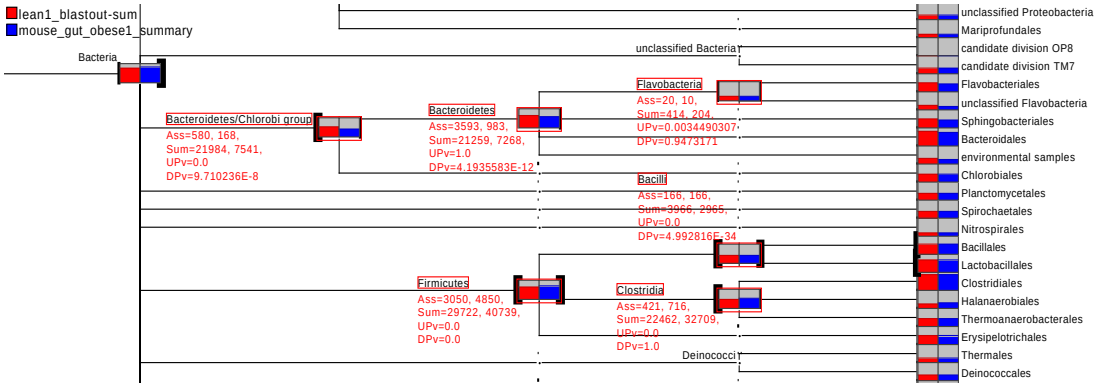

Figure 4.9: Same tree as Figure 4.7, here the p-values are computed using the Holm-Bonferroni correction.

be detected by our method. We will refer to these as the Soil and Sea datasets. For this experiment, we took 20 random subsamples (with replacement) from both datasets, each containing 20% of the original data. In this way, we got 20 Sea datasets (approx. 28,000 reads each) and 20 Soil datasets (approx. 30,000 reads each). We then conducted a Sea vs. Sea comparison, a Sea vs. Soil comparison and a Soil vs. Soil comparison, focusing our attention on particular bacterial nodes, namely Gammaproteobacteria, the Bacteroidetes/Chlorobi group and Firmicutes. Here all p-values are computed with no correction as we want to find out every possible differences and correction methods adjust the p-values. A detailed result including all the up and down p-values for all comparisons can be found in supplementary Tables C.1.1, C.1.2, C.1.3. We have summarized the results in Box-and-Whisker plots for each of the three bacterial nodes (Figure 4.10).

If we take "a proportional similarity of reads assigned to a particular node between two datasets" as a null hypothesis, then in the Soil vs. Soil comparisons and Sea vs. Sea comparisons, the up p-values (UPv) lie above the significance level (0.01) in more than 99% of the cases for all three bacterial nodes (Figure 4.10). Hence, 99% of all cases are consistent with the null hypothesis, that is, we cannot reject the null hypothesis. On the other hand, the up p-value (UPv) is close to zero (less than the significance level 0.01) in more than 95% cases of the Sea vs. Soil comparisons, reflecting a highly significant difference in the proportion of these three bacterial groups, between subsamples. In Figure 4.10 white boxes represent the up p-values (UPv) for Gammaproteobacteria, the Bacteroidetes/Chlorobi group and Firmicutes in all three comparisons.

Moreover, if we take "a proportional similarity between distribution of reads among the children of a particular node between two datasets" as a null hypothesis, then in the Soil vs. Soil comparisons and Sea vs. Sea comparisons, the down p-values (DPv) lie above the significance level (0.01) in more than 99% of the cases (Figure 4.10). Hence, in 99% of all cases, the data sets are consistent with the null hypothesis that the distribution of reads in the children of the Gammaproteobacteria, of the Bacteroidetes/Chlorobi group and of the Firmicutes is similar in the two datasets (Soil and Sea). For the Soil vs. Sea comparisons (Figure 4.10) for Gammaproteobacteria and the Bacteroidetes/Chlorobi group nodes, the down p-value (DPv) is close to zero (less than the significant level 0.01) in more than 99% cases, reflecting a highly significant difference in the distribution of reads among the children of these nodes. For Firmicutes, the down p-values (DPv) are close to zero (less than the significance level 0.01) in only 40% of the cases, reflecting that the distribution of reads is significantly different among the children of this node (Firmicutes) between the two datasets only in 40% of the cases. This may be because Firmicutes are common Gram-positive bacteria present in both marine and land-based environments [Fierer et al., 2007, Yooseph et al., 2007]. In

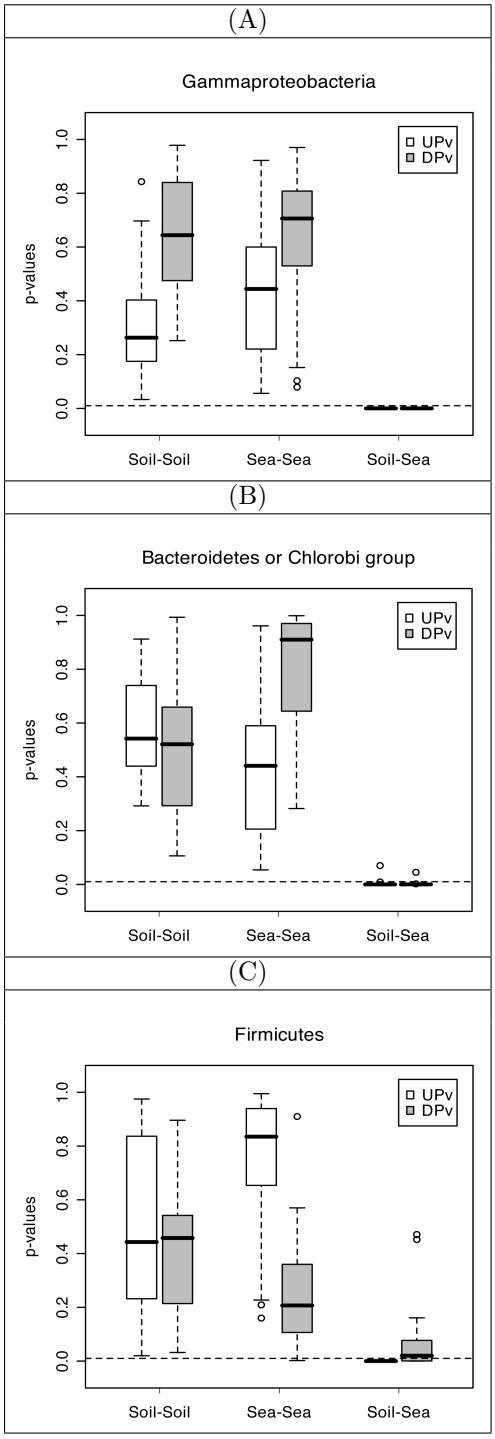

Figure 4.10: Box-and-Whisker plots summarizing the up p-values (UPv: white boxes) and down p-values (DPv: gray boxes) for (A) Gammaproteobacteria, (B) the Bacteroidetes/Chlorobi group and (C) Firmicutes in a Soil vs. Soil comparison, a Sea vs. Sea comparison and a Soil vs. Sea comparison. Each comparison is based on 20 independent pairs of subsamples. Dashed line indicates $p = 0.01$ threshold.

many cases, the proportional distribution of reads among child nodes of Firmicutes can be similar in different Soil and Sea subsamples. In Figure 4.10 gray boxes are representing the down p-values (DPv) for Gammaproteobacteria, the Bacteroidetes/Chlorobi group and Firmicutes in all three comparisons. In the case of Soil vs. Sea comparisons, most of the time the values are very close to zero reflecting a highly significant difference between the two subsamples.

This illustrates how the "Directed Homogeneity test" can provide an initial statistical comparison.

4.3 Discussion

Comparative metagenomics is a rapidly growing field. Fast and user-friendly tools are needed to analyze metagenomic datasets. In this chapter, we have introduced some visual comparison techniques for multiple comparison and statistical approaches for comparing metagenome datasets in a pair. These results are implemented in MEGAN 3 (or later) and can help users to get a first impression of the similarity between multiple metagenomes and allow close comparison of two datasets at a time. Not only metagenomes, these tests are also applicable to metatranscriptome samples. Our next goal is to support sophisticated comparative analysis of multiple metagenome datasets.

Chapter 5

Multiple Metagenome Comparison using Networks

As mentioned in Chapter 2, there is a need for the development of new methods for analysing, comparing and visualizing multiple metagenome datasets simultaneously. In this chapter we present a novel approach that combines the use of taxonomic analysis, ecological indices and non-hierarchical clustering to provide a network representation of the relationships between different metagenome datasets. Explicitly, a tool that combines the visualization of relationships with a metric of distance in a single package which includes appropriate ecological indices without the need to fit metagenomic data to a root evolutionary dendrogramatic relationship. The goal of this chapter is to solve the question:

- *How can multiple metagenome datasets be compared?*

5.1 Theory and Background

Here we describe a few fundamental aspects of numerical ecology, phylogenetic and non-hierarchical clustering techniques. Besides the description of some methods, the focus is primarily kept on concepts that are important for this thesis.

5.1.1 Ideas from Ecology

"For almost a century, ecologists have collected quantitative observations to determine the resemblance between either the objects under study (sites) or the variables describing them (species or other descriptors)"

[Legendre and Legendre, 1998]. As we discussed in chapter 3 in the study of metagenomics, we obtain an estimate of the taxonomical content of a sample using MEGAN [Huson et al., 2007]. We can use these descriptors from multiple metagenomes to get an estimate of similarity and/or dissimilarity between the datasets based on the ideas from ecology.

The field of mathematical ecology provides a number of different distance measures for comparing the population structure of different habitats. We have tested a broad range of such distance measures to determine the best possible method for multiple metagenome comparison. After reviewing 27 different ecological measures (listed in [Legendre and Legendre, 1998]), six different distance measures were selected and tested in this study, namely the Euclidean distance, three quantitative coefficients (Kulczynski [Odum, 1950], Bray-Curtis [Bray and Curtis, 1957] and Hellinger [Rao, 1995]) and two probabilistic coefficients (Chi-square [Lebart et al., 1979] and Goodall's index [Goodall, 1964, Goodall, 1966]).

The basic metric measure is the *Euclidean distance*, which is computed using *Pythagoras' formula*. By determination of the Euclidean distance, the distance (D) between two metagenome samples (X, Y) can be calculated using

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\,, \tag{5.1}$$

where x_i and y_i are the read counts for the i^{th} taxon of the respective metagenome samples X and Y, which are positioned in an n-dimensional *Euclidean space*.

The Euclidean distance does not have an upper limit, its value can increase indefinitely with the number of descriptors and depends on the scale of each descriptor. Czekanowski (1909), Pearson (1926), Mahalanobis (1936), Whittaker (1952), Williams and Stephenson (1973), Orlóci (1978) and many other mathematicians discussed the drawbacks of Euclidean distance and proposed different ways. In general double-zero cases[1] lead to reduction of distances. It is thus preferable to abstain from drawing any ecological conclusion from the absence of a species in two datasets.

In numerical terms, this means skipping double zeros in computations. Coefficients of this type are called *asymmetrical coefficients*. Many coefficients are available for comparing sites using species presence-absence data such as 'Jaccard' (1900, 1901, 1908), 'Kulczynski' (1928), 'Russell and Rao' (1940), 'Steinhaus' (1947), 'Sørensen' (1948) and many others. Steinhaus's measure has been

[1] In comparison of more than two datasets, if a species is absent in two datasets but present in the other, then this is a double-zero case.

rediscovered and modified a number of times as 'Odum' (1950) or as 'Bray and Curtis' (1957). Among these measures we select 'Kulczynski' (5.2) and 'Bray and Curtis' (5.3) distances as the representative of this group for our study, which are defined as:

$$D(X,Y) = 1 - \frac{1}{2}\left(\frac{\sum_{i=1}^{n}\min(x_i,y_i)}{\sum_{i=1}^{n}x_i} + \frac{\sum_{i=1}^{n}\min(x_i,y_i)}{\sum_{i=1}^{n}y_i}\right) = 1 - \frac{1}{2}\left(\frac{W}{A} + \frac{W}{B}\right) \quad (5.2)$$

and

$$D(X,Y) = 1 - 2\frac{\sum_{i=1}^{n}\min(x_i,y_i)}{\sum_{i=1}^{n}(x_i+y_i)} = 1 - \frac{2W}{A+B}\,, \quad (5.3)$$

where W is the sum of the minimum abundances of the various species, this minimum being defined as the abundance in the dataset where the species is the rarest. A and B are the sums of the abundances of all species at each of the two datasets or, in other words, the total number of species observed or captured at each dataset, respectively [Legendre and Legendre, 1998].

There are other measures to calculate the distance among sites using species abundances such as the χ^2 metric by Roux and Reyssac (1975), χ^2 distance by Lebart and Fŕnelon (1971) and Hellinger distance by Rao (1995). χ^2 distance differs from the χ^2 metric in that the terms of the sum of squares are divided by the probability (relative frequency) of each row in the overall table instead of its absolute frequency. Among these measures we select χ^2 (5.4) and 'Hellinger' (5.5) distances for our study, which are defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}\frac{(\hat{x}+\hat{y})}{(x_i+y_i)}\left(\frac{x_i}{\hat{x}} - \frac{y_i}{\hat{y}}\right)^2}\,, \text{with } \hat{y} = \sum_{i=1}^{n} y_i \quad (5.4)$$

and

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}\left(\sqrt{\frac{x_i}{\hat{x}}} - \sqrt{\frac{y_i}{\hat{y}}}\right)^2}\,, \text{with } \hat{y} = \sum_{i=1}^{n} y_i\,. \quad (5.5)$$

The functions so far discussed measure the likeness and/or unlikeness of two samples X and Y as a composite of n individual quantities, each representing a similarity and/or dissimilarity value $D(X,Y)_i$ of the two samples with respect to the i^{th} taxon of the n taxons in the samples.

A different approach is Goodall's probabilistic coefficient [Goodall, 1964, Goodall, 1966]. It takes into account the frequency distribution of the various states of each descriptor in the whole set of samples. Indeed, it is less likely for two datasets to both contain the same rare taxon (or species) than a more frequent taxon. In this sense, likeliness for a rare species should be given more

importance than for a common species, when estimating the similarity between datasets.

In a comparison of multiple datasets, first a partial similarity measure (Gower's matrix) for all possible pair combinations (s_{pair}) is calculated for each species i, as:

$$s_{pair_i} = 1 - \left[\frac{|x_i - y_i|}{R_i}\right],$$

where R_i is the range (largest difference) of abundance for the i^{th} species across all the datasets in the study. Then for each pair of datasets, one computes the proportion of partial similarity values belonging to species i that are larger than or equal to the partial similarity of the pair of datasets being considered. These proportions (p_i) are combined for the n species by computing the product[2] of the values relative to various species as:

$$\prod = \prod_{i=1}^{n} p_i$$

Finally the similarity (S) between two datasets (X, Y) can be obtained as the proportion of the products ($\prod$) that are larger than or equal to the product of the pair of datasets ($\prod_{pair}$) considered. The equation is given by

$$S(X,Y) = \frac{\sum_{pairs} d}{\frac{n(n-1)}{2}}, \text{where } d = \begin{cases} 1 & \text{if } \prod \geq \prod_{pair} \\ 0 & \text{if } \prod < \prod_{pair} \end{cases} . \tag{5.6}$$

For better understanding, computation of Goodall's index is explained in supplementary section C.2 with an example. Further details are available in [Goodall, 1964, Goodall, 1966, Legendre and Legendre, 1998].

Some distance measures do not fulfill the triangle inequality axiom, thus they are not a metric. As a consequence, they do not allow a proper ordination of sites in a full Euclidean space. Sørensen's coefficient is one of them, whereas Kulczynski, Odum and Bray-Curtis, Goodall are semimetric measures. For both metric and semimetric measures we can use $D = 1 - S$, where D and S stand for 'Distance' or 'Difference' and 'Similarity'.

There are other probabilistic similarity coefficients, such as, Raup and Crick (1979), McCoy (1986) etc. But these measures only consider species presence-absence data and so they are not considered in our study.

[2]For our purpose we have considered this product in a logarithmic scal, as there are many taxa in a metagenome datasets and this product can become very small.

5.1.2 Ideas from Phylogeny

There are two popular ways to produce a graphical representation of genetic distance matrices. The first approach is widely applied in ecological studies, which includes using a principle component analysis (PCA) or non-metric multidimensional scaling (NMDS) to obtain a two-dimensional layout. The second is widely used in evolutionary studies, that involves using rooted trees computed by a hierarchical clustering method [Rusch et al., 2007]. The advantage of a tree representation is that it explicitly provides clusters of closely related datasets. However, metagenomes do not evolve along a tree, and there are numerous environmental factors that may affect dataset composition resulting in distances that reflect incompatible signals. While ordination methods do not suffer from this problem, they do not link data points into explicit clusters and provide no metric against which to determine the distance between datasets. Hence, we propose to use the 'Neighbor-Net' method [Bryant and Moulton, 2004], which extends the 'Neighbor-Joining' (NJ) algorithm [Saitou and Nei, 1987], to compute an unrooted phylogenetic network that enjoys the advantages of both methods. This is a special type of phylogenetic network that simultaneously represents both groupings in the data and evolutionary distances between taxa. Here we use these networks for clustering multiple datasets. Such networks are not restricted to being a tree and are able to display an incompatible, that is non-hierarchical, clustering of the data.

5.2 Multiple Comparison

We combined the different ideas of ecology and phylogeny for comparing multiple metagenome datasets. First, a taxonomic profile is computed for each dataset. Second, a matrix of pairwise distances is determined using one of the six ecological distances described in 5.1.1. Finally, neighbor-net is applied to the matrix and the network is represented using an appropriate visualization technique [Dress and Huson, 2004]. We apply the approach to marine metagenomes or metatranscriptomes from three types of studies; a mesocosm experiment [Gilbert et al., 2008a], a spatially structured dataset (the Global Ocean Survey) [Rusch et al., 2007] and a time-series dataset [Gilbert et al., 2009]. Our work suggests that the approach is robust as it produces networks that are very similar across all ranks of the NCBI taxonomy and, to a lesser extent, across different ecological indices (see [Mitra et al., 2010a]).

5.2.1 Methods

In general first we processed all metagenomes and metatranscriptomes aligning against the NCBI-NR database using the BLASTX tool. The results were imported then into MEGAN using the 'Import from BLAST' option. We then performed multiple comparisons using various ecological indices and constructed networks using the neighbor-net algorithm [Bryant and Moulton, 2004], as implemented in version 4 of MEGAN. We conducted six case studies as exemplary to establish our method.

Case 1:

At first, we compared eight published marine datasets consisting of four metagenomes (DNA) and four metatranscriptomes (cDNA), from a controlled costal ocean mesocosm study (Bergen, Norway) involving an induced phytoplankton bloom. The samples were taken at two time points, at the peak (Time1 or 13th May) and immediately after the collapse of the bloom (Time2 or 19th May) and named these eight samples as follows: 1-Time1-Bag1-DNA, 2-Time1-Bag6-DNA, 3-Time2-Bag1-DNA, 4-Time2-Bag6-DNA, 5-Bag1-13May-cDNA, 6-Bag1-19May-cDNA, 7-Bag6-13May-cDNA and 8-Bag6-19May-cDNA (please refer to [Gilbert et al., 2008a] for details). Further we will refer to these as PML-Bergen datasets.

All datasets were randomly re-sampled to the smallest data set size to allow inter-comparison. After opening all the datasets in MEGAN, the 'Compare' menu item was used to generate a new document that contains a comparison of all datasets. We compared the taxonomical profiles (as MEGAN files), which contain taxon abundance counts, of these eight datasets. Multiple comparisons of the datasets were then performed using six different ecological distance measures (Euclidean, Kulczynski [Odum, 1950], Bray-Curtis [Bray and Curtis, 1957], Hellinger [Rao, 1995], Chi-square [Lebart et al., 1979] and Goodall's index [Goodall, 1964, Goodall, 1966]) at each of seven taxonomic ranks ('kingdom', 'phylum', 'class', 'order', 'family', 'genus' and 'species') to create a total of 42 networks (Figures 5.1, C.2.1, C.2.2, C.2.3). The distances were processed by the neighbor-net algorithm [Bryant and Moulton, 2004] to obtain a collection of unrooted phylogenetic networks.

Case 2:

In a second study, we used one random sub-sample of the Sargasso Sea data [Venter et al., 2004] and one sub-sample from the Sorcerer II Global Ocean Sam-

pling expedition data (GOS) [Rusch et al., 2007] and the data and setup from the previous case study. We wanted to visualize the comparison of multiple marine metagenomes from different environments processed using different sampling and sequencing strategies. All ten datasets were randomly re-sampled to allow inter-comparison of taxonomic abundances. In a similar way to the previous study we compared these ten datasets using four of the distances (Goodall's index, Euclidean distance, Hellinger distance and Chi-squared distance) at each of seven taxonomic ranks to create 28 additional networks. We observed that the networks obtained using the Kulczynski and Bray-Curtis distances looked very similar to the networks obtained using Euclidean distances in the previous study (Figures 5.2, C.2.4, C.2.5). So we decided to drop the Kulczynski and Bray-Curtis distances from subsequent experiments.

Case 3:

Furthermore we believe that the bacterial taxa are more important for metagenomic studies than the eukaryotes and viruses. So we performed multiple comparisons using four of the indices (Goodall's index, Euclidean distance, Hellinger distance and Chi-squared distance) considering only bacterial taxa at six taxonomic ranks. First we did this for eight PML-Bergen datasets and then together with the sub-sample of the Sargasso Sea and the GOS dataset (i.e. 10 datasets together). This experiment results further 24 networks for each case (Figures 5.3, C.2.6, C.2.7, 5.4, C.2.8 and C.2.9). Unlike Euclidean distances and Goodall's index, two other distances can be applied to raw data. Thus for the Goodall's index and Euclidean distances, we randomly normalized the numbers of bacterial sequences, to standardize the apparent sequencing effort.

Case 4:

Here we investigated the effect of excluding rare taxa from the taxonomical profiles. For example, we considered the data at the class rank of the NCBI taxonomy. The six metagenomes (four Bergen metagenomes, one Sargasso Sea sample and one GOS sample from the previous study) were duplicated. Then we excluded all taxa that have an abundance of less than 0.025% of the total community abundance from each duplicated dataset. In Table 5.1 the details of the community change can be seen. We then compared these six truncated metagenomic datasets using all six indices, resulting six networks at the 'class' level (Figure 5.5).

Case 5:

In a fifth study we compared all 41 samples of spatially structured GOS data. We downloaded the GOS data, from the CAMERA website [Seshadri et al., 2007]. All datasets were blasted against the NCBI-NR database and the result was imported into MEGAN. As for the Bergen samples, we computed taxonomic profiles as MEGAN files. We then normalized the datasets to the smallest size to allow inter-comparison of taxonomic abundances. Next we performed the comparison using Goodall's index at the class rank considering all the sites together (Figure 5.7.B). We assume that the coastal sites may harbor a more diverse microbiota than the open ocean sites. So we further compared only the coastal and open ocean sites (Figure 5.7.C) to illustrate biogeographic clustering.

Case 6:

Finally we studied and analyzed the correlation between 12 16S rRNA V6 tag-pyrosequencing datasets spanning 12 months of 2007 at a continually monitored sampling site, L4, in the Western English Channel [Gilbert et al., 2009]. As before, to allow inter-comparison, we re-sampled these 12 samples to identical sequencing depth. The OTU abundance matrix was prepared by adding zeros where there were no representatives for that sample.

First we compared samples using Goodall's index in combination with neighbor-net based on all unique OTUs (Figure 5.8.A), then excluding OTUs found on only one occasion (Figure 5.8.B). Finally comparison is performed considering only the OTUs found every time (Figure 5.8.c). Beside this we prepared the principle component analysis (PCA) and non-parametric multidimensional scaling (NMDS) plots using the same OTU data i.e. for OTUs present in two or more occasions. For the PCA analysis we used the raw data and for the NMDS calculation we used a computed Bray-Curtis matrix (Figure 5.9).

5.2.2 Implementation

A program for computing ecological indices from taxonomical profiles (called MEG2DIST, written in java) is available as open source from the website:
`www-ab.informatik.uni-tuebingen.de/software/megan/meg2dist`.
We implemented all six ecological indices (Euclidean, Kulczynski [Odum, 1950], Bray-Curtis [Bray and Curtis, 1957], Hellinger [Rao, 1995], Chi-square [Lebart et al., 1979] and Goodall's index [Goodall, 1964, Goodall, 1966]) in our program. In addition the code is completely integrated into version 4 of MEGAN, which is available from the website:

`www-ab.informatik.uni-tuebingen.de/software/megan`.
After comparing all the metagenomes, the user can directly compute ecological indices and visualize networks by choosing 'Compare Datasets Using Networks' from the 'Options' menu item.

5.2.3 Results and Discussions

Case 1: Comparison of eight marine samples (metagenome and metatranscriptomes) from an ocean acidification study. For the analysis of PML-Bergen samples, all six selected ecological indices produce almost identical placements of the eight samples within a neighbor network. Only minor differences are visible in the distances between samples (see Figure 5.1 and supplementary information Figures C.2.1, C.2.2, C.2.3). The placement of these PML-Bergen samples conforms to reported biological and experimental relationships [Gilbert et al., 2008a], with the metagenomes being well-separated from the metatranscriptomes. Moreover and the samples from the peak of the induced phytoplankton bloom (Time1 or 13th May) appears more separated from the samples following the collapse of the phytoplankton bloom (Time2 or 19th May) than each group is to itself. Interestingly, for the 'Time 2' or '19th May' metagenomes, the

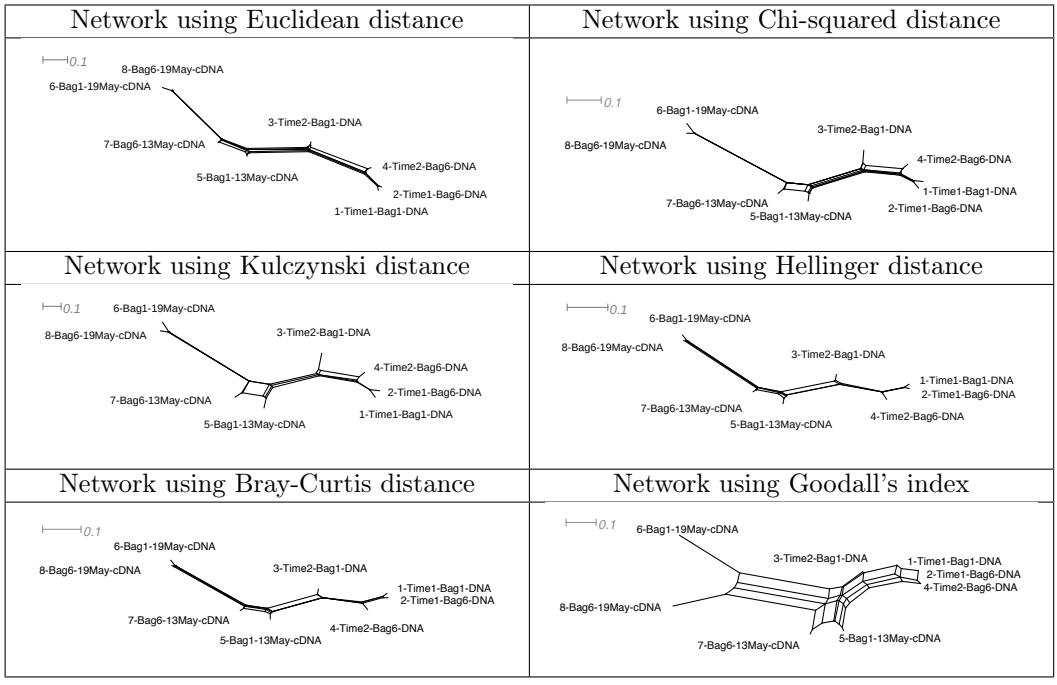

Figure 5.1: Neighbor-net based network obtained using six ecological indices showing the comparison of eight PML-Bergen samples (four metagenomes and four metatranscriptomes) considering all nodes at the class rank of the NCBI taxonomy.

opposite is true. i.e. the differences between these being greater than their similarity to samples within the 'Time 1' metagenomes. This indicates the extremely different ecology of the mesocosm samples that existed following the collapse of the bloom. This was brought about by the experimental methodology used, where immediately following the collapse of the bloom 'Bag1' was re-bubbled with CO2 and Bag6 was re-bubbled with air. This significantly altered the community composition and hence forced these samples apart (for more information refer to [Gilbert et al., 2008a]).

Case 2: Comparison of multiple marine metagenomic samples from different studies. To confirm that the Bergen-PML network was robust in the presence of additional marine samples, we added two additional marine metagenomes as "decoys". The first was a subset of reads taken from the pooled Sargasso Sea study [Venter et al., 2004] and the second was a subset of the larger Global Ocean Survey (GOS) [Rusch et al., 2007]. To allow an accurate comparison, a random subset of 96,201 sequences (the size of the smallest PML-Bergen dataset [Gilbert et al., 2008a]) was extracted from each study. After computing networks with four indices (Figures 5.2, C.2.4, C.2.5), we confirmed that the eight PML-Bergen samples remain in their original groupings and that the two decoys are placed at a distance from them. Interestingly, there are clear differences be-

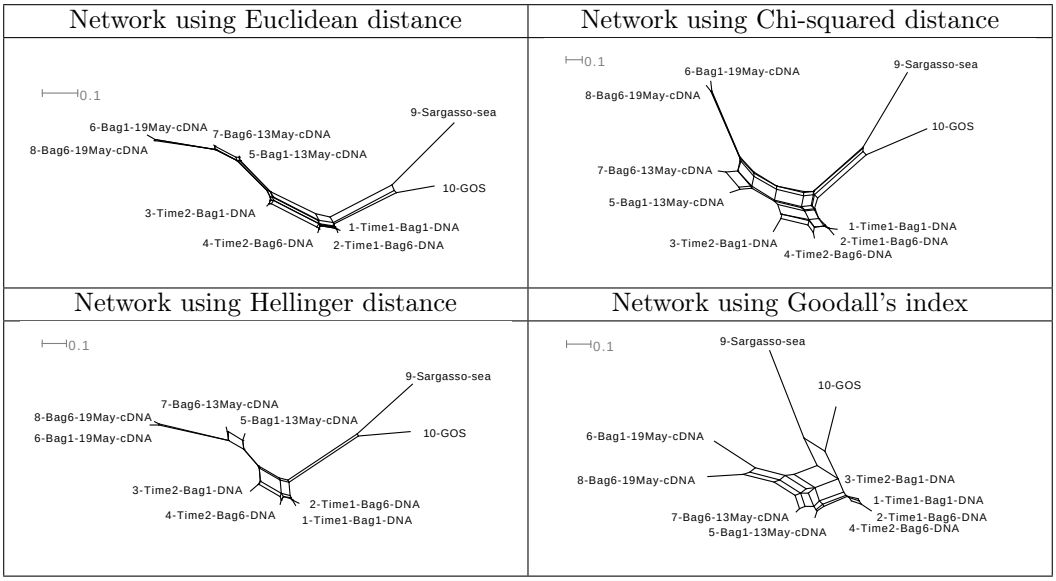

Figure 5.2: Network obtained using four ecological indices showing the comparison of ten marine samples (randomly resampled Sargasso Sea and GOS samples together with the eight PML-Bergen samples) considering all nodes at the class rank of the NCBI taxonomy.

tween the networks based on the Euclidean distance, wherein the decoys are much more distantly related to the PML-Bergen samples than that for the Goodall's in-

dex (Figures 5.2, C.2.4). We hypothesize that this is due to the biases induced by the vast rare biosphere and the way each index handles low-abundance sequences. Goodall's index provide more importance to the likeliness or similarity for a rare species than for a common species, whereas Euclidean distance is dominated by the abundant species. The networks based on the Hellinger and Chi-squared distances (Figures 5.2, C.2.5) are also similar. The GOS sample appears to cluster more closely to the PML-Bergen samples than the Sargasso Sea sample, as the GOS sample (random sub-sample of all GOS samples) is heavily enriched from coastal study sites, whereas the Sargasso Sea is an oligotrophic open ocean [Venter et al., 2004].

