Universität Bielefeld Technische Fakultät

Taxonomic Classification of Metagenomic Sequences

Ph. D. Thesis

submitted to the
Faculty of Technology, Bielefeld University, Germany
for the degree of Dr. rer. nat.

by

Wolfgang Gerlach

February 2012

Supervisor: Prof. Dr. Jens Stoye

Referees: Prof. Dr. Robert Giegerich
Prof. Dr. Jens Stoye

Gedruckt auf alterungsbeständigem Papier nach DIN-ISO 9706.
Printed on non-aging paper according to DIN-ISO 9706.

Zusammenfassung

Bakterien, Archaeen und eukaryotische Mikroorganismen sind in fast jedem Habitat auf der Erde zu finden, insbesondere im Erdreich, in Sedimenten und in Gewässern. Typischerweise leben sie in komplexen Gemeinschaften mit verschiedenen Formen von symbiotischen Assoziationen, insbesondere in Beziehungen mit größeren Organismen wie Tieren und Pflanzen.

Die große Mehrheit solcher Mikroorganismen ist nicht kultivierbar und kann daher nicht mit traditionellen Methoden sequenziert werden. Die relativ junge Disziplin der Metagenomik bietet verschiedene *in vivo*- und *in silico*-Werkzeuge um dieses Hindernis zu überwinden. Insbesondere Hochdurchsatz-Sequenziertechnologien, wie 454 oder Solexa-Illumina, ermöglichen es, solche Mikroorganismen zu untersuchen, indem komplette natürliche mikrobielle Gemeinschaften einschließlich ihrer biologischen Diversität, sowie der zugrundeliegenden metabolischen Pfade analysiert werden. Eine gegenwärtige Beschränkung solcher Technologien ist, daß nur DNA Fragmente begrenzter Länge sequenziert werden können. Ein zusätzliches Problem stellt dar, daß die sequenzierten Fragmente einer mikrobiellen Gemeinschaft nicht ohne weiteres ihren jeweiligen Spezies aus der Gemeinschaft zugeordnet werden können.

In den letzten Jahren wurden verschiedene Methoden entwickelt, deren Ziel es ist, einzelne metagenomische Sequenzen sowohl taxonomisch als auch funktionell zu klassifizieren. Trotzdem stellen insbesondere die taxonomische Klassifikation metagenomischer Sequenzen, die von bisher unbekannten Spezies stammen und für die auch keine Sequenzen von anderen nahe verwandten Spezies in den biologischen Sequenzdatenbanken vorliegen, eine besondere Herausforderung dar. In solchen Fällen machen die bisher existierenden Methoden auf den niedrigeren taxonomischen Ebenen viele falsche Vorhersagen.

In dieser Doktorarbeit präsentieren wir CARMA3, eine neue Methode zur taxonomischen Klassifikation von assemblierten und unassemblierten metagenomischen Sequenzen, die sowohl BLAST, als auch HMMER3, verwenden kann. CARMA3 akzeptiert Protein-kodierende DNA Sequenzen, Protein Sequenzen und 16S-rDNA Sequenzen als Eingabe. Zusätzlich stellen wir WebCARMA vor, eine Web-Anwendung für die Analyse von Protein-kodierender DNA mit CARMA3, die eine lokale Installation von CARMA3 unnötig macht.

Wir evaluieren unsere Methode in verschiedenen Experimenten mit simulierten und echten Metagenomen und zeigen daß CARMA3 weniger falsche taxonomische Vorhersagen macht (bei gleicher Sensitivität) als andere BLAST-basierte Methoden. In unserem letzten Experiment zeigen wir, daß auch sehr kurze DNA Fragmente benutzt werden können – zumindest prinzipiell – um die taxonomische Zusammensetzung eines Metagenoms zu bestimmen.

Abstract

Bacteria, archaea and microeukaryotes can be found in almost every habitat present in nature, in particular in soil, sediments and sea water. They typically live in complex communities with different kinds of symbiotic associations which include relationships with larger organisms like animals or plants. Examples are microbial communities in the gut or on the skin of animals and humans, or bacteria that live in symbiosis with plants.

The vast majority of such microbes are unculturable and thus cannot be sequenced by means of traditional methods. The recently upcoming discipline of metagenomics provides various *in vivo*- and *in silico*-tools to overcome this limitation. In particular, high-throughput sequencing techniques like 454 or Solexa-Illumina make it possible to explore those microbes by studying whole natural microbial communities and analysing their biological diversity as well as the underlying metabolic pathways. A current limitation of theses technologies is that they can sequence only DNA fragments of a limited length. With this limitation it is usually not possible to recover complete microbial genomes. In addition, the DNA fragments are drawn randomly from the microbial communities and the exact species of origin is unknown.

Over the past few years, different methods have been developed for the taxonomic and functional characterization of metagenomic shotgun sequences. However, the taxonomic classification of metagenomic sequences from novel species without close homologues in the biological sequence databases poses a challenge due to the high number of wrong taxonomic predictions on lower taxonomic ranks.

In this thesis we present CARMA3, a novel method for the taxonomic classification of assembled and unassembled metagenomic sequences that has been adapted to work with both BLAST and HMMER3 homology searches. CARMA3 accepts protein-encoding DNA sequences, protein sequences, and 16S-rDNA sequences as input. In addition, we present WebCARMA, a web application for the analysis of protein-encoding DNA sequences with CARMA3 without the need for a local installation.

We evaluate our novel method in different experiments using simulated and real shotgun metagenomes and show that CARMA3 makes fewer wrong taxonomic predictions (at the same sensitivity) than other BLAST-based methods. In the last experiment we show that also very short reads can, in principle, be used to describe the taxonomic content of a metagenome.

Contents

Chapter 1

Introduction

When life on Earth arose about 3.5 billion years ago, it solely consisted of microbial life [172]. Still today, microbial life dominates Earth in many aspects. With an estimated population of 5×10^{30}, prokaryotes are the most numerous organisms on Earth and constitute a huge diversity [218]. They have been estimated to comprise 10^6 to 10^8 separate genospecies [177]. The total carbon of prokaryotes constitutes about $60 - 100\%$ of the total carbon found in plants [211]. Bacteria, archaea and microeukaryotes can be found in almost every habitat present in nature, in particular in soil, sediments and sea water. Microbes can also be found in rather hostile environments like the Arctic [18], deserts [48, 81], hot springs [141] and in rocks as much as 7 kilometers below the Earth's surface [192]. They typically live in complex communities with different kinds of symbiotic associations which include relationships with larger organisms like animals or plants. Examples are microbial communities in the gut and rumen or on the skin of animals, or bacteria that live in symbiosis with plants. Figure 1.1 depicts a cow that has been surgically modified such that researchers have direct access to the cow rumen and its microbial community.

Microbes are important for us as they are involved in the distribution and cycling of nutrients in the ecosystem, degradation of different compounds, and they also have substantial impact on global climate. Furthermore, they account for the human micro flora — the human body contains about 100 times more bacterial cells than human cells [14, 165]. Considering the huge influence of microbes on the human body, it is clear that a more thorough understanding of the human condition also requires understanding of the diversity and functions of the microbes in our body [132, 200, 218]. The MetaHIT project [144] and the Human Microbiome Project [199] are examples of recent efforts to improve knowledge about the human micro flora. A prominent example for the importance of the human gut microbiome to human health is the association of large-scale alterations in the phylogenetic composition of gut microbiota with obesity [100].

Understanding microbes at the biochemical and genomic level is important as this will improve our ability to use microbes and their genetic potential to produce useful materials and products. In fact, we have been using microbes for more than 5000 years for food preservation and to enrich our diet. Beverages like beer and wine, cheese, bread and a variety of

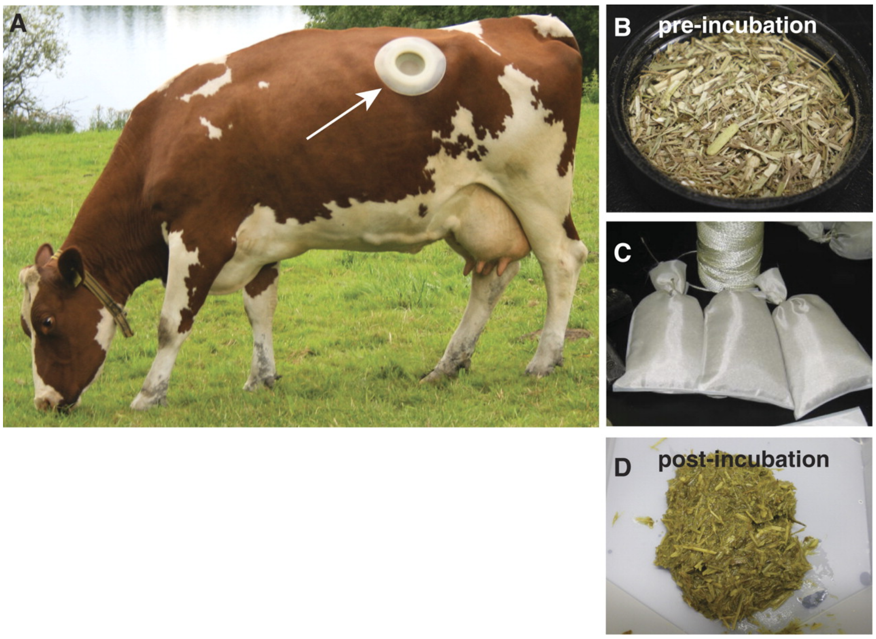


Figure 1.1: A fistulated cow whose rumen microbiome has been searched for biomass-degrading genes and genomes. From [70]. Reprinted with permission from ASSS.

other fermented foods form a significant proportion — about one third — of human food consumption [23]. Whereas in ancient times microbes were used unknowingly, today microbes are actively used and modified to produce, for example, food flavoring and preservative agents [58]. But microbes are also useful in other fields. Microbial biotechnology makes it possible to use microbes and their enzymes for the production of chemicals for industrial and pharmaceutical applications like vaccines, antibiotics and other health-care products for medical purposes [36]. Another application of microbes is the recovery of metals from certain types of copper, uranium, and gold-bearing minerals and the recycling of metals from industrial waste to avoid excessive and environmentally harmful mining practices [21, 148]. A somewhat related application is the removal of toxic metals or organic pollutants from soil or water with the help of microbes [28, 52].

Dwindling natural resources on Earth, in particular of fossil fuels, pose momentous challenges for humanity. At the same time, the world population is growing and developing countries demand higher living standards involving further increase in consumption of energy and natural resources [203]. Environmental problems related to the usage of fossil fuels, for example greenhouse gas emissions, are another argument to search for alternative sustainable energy sources. Among such alternatives are many technologies that involve the usage of microbes. Complex microbial communities are used to produce methane by anaerobic degradation of biomass, typically either high-energy biomass that has been grown for this purpose or agricultural biomass waste [169]. Microbes are also used to produce ethanol from corn glucose or sucrose by fermentation [102]. A promising technology, albeit still in its infancy, includes microalgae to produce fuel by directly converting sunlight into hydrogen [59, 120].

Knowledge about how microbes function is an essential prerequisite to improve efficiency of such microbe-dependent technologies. The development of efficient technologies is of high importance in order to meet the economic and ecological challenges of our society. Nevertheless, the study of microbes is hampered by the problem that most microbes in nature live in complex microbial communities and therefore are not accessible by means of traditional culturing methods. The recently upcoming discipline of metagenomics provides *in vivo*- and *in silico*-tools to overcome this limitation. In particular, the development of new sequencing technologies has provided the possibility to gather genomic information from microbial communities. A current limitation of theses technologies is that they can sequence only DNA fragments of a limited length. With this limitation it is usually not possible to recover complete microbial genomes. In addition, the DNA fragments are drawn randomly from the microbial communities and the exact species of origin is unknown. Therefore, one of the problems in metagenomics is to determine the species of origin and the function of a DNA fragment. These questions can be answered, at least to some extent, using computational methods for the taxonomic and functional classification of DNA fragments.

1.1 DNA

Prior to the advent of modern analysis techniques in the second half of the 19th century, the scientific study of living organisms was mainly restricted to their phenotype, observable characteristics like morphology or behavior. A deeper understanding and characterization of a living organism can be obtained by knowledge of its genotype, the total genetic information of the organism. The genotype is the genetic blueprint of an organism and thus mainly determines its phenotype. The genetic information is inherited from one or two parents, but smaller amounts can also be obtained through horizontal gene transfer from other organisms. Most of the information that accounts for the genotype is encoded at the molecular level by the *deoxyribonucleic acid* (DNA), a double helix of two polymer strands. A schematic overview of the DNA is depicted in Figure 1.2. Each of the strands consists of sugars and phosphate groups that serve as backbones. Each sugar has one of four possible bases attached to it, either one of the two purine bases adenine (A) and guanine (G), or one of the two pyrimidine bases thymine (T) and cytosine (C). The two strands of a double helix are connected by hydrogen bonds between the complementary bases: A pairs with T and G pairs with C. The order of these bases in a strand defines the DNA sequence. As bases in a strand can be read in two directions, either right-to-left or left-to-right, it is convention to read the bases in 5'-to-3' order, a convention based on the labeling of carbon atoms in the sugar group of the backbone of the strand. Due to the complementarity in this base-pairing, the two strands are anti parallel, i.e., the DNA sequence of one strand is the same as the reverse complement of the DNA sequence of the other strand. The total DNA in a living cell constitutes the *genome* which is partitioned into one or several long DNA molecules, the *chromosomes*. Archaeal and bacterial cells typically contain one circular chromosome, whereas eukaryotic cells contain several linear chromosomes [3]. In addition to the DNA sequence there are also epigenetic factors that can be inheritable and do influence the phenotype, for example methylation of DNA [209] or paramutation [25].

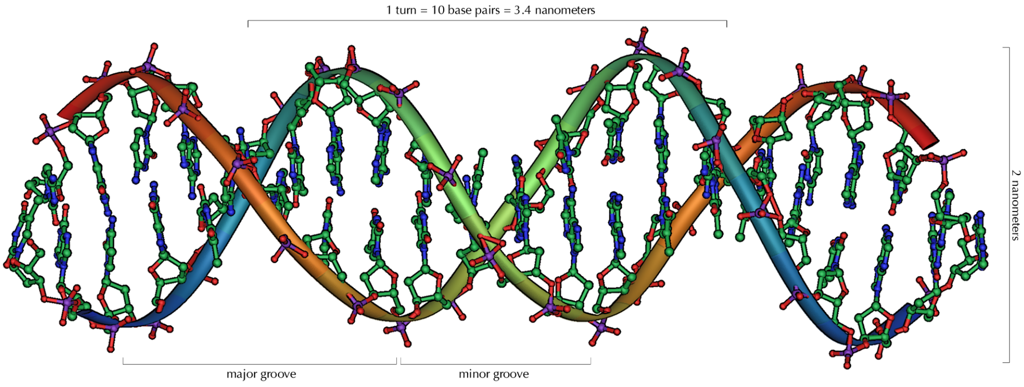


Figure 1.2: DNA Overview, Source: [190]

DNA Sequencing

One of the most important steps in exploring DNA was the detection of the double-helix nature of DNA by Watson and Crick in 1953 [210]. In the following years, different methods have been developed with the goal to determine the exact sequence of nucleotides in a given piece of DNA [121]. The DNA sequencing method developed by Sanger in 1977 [164] (also called *Sanger sequencing*) was the most commonly used method for many years.

Sequencing machines can sequence DNA fragments only up to a certain length, e.g., Sanger sequencing achieves a length of 800–1000 bp (base pairs). These DNA sequences are called *reads*. In recent years, newer high-throughput sequencing (also next- or second-generation sequencing) technologies such as Roche's 454 or Illumina's Genome Analyzer have been developed, which produce much more data at lower cost than traditional Sanger sequencing. A drawback of these technologies is that they produce rather short reads (35 bp–400 bp) compared to Sanger technology [110, 212].

The newest step in the evolution of DNA sequencers are the third-generation sequencing platforms. Most of these technologies are still under development and promise features including "single-molecule templates, lower cost per base, easy sample preparation, significantly faster run times and simplified primary data analysis", while at the same time having the potential to overcome the short read lengths of second-generation technologies [125]. The exponential increase of data produced by the various DNA sequencing platforms in the last 30 years, depicted in Figure 1.3, has provided the capability to study new genomic aspects of microbial and multi-cellular life.

Polymerase Chain Reaction

Another technique for the analysis of DNA is the polymerase chain reaction (PCR) [10]. It was developed in 1983 to amplify specific pieces of DNA. The method is based on the use of primers, which serve as anchors for amplification. Primers are short synthetic oligonucleotides designed such that they show complementarity to specific regions up-stream and down-stream of the piece of DNA to be amplified.