Case 3: Multiple metagenome/metatranscriptome comparisons considering only bacterial nodes. When only bacterial taxa of eight PML-Bergen samples are considered, the samples come more close to each other, while remaining mostly in their original grouping (Figures 5.3, C.2.6, C.2.7). This is because the samples were taken under very different conditions, thus while the bacterial populations remain unchanged the Eukaryotic populations were quite different so that we see a lot more Eukaryotic DNA in the mid-bloom (time 1) samples than in the post-bloom (time 2) samples. As a result, inclusion of all taxa leads to bigger differences than for the bacteria, but far more subtly.

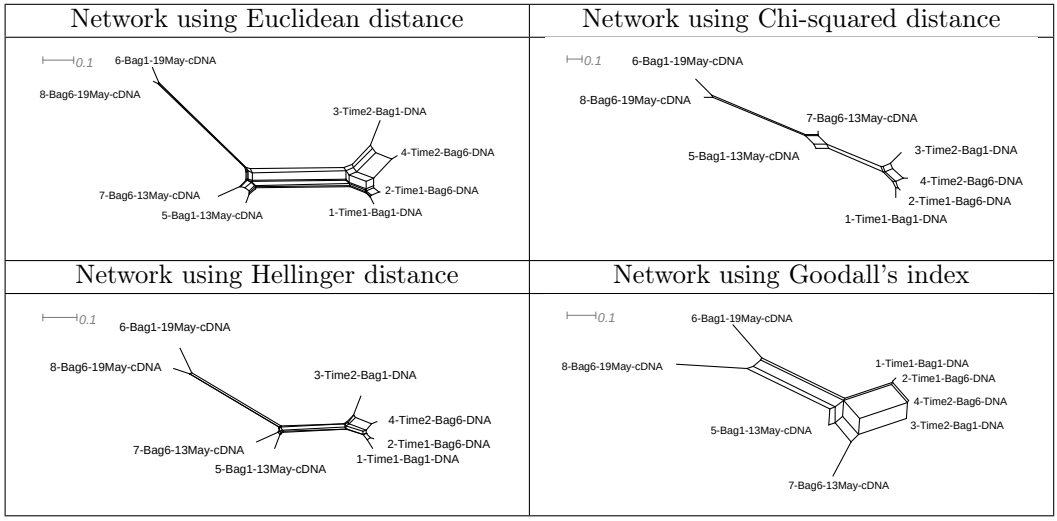

Figure 5.3: Network obtained using four ecological indices showing the comparison of eight PML-Bergen samples (four metagenomes and four metatranscriptomes) considering only bacterial nodes at the class rank of the NCBI taxonomy.

For multiple comparison of ten datasets (data similar to the second study) the Sargasso Sea dataset appears to be more similar to the other datasets than it does when all taxa are considered (Figures 5.4, C.2.8 and C.2.9). This is because the Sargasso Sea sample contains a much smaller number of eukaryotic

reads compared to the other datasets. This reflects the similar water sampling procedures (e.g. filter size) for the GOS [Rusch et al., 2007] and PML-Bergen [Gilbert et al., 2008a] datasets, resulting in organisms of a similar size range being analysed; whereas the Sargasso Sea study used a different sampling procedure [Venter et al., 2004] which excluded micro-eukaryotes.

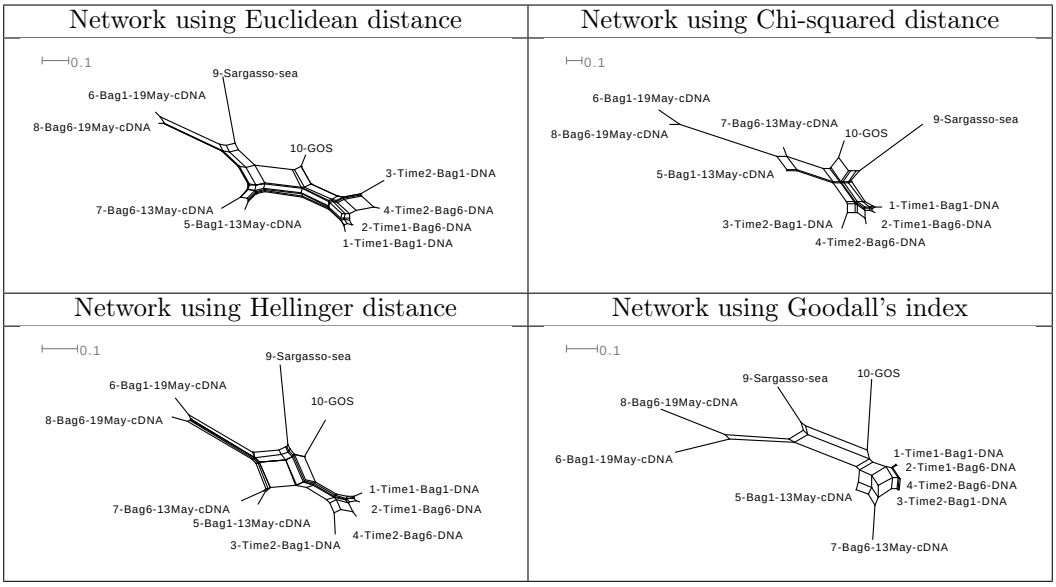

Figure 5.4: Network obtained using four ecological indices showing the comparison of ten marine samples (randomly resampled Sargasso Sea and GOS samples together with the eight PML-Bergen samples) considering only bacterial nodes at the class rank of the NCBI taxonomy.

In this study the networks computed using Goodall's index and Hellinger distance (Figures 5.3, 5.4 and right column of C.2.6, C.2.7, C.2.8 and C.2.9) maintain a very similar layout over all ranks of the NCBI taxonomy for the ten metagenome datasets, whereas the networks using Euclidean distance and Chi-squared distance (Figures 5.3, 5.4 and left column of C.2.6, C.2.7, C.2.8 and C.2.9) exhibit more variability. Strikingly, unlike the first and second studies, the PML-Bergen metagenomes tend to group together by time, with the 'Time 1' (13th May) being more similar to each other than to the 'Time 2' (19th May), and *vice versa*. This suggests that the post-bloom bubbling treatment of these bags had a greater impact on the eukaryotic and archaeal communities than on the bacterial communities. This is possibly a result of the bubbling-induced lysis of eukaryotic cells.

Case 4: The effect of rare taxa. To study the effect of rare taxa on such analyses, we excluded all taxa having an abundance of less than 0.025% from each of the six metagenomes examined above (now excluding the four meta-transcriptomes). The resulting truncated data sets were then compared with the

original full datasets. We observe that the placement of the original metagenomes remains the same in all the networks computed. The networks based on the Euclidean, Kulczynski and Bray-Curtis distances are unable to distinguish between the original and truncated metagenomes, placing them at identical locations in the network (Figure 5.5; left column). Networks obtained using the Chi-Squared and Hellinger distances place the truncated samples close to the original metagenomes, but on separate branches (Figure 5.5; right column). Only the network based on

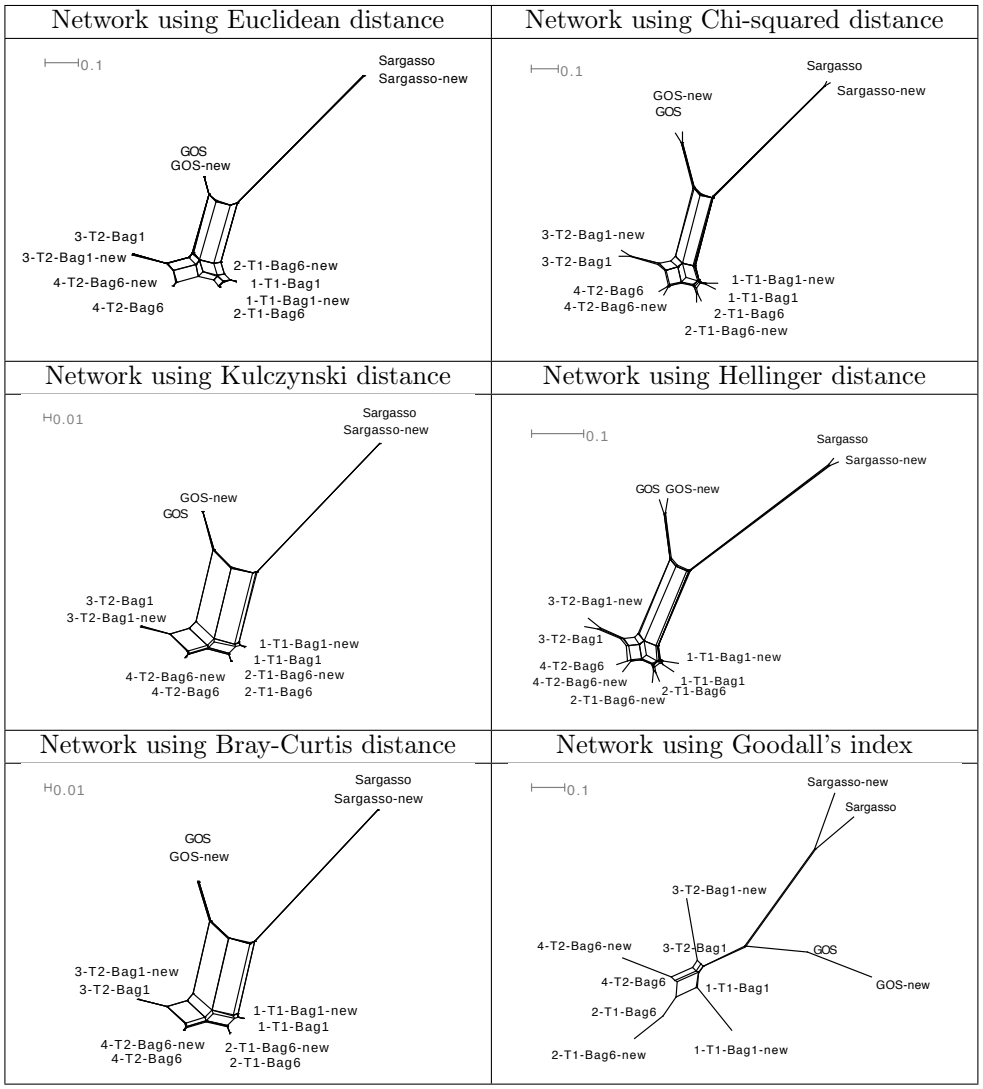

Figure 5.5: Comparison of six marine metagenomes with six truncated copies in which all rare taxa were excluded, analyzed at the 'Class' level of the NCBI. The displayed networks were obtained using Euclidian, Kulczynski and Bray-Curtis distances (left column) and Chi-squared, Hellinger distances and Goodall's index (right column).

Goodall's index was able to represent the correct branching within the datasets

(Figure 5.5).

Interestingly, we observed that the distances between the original and the truncated datasets are roughly proportional to the percentages of community change. This can be further understood by combining the result from Table 5.1 and Figure 5.6. For example in '1-T1-Bag1' dataset the total number of identified 'class' level taxa was 96. Out of those 96 taxa, 27 taxa had an abundance of read less than 0.025% of the total community abundance (Table 5.1).

Dataset name	Total number of taxon identified at 'Class' level in original dataset	Total number of taxon excluded from original dataset (abundance < 0.025%)	% of community (taxon) change in new dataset
1-Time1-Bag1-DNA	87	27	31.03 %
2-Time1-Bag6-DNA	77	16	20.78 %
3-Time2-Bag1-DNA	89	28	31.46 %
4-Time2-Bag6-DNA	94	26	27.66 %
Sargasso	81	12	14.81 %
GOS	96	34	35.42 %

Table 5.1: Detail numbers of identified and excluded taxa and resulting community change in truncated datasets.

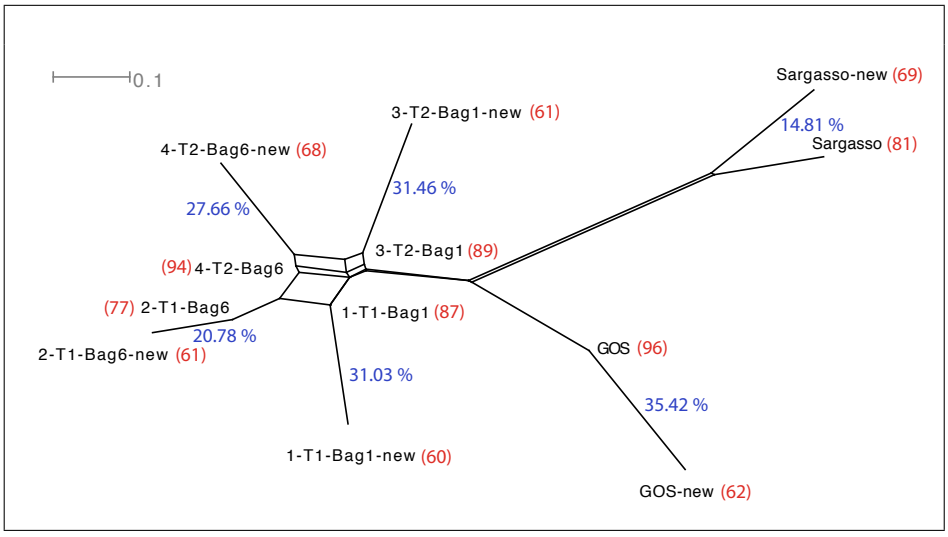

Figure 5.6: Network using Goodall's index from a control study to analyze the effect of excluding rare taxa. Comparison of six marine metagenomes with six truncated copies in which all rare taxa were excluded, analyzed at the 'Class' level of the NCBI. Number of taxon (Class) present in each data is reported in red and % of taxa or community (Class) changes in each control data is reported in blue.

After excluding all the reads from these 27 taxa, the new datasets ('1-T1-

Bag1-new') had 31.03% of community change. From the Figure 5.6 we can easily visualize that the distance between these two datasets is roughly proportional to 31.03% in relative sense when compared to other differences and proportions.

Case 5: Comparison of the 41 GOS datasets. We applied our approach to the geospatially-structured GOS data [Rusch et al., 2007] and computed two networks using Goodall's index, one considering all 41 sites and the second considering only the open ocean and coastal sites (Figure 5.7). Both networks exhibit a star-like structure, reflecting a high level of diversity in the data. Spatially related samples tend to cluster together, with the open ocean samples showing apparently fewer sample-specific taxa than the coastal ones.

Case 6: Comparison of 16S rRNA time series data from Western English Channel. To demonstrate the use of our method on 16S rRNA tag-pyrosequencing datasets, we applied it to the operational taxonomic units (OTUs) obtained from a continually monitored sampling site in the Western English Channel spanning Feb-Dec 2007 [Gilbert et al., 2009]. A comparison based on all $12,393$ OTUs from this time-series data set using Goodall's index leads to a highly unresolved network (Figure 5.8.A), which reflects the high abundance of rare taxa in the data across monthly samples. A more informative network can be obtained by excluding the OTUs found on only one occasion (considering 2666 OTUs, 22%) from the analysis (Figure 5.8.B). A network based only on those OTUs present in all data (71 OTUs, 0.5%) exhibits similar clusters, but as a result, a proportion of the distance information is lost (Figure 5.8.C). This network visually captures both the relationships between the samples and the seasonality of the dataset as previously described less adequately using traditional NMDS methods [Gilbert et al., 2009]. This analysis highlights the robust nature of Goodall's index in marker-based metagenomic studies, and also the importance of identifying rare taxa in these datasets.

Finally to establish the benefits of using this network representation, we prepared PCA and NMDS plot based only on those OTUs present in more than one time points (Figure 5.9). Unlike the NMDS plot (Figure 5.9.B), the network representation (Figure 5.8.B) provides a clear visualization of the distances between the different datasets, and unlike the PCA analysis (Figure 5.9.A) it suggests possible sample groupings. An obvious direct benefit is that the network representations provide a mix of the visual sensitivity of NMDS and PCA with the quantitative nature of classical dendrograms.

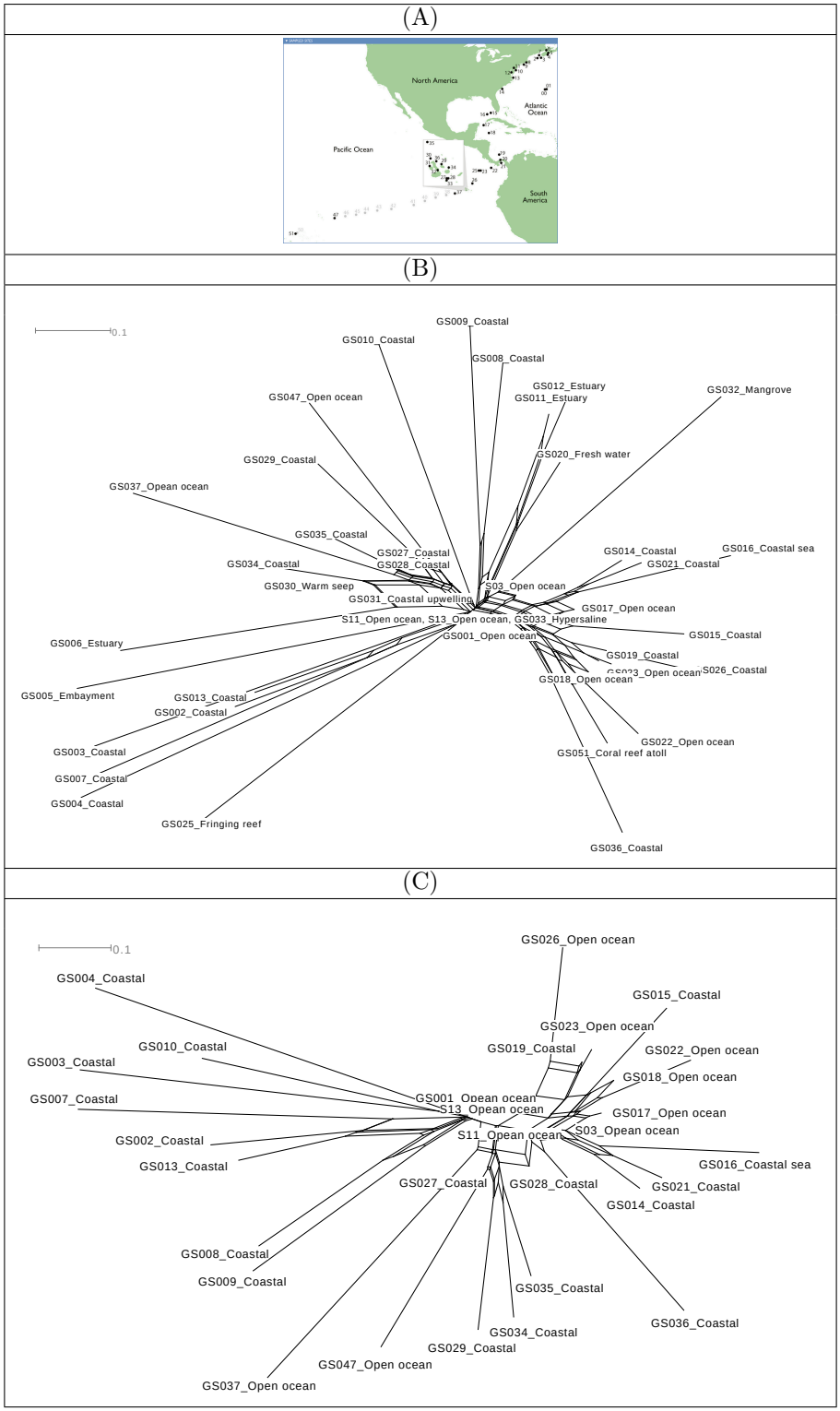

Figure 5.7: (A) The Global Ocean Sampling Survey (GOS) map (from [Rusch et al., 2007]), (B) the network considering all 41 sampling sites of GOS data and (C) the network considering only open ocean and costal sites using Goodall's index at the class rank of the NCBI taxonomy.

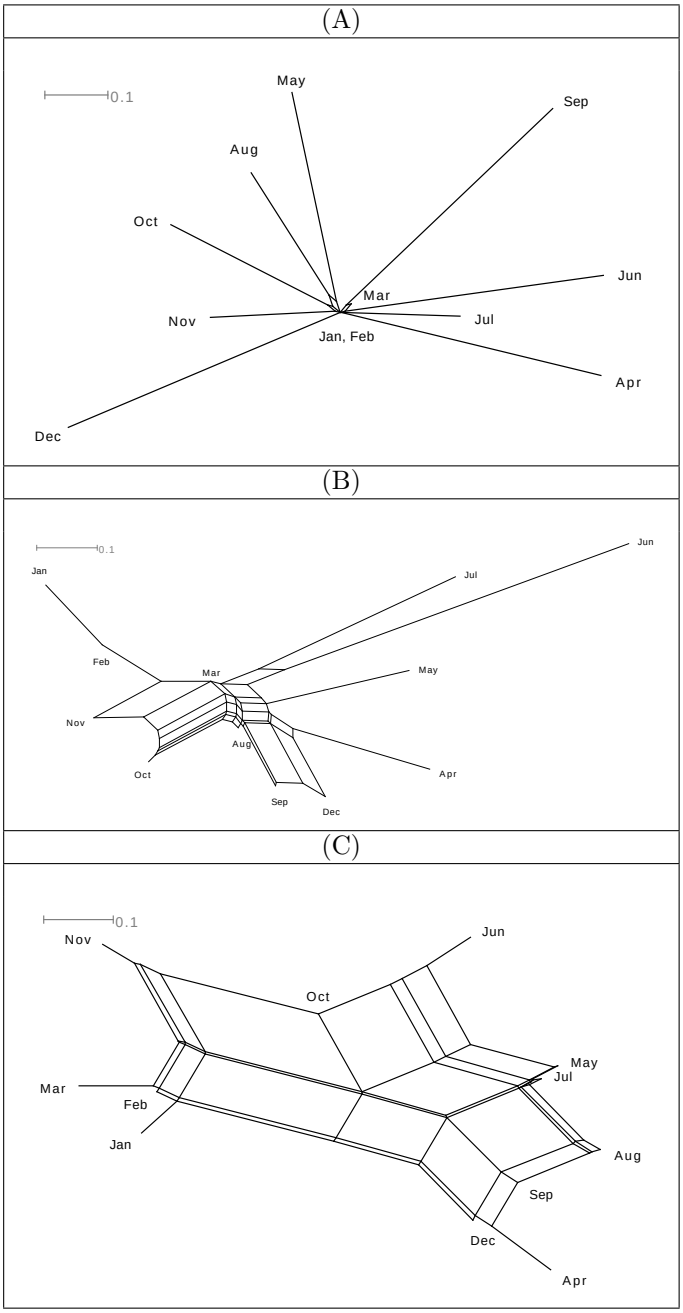

Figure 5.8: Comparison of 16S rRNA time series data from Western English Channel. Network using Goodall's index (A) considering all 12,393 OTUs, (B) Considering OTUs found at more than one time point (22%), (C) Considering OTUs found at all time points (0.5%).

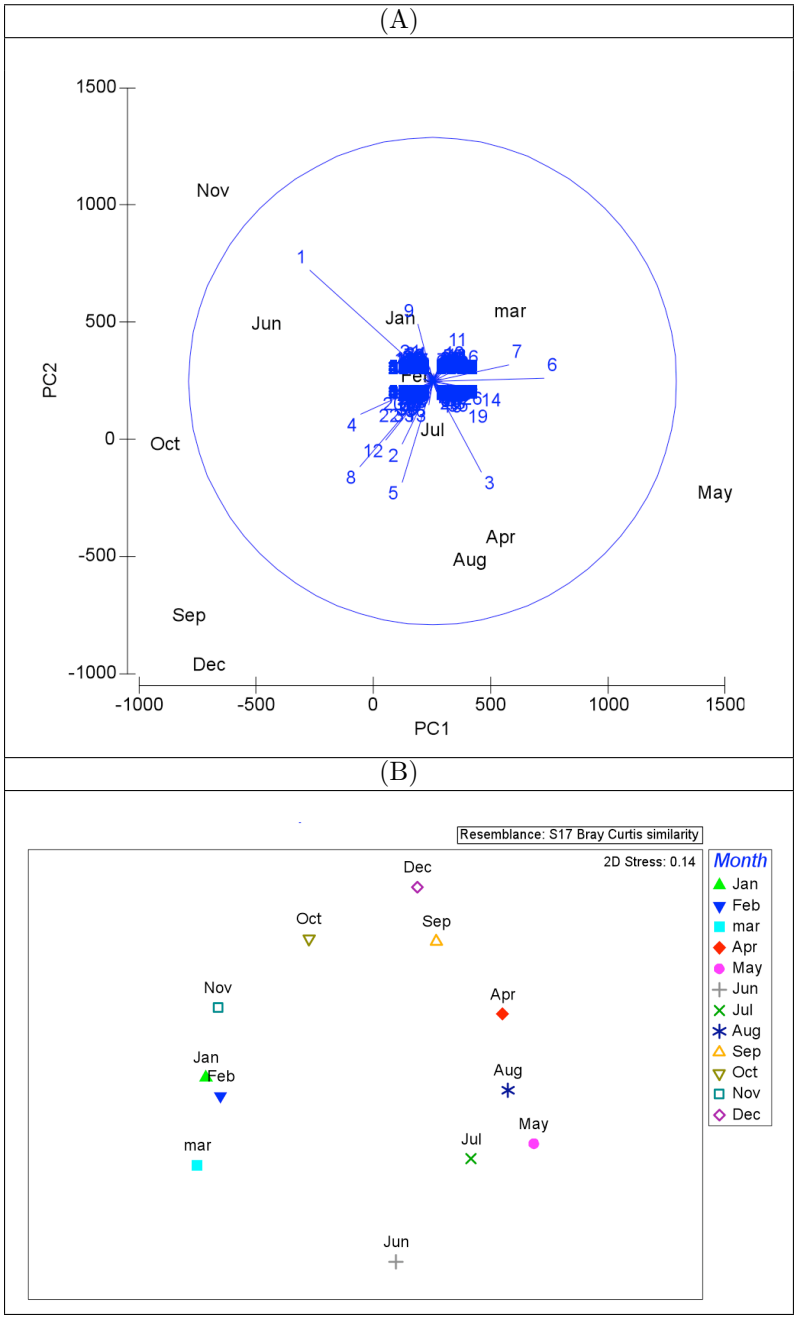

Figure 5.9: Comparison of 16S rRNA time series data from Western English Channel. (A) PCA Plot considering OTUs found at more than one time point (22%), (B) NMDS plot calculated from Bray-Curtis similarity matrix, obtained from OTUs found at more than one time point (22%)

5.3 Multiple Comparison of Functional Content using Networks

As the comparison of taxonomic content, the functional content of a collection of datasets can also be compared using six different ecological indices in a similar fashion. In Figure 5.10 we have compared eight above mentioned PLM-Bergen marine samples based on their functional content using six ecological indices.

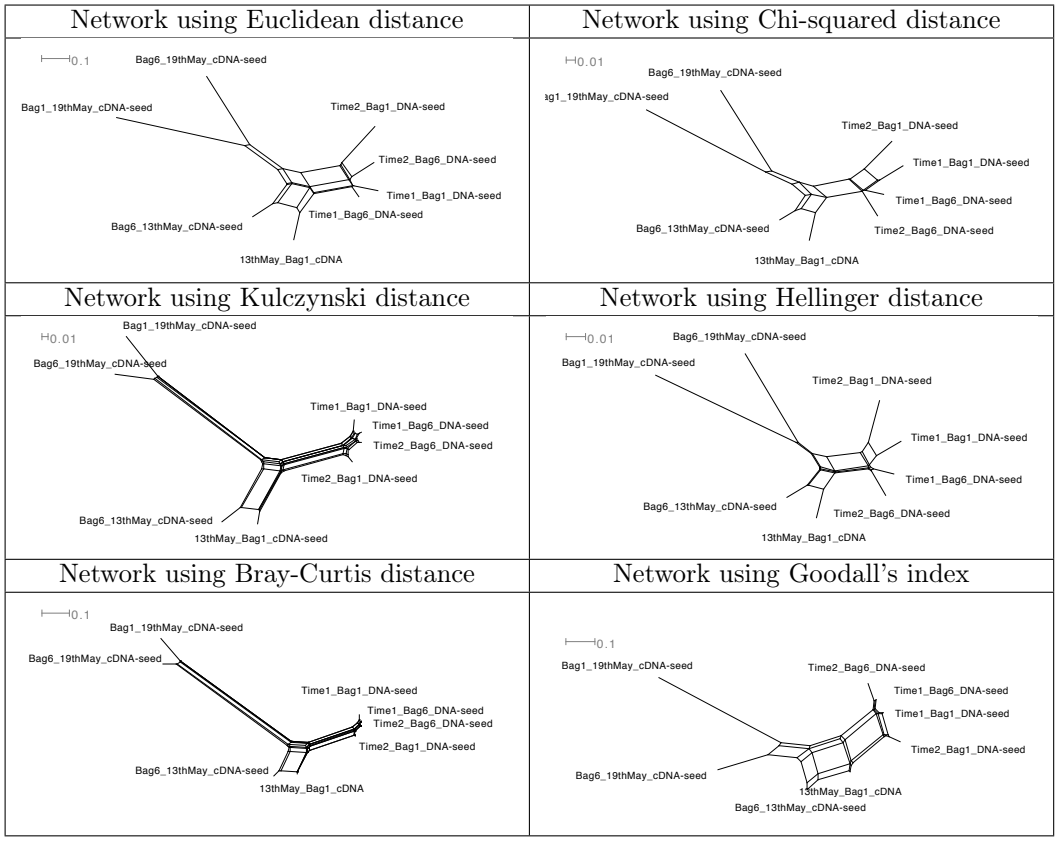

Figure 5.10: Network obtained using six ecological indices showing the functional comparison of eight PML-Bergen samples (four metagenomes and four metatranscriptomes) using SEED subsystems.

In these networks, the eight PML-Bergen samples demonstrated very similar clustering patterns using the functional data for analysis, compared with that obtained using just the taxonomic content. Therefore, this suggests that the chosen methodology is extremely robust for this type of comparative metagenomics.

5.4 Conclusion

Upcoming sequencing technologies are fueling a vast increase in the number and scope of metagenome projects. There is a great need for the development of new methods for visualizing the relationships between multiple metagenomic data sets. To address this, we have introduced a novel approach in this chapter that combines the use of taxonomic analysis, ecological indices and non-hierarchical clustering to provide a network representation of the relationships between different metagenome data sets. The approach was applied to different types of published data sets, including metagenomes, metatranscriptomes and 16S ribosomal profiles. Application of the approach to the same data using gene content summarized at different taxonomic levels (and also with functional content) gives rise to remarkably similar networks, indicating that the analysis is very robust. Importantly, the networks provide both a visual definition and metric quantification of the non-rooted relationship between samples, combining the desirable characteristics of other tools into one.

Chapter 6

Comparison of Sequencing Technologies for Metagenomics

From Section 1.4 we already know the theoretical background and advances of sequencing technologies. This chapter is devoted to solve the previously mentioned (in Chapter 2) question:

- *Which technology is most suitable for a particular metagenomic project?*

6.1 Overview

Most metagenome sequencing projects so far are based on Sanger sequencing [Venter et al., 2004, Rusch et al., 2007, Woyke et al., 2009]. The main advantage of Sanger sequencing is that the reads can be up to 1000 bp in length. Such long reads are desirable for a number of reasons. First, longer reads usually help to achieve longer and better matches to reference sequences, and so such reads can be assigned to specific taxa with higher confidence. Second, reads of this length can contain whole open reading frames and thus are very useful for finding new genes. Finally, the problem of assembling the most abundant species in a metagenome, when desired, is easier for longer reads. The main drawback of Sanger sequencing is the high price per base pair.

The first of the so-called "next generation" sequencing technologies, Roche-454 sequencing [Margulies et al., 2005], has become more and more popular as an alternative to Sanger sequencing. Now a read length of over 400 bp is possible, for a much lower price per base pair than Sanger sequencing. Short-read

sequencing technologies do have significant utility in whole genome sequencing projects because of their low cost and high throughput.

Until quite recently, the second next-generation sequencing technique to become commercially available, Illumina sequencing [Bentley, 2006], was not considered suitable for metagenomic studies because of its short read length in the range of 35 bp. Recent improvements support a read length of 75 bp, and such reads can now be collected in a paired-read protocol (see 1.4.4 for more details). Illumina sequencing has become an even cheaper option for metagenome sequencing among the currently available ones (see Table 1.1).

In this chapter the performance of two second-generation sequencing technologies are compared by simulating metagenomes. In particular the problem of taxonomic analysis of paired reads is addressed. In the simulation study we investigate the use of Illumina paired-sequencing in a taxonomical analysis and compare the performance of single reads, short clones and long clones. Because of the rather short reads (max 50 *bp*) [Metzker, 2010] SOLiD sequencing technology is still not suitable for metagenomics, so we exclude this from the comparison. In addition, we also compare against simulated Roche-454 sequencing runs. As the Roche-454 pair-end protocol requires an additional cloning step, it is not usually used in current metagenome projects. Therefore here we only consider Roche-454 single reads for comparison. The main hypothesis in our investigation is that, although the Illumina technology generates shorter sequences, the presence of paired reads will produce more specific taxonomical assignments when used with the "LCA-gene content" (lowest common ancestor) algorithm of MEGAN (see section 3.2 for details). We have supported our hypothesis through a large number of experimental results on three different metagenomes of different complexities. For the publication associated with this work please refer to [Mitra et al., 2010b].

6.2 Theory and Background

Here we describe a few details about different kinds of metagenome datasets, the program MetaSim [Richter et al., 2008], used for simulation and some fundamental technical knowledge about BLAST [Altschul et al., 1990].

6.2.1 Complexity of Metagenome Datasets

As described in 6.3.1, metagenomes can be classified into three different groups. Microbial communities, represented by a dominant population and flanked by low-abundance ones, are called 'low complexity' (LC) metagenomes. This type of metagenome can be found in bioreactors [Strous et al., 2006,

Garcia Martin et al., 2006]. 'medium complexity' (MC) communities have more than one dominant population, also flanked by low abundance ones, as seen in an acid mine drainage biofilm or the *Olavius algarvenis* symbionts [Tyson et al., 2004, Woyke et al., 2006]. If no dominant population is available, for example in the agricultural soil [Tringe et al., 2005], the community is called 'high complexity' (HC).

6.2.2 MetaSim – Metagenome Simulator

MetaSim is a program to generate collections of synthetic reads that reflect the diverse taxonomical composition of typical metagenome data sets. Based on a database of given genomes, this program allows the user to design a metagenome by specifying the number of genomes present at different levels of the NCBI taxonomy, and then to collect reads from the metagenome using a simulation of a number of different sequencing technologies.

As input MetaSim takes a set of known genome sequences and an abundance profile. This profile determines which genome sequences are selected for the simulation and the relative abundance of each genome sequence in the dataset. An 'induced tree view" of the NCBI taxonomy is then integrated. Furthermore users have the possibility to choose among different (adaptable) error models of current sequencing technologies, for example Sanger, Roche's 454 and Illumina. We used MetaSim for simulating metagenome datasets of different complexity. For more details of the program please refer to [Richter et al., 2008].

6.2.3 Basic Local Alignment Search Tool (BLAST)

Basic Local Alignment Search Tool, or BLAST [Altschul et al., 1990], is a tool which finds regions of local similarity between biological sequences, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables researchers to compare a query sequence with a library or sequence databases by identifying library sequences that resemble the query sequence above a certain threshold and calculates the statistical significance of matches.

6.3 Methods and Analysis

6.3.1 Simulation of Metagenomes and Sequencing

We used the MetaSim simulator to simulate the sequencing of three different synthetic metagenomes of different complexities using Roche-454 sequencing, Illumina paired-end sequencing of short clones and Illumina pair-end sequencing of long clones, as described in more detail below. For a fair comparison, the ratio of the total number of base pairs simulated for the Roche-454 and Illumina technologies was 1:10, based on the assumption that the price ratio between Roche-454 sequencing and Illumina paired-end sequencing is roughly of that order.

The three synthetic metagenomes were generated using whole-genome prokaryotic sequences downloaded from the NCBI website (April 2009), in accordance with the three profiles described in [Mavromatis et al., 2007]. In more detail, the three metagenomes are:

- A low complexity (LC) metagenome, consisting of 104 species and featuring the highly abundant species *Rhodopseudomonas palustris*;
- A medium complexity (MC) metagenome, consisting of the same 104 species including six highly abundant species: *Xylella fastidiosa Dixon, Rhodopseudomonas palustris BisB5, Bradyrhizobium sp. BTAi1, Xylella fastidiosa Ann-1, Rhodopseudomonas palustris BisB18* and *Rhodospirillum rubrum ATCC 11170*;
- A high complexity (HC) metagenome, consisting of the same 104 species, all at similar levels of abundance.

(Nine taxa mentioned in [Mavromatis et al., 2007] were not found in the NCBI database and thus were omitted from our analysis. Their taxon ids are: 155920, 155919, 165597, 332415, 322710, 286604, 321955, 333146 and 333849.)

We simulated Roche-454 reads for each of these three datasets with MetaSim using a setting of 98 flow cycles to obtain reads that are ≈ 250 bp in length. MetaSim models the basic base-calling procedure of Roche-454 sequencing. However, additional corrective post-processing is not simulated and so the errors reported here may be higher than what one would encounter in practice. For each dataset, we produced $6,000$ (non-paired) reads (see Table 6.1).

For each of the three synthetic metagenomes, we produced two different sets of Illumina reads with the goal of simulating the sequencing of both short clone and long clone libraries. The latest release of the MetaSim software provides an error profile for Illumina reads of length 36 bp. To obtain an error profile for

	LC-454	MC-454	HC-454
Simulated reads	6,000	6,000	6,000
Simulated base pairs	1,548,902 bp	1,541,252 bp	1,573,651 bp
Average read length	258.15 bp	256.88 bp	262.28 bp
Insertions	35,796 (2.3%)	35,425 (2.3%)	36,013 (2.3%)
Deletions	8,911 (0.5%)	8,839 (0.5%)	9,208 (0.5%)
Substitutions	0	0	0

Table 6.1: **Roche-454 reads statistics:** Summary of the Roche-454 reads generated by MetaSim for each of the three synthetic metagenome datasets LC, MC and HC.

longer reads of length 75 bp, we applied a non-linear regression ($f(x) = a \cdot e^{b \cdot x} + c$) to produce the best fitted error model ($a = 3.957e-4, b = 1.319e-1$ and $c = 5.362e-3$).

For the short clone library (S), we set MetaSim to generate clones according to a normal distribution with $\mu = 200$ bp and $\sigma = 20$ bp (see Table 6.2). For the long clone library (L), we set MetaSim to generate clones according to a normal distribution with $\mu = 1,900$ bp and $\sigma = 300$ bp (see Table 6.3). In total, we produced nine datasets of simulated reads:

- Roche-454 reads: LC-454, MC-454 and HC-454;
- Illumina reads, short clones: LC-ilm-S, MC-ilm-S and HC-ilm-S;
- Illumina reads, long clones: LC-ilm-L, MC-ilm-L and HC-ilm-L.

	LC-ilm-S	MC-ilm-S	HC-ilm-S
Simulated reads	200,000	200,000	200,000
Read length	75 bp	75 bp	75 bp
Clone length	200 bp	200 bp	200 bp
Simulated base pairs	15,000,000	15,000,000	15,000,000
Insertions	0	0	0
Deletions	0	0	0
Substitutions	227,913 (1.5%)	228,516 (1.5%)	227,279 (1.5%)

Table 6.2: **Illumina short-clone reads statistics:** Summary of the Illumina short-clone reads generated by MetaSim for each of the three synthetic metagenome datasets LC, MC and HC.

To be able to estimate the robustness of the results reported below, we additionally produced five replicates for each of the described datasets. Due to time constraints, in these replicates each Illumina datapoint was simulated using only 10,000 clones.

	LC-ilm-L	MC-ilm-L	HC-ilm-L
Simulated reads	200,000	200,000	200,000
Read length	75 bp	75 bp	75 bp
Clone length	1,900 bp	1,900 bp	1,900 bp
Simulated base pairs	15,000,000	15,000,000	15,000,000
Insertions	0	0	0
Deletions	0	0	0
Substitutions	227,880 (1.5%)	228,256 (1.5%)	228,262 (1.5%)

Table 6.3: **Illumina long-clone reads statistics:** Summary of the Illumina long-clone reads generated by MetaSim for each of the three synthetic metagenome datasets LC, MC and HC.

6.3.2 Sequence Similarity Search and MEGAN Analysis

We performed a MEGAN analysis of all nine datasets with their replicates. First, each of the datasets was compared against the NCBI-nr database (April 3, 2009 version) using BLASTX. Each of the nine BLASTX output files was then parsed and analyzed using MEGAN, as described in more detail in results (Section 6.4).

6.3.3 Processing paired reads in MEGAN

Reads from metagenomic datasets are usually processed in isolation (unless an assembly is attempted). MEGAN filters the BLAST matches obtained for a read by bit score. First, only matches that exceed a minimal bit score of 35, say, are kept (this is called the *min score* filter). Second, the hits are filtered further so that only those that attain a score that is within 10% (say) of the best score seen for the given read are kept (the *top percent* filter). For each hit that passes these two filters, MEGAN determines the corresponding species and then assigns the read to the LCA of the species of all hits, as outlined in section 3.2. A third filter, called the *min support* filter is then applied which removes all taxa from the reported result that were not hit by a specific number of reads.

To accommodate paired reads (see 1.4.4 for details), we have implemented a new *paired-reads mode* in MEGAN. After importing all reads, MEGAN processes each pair of reads in turn. In more detail, matches to the same organism from the two different reads are treated as one match. To give one of these paired matches more weight, we propose to combine the bit scores s_1 and s_2 from the two reads using the following equation:

$$s_{sum} = \sum_{i=1}^{r} s_i + \frac{r \cdot \ln(k) - \ln(km'n') - r(r-1) \cdot (\ln(k) + 2\ln(g)) - \log(r!)}{\ln(2)} \quad (6.1)$$

with $r = 2$, $k = 0.041$ (database parameter reported by BLAST), gap size $g = 50$, effective length of the query $m' = \max\left(\frac{1}{k}, m - h\right)$ (For BLASTX $m' = \max\left(\frac{1}{k}, \frac{m}{3} - h\right)$), query length m, effective HSP h, effective length of the subject $n' = \max\left(\frac{1}{k}, n - h\right)$ and subject length n. For more details on these parameters, see [Korf et al., 2003].

The number of organisms that are hit by both reads of a pair will often be smaller than the number of different organisms that are hit by either of the reads on their own. The modified bit score of two combined hits will often be more than 10% higher than the score of uncombined hits and so, in many cases, only the combined matches will pass the 10% filter. In consequence, the resulting LCA placement should be more specific.

MEGAN 4 is able to process sequencing reads in pairs and makes assignments of such reads based on the combined bit scores of their matches to reference sequences.

6.4 Results and Discussion

6.4.1 Short clones or long clones?

As mentioned before, the LCA-gene-content approach implemented in MEGAN suffers primary from a lack of resolution. A read that has a highly significant match to a sequence in the NCBI-nr database will often match with similar sequences from other organisms, as well, and thus may be placed on a higher-level taxon.

Assume that we have a set of reads collected using a paired-read protocol. If we process two reads (A and B) from the same clone simultaneously, then the distance between the two reads in the source genome (i.e. the length of the clone from which they were sequenced) will affect the performance of the LCA-gene content algorithm: If the two reads are close together, as in the case of short clones, then it is more likely that the two reads will come from the same gene and thus will display the same pattern of hits among species. If, on the other hand, the two reads lie much further apart in the source genome, as in the case of long clones, then it is more likely that the reads will come from two different genes, and these might show quite different patterns of conservation among species (see Figure 6.1). A main hypothesis is that more species will hit (that is, contain sequences that align to) both A and B if the reads come from a short clone than would be the case if the two reads come from a long clone.

Indeed, in the simulations reported below, we observed that whenever the

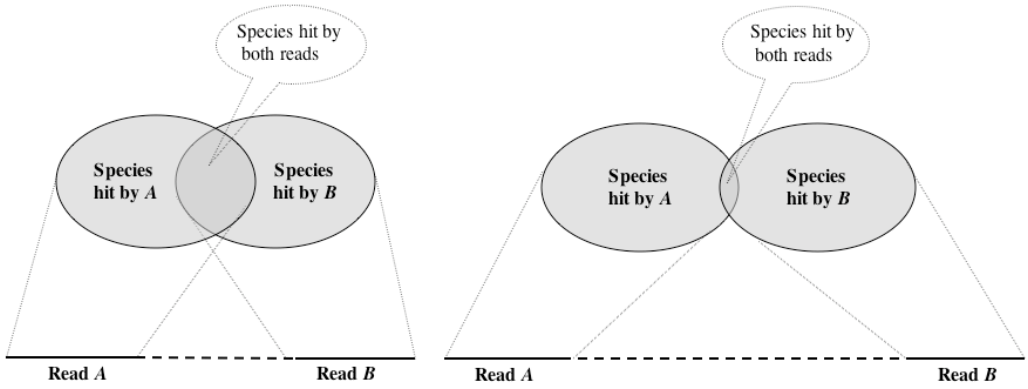

Figure 6.1: We assume that the intersection of the species hit by two reads A and B will be larger for pair reads obtained from short clones than for ones obtained from long clones. If this is the case, then use of the long clones in metagenome projects should lead to a more specific assignment of reads.

two reads of a short clone matched the same taxon, then this is due to matches to the same gene in over 80% of the cases, whereas for long clones this is true for just under 12%.

Thus, if we modify the LCA-gene content algorithm to place more weight on those species that are hit by both reads, then it should be the case that using long clones will give rise to more specific taxonomical assignments than when using short clones, without increasing the number of false-positive assignments. Moreover, it should, of course, be the case that processing both reads of a pair together will provide better results than processing each read in isolation.

6.4.2 Analysis of Roche-454 reads

The MEGAN analysis of the three different Roche-454 datasets, LC-454, MC-454 and HC-454, using the full NCBI-nr reference database, produced very few false negative species. Less than 6% of all species present in the synthetic metagenomes were not detected. Because of the low number of reads in each of the datasets (6,000 each), it is not surprising that some species of low abundance were missed. The false positive rate was zero for the LC-454 and HC-454 datasets, and less than 2% for the MC-454 dataset. Of course, the number of false negatives and the number of false positives both depend on the parameters applied, and the usual trade-off between false positives and false negatives can be observed. For these datasets, the best settings are $min\,score = 50$, $top\,percent = 10$ and the $min\,support = 3$ (see Figure 6.2 for the distribution of bit scores for each of the

three Roche-454 datasets).

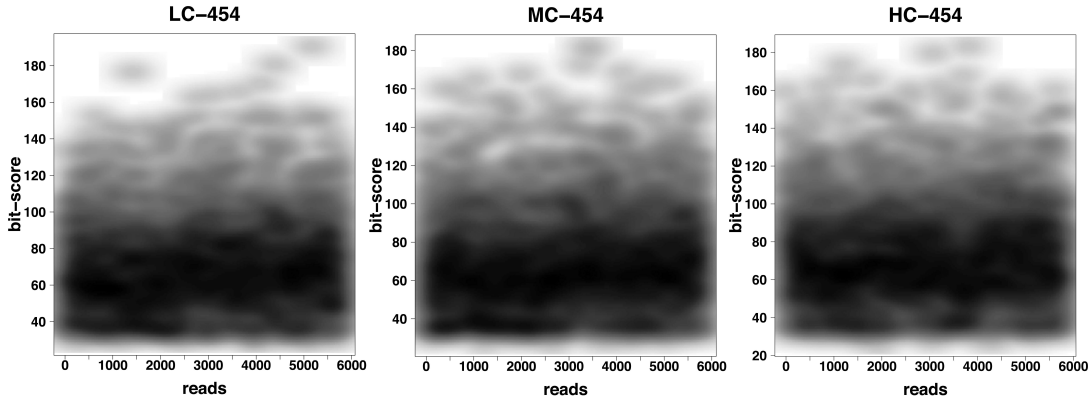

Figure 6.2: For each of the three simulated Roche-454 datasets, LC-454, MC-454 and HC-454, we plot the highest bit scores for all 6,000 reads.

Note that this analysis addresses only the problem of detecting specific species in the dataset, not whether individual reads have been correctly assigned. To obtain an indication of how well the individual reads are assigned to the correct species, in Figure 6.3, we compare the number of reads assigned to specific species against the number of reads actually simulated for each species. Here we also have normalized the data for this comparison. (The corresponding values for the five replicate datasets differ by between 10% (LC dataset) to 25% (HC dataset)).

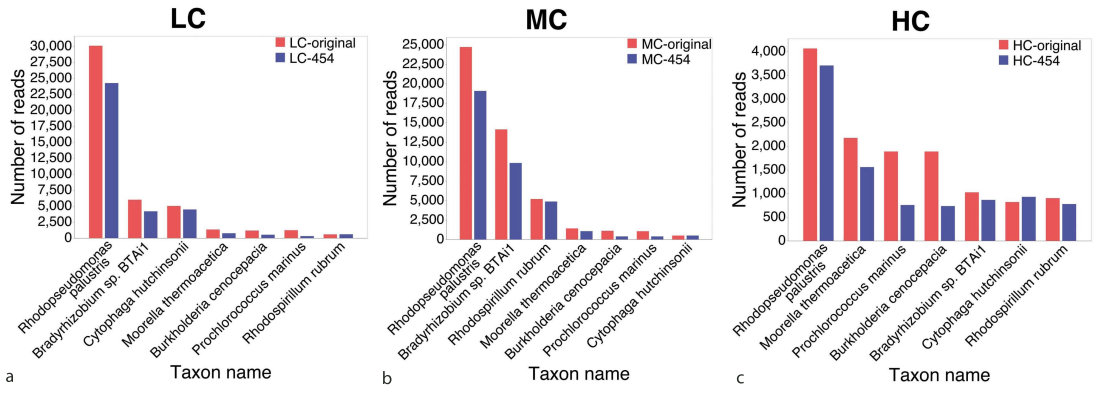

Figure 6.3: Blue bars indicate how many Roche-454 reads were assigned to seven different taxa, for each of the synthetic datasets LC (a), MC (b) and HC (c). Red bars indicate how many Roche-454 reads were actually simulated for each of the taxa. For ease of comparison, we have normalized the counts to a total of 100,000.

6.4.3 Analysis of Illumina reads

All six files containing simulated Illumina reads, LC-ilm-S, MC-ilm-S, HC-ilm-S, LC-ilm-L, MC-ilm-L and HC-ilm-L, were compared against the NCBI-nr database using BLASTX and then analyzed using MEGAN's paired-read mode. In addition, to simulate single Illumina reads, we used the reads from our Illumina long clone files and processed them with MEGAN as single reads.

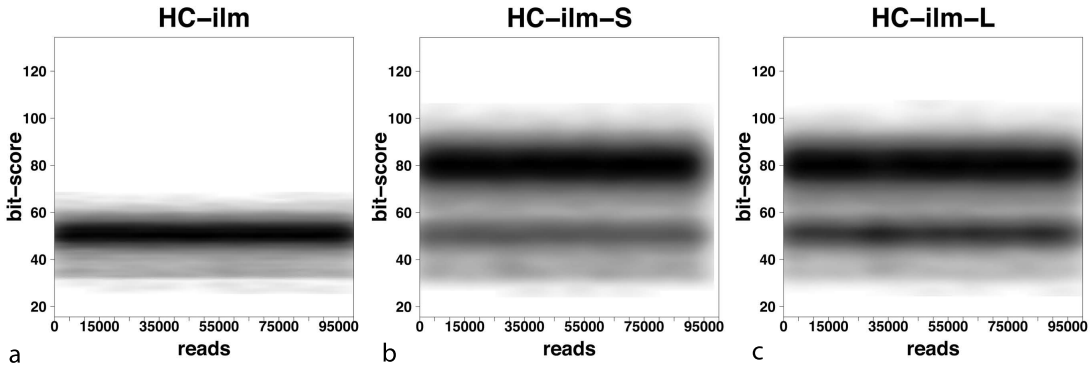

Figure 6.4: For each of 100,000 (normalized) reads sampled from the HC synthetic metagenome, we plot the highest bit score attained for (a) Illumina single reads (HC-ilm), (b) Illumina short-clone pairs of reads (HC-ilm-S) and (c) Illumina long-clone pairs of reads (HC-lim-L). The later two charts include the combined bit scores computed using equation (6.1). The plots for the LC and MC datasets look very similar and are therefore omitted

In Figure 6.4, we show the distribution of BLASTX bit scores for (a) single Illumina reads of the synthetic HC dataset and compare it with the distribution of bit scores for both the (b) short-clone and (c) long-clone libraries. In the latter two charts, the dark bands centered at 80 bits are scores obtained by combining the scores of paired reads using equation 6.1, as implemented in MEGAN's paired-read mode. These plots clearly show the effect of combining matches from paired reads. The attained bit scores are much higher and it is clear that using a top percentage filter setting of 10% will make MEGAN use only those species that are hit by both reads of a pair in the LCA computation, whenever such hits are present. While the average combined bit scores are not as high as the bit scores reported for the simulated Roche-454 reads (see Figure 6.2), they are nevertheless much higher than the Illumina single read scores (centered at 52 bits (Figure 6.4).

To obtain an indication of how well the individual reads are assigned to the correct species, in Figure 6.5 we compare the number of reads assigned to specific species with the number of reads actually simulated for each species, for the same species as above. Here also we have normalized the data for comparison. (The corresponding values for the five replicate datasets differ by between 5% (LC

dataset) to 25% (HC dataset).) In most cases, the number of assigned reads from long clones is larger than the number from short clones, which in turn is larger than the number of assigned single reads. In general, the number of false positive assignments is very small, except in the case of the HC dataset, where about 8% of the long-clone reads were falsely assigned to *Rhodopseudomonas palustris*.

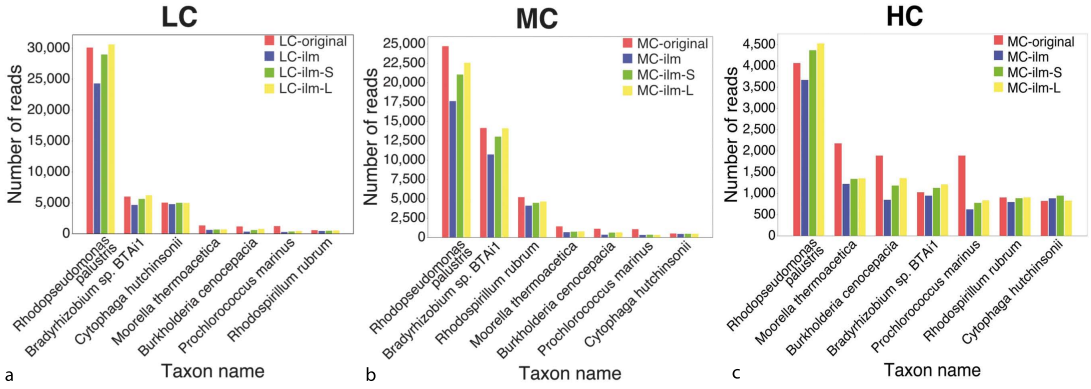

Figure 6.5: For seven key species, we indicate the number of simulated reads (red), along with the number of simulated Illumina single reads (blue), short-clone reads (green) and long-clone reads (yellow), assigned to the species by the LCA-gene content algorithm, for each of the three synthetic metagenome datasets LC (a), MC (b) and HC (c). (All values normalized to 100,000)

Reads are assigned to nodes at different ranks of the NCBI taxonomy, depending on how conserved their sequence is across species. In Figure 6.6, we show the number of reads assigned to nodes at different ranks of the NCBI taxonomy, from the phylum level down to the species level.

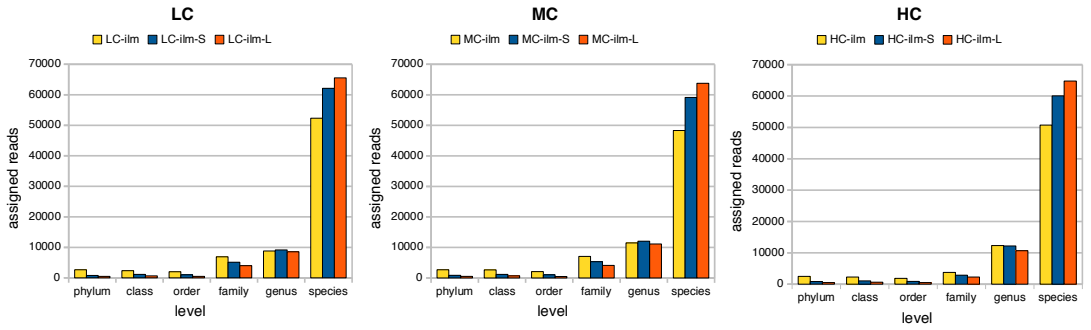

Figure 6.6: The number of reads assigned to nodes at different ranks of the NCBI taxonomy, from the phylum level down to the species level. These numbers are reported for Illumina single reads (yellow), short-clone reads (blue) and long-clone reads (red), for each of the three synthetic metagenome datasets. All datasets normalized to 100,000 for ease of comparison.

The rate of false positive assignments to nodes of the different levels is very close to zero, and so we do not distinguish between correctly and falsely assigned reads in this figure. These charts indicate that the assignment of reads to taxa is most specific for Illumina long-clone reads, slightly less specific for short-clone reads and even less specific for single reads.

6.4.4 The effect of unknown species

The study described so far simulates the situation in which all organisms in the metagenome are represented by sequences in the reference database. In practice, a metagenome usually contains a significant percentage of unknown organisms, which are not represented in the reference database. To mimic this situation, we decided to rerun the analysis while ignoring all BLAST matches to any taxon in the genus of *Rhodopseudomonas*. In Figure 6.7, we show the performance of MEGAN for both short clones and long clones. When using the whole of the NCBI-nr database as a reference, we can assign $60,000 - 65,000$ reads at the species level, with a very small number of false positive assignments.

When we remove the genus of *Rhodopseudomonas* from the reference database, the percentage of reads assigned to species drops by a number roughly proportional to the number of reads that were actually sampled from the genus. In this case, the number of false positive assignments rises to about 2.5%, while most of the reads that were sourced from the "unknown" genus are classified as unassigned and are thus considered false negatives. This confirms that the LCA-gene content method for taxonomical analysis is indeed quite conservative in that unknown sequences are much more likely to produce false negatives than they are to produce false positives.

6.4.5 Choice of MEGAN parameters

This taxonomical analysis of simulated Illumina reads was performed using the following MEGAN parameters: $min\,score = 50$, $top\,percent = 10$ and $min\,support = 50$. The most crucial parameter is the *min score*, which prescribes the minimal bit score that a match must achieve to be considered in the analysis. For single reads of short length, the program's recommended setting of this parameter is 35 bits. Figure 6.4 indicates that a *min score* of 40 or 45 might be more suitable, as it will be more specific, while still allowing most reads to be placed.

For paired-reads, Figure 6.4 suggests that using only those BLAST matches whose bit score exceeds 50 should perform very well. With this setting, for any

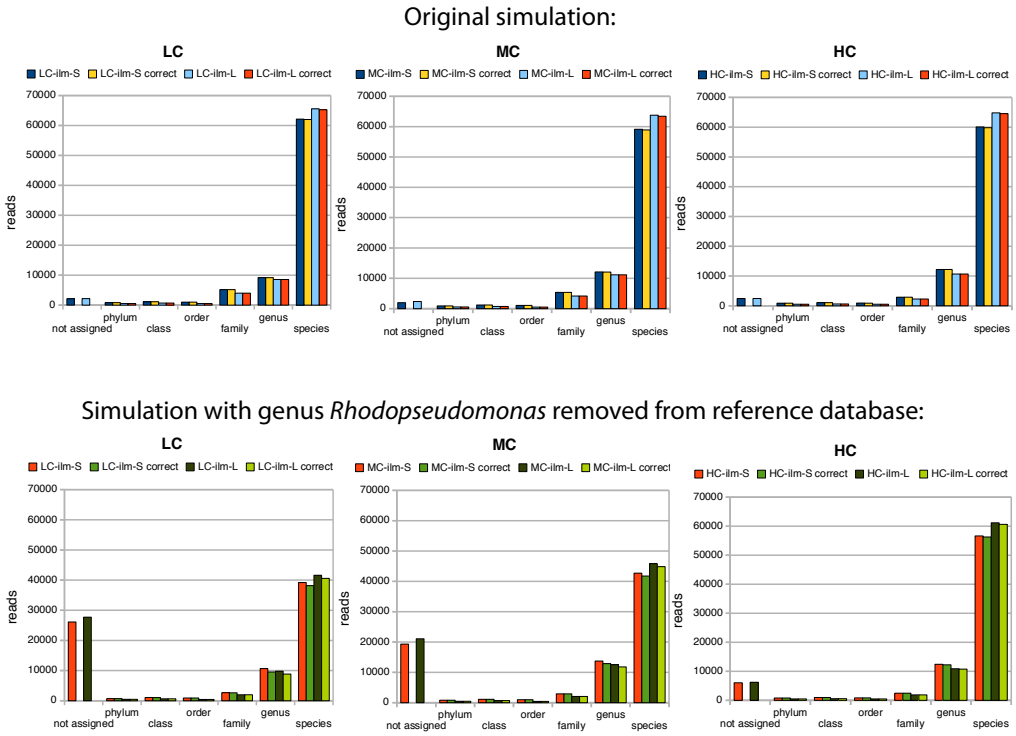

Figure 6.7: In the top row, we show the number of assigned and correctly assigned Illumina short- and long-clone reads at different taxonomical ranks. In the bottom row, we show the same quantities for a taxonomical analysis performed with the entire *Rhodopseudomonas* removed from the reference database.

pair of reads that has combined matches, only the combined matches will be used, as the bit scores of single-read matches will not pass the *top percent* filter. In cases where a pair of reads does not give rise to a pair of combined matches, then only very high-scoring single-read matches will be used.

To determine a recommended setting for the *min support* filter for Illumina paired reads, we studied the number of false positive and false negative assignments for both short-clones and long-clones for a number of different settings. All three synthetic metagenomes, LC, MC and HC, gave similar results, and so we only show the results for the HC dataset in Figure 6.8. Our studies suggest that $min\ support = 50$ is a good choice, as it minimizes both the number of false positives and false negatives while giving higher or conservative support value. However choice of parameters always depends on the kind of the study.

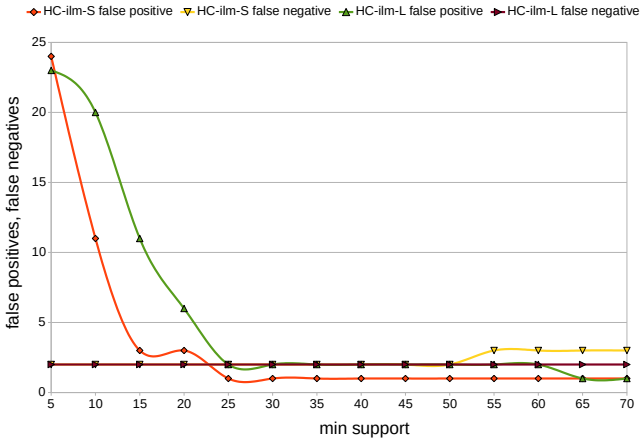

Figure 6.8: For the synthetic HC metagenome, we report the number of false positives for the short clones (red) and long clones (green), and the number of false negatives for the short-clones (yellow) and the long clones (brown), as a function of the minimal number of hits required for a species to be considered detected.