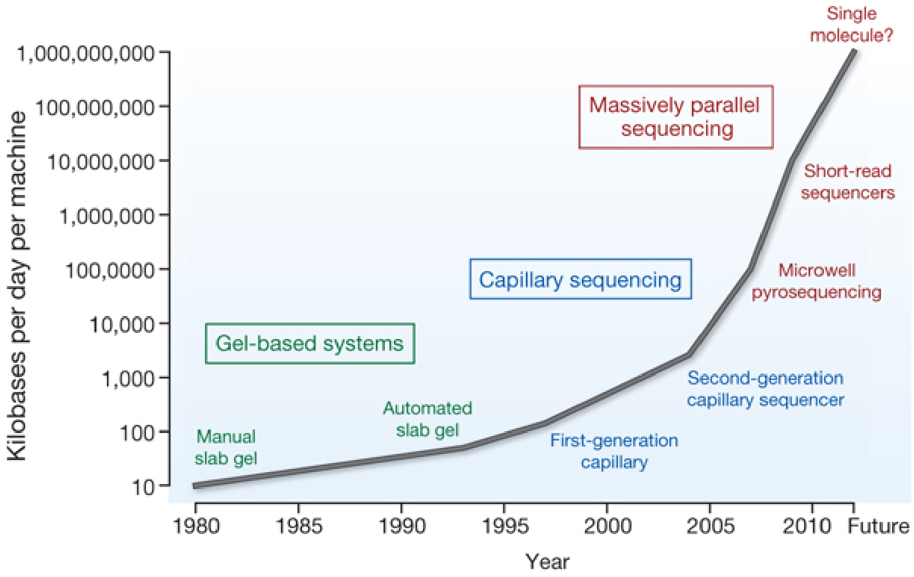


Figure 1.3: The rate of DNA sequencing over the past 30 years and into the future. Reprinted by permission from Macmillan Publishers Ltd: Nature 458, 719–724, copyright (2009) [188]

PCR starts with denaturation of the DNA at high temperatures such that complementary strands of the DNA separate. Then, at lower temperatures, the primers bind to their specific regions on the strands. In the next step, the annealed primers serve as starting points for the *Taq polymerase*, a heat-stable DNA polymerase, which successively fills-up complementary nucleotides in order to synthesize the complementary strand. Repeating these two steps several times thus allows to create million copies of a single DNA double strand [88].

Applications of PCR in metagenomics involve amplification of marker genes for the phylogenetic characterization of metagenomes as described in Section 1.2.1 and screening of metagenomes for new genes from known gene families as discussed in Section 1.2.4.

1.2 Metagenomics

In traditional genomics, sequencing new microbial genomes requires the cultivation of microbes in a monoculture. It has been shown that only a very small fraction of the microbes in an environment can be grown in a culture, and therefore most microbes are not accessible by means of complete genome recovery [5]. Although new cultivation techniques for microbes that were believed to be unculturable have been developed [79, 147], for the vast majority of microbes it is still unknown how to cultivate them. As the term *unculturable microbes* is misleading, we will refer to them as *uncultured microbes*.

Metagenomics, or *environmental genomics*, is a new field of research on natural microbial communities containing uncultured microbes. Culture independent methods are used to obtain information about the genetic diversity, population structure, and ecological roles of members of the communities. These methods complement or even replace culture-based ap-

Table 1.1: Milestones of (Meta-)genomics

Year	Milestone	Ref.
1977	Sanger *et al.* sequence bacteriophage Phi X 174	[163]
1977	Woese and Fox assess evolutionary relationships of organisms by studying the 16S and 18S RNA	[216]
1983	Mullis introduces the polymerase chain reaction	[29]
1984	Stahl *et al.* sequence clones from a 5S rRNA cDNA library from a symbiontic community within the tube worm *Riftia pachyptila*	[182]
1985	Lane *et al.* describe a PCR protocol with universally applicable primers to access 16S-rRNA sequences for phylogenetic characterizations without isolation of the 16S-rRNA or cloning of its gene	[96]
1985	Pace *et al.* suggest the concept of cloning DNA directly from the environment	[133]
1987	Woese proposes a 16S-rRNA-based phylogeny	[215]
1991	Schmidt *et al.* isolate and clone bulk DNA from seawater using λ phages and screen for 16S-rRNA genes	[170]
1992	Introduction of BAC and Fosmid cloning vectors	
1995	Fleischmann *et al.* sequence a bacterial genome	[50]
1996	Stein *et al.* sequence and reconstruct a 40 kb long fragment from an uncultured marine archaeon using a fosmid library	[185]
1998	Handelsmann *et al.* coin the term “metagenome”	[63]
2000	Béjà *et al.* construct first BAC library from marine environment and sequence one BAC insert of size 60 kb	[22]
2000	Rondon *et al.* clone the soil metagenome using BACs	[155]
2002	Breitbart *et al.* clone and sequence two uncultured marine viral communities	[20]
2004	Venter *et al.* clone and sequence the metagenome of the Sargasso Sea in large scale	[205]
2005	Margulies *et al.* introduce 454-sequencing technology	[112]
2005	Uchiyama *et al.* introduce substrate-induced gene expression screening (SIGEX)	[202]
2006	Edwards *et al.* use 454-technology to sequence a deep mine microbial community without cloning	[44]

proaches and bypass some of their limitations [177]. The term *metagenome* was coined in 1998 by Handelsmann *et al.* [63] in the context of soil as a microbial habitat and was defined as “the collective genomes of soil microflora”. The term *metagenome* is also being used to denote the *in silico* representation of a metagenome which usually is an incomplete representation of the actual metagenome. Due to technical limitations of currently available high-throughput sequencing (HTS) technologies, the sequences in an *in silico* metagenome represent only short randomly drawn fragments which also do not necessarily cover all genomes in the metagenome. With the ongoing development of sequencing technologies, this discrepancy will most likely become smaller.

A new perspective on microbial diversity was provided in 1977 by Carl Woese [216] who used 16S and 18S ribosomal RNA to assess the evolutionary relationships of different organisms. Based on this technique, he proposed the archaea as a separate group of prokaryotes and introduced a separation of life into three domains: Archaea, bacteria and eukaryotes. This was a departure from previous taxonomies that were based on the analysis and comparison of phenotypic characteristics of organisms. Although highly controversial in the beginning, nowadays the three-domain system of Carl Woese is widely accepted.

By suggesting the concept of cloning DNA directly from the environment in 1985, Pace *et al.* [133] paved the way for the exploration of the diversity of uncultured microbes. Schmidt *et al.* realized this concept six years later by isolating and subsequently cloning bulk DNA from seawater, using a λ phage library [170]. In such a clone library, each clone carries a piece of environmental DNA. This transfer of DNA pieces from uncultured microbes into culturable hosts made previously unaccessible DNA available for further analyses. Different techniques can be used to screen the clones for functionally active environmental genes or other sequences of interest, for example phylogenetic markers.

Although the usage of environmental clone libraries was the crucial step towards metagenomics, construction and screening of clone libraries remained laborious. Furthermore, the clone libraries are known to exhibit significant bias [46, 195]. The introduction of HTS technologies like Roche's 454-sequencing, ABI's SOLiD or Illumina's Genome Analyzer has allowed for sequencing metagenomic samples without a prior cloning step. The first sequencing technology being used for shotgun sequencing of a metagenomic sample, similar to the sequencing of whole genomes, was 454-sequencing. One of the first metagenomic samples being sequenced was a deep mine microbial community in 2006 [44]. An overview of the milestones of (meta-)genomics is given in Table 1.1.

Figure 1.4 depicts a schematic overview of possible steps in a metagenomic workflow. Preparation steps, necessary for the extraction of DNA from a metagenomic sample, depend on the sample and can require, for example, filtering steps for seawater or sieving steps for soil. Given the extracted and purified DNA, three possible approaches are common: (a) amplification with PCR, typically using 16S primers, followed by sequencing of the amplified DNA, (b) the construction of a metagenomic library, or (c) direct shotgun sequencing. A more detailed description for each of these variants is given in the following Sections 1.2.1 to 1.2.5.

1.2.1 16S-based PCR Amplification

For accurate assessment of the population structure of a microbial ecosystem, different gene markers can be used. The most common and established marker is the 16S-rDNA in prokaryotes, but mitochondrial and chloroplastic rRNA are also used [204, 123]. The 16S-rDNA is a gene that encodes for the 16S-rRNA, a component of prokaryotic ribosomes. Ribosomes are responsible for the translation of genes into proteins. They consist of different RNAs and proteins, but the key catalytic activity is provided by the RNAs [24]. Due to their essential function in all living cells, ribosomal RNAs are highly conserved. Carl Woese found that

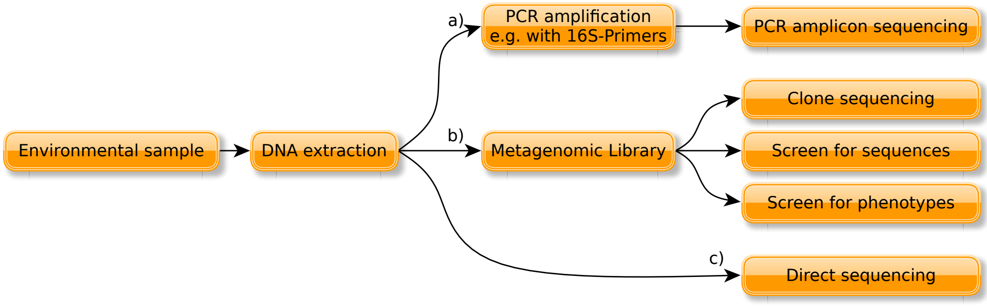


Figure 1.4: Possible steps in a metagenomic workflow.

in particular the small subunit 16S ribosomal RNA is suited to serve as a genetic marker in prokaryotes [216]. The 16S-rRNA consists of regions that are highly conserved between different bacterial and archaeal species, and regions that are highly variable. The former can be used as anchors for the detection of the 16S-rDNA using universal primers in new phylogenetically remote sequences. The latter, the highly variable regions, can provide species-specific signature sequences.

16S-rDNA sequences can be obtained by sequencing of clones from a metagenomic library, yet this approach requires significant manual labour, and only a few population constituents can be phylogenetically characterized. Using next-generation sequencing technologies and primer directed PCR amplification, it is possible to obtain 16S-rDNA sequences more directly. By focusing the sequence coverage on 16S-rDNA, especially the V6 hyper-variable region [75, 77] — or theoretically any other suited genetic marker — it is possible to get a much more detailed view of the phylogenetic diversity and low abundant species of highly complex natural communities [179]. Another possibility is the random shotgun sequencing of a metagenomic sample with next-generation sequencing technologies and successive screening for fragments that encode for 16S-rDNA, but the yield of 16S-rDNA can be below 0.5% in this approach [94].

For taxonomic classification of complete and partial 16S-rDNA sequences, different methods are available. For an overview see Liu *et al.* [106]. One of the most commonly used methods is the RDP classifier, a naïve Bayesian classification method [208]. It uses a feature space consisting of all possible RNA sequences (words) of length eight. Words occurring in the query sequence of unknown taxonomic origin are used to compute the probability of the query sequence to be a member of a group of reference sequences from a certain taxonomic clade. Within each taxonomic rank, the query is assigned to the clade that gives the highest probability score.

The taxonomic classification of 16S-rDNA sequences is used for assessment of the microbial diversity, evenness, and community structure. Rarefaction curves can be used to estimate the completeness of a sample and the species richness of a microbial community [168]. For example, Sogin *et al.* [179] pyrosequenced 16S-rRNA PCR amplicons of different microbial communities from the North Atlantic and showed that the communties are one to two orders of magnitude more complex than previously reported for any microbial environment. It was observed that the microbial communities consist of small numbers of dominating populations and thousands of low-abundance populations that account for most of the phylogenetic diversity, therefore termed by the authors as the “rare biosphere”. Subsequently, however, there has been a controversy about the impact of sequencing errors that might have introduced an overestimation of the species richness [95, 145].

Despite the known problems of primer bias [9, 76] and horizontal gene transfer [2], 16S-rDNA-based metagenomics is still the method of choice for analyses with a sole focus on taxonomic composition. On the other side, information about the functional potentials of a microbial community can only be inferred by the assumption that species in the community that are similar to already known (cultured) species also share most of their metabolic functions [74, 134, 170].

1.2.2 Metagenomic Libraries

To analyse and sequence a DNA fragment, it needs to be isolated and amplified. This amplification step is also called *cloning*, either performed *in vitro* using PCR, or *in vivo* using living cells. For the construction of a metagenomic library, the cell-based DNA cloning is used which takes advantage of the natural propagation via cell division of unicellular organisms like bacteria. A set of clones that carry a foreign DNA fragment is called a *clone library*. A clone library that carries environmental DNA is called a *metagenomic library*, but other names like *environmental DNA library*, *zoolibrary*, *soil library*, and many other names are also used [153].

The cloning strategy consists of four steps. The first step is the construction of recombinant DNA molecules by first cutting the DNA fragment and a replicon — a sequence where the replication of DNA is initiated — with specific restriction endonucleases. Then, as the second step, the DNA fragment and the replicon are ligated together using the enzyme DNA ligase. Because the replicon serves as a carrier of the DNA fragment, it is also called *vector*. Common vectors are bacteriophages, cosmids or bacterial artificial chromosomes (BACs). The latter two allow the cloning of rather long fragments, 35-40 kb for cosmids and up to 200 kb for BACs [76]. In the third step these recombinant molecules are transferred into the surrogate host cells, typically into *Escherichia coli* via electroporation [68]. The transformed cells are plated out by spreading over the surface of nutrient agar in a petri dish. Colonies grow consisting of clones that are all identical to an ancestral single cell. Individual colonies are picked from the plate for subsequent growth in liquid culture. In the last step, the transformed and enriched cells are lysed and their DNA is extracted and purified. The differences between the recombinant DNA and the host chromosomal DNA allow to distinguish between both, and the recombinant DNA can finally be recovered [187].

Depending on the kind of habitat, the DNA extraction process often involves different mechanical filtering steps. Seawater can be filtered such that bigger eukaryotic inhabitants and smaller viral particles are discarded. This ensures that in the following DNA extraction step mainly microbial DNA of interest is obtained. Also sieving of soil is often performed to obtain enough DNA of the favored soil community [189].

The extraction of DNA from a natural environment can be a challenging task. For example, polyphenolic compounds from decaying plant material, which are difficult to remove, often contaminate the purified DNA from soil [189]. Another issue is the potential bias of the DNA extraction protocol. It has been shown that different methods for the extraction lead to different community compositions [113].

If a certain subpopulation with low abundance exhibits genes of interest, it is possible to perform a pre-enrichment of the metagenomic sample, such that the metagenomic library finally contains more of these genes. Different enrichment methods are reviewed for example in [30].

Although the construction of a metagenomic library is quite laborious and only few sequences can be obtained compared to recent approaches of direct sequencing with HTS technologies like 454 or Illumina (Section 1.2.5), using metagenomic libraries has some advantages. The obtainable sequence length is not limited by the read length since the metagenomic fragments in the library can be sequenced with the shotgun [6, 181] or chromosome-walking [27] approach. Therefore, metagenomic libraries allow to sequence complete genes. Still, large gene clusters or operons cannot be captured by this approach.

Besides sequencing of fragments from a metagenomic library, the search for interesting genes and functions within such a library using highly automated function- and sequence-based screening techniques are other applications of metagenomic libraries, which are also described in the following subsections.

1.2.3 Function-Based Screening

A metagenomic library consists of host organisms that hold foreign metagenomic DNA fragments. If a fragment contains a complete gene, it is possible that the gene is expressed. Functional metagenomics takes advantage of this *heterologous expression* of metagenomic genes. This allows to find completely new classes of genes in a metagenomic community, also if the new genes exhibit no sequence similarity to any other previously known gene sequence [66, 149]. To find such new genes, individual clones from a metagenomic library are searched for certain enzymes or other bioactivities [83].

The detection of novel gene products not only deepens our understanding of the biology and biochemistry of the microbial community within its habitat, but novel gene products like biocatalysts can also improve or enable applications in different industrial applications [83, 107]. Examples for the discovery of new gene products as a result of heterologous expression in metagenomic libraries are antibiotics [56, 109] and antibiotic resistance genes [152].

Typically, only a few clones out of thousands in a library express the gene product that is searched for [62]. To also find low abundant genes in a metagenome, sufficiently large libraries have to be screened, which is often done with highly automated high-throughput picking and pipetting robots. The function-based screening is also called enzyme-activity-based or phenotype-based screening [69, 189]. To detect the expression of a certain function in a clone, the following techniques are used.

(1) Heterologous complementation of host strains or mutants

In this approach, mutants are used as host organisms. These mutants do not grow under normal conditions, they require certain selective conditions. If a foreign gene compensates the inactive gene of the mutant, the mutant can also grow under normal conditions. An example are *E. coli* mutants with a cold-sensitive mutation in a domain of the DNA polymerase I, which is lethal at temperatures below 20 °C. By exposing a metagenomic library of mutant clones to temperatures below this temperature, only such clones survive and thus indicate a potentially novel gene which carry and express a metagenomic gene for the DNA polymerase I [178, 177].

(2) Direct detection of specific phenotypes of individual clones

Specific substrates or indicator dyes which can interact with the desired gene product are incorporated into the growth medium. Clones that express the gene product can then be detected in a screen due to a color change in the growth medium of the individual clones [62, 177]. An example using skim milk agar plate is depicted in Figure 1.5.