6.4.6 Comparison between Roche-454 and Illumina

How do reads of length 250 bp compare against paired reads of length 75 bp? In Figure 6.9, for each of the three synthetic metagenomes, we report the number of reads that were correctly assigned on the species level, for Roche-454 sequencing, and Illumina single reads, short-clone reads and long-clone reads. In all cases, the number of falsely assigned reads is close to zero. The rest of the correctly assigned reads are usually assigned to the higher level in the taxonomy. Our study suggests that a higher percentage ($\approx 8\%$) of Illumina paired reads than of Roche-454 single reads are correctly assigned to species.

As we indicate above, long clones are more specific than short clones, because they lead to placements based on well-separated reads. This argument carries over to Roche-454 reads as well: While the reads are longer and thus support longer and more significant BLAST matches, the matches will usually reflect the gene content pattern of only one gene, rather than two.

How much of the difference between the results for the Roche-454 and the Illumina long-clones sequences is due to the different types of errors produced by the two different sequencing technologies? To investigate this, we generated an additional dataset covering all read lengths and clone lengths described above, but without applying any sequencing error models. Figure 6.10 shows the assignments for these error-free reads. This analysis is analogous to Figure 6.9 and exhibits a slightly different ranking of protocols by increasing performance, namely first short single reads, then short clones, then long single reads and then long clones.

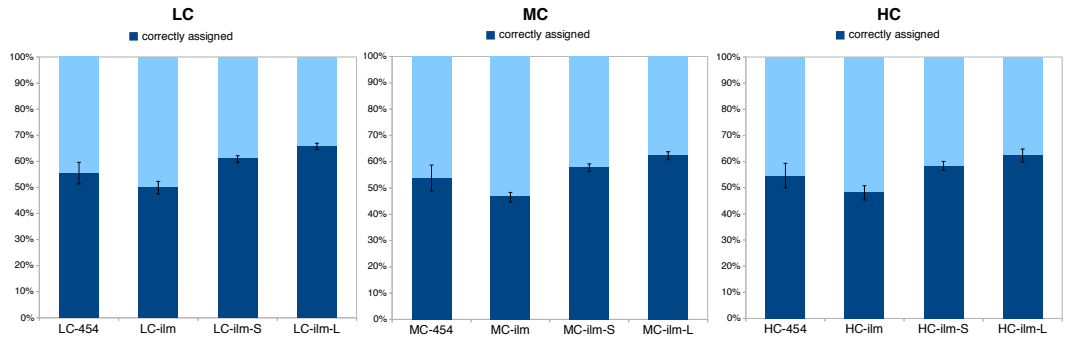

Figure 6.9: The percentage of correctly assigned reads (dark blue) to nodes at the species level of the NCBI taxonomy, averaged over the five replicate datasets, with error bars indicating the range of all five values. These numbers are reported for Roche-454 single reads (labeled 454), Illumina single reads (ilm), Illumina short-clone reads (ilm-S) and Illumina long-clone reads (ilm-L).

The gain of long-clone data (75 bp paired reads) over long single-read data (250 bp reads) is still significant at $\approx 4\%$.

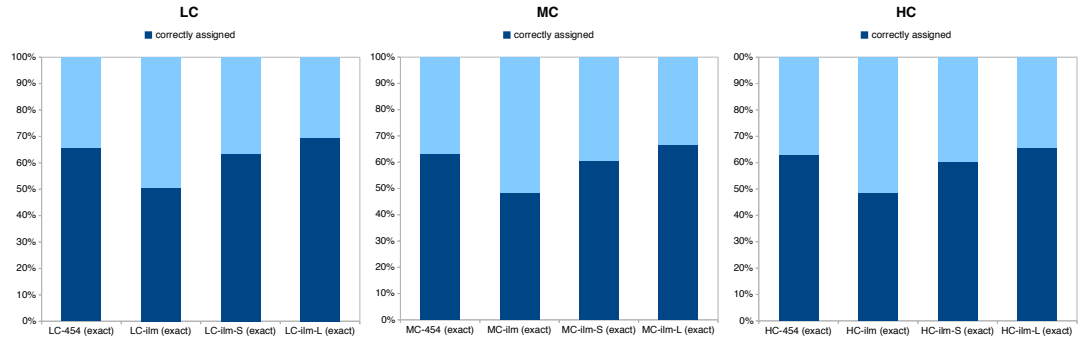

Figure 6.10: The same notations as Figure 6.9, but without applying any sequencing error models.

6.5 Conclusion

In this chapter, we have investigated the question whether the taxonomical analysis of metagenomic datasets can be performed by Illumina paired reads, and, if so, whether short clones or long clones should be used. Our simulation study suggests that Illumina paired reads are well-suited for this task and that long clones are more specific, even compared to much longer Roche-454 reads (of length 250 bp),

when using the LCA-gene content algorithm. We have represented that this is due to the fact that the placement of reads from long clones are based on the gene-content pattern of two different genes, rather than just one. This is a general observation that will probably affect other analysis methods that consider paired reads, as well.

Because Illumina sequencing is much cheaper than Roche-454 sequencing, it is clear that future metagenomics projects will use Illumina sequencing, as well as Sanger and Roche-454 sequencing.

Chapter 7

Application in Metagenomic Projects

This chapter describes the application of previously mentioned methods in four metagenome projects: detection and diversity of pathogenic Vibrio in coastal Fiji waters, ocean acidification study, seasonal and diel marine bacterial function and mammoth microbiome. Here I will briefly describe my contributions in these projects.

7.1 Detection and diversity of pathogenic Vibrio in coastal Fiji waters

This study[1] describes the metagenomic analyses to examine the diversity of Vibrio species in the coastal waters around Suva, Fiji.

7.1.1 Overview

Members of the genus Vibrio are gram-negative, motile rods. These bacteria are ubiquitous in marine environments where they form associations with a wide array of eukaryotes. Strains of several Vibrio species are clinically important

[1] *The content of this section is submitted for publication as a part of the study done by Reema Singh, Vinay Narayan, Patricia McLenachan, Richard C. Winkworth, Suparna Mitra, Peter J. Lockhart, Lorraine Berry, Abdulla M.Hatha, William Aalbersberg and Dhana Rao*

human pathogens. Commonly, disease-causing forms are associated with gastrointestinal infections (e.g., V. cholerae, V. parahaemolyticus and V. vulnificus). Perhaps clinically most important is V. cholerae. Toxigenic strains of this species are the causative agent of cholera, a disease that claimed millions of lives during the 19th century and currently affects $3 - 5$ million people annually [Thompson and Klose, 2005, Hunt et al., 2008].

Little is known about the distribution and prevalence of disease-causing Vibrio species in the Pacific Islands. During the late 1970s the O1 El Tor strain was responsible for a cholera outbreak in the South Pacific. More recently the identification of non-toxigenic O1 strains in Fiji suggests a silent reservoir of V. cholerae remains [Nair et al., 2006]. Interestingly, Fijian public health records indicate an increasing incidence of diarrhea (Ministry of Health Bulletin, 2009). Although there is no data linking this increase to food-borne disease, given that fish and shellfish are common vectors and raw fish is frequently eaten in Fiji, it is possible that these diseases may be linked to Vibrio infections.

The goal of this study is to investigate the occurrence of pathogenic Vibrio species on fish available for consumption and in the coastal waters around Suva, Fiji. Biochemical tests were used to screen fish sold at retail outlets in Suva for the presence of Vibrio. These tests suggest the presence of a moderately diverse community of these bacteria. Phylogenetic analyses of three markers (i.e. $16S$, recA and pyrH) confirm the presence of V. parahaemolyticus and suggest this species is represented by several genotypes. Both clinical and non-clinical species were associated with the sampled fish. Illumina GAII sequencing and MEGAN analyses were used to detect Vibrio in seawater samples; this approach identified several Vibrio species. Consistent with the fish screening we detect the pathogenic V. cholerae and V. parahaemolyticus in the coastal water column. This section describes the metagenomic analysis which is a part of study to examine the diversity of Vibrio species in the coastal waters around Suva, Fiji.

Methods for Metagenomic Analysis of Seawater Samples

Seawater samples of 60 L were collected at a depth of 5 m from two open water sites, one in the Suva Harbour and the other near Beqa Island. Samples were pre-filtered through $5.0\,\mu m$ membrane disc filters (Milipore) to remove debris and microorganisms collected by pressure filtration through $0.8\,\mu$m and $0.22\,\mu$m filters (Pall Life Sciences). Total community DNA was extracted using a modified version of the protocol described by [Venter et al., 2004]; DNA was extracted separately from $0.8\,\mu$m and $0.22\,\mu$m filters with the DNA pooled following extraction.

A paired-end genomic DNA library was prepared by (i) fragmenting puri-

fied genomic DNA using an Invitrogen nebulisation kit, (ii) ligation of paired-end index Illumina adaptors and (iii) fragment enrichment using the Illumina Multiplex Paired End Genomic DNA library preparation protocol (18 cycles of PCR were used). Enriched libraries were quantified and quality checked then diluted to 10 nM using EB buffer (Qiagen) and quantified for optimal cluster density. Libraries were amplified in a single flow cell lane on the Illumina Cluster Station instrument using the Illumina Paired End Cluster Generation kit $v2$ with a cluster density of 140,000 per tile and a molarity of 13 pM. Amplified libraries were sequenced using a 75 base paired-end indexed run on an Illumina GAII instrument; sequencing reactions used the Illumina 36 cycle SBS sequencing kit $v3$ with Multiplex Sequencing Primers and PhiX control kit $v2$ (Illumina). After sequencing, the images were analysed using the Illumina pipeline (version 1.3).

For metagenomic analysis a random sub-sample of 100,000 pair-end fragments was drawn from the full set of fragments generated by Illumina sequencing. Fragments were first aligned against the NCBI-NR (non-redundant protein; February 2010 version) database using BLASTX (translated DNA to protein; Altschul et al., 1990). Taxonomic assignment of fragments was then made using the conservative lowest common ancestor (LCA) algorithm implemented in MEGAN version 4.0 alpha1 [Huson et al., 2007]. Analyses used a bit score threshold of 35 and the February 2010 version of the NCBI-database [Benson et al., 2005]. To examine assignment sensitivity a pair of MEGAN analyses were performed. The first analysis, conducted using the standard algorithm, treated paired end fragments as two single reads (i.e., the 100,000 paired-reads are treated as 200,000 individual reads). The second analysis treated pairs of reads together, with matches found using both given greater weight (for more details of paired read mode see [Mitra et al., 2010b]). Final species profiles were inferred from preliminary lists by excluding taxa with only one assigned read.

7.1.2 Result and Discussion

BLAST searches of the random sub-sample of reads (100,000 pairs) resulted in 1,942,041 significant matches to sequences in the NCBI-NR database (Benson et al., 2005). Analyzing the BLAST output using LCA algorithm (with a bit score threshold 35) MEGAN assigned 23,561 reads to taxonomic groups. Of the remaining reads, 23,216 were unassigned because the bit score for matches was below the threshold and 153,223 were unmatched to sequences in the NCBI-NR database; the "Not assigned" and "No hits" categories, respectively. Of assigned reads, 138 corresponded to the Vibronales clade; 90 of those to the genus Vibrio and 7 to Photobacterium. The analysis assigned reads to five Vibrio species - V. cholerae, V. harveyi, V. parahaemolyticus, V. shilonii and V. vulnificus (each

represented by $7-14$ reads). Further, reads were assigned to strain level within three species (Figure 7.1.1.A).

Paired-read analysis assigned $42,354$ reads to taxonomic groups; $15,256$ reads were not assigned and $142,390$ had no hits. In general, the number of reads assigned to a given node in the paired end analysis is close to double that of the corresponding node in the single read analysis (Figure 7.1.1.B). Paired read analysis assigned a total of 284 reads to the Vibronales clade and identified three genera of the Vibrionaceae -Aliivibrio (6 reads), Photobacterium (12 reads) and Vibrio (185 reads). Within Vibrio, six species (those identified in the single read analysis plus V. splendidus) and strains within four species were identified using the paired end approach. Species and strain identifications were based on $6-29$ reads.

Single and paired end MEGAN analyses assigned GAII sequencing reads to five and six Vibrio species, respectively (Figure 7.1.1). These analyses indicate both clinical (e.g., V. cholerae and V. parahaemolyticus) and non-clinical (e.g., V. harveyi and V. splendidus) species are present in the water column close to Suva. As with the fish-associated community the presence of clinically important species represents a potential health risk. All of the species identified in this analysis were also identified as potential members of the fish-associated community. This result suggests that water-column and fish-associated communities are linked. A result that is perhaps not unexpected. However, if we assume that the water column species is present when it is among the set of species compatible with the biochemical results then several species appear to be missing from the water column. This absence might be explained in several ways. One possibility is that it reflects real differences between the communities. Perhaps some vibrios are simply rare in the water column but commonly associated with eukaryotes. Alternatively these differences may reflect sampling issues. In this context it may be that our original water samples did not contain representatives of all vibrio species present in the water column. This seems be likely if certain taxa are rare; certainly the low number of reads associated with the detected species suggests they are not common community members. Also by sub-sampling reads we may have influenced the identification of taxa by MEGAN. Again this would likely have had the most affect on rare community members. Repeating MEGAN analyses using alternative sub-samples could be used to examine whether this approach significantly changes the results.

Our analyses suggest a moderately complex Vibrio community is associated with marine environments close to Suva, Fiji. Although our analyses cannot fully characterize the community it is clear that clinically important species, including V. cholerae and V. parahaemolyticus, are among those species present. Further investigation is needed in order to assess whether the strains present pose direct

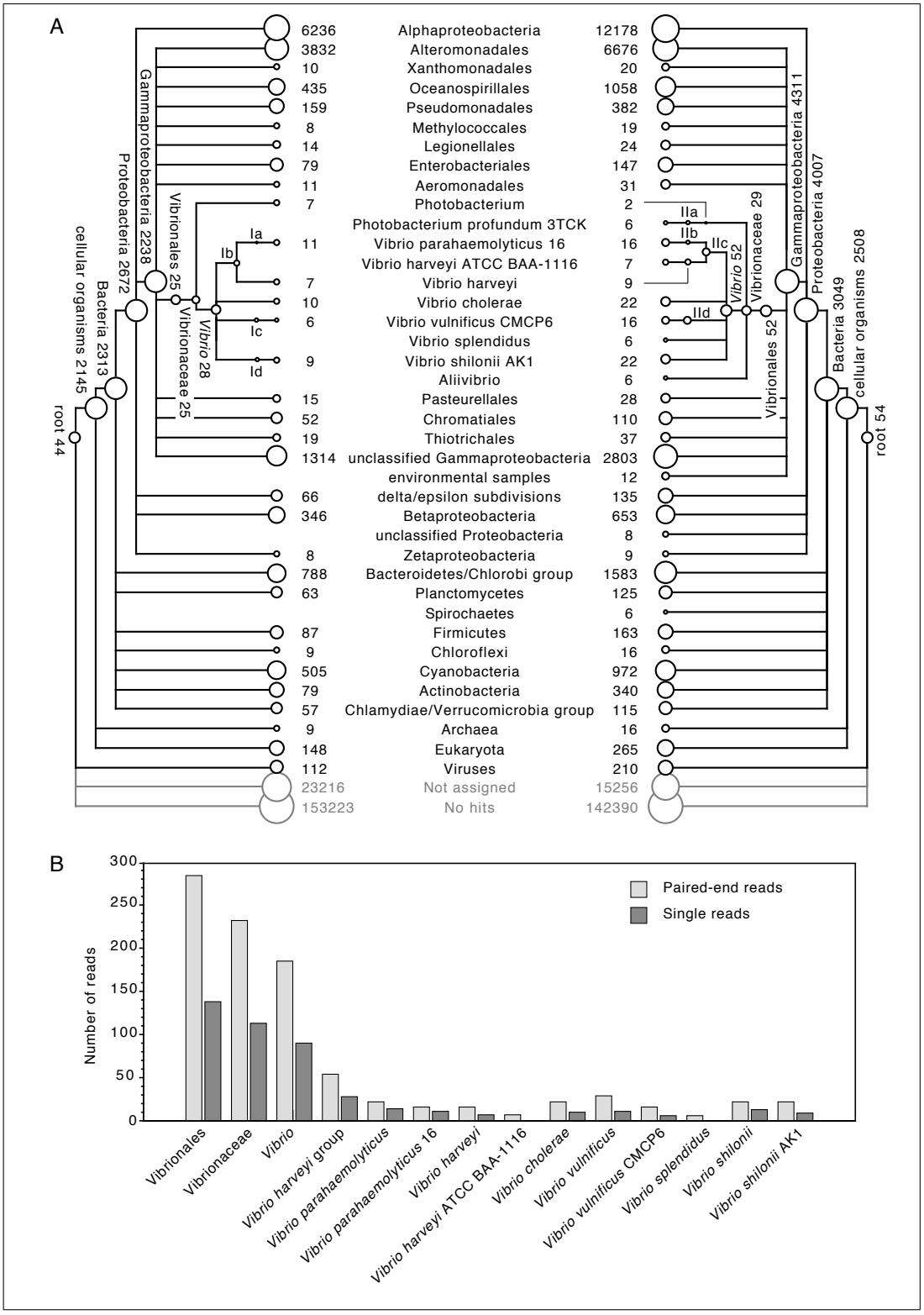

Figure 7.1.1: (A) left hand side - MEGAN taxonomic assignment of single reads; right hand side - MEGAN taxonomic assignment of pair-end reads. (B) bar plot comparing number of assigned reads to vibrios for single and pair-end analysis.

health risks. Given the potentially serious health implications of Vibrio infection monitoring these bacteria in the environment may provide an important tool for managing health risks. Towards this end we investigated whether Illumina GAII sequencing in conjunction with MEGAN analyses could be used to monitor bacterial communities. Our results suggest that this approach may offer an effective and efficient method for biomonitoring microorganisms in coastal waters and presumably other environments.

7.2 Ocean Acidification Study

In this study[2] we tried to investigate the impact of ocean acidification on diversity of *Alphaproteobacteria* in a coastal mesocosm. The hypothesis of this study is that *Alphaproteobacteria* diversity within the community will be altered by reduced pH conditions as predicted for the year 2100.

7.2.1 Overview

Ocean acidification is the name given to the ongoing decrease in the pH of the Earth's oceans. The anthropogenic carbon dioxide (CO_2) is absorbed by the global marine ecosystem, causing an increase in carbonic acid and hence a reduction in pH. The pH level is expected to decrease by 0.4 pH units from 8.2 in the present-day to 7.8 in the year 2100 [Nakicenovic et al., 2001]. Since the 19^{th} century, anthropogenic contribution has raised the atmospheric concentration to nearly 380 parts per million (ppm), which have remained stable at 200 to 280 ppm over the last $400,000$ years. This is mainly due to the uncontrolled burning of fossil fuels, industry and land use [Feely et al., 2004]. It is extremely important to understand the impact that this acidification will have on marine biodiversity and ecosystem functioning. The Royal Society [Riebesell et al., 2007] report recommended the use of mesocosm experiments, among others, to investigate the effects of ocean acidification on marine bacterioplankton. Members of the sub-phylum *Alphaproteobacteria* are key microbial components of the marine ecosystem and constitute as much as 50% of the total bacterial abundance, with global distribution. They play a major role in key biogeochemical cycles, especially sulphur and carbon cycling.

This study aims to assess the phylogenetic diversity of the bacterial subphylum *Alphaproteobacteria* in mesocosms exposed to high CO_2 conditions and

[2] *The content of this section will be submitted for publication as a part of the study done by Jack A. Gilbert, Samantha Craven, Ana Stores-Fernandez, Suparna Mitra, Ben Temperton, Colin Munn and Ian Joint.*

present-day conditions using culture-independent metagenomic approaches. The goal is to investigate the potential impact that ocean acidification will have on the diversity of this class through the use of three different molecular techniques, metagenomic pyrosequencing, fosmid clone libraries and 16S rDNA clone libraries.

7.2.2 Methods

An ocean acidification mesocosm experiment was run by Plymouth Marine Laboratory during May 2006 at the large mesocosm facility near Bergen, Norway [Gilbert et al., 2008a]. Out of six mesocosms, anthropogenic CO_2 was induced to create ocean acidification condition into mesocosms $1-3$ to reduce pH by increasing the pressure of atmospheric CO_2 (pCO_2) concentration; for the control, air was bubbled into mesocosms $4-6$. A phytoplankton bloom was induced by adding phosphate and nitrate to all six mesocosms. Mesocosms 1 and 6 were used as experimental bags for this study. The samples were taken at two time points, at the peak (Time 1 or 13^{th} May) and immediately after the collapse of the bloom (Time 2 or 19^{th} May) (please refer to [Gilbert et al., 2008a] for details of the samples). Sampling strategy, DNA extraction and sample preparation for the four metagenomic sequence databases are detailed in [Gilbert et al., 2008a].

As a part of analysis, the MEGAN software [Huson et al., 2007] was used to refine and cluster taxonomic output from BLASTX comparison. The 'Directed Homogeneity test' of MEGAN [Mitra et al., 2009], gives an impression of significant differences in taxon abundance caused by the induced acidification (see Section 4.2 for details of this method). Changes in the diversity of the *Alphaproteobacteria* are shown to be significant in a comparison between high CO_2 and present-day CO_2 treatments in a coastal marine ecosystem mesocosm.

7.2.3 Result and Discussion

A comparative metagenomic species profile is depicted in Figure 7.2.1 and 7.2.2. In Figure 7.2.1 we see that for the 13^{th} May datasets the differences between the communities in each treatment mainly lie in the *Alphaproteobacteria* node. In *Alphaproteobacteria* the up p-value ($UPv = 2.78E-14$) value demonstrates that the difference in the proportions of occurrences of this node in Proteobacteria was statistically significant between these the datasets. The down p-value ($DPv = 0.0$; all p-values $< 1E-37$ are 0.0) suggests that the observed difference is not only due to this node but is also a result of the abundances of low hierarchical taxa (orders, families, genera and species). Those nodes which most contribute to this difference are Rhodobacterales ($UPv = 0.0, DPv = 4.89E-6$) and Rickettsiales ($UPv = 0.0, DPv = 0.05$) (Figure 7.2.1). For Rhodobacterales, the DPv again

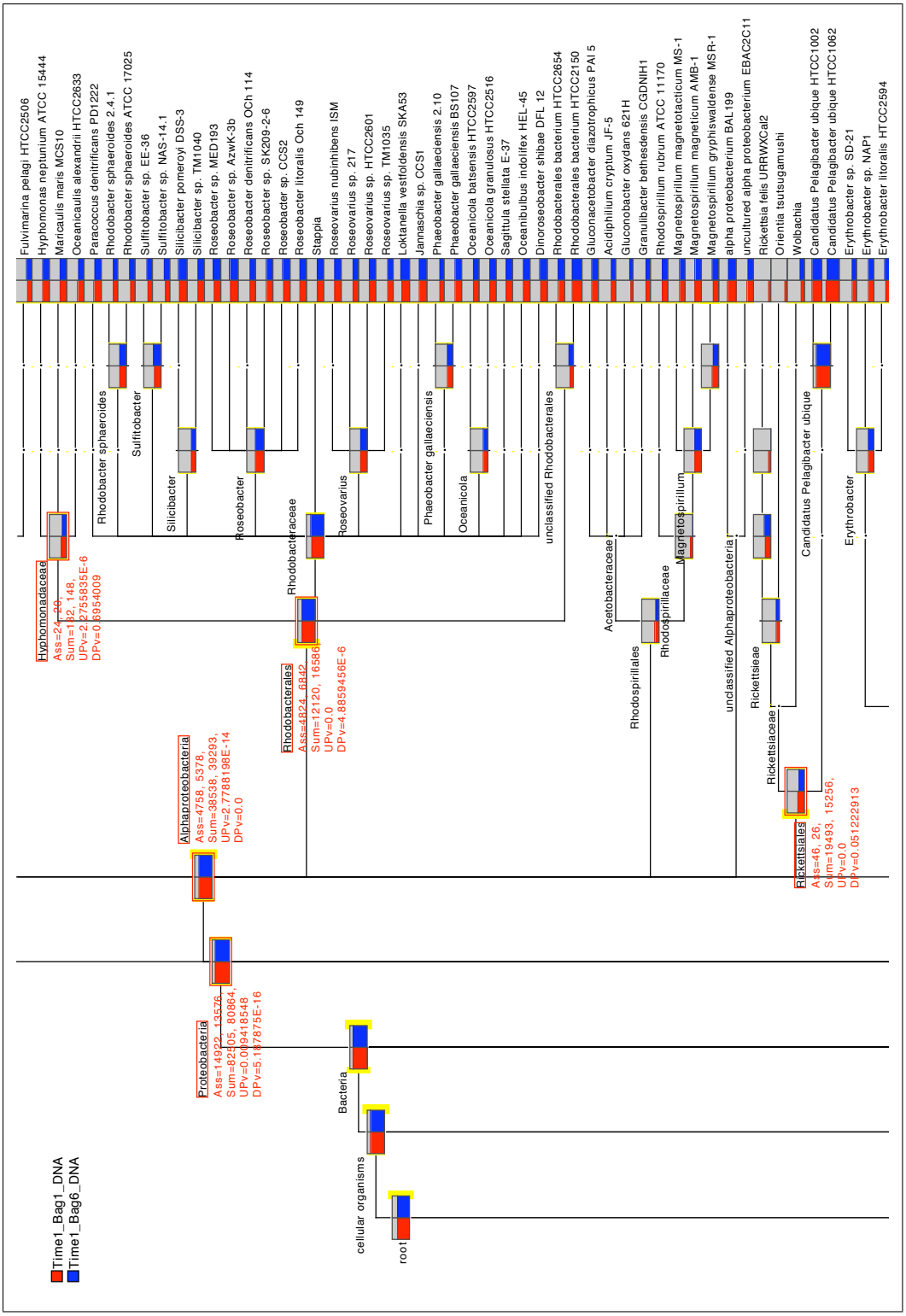

Figure 7.2.1: Pairwise comparison of two mesocosm samples at the peak of the bloom (Time 1 or 13^{th}). One sample in red (Bag 1 mesocosm) was enriched with CO_2 and the second in blue (Bag 6) was bubbled with air.

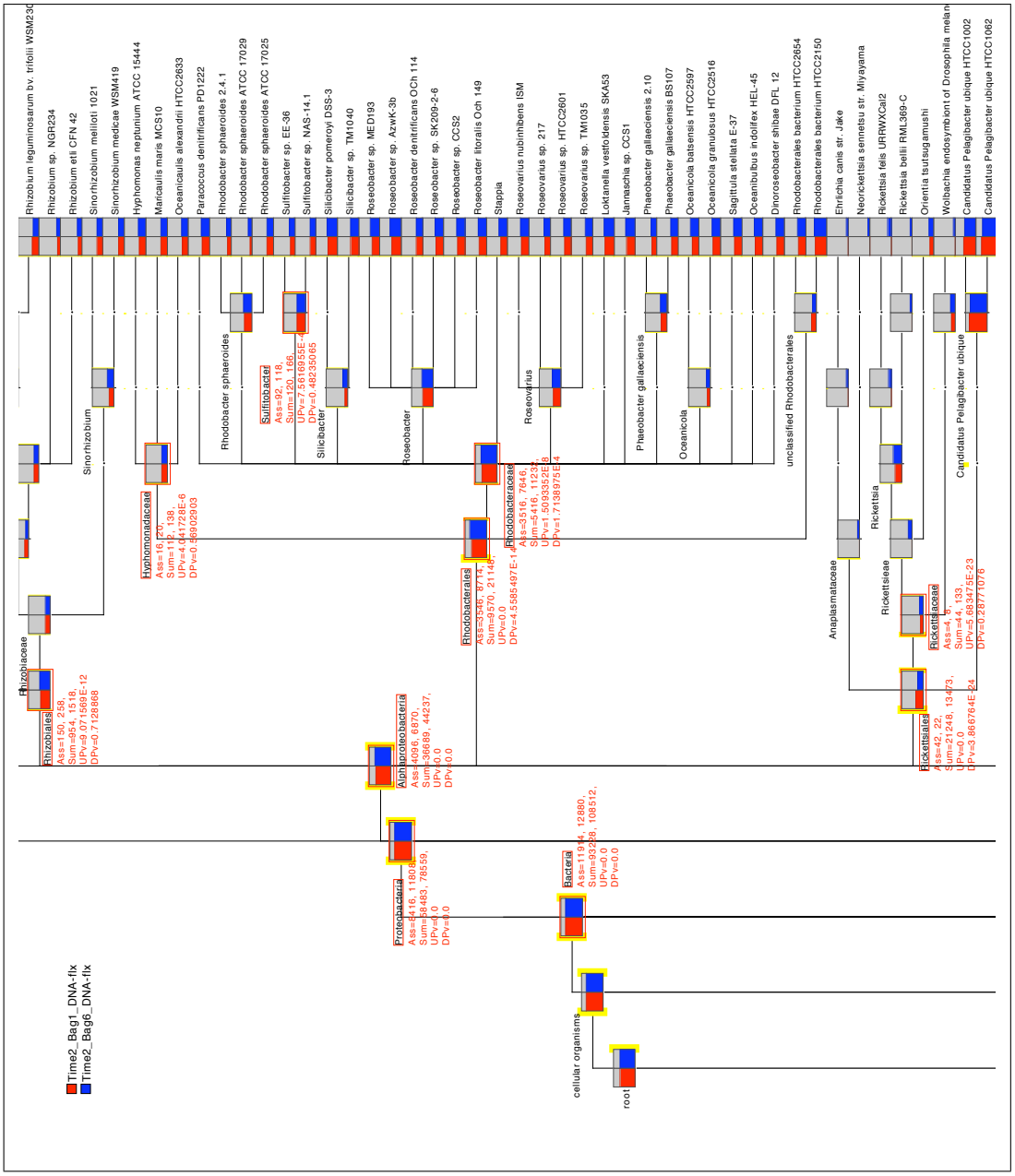

Figure 7.2.2: Pairwise comparison of two mesocosm samples, taken immediately after the collapse of the bloom (Time 2 or 19^{th}). One sample in red (Bag 1 mesocosm) was enriched with CO_2 and the second in blue (Bag 6) was bubbled with air.

indicates that the sub-taxa within this order are in part responsible of the observed difference. However, for Rickettsiales, the DPv is not below the critical level (0.01) and so the observed difference can be explained entirely based on this node. When we drill down into Rhodobacterales the observed difference is mostly

caused by the difference for Hyphomonadaceae ($UPv = 2.27E{-}6, DPv = 0.695$).

In Figure 7.2.2 we see the result of a pair-wise comparison between the Time2-Bag1-DNA (High CO2, 19^{th} May) and Time2-Bag6-DNA (Present Day, 19^{th} May) datasets. As with the 13^{th} May datasets there is the observed significant difference between the datasets at the *Alphaproteobacteria* node ($UPv = 0.0$). Unlike the 13^{th} May datasets, the major differences in the *Alphaproteobacteria* node can be explained by differences in the *Rhizobiales* ($UPv = 9.07E-12, DPv = 0.71$) *Rhodobacterales* ($UPv = 0.0, DPv = 4.55E-14$) and *Rickettsiales* ($UPv = 0.0, DPv = 3.87E-24$). For both *Rhodobacterales* and *Rickettsiales* the DPv again indicate that lower taxa also contribute. For the *Rhodobacterales* node the *Hyphomonadaceae* ($UPv = 4.04E-6, DPv = 0.56$) and *Rhodobacteraceae* ($UPv = 1.51E-8, DPv = 1.71E-4$) both contribute significantly. Yet, while the *Hyphomonadaceae* is again solely responsible for this difference, the differences in the *Rhodobacteraceae* can be explained by differences in *Sulfitobacter* ($UPv = 7.6E-4, DPv = 0.48$). For *Rickettsiales* the difference is due to *Rickettsiaceae* ($UPv = 5.6E-23, DPv = 0.29$) only.

Thus the 'Directed Homogeneity test' of MEGAN helped to visualize and identify significant different nodes between the acidified (Bag1) and the control (Bag6) samples at both time points (T1 and T2).

7.3 Seasonal and Diel Marine Bacterial Function

This study[3] is devoted to investigate seasonal and diel structured functional profiles of marine bacteia. Moreover we hypothesize that bacterial diversity will show little variability between day and night at any given time point during the annual cycle and the metatranscriptomic profile of the microbial community will vary between day and night as a direct influence of environmental factors.

7.3.1 Overview

It is expected that the functional diversity of ecosystems will be vast. This has been well characterized in various biogeographic studies (e.g. [Rusch and *et al.*, 2007, DeLong et al., 2006]) which have highlighted the huge number of microbial proteins present in the marine community, e.g. Rusch and colleagues [Rusch and *et al.*, 2007] demonstrated approximately 4.4 million

[3] *The content of this section will be submitted for publication as a part of the study done by Jack A. Gilbert, Dawn Field, Paul Swift, Suparna Mitra, Simon Thomas, Denise Cummings, Ben Temperton, Sue Huse, Margaret Hughes, Ian Joint, Paul Somerfield, Martin Mühling.*

unique genetic fragments from a study of 7.7 million sequences. Such cultivation-independent genomic surveys have proved a useful approach for characterizing the genetic potential of microbial communities [Handelsman, 2004, DeLong, 2005, DeLong et al., 2006]. However, there have been very few studies to determine how microbial function varies over time, especially using high-throughput metagenomic and metatranscriptomic studies.

Marine bacteria demonstrate seasonal patterns in diversity with on average a higher diversity during the winter than the summer [Murray et al., 1998, Fuhrman et al., 2006, Gilbert et al., 2009] in pelagic ecosystems. Numerous environmental factors have been suggested to influence this diversity (e.g. temperature and nutrients). This change in community structure over time has to date only been characterized using taxonomic profiling, and yet it stands to reason that if the community changes then so must the functional potential of that community.

To test several hypotheses mentioned above, multiple datasets were generated from pelagic water samples taken during the day and night at 3 annual time points, January, April and August, representing Winter, Spring and Summer. To determine whether bacterial and archaeal DNA taxonomy changes between seasons and diel time points $16S$ rDNA $V6$ profiling (e.g. [Gilbert et al., 2009]) was employed. Additionally, metagenomic shotgun DNA sequencing was also employed to investigate the impact of seasons and day and night on the functional potential of the community. Additionally, the impact of season and day/night was determined by sequencing the metatranscriptome of the community at each time point. Through provision of this dataset it is intended to show that changes in bacterial taxonomy relate to change in the functional potential of the community and additionally the relative impact of seasonal community change.

7.3.2 Methods for Metagenomic Analysis

All samples were collected from the surface water ($0 - 2\,\mathrm{m}$) of the L4 sampling station ($50.2518\,N, 4.2089\,W$) which is part of the Western Channel Observatory[4]. The sampling dates were January 28th, April 22^{nd}, August 26^{th} and August 27^{th}. During January a sample was taken at 15:00 at the $L4$ station at which point a minimal impact surface buoy with a $7m$ current drogue was deployed to track the surface currents for Lagrangian drift sampling. Approximately one hour post sun-down at 19:00 a second sample was taken at $50.2611N$: $4.2435W$. During April a sample was taken at 16:00 at the $L4$ station and following a Lagrangian drift a second sample was taken at 22:00 at $50.253N$: $4.1875W$. During August four samples were taken over a 24 hour period following a Lagrangian drift, with

[4] `http://www.westernchannelobservatory.org.uk`

the first sample at 16:00 on the 26^{th} at $L4$, the second sample at 22:00 on the 26^{th} at $50.2545N$: $4.199W$, a third at 04:00 on the 27^{th} at $50.2678N$: $4.1723W$, and a fourth at 10:00 on the 27^{th} at $50.2665N$: $4.1486W$. We will not describe the details of sample preparation, nucleic acid extraction and dataprocessing here in this section, but rather will emphasis the metagenome analysis using MEGAN [Huson et al., 2007] . All raw fasta files taken from pyrosequencing analysis were run through a metatranscriptomic data processing pipeline, to provide data for statistical analysis and annotation. As a part of this analysis, the MEGAN software was used to refine and cluster the taxonomic output from the BLASTX comparison.

7.3.3 Result and Discussion

Previous studies [Gilbert et al., 2009, Craft et al., 2010] demonstrate that bacterial communities in the Western English Channel exist within seasonally structured communities, exhibiting distinct Winter, Summer and Spring composition profiles. The aim of this study was to determine whether bacteria demonstrate seasonal-specific metagenomic and metatranscriptomic profiles, i.e. that the functional profile of the community was also seasonally structured between winter, spring and summer. A second axis of investigation was used to elucidate the response of the bacterial community to light availability by sampling this community in the same water mass between day and night. Bacteria and Archaea communities are seasonally structured but maintain stable community composition between day and night.

Metagenomic characterization of community composition and functionality between samples: Metagenomic taxonomic assignment of functional processes through comparison against the non-redundant NCBI protein database using the MEGAN platform (as described in 3.2) demonstrates that microbial communities show both seasonal and day/night induced changes in their composition. To perform the metagenomic analyses, random subsample of 50,000 metagenomic fragments was processed from each sample. This involved comparing each fragment against the NCBI-NR database using BLASTX and then identifying the closest taxonomic affiliation for each sample using the MEGAN protocol. In this way the taxonomic profile of the community could be visually compared between multiple samples for each taxon. MEGAN processing was performed for all samples together against the entire NR database. We then performed comparisons of each diel sample for January, April and pair wise for the four August time points, to observe the difference between the dataset in as much detail as possible. With the visual comparison technique of MEGAN, the relative abundance of each annotated taxon in this study was investigated. After that

'Directed Homogeneity Test (4.2.2) of MEGAN had been employed to investigate the significant diel responses of the taxonomic lineages of the most abundant bacterial phylum, the *Alphaproteobacteria*. To provide statistical significance for the differences in abundance of each node of the alphaproteobacterial tree in each sample pair-wise comparisons between each diel system have also been produced (Figures 7.3.1 – 7.3.5).

For January (Figure 7.3.1) the abundance of *Alphaproteobacteria* within *Proteobacteria* is significant different between day and night ($UPv = 3.4e - 7$) and this can be explained by differences in the lower nodes ($DPv = 0.0$; p-values $\leq 1E - 37 = 0.0$). *Rhizobiales* is more abundant during the day and this is mainly due to differences in homologues of the *Bradyrhizobiaceae* species *Rhodopseudomonas palustris*. The *Rhodobacteraceae* are also significantly different in their abundance between day and night, generally due to a greater abundance of *Roseobacter* clade species during the day. The *Rhodospirillaceae* are also more significantly abundant in the day, as are the *Rickettsiales*, the latter is mainly due to an increased abundance in functional homologues of *Pelagibacter ubique* genes (Figure 7.3.1).

During April (Figure 7.3.2) the exact same trend exists, virtually all Alphaproteobacterial taxa show a statistically significant increased abundance in the day. Of note is a significant increase in *Roseovarius* and *Sphingomonas* abundance during the day.

During August the difference between 4 pm and 10 pm on the 27^{th} is less pronounced with mostly higher-level taxa showing significant difference, one exception is *Pelagibacter* (Figure 7.3.3); while between 10 pm and 4 am we see only small changes in higher level taxa, e.g. *Rhodobacterales, Rickettsiales and Sphingomonadales* (Figure 7.3.4). The differences between 4 am and 10 am (Figure 7.3.5) virtually mirror those for 4 pm-10 pm. Overall this demonstrates that there are more statistically significant functional taxonomic differences in the distribution for the *Alphaproteobacteria* during the January day/night cycle than for April, while August shows very little change apart from a small but significant increase in *Pelagibacter* abundance during the day.

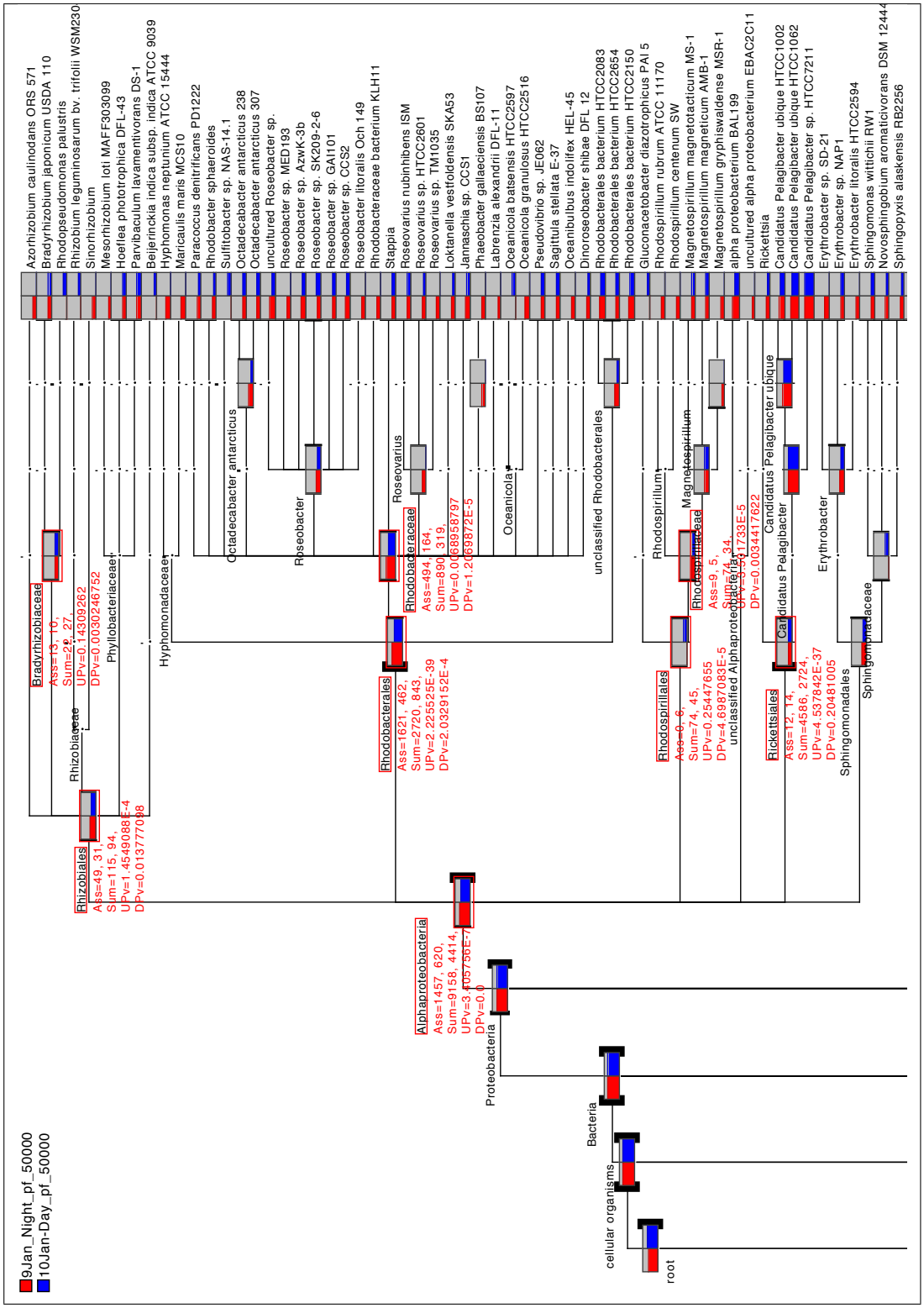

Figure 7.3.1: A part of a comparison of the alphaproteobacterial tree between 'day' and 'night' -water samples in the month of January.

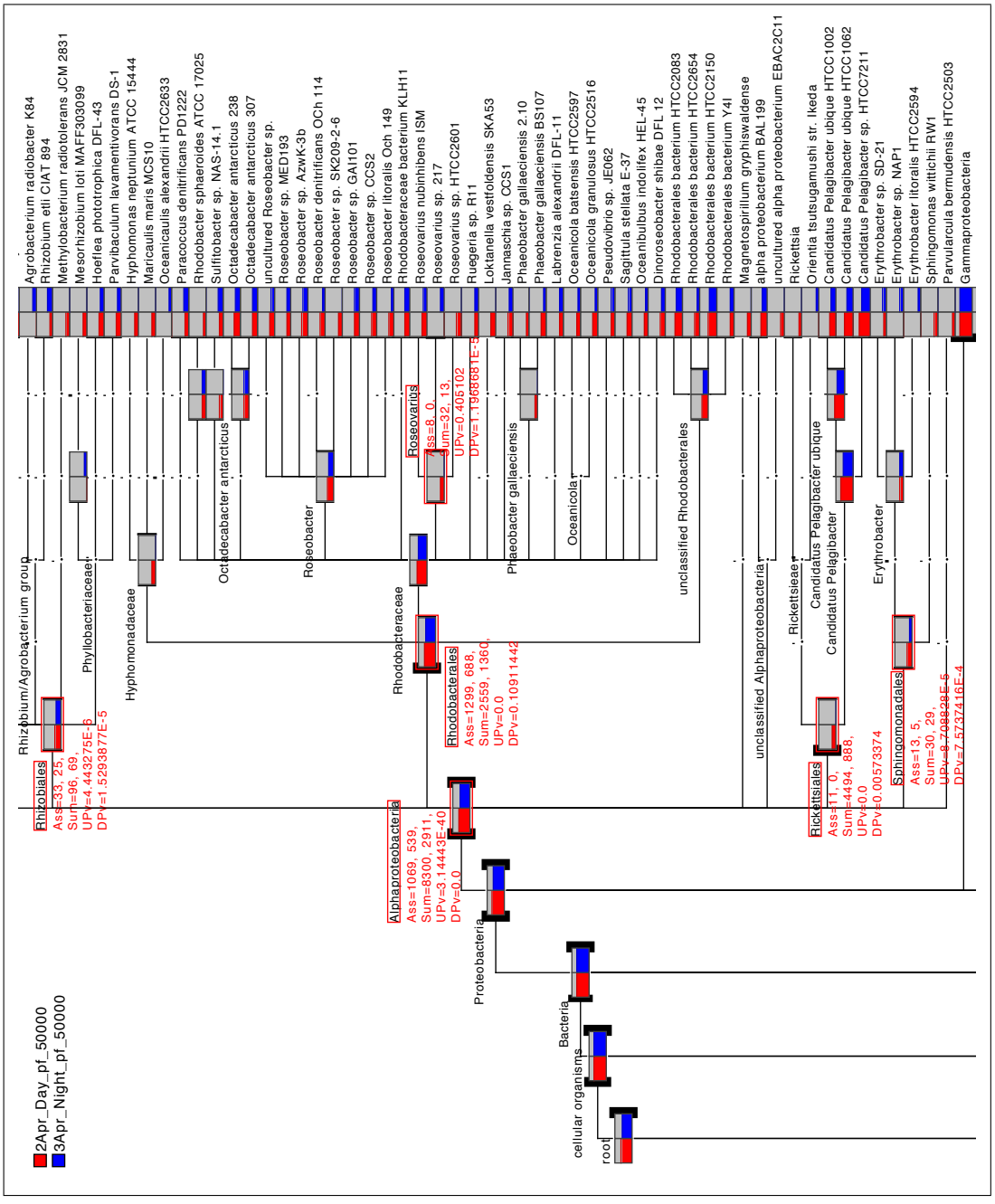

Figure 7.3.2: A part of a comparison of the alphaproteobacterial tree between 'day' and 'night' water samples in the month of April.

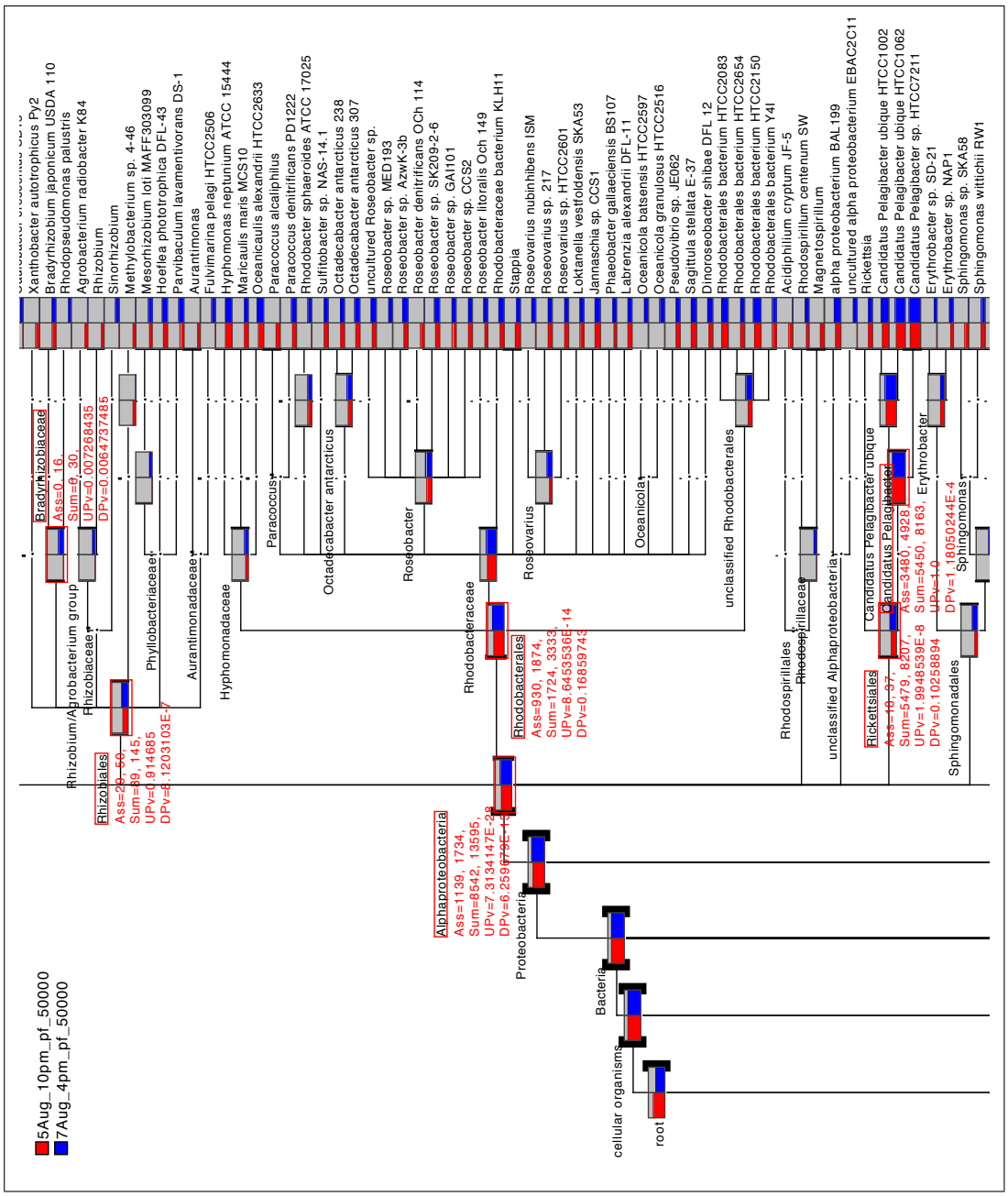

Figure 7.3.3: A part of a comparison of the alphaproteobacterial tree between 4 pm and 10 pm water samples in the month of August.

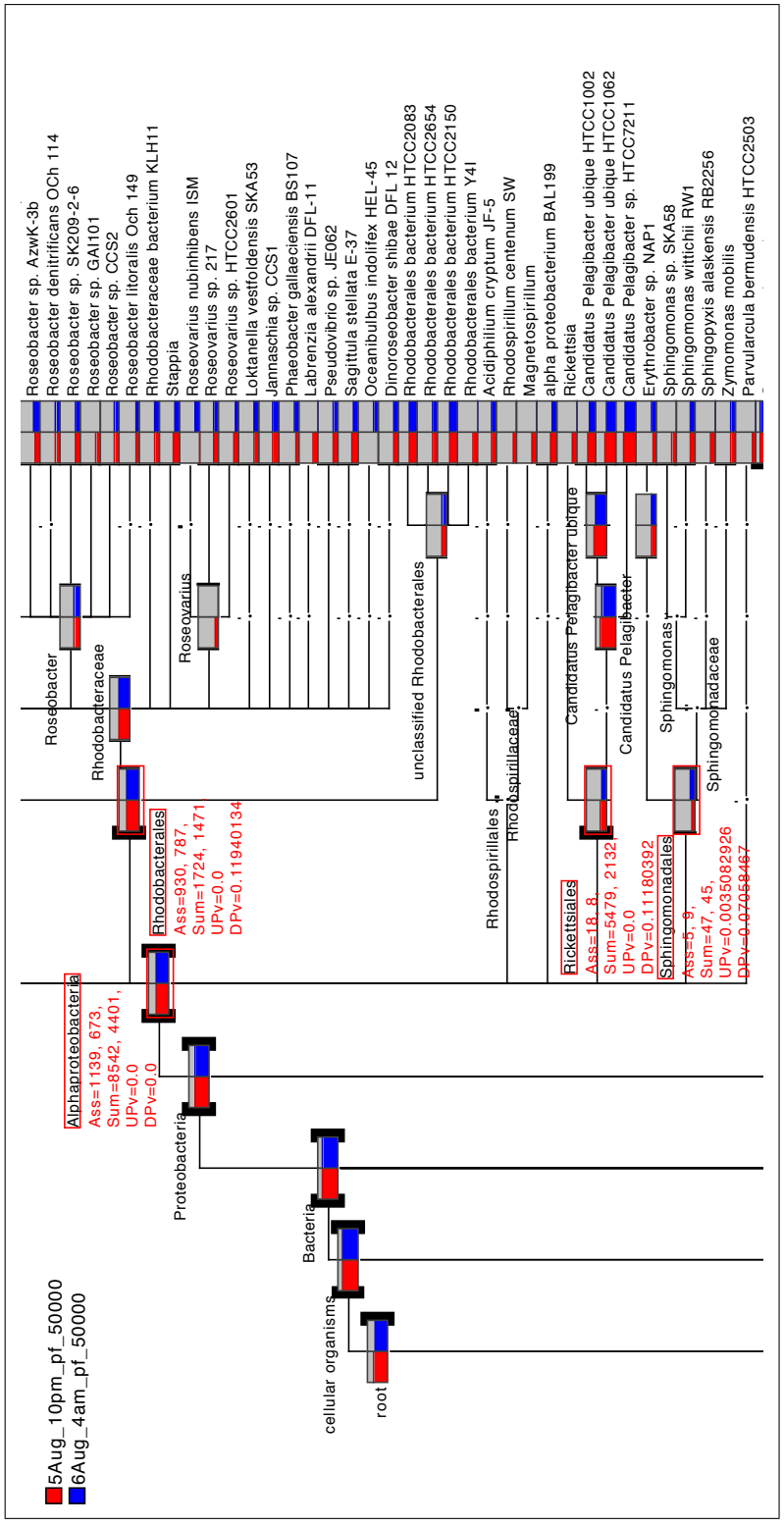

Figure 7.3.4: A part of a comparison of the alphaproteobacterial tree between 4 pm and 10 am water samples in the month of August.

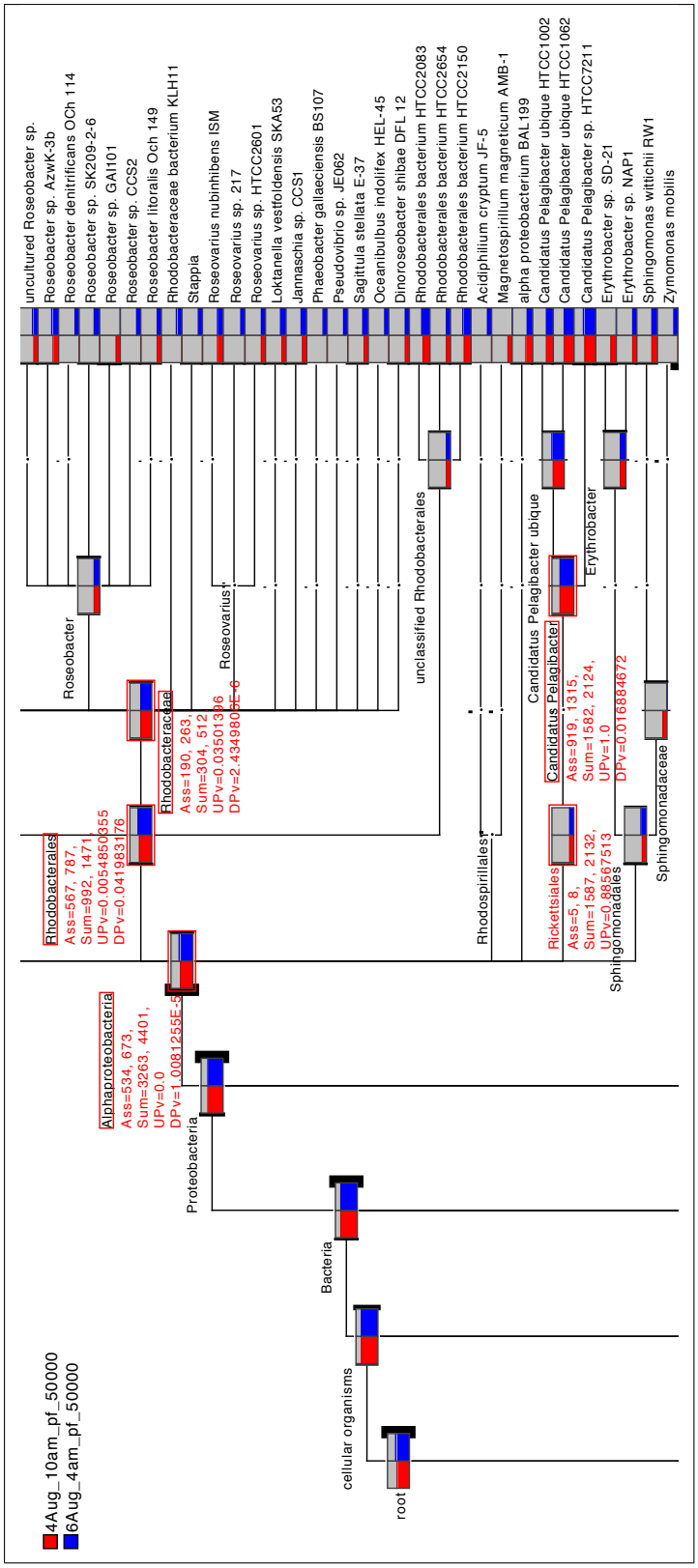

Figure 7.3.5: A part of a comparison of the alphaproteobacterial tree between 4 am and 10 am water samples in the month of August.

Network analyses of multiple samples: Furthermore, we have performed network analyses of all samples simultaneously, including January (day and night), April (day and night) and of four time points of August samples (Figure 7.3.6). Here we can easily observe for January and April that the samples from day and night are very different (not even clustering together). However, in August the difference between the samples taken at 4 pm and 10 pm, are less prominent and they are clustering together. Similar nature can be observed between the samples taken at 4 am and 10 am. These results also agree with the observed facts from pair-wise comparison of the samples.

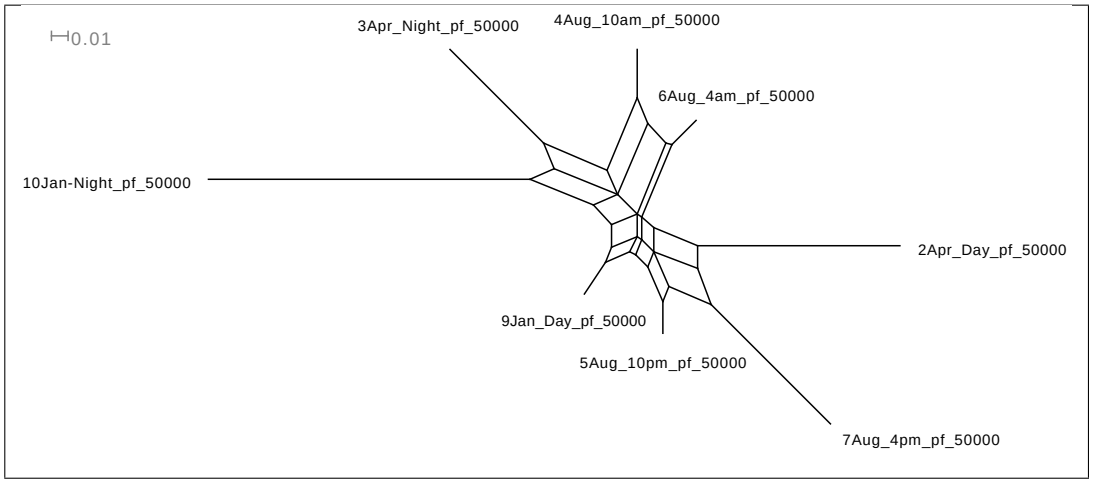

Figure 7.3.6: Network comparison of genomic DNA samples taken at six time points including January (day and night), April (day and night) and from four time points of August.

Thus 'Directed Homogeneity test' and 'multiple comparison with networks' helped to analyze the diversity of the microbial population in the samples between day and night at different time points during the annual cycle.

7.4 Analysis of the Mammoth Microbiome

This study[5] describes the diversity and community structure of microbes isolated from unique permafrost samples, 13 hair and 1 bone specimen from woolly mammoths. Comparisons between the mammoth biomes and modern microbial communities would shed light on not only the ancient microorganisms associated with woolly mammoths but also on those inhabit the permafrost.

7.4.1 Overview

Numerous microbial communities have been identified in the permafrost or glacial ice environments. While several studies have focused on permafrost microbial communities, few have analyzed this habitat in detail using metagenomics.

The woolly mammoth (*Mammuthus primigenius*) died out several thousand years ago. Since these animals lived and died in a subzero climate, some woolly mammoth remains are well preserved. The mammoth biome represents a unique community for research on microbial evolution, low temperature adaptation, and helps elucidate a core set of ancient microbes that coexisted with woolly mammoths. For this study samples are taken from diverse locations around the Siberian permafrost.

7.4.2 General features of metagenomic data

In the mammoth genome sequencing project, ~ 1 Gb metagenomic data derived from environmental organisms (eg. bacteria, fungi, virus, etc) were generated. Most of these data were derived from hair shaft samples (M-series; eight samples: M6, M7, M11-M14, M16, M20) and a small proportion (~ 9 %) was derived from mammoth bone (Poinar).

7.4.3 Multiple Comparison of Mammoth Samples

The average read length for the M-series samples (172 bp) is much longer than that of Poinar sample (102 bp), largely because of updating from a GS 20 to GS-FLX sequencer (454 Life Sciences, Branford, CT).The mammoth library was prepared according to the Illumina paired-end library preparation protocol were

[5] *The content of this section is submitted for publication as a part of the study done by Fangqing Zhao, Ji Qi, Daniel C. Richter, Anne Buboltz, Daniela Drautz, Suparna Mitra, Daniel H. Huson, Stephan C. Schuster.*

processed using the Illumina pipeline software. For phylogenetic assignment of metagenomic sequences and network analysis long 454 reads ($\geq$ 120 bp) were compared against the nonredundant NCBI protein database (12/6/2008) using BLASTX [Altschul et al., 1990]. Then the MEGAN software [Huson et al., 2007] was used to assign reads to taxa of the NCBI taxonomy using the parameters: MinScore = 35, TopPercent = 10, and MinSupport = 5. After collapsing all the reads to 'Genus' level of NCBI taxonomy, we selected only the bacterial leaf nodes and compared mammoth biome samples using the 'Compare Datasets Using Network' option of MEGAN (See Section 5.2 for details of this method). This comparison results an unrooted phylogenetic network (using the 'Neighbor-Net' method [Bryant and Moulton, 2004]) for the mammoth samples.

Reads putatively mapping to rare taxa were removed from all nine datasets. Therefore, the differences coming from 'abundant taxa' is more robust to measure the distance among various mammoth biomes. In this study we considered three distance measures ('Bray-Curtis', 'Chi-square' and 'Hellinger') to observe the similarity values between all possible pair of samples (pair-combinations) with respect to each species.