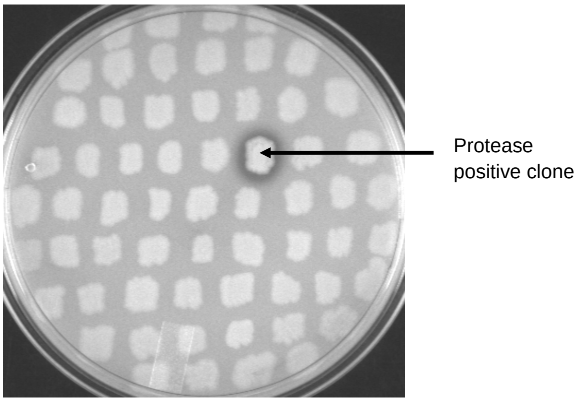


Figure 1.5: Functional screening of a metagenomic library for protease activity on a skim milk agar plate. The positive clone showing zone of clearance in skim milk agar plate is indicated by an arrow. Source: [143]

(3) Substrate- and metabolite-induced gene expression screening methods

Substrate-induced gene expression screening (SIGEX) is a method for the detection of catabolic genes. The expression of catabolic genes is usually induced by certain substrates or metabolites and is often controlled by regulatory elements (promoters) which are located close to the catabolic genes. For the SIGEX method, an operon-trap expression vector is used, which harbors the gene for a green fluorescent protein (gfp) but not a promoter. If a substrate is added to growth medium of the metagenomic library, all clones are expressed that carry a metagenomic catabolic gene with a promotor that is specific to the substrate. As the green fluorescent protein is also located on the vector, it is subsequently coexpressed. Finally, fluorescence activated cell sorting (FACS) is used to separate expressing from non-expressing clones [202]. Another method is metabolite-regulated expression (METREX) which is able to detect small molecules. It makes use of *quorum sensing*, a natural process used by many microbes for the cell-to-cell communication. Quorum sensing is mediated by small signal molecules and allows the cells to determine their own cell density. In addition to the genome and the vector for the metagenomic fragment, the host cells contain plasmids which serve as biosensors that contain genes for the quorum sensing and the green fluorescent protein. The *E. coli* host strain cells themselves do not produce detectable quorum-sensing inducing compounds. When a metagenomic gene in a clone is expressed and its gene product produces enough quorum-sensing inducing metabolites to reach the required quorum of the biosensor, the green fluorescent protein also becomes expressed and the clone can be detected [214].

To a certain extent, metagenomic libraries can also be used to obtain a linking of functions and phylogeny in metagenomes. Clones that express a metagenomic gene of interest can be sequenced and searched for flanking phylogenetic anchors like rRNA genes. Another alternative is the usage of a 16S-rRNA gene library as follows. After sequencing and taxonomically classifying the 16S-rRNA genes, it is possible to search in the sequences of a particular phylogenetic group for flanking genes that encode a function. In both cases, this approach has the disadvantage that phylogenetic anchors and genes that encode for a certain function are only occasionally located in proximity to each other [153, 155].

The detection of novel genes using functional metagenomics is often limited by the inability of the hosts to express the metagenomic genes. Different promoter regions or different codon usage keep the expression at a low level or may even inhibit the expression of the metagenomic gene. Even if a metagenomic gene is transcribed and translated correctly, its folding might be incorrect if required chaperones do not exist in the host cell. The same holds for cofactors which might be absent in the host cell or cannot be correctly incorporated into the final protein. Furthermore, if a function of interest is the result of a cascade of genes, it cannot be detected because of the limited length of inserts in metagenomic libraries. Another problem are metagenomic genes which are lethal for the host. These genes cannot be cloned. Approaches to improve the expression of foreign genes in metagenomic libraries can either be the enhancement of the capabilities of existing host strains like *E. coli* or the usage of new host strains that are better suited for the expression of metagenomic genes from other uncultured taxonomic clades [189].

1.2.4 Sequence-Based Screening

Sequence-based screening methods include sequencing of clones and different PCR- and hybridization-based methods. Sequencing of clones will be discussed separately in the next section. Both PCR- and hybridization-based methods use sequence similarity to known sequences to detect new sequences. Sequence-based screening has the advantage over functional metagenomics of not requiring heterologous expression. Primers and hybridization probes are designed such that they target conserved DNA regions, which will increase the chance to find new members of the same gene family in the metagenome. The disadvantage of these methods is that only members of known families can be detected [34, 177]. They are described in the following two paragraphs.

PCR-based methods The screening for 16S-rDNA genes using PCR, as already discussed in Section 1.2.1, is a special case of the sequence-based screening methods as it concentrates on the recovery of one specific marker gene and has become the standard method for the phylogenetic characterization of metagenomes. Besides this, PCR can also be used to discover function-encoding genes. Examples of newly discovered functional enzymes using PCR are chitinases from aquatic environments [98], alcohol oxidoreductases [87], and diol dehydratases [86], the latter two from various sample sites.

One of the disadvantages of the standard PCR method is that it usually amplifies only fragments of limited length. To access full-length genes from metagenomes, there exist several technically more involved variants of PCR. These include universal fast walking [126], panhandle PCR [118], random primed PCR [105], inverse PCR and adaptor ligation PCR [131], pre-amplification inverse-PCR (PAI-PCR) [223], PCR with highly degenerated consensus primers [11], and gene cassette PCR [186]. Some of these PCR variants also provide library-independent approaches for the recovery of novel genes [30].

Besides the inherent limitation of PCR regarding the amplification and detection of completely novel gene familes, several studies have shown that primer bias [150, 191] and co-amplification of homologous genes that generate chimeric sequences [207] reduce the quality of PCR amplified sequences, and thus prevent full recognition of the microbial diversity [206].

Microarrays The microarray technology can also be used for the taxonomic and functional characterization of metagenomes. For the taxonomic characterization of a metagenome, they are created by arraying oligonucleotides on the microarray surface that are complementary to 16S-rRNA sequences. These oligonucleotides are derived from known 16S-rDNA of various cultured and uncultured species. By specific hybridization of the 16S-rRNAs (the queries) in a metagenome to their targets on the microarray, a taxonomic profile of the metagenome can be obtained [61, 224]. Similarly, sequences of other known genes as targets can be used to create a functional profile of the metagenome [220].

An alternative strategy involves the spotting of metagenomic fragments derived from clone libraries onto the microarray. In contrast to the previous two settings, the metagenomic DNA now serves as target, while DNA from different sources, e.g., metagenomic isolates, reference strains, and complete communities, is used to probe the microarray. The different hybridization patterns provide a characterization of the metagenomic clones. For example, fragments on the microarray that show hybridization with multiple related species are likely to indicate conserved genes [135, 175].

Although environmental microarrays allow for fast identification and characterization of many clones, they have the disadvantage that the hybridization-based approach does not allow for the detection of sequences from novel, distantly related species. In addition, compared to PCR-based approaches, microarrays exhibit a 100 to 10 000-fold lower sensitivity for the detection of gene sequences [34].

1.2.5 Shotgun Metagenomics

The development of BAC libraries has made it possible to isolate larger fragments of environmental DNA from uncultured species for further investigation. In 2000, Béjà *et al.* [22]

created a BAC library of a marine environment with insert sizes of up to 150 kb and successively sequenced a 60 kb archaeal genome fragment. In the following years, decreasing sequencing costs faciliated further studies involving the sequencing of random genomic DNA fragments from natural microbial communities [20, 201]. In particular, the pioneering survey of the Sargasso Sea in 2004 by Venter *et al.* [205], which included cloning and sequencing of about two million DNA inserts, marked a new era in metagenomics, as this was the first attempt to sequence the entire genomic content of an environmental community.

With the development of 454 pyrosequencing [112], it became possible to do "shotgun metagenomics", i.e., the sequencing of huge amounts of DNA directly from a microbial community without the need for prior laborious cloning steps [44]. With 454 and other more recent HTS technologies that can sequence several orders of magnitude more DNA than the Sanger technology at the same cost, shotgun metagenomics has become a standard technique for the analysis of biological diversity in microbial communities and the underlying metabolic pathways. In contrast to the 16S-rDNA-based studies, which use "universal" primers for the rDNA genes for amplification and therefore are inherently biased [111], the shotgun sequencing is an undirected approach where primer-induced biases are avoided.

A limitation of the currently available HTS technologies is that they produce rather short reads (35–400 bp) and thus cover genes only partly [110]. A common strategy to handle the short reads is to assemble them into longer fragments in order to reconstruct full-length genes. Since this approach is very limited in a metagenomic context and may lead to wrong assemblies, one can try to infer information of a microbial community from the short reads without prior assembly steps. Comparing the short metagenomic sequences with sequences of known taxonomic origin and function allows to directly infer the taxonomic profile and the functional potentials of the metagenome.

Therefore, one of the differences between shotgun metagenomics and traditional genomics is that the analysis and annotation of a metagenome is performed on a set of short DNA sequences, rather than on complete genomic sequences. Higher-order genomic analyses beyond the size of single fragments are not possible without knowledge of the gene order in a genome.

In the following, we will discuss strategies and limitations of the assembly of shotgun reads, and the functional and taxonomic classification of assembled and unassembled metagenomic fragments.

Assembly of Shotgun Reads

Reconstruction of complete genome sequences by assembling the metagenomic reads is usually not possible. One problem is that complex microbial communities with large species richness require a huge sequencing depth in order to capture complete genomes of low-abundant species with sufficient coverage. Even if enough reads of an uncultured genome have been sequenced, the assembly process is hampered by the fact that often homologous sequences from different species are so similar on the nucleotide level that reads from different species are assembled together and chimeric contigs are produced [146]. Inserted phages might also contribute to such chimeric assemblies [161]. This problem is further increased by the problem that sequencing errors make it difficult to decide if two nearly perfect overlapping fragments represent two different species or only one species where the fragments differ in the overlapping regions because of sequencing errors. In addition, it has been shown that a high frequency of polymorphisms and genome variations can be found even up to the subspecies level [80, 146].

For metagenomes with a low biodiversity, it has been shown that it is possible to reconstruct near-complete genomes of the dominant species. Tyson *et al.* recovered two near-complete microbial genomes and three other genomes partially from an acid mine drainage metagenome [201]. Over 100 000 reads with average length of 737 bp were obtained by shotgun sequencing a small insert plasmid library. After assembly of the reads into longer fragments, the fragments were assigned to the different species based on G+C content and different sequence coverage. The resulting near-complete genomes were validated by comparison with closely related reference sequences.

If a particular uncultured species and not the whole metagenome is of interest, it might be possible to grow the metagenomic community under controlled conditions and to adapt the parameters such that the species of interest gets selectively enriched. This approach was followed by Ettwig *et al.* [45] who enriched denitrifying methanotrophic bacteria, a population of strains of a ditch sediment microbial community in a bioreactor. The enriched bacteria were sequenced with Illumina sequencing, yielding over six million 32-nucleotide reads. Dutilh *et al.* [41] used these reads, mapped them against a related reference genome and then assembled them into a consensus sequence. This assembly again served as a reference for a second assembly step and the process was iteratively repeated a few times until convergence of the consensus sequence was achieved. The advantage of this assembly strategy is that, with each iteration, the assembly becomes less dependent on the reference genome. The final assembly is a consensus of the multi-strain population. A similar approach has also been described in [137].

Increasing fragment length in order to reconstruct highly abundant genes in metagenomes is a common approach. In particular, genes that are shared by a large fraction of the individuals within a microbial community are likely to have a sufficiently high sequencing coverage that allows for an assembly. As this often holds for genes that encode functions which are specific to this metagenome, this approach can provide insights into the microbial community and its functions [70].

In the past, single genome assemblers have been used for the assembly of metagenomic reads into longer contigs [101, 139, 226]. Since these assemblers are not designed to cope with multiple species, they do not perform well on metagenomic data sets [115, 138]. Polymorphisms of genes that are shared by several species or strains and uneven species abundance ratios hamper the construction of longer contigs. New metagenomic *de novo* assemblers produce longer contigs at accuracies similar to the single genome assemblers [97, 138].

Shotgun Reads – Gene detection

One of the important tasks in metagenomic sequencing projects is to find reads that encode for proteins. This problem is closely related to the traditional gene detection problem in the single genome scenario. Computational gene detection is typically performed either with a homology search or with a *de novo* method. Homology-based methods, that use BLAST or HMMER [4, 43, 91], work only for genes for which already known homologues with high sequence similarity exist. Since this is typically the case for only a fraction of the genes in a metagenomic sample, *de novo* methods can be used that allow to find also novel genes. *De novo* methods use extrinsic information like start and stop codons, and oligonucleotide patterns combined with statistical models [17, 160]. Traditional microbial *de novo* gene finders are not optimal in the context of unassembled short metagenomic reads as they expect to find full-length ORFs including start and stop codons, whereas a metagenomic read can only encode a fragment of an ORF due to its limited length. Recently, new *de novo* gene detection methods like MetaGene [129], MetaGeneAnnotator [130], Orphelia [72] and GeneMark.hmm [228] have been published that use adapted statistical models, and thus can also detect incomplete ORFs and show a higher gene detection sensitivity and accuracy than previous traditional gene detection methods in the context of metagenomic data.

Shotgun Reads – Functional Classification

Since *de novo* gene detection methods usually do not provide a functional annotation of the predicted genes, BLAST- or HMMER-based homology searches are mostly used to detect and annotate genes [176]. For homology-based functional annotation, the unannotated sequence is searched against a database of sequences with known functions. A sequence with unknown function that shows a high similarity towards a database sequence with known function is then considered to likely have the same or a very similar function as the database sequence [82].

This approach can in principle also be applied on unassembled metagenomic reads [53]. For bigger metagenomic datasets with hundreds of gigabases, the homology search can become a computationally expensive or even intractable task [54]. A way to handle this problem is to apply various data reduction methods, like assembly of reads into longer fragments, gene detection with a *de novo* gene finder to detect ORFs, clustering of highly similar ORFs, and translation of the non-redundant ORFs into protein sequences [70, 144]. Such procedures can reduce the amount of sequences and therefore make a homology search computationally feasible.

Shotgun Reads – Taxonomic Classification

One way to examine the phylogenetic diversity and to create a taxonomic profile of a shotgun metagenome is to analyze special marker genes. Typically, 16S-rDNA genes, as discussed in Section 1.2.1, but also other genetic markers [108], are used. Nevertheless, these genes constitute only a small fraction of the total DNA within genomes. In an analysis that is based on 16S-rDNA genes, only 0.07–0.3% of all metagenomic reads can finally be used for a phylogenetic assignment [92, 205]. In contrast, composition-based taxonomic classification methods try to classify all metagenomic reads. But similarly, they have the disadvantage of not providing a direct linking between function and phylogeny. Therefore, another alternative is the taxonomic classification of fragments that encode for proteins in a metagenome using comparison-based methods. A high similarity between a metagenomic read and known DNA or protein sequences can be used to infer some information about phylogeny of the metagenomic sequences. An overview of the various composition- and comparison-based methods for the taxonomic classification of metagenomic sequences is given in Chapter 2.

In contrast to most of these methods, that create taxonomic profiles by classifying individual sequencing reads, Taxy [119] uses the total oligonucleotide composition of a metagenome to create a taxonomic profile.

Also closely related to the taxonomic classification is the clustering or *binning* of metagenomic sequences. Here the sequences are assigned into different bins, where each bin represents one species of the metagenome. In particular, if no closely related species is available as reference, and taxonomic classification methods therefore only can make predictions at higher taxonomic ranks, binning algorithms allow to estimate the species abundance in a metagenome without knowledge of the actual taxonomy of the underlying species. In addition, species-specific bins may facilitate assembly of metagenomic sequences. Examples for binning algorithms are LikelyBin [85], CompostBin [26], cBar [227], AbundanceBin [222], and MetaCluster [99].

1.2.6 Metatranscriptomics & Metaproteomics

The methods employed in metagenomics allow to reveal the genetic potential of microbial communities. They do not provide information about which genes are actually active at a specific time and place, or how those activities change in response to different environmental forces. In metatranscriptomics, this information can be obtained by investigating the abundance of messenger RNA (mRNA) in the environmental sample.

The first step in the metatranscriptomic sequencing of a microbial community is the extraction of RNA from the sample. Most of the extracted RNA consists of the more abundant and stable rRNAs, which are often selectively removed from the total RNA pool in order to increase the yield of sequenced mRNA later on. After the linear amplification of the mRNA with PCR, the mRNA is converted to cDNA, which can then be directly sequenced with the new pyrosequencing techniques.

Although this method is capable of detecting gene transcription activity in a microbial sample, it has the problem that the abundance of mRNA is not a perfect indicator of protein activity. Transcriptional regulation may prohibit translation of an mRNA into a protein and the protein activity can also be regulated after translation.

Metaproteomics, which includes protein extraction, separation and identification of proteins promises to overcome the restriction of the metatranscriptomic approach, but is still considered to be a more onerous approach [124].

1.3 Overview of the Thesis

In this chapter we have given a general overview of different metagenomic methods. We have shown that various PCR-based strategies, function and sequence-based screening methods for metagenomic libraries, and shotgun sequencing of microbial communities can be used to explore uncultured microbes. Since the main focus of this work is the taxonomic classification of short microbial DNA fragments, we will review existing methods that are used for that purpose in Chapter 2.

The main part of this thesis consists of the presentation of a novel algorithm for the taxonomic classification of DNA fragments in Chapter 3. This algorithm has been realized in CARMA3 and can also be used for the taxonomic classification of protein sequences and 16S-rDNA/RNA sequences.

In Chapter 4 we present WebCARMA, a web application that allows the usage of CARMA3 without the need for a local installation. Chapter 5 contains experiments for a comparative evaluation of CARMA3 on simulated metagenomes. Furthermore, we analyze several real metagenomes with CARMA3 and compare the results with 16S-rDNA based classifications. In the last experiment we show that also very short reads can, in principle, be used to describe the taxonomic content of a metagenome. Finally, the last chapter consists of a conclusion and ideas for the further development of CARMA3.

1.4 Acknowledgements

I thank my advisor Jens Stoye for making this dissertation possible and for his advice and suggestions during the course of this work. I also thank Daniel Dörr, Nils Hoffmann, Eyla Willing, and Roland Wittler for several helpful discussions and proofreading of my thesis. In particular, I thank my wife Chan for her understanding and love during the past few years.

Finally, I would like to acknowledge the DFG Graduiertenkolleg Bioinformatik for three years of scholarship and the NRW Graduate School in Bioinformatics and Genome Research for financial support.

Chapter 2

Methods for Taxonomic Classification of Metagenomic Reads

The most common taxonomy used to describe the origin of biological sequences is the NCBI taxonomy [12, 167]. It is a phylogenetic taxonomy with a tree structure that approximates the evolutionary relationships among organisms [116]. Leaves in the taxonomy tree usually refer to living organisms. In many cases, internal nodes in the NCBI taxonomy are multifurcating nodes indicating unresolved ancestral relationships. Each node is assigned a taxonomic rank, representing their relative position in the taxonomic hierarchy. The most commonly used ranks are superkingdom, phylum, class, order, family, genus, and species. Examples for taxa at various taxonomic ranks of human and *E. coli* are given in Table 2.1.

The purpose of taxonomic classification of metagenomic sequences is to find the exact species of origin. However, this is often not possible, in particular if this species has not been sequenced before. Furthermore, species closely related to the species of origin are often also not available as reference. In such a case, a taxonomic classification should make a prediction at a higher taxonomic rank, corresponding to the lowest known ancestor of the species of origin, and thus achieving highest possible sensitivity while avoiding false positives

Table 2.1: Taxa of human and the common bacterium *E. coli* at various taxonomic ranks according to NCBI taxonomy.

Taxonomic rank	Human	*E. coli*
Superkingdom	*Eukaryota*	*Bacteria*
Phylum	*Chordata*	*Proteobacteria*
Class	*Mammalia*	*Gammaproteobacteria*
Order	*Primates*	*Enterobacteriales*
Family	*Hominidae*	*Enterobacteriaceae*
Genus	*Homo*	*Escherichia*
Species	*Homo sapiens*	*Escherichia coli*

on lower taxonomic ranks. Sensitivity and specificity are measured for each taxonomic rank independently. The definitions of these measures slightly differ from the normal definitions since taxonomic classification is a multiclass classification problem in which a metagenomic sequence can either be assigned to a taxon at a given taxonomic rank, or it can be assigned the artificial taxon "unknown" [37]. In an evaluation each taxonomic assignment of a sequence at a given rank is considered to be either a correct classification and counts as a True Positive (TP), a wrong classification and counts as False Positive (FP), or it has been assigned the status "unknown" and it counts as an Unknown (U). Thus, in context of metagenomic classification, sensitivity and specificity are commonly given by:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{U}} \tag{2.1}$$

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.2}$$

In this section, we briefly review several methods for the taxonomic classification of metagenomic sequences. Further details on the various methods can be found in the corresponding publications. A systematic comparison in terms of sensitivity and specificity is not given here since the evaluations in the corresponding publications are usually not comparable. In general, they are either done on real or on simulated metagenomes. In case of real metagenomes, the predicted taxonomic profile is compared with a profile that has been obtained by other means, e.g. 16S analysis. In contrast, the usage of simulated metagenomes has the advantage that the taxonomic origin of each single sequence is known and can be compared with its prediction. Evaluations with simulated metagenomes usually consist of a test set, i.e., the simulated metagenomic sequences, and a training set, the reference dataset. A comparison of different evaluations can be problematic, for example, because evaluation settings can differ in the training sets, e.g., protein vs. genome sequences, or complete sequence databases vs. smaller marker gene databases. In addition, the simulated shotgun reads in the test sets may have different read lengths or were simulated under different error models.

Furthermore, evaluations can differ in how test and training sets are related to each other. *All-in* experiments or various variants of *cross-validation* experiments are used, representing different degrees of severity of the evaluation:

- *All-in:* The test set remains in the training set.
- *Leave-one-out with strains:* The test set is removed from the training set, but it is chosen such that each member of the test set has at least n strains of the same species in the training set. A typical value of n is 5.
- *Leave-one-out:* The test set is removed from the training set.
- *Leave-one-species-out:* The test set and all sequences that belong to the same species as those in the test set are removed from training set.

- *Leave-one-clade-out:* The test set and all sequences that belong to the same clade as those in the test set are removed from training set. Here a clade is defined by the set of all organisms that belong to the subtree induced by the ancestor of the test set member at a certain taxonomic rank.

Since real metagenomes represent a composition of different species to which often no closely related reference species is available, evaluation experiments should be designed in leave-one-clade-out manner, performed for different taxonomic ranks, to cover this scenario. This ensures that the evaluation provides a more realistic estimate on how a taxonomic classification method performs on real metagenomes in terms of sensitivity and specificity. Due to the above mentioned differences in the evaluations we provide only an overview in this chapter, rather than a systematic comparison.

In principle, two kinds of methods for the taxonomic classification of metagenomic shotgun sequences can be distinguished. *Composition-based* methods first extract sequence features and then perform a comparison on these features, whereas *comparison-based* methods compare metagenomic sequences with the reference sequences directly at the sequence level. We review these methods in the following two sections. An overview of these methods is given in Table 2.2.

2.1 Composition-based Methods

Composition-based methods extract sequence features like GC content [51], codon usage [129] or k-mer frequencies [117, 194], and compare them with features computed from reference sequences with known taxonomic origin. In detail, different techniques like the calculation of correlation coefficients between oligonucleotide patterns [194], Self-Organizing Maps (SOMs) [1], or Support Vector Machines (SVMs) [37] can be used to classify the metagenomic fragments. A disadvantage is that rather long sequences are required to obtain a reasonable classification accuracy. One of the advantages of composition-based methods over comparison-based method is that these methods usually are much faster because they do not require a time consuming homology search. In the following we review several composition-based methods.

2.1.1 TETRA (2004)

Tetranucleotide usage patterns can serve as species-specific intrinsic DNA-signatures in metagenomic fragments. TETRA [193, 194] uses the distribution of tetranucleotides of different metagenomic fragments to compute pairwise correlation coefficients, which can be used as an estimation for the likelihood that two metagenomic fragments originate from the same genome. A comparison of sequences from a metagenome with the distribution of tetranucleotide patterns of sequences with known taxonomic affiliations thus provides a taxonomic characterization of the metagenome. The authors report that this method requires sequences in the range of 40 kb to work well, while sequences below 1 kb are not suited for the analysis.

Table 2.2: An overview of the various methods for taxonomic classification reviewed in this chapter. Checkmarks indicate whether the methods are composition- or comparison-based and if they use a Bayesian Classifier or Maximum Likelihood (ML).

		Basis				
Method	Year	Composition	Comparison	Bayesian	ML	Ref.
TETRA	2004	✓				[193, 194]
PhyloPythia	2007	✓				[117]
PhyloPythiaS	2011	✓				[136]
TACOA	2009	✓				[37]
RAIphy	2011	✓				[127]
NBC	2008	✓		✓		[156]
Phymm	2009	✓		✓		[19]
PhymmBL	2009	✓	✓	✓		[19]
GSTaxClassifier	2009	✓		✓		[225]
BLAST	1990		✓			[4, 57]
MG-RAST	2008		✓			[122]
MEGAN	2007		✓			[78]
CARMA1	2008		✓			[93]
AMPHORA	2008		✓			[221]
SOrt-ITEMS	2009		✓			[65]
Sphinx	2010	✓	✓			[64]
DiScRIBinATE	2010		✓			[55]
MetaPhyler	2010		✓			[104]
MARTA	2010		✓			[73]
EPA	2009		✓		✓	[15, 16]
Pplacer	2010		✓		✓	[114]
MLTreeMap	2010		✓		✓	[184]
Treephyler	2010		✓		✓	[173]

2.1.2 PhyloPythia (2007)

For each sequence, PhyloPythia [117] uses oligonucleotides of a certain length (typically 4–6 bp) to represent training sequences from reference genomes as vectors that contain the abundance of each oligonucleotide, normalized by the total number of oligonucleotides in the corresponing sequences. These vectors are used to train a collection of Support Vector Machines (SVMs) for each taxonomic rank and each clade that is represented by at least three genomes. As the SVM is intrinsically a binary classifier, an all-versus-all technique is applied to perfom a multiclass classification for the different possible clades at each taxonomic rank. Using a voting mechanism, metagenomic fragments are assigned to a clade, which is re-evaluated with a classifier that has been trained to discriminate between training sequences of this clade and all other training sequences (one-versus-all approach). The authors report that their method provides an accurate classification for longer metagenomic fragments, but for fragments shorter than 3–5 kb the sensitivity decreases strongly.

PhyloPythiaS (2011)

PhyloPythiaS [136] is an improved successor of PhyloPythia that uses an ensemble of linear models instead of multiclass SVMs. The parameters of the linear models are obtained using the paradigm of SVMs with structured output spaces in order to represent composition-based specifics of each clade in the taxonomic hierarchy.

The authors show that their algorithm can outperform the comparison-based algorithms MEGAN (Section 2.2.2) and PhymmBL (Section 2.1.5) in a special scenario where 100 kb fragments of the novel species are available in the training phase. Results for the usual scenario without such additional information from 100 kb fragments are not given.

2.1.3 TACOA (2009)

TACOA [37] is based on the k-nearest neighbor (k-NN) approach combined with a Gaussian kernel function. Each genomic sequence is represented as a vector that stores for each oligonucleotide the ratio between its frequency and its expected frequency given the GC-content of the sequence. These vectors are computed, once in a preprocessing step, for all sequences from the set of reference genomes, and for each metagenomic sequence with unknown taxonomic affiliation. Each metagenomic sequence is assigned to one taxon of each taxonomic rank from superkingdom to genus. For each taxonomic rank, the taxon is chosen for which a discriminant function provides the maximal value. If the taxon with the second highest value of the discriminant function is too close to the highest value, the metagenomic fragment is instead assigned to “unknown” at this taxonomic rank. The authors show that their method is able to classify genomic fragments of length 800 bp to 1,000 bp with high accuracy for rank class or higher. For longer sequences the method provides accurate predictions also at lower taxonomic ranks like order or genus.

2.1.4 RAIphy (2011)

RAIphy [127] is a composition-based semisupervised binning algorithm which uses a so-called *Relative Abundance Index* (RAI). This index is computed for each k-mer and indicates the over- or underabundance of a k-mer within a taxon. It is computed using a sequence of fixed-length Markov models and log-odds ratios between the observed and expected frequencies of the k-mers in each taxon. To assign a metagenomic fragment to a taxon, the RAI is computed for each k-mer in the metagenomic fragment and for each taxon. A taxon membership score is then obtained for each taxon by summing up the values for all k-mers. Finally, the metagenomic fragment is assigned to the taxon that yields the highest membership score. An additional refinement phase can be used to further improve the taxonomic assignment. For a leave-one-out cross-validation, the authors report a sensitivity of 38% to 81% for sequences of length between 100 bp and 1,000 bp.

2.1.5 Bayesian Classifiers

The general principle of using a Bayesian classifier in the context of taxonomic classification of bacterial genomic sequences was first proposed and evaluated by Sandberg *et al.* in 2001 [162]. Their method was able to correctly classify 400 bp-long reads with a sensitivity of 85%. In 2006, Dalevi *et al.* [32] presented a similar algorithm that combined the Bayesian classifier with fixed higher order- and variable length Markov Models in order to predict horizontal gene transfer. The RDP classifier, described in detail in Section 1.2.1, is also based on this method.

NBC (2008)

NBC by Rosen *et al.* [156] is a naïve Bayesian classification method. To classify a metagenomic fragment, NBC compares the k-mer frequency profile of the metagenomic fragment with all k-mer frequency profiles from the set of the microbial reference genomes. The naïve Bayesian classifier is then used to calculate the posterior probability of each taxonomic clade. The metagenomic fragment is finally assigned to the taxonomic clade with the highest probability. In addition, the NBC uses an optimized algorithm for efficiently counting k-mer frequencies. The authors report that, in a cross-validation with each species represented by at least four strains, their method achieves a species-sensitivity of 90% and more. NBC is also available as a webserver [157].

Phymm and PhymmBL (2009)

Phymm [19] uses interpolated Markov models (IMMs) to compare variable-length oligonucleotide usage patterns of the query sequence and the reference sequences. A Bayesian decision machine is used to compute the most likely taxonomic origin of the query. PhymmBL is a variant of Phymm which additionally incorporates BLAST results by using a weighted combination of scores from Phymm and the best BLAST hit. In a leave-one-clade-out evaluation, Phymm performs similar to BLAST in terms of sensitivity, whereas the hybrid method PhymmBL slightly outperforms both Phymm and BLAST. An interpretation of the results of the evaluation is difficult because the authors do not report the specificity of the compared methods. Furthermore, the numbers of wrong predictions at lower taxonomic ranks that have been removed are not given.

GSTaxClassifier (2009)

The GSTaxClassifier [225] is a slightly modified variant of the Bayesian method proposed by Sandberg *et al.* [162]. In a leave-one-out evaluation with bacterial 400 bp reads the authors report a taxonomic assignment sensitivity of 63–95% at ranks order to kingdom.

2.2 Comparison-based Methods

In contrast to the composition-based methods, comparison-based methods rely on homology information obtained by database searches. Databases used in this context can contain nucleotide sequences, e.g. complete genomes, or protein sequences with known taxonomic origin. Since protein sequences are more conserved than nucleotide sequences, they are better suited for detection of remote homologies. Microbial communities mainly consist of uncultured microbes. Therefore, a high sensitivity is required for most metagenomic sequences in order to find the closest homologues. If protein sequences are used as references and metagenomic reads are found to encode for known proteins, they are called *environmental gene tags* (EGTs) [198]. It is in principle possible to use reads much shorter than 100 bp, but then only a small fraction of all reads actually provides information about their function and taxonomic origin as shown in the experiment in Section 5.9.

Usage of protein sequences as reference has the disadvantage that, for metagenomic fragments which contain only non-coding DNA, no homologies can be found. Nevertheless, in practice only a small fraction of the bacterial and archaeal metagenomic sequences are affected by this. This is discussed in more detail in the experiment described in Section 5.6. Another disadvantage of protein references is that the metagenomic DNA fragments have to be translated into all six reading frames, which increases computation time of the homology search.

In general, comparison-based methods can be further subdivided into methods that are based on Hidden Markov Model (HMM) homology searches [43] and those that are based on BLAST homology searches [4, 57]. CARMA1 [93] and various maximum likelihood methods detailed in Section 2.2.8 belong to the HMM-based methods. In contrast, algorithms like MEGAN [78] and SOrt-ITEMS [65] use BLAST for the homology search. The method CARMA2 [53] which we will introduce in Chapter 3 of this thesis is a HMM-based method. CARMA3 [54] which we also introduce in the same Chapter is available in two variants, a HMM-based variant and a BLAST-based variant.

2.2.1 Best BLAST Hit, MG-RAST (2008)

For the taxonomic classification of metagenomic reads, the most basic and widely used method is probably BLAST [7, 31, 35, 197]. This approach is also followed in MG-RAST [122]. The idea is to search for the best BLAST hit in a database of sequences with known origin and to use the taxonomy of the sequence that produced the best hit as reference for the metagenomic read. Since the evolutionary distance between the source organisms of the metagenomic fragment and the database sequence is unknown, a classification result solely based on a best BLAST hit has to be interpreted carefully. In general, such a classification is more reliable on higher taxonomic levels (e.g., superkingdom or phylum) than on lower taxonomic levels (e.g., genus or species), but it is difficult to decide which taxonomic level is reliable enough, as this strongly varies for each metagenomic fragment. Usually only sequences that have BLAST hits with good (low) E-values are trusted. Other homology search algorithms, like FASTA [103] or BLAT [84], can also be used in principle.

2.2.2 MEGAN (2007)

The program MEGAN [78] is based on the lowest common ancestor (LCA) approach: A BLAST search is performed, and all BLAST hits that have a bit score equal or higher than

90 % of the bit score of the best hit are collected. This percentage value is a parameter that allows a trade-off between sensitivity and specificity. The metagenomic fragment is then classified by computing the LCA of all species in this set. One of the reasons for the improved classfication accuracy of this approach compared to using only the best BLAST hit is that fragments with ambiguous hits are assigned at higher taxonomic levels. All-in and leave-one-out evaluation experiments have been performed on simulated reads of different lengths from two individual genomes. The authors conclude that their results demonstrate the robustness of the LCA algorithm. The LCA-method has also been adopted in other analysis scenarios, e.g. [33].

2.2.3 CARMA (2008)

Within the DNA reads that next-generation sequencing technologies produce, CARMA 1.2 by Krause *et al.* [93] detects those that encode for known proteins. These EGTs are then assigned in a second step to taxa from six taxonomical ranks: superkingdom, phylum, class, order, genus and species. The set of classified EGTs provides a taxonomical profile for the microbial community.

In detail, BLASTx is used to search within the set of reads for candidate EGTs that encode for protein sequences contained in the Pfam database [49]. A rather relaxed E-value of 10 and frameshift option `-w 15` are used. Each read that has a match to a protein family member is translated according to BLASTx reading frame and frameshift predictions. The final determination of EGTs is done by matching the candidate EGTs against their matching protein families with the corresponding Pfam Hidden Markov Models [40]. For this purpose, `hmmpfam` from the HMMER package [43] is used. Only candidate EGTs with an `hmmpfam` E-value match of 0.01 or lower are accepted as EGTs.