7.4.4 Result and Discussion

We firstly used the similarity and dissimilarity of phylogenetic profiles to assess genetic distance among these biomes. As shown in Figure 7.4.1, the three different measurements give rise to quite similar topologies of the phylogenetic network. As expected, the Poinar datasets is separated from the other hair metagenomes (M series) by a nearly two-times longer distance. It is notable that the genetic distances between these metagenomes do not necessarily reflect their geographic distances. For example, M2 and M3 are geographically closest among the listed samples, whereas their genetic distance is rather longer than the distance between M2 and M18.

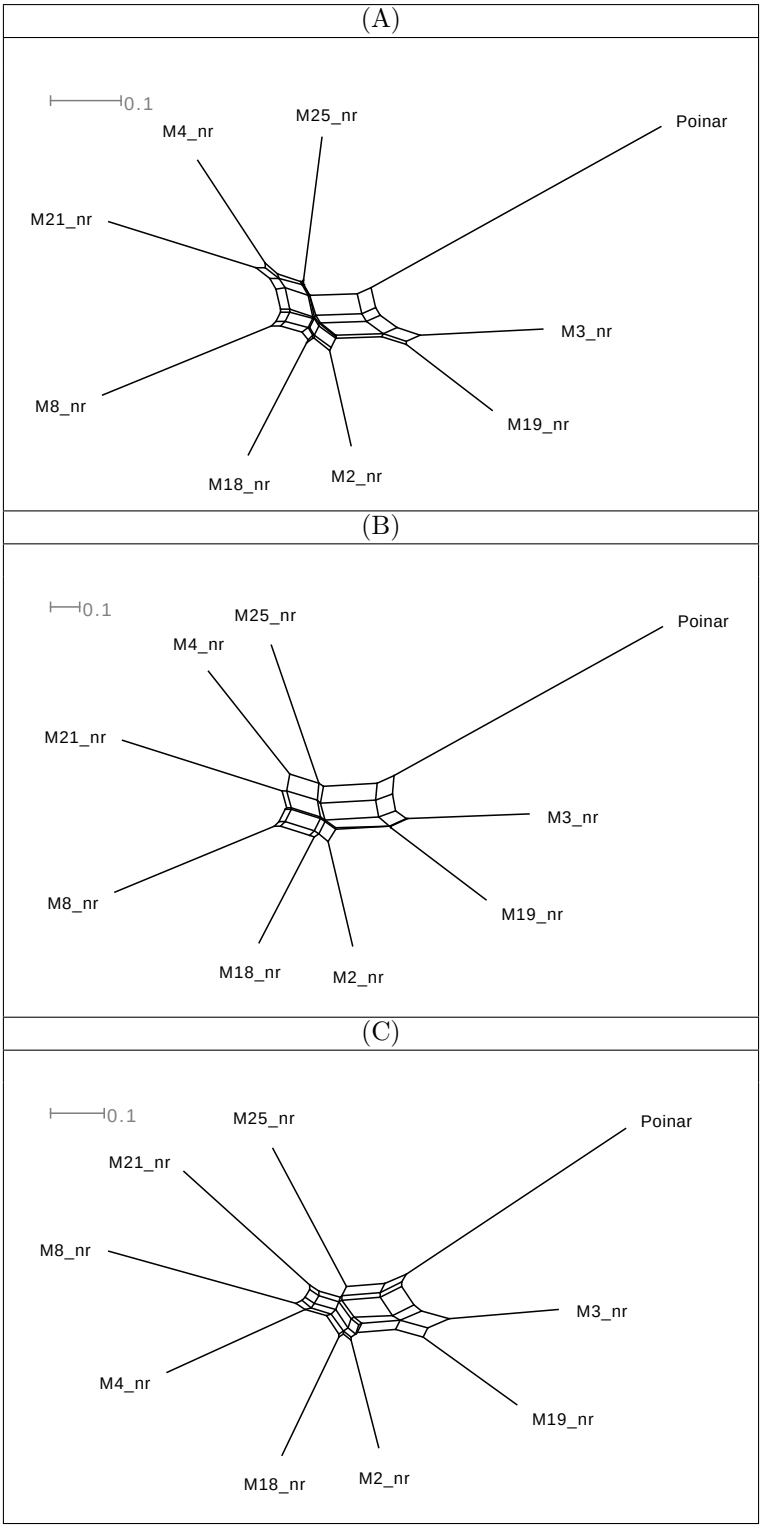

Figure 7.4.1: Comparison of eight hair samples ($M2, M3, M4, M8, M18, M19, M21$ and $M25$) and one bone (*Poinar*) sample from 'Wooly mammoth' at 'Genus' level of NCBI taxonomy. Networks obtained using (A) 'Bray-Curtis distance', (B) 'Chi-square distance' and (C) 'Hellinger distance' measures.

Chapter 8

Summary and Outlook

The work described in this thesis contributes to the emerging field of metagenomic research. Several approaches have recently been developed to deal with the taxonomic and functional analyses of metagenomic sample. However there is a lack of easy, powerful and readily available tool for such analysis.

Researchers are studying our microbial world from different angles, uding metagenomics, metatranscriptomics, metaproteomics, metabolomics. New sequencing technologies have made DNA sequencing feasible at an affordable cost, and this has boosted the number and size of metagenome projects. An overwhelming quantity of DNA sequences are being deposited in the databases. Third-generation sequencing technologies are on the horizon with their great power, which will initiate an abrupt change in these research fields. Fast and user-friendly tools are necessary to analyze multiple metagenomic datasets. MEGAN attempts to fill this gap.

The initial goal of metagenomic studies is to obtain a vision of the microbial community, both surrounding us and within us. In a quest for better understanding the silent rulers of different communities, a main challenge is to compare multiple datasets. While pursuing taxonomic and functional analyses of metagenome samples, the exploratory work of this thesis is devoted to metagenome comparison. Initial work has been performed to compare the contents of metagenome samples considering statistical aspects and confidence. The method allows close comparison of two metagenome datasets, at each node within a tree hierarchy.

To compare multiple metagenomes simultaneously, a novel approach is presented that combines the use of taxonomic or functional analysis with ecological indices and non-hierarchical clustering techniques. This method provides a net-

work representation of the relationships between different metagenome datasets. In a network it is easy to identify datasets with similar content as they cluster together. Besides metagenome samples, these approaches are also applicable on metatranscriptome or 16S rRNA profiles.

Several collaborations which flourished during this research period made it possible to apply all of these methods to real biological data. The application results presented at the end of this thesis help serve to demonstrate the impact of the methods in respective studies. All the methods and ideas described in this thesis are implemented in MEGAN and are easily available for further use.

The method presented in this thesis for multiple comparison with simultaneous visualisation, is the first attempt in this field towards answering such question. As an outlook of this research, sophisticated statistical methods can be applied to understand the significance of each edge in a multiple comparison network more closely. It would be desirable to understand which biological features are responsible for the distances. This method can be further motivated for finding different disease causing genes, when compared to healthy reference samples, considering functional pathway together with taxonomic profiles. With the advances of new cost effective and high throughput sequencing technologies, many research projects will be performed for better understanding the biological diversity of different communities. The bias towards known organisms will be improved with more knowledge. Comparative metagenomics will play an important role in better understanding the community structure and thus unveiling our microbial planet.

Appendix A

Publications

A.1 Published Manuscripts

1. Daniel H. Huson, Daniel C. Richter, Suparna Mitra, Alexander F. Auch and Stephan C. Schuster. **Methods for comparative metagenomics.** *BMC Bioinformatics* 2009, 10 (Suppl 1):S12.

Background: Metagenomics is a rapidly growing field of research that aims at studying uncultured organisms to understand the true diversity of microbes, their functions, cooperation and evolution, in environments such as soil, water, ancient remains of animals, or the digestive system of animals and humans. The recent development of ultra-high throughput sequencing technologies, which do not require cloning or PCR amplification, and can produce huge numbers of DNA reads at an affordable cost, has boosted the number and scope of metagenomic sequencing projects. Increasingly, there is a need for new ways of comparing multiple metagenomics datasets, and for fast and user-friendly implementations of such approaches.
Results: This paper introduces a number of new methods for interactively exploring, analyzing and comparing multiple metagenomic datasets, which will be made freely available in a new, comparative version 2.0 of the stand-alone metagenome analysis tool MEGAN.
Conclusion: There is a great need for powerful and userfriendly tools for comparative analysis of metagenomic data and MEGAN 2.0 will help to fill this gap.

2. Suparna Mitra, Bernhard Klar and Daniel H. Huson. **Visual and statistical comparison of metagenomes.** *Bioinformatics* 2009, 25 (15):1849-55.

   Background: Metagenomics is the study of the genomic content of an environmental sample of microbes. Advances in the through-put and cost-efficiency of sequencing technology is fueling a rapid increase in the number and size of metagenomic datasets being generated. Bioinformatics is faced with the problem of how to handle and analyze these datasets in an efficient and useful way. One goal of these metagenomic studies is to get a basic understanding of the microbial world both surrounding us and within us. One major challenge is how to compare multiple datasets. Furthermore, there is a need for bioinformatics tools that can process many large datasets and are easy to use.
   Results: This article describes two new and helpful techniques for comparing multiple metagenomic datasets. The first is a visualization technique for multiple datasets and the second is a new statistical method for highlighting the differences in a pairwise comparison. We have developed implementations of both methods that are suitable for very large datasets and provide these in Version 3 of our standalone metagenome analysis tool MEGAN.
   Conclusion: These new methods are suitable for the visual comparison of many large metagenomes and the statistical comparison of two metagenomes at a time. Nevertheless, more work needs to be done to support the comparative analysis of multiple metagenome datasets.

3. Suparna Mitra, Max Schubach and Daniel H. Huson. **Short clones or long clones? A simulation study on the use of paired reads in metagenomics.** *BMC Bioinformatics* 2010, 11(Suppl 1):S12.

   Background: Metagenomics is the study of environmental samples using sequencing. Rapid advances in sequencing technology are fueling a vast increase in the number and scope of metagenomics projects. Most metagenome sequencing projects so far have been based on Sanger or Roche-454 sequencing, as only these technologies provide long enough reads, while Illumina sequencing has not been considered suitable for metagenomic studies due to a short read length of only $35bp$. However, now that reads of length $75bp$ can be sequenced in pairs, Illumina sequencing has become a viable option for metagenome studies.
   Results: This paper addresses the problem of taxonomical analysis of paired reads. We describe a new feature of our metagenome analysis software MEGAN that allows one to process sequencing reads in pairs

and makes assignments of such reads based on the combined bit scores of their matches to reference sequences. Using this new software in a simulation study, we investigate the use of Illumina paired-sequencing in taxonomical analysis and compare the performance of single reads, short clones and long clones. In addition, we also compare against simulated Roche-454 sequencing runs.
Conclusion: This work shows that paired reads perform better than single reads, as expected, but also, perhaps slightly less obviously, that long clones allow more specific assignments than short ones. A new version of the program MEGAN that explicitly takes paired reads into account is available from our website.

4. Suparna Mitra, Jack A. Gilbert, Dawn Field and Daniel H. Huson. **Comparison of multiple metagenomes using phylogenetic networks based on ecological indices.** *ISME J* 2010, 4:1236-1242.

   Second-generation sequencing technologies are fuelling a vast increase in the number and scope of metagenome projects. There is a great need for the development of new methods for visualizing the relationships between multiple metagenomic datasets. To address this, a novel approach is presented that combines the use of taxonomic analysis, ecological indices and non-hierarchical clustering to provide a network representation of the relationships between different metagenome datasets. The approach is illustrated using several published data sets of different types, including metagenomes, metatranscriptomes and 16S ribosomal profiles. Application of the approach to the same data summarized at different taxonomical levels gives rise to remarkably similar networks, indicating that the analysis is very robust. Importantly, the networks provide the both visual definition and metric quantification for the non-rooted relationship between samples, combining the desirable characteristics of other tools into one.

5. Suparna Mitra, Paul Rupek, Daniel C. Richter, Tim Urich, Jack A. Gilbert, Folker Meyer, Andreas Wilke, Daniel H. Huson. **Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG**. *BMC Bioinformatics* 2011, 12 (Suppl 1):S12.

   Background: Metagenomics is the study of microbial organisms using sequencing applied directly to environmental samples. Technological advances in next-generation sequencing methods are fueling a rapid increase in the number and scope of metagenome projects. While metagenomics provides information on the gene content, metatranscriptomics aims at understanding gene expression patterns in micro-

bial communities. The initial computational analysis of a metagenome or metatranscriptome addresses three questions: (1) Who is out there? (2) What are they doing? and (3) How do different datasets compare? There is a need for new computational tools to answer these questions. In 2007, the program MEGAN (MEtaGenome ANalyzer) was released, as a standalone interactive tool for analyzing the taxonomic content of a single metagenome dataset. The program has subsequently been extended to support comparative analyses of multiple datasets.
Results: The focus of this paper is to report on new features of MEGAN that allow the functional analysis of multiple metagenomes (and metatranscriptomes) based on the SEED hierarchy and KEGG pathways. We have compared our results with the MG-RAST service for different datasets.
Conclusions: The MEGAN program now allows the interactive analysis and comparison of the taxonomical and functional content of multiple datasets. As a stand-alone tool, MEGAN provides an alternative to web portals for scientists that have concerns about uploading their unpublished data to a website.

A.2 Published Book Chapter

1. Daniel H. Huson and Suparna Mitra. **Comparative metagenome analysis using MEGAN**, *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, De Bruijn FJ, ed., Wiley Blackwell Publishers, ppXXXX, 2010.

In metagenomics, random shotgun sequencing is used to study a community of microbes. The first three computational questions are: What is the taxonomical content of a sample? What is the functional content of a sample? How do different samples compare? All three questions can be addressed using the program MEGAN. The result of comparing a metagenome dataset against a database of reference sequences obtained, for example, by using BLASTX against NCBI-NR, is parsed by the program and then both a taxonomical and functional analysis are performed. Multiple datasets can be opened simultaneously and compared. This chapter demonstrates how to perform such analyses using a number of published mouse gut datasets.

A.3 Submitted Manuscripts

1. Simon Domke, Suparna Mitra, Nico Weber, Stephan C. Schuster, Thomas Rattei and Daniel H. Huson. **MEGAN-DB – The MEtaGenome ANalyzer DataBase**. (Submitted to *Nucleic Acids Research- Database issue*), 2010.

   The sequencing of genomic or transcribed DNA from environmental samples ("metagenomics") allows investigating the structure, function and metabolism of environmental communities on the molecular level. Comparative analyses between multiple metagenomes is becoming increasingly informative due to the large, rapidly growing number of available metagenomic sequences.

   MEGAN-DB provides a comprehensive repository of metagenomes and their taxonomic and functional profiles. Exploration of the database contents as well as comparative analyses are facilitated by a user-friendly web portal, providing versatile tools and an integrated interface to the MEGAN software. MEGAN-DB offers direct data access to the meta-data through web services and to MEGAN input files via FTP. Users may upload new metagenomes for permanent integration into MEGAN-DB or for single conversion into MEGAN files.

   Frequent updating of MEGAN-DB to integrate new metagenomes as well as new reference sequences will facilitate comparative taxonomic and functional interpretation of metagenomes considering the whole publicly available metagenomic sequence space. MEGAN-DB is freely available to academic and commercial users via `http://megan-db.org`.

2. Fangqing Zhao, Ji Qi, Daniel C. Richter, Anne Budoltz, Daniela Drautz, Suparna Mitra, Daniel H. Huson and Stephan C. Schuster. **Metagenomic analysis of the microbiome associated with the hairs of extinct woolly mammoths**. (Submitted to *PLoS ONE*), 2010.

   Background: Metagenomics based on random sequencing of microbial community DNA offers the opportunity to understand the phylotypic diversity and the functional potential present in microbial communities. Till now, extensive studies have examined a wide range of microbial habitats, including deep sea, soil, biofilm, human skin and human gut. However, few studies have been reported on microbial communities associated with paleo specimen isolated from per-

mafrost. The most significant difference between paleo-metagenomic analysis and others is that paleo samples likely include a collection of continuously aged microorganisms, from ancient bacteria to modern bacteria, as the extremely cold temperature in permafrost provides for the preservation of ancient DNA.
Methodology/Principal Findings: We studied the diversity and community structure of microbes isolated from unique permafrost samples, 13 hair and 1 bone specimen from woolly mammoths. After analyzing environmental sequences from these specimens, we determined that the bone sample contained more soil bacteria, suggesting that the taxonomic composition of bacterial assemblages varies greatly between these samples. A number of putative ancient bacteria were identified, as they had an elevated DNA damage rate compared to modern strains, such as strains from Streptomyces, Caulobacter and Propionibacterium. A small proportion of fungal sequences were found in the mammoth biomes, with six of the ten most abundant species being plant pathogens. Psychrophilic Flavobacterium spp., Psychroflexus spp. and Psychrobacter spp. dominate the cold adapted bacteria.
Conclusions/Significance: Comparisons between the mammoth biomes and modern microbial communities shed light on both the ancient microorganisms associated with woolly mammoths and those that inhabit the permafrost.

3. Reema Singh, Vinay Narayan, Patricia McLenachan, Richard C. Winkworth, Suparna Mitra, Peter J. Lockhart, Lorraine Berry, Abdulla M. Hatha, William Aalbersberg and Dhana Rao. **Detection and diversity of pathogenic *Vibrio* from Fiji**. (Submitted to *Environmental Microbiology*), 2010.

   The present study investigates the diversity of pathogenic *Vibrio* species on fish available for consumption and in the coastal water column close to Suva, Fiji. We used biochemical tests to screen fish sold at retail outlets in Suva; these analyses were consistent with the presence of both clinical and nonclinical *Vibrio* species on fish. Phylogenetic analyses of three markers (i.e. $16S$, *recA* and *pyrH*) confirmed the identity of several isolates as *V. parahaemolyticus* and suggest this clinically important species is represented by several genotypes. We used MEGAN analyses of Illumina GAII single and paired end sequencing to investigate the potential of short read DNA sequencing for identifying *Vibrio* species in the coastal water column. While both approaches identified several species, paired end sequencing resulted in substantially more taxonomic assignments. These analyses suggest

that a moderately complex *Vibrio* community containing both clinically important and non-clinical species occurs in the coastal water column in this area. Our results are very encouraging, it appears that with even short reads Illumina GAII paired end sequencing may offer an effective and efficient method for monitoring microorganisms in the environment.

Appendix B

Contribution

This thesis illustrates several algorithms and contains implementations and applications of those algorithms in different current research areas. This is the outcome of my research while pursuing my PhD. At this point, I would like to distinguish the contributions of my colleagues or collaborators from my own work.

Chapter 3: *M*etagenome Analysis using MEGAN &

Chapter 4: *V*isual and Statistical Comparison of Metagenomes:

> D. Huson developed and implemented the described taxonomic and functional assignment and visual comparison techniques. I developed the extension of the statistical technique of [Rodriguez-Brito et al., 2006] for comparing two metagenomes. D. Huson wrote the manuscript [Huson et al., 2009], where I contributed in writing and performing the statistical comparison.
>
> Later with the help of B. Klar, I found the drawbacks of the Rodriguez-Brito's method and developed a sophisticated approach, the 'Directed Homogeneity Test' (which includes the up- and down-tests) and implemented the methods. D. Huson integrated the statistical methods into MEGAN. I wrote the manuscript [Mitra et al., 2009], selected the journal and interacted with the editor and reviewers, whereas B. Klar helped me in statistical aspects and D. Huson contributed many useful comments.

Chapter 5: *Multiple Metagenome Comparison using Networks:*

> I searched a huge amount of literatures in traditional statistics, ecology and phylogenetics and developed the technique of multiple metagenome comparison. I implemented the methods with the help of W. Wu and D. Huson

integrated this into MEGAN. I rote the manuscript [Mitra et al., 2010a] and interacted with the editor and reviewers, whereas J. Gilbert, D. Field and D. Huson contributed many useful comments.

Chapter 6: ***Comparison of Sequencing Technologies for Metagenomics:***

I and D. Huson designed the project and wrote the manuscript [Mitra et al., 2010b]. M. Schubach performed the simulation study and did the analysis with me. I supervised him in each step of this process and did several BLAST runs for the study. M. Schubach and D. Huson wrote necessary scripts for combining paired reads and D. Huson integrated this into MEGAN.

Chapter 7: ***Application in Metagenomic Projects:***

I contacted many researchers working in the field of metagenomics, communicated with them regarding their samples and performed all the necessary steps (such as data analysis planning, BLAST runs and analyses of results) related to metagenomic analyses of various sample. My analyses (as described in Chapter 7) gave extra importance to their study. These results will be part of different separate publications.

Appendix C

Supplementary Material

C.1 Chapter 4: Visual and Statistical Comparison of Metagenomes

This section includes the supporting tables for Chapter 4.

Data	Gammaproteobacteria	Bacteroidetes/ Chlorobi Group	Firmicutes
Soil vs Soil	UPv, DPv	UPv, DPv	UPv, DPv
20-1-2cmp.megan	0.033, 0.942	0.716, 0.638	0.233, 0.458
20-1-3cmp.megan	0.689, 0.436	0.459, 0.521	0.020, 0.458
20-1-4cmp.megan	0.136, 0.708	0.431, 0.131	0.033, 0.107
20-1-5cmp.megan	0.044, 0.697	0.448, 0.404	0.443, 0.070
20-1-6cmp.megan	0.263, 0.830	0.843, 0.117	0.385, 0.704
20-1-7cmp.megan	0.297, 0.954	0.684, 0.993	0.400, 0.563
20-1-8cmp.megan	0.333, 0.597	0.705, 0.521	0.888, 0.513
20-1-9cmp.megan	0.035, 0.644	0.431, 0.869	0.910, 0.527
20-1-10cmp.megan	0.123, 0.708	0.292, 0.453	0.975, 0.412
20-1-11cmp.megan	0.434, 0.948	0.490, 0.344	0.315, 0.298
20-1-12cmp.megan	0.843, 0.313	0.655, 0.680	0.808, 0.237
20-1-13cmp.megan	0.697, 0.283	0.426, 0.524	0.865, 0.371
20-1-14cmp.megan	0.372, 0.252	0.303, 0.503	0.129, 0.191
20-1-15cmp.megan	0.285, 0.491	0.458, 0.523	0.231, 0.032
20-1-16cmp.megan	0.238, 0.606	0.912, 0.194	0.790, 0.698
20-1-17cmp.megan	0.260, 0.459	0.542, 0.106	0.654, 0.507
20-1-18cmp.megan	0.214, 0.850	0.783, 0.836	0.871, 0.557
20-1-19cmp.megan	0.251, 0.590	0.765, 0.867	0.126, 0.896
20-1-20cmp.megan	0.040, 0.978	0.763, 0.241	0.696, 0.161

Table C.1.1: Comparison within soil data subsamples. First soil subsample is compared with 20 other subsamples.

Data	Gammaproteobacteria	Bacteroidetes/ Chlorobi Group	Firmicutes
Sea vs Sea	UPv, DPv	UPv, DPv	UPv, DPv
20-1-2cmp.megan	0.620, 0.530	0.340, 0.960	0.880, 0.570
20-1-3cmp.megan	0.386, 0.840	0.371, 0.988	0.160, 0.910
20-1-4cmp.megan	0.580, 0.970	0.167, 0.830	0.990, 0.380
20-1-5cmp.megan	0.530, 0.830	0.244, 0.910	0.956, 0.169
20-1-6cmp.megan	0.134, 0.867	0.090, 0.966	0.568, 0.450
20-1-7cmp.megan	0.155, 0.745	0.612, 0.864	0.851, 0.111
20-1-8cmp.megan	0.056, 0.786	0.695, 0.817	0.818, 0.478
20-1-9cmp.megan	0.444, 0.530	0.054, 0.998	0.924, 0.279
20-1-10cmp.megan	0.468, 0.706	0.568, 0.999	0.916, 0.266
20-1-11cmp.megan	0.302, 0.548	0.634, 0.282	0.788, 0.072
20-1-12cmp.megan	0.457, 0.103	0.137, 0.924	0.790, 0.340
20-1-13cmp.megan	0.722, 0.841	0.458, 0.968	0.955, 0.215
20-1-14cmp.megan	0.835, 0.626	0.815, 0.558	0.209, 0.065
20-1-15cmp.megan	0.202, 0.152	0.114, 0.286	0.995, 0.066
20-1-16cmp.megan	0.324, 0.703	0.961, 0.359	0.983, 0.171
20-1-17cmp.megan	0.922, 0.079	0.270, 0.588	0.585, 0.207
20-1-18cmp.megan	0.212, 0.245	0.485, 0.972	0.722, 0.002
20-1-19cmp.megan	0.229, 0.740	0.441, 0.700	0.835, 0.204
20-1-20cmp.megan	0.914, 0.753	0.473, 0.991	0.227, 0.102

Table C.1.2: Similar comparison within Sea subsamples.

Data	Gammaproteobacteria	Bacteroidetes/ Chlorobi Group	Firmicutes
Soil vs Sea	UPv, DPv	UPv, DPv	UPv, DPv
20-1-1cmp.megan	0.0, 0.0	0.002, 0.0	0.0, 0.135
20-2-2cmp.megan	0.0, 0.0	2.29e-4, 0.0	0.0, 0.452
20-3-3cmp.megan	0.0, 0.0	0.002, 0.0	0.0, 0.018
20-4-4cmp.megan	0.0, 0.0	0.0, 0.0	0.0, 0.029
20-5-5cmp.megan	0.0, 0.0	5.05e-4, 0.0	0.0, 0.001
20-6-6cmp.megan	0.0, 0.0	0, 0.0	0.0, 0.12
20-7-7cmp.megan	0.0, 0.0	4.4e-5, 0.0	0.0, 0.016
20-8-8cmp.megan	0.0, 0.0	8.72e-5, 2.52e-4	0.0, 0.023
20-9-9cmp.megan	0.0, 0.0	0.0, 0.0	0.0, 0.471
20-10-10cmp.megan	0.0, 0.0	0.009, 0.0	0.0, 0.161
20-11-11cmp.megan	0.0, 0.0	1.52e-5, 0.002	0.0, 0.006
20-12-12cmp.megan	0.0, 0.0	0.0, 1.57e-4	0.0, 0.025
20-13-13cmp.megan	0.0, 0.0	0.0, 0.0	0.0, 0.0
20-14-14cmp.megan	0.0, 0.0	0.071, 1.76e-4	0.0, 0.001
20-15-15cmp.megan	0.0, 0.0	0.0, 4.06e-4	0.0, 7.61e-4
20-16-16cmp.megan	0.0, 0.0	0.002, 0.0	0.0, 9.57e-4
20-17-17cmp.megan	0.0, 0.0	0.0, 0.045	0.0, 0.034
20-18-18cmp.megan	0.0, 0.0	4.26e-4, 0.0	0.0, 2.74e-4
20-19-19cmp.megan	0.0, 0.0	0.0, 0.0	0.0, 0.022
20-20-20cmp.megan	0.0, 0.0	3.21e-5, 0.0	0.0, 1.32e-4

Table C.1.3: Comparison between 20 Soil and 20 Sea subsamples.

C.2 Chapter 5: Multiple Metagenome Comparison using Networks

This section includes the additional materials from Chapter 5.

Example of computing Goodall's index: The following numerical example illustrates the computation of Goodall's index for a small dataset. In this example, five datasets are characterized by the abundances of eight taxa. Small numbers are considered for an easy example. Example taken from [Legendre and Legendre, 1998] and modified for our purposes.

a) The original data:

Data	D1	D2	D3	D4	D5	Range R_i
Taxa-1	3	3	0	0	0	3
Taxa-2	0	0	2	2	0	2
Taxa-3	0	2	3	0	2	3
Taxa-4	0	0	4	3	3	4
Taxa-5	4	4	0	0	0	4
Taxa-6	0	2	0	3	3	3
Taxa-7	0	0	0	1	2	2
Taxa-8	3	3	0	0	0	3

b) These five datasets have $n(n-1)/2 = 10$ pair combinations to compare with each other. Now we compute a partial similarity measure (s_{pair_i}) for all possible pair combinations for each 'Taxa' (i) resulting in a matrix called Gower's matrix. The matrix has 6 rows (for 6 'Taxa') and 10 columns which correspond to the 10 pairs of datasets.

Pair combination of datasets

Data	D1 –D2	D1 –D3	D1 –D4	D1 –D5	D2 –D3	D2 –D4	D2 –D5	D3 –D4	D3 –D5	D4 –D5
Taxa-1	1	0	0	0	0	0	0	0	0	0
Taxa-2	0	0	0	0	0	0	0	1	0	0
Taxa-3	0.33	0	0	0.33	0.67	0.33	1	0	0.67	0.33
Taxa-4	0	0	0.25	0.25	0	0.25	0.25	0.75	0.75	1
Taxa-5	1	0	0	0	0	0	0	0	0	0
Taxa-6	0.33	0	0	0	0.33	0.67	0.67	0	0	1
Taxa-7	0	0	0.50	0	0	0.50	0	0.50	0	0.50
Taxa-8	1	0	0	0	0	0	0	0	0	0

c) Now we compute, for each pair of dataset and each row (taxa), the proportion of partial similarity values in the row that are larger than or equal to the partial similarity of the pair of datasets being considered. The value under consideration is itself included in the proportion. For example, for the pair of datasets (D1–D4), the fourth taxa has a similarity of 0.25. In the fourth row, there are 7 values out

of 10 that are larger than or equal to 0.25. Thus the ratio associated with the pair (D1–D4) in the table is 0.7.

Pair combination of datasets

Data	D1 –D2	D1 –D3	D1 –D4	D1 –D5	D2 –D3	D2 –D4	D2 –D5	D3 –D4	D3 –D5	D4 –D5
Taxa-1	0.1	1	1	1	1	1	1	1	1	1
Taxa-2	1	1	1	1	1	1	1	0.1	1	1
Taxa-3	0.7	1	1	0.7	0.3	0.7	0.1	1	0.3	0.7
Taxa-4	1	1	0.7	0.7	1	0.7	0.7	0.3	0.3	0.1
Taxa-5	0.1	1	1	1	1	1	1	1	1	1
Taxa-6	0.5	1	1	1	0.5	0.3	0.3	1	1	0.1
Taxa-7	1	1	0.4	1	1	0.4	1	0.4	1	0.4
Taxa-8	0.1	1	1	1	1	1	1	1	1	1

d) Finally in the next table, a Dataset $\times$ Dataset symmetric matrix is computed, that records the products of the terms in each column of the previous table.

Data	D1	D2	D3	D4	D5
D1					
D2	0.00035	-			
D3	1.00000	0.15000	-		
D4	0.28000	0.05880	0.01200	-	
D5	0.49000	0.02100	0.09000	0.00280	-

From the above mentioned steps we get Goodall's index. Now as Goodall's index is a semimetric measures we can use $Distance = 1 - Similarity$.

Additional networks: All the additional networks from Chapter 5 are shown on following pages.

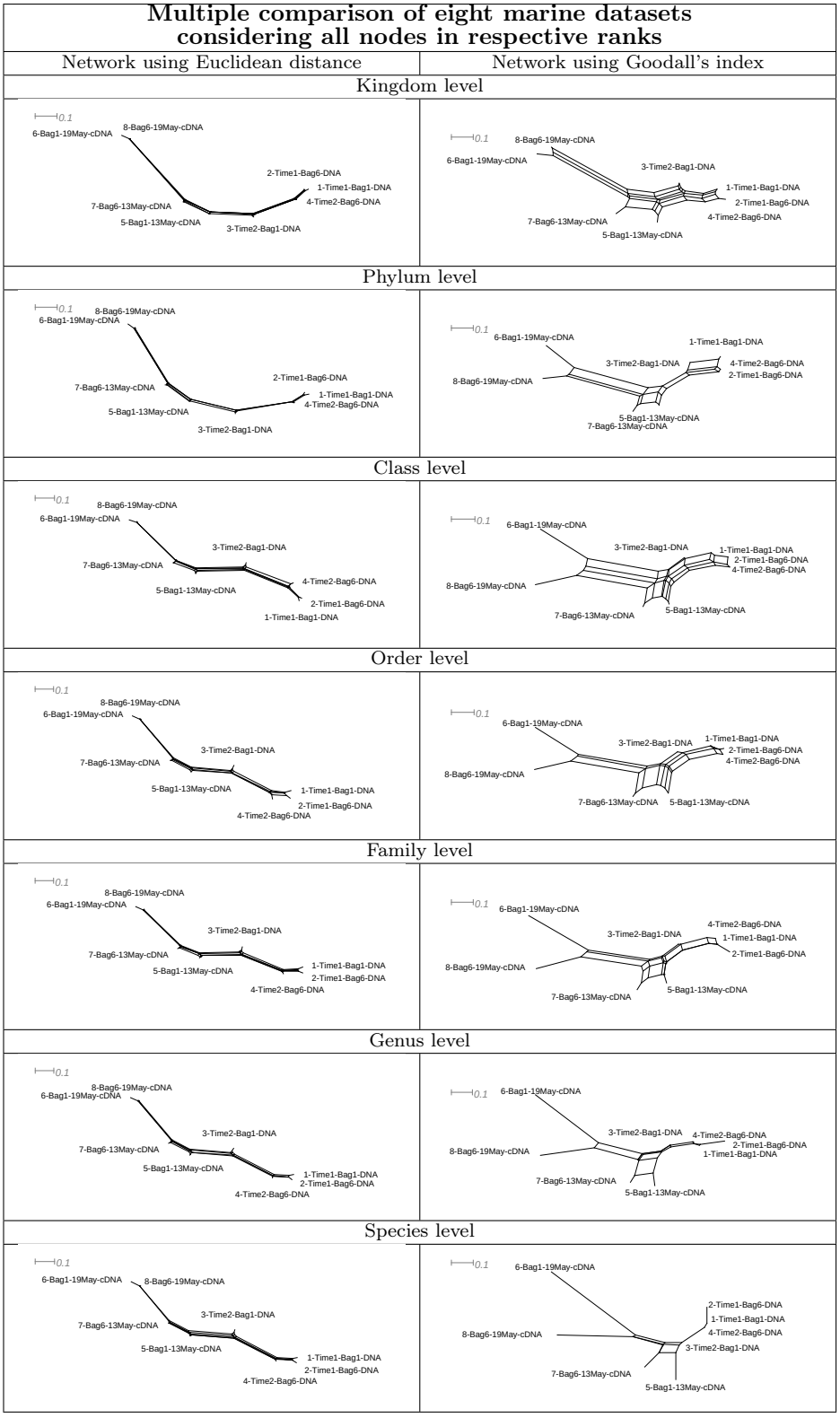

Figure C.2.1: Networks obtained using Euclidean distances (left column) and Goodall's index (right column), showing the comparison of eight Bergen marine samples (four metagenomes and four metatranscriptomes) considering all nodes at the indicated ranks of the NCBI taxonomy.

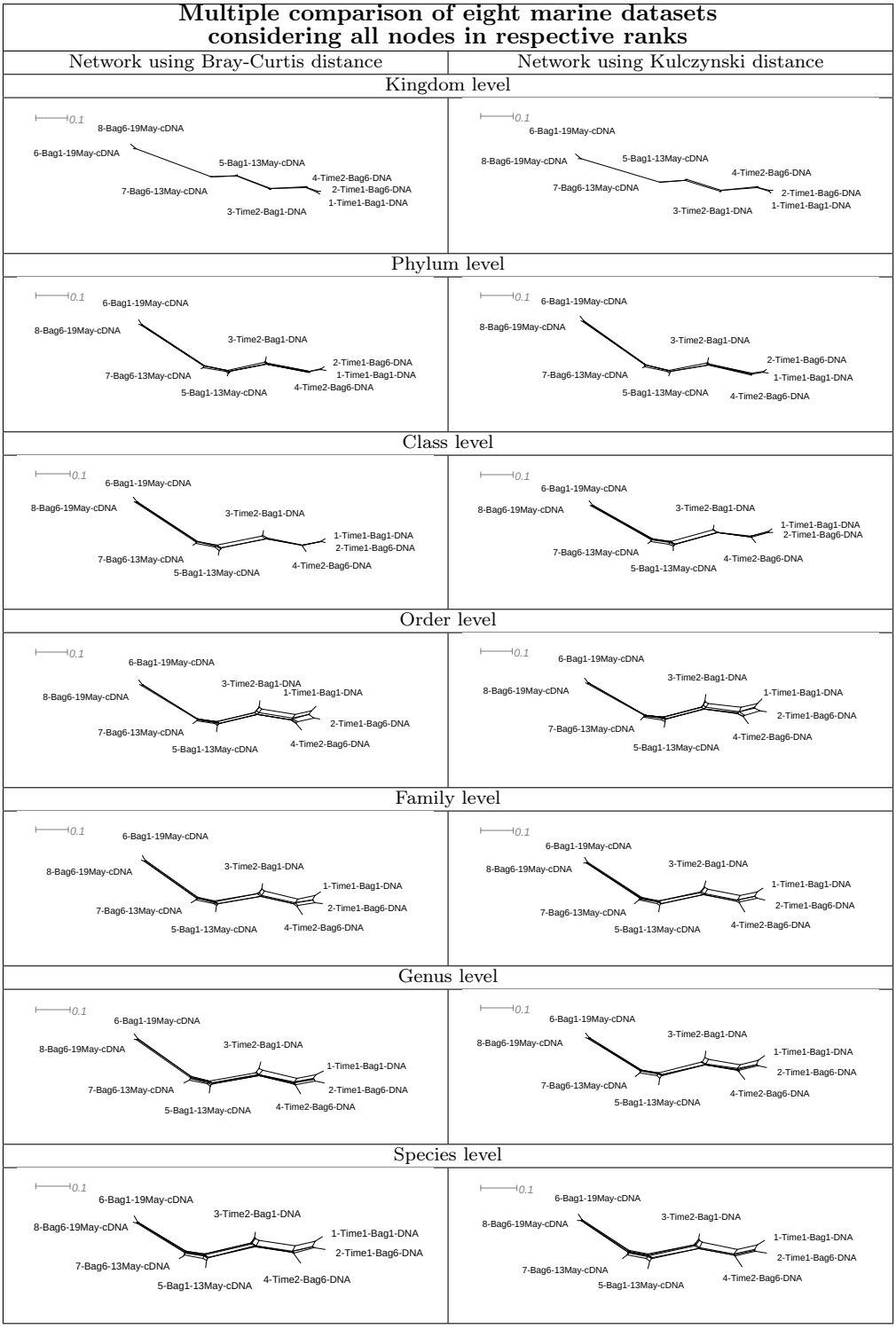

Figure C.2.2: Networks obtained using Bray-Curtis (left column) and Kulczynski distances (right column), showing the comparison of eight Bergen marine samples (four metagenomes and four metatranscriptomes) considering all nodes at the indicated ranks of the NCBI taxonomy.

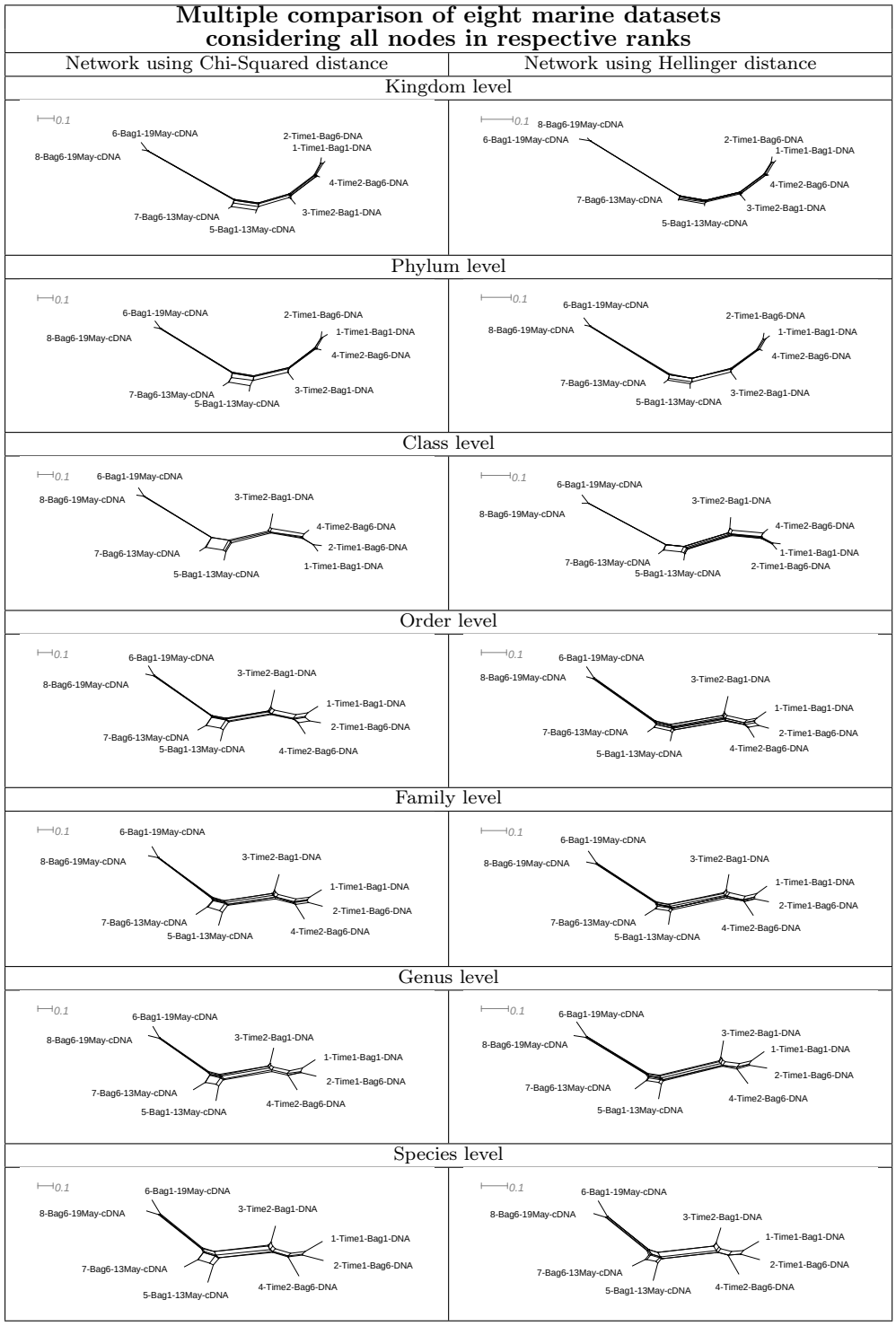

Figure C.2.3: Networks obtained using Chi-Squared (left column) and Hellinger distances (right column), showing the comparison of eight Bergen marine samples (four metagenomes and four metatranscriptomes) considering all nodes at the indicated ranks of the NCBI taxonomy.

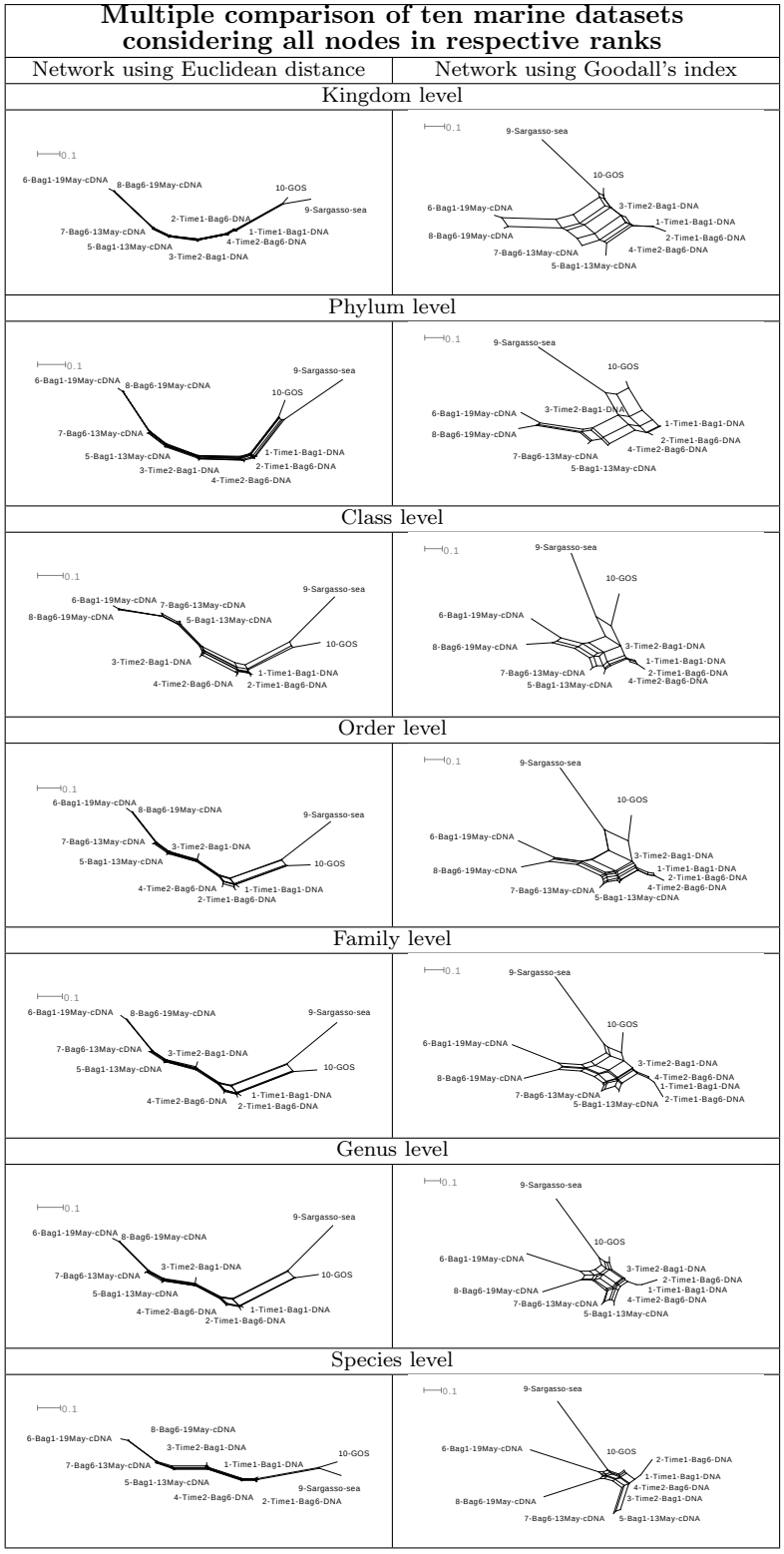

Figure C.2.4: Networks obtained using Euclidean distances (left column) and Goodall's index (right column), showing the comparison of ten marine samples (Sargasso Sea and GOS samples together with eight Bergen marine samples) considering all nodes at the indicated ranks of the NCBI taxonomy.

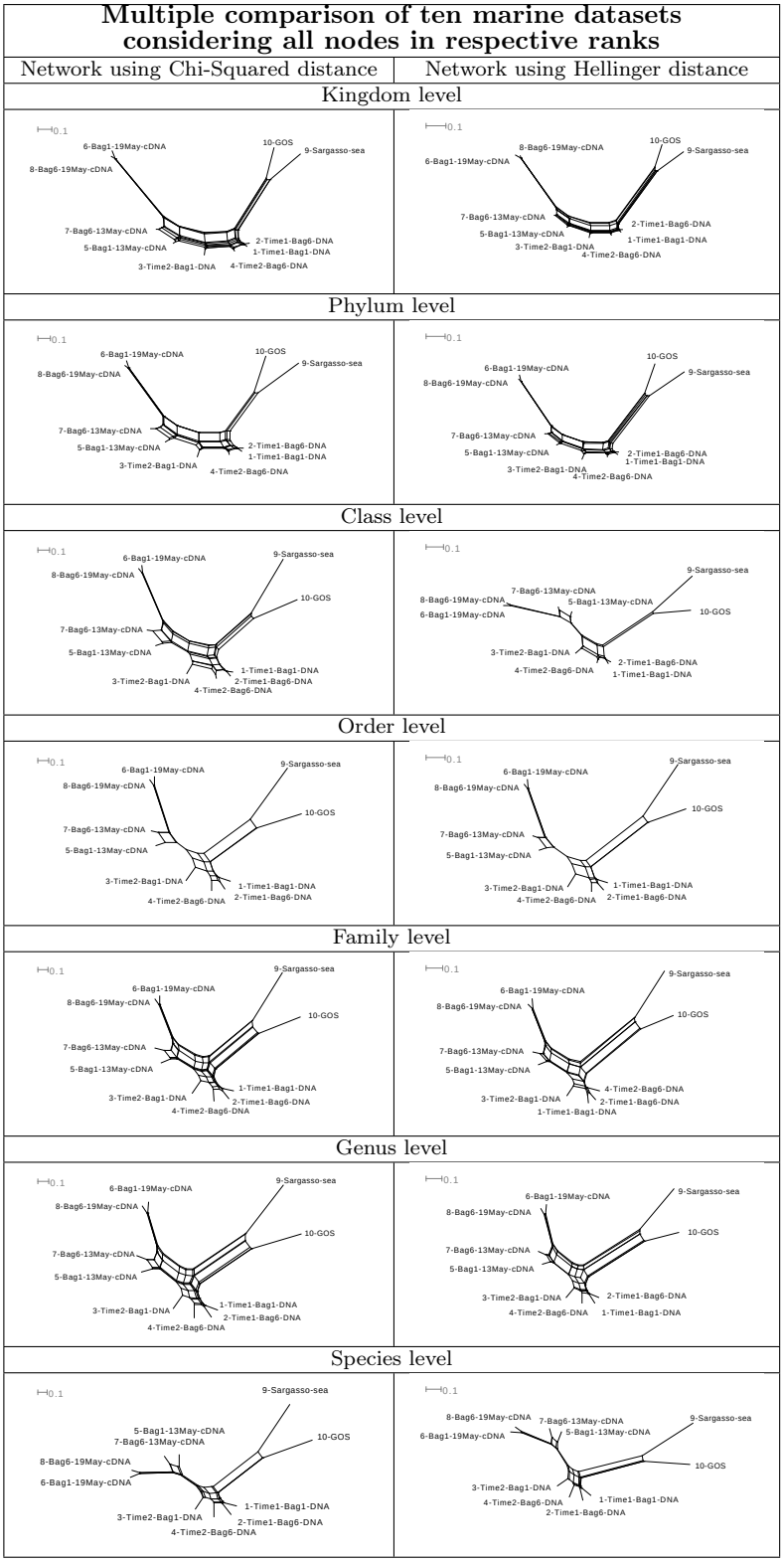

Figure C.2.5: Networks obtained using Chi-Squared distance (left column) and Hellinger distance (right column), showing the comparison of ten marine samples (Sargasso Sea and GOS samples together with eight Bergen marine samples) considering all nodes at the indicated ranks of the NCBI taxonomy.

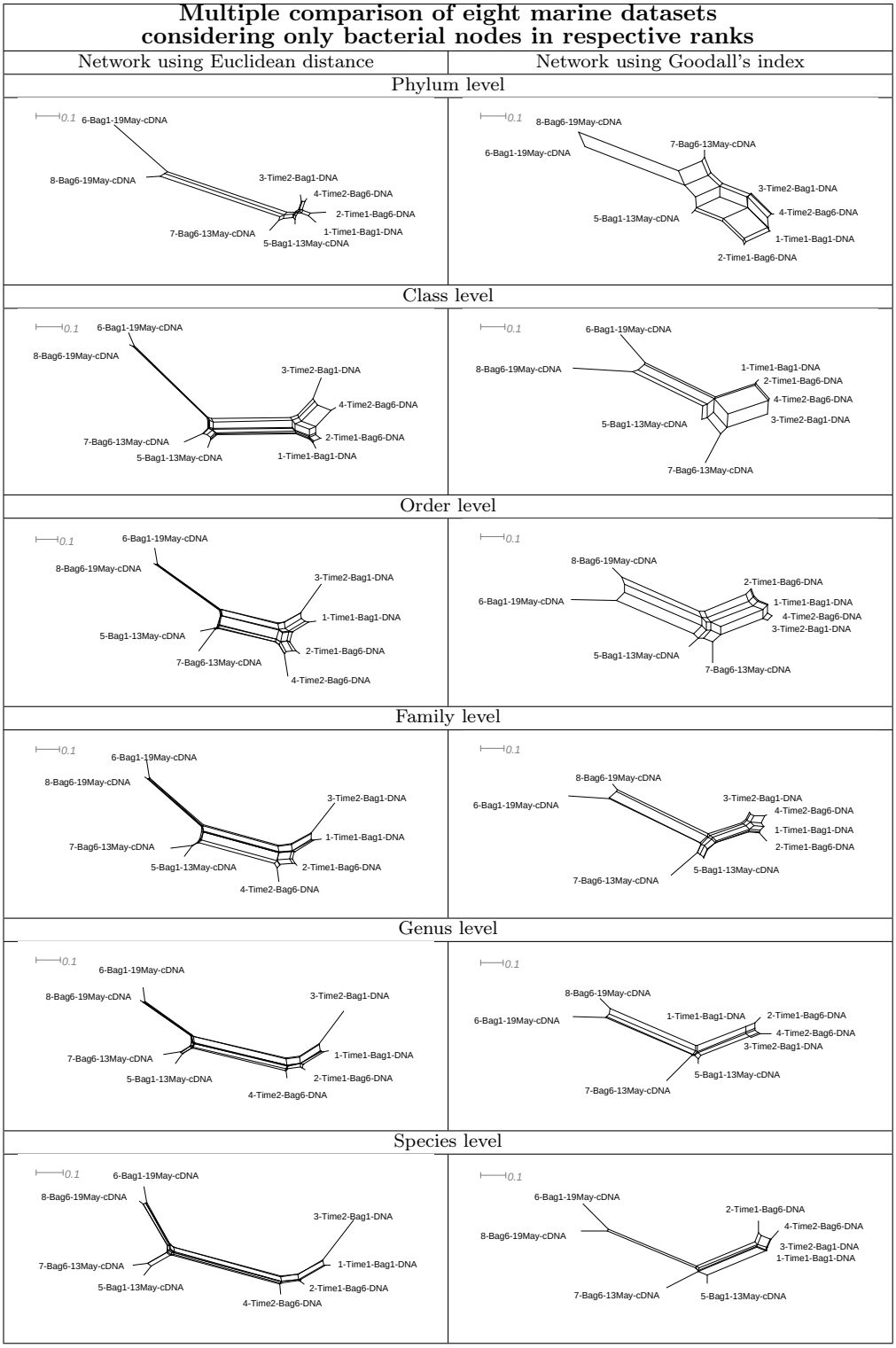

Figure C.2.6: Networks obtained using Euclidean distances (left column) and Goodall's index (right column), showing the comparison of eight Bergen marine samples (four metagenomes and four metatranscriptomes) considering only bacterial nodes at the indicated ranks of the NCBI taxonomy.

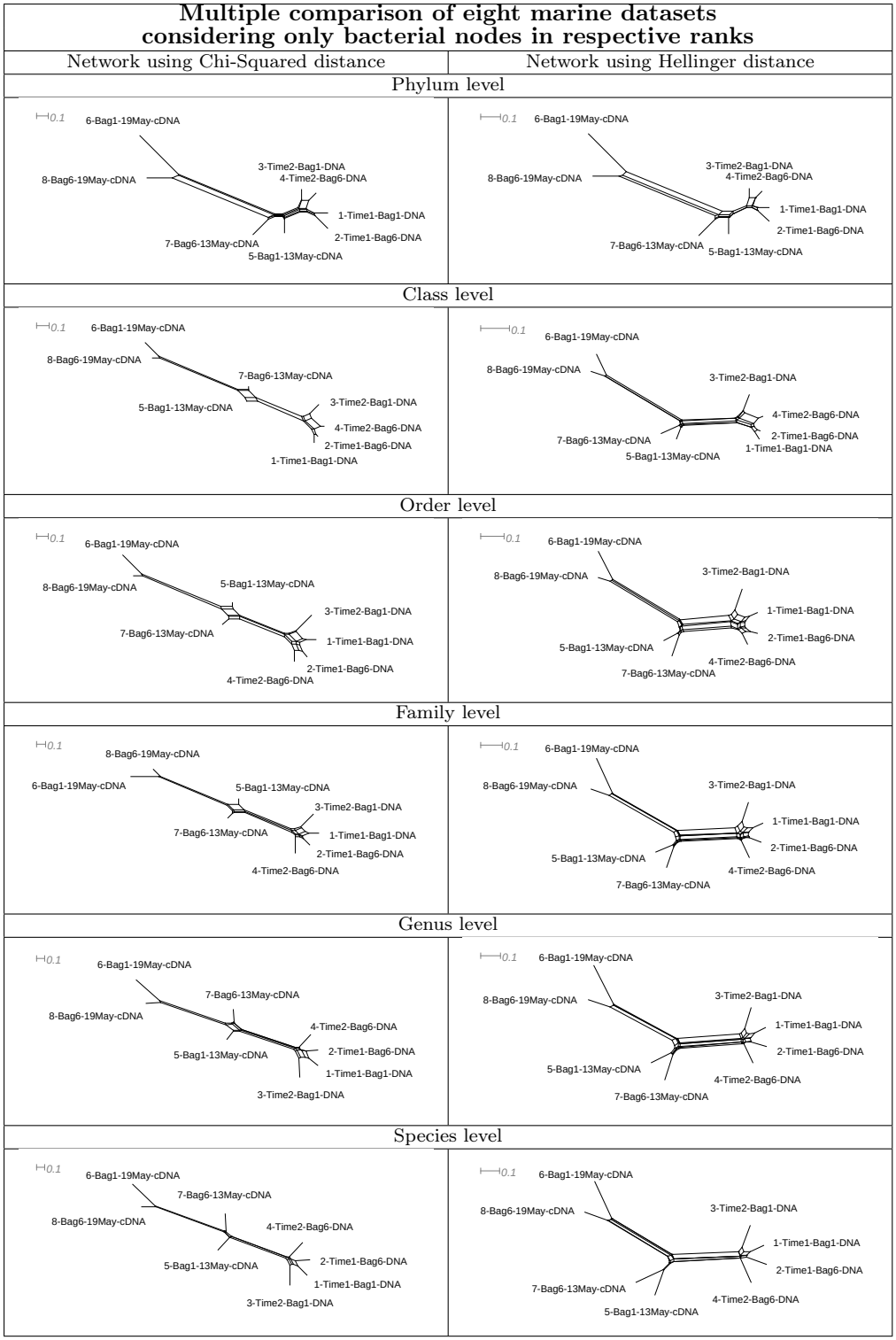

Figure C.2.7: Networks obtained using Chi-Squared distances (left column) and Hellinger dintances (right column), showing the comparison of eight Bergen marine samples (four metagenomes and four metatranscriptomes) considering only bacterial nodes at the indicated ranks of the NCBI taxonomy.

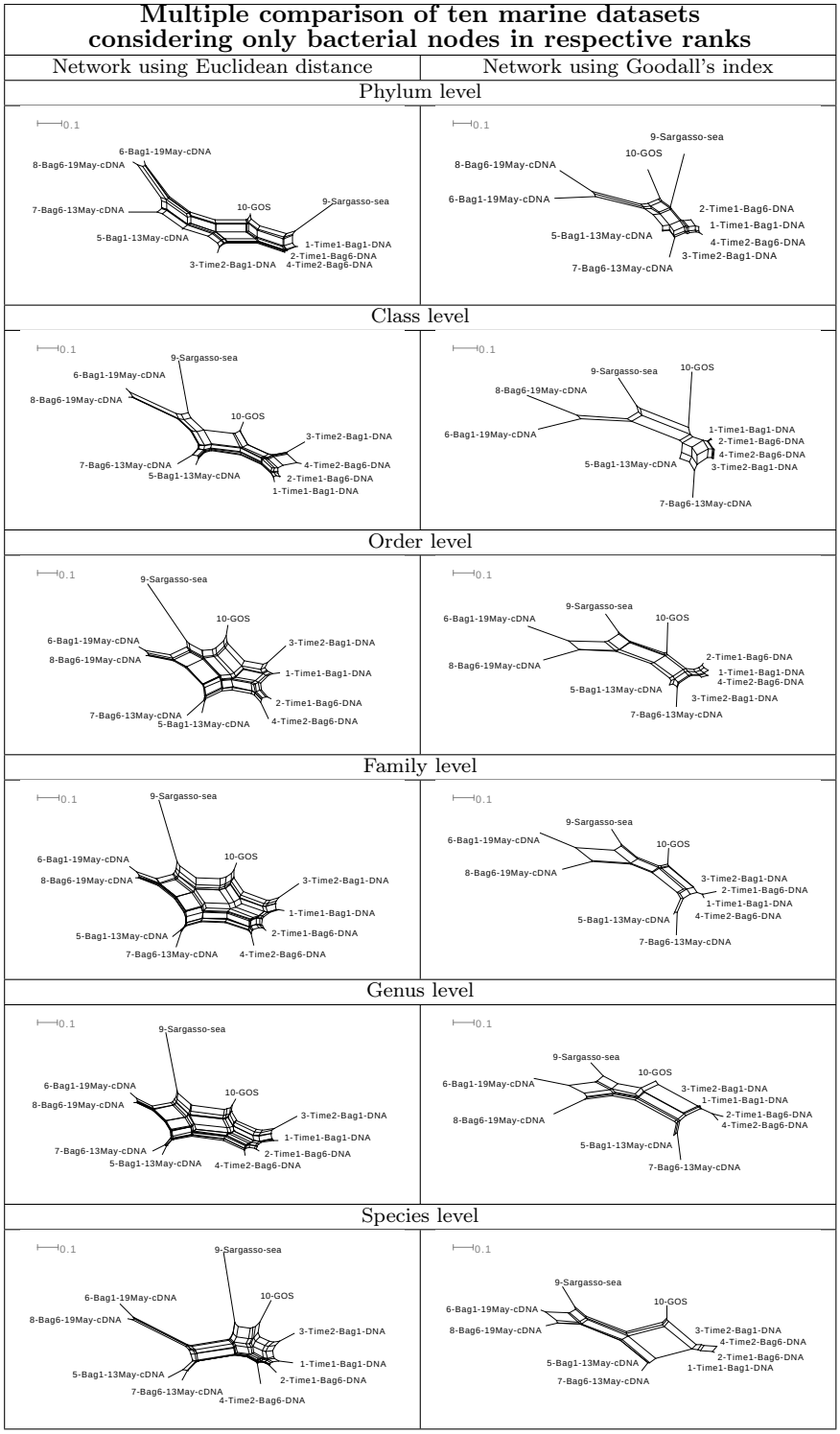

Figure C.2.8: Networks obtained using Euclidean distances (left column) and Goodall's index (right column), showing the comparison of ten marine samples (Sargasso Sea and GOS samples together with eight Bergen marine samples) considering only bacterial nodes at the indicated ranks of the NCBI taxonomy.