After the EGTs are identified, they are taxonomically classified: Each EGT is aligned against the multiple alignment of its family with `hmmalign` (also contained in the HMMER package). From this new alignment, the pairwise sequence distance is computed for all pairs of sequences, based on the fraction of identical amino acids. This produces a pairwise distance matrix which is then used to compute a phylogenetic tree with the neighbor-joining method [159]. After this step, the EGT is classified depending on its position within this tree. If the EGT is localized within a subtree of family members all sharing the same taxon, then the EGT is classified with the same taxon. For example, if the EGT is localized in a subtree with the three members *Bacteria Cyanobacteria Synechococcales Prochlorococcus*, *Bacteria Cyanobacteria Chroococcales Synechococcus* and *Bacteria Cyanobacteria Nostocales Nostoc*, the EGT is classified as *Bacteria Cyanobacteria*. For a more formal definition and further details, see [93].

CARMA is evaluated on a synthetic metagenome with 80–120 bp long reads in a leave-one-species-out strategy. For taxonomic rank order the authors report a specificity of about 60% at a sensitivity of about 90%.

More recent versions of CARMA, notably CARMA2 and CARMA3, that also belong the comparison-based methods, are introduced in Chapter 3.

2.2.4 AMPHORA (2008)

AMPHORA [221] uses a set of 31 protein-encoding marker genes for the taxonomic characterization of metagenomic samples. These marker genes are housekeeping genes, mostly single-copy genes, that are universally distributed in bacteria. For each marker gene, the corresponding homologous protein sequences from a set of reference genomes are aligned using CLUSTAL W [196]. The alignments are concatenated to obtain a long alignment which is used to create a maximum likelihood tree with PHYML [60]. This tree serves as a reference tree in the following. In addition, local profile Hidden Markov Models [43] are created for each individual marker gene alignment. The Hidden Markov Models are used to search the metagenomic sequences for those that encode for the marker genes, and to align them against the marker gene multiple alignments. Using a maximum parsimony method of RAxML [183], the metagenomic sequence is placed into the reference tree. For more robust results, a further refinement is performed including additional bootstrap replicates leading to the final taxonomic assignment. In a comparative evaluation against MEGAN, AMPHORA yields a higher sensitivity at similar specificity.

2.2.5 SOrt-ITEMS (2009)

The SOrt-ITEMS [65] method extends the LCA method and uses additional techniques to reduce the number of false positive predictions. One approach is the reduction of the number of hits by using a reciprocal BLAST search step. Another technique used is the adaptation of the taxonomic assignment level for all hits, based on different alignment parameters like sequence similarity between the metagenomic fragment and the aligned database sequence.

SOrt-ITEMS was evaluated in a leave-one-clade-out manner, for taxonomic ranks species and genus, with simulated 454 and Sanger reads. Results show that SOrt-ITEMS consistently makes significantly fewer false predictions than MEGAN in all evaluation scenarios. The numbers of false predictions at lower taxonomic ranks are not given.

In the following, two variants of SOrt-ITEMS are described.

Sphinx (2010)

Sphinx [64] is a binning algorithm that extends the above mentioned SOrt-ITEMS algorithm by introducing a k-mer-based filter step and therefore combines composition- and comparison-based strategies. In a pre-processing step, protein encoding sequences from microbial genomes are clustered based on their tetra-nucleotide frequencies with a k-means clustering approach. For each cluster a centroid is computed and the sequences are translated into protein sequences. The first step in the taxonomic classification consists of computing the distance of the metagenomic fragment to all cluster centroids. The fragment is then assigned to the cluster whose centroid has the smallest distance. After a BLASTx search of the metagenomic fragment against the translated sequences in this cluster, the SOrt-ITEMS algorithm is used for the final classification. The authors report a significant improvement in speed with little loss in sensitivity compared to SOrt-ITEMS.

DiScRIBinATE (2010)

DiScRIBinATE [55] is another modified variant of SOrt-ITEMS. The method involves a reclassification of reads that have been assigned to taxa with few assigned reads. An evaluation shows a reduced running time and slightly higher sensitivity in comparison to SOrt-ITEMS.

2.2.6 MetaPhyler (2010)

MetaPhyler [104] restricts the taxonomic classification to 31 marker genes similar to those of AMPHORA to create a phylogentic profile. It reduces the taxonomic assignment of the best BLAST hit to a higher taxonomic level depending on parameters like the bit score and length of the high-scoring segment pairs of the hit. The thresholds for the taxonomic level are obtained in a preprocessing step from the reference database for each reference gene individually. This approach is more flexible than a universal BLAST threshold as it takes into account that some genes are more conserved than others. Conceptually, this approach is similar to SOrt-ITEMS (or CARMA3, see Section 3.2) since it uses adapted thresholds to reduce the level of the taxonomic assignment of the best BLAST hit. In a comparative leave-one-out evaluation with MetaPhyler, MEGAN, and PhymmBL, MetaPhyler shows a higher sensitivity at nearly the same specificity level compared to MEGAN or PhymmBL. A leave-one-clade-out evaluation is also performed, but only for MetaPhyler. The results indicate a high specificity for all taxonomic ranks, while sensitivity falls significantly if no closely related species are available. Numbers for wrong classifications at lower taxonomic ranks are not given.

2.2.7 MARTA (2010)

The taxonomic classification of MARTA [73] is based on the taxonomy of the best BLAST hit and uses alignment parameters (e.g. percent identity) as thresholds to make assignments at higher taxonomic ranks if necessary. MARTA can be used to classify metagenomic shotgun sequences as well as 16S-rDNA sequences. In the evaluation on a set of reference 16S-rDNA sequences, the authors compare their method with the RDP Classifier [208] and yield similar results in terms of sensitivity and specificity. Notably, WebCARMA (CARMA2.1) also was included in this evaluation, although it is a classifier for protein-encoding DNA sequences rather than 16S-rDNA sequences.

2.2.8 Maximum Likelihood Methods

Maximum Likelihood is a commonly used approach for the reconstruction of phylogenetic trees. Given a multiple alignment of sequences from different species, a model of evolution and a tree topology, one can compute the likelihood that the multiple alignment was produced under this model of evolution and tree topology. The idea is that a tree topology maximizing this likelihood is at least a good approximation of the real phylogenetic tree [47].

This method of reconstructing phyogenetic trees can also be used to place a new metagenomic sequence with unknown taxonomy into a reference taxonomy. The metagenomic sequence is aligned against the corresponding reference alignment and for each possible location of the metagenomic sequence in the reference tree, the likelihood of the new tree topology is computed. The placement of the metagenomic sequence that yields the highest likelihood finally determines its taxonomic assignment. Four methods that are based on this principle are introduced in the following.

EPA (2009, 2011)

The evolutionary placement algorithm (EPA) [15, 16] starts with a given reference tree and reference alignment. Using RAxML [183] it optimizes the Maximum Likelihood model parameters and branch lengths. Then, for each query, the tree is traversed once. At each edge, the likelihood score of the complete tree that is obtained by inserting the query into the current edge is computed. The scores for all queries and all insertion points are stored in a table. A heuristic is applied to optimize the edge lengths. Finally, for each query, the edge that yields the best insertion score is assigned.

The authors perform a leave-one-out evaluation where for each query the node distance between original taxon and assigned taxon is measured. The evaluation results show that the EPA yields on average a significantly lower node distance than an assignment based on BLAST.

Pplacer (2010)

Pplacer [114] can be run in Maximum Likelihood (ML) mode or in Bayesian mode to place a query sequence into a reference tree. Unlike e.g. MEGAN, pplacer does not provide a taxonomic labeling of individual metagenomic fragments. In ML mode, pplacer is conceptually quite similar to EPA. In an evaluation using 16S sequences, the authors show that their algorithm performs similar to EPA regarding speed and sensitvity.

MLTreeMap (2010)

MLTreeMap [184] uses a set of 40 reference protein families as markers, similar to AMPHORA. In a first step, the metagenomic sequences are searched for marker genes using BLASTx. Sequences that encode for marker genes are extracted and translated using GeneWise and aligned against the reference protein families using `hmmalign`. Finally, the metagenomic fragments are placed in their most likely position using RAxML.

The results of a leave-one-out evaluation show a slight advantage in sensitivity for the Maximum Likelihood approach of MLTreeMap compared to the Maximum Parsimony approach of AMPHORA.

Treephyler (2010)

Treephyler [173] uses profile Hidden Markov Models to assign metagenomic sequences to Pfam families. For each Pfam family and its assigned metagenomic sequences, a phylogenetic tree is computed using FastTree, a minimum evolution heuristic with an sensitivity closely to that of Maximum Likelihood methods [140]. The evaluation using a real dataset shows a high consistency between the results of Treephyler and CARMA1.

Chapter 3

CARMA2 and CARMA3

In Chapter 1, we have shown several metagenomic techniques that have been developed to explore microbes that cannot be cultivated in a monoculture. Of these approaches, we are particularly interested in computational methods for the taxonomic classification of metagenomic sequences. An overview of a variety of existing methods has been given in Chapter 2.

In this chapter, we present our contributions to the development of improved methods for the taxonomic classification of metagenomic sequences. In the first section of this chapter, we present CARMA2, a slightly improved version of CARMA1, which we also use to evaluate the applicability of short reads in metagenomics. In the second section of this chapter, we present CARMA3, our major contribution, which implements a novel method for the taxonomic classification of different kinds of metagenomic sequences. Both versions of CARMA can be downloaded from `http://webcarma.cebitec.uni-bielefeld.de`.

3.1 CARMA2

As reported in Section 2.2, comparison-based methods can be subdivided into methods that are based on Hidden Markov Model homology searches and those that are based on BLAST homology searches. Version 1 of CARMA (reviewed in Section 2.2.3) belongs to the former since it uses HMMER in combination with the Pfam database.

We have reimplemented large parts of CARMA1, including a faster construction of the phylogenetic trees by caching the pairwise distances between Pfam family members. The CARMA results now include for each EGT the corresponding hmmpfam E-values and a list of GO-Ids (Gene Ontology Identifiers) [8] associated with the corresponding Pfam family. The Gene Ontology provides a controlled vocabulary for gene products, distinguishing between their associated biological processes, cellular components and molecular functions, and can therefore be used to create a functional profile of the metagenome.

A major modification in CARMA2 is the usage of the NCBI taxonomy database [12, 167] instead of the Pfam nomenclature. The NCBI taxonomy database currently indexes over 200,000 species [128], which are classified in a hierarchical tree structure. Each taxon from the taxonomy is represented as a node in the tree with a unique identifier (`tax_id`) and its taxonomic rank ranging from “superkingdom” to “subspecies”. For compatibility with other applications and databases, the output files of CARMA contain for each classification the taxon name and NCBI `tax_id`. A detailed description of the output formats is given in Section A.1.2.

We have used CARMA2 to evaluate the applicability of short reads in a metagenomic analysis. The experiments can be found in Section 5.9. Furthermore, CARMA2 was the initial version that was taken as the back end for WebCARMA, a web application for the taxonomic and functional classification of metagenomic DNA sequences, which we will introduce in Chapter 4 of this thesis.

3.2 CARMA3

The method employed by CARMA3 can be seen as the result of an evolution of several BLAST-based methods. As already pointed out in Section 2.2, the taxonomic assignment provided by a best BLAST hit is less reliable on lower taxonomic ranks. The LCA method of MEGAN from 2007 (Section 2.2.2) avoids many such false taxonomic predictions at low ranks by assigning metagenomic fragments with ambiguous hits at higher taxonomic ranks. In 2009, SOrt-ITEMS (Section 2.2.5) was developed by extending the LCA method. The introduction of a reciprocal search step significantly decreased the number of wrong predictions compared to MEGAN.

Inspired particularly by the reciprocal search step of SOrt-ITEMS, we have developed a new algorithm that further improves the accuracy of the taxonomic classification. Our method makes explicit use of the assumption of a model of evolution where different gene families have different rates of mutation, but within each family this rate does not change too much. It also accounts for variably conserved regions within genes, e.g. functional domains vs. evolutionarily less conserved regions. In contrast to the previous CARMA versions 1 and 2 that

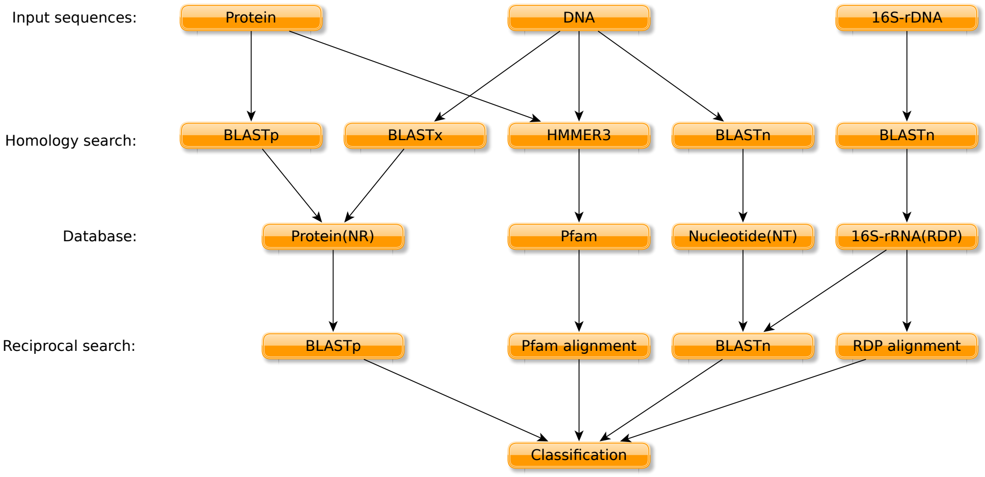


Figure 3.1: Overview of the CARMA3 pipeline showing the possible processing paths for different input sequences.

use HMMER2 for the homology search, we haved adapted CARMA3 to work with different homology search methods, namely BLAST and HMMER3.

CARMA3 accepts protein sequences, protein-encoding DNA sequences and 16S-rDNA sequences as input. Depending on the input sequences and the chosen reference database for homology search, the CARMA3 pipeline uses different processing paths, as depicted in Figure 3.1.

In the following we first introduce the BLASTx-based variant of our method, and then we detail the adaptations necessary for the HMMER3 variant. The BLASTx-based variant corresponds to the path "DNA → BLASTx → Protein(NR) → BLASTp → Classification" in Figure 3.1, while the HMMER3 variant corresponds to the path "DNA → HMMER3 → Pfam → Pfam alignment → Classification". After these two variants we also introduce the other BLAST variants and the 16S variant.

Definitions

For a given BLAST hit h, let $\mathrm{q}(h)$ be the aligned query sequence without gap and frameshift characters. In case of BLASTx, $\mathrm{q}(h)$ is a translated substring of the DNA query sequence. Similarly, $\mathrm{s}(h)$ is the substring of the database sequence used in the alignment of h. Furthermore, $\mathrm{score}(h)$ is the bit score of the alignment of h and $\mathtt{tax}(h)$ is the taxonomic assignment of the database sequence of h. Given two taxa a and b, $\mathrm{lca}(a, b)$ is the lowest common ancestor of a and b. Let RANKS be the set of the taxonomic levels {unknown, superkingdom, phylum, class, order, family, genus, species}, with the underlying taxonomic ordering relation unknown $>$ superkingdom $> \ldots >$ species. For a given taxon a, $\mathrm{rank}(a)$ is the taxonomic rank of taxon a. The *lineage* of some taxon a denotes the set of taxa on the path from the root to a in the taxonomy tree. For a given rank k, $\mathrm{ancestor}(k, a)$ defines the taxon at rank k in the lineage of a. In the remainder of this section, let query q be an metagenomic sequence with unknown taxonomic affiliation.

Reciprocal Search

The basic idea of using a reciprocal search in the context of the taxonomic classification of metagenomic reads as described in the following goes back to SOrtITEMS (Section 2.2.5). The first step of our method is to use BLASTx to search for homologs of q in the NCBI NR protein database. BLASTx hits with taxonomic assignment *Other* or *Unclassified* and hits without any taxonomic assignment are discarded. Furthermore, hits that have bit scores or alignment lengths that are below certain thresholds, are also discarded. Let $B = \{h_1, \ldots, h_{|B|}\}$ be the set of BLAST hits of q with $\mathrm{score}(h_1) \geq \ldots \geq \mathrm{score}(h_{|B|})$. If B is empty, then q is classified as "unknown". Otherwise, the next step is the construction of a new BLAST database consisting of $\{\mathrm{q}(h_1), \mathrm{s}(h_1), \ldots, \mathrm{s}(h_{|B|})\}$. Then, BLASTp is used to search for hits of $\mathrm{s}(h_1)$ in the new database. The result of this reciprocal search is (r_{query}, R), where r_{query} denotes the hit obtained by the alignment between $\mathrm{s}(h_1)$ and $\mathrm{q}(h_1)$, and $R = \{r_1, \ldots, r_{|R|}\}$ denotes the set of hits with known taxonomic affiliation with $\mathrm{score}(r_1) \geq \ldots \geq \mathrm{score}(r_{|R|})$. In addition, we require that, in case of co-optimal results with the same highest score, $r_1 \in R$ denotes the hit obtained by the alignment of $\mathrm{s}(h_1)$ with itself. Let $x = \mathrm{tax}(r_{\text{query}})$ and $t_i = \mathrm{tax}(r_i)$ for all $r_i \in R$. Determining x, the species of the metagenomic fragment, is usually not possible if the species has not been sequenced before. The purpose of this method is to approximate $y = \mathrm{lca}(x, t_1)$, which is the best possible classification, assuming t_1 is the phylogenetically closest known homolog of x. For each $r \in R$, $\mathrm{p}(r) = \mathrm{rank}(\mathrm{lca}(\mathrm{tax}(r), t_1))$ denotes the projection of r onto the lineage of t_1. For each $k \in$ RANKS, let $\mathrm{P}_k = \{r \in R \mid \mathrm{p}(r) = k\}$. If $\mathrm{P}_k \neq \emptyset$, let $\mathrm{Pmin}_k = \min\{\mathrm{score}(r) \mid r \in \mathrm{P}_k\}$ and $\mathrm{Pmax}_k = \max\{\mathrm{score}(r) \mid r \in \mathrm{P}_k\}$, otherwise $\mathrm{Pmin}_k = \mathrm{Pmax}_k = 0$. Pmin_k and Pmax_k define intervals for each taxonomic rank k.