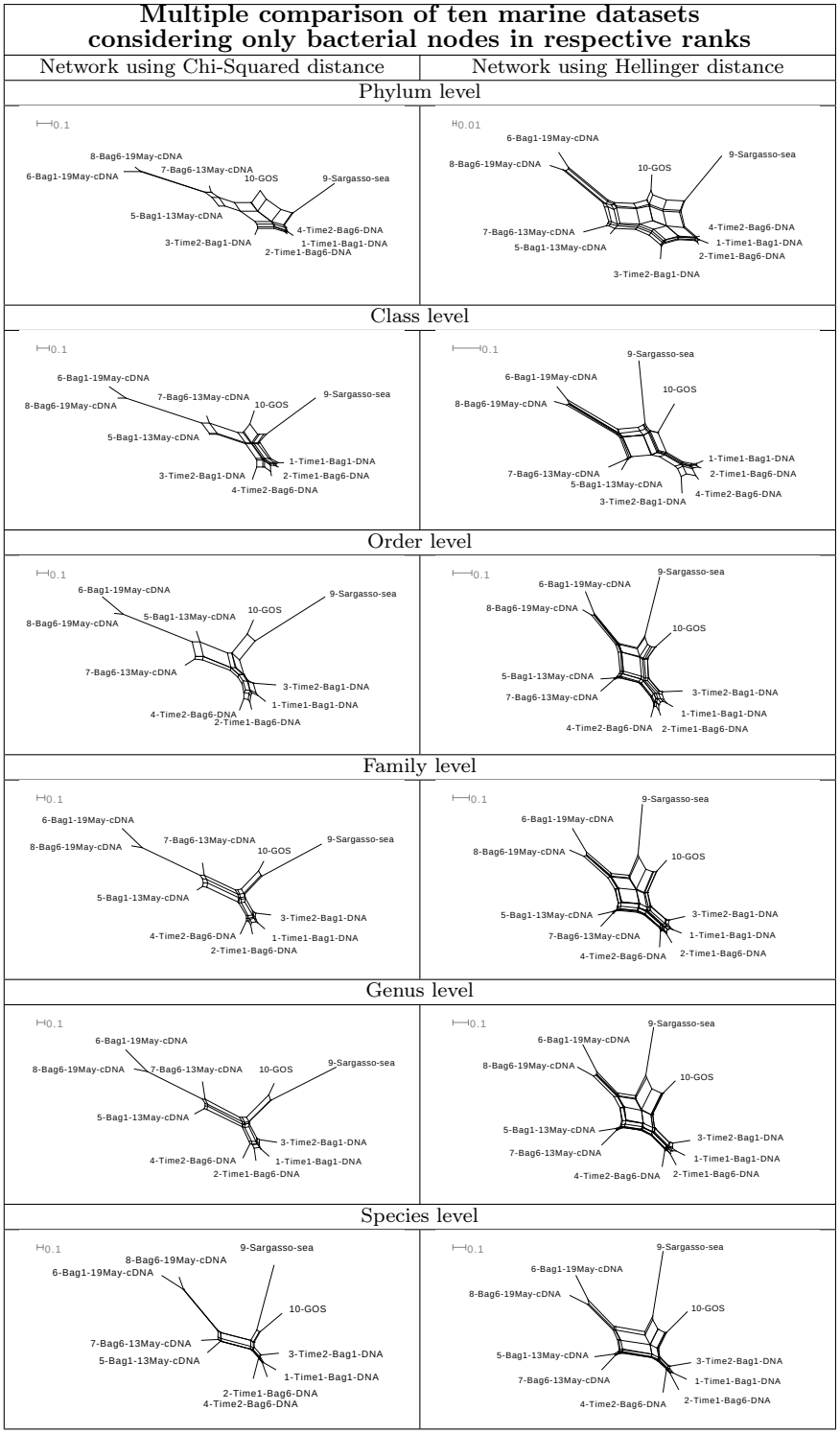

Figure C.2.9: Networks obtained using Chi-Squared distances (left column) and Hellinger distance s(right column), showing the comparison of ten marine samples (Sargasso Sea and GOS samples together with eight Bergen marine samples) considering only bacterial nodes at the indicated ranks of the NCBI taxonomy.

Appendix D

Internet Resources

- BLAST (`http://blast.ncbi.nlm.nih.gov/Blast.cgi`)
- COG database (`http://www.ncbi.nlm.nih.gov/COG/`)
- GO (`http://www.geneontology.org/`)
- GO slims (`http://www.geneontology.org/GO.slims.shtml`)
- GOLD: Genomes OnLine Database (`http://www.genomesonline.org/`)
- KEGG (`http://www.genome.jp/kegg/`)
- MEGAN DB (`http://www.megan-db.org`)
- MEGAN software (`http://www-ab.informatik.uni-tuebingen.de/software/megan`)
- NCBI-NR/NT database (`ftp://ftp.ncbi.nih.gov/blast/db/`)
- Refseq (`http://www.ncbi.nlm.nih.gov/RefSeq`)
- SEED (`ftp://ftp.theseed.org`)
- SEED to NCBI mapping file `ftp://ftp.theseed.org/misc/Data/idmapping/seed2ncbi.gz`.

Bibliography

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., *et al.*, 1990. Basic local alignment search tool. *J Mol Biol*, **215**:403–410.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., *et al.*, 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1):25–29.

[Astbury, 1937] Astbury, W., 1937. Nucleic acid. *Symp SOC Exp Bbl*, **1**(66).

[Avery et al., 1944] Avery, O., MacLeod, C., and McCarty, M., 1944. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *J Exp Me*, **79**:137–158.

[Azam and Malfatti, 2007] Azam, F. and Malfatti, F., 2007. Microbial structuring of marine ecosystems. *Nat Rev Microbiol*, **5**:782–791.

[Benson et al., 2005] Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D., *et al.*, 2005. Genbank. *Nucleic Acids Res*, **1**(33 (Database issue)):D34–38.

[Bentley, 2006] Bentley, D., 2006. Whole-genome re-sequencing. *Curr Opin Genet Dev*, **16**:545–552.

[Branin and Case, 1998] Branin, J. J. and Case, M., 1998. Reforming scholarly publishing in the sciences: A librarian perspective. *Notices Amer Math Soc*, **45**(4):475–486.

[Braslavsky et al., 2003] Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S. R., 2003. Sequence information can be obtained from single dna molecules. *PNAS*, **100**(7):3969–3964.

[Bray and Curtis, 1957] Bray, R. J. and Curtis, J. T., 1957. An ordination of the upland forest communities of southern wisconsin. *Ecol Monogr*, **27**:325–349.

[Breitbart et al., 2002] Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., *et al.*, 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A*, **99**(22):14250–5.

[Bryant and Moulton, 2004] Bryant, D. and Moulton, V., 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol*, **21**(2):255–265.

[Chargaff, 1951] Chargaff, E., 1951. Structure and function of nucleic acid as cell constituent. *Fed Proc*, **10**:654–659.

[Chargaff et al., 1949] Chargaff, E., Vischer, E., Doniger, R., Green, C., and Misani, F., *et al.*, 1949. The composition of the desoxypentose nucleic acid of thymus and spleen. *J Biol Chem*, **177**:405–416.

[Check Hayden, 2009] Check Hayden, E., 2009. Genome sequencing: the third generation. *Nature*, **457**(7231):768–769.

[Committee on Metagenomics, 2007] Committee on Metagenomics, Challenges and Functional Applications, National Research Council, 2007. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press.

[Cox-Foster et al., 2007] Cox-Foster, D. L., Conlan, S., Holmes, E. C., Palacios, G., Evans, J. D., *et al.*, 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*, **308**(5848):283–287.

[Craft et al., 2010] Craft, J. A., Gilbert, J. A., Temperton, B., Dempsey, K. E., Ashelford, K., *et al.*, 2010. Pyrosequencing of mytilus galloprovincialis cDNAs: tissue-specific expression patterns. *PLoS One*, **5**(1):e8875.

[Culley et al., 2006] Culley, A. I., Lang, A. S., and Suttle, C. A., 2006. Metagenomic analysis of coastal rna virus communities. *Science*, **312**(5781):1795–8.

[Dahm, 2008] Dahm, R., 2008. Discovering DNA: Friedrich miescher and the early years of nucleic acid research. *Hum Genet*, **122**(6):565–81.

[Davison and Hinkley, 1997] Davison, A. and Hinkley, D., 1997. *Bootstrap methods and their application*. Cambridge University Press.

[DeLong, 2004] DeLong, E., 2004. Microbial population genomics and ecology: the road ahead. *Environ Microbiol*, **6**(9):875–878.

[DeLong, 2005] DeLong, E. F., 2005. Microbial community genomics in the ocean. *Nat Rev Microbiol.*, **3**(6):459–69.

[DeLong, 2007] DeLong, E. F., 2007. Sea change for metagenomics? *Nature*, **5**:326.

[DeLong et al., 2006] DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., *et al.*, 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, **311**(5760):496–503.

[Dress and Huson, 2004] Dress, A. W. M. and Huson, D., 2004. Constructing splits graphs. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, **1**(3):109–115.

[Droege and Hill, 2008] Droege, M. and Hill, B., 2008. The genome sequencer flx system–longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol*, **136**(1-2):3–10.

[Dutilh et al., 2008] Dutilh, B. E., He, Y., Hekkelman, M. L., and Huynen, M. A., 2008. Signature, a web server for taxonomic characterization of sequence samples using signature genes. *Nucleic Acids Res*, **36**(Web Server issue):W470–W474.

[Eckburg et al., 2005] Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., *et al.*, 2005. Diversity of the human intestinal microbial flora. *Science*, **308**(5728):1635–1638.

[Edwards and Caskey, 1991] Edwards, A. and Caskey, T., 1991. Closure strategies for random dna sequencing. *Methods.*, **3**(1):41–47.

[Feely et al., 2004] Feely, R. A., Sabine, C. L., Lee, K., Berelson, W., Kleypas, J., *et al.*, 2004. Impact of anthropogenic CO_2 on the $CaCO_3$ system in the oceans. *Science*, **305**(5682):362–366.

[Fierer et al., 2007] Fierer, N., Bradford, M. A., and Jackson, R. B., 2007. Toward an ecological classification of soil bacteria. *J Ecol*, **88**(6):1354–4.

[Fuhrman et al., 2006] Fuhrman, J. A., Hewson, I., Schwalbach, M. S., Steele, J. A., Brown, M. V., *et al.*, 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *PNAS*, **103**(35):13104–9.

[Fullwood et al., 2009] Fullwood, M. J., Wei, C.-L., Liu, E. T., and Ruan, Y., 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*, **19**(4):521–532.

[Garcia Martin et al., 2006] Garcia Martin, H., Ivanova, N., Kunin, V., Warnecke, F., Barry, K. W., *et al.*, 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol*, **24**(10):1263–9.

[Gianoulis et al., 2009] Gianoulis, T. A., Raes, J., Patel, P. V., Bjornson, R., Korbel, J. O., *et al.*, 2009. Quantifying environmental adaptation of metabolic pathways in metagenomics. *PNAS*, **106**(5):1374–9.

[Gilbert et al., 2009] Gilbert, J., Field, D., Swift, P., Newbold, L., Oliver, A., *et al.*, 2009. The seasonal structure of microbial communities in the western english channel. *Environ Microbiol*, **11**(12):3132–9.

[Gilbert et al., 2008a] Gilbert, J. A., Field, D., Huang, Y., Edwards, R., Li, W., *et al.*, 2008a. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*, **3**:e3042.

[Gilbert et al., 2008b] Gilbert, M., Tomsho, L., Rendulic, S., Packard, M., Drautz, D., *et al.*, 2008b. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science*, **322**(5903):857.

[Gilbert et al., 2007] Gilbert, M. T. P., Tomsho, L. P., Rendulic, S., Packard, M., Drautz, D. I., *et al.*, 2007. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science*, **317**(5846):1927–30.

[Gilbert, 1980] Gilbert, W., 1980. DNA sequencing and gene structure. *Nobel lecture*, **8**.

[Gill et al., 2006] Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., *et al.*, 2006. Metagenomic analysis of the human distal gut microbiome. *Science*, **312**(5778):1355–9.

[Good, 2004] Good, P., 2004. *Permutation, parametric, and bootstrap tests of hypotheses.* Springer; 2nd edition.

[Goodall, 1964] Goodall, D. W., 1964. A probabilistic similarity index. *Nature*, **203**:1098.

[Goodall, 1966] Goodall, D. W., 1966. A new similarity index based on probability. *Biometrics*, **22**:882–907.

[Green et al., 2006] Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., *et al.*, 2006. Analysis of one million base pairs of neanderthal DNA. *Nature*, **444**(7117):330–336.

[Handelsman, 2004] Handelsman, J., 2004. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, **68**(4):669–685.

[Handelsman et al., 1998] Handelsman, J., Rondon, M. R., Brady, S. G., Clardy, J., and Goodman, R., *et al.*, 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, **5**:245–249.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. H., 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer; 2nd edition.

[Hershey and Chase, 1952] Hershey, A. and Chase, M., 1952. Independent functions of viral proteins and nucleic acid in growth of bacteriophage. *J Gen Physiol*, **36**:39–56.

[Hofreiter, 2008] Hofreiter, M., 2008. Paleogenomics. *C R Palevol*, **7**:113–124.

[Holm, 1979] Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scand J Stat*, **6**:65–70.

[Hong, 1981] Hong, G. F., 1981. A method for sequencing single-stranded cloned DNA in both directions. *Biosci Rep*, **1**(3):243–252.

[Human Microbiome Jumpstart Reference Strains Consortium et al., 2010] Human Microbiome Jumpstart Reference Strains Consortium, Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., *et al.*, 2010. A catalog of reference genomes from the human microbiome. *Science*, **328**(5981):994–999.

[Hunt et al., 2008] Hunt, D., David, L., Gevers, D., Preheim, S., Alm, E., *et al.*, 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science*, **320**:1081–5.

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

[Huson et al., 2009] Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F., and Schuster, S. C., *et al.*, 2009. Methods for comparative metagenomics. *BMC Bioinformatics*, **10 Suppl 1**:S12.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S., 2000. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**(1):27–30.

[Kasianowicz et al., 1996] Kasianowicz, J. J., Brandin, E., Branton, D., and Deamer, D. W., 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci USA*, **93**:13770–3.

[Kircher and Kelso, 2010] Kircher, M. and Kelso, J., 2010. High-throughput dna sequencing–concepts and limitations. *Bioessays*, **32**(6):524–36.

[Korbel et al., 2007] Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., *et al.*, 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**(5849):420–6.

[Korf et al., 2003] Korf, I., Yandell, M., and Bedell, J., 2003. *BLAST*. O'Reilly.

[Krause et al., 2008] Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., *et al.*, 2008. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res*, **36(7)**:2230–9.

[Kristiansson et al., 2009] Kristiansson, E., Hugenholtz, P., and Dalevi, D., 2009. ShotgunfunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics*, **25**(20):2737–8.

[Kuever et al., 2005] Kuever, J., Rainey, F. A., and Widdel, F., 2005. *Bergey's Manual of Systematic Bacteriology*. Springer.

[Lambert et al., 1967] Lambert, R., Chassignol, S., Sedallian, A., Descos, L., and Martin, F., *et al.*, 1967. Influence of gastrectomy and by-passing of the stomach on the intestinal flora of the rat. *J Pathol Bacteriol*, **94**(1):183–189.

[Lang-Unnasch et al., 1998] Lang-Unnasch, N., Reith, M. E., Munholland, J., and Barta, J. R., 1998. Plastids are widespread and ancient in parasites of the phylum apicomplexa. *Int J Parasitol*, **28**(11):1743–54.

[Lebart et al., 1979] Lebart, L., Morineau, A., and Félon, J. P., 1979. *Traitement des donndées statistiques - Méthodes et programmes.* Dunod. Paris.

[Legendre and Legendre, 1998] Legendre, P. and Legendre, L., 1998. *Numerical Ecology*. Elsevier Science Publishers B. V.

[Levene et al., 2003] Levene, M., Korlach, J., Turner, S., Foquet, M., Craighead, H., *et al.*, 2003. Zero-mode waveguides for single molecule analysis at high concentrations. *Science*, **299**:682–686.

[Levene, 1919] Levene, P., 1919. The structure of yeast nucleic acid. *J Biol Chem*, **40**(2):415–424.

[Li et al., 2010] Li, L., Victoria, J. G., Wang, C., Jones, M., Fellers, G. M., *et al.*, 2010. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *J Virol*, **84**(14):6955–65.

[Lozupone et al., 2006] Lozupone, C., Hamady, M., and Knight, R., 2006. Unifrac - an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, **7**:371.

[Mardis, 2008] Mardis, E. R., 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, **9**:387–402.

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

[Markowitz et al., 2008] Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., *et al.*, 2008. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res*, **36**:D534–D538.

[Markowitz et al., 2006] Markowitz, V. M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., *et al.*, 2006. The integrated microbial genomes (IMG) system. *Nucleic Acids Res*, **34**:344–348. Database-Issue.

[Mavromatis et al., 2007] Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., *et al.*, 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, **4**(6):495–500.

[Maxam and Gilbert, 1977] Maxam, A. and Gilbert, W., 1977. A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, **74**(2):560–564.

[McHardy et al., 2006] McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I., *et al.*, 2006. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods.*, **4**(1):63–72.

[Metzker, 2010] Metzker, M. L., 2010. Sequencing technologies - the next generation. *Nat Rev Genet*, **11**(1):31–46.

[Meyer et al., 2008] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., *et al.*, 2008. The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**:386.

[Miller, 1981] Miller, R. G. J., 1981. *Simultaneous Statistical Inference*. Springer; 2nd edition.

[Miller et al., 2009] Miller, W., Drautz, D. I., Janecka, J. E., Lesk, A. M., Ratan, A., *et al.*, 2009. The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*) . *Genome Res*, **19**(2):213–220.

[Mitra et al., 2010a] Mitra, S., Gilbert, J. A., Field, D., and Huson, D. H., 2010a. Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *ISME J*, **4**:1236–1242.

[Mitra et al., 2009] Mitra, S., Klar, B., and Huson, D. H., 2009. Visual and statistical comparison of metagenomes. *Bioinformatics*, **25**(15):1849–55.

[Mitra et al., 2010b] Mitra, S., Schubach, M., and Huson, D. H., 2010b. Short clones or long clones? a simulation study on the use of paired reads in metagenomics. *BMC Bioinformatics*, **11 Suppl 1**:S12.

[Murray et al., 1998] Murray, A. E., Preston, C. M., Massana, R., Taylor, L. T., Blakis, A., *et al.*, 1998. Seasonal and spatial variability of bacterial and archaeal assemblages in the coastal waters near anvers island, antarctica. *Appl Environ Microbiol*, **64**(7):2585–95.

[Nair et al., 2006] Nair, G., Safa, A., Bhuiyan, N., Nusrin, S., Murphy, D., *et al.*, 2006. Isolation of *vibrio cholerae* O1 strains similar to pre-seventh pandemic El tor strains during an outbreak of gastrointestinal disease in an island resort in Fiji. *J Med Microbiol*, **55**:1559–62.

[Nakicenovic et al., 2001] Nakicenovic, N., Alcamo, J., Davis, G., de Vries, B., Fenhann, J., *et al.*, 2001. *Special Report on Emissions Scenarios*. Number ISBN 0521804930. Cambridge University Press.

[Odum, 1950] Odum, E. P., 1950. Bird populations of the highlands (North Carolina) plateau in relation to plant succession and avian invasion. *Ecology*, **31**:587–605.

[Olsen et al., 1986] Olsen, G., Lane, D., Giovannoni, S., Pace, N., and Stahl, D., *et al.*, 1986. Microbial ecology and evolution: a ribosomal rna approach. *Annu Rev Microbiol*, **40**:337–365.

[Overbeek et al., 2005] Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., *et al.*, 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, **33**(17):5691–02.

[Pace et al., 1985] Pace, N., Stahl, D., Olsen, G., and Lane, D., 1985. Analyzing natural microbial populations by rRNA sequences. *American Society for Microbiology News*, **51**:4–12.

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

[Parson, 2005] Parson, A., 2005. Craig Venter, of human genome fame, mines the oceans for genetic riches. *The San Diego Union Tribune*, **11/9/05**.

[Pennisi, 2006] Pennisi, E., 2006. Genomics. on your mark. get set. sequence! *Science*, **314**(5797):232.

[Poinar et al., 2006] Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R. D. E., *et al.*, 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, **311**(5759):392–394.

[Pruitt et al., 2009] Pruitt, K., Tatusova, T., Klimke, W., and Maglott, D., 2009. NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*, **37**(Database issue):D32–6.

[Pushkarev et al., 2009] Pushkarev, D., Neff, N. F., and Quake, S. R., 2009. Single-molecule sequencing of an individual human genome. *Nat Biotechnol*, **27**(9):847–52.

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

[Ramérez et al., 2009] Ramérez, O., Gigli, E., Bover, P., Alcover, J. A., Bertranpetit, J., *et al.*, 2009. Paleogenomics in a temperate environment: shotgun sequencing from an extinct mediterranean caprine. *PLoS One*, **4**(5):e5670.

[Rao, 1995] Rao, C. R., 1995. A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Qüestiió (Quaderns d'Estadística i lnvestigació 0perativa)*, **19**:23–63.

[Rhee et al., 2008] Rhee, S., Wood, V., Dolinski, K., and Draghici, S., 2008. Use and misuse of the gene ontology annotations. *Nat Rev Genet*, **9**(7):509–515.

[Richter et al., 2008] Richter, D. C., Ott, F., Auch, A. F., Schmid, R., and Huson, D. H., *et al.*, 2008. Metasim—a sequencing simulator for genomics and metagenomics. *PLoS One*, **3**(10):e3373+.

[Riebesell et al., 2007] Riebesell, U., Schulz, K. G., Bellerby, R. G. J., Botros, M., Fritsche, P., *et al.*, 2007. Enhanced biological carbon consumption in a high CO_2 ocean. *Nature*, **450**(7169):545–548.

[Rodriguez-Brito et al., 2006] Rodriguez-Brito, B., Rohwer, F., and Edwards, R. A., 2006. An application of statistics to comparative metagenomics. *BMC Bioinformatics*, **7**:162.

[Rusch and *et al.*, 2007] Rusch, D. B. and *et al.*, 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*, **5**(3):e77.

[Rusch et al., 2007] Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., *et al.*, 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*, **5**(3):e77.

[Rusk, 2009] Rusk, N., 2009. Cheap third-generation sequencing. *Nature Methods*, **6**(4):244.

[Saitou and Nei, 1987] Saitou, N. and Nei, M., 1987. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. *Mol Biol and Evol*, **4**:406–425.

[Salyers and Whitt, 2000] Salyers, A. A. and Whitt, D. D., 2000. *Microbiology: Diversity, Disease and the Environment*. Fitzgerald Science Press.

[Sanger, 1980] Sanger, F., 1980. Determination of nucleotide sequences in DNA. *Nobel lecture*, **8**.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R., 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, **74**(12):5463–7.

[Sapkota et al., 2010] Sapkota, A. R., Berger, S., and Vogel, T. M., 2010. Human pathogens abundant in the bacterial metagenome of cigarettes. *Environ Health Perspect*, **118**(3):351–6.

[Schloss and Handelsman, 2006a] Schloss, P. and Handelsman, J., 2006a. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol*, **72**:6773–79.

[Schloss and Handelsman, 2005] Schloss, P. D. and Handelsman, J., 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol*, **71**(3):1501–6.

[Schloss and Handelsman, 2006b] Schloss, P. D. and Handelsman, J., 2006b. Introducing treeclimber, a test to compare microbial community structures. *Appl Environ Microbiol*, **72**(4):2379–84.

[Schloss and Handelsman, 2006c] Schloss, P. D. and Handelsman, J., 2006c. Toward a census of bacteria in soil. *PLoS Comput Biol*, **2**(7):e92.

[Schloss et al., 2004] Schloss, P. D., Larget, B. R., and Handelsman, J., 2004. Integration of microbial ecology and statistics: a test to compare gene libraries. *Appl Environ Microbiol*, **70**:5485–92.

[Schloss et al., 2009] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., *et al.*, 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities$^{\nabla}$. *Appl Environ Microbiol*, **75**(23):7537–41,.

[Schwartz et al., 2003] Schwartz, S., Kent, W., Smit, A., Zhang, Z., Baertsch, R., *et al.*, 2003. Human-mouse alignments with BLASTZ. *Genome Res.*, **13**:103–107.

[Seshadri et al., 2007] Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., and Frazier, M., *et al.*, 2007. CAMERA: A community resource for metagenomics. *PLoS Biology*, **5**(3).

[Shaffer, 1995] Shaffer, J. P., 1995. Multiple hypothesis testing. *Annu Rev Psychol*, **46**:561–584.

[Shaw, 2007] Shaw, J., 2007. The undiscovered planet: Microbial science illuminates a world of astounding diversity. *Harvard Magazine*, **Nov-Dec**.

[Shendure and Ji, 2008] Shendure, J. and Ji, H., 2008. Next-generation DNA sequencing. *Nat Biotechnol*, **26**(10):1135–45.

[Shendure et al., 2005] Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., *et al.*, 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**(5741):1728–32.

[Simon et al., 2009] Simon, C., Wiezer, A., Strittmatter, A. W., and Daniel, R., 2009. Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Appl Environ Microbiol*, **75**(23):7519–26.

[Singleton et al., 2001] Singleton, D. R., Furlong, M. A., Rathbun, S. L., and Whitman, W. B., 2001. Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. *Appl Environ Microbiol*, **67**:4374–4376.

[Strous et al., 2006] Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., *et al.*, 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*, **440**(7085):790–794.

[Tatusov et al., 2003] Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., *et al.*, 2003. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**:41.

[Tatusov et al., 1997] Tatusov, R. L., Koonin, E. V., and Lipman, D. J., 1997. A genomic perspective on protein families. *Science*, **278**(5338):631–637.

[Teeling et al., 2004] Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F. O., *et al.*, 2004. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, **5**:163.

[Thompson and Klose, 2005] Thompson, F. and Klose, K., 2005. Vibrio: The first international conference on the biology of vibrios. *J Bacteriol*, **188**:4592–6.

[Tringe et al., 2005] Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., *et al.*, 2005. Comparative metagenomics of microbial communities. *Science*, **308**:554–557.

[Turcatti et al., 2008] Turcatti, G., Romieu, A., Fedurco, M., and Tairi, A., 2008. A new class of cleavable uorescent nucleotides: synthesis and optimization as reversible terminators for dna sequencing by synthesis. *Nucleic Acids Res*, **4**(e25).

[Turnbaugh et al., 2007] Turnbaugh, P., Ley, R., Hamady, M., Fraser-Liggett, C., Knight, R., *et al.*, 2007. The human microbiome project. *Nature*, **449**(7164):804–810.

[Turnbaugh et al., 2008] Turnbaugh, P. J., Backhed, F., Fulton, L., and Gordon, J. I., 2008. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe*, **3**(4):213–223.

[Turnbaugh et al., 2006] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., *et al.*, 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**(7122):1027–31.

[Tyson et al., 2004] Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., *et al.*, 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**:37–43.

[Urich et al., 2008] Urich, T., Lanzen, A., Qi, J., Huson, D. H., Schleper, C., *et al.*, 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One*, **3**:e2527.

[van Elsas et al., 2008] van Elsas, J. D., Costa, R., Jansson, J., Sjoling, S., Bailey, M., *et al.*, 2008. The metagenomics of disease-suppressive soils - experiences from the metacontrol project. *Trends Biotechnol*, **26**(11):591–601.

[Venter et al., 2004] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., *et al.*, 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**(5667):66–74.

[Vogel et al., 2009] Vogel, T. M., Simonet, P., Jansson, J. K., Hirsch, P. R., Tiedje, J. M., *et al.*, 2009. Terragenome: a consortium for the sequencing of a soil metagenome. *Nat Rev Microbiol*, **7**(4):252–252.

[von Mering et al., 2007] von Mering, C., Hugenholtz, P., Raes, J., Tringe, S. G., Doerks, T., *et al.*, 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**(5815):1126–30.

[Warnecke et al., 2007] Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., *et al.*, 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, **450**(7169):560–5.

[Watson and Crick, 1953a] Watson, J. and Crick, F., 1953a. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**:964–967.

[Watson and Crick, 1953b] Watson, J. and Crick, F., 1953b. A structure for deoxyribose nucleic acid. *Nature*, **171**:737–738.

[Wells et al., 2007] Wells, C. L., Johnson, M.-A., Henry-Stanley, M. J., and Bendel, C. M., 2007. Candida glabrata colonizes but does not often disseminate from the mouse caecum. *J Med Microbiol*, **56**(Pt 5):688–693.

[White et al., 2009] White, J., Nagarajan, N., and Pop, M., 2009. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol*, **5**(4):e1000352.

[Whiteman, 2008] Whiteman, L., 2008. Microbes to people: Without us, you're nothing. *National Science Foundation*, **8/04/08**.

[Whitman et al., 1998] Whitman, W. B., Coleman, D. C., and Wiebe, W. J., 1998. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A*, **95**(12):6578–83.

[Williamson et al., 2008] Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., *et al.*, 2008. The sorcerer ii global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One*, **3**(1):523–537.

[Woyke et al., 2006] Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., *et al.*, 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, **443**(7114):950–955.

[Woyke et al., 2009] Woyke, T., Xie, G., Copeland, A., Gonzalez, J. M., Han, C., *et al.*, 2009. Assembling the marine metagenome, one cell at a time. *PLoS One*, **4**(4):e5299.

[Wu and Eisen, 2008] Wu, M. and Eisen, J. A., 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*, **9**(10):R151.

[Yooseph et al., 2007] Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S., *et al.*, 2007. The Sorcerer II Global Ocean Sampling expedition: Expanding the Universe of Protein Families. *PLoS Biol*, **5**(3):e16.

List of Figures

List of Tables

Index

Curriculum Vitae

Suparna Mitra

Tübingen University
Algorithms in Bioinformatics
ZBIT Center for Bioinformatics
Sand 14, 72076 Tübingen
Germany
Date of Birth: 25/11/1980
Citizenship: India

Phone: ++49-7071-29 70453 (O)
Fax: ++49-7071-29 5148 (O)
Email: mitra@informatik.uni-tuebingen.de
or suparna.mitra@gmail.com

Education and Research Experience

1986 - 1996	- School education in Burdwan Municipal Girls' High School, Burdwan, India.
12/1997	- Secondary Examination. *Obtained grade:* **1st division**.
1997 - 1999	- Higher Secondary education in Burdwan Municipal Girls' High School, Burdwan, India
10/1999	- Higher Secondary Examination. *Obtained grade:* **1st division**.
2000 - 2003	- Bachelor degree course in Vivekananda Mahavidyalaya, Burdwan, India; *Honors:* **Mathematics**; *Minors:* **Physics** and **Chemistry**.
07/2003	- Degree: **B. Sc. Mathematics**, *Obtained grade:* **2nd class**.
2003 - 2005	- Master degree course in Burdwan University, India. *Project title:* "A Simple Approach to Phylogeny 'R' Code for Molecular Data Analysis".
10/2005	- Degree: **M. Sc. Statistics**, *Obtained grade:* **1st class**.

02/2006 - 03/2007	- Worked as Research Assistant in a project "Influence of parasite infection on wild life population" at *Institute for Mathematical Stochastic*, University of Karlsruhe, Germany.
04/2006 - 06/2007	- Worked as a Research Staff in *Genetic Epidemiology* department, University of Ulm, Germany.
from 07/2007	- Pursuing **Ph.D.** and working as Research and Teaching Staff, at *Algorithms in Bioinformatics*, ZBIT Center for Bioinformatics, University of Tübingen; under Prof. Dr. Daniel Huson.
11/2010	- Degree: **Ph.D. Bioinformatics**, University of Tübingen, Germany, *Obtained grade:* **Outstanding (Summa cum laude)**.

Academic Teaching Experience

SS 2007	Practical Course: "Bioinformatics Software Tools" for Master and Diploma students.
WS 2007/08	Seminar: "Genomics and Methagenomics" for Master and Diploma students.
SS 2008	Seminar: "Genomics and Methagenomics" for Bachelor Bioinformatic students.
WS 2008/09	Seminar: "Metagenomics" for MSc Bioinformatic students.
SS 2009	Seminar: "Genomics and Methagenomics" for Bachelor Bioinformatic students.
WS 2009/10	Seminar: "Metagenomics" for BSc and MSc Bioinformatic students.
SS 2010	Proseminar: "Grundlagen der Bioinformatik" for BSc students and Seminar: "Sequence Analysis" for MSc students.

Supervised Bachelor/Master Thesis:

2008	Localisation of reads in a specific genome; *Student: Joerg Vetter*
2009	Simulation study of metagenome analysis methods; *Student: Max Schubach*
2010	Functional analysis of metagenome samples with KEGG orthologies; *Student: Paul Rupek*
2010	Metagenome analysis with 16S rRNA sequences; *Student: Mario Stärk*
2010	Comparing different paired-end protocols with 16S rRNA metagenome samples; *Student: Hannelore Clément*