Figure 3.2(a) depicts an example with projections of phylogenetic affiliations $t_2, \ldots, t_8$ of reciprocal BLAST hits $r_2, \ldots, r_8$ onto the lineage of t_1. Note that this tree is not a phylogenetic tree. For example, the species t_8, t_7 and t_6 share a common ancestor at taxonomic level "order" with t_1, but this is not necessarily the last common ancestor of t_8, t_7 and t_6. The dashed edges represent the projections of the hitherto unknown phylogenetic affiliations x and x' of metagenomic sequences q and q', respectively.

Figure 3.2(b) shows intervals defined by Pmin_k and Pmax_k that were obtained from the reciprocal scores in Figure 3.2(a). For example, the species t_8, t_7 and t_6 define the interval $(50, 75)$ at taxonomic rank "order" and "species"; t_4 and t_2 define the interval $(95, 120)$ at taxonomic rank "genus".

Polishing

Under ideal conditions, one would expect reciprocal hits that are phylogenetically further away from t_1 having a lower bitscore. Thus, one would expect that for each taxonomic rank $k \in$ RANKS $\setminus$ $\{\text{unknown}\}$, $\mathrm{Pmax}_k \geq \mathrm{Pmax}_{k+1}$ holds. As this is not always the case for real data, Pmax_k is set to zero for all ranks k with $\mathrm{Pmax}_k < \mathrm{Pmax}_{k+1}$.

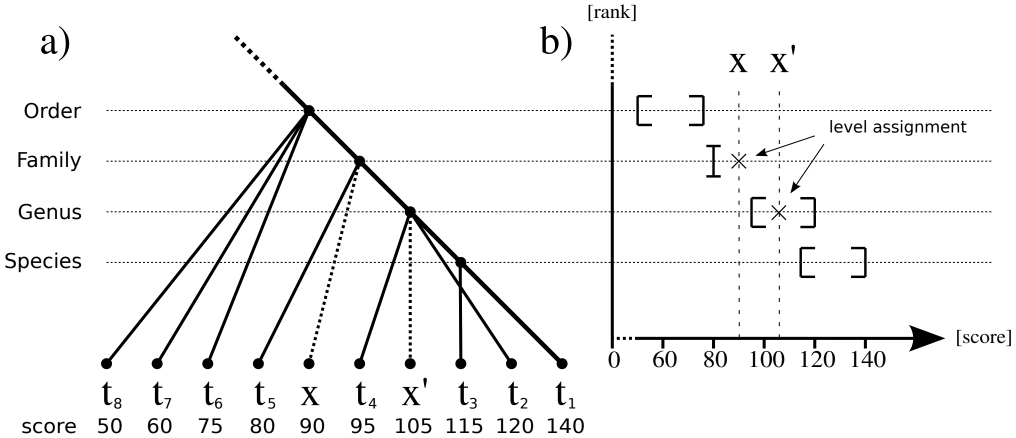


Figure 3.2: (a) Projections of BLAST hits obtained from reciprocal search onto the lineage of t_1. The dashed edges represent projections of unkown phylogenetic affiliations x and x' of metagenomic sequences q and q', respectively. (b) Intervals given by Pmin_k and Pmax_k for each taxonomic rank k and level assignments of x and x' based on their score.

Values of Pmax_k that are zero, because there was no hit at this taxonomic rank or because they have been set to zero in the previous step, can be approximated by a linearly interpolated score if there exists at least one higher and one lower taxonomic rank for which Pmax is non-zero. Note that there always exists some lower taxonomic rank with Pmax $\neq 0$ since r_1 provides a lower bound at taxonomic rank “species”. Thus, if a higher taxonomic rank with Pmax $\neq 0$ exists, the smallest rank $k_h > k$ with $\text{Pmax}_{k_h} \neq 0$ and the largest rank $k_l < k$ with $\text{Pmax}_{k_l} \neq 0$ are used as anchors for the linear interpolation. If $\text{Pmin}_k = 0$, Pmin_k is set to Pmax_k. If no k_h exists, an interpolation is not possible.

Classification

Another formulation of the best possible classification $y = \text{lca}(x, t_1)$ is $y = \text{ancestor}(k, t_1)$, assuming that rank $k = \text{rank}(y)$ is given. Similarly, y_{approx}, an approximation of the best possible classification, can be obtained by $\text{ancestor}(k_{\text{approx}}, t_1)$ if rank k_{approx} is given. Therefore, the goal of our method is to find such an approximation k_{approx}. This step requires that there exists some reciprocal BLAST hit $r \in R$ with $\text{score}(r) \leq \text{score}(r_{\text{query}})$. If this is not the case, a fall-back method, which is described below, will be used. Otherwise, we obtain k_{approx} by $\min\{k \in \text{RANKS} \mid \text{Pmin}_k \leq \text{score}(r_{\text{query}}) \text{ and for all } l > k : \text{Pmax}_l < \text{score}(r_{\text{query}})\}$. The algorithm for this works as follows: Starting at taxonomic rank $k =$ “unknown”, k is decreased until $\text{Pmax}_{k-1} \geq \text{score}(r_{\text{query}})$. If k is above the taxonomic rank “species” and $\text{score}(r_{\text{query}}) \geq \text{Pmin}_{k-1}$, then k will be decreased once again. The rank k_{approx} is then given by k.

Two examples for the taxonomic classification are given in Figure 3.2(b). The metagenomic read q with unknown phylogenetic affiliation x has a reciprocal score of 90 and k is decreased until $\text{Pmax}_{k-1} \geq 90$. Since the interval at taxonomic rank “genus” contains a reciprocal hit (t_2) with a score of 120, which is higher than that of q, k is set to rank “family”.

Because the score of q is also smaller than the lowest score Pmin_{k-1} of any reciprocal hit in the interval at rank "genus", k remains at its last rank and k_{approx} is set to "family". For the metagenomic read q' with reciprocal score of 105, k is similarly placed at taxonomic rank "family" in the first phase, but in contrast to q, its score is higher than the lowest score in the interval at taxonomic rank "genus". Therefore k_{approx} is set to "genus" for metagenomic read q'.

Fall-back

As mentioned before, the previous step will only work if there exists some reciprocal BLAST hit $r \in R$ with $\text{score}(r) \leq \text{score}(r_{\text{query}})$. If there is no such r, the highest taxonomic rank k_{low} with $\text{P}_{k_{\text{low}}} \neq \emptyset$ will only provide a lower bound for the approximation of y. As a fall-back method for this case, the lower bound prediction k_{low} will be combined with a technique introduced in SOrt-ITEMS [65] that is based on the assumption of a uniform rate of evolution. Different BLASTx alignment parameters, e.g. percent identity, are used to estimate the taxonomic rank of the lowest common ancestor of the metagenomic sequence and the database sequence. A high similarity between both sequences will result in the estimation of a lower taxonomic rank and a lower similarity will result in a higher taxonomic rank, respectively. For example, a metagenomic read with a BLAST hit h to some database sequence, with $\text{length}(\text{q}(h)) = 200\,\text{bp}$ and $\text{percent identity} = 60$, is assigned at the taxonomic rank "family" of the database sequence. In contrast, the same metagenomic read with an alignment with a percent identity of only 55 will be assigned at the higher taxonomic rank "order", as it is assumed to be evolutionarily further away from the database sequence. For reasons of comparability, the thresholds for the alignment parameters used in this method are the same as in SOrt-ITEMS. Let k_{uni} be the taxonomic rank obtained by this technique using the alignment parameters of the best BLAST hit h_1 from the initial BLAST search. Both predictions are combined by taking the maximum, i.e., $k_{\text{max}} = \max(k_{\text{low}}, k_{\text{uni}})$. The final classification y_{approx} is then given by $\text{ancestor}(k_{\text{max}}, t_1)$.

Parameter p

Except for the homology search thresholds and the fall-back method, our classification algorithm is parameter-free. For evaluation and comparison purposes, we introduce a parameter p to trade off sensitivity against specificity of the taxonomic classification. It is used to artificially increase or decrease the score of the metagenomic sequence in the reciprocal phase, i.e., $\text{score}_{\text{new}}(r_{\text{query}}) = \min(p \cdot \text{score}(r_{\text{query}}),\ \text{score}(r_1))$. For example, values of $p > 1$ will increase sensitivity and decrease specificity of the classifications. The parameter is suited only for small changes in the sensitivity-specificity trade-off because the fall-back method is not affected by the parameter.

HMMER Variant

It is also possible to apply the same classification technique within the context of HMMER3-based homology searches against the Pfam database [49].

For convenience, some of the previous notations are reused. Let h be a pairwise alignment, $\mathrm{q}(h)$, $\mathrm{s}(h)$ and $\mathrm{tax}(h)$ are defined analogously. The value $\mathrm{score}(h)$ is given by computing a similarity score over the pairwise alignment with the BLOSUM62 score matrix [67]. The first step is to translate all six reading frames of the metagenomic sequence into protein sequences and to search them against Pfam-A using `hmmscan`. If there is no significant match, the metagenomic sequence is classified as "unknown". Otherwise, let $\hat{q}$ be the aligned sequence of the match with the lowest Pfam-HMM E-value. Then, $\hat{q}$ is aligned against the full multiple alignment of the Pfam family using `hmmalign`. Let q^* be the alignment row corresponding to $\hat{q}$ and let $F = \{f_1, \dots f_{|F|}\}$ be the set of alignment rows of the Pfam family members of the full multiple alignment.

The next step is similar to the BLAST approach, where the closest homolog of the (translated) metagenomic sequence $\hat{q}$ is searched for: For each pair in $\{(q^*, f) \mid f \in F\}$, a pairwise alignment is obtained where columns that correspond to leading and trailing gaps of q^* as well as columns in which both sequences have a gap are discarded. Pairwise alignments that are too short or have too low a score will not be considered for further processing.

Let $B = \{h_1, \dots, h_{|B|}\}$ be the set of all these pairwise alignments, such that $\mathrm{score}(h_1) \geq \dots \geq \mathrm{score}(h_{|B|})$. The reciprocal search is performed by computing the pairwise similarity between $\mathrm{s}(h_1)$ and all other Pfam family members. The following steps, the creation of intervals and the classification are performed in the same way as for the BLAST variant. The alignment parameters that are needed for the fall-back method can easily be computed by counting the number of identities, positives and gaps in the alignment.

Since HMMER3 does not support DNA to Protein alignments yet, frameshifts cannot be detected directly. This decreases both, the sensitivity of homology detection and the classification accuracy. In order to incorporate frameshifts, it is possible to add to the default six reading frame translations the BLASTx-based translation $\mathrm{q}(h_1)$ if available. In this case, seven translations, instead of six, are searched against Pfam-A.

Functional Classification

An important feature of the HMMER variant is the functional classification of metagenomic reads based on Gene Ontology Identifiers (GO-Ids) [8]. The Gene Ontology provides a controlled vocabulary for gene products, distinguishing between their associated biological processes, cellular components and molecular functions. A metagenomic sequence that has a significant match to some Pfam family can then be classified by the set of GO-Ids that are assigned to this Pfam family.

Taxonomic Classification using the NCBI-NT Database

The BLAST variant as described above can also be performed using the NCBI NT nucleotide database instead of using the NCBI NR protein database. In this case, BLASTn is used for the homology search as well as the reciprocal search. An advantage of this variant is that also non-protein encoding metagenomic sequences can be classified if homologous sequences are available as reference. This variant is indicated in Figure 3.1 as path "DNA → BLASTn → Nucleotide(NT) → BLASTn → Classification".

Taxonomic Classification of Amino Acid Sequences

Both, the BLAST and the HMMER variants of CARMA3 can also be used for the taxonomic classification of amino acid sequences. In the case of the BLAST variant of CARMA3, BLASTx is replaced by BLASTp. In the HMMER variant, the amino acid sequences are passed directly to HMMER3, in contrast to DNA that first requires translation into six reading frames. These variants are represented in Figure 3.1 by the paths "Protein → BLASTp → Protein(NR) → BLASTp → Classification" and "Protein → HMMER3 → Pfam → Pfam alignment → Classification".

Taxonomic Classification of 16S-rDNA Sequences

CARMA3 uses 16S-rRNA sequences from the Ribosomal Database Project (RDP) [76] as reference for the taxonomic classification of 16S-rDNA and 16S-rRNA sequences. Similar to the BLAST variant of CARMA3 for protein-encoding DNA sequences, the 16S-rDNA query sequence is searched against the RDP database using BLASTn. The reciprocal search step is conceptually the same, but a fall-back method is not available here. In addition to the reciprocal search step via BLASTn, we have implemented a reciprocal search step that is based on 16S-rRNA alignments provided by RDP. This is similar to the HMMER variant of CARMA3, where pairwise distances between the most similar alignment sequence $s(h_1)$ and all other sequences in the alignment are computed. The scores for a match (1), mismatch (-3), gap opening (-1), and gap extension (-1) used for the computation of the pairwise distances between the 16S-rDNA sequences are the same as the default scores used by BLASTn. These two variants refer to the paths "16S-rDNA → BLASTn → 16S-rRNA(RDP) → BLASTn → Classification" and "16S-rDNA → BLASTn → 16S-rRNA(RDP) → RDP alignment → Classification" in Figure 3.1.

3.3 Conclusion

In this chapter, we have introduced CARMA2 and CARMA3. The complete source code of CARMA2 (Perl) and CARMA3 (C/C++) has been released under the GPL and is available for download at the WebCARMA homepage. CARMA3 is a novel method for the taxonomic classification of metagenomic sequences and is used by WebCARMA, which is described in the next chapter.

Our new method implemented in CARMA3 is inspired by techniques introduced by SOrt-ITEMS, e.g., the reciprocal search step and the adaptation of the taxonomic assignment level based on various BLAST alignment parameters. The evaluation experiments in Chapter 5 show that CARMA3 outperfoms SOrt-ITEMS regarding the accuracy of taxonomic classifications. The reason for this is that CARMA3 combines these techniques in a different way than SOrt-ITEMS. We believe that our method works because reciprocal hits provide a reasonable estimation of the last common ancestor of the metagenomic sequence and its best hit in the sequence database. In contrast to SOrt-ITEMS and MEGAN, our method is not based on the LCA and therefore does not discard reciprocal hits that can provide valuable information for the taxonomic classification. In addition, SOrt-ITEMS always applies the method of adaptation of the taxonomic assignment level based on various BLAST alignment parameters, although this is a method which does not make differences between highly conserved and highly variable genes. Both, this method and the LCA method, limit the discriminative power of the reciprocal search for finding a good taxonomic assignment. Nevertheless, we think that the method of adaptation of the taxonomic assignment level of SOrt-ITEMS is justified in cases where the reciprocal search is not able to make a taxonomic classification. Therefore we have implemented this technique as a fall-back scenario in CARMA3.

In addition to the taxonomic classification of metagenomic shotgun sequences, we have implemented two variants of CARMA3 for the taxonomic classification of 16S-rDNA/RNA sequences. In both cases, we used RDP sequences as references, once as a BLAST database and once as RDP alignments. First preliminary evaluation results indicate that the accuracy of these variants do not differ much. A comparison of these variants of CARMA3 with the RDP Classifier (Section 1.2.1) revealed that the latter achieves a significantly better performance regarding accuracy of the taxonomic classifications. We manually compared several individual taxonomic assignments of CARMA3 and the RDP Classifier and found that in most cases the prediction of the RDP Classifier was much closer to the nearest neighbor of the query sequence than the taxonomy of the best BLAST hit, or in case of the alignment variant, the taxonomy of the most similar alignment sequence.

It is known that the best BLAST hit is often not necessarily the nearest neighbor [89]. To our knowledge, the extent of this has not been evaluated for highly conserved sequences like 16S-rDNA. On these data, the RDP Classifier seems to perform better than BLAST, probably due to the k-mer strategy that makes genus assignments based on an averaged k-mer usage of individual species within one genus.

For the future, we can imagine to use a novel homology search step in CARMA3 for 16S-rDNA sequences that is based on a similar concept as used by the RDP Classifier. Given such a homology search step in CARMA3 which performs as accurately as the genus assignment of the RDP Classifier, we believe it is possible that the reciprocal search step, and thus the final taxonomic assignment, is competitive to or better than the Bayesian Classifier method used by the RDP Classifier.

Chapter 4

WebCARMA

A local installation of CARMA3 requires several bioinformatics tools, like BLASTX or the HMMER3 package. Because of these, CARMA3 has high computational demands which make the usage of a high-performance grid inevitable. Therefore, we introduce WebCARMA, a platform-independent web application for the taxonomic and functional classification of unassembled and assembled metagenomic DNA sequences that makes CARMA3 easily accessible to the scientific community.

4.1 Implementation

Originally, WebCARMA was published using CARMA2 as back end. Along with the publication of CARMA3, WebCARMA has been updated to this new version.

The WebCARMA website is built upon an Apache Web Server using Perl and CGI. The CARMA3 pipeline is executed on the compute cluster of the Bielefeld University Bioinformatics Resource Facility at the Center for Biotechnology (CeBiTec) using Sun Grid Engine `http://gridengine.sunsource.net/`.

As depicted in Figure 4.1, the WebCARMA pipeline takes a FASTA file as input and successively calls the BLASTx and HMMER variants of CARMA3. The output files of CARMA3 are further processed and the results are visualized as histograms using gnuplot [213]. Finally, all output files are collected for download in a compressed archive file. The output files as well as the processing scripts are described in the next section.

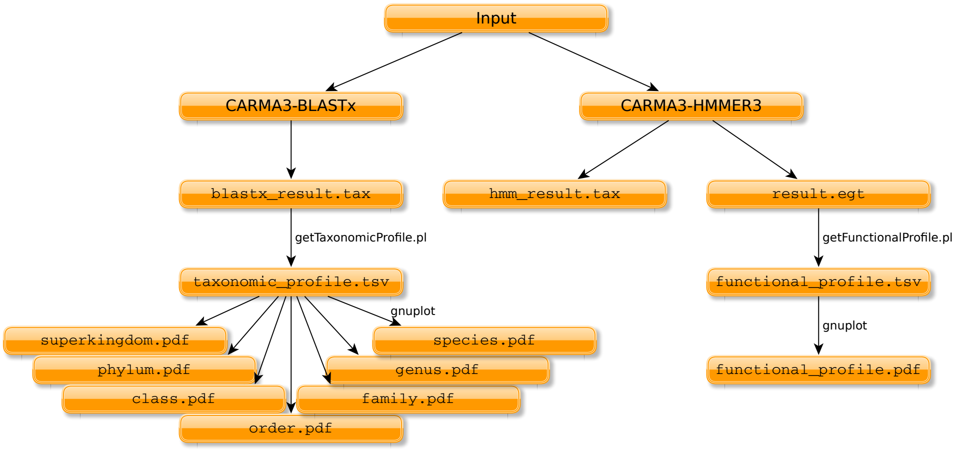
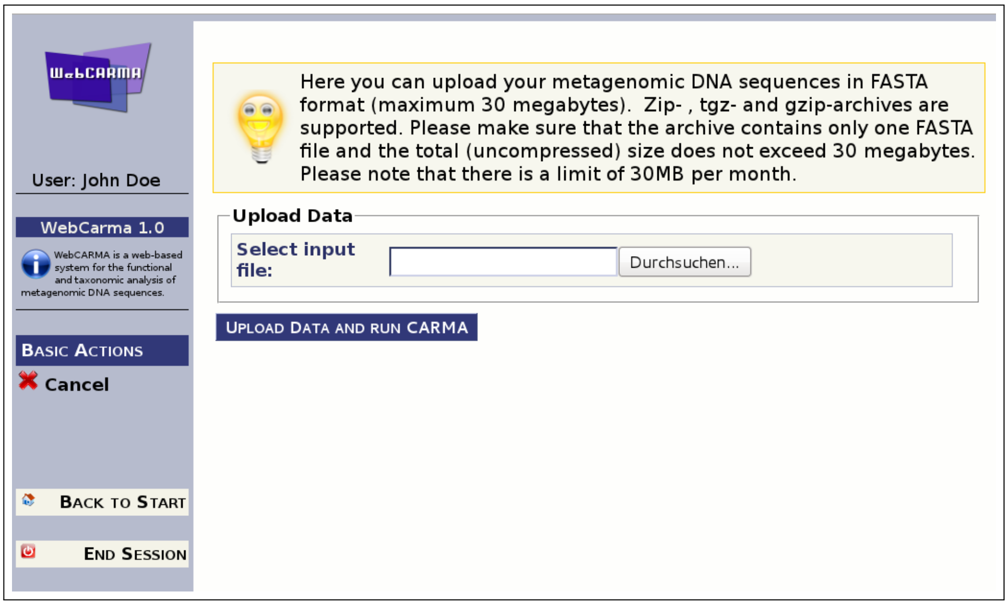


Figure 4.1: Overview of the WebCARMA pipeline.



WebCARMA is a web-based system for the functional and taxonomic analysis of metagenomic DNA sequences.

Here you can upload your metagenomic DNA sequences in FASTA format (maximum 30 megabytes). Zip- , tgz- and gzip-archives are supported. Please make sure that the archive contains only one FASTA file and the total (uncompressed) size does not exceed 30 megabytes. Please note that there is a limit of 30MB per month.

Select input file:

Figure 4.2: Screenshot of the WebCARMA upload form.

4.2 Usage of WebCARMA

In order to use WebCARMA and to upload metagenomic sequences, a user has to register with his e-mail address. An upload form allows to upload the metagenomic sequences (see Figure 4.2). After the uploads are finished, CARMA3 starts with the search for EGTs and the taxonomical classification. By the time the jobs are completed, the user receives an e-mail with a download address pointing to the results.

We provide several data formats that allow the user to explore the results in different ways:

- The translated EGTs with additional information about the name of the original metagenomic sequence, reading frame, Pfam family, HMMER3 E-value and a list of Gene Ontology Identifiers.
- A Gene Ontology term profile in two variants, as a text data file and visualized as a histogram in PDF format.
- The taxonomic classification results of the BLAST and HMMER variants as text data files.
- A taxonomic profile based on the BLAST variant of CARMA3, once as a text data file and once visualized by histograms for each taxonomic rank in PDF format.

The profile data files as well as the classification results are provided in TSV-format (Tab Separated Values), which makes it easy to import the data into other programs (e.g. spreadsheet) for different visualization types or any other further processing. The functional and taxonomic profiles are available in text format and as histograms. More details on the input and output files are given in Appendix A.1.

4.3 Availability

WebCARMA is available under `http://webcarma.cebitec.uni-bielefeld.de`. There are no restrictions to use by non-academics. The upload is restricted to a maximum of 30 MB of FASTA file per user per month.

To date, over 400 users from all over the world have registered at the WebCARMA site and have uploaded over 1,000 metagenomic data sets.

Chapter 5

Experiments

In this chapter we present experiments we have performed to evaluate CARMA3 and the applicability of short reads in metagenomics. In the following, CARMA3$_{\text{BLASTx}}$ denotes the BLASTx variant and CARMA3$_{\text{HMMER3}}$ denotes the HMMER3 variant of CARMA3.

In the first experiment we use simulated metagenomes to compare CARMA3$_{\text{BLASTx}}$, CARMA3$_{\text{HMMER3}}$ and their predecessor CARMA2.1$_{\text{HMMER2}}$ to each other, in the second experiment CARMA3$_{\text{BLASTx}}$, SOrt-ITEMS [65] and MEGAN [78]. In these first two experiments we perform different leave-one-clade-out evaluations (see Chapter 2). In the third and fourth experiment we use different kinds of real metagenomes to evaluate CARMA3. In the fifth experiment we evaluate the applicability of short metagenomic reads for taxonomic classification.

5.1 Evaluation Measures

The taxonomic classification methods assign to a metagenomic sequence one taxon and therefore also one taxonomic rank. This taxon implicitly provides a taxonomic classification also for the higher taxonomic ranks. For example, the taxon *Gammaproteobacteria* at the taxonomic rank class, implicitly provides the taxonomic classification *Bacteria* at the taxonomic rank superkingdom. The taxonomic ranks below the predicted taxon can be considered to be classified as “unknown”. Therefore, for each taxonomic rank a metagenomic sequence can either be correctly classified and counts as a true positive (TP), can be wrongly classified and counts as a false positive (FP), or it is not classified and counts as unknown (U).

In a leave-one-clade-out evaluation, where species below a certain taxonomic rank have been filtered away, it is not possible to obtain true positives below this taxonomic rank. The specificity measure, as described in Chapter 2, is always zero for these ranks and thus cannot be used to measure the ability of a method to avoid false positives. Therefore, we report the raw numbers, TP and FP, instead of the measures sensitivity and specificity in the results of our experiments. As for each taxonomic rank the numbers TP, FP and U sum up to the total number N of reads used in the evaluation, and thus U equals $N - TP - FP$, U will not explicitly be given in the results.

5.2 Metagenomes

For the evaluation of CARMA3 a synthetic and two real data sets were used. The synthetic metagenome (see Appendix Section A.2, Table A.1) was constructed consisting of 25 randomly chosen bacterial genomes from the NCBI ftp site (`ftp://ftp.ncbi.nih.gov/genomes/Bacteria/`). $N = 25,000$ metagenomic reads were simulated using MetaSim [151] with the default 454 sequencing error model resulting in an average read length of 265 bp. The real data set used in Experiment 3 consists of over 600,000 unassembled reads from a biogas plant microbial community [169]. The reads were sequenced with the 454 *Genome Sequencer FLX* system (Roche Applied Science) and have an average length of 230 bp. The real data set used in Experiment 4 consists of 3.3 million non-redundant microbial genes of the gene catalogue of the human gut microbiome [42]. Faecal samples from different individuals were sequenced with the Illumina Genome Analyser (GA) which yielded 576.7 Gb of sequence. The reads were assembled into longer contigs and a gene finder was used to detect open reading frames (ORFs). Similar ORFs were clustered to obtain the final non-redundant gene set. We downloaded this gene set and translated the ORFs into protein sequences using the NCBI Genetic Code 11. The simulated metagenomes and the results of the CARMA3 analyses of the real metagenomes used in the evaluation are available for download at the WebCARMA homepage.

In order to evaluate the applicability of short and ultra-short reads ($\geq$ 35 bp) in metagenomics in Experiment 5, we used the 454 real data set from the biogas plant microbial community described above to create several realistically simulated data sets. In detail, we simulated the short and ultra-short reads by clipping off suffixes of the 454-reads to get the desired read lengths. We generated nine data sets, each consisting of reads of one of the lengths 35 bp, 40 bp, 50 bp, 60 bp, 70 bp, 100 bp, 150 bp, 200 bp, and 250 bp, respectively.

5.3 Databases

To evaluate the different BLAST-based methods regarding their ability to classify sequences of unknown source organism, three BLAST NR protein databases were created: "order filtered", without sequences from species that share the same order as any of the species from the synthetic metagenome, "species filtered", without sequences from species in the synthetic metagenome, and "all", the complete NR database.

Similarly, for $\text{CARMA3}_{\text{HMMER3}}$, the curated Pfam-A database from Pfam 24.0 was used to create the three databases, "order filtered", "species filtered" and "all", by removing corresponding sequences from the full multiple alignments.

5.4 Parameter Settings

The BLASTx runs for $\text{CARMA3}_{\text{BLASTx}}$, SOrt-ITEMS and MEGAN were performed with default E-value threshold (`-e 10`), soft sequence masking (`-F "m S"`), and frameshift penalty 15 (`-w 15`). To ensure comparability, $\text{CARMA3}_{\text{BLASTx}}$ used the same thresholds as SOrt-ITEMS regarding the BLASTx hits, a minimal bit score of 35 and a minimal alignment length of 25. For our first experiment, the CARMA3 parameter p was set to 1. For the second experiment, p was set differently for each of the three databases, since for $p = 1$, CARMA3$_{\text{BLASTx}}$ has fewer true positives and fewer false positives than SOrt-ITEMS (except for taxonomic rank superkingdom). In order to be comparable, p was chosen for the order and the species filtered databases such that CARMA3$_{\text{BLASTx}}$ had about the same number of true positives as SOrt-ITEMS on the lowest taxonomic rank that had not been filtered. For the unfiltered database (all), SOrt-ITEMS gave no classifications on the taxonomic rank species. Therefore p was chosen with respect to the taxonomic rank genus. The values of p were 1.024 for order filtered, 1.033 for species filtered and 1.15 for the unfiltered database.

The parameter for the minimal number of reads that are required to report a taxon in SOrt-ITEMS and MEGAN was set to 1 in all experiments. To ensure comparability of MEGAN with the other two BLAST-based methods, the `top percent` parameter was increased from 10 (default) to 15 resulting in more conservative predictions.

CARMA3$_{\text{HMMER3}}$ was run with an E-value of 0.1 for `hmmscan`, a minimal alignment length of 25 and a minimal score of 30 for the pairwise alignments. In Experiment 1 CARMA2.1$_{\text{HMMER2}}$ was run with an E-value of 0.0001 for `hmmpfam`, whereas in Experiment 5 it was run with the default E-value of 0.1 to ensure a high sensitivity for the ultra-short reads.

5.5 Experiment 1: CARMA3$_{\text{BLASTx}}$ vs CARMA3$_{\text{HMMER3}}$

In the first experiment CARMA3$_{\text{BLASTx}}$ and CARMA3$_{\text{HMMER3}}$ were compared with each other in order to see which of both variants provides better taxonomic classification results (Table 5.1). As a third variant the older version CARMA2.1$_{\text{HMMER2}}$ was also included in the comparison.

For the **order filtered database**, CARMA3$_{\text{BLASTx}}$ has more true positives but also more false positives than CARMA3$_{\text{HMMER3}}$ at all taxonomic ranks. In the **species filtered database**, CARMA3$_{\text{BLASTx}}$ has more true positives than CARMA3$_{\text{HMMER3}}$ as before, but this time it also has fewer false positives than CARMA3$_{\text{HMMER3}}$. Similar results are provided for the **unfiltered database**: CARMA3$_{\text{BLASTx}}$ has significantly more true positives and at the same time considerably fewer false positives than CARMA3$_{\text{HMMER3}}$ at all taxonomic ranks. While for the order filtered database it is not obvious which variant should be preferred over the other, for the species filtered and unfiltered databases CARMA3$_{\text{BLASTx}}$ clearly outperforms CARMA3$_{\text{HMMER3}}$.

The comparison of CARMA3$_{\text{HMMER3}}$ and CARMA2.1$_{\text{HMMER2}}$ using the unfiltered database shows that CARMA3$_{\text{HMMER3}}$ is superior to CARMA2.1$_{\text{HMMER2}}$ on all taxonomic ranks from class to genus.

Fraction of fall-back method on the overall classification About $10 - 20\%$ of all metagenomic reads that have been classified with CARMA3$_{\text{BLASTx}}$ were classified using the fall-back method (see Table 5.2). Of these, about one half of the reads were classified with the fall-back method because they had only one BLAST hit in the corresponding database.

Table 5.1: Comparison of the taxonomic classification accuracy of the different CARMA variants CARMA3$_{\mathrm{BLASTx}}$, CARMA3$_{\mathrm{HMMER3}}$ and CARMA2.1$_{\mathrm{HMMER2}}$. Table entries "–" indicate taxonomic ranks where the corresponding species have been filtered away and therefore true positives are not possible.

	order filtered				species filtered				all					
	C3$_{\mathrm{BLASTx}}$		C3$_{\mathrm{HMMER3}}$		C3$_{\mathrm{BLASTx}}$		C3$_{\mathrm{HMMER3}}$		C3$_{\mathrm{BLASTx}}$		C3$_{\mathrm{HMMER3}}$		C2.1$_{\mathrm{HMMER2}}$	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
superkingdom	12282	799	6668	660	20059	113	9563	516	22725	31	11276	544	6099	140
phylum	8532	1094	4194	657	18968	183	8065	377	22626	17	10255	345	5724	238
class	3700	1257	1983	721	15793	274	6329	322	20584	25	8822	223	4969	278
order	–	2019	–	1158	14829	275	5084	367	20869	30	8066	220	4756	385
family	–	926	–	531	11126	239	3400	324	18301	25	6485	223	4149	346
genus	–	144	–	175	6897	427	1852	517	16025	107	5366	303	3487	746
species	–	9	–	25	–	142	–	214	1142	31	809	176	2092	1135

Table 5.2: The total number of reads ("total") classified with CARMA3$_{\text{BLASTx}}$ and the number of reads classified with the fall-back method ("fall-back"). "single" represents the number of metagenomic reads that had only one BLAST hit and "multiple" represents the number of reads with two or more BLAST hits.

	total	fall-back	fall-back	
			single	multiple
order filtered	13081	2668	1397	1271
species filtered	20172	1907	878	1029
all	22756	2203	1159	1044

Performance on different read lengths The performance of CARMA3$_{\text{BLASTx}}$ was also evaluated for other read lengths and different error models. To simulate a metagenome sequenced with 454 GS FLX Titanium, reads were created with the default 454 error model of MetaSim with an average read length of 400 bp. For the 454-GS20 and Illumina sequencing technology, reads of length 80 bp were simulated. The error model for the Illumina reads (`errormodel-80bp.mconf`) was downloaded separately from the MetaSim homepage. As no error model for Illumina reads longer than 80 bp was available, the 454-GS20 reads were adapted to this length. Each of the three simulated metagenomes (454-400 bp, 454-80 bp and Illumina-80 bp) was analysed using the order-, species- and unfiltered protein databases. The results are given in Appendix Section A.3.

In general, the 400 bp reads provide more classifications than the 265 bp. In addition, in many cases the 400 bp reads account for more true positives and fewer false positives than the 265 bp reads. This is the case in the species-filtered database at taxonomic ranks class to family, but also for the unfiltered database at taxonomic ranks superkingdom, family and genus. As expected, the shorter 454-80 bp reads perform worse than the 454-265 bp reads. This is clearly shown for the species-filtered database at taxonomic rank family and the unfiltered database at taxonomic ranks phylum to family.

The comparison of the 454-80 bp and Illumina-80 bp reads shows that Illumina reads are about twice as often classified as the 454 reads for all databases. For the species filtered database at taxonomic rank superkingdom and the unfiltered database at taxonomic ranks superkingdom to genus the Illumina error model clearly outperfoms the 454 error model in terms of accuracy. A comparison of the simulated reads revealed that the 454 error model has produced many more base substitutions than the Illumina error model. In addition, the 454 error model accounts for insertions and deletions, which the Illumina error model does not. It is unclear to the authors how representative the MetaSim default error models are for the currently available sequencers by 454 and Illumina. Therefore, rather than as a comparison of two different sequencing technologies, the comparison of both error models should be understood as a demonstration of the influence of sequencing errors on the accuracy of the taxonomic classification.

5.6 Experiment 2: CARMA3 vs SOrtITEMS vs MEGAN

In the second experiment our new method CARMA3$_{\text{BLASTx}}$ was compared to the two other BLASTx-based methods, SOrt-ITEMS and MEGAN (Table 5.3).

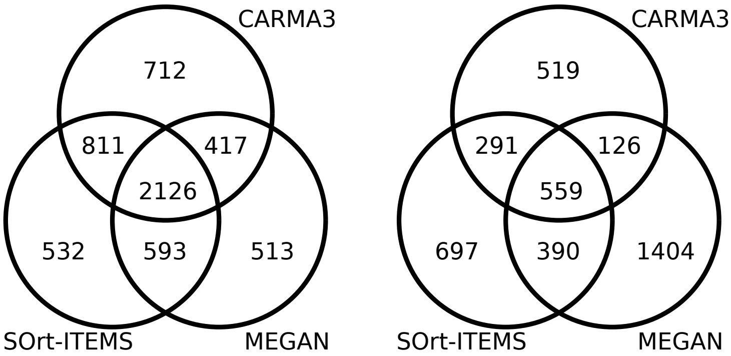
While for the **order filtered database** CARMA3 performs better than SOrt-ITEMS at rank class, for the ranks superkingdom and phylum it is not clear which method is better. At the taxonomic ranks order to genus, where the metagenomic sequences have been filtered away, CARMA3 has much fewer ($\approx$ 37% $-$ 74%) false positives than SOrt-ITEMS. CARMA3 has better results than MEGAN at all taxonomic ranks, while SOrt-ITEMS has better results than MEGAN at all taxonomic ranks below superkingdom. For the **species filtered database** CARMA3 has better results than SOrt-ITEMS and MEGAN at taxonomic rank genus. For the other taxonomic ranks the results of CARMA3 and SOrt-ITEMS are not comparable, since SOrt-ITEMS has more true positives and more false positives. Only at taxonomic rank species CARMA3 has false positives which SOrt-ITEMS does not have. The reason for this is that SOrt-ITEMS requires a minimal alignment length of 550 bp in order to make classifications at the taxonomic rank species, but the simulated metagenome contains only reads with an average length of 265 bp. The advantage of avoiding false positives at rank species in the order and species filtered databases is traded off against the disadvantage of not detecting species in the unfiltered database. CARMA3 performs better than MEGAN at all taxonomic ranks, except superkingdom, where the results are not comparable. To provide comparability between the methods also for the **unfiltered database** we tried to increase the number of true positives of CARMA3 at the taxonomic rank genus. We were able to increase the number of true positives by 4,405 from 16,025 (Table 5.1) to 20,430, but not higher. The reason for this is that classifications of reads from the fall-back method can not be changed with the parameter p. Although CARMA3 performs worse than SOrt-ITEMS at three taxonomic ranks (superkingdom, order and family) in the unfiltered data set, the corresponding TP and FP numbers at each taxonomic rank except species are quite similar. CARMA3 is able to detect many species where SOrt-ITEMS does not detect any. On all taxonomic ranks, except ranks genus and species, CARMA3 and SOrt-ITEMS perform better than MEGAN.

Non-protein coding sequences Assuming that about 10% to 20% of microbial genomes are non-protein coding sequences [154], it is clear that many of the metagenomic reads can not be classified using protein homology information. But because many of these reads do overlap at least partly with a coding region, it can be observed that 92% to 96% of the reads are correctly assigned to bacteria by the BLASTx based methods using the unfiltered database.

Overlap Figure 5.1 shows Venn diagrams for the overlap of (a) correct and (b) wrong classifications for the order-filtered data set at taxonomic rank class. Although each method has about 3,600-4,000 correct classifications, only about 2,100 reads have been correctly classified by every method. In this particular case each of the three compared methods correctly classifies a significant proportion of the reads, which the other methods do not. However, for higher taxonomic ranks and the species- and unfiltered data set the overlap of correct

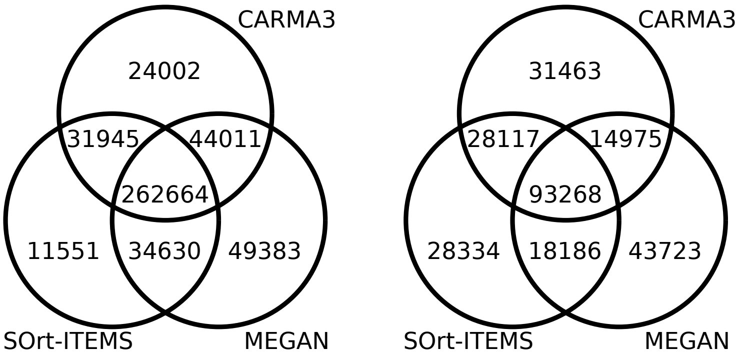
Table 5.3: Comparison of the taxonomic classification accuracy of the different BLASTx-based methods CARMA3$_{\text{BLASTx}}$, SOrt-ITEMS and MEGAN. In the table CARMA3 refers to the BLAST variant CARMA3$_{\text{BLASTx}}$.

	order filtered						species filtered						all					
	CARMA3		SOrt-ITEMS		MEGAN		CARMA3		SOrt-ITEMS		MEGAN		CARMA3		SOrt-ITEMS		MEGAN	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
superkingdom	12696	861	12576	786	12626	1849	20266	118	20345	128	20840	453	22890	36	23979	30	23900	105
phylum	8989	1224	9254	1736	8079	1985	19268	227	19466	356	19010	535	22832	30	23909	43	23607	91
class	4066	1495	4062	1937	3649	2479	16206	349	16259	401	15921	735	20932	38	21912	41	21418	107
order	–	2507	–	4011	–	4975	15671	366	15684	535	15105	954	21994	65	22871	58	21543	155
family	–	1186	–	2565	–	4087	12117	345	13104	606	11625	1101	20089	62	20864	59	18937	143
genus	–	210	–	798	–	4041	8328	752	8299	1112	8031	1889	20430	314	21124	483	17758	263
species	–	23	–	0	–	3544	–	995	–	0	–	4346	15232	685	0	0	11786	550



(a) True positives (b) False positives

Figure 5.1: Overlap of 25,000 simulated metagenomic reads classified by CARMA3, SOrt-ITEMS and MEGAN for the order-filtered data set at taxonomic rank class.



(a) Superkingdom (b) Class

Figure 5.2: Venn diagrams for a biogas plant metagenome with over 600,000 reads. The subset sizes depict the numbers of reads being classified with CARMA3, SOrt-ITEMS and MEGAN at taxonomic ranks superkingdom and class.

classifications is much higher and therefore the differences between the methods are smaller. Figure 5.1(b) shows that the overlap of wrong classifications is relatively smaller than that of the correct classifications. As expected, a high number of wrong classifications are unique to MEGAN. For the Venn diagrams of the other taxonomic ranks and data sets see Appendix Sections A.4–A.6.

5.7 Experiment 3: CARMA3 on 454 Biogas Metagenome

For the evaluation on a real data set of unassembled 454 reads, the metagenome of a biogas plant microbial community was analysed with CARMA3$_{\mathrm{BLASTx}}$, SOrt-ITEMS and MEGAN. Figure 5.2 shows Venn diagrams for the number of reads being classified at taxonomic ranks superkingdom and class (see Appendix Section A.7 for the other ranks). Reads that are classi-

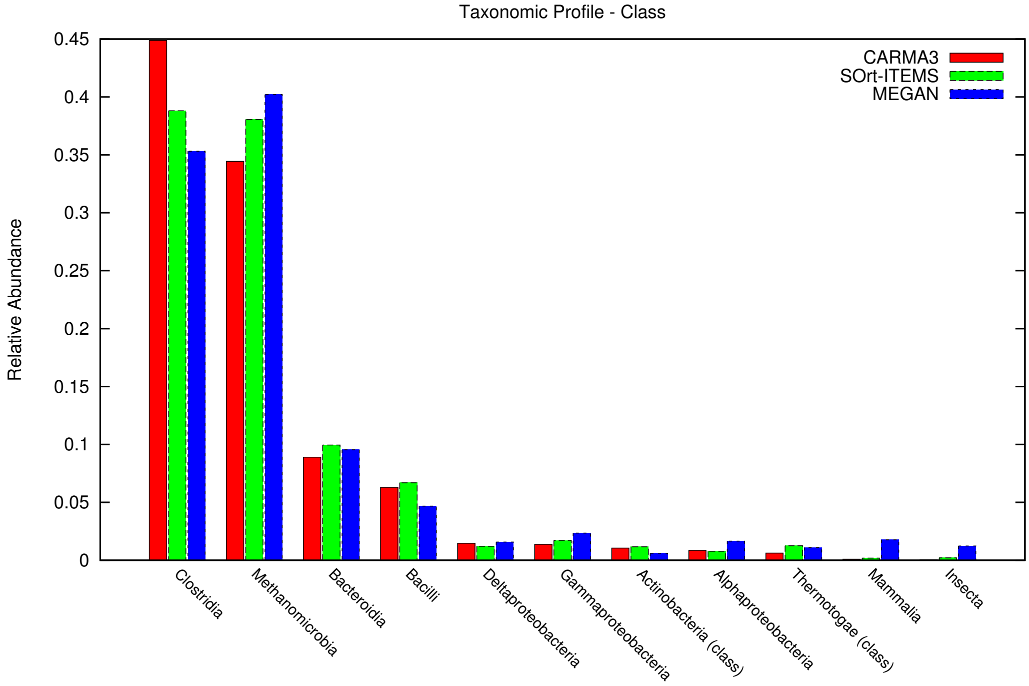


Figure 5.3: Comparative taxonomic profile of a biogas plant metagenome analysed with CARMA3, SOrt-ITEMS and MEGAN at taxonomic rank class.

fied by two or all methods are not necessarily assigned to the same taxon. The Venn diagrams show that the fraction of reads that are classified by all methods is bigger at higher taxonomic ranks than on lower taxonomic ranks. For a qualitative comparison of the taxonomic classifications of the three methods, comparative taxonomic profiles for each taxonomic rank have been created. Figure 5.3 shows the profile for taxonomic rank class, Appendix Section A.8 contains the full set of profiles. In order to restrict the number of taxa in the taxonomic profiles to the most abundant ones, all taxa with a relative abundance smaller than 0.01 were discarded. Taxa for which any of the classification methods predicted an abundance of 0.01 or higher were not discarded. After this threshold was applied, the remaining taxa were normalised such that the relative abundances sum up to one for each of the methods ensuring comparability between the methods. In contrast to the profiles of the other taxonomic ranks, the profile of taxonomic rank superkingdom includes the relative abundance of reads that have been classified as “unknown”.

The comparative taxonomic profiles reveal a strong consistency between the compared methods regarding the relative abundances of the most abundant taxa. Only at taxonomic ranks genus and species, bigger differences can be found: CARMA3 predicts more *Clostridia*, SOrt-ITEMS more *Methanocullei* and MEGAN predicts more *Cloacamonas*. The reason for the high consistency between the three methods above taxonomic rank genus is that low abundant species have been filtered away. Filtering of low abundant taxa provides a trade-off between filtering noise produced by false positives and the detection of low abundant true positive taxa. Table A.5 in Appendix Section A.8 shows how many reads of each method have

Table 5.4: Running times for the homology searches (BLASTx and HMMER3) and the taxonomic classifications with CARMA3, SOrt-ITEMS and MEGAN.

	CARMA3	SOrt-ITEMS	MEGAN
BLASTx	54 h 15 min	54 h 15 min	54 h 15 min
-classification	52 min 22 s	12 min 36 s	3 min 4 s
HMMER3	6 h 20 min	-	-
-classification	41 min 8 s	-	-

been filtered away. For example at taxonomic rank order, about 7% of all reads classified by CARMA3, 11% of all reads classified by SOrt-ITEMS and about 28% of all reads classified by MEGAN have been filtered away. This effect and the differences between the methods are even stronger at lower taxonomic ranks. The results of the evaluation in Experiment 2, showing that SOrt-ITEMS and in particular MEGAN produce more false positives than CARMA3, are an indication that most of the filtered taxa in this data set are actually wrong predictions rather than truly low abundant taxa.

This biogas plant metagenome has formerly been analysed using two different approaches, (a) construction of bacterial and archaeal 16S-rDNA amplicon libraries and (b) screening for reads in the 454 data set that encode for 16S-rDNAs [94]. Both 16S-rDNA approaches and our results coincide in the identification of the main abundant taxa. For example at taxonomic rank order, the archaea *Methanomicrobiales* and the bacteria *Clostridiales* and *Bacteroidales* have by all approaches been predicted as the main abundant taxa. Apart from these consistent predictions, the differences in the relative abundances of the other taxa might also be explained by various biases that are inherent to the compared methods. For example, the database reference sequences come mainly from culturable species and therefore are biased towards certain bacterial phyla [76]. On the other side, the oligonucleotide primers that are used to amplify the 16S-rDNA can exhibit substantial variations in their specificity towards different clades [9]. Considering these potential biases, the taxonomic classifications of the BLASTx-based methods show a high consistency with the results of the 16S-rDNA analyses.

Running times To determine the running time of our method 10,000 metagenomic reads from the biogas plant metagenome with the complete CARMA3 pipeline were analysed. For comparative purposes, the running times of SOrt-ITEMS and MEGAN were also measured. The computation was conducted on a 2.5 GHz Intel Core 2 Duo processor with 8 GB RAM, running Linux (64-bit Ubuntu 10.04, kernel version 2.6.35.23). The observed running times, measured with the GNU time command (user+sys), are given in Table 5.4.

The results show that for the BLAST-based classifications, the BLAST homology search accounts for more than 98% of the total running time. Among the three BLAST-based classification methods, MEGAN is the fastest method, more than 4 times faster than SOrt-ITEMS. SOrt-ITEMS in turn is about 4 times faster than CARMA3$_{\text{BLASTx}}$. In contrast to MEGAN, CARMA3$_{\text{BLASTx}}$ and SOrt-ITEMS spend additional time on performing reciprocal BLAST searches and therefore are slower. CARMA3$_{\text{BLASTx}}$ is slower than SOrt-ITEMS because it does not use a top-percent filter and therefore creates bigger BLAST databases in the reciprocal search step.

To measure the time needed to run BLASTx on shorter Illumina reads, 10,000 75 bp-reads sequenced with Illumina Genome Analyser (GA) from a human gut microbial community [144] were searched against the full NR protein database. The running time of about 14.5 hours for the BLASTx run is in terms of bases per second quite similar to the running time of the BLASTx analysis of the 454 data. While a BLASTx analysis of a complete 454 run is feasible on a compute cluster in the order of hours or a few days, this approach seems to be less practical for the analysis of all unassembled reads produced by a complete run of an Illumina sequencing machine that produces one to two orders of magnitudes more bases in total than a 454 sequencing machine in a single run. The usage of data reduction techniques, as shown in Experiment 4, can be a way to overcome this limitation.

5.8 Experiment 4: CARMA3 on Assembled Illumina Data

Data reduction techniques are a common method to handle the amount of data produced by Illumina sequencing machines [144, 70]. Typical steps involve the assembly of reads into longer fragments, gene detection with a gene finder to detect open reading frames (ORF), clustering of highly similar ORFs, and translation of the non-redundant ORFs into protein sequences. Such a metaproteome has, in contrast to the full set of unassembled Illumina reads, a size that makes the analysis with the BLASTp variant of CARMA3 possible on a compute cluster in the order of hours or a few days.

To evaluate the applicability of CARMA3 on amino acid sequences derived from assembled Illumina reads, the BLASTp variant of CARMA3 was used to analyse the gene catalogue of the human gut microbiome [144]. Figure 5.4 shows the results of this analysis at taxonomic rank genus. The profiles of the other taxonomic ranks can be found in Appendix Section A.9, Figures A.16–A.22. These results were compared to the taxonomic classification of another study of the human intestinal microbial flora based on 13,355 prokaryotic 16S ribosomal RNA gene sequences [42].

Both methods, the 16S-rDNA analysis and CARMA3, identify *Firmicutes* and *Bacteroidetes* as the most abundant phyla, followed by *Proteobacteria*, *Actinobacteria*, *Verrucomicrobia*, and *Fusobacteria*. Also, in both analyses the phylum *Firmicutes* consists mainly of the class *Clostrida*. Nearly all genera of the *Clostridia* that have been predicted by the 16S-rDNA analysis, like *Eubacterium*, *Ruminococcus*, *Dorea*, *Butyrivibrio*, and *Coprococcus*, have also been predicted by CARMA3 (Appendix Figure A.23). Also most of the species of *Clostridia* like *E. rectale*, *E. hallii*, *R. torques*, *R. gnavus*, *F. prausnitzii*, *D. formicigenerans*, and *D. longicatena* that are found by the 16S-rDNA analysis could be confirmed by CARMA3 (Appendix Figure A.24). However, the species *E. hadrum* and *R. callidus* that have been found by 16S-rDNA were not found by CARMA3. The genus *Clostridium* which is the taxon found by CARMA3 to have the highest abundance in the class *Clostrida* is not reported by the 16S-rDNA analysis. The reason for this might be that the 16S-rDNA sequence of *Clostridium bartlettii*, which mostly contributes to the genus *Clostridium* and is known to be found in human faeces, might not have been available at the time of the 16S-rDNA analy-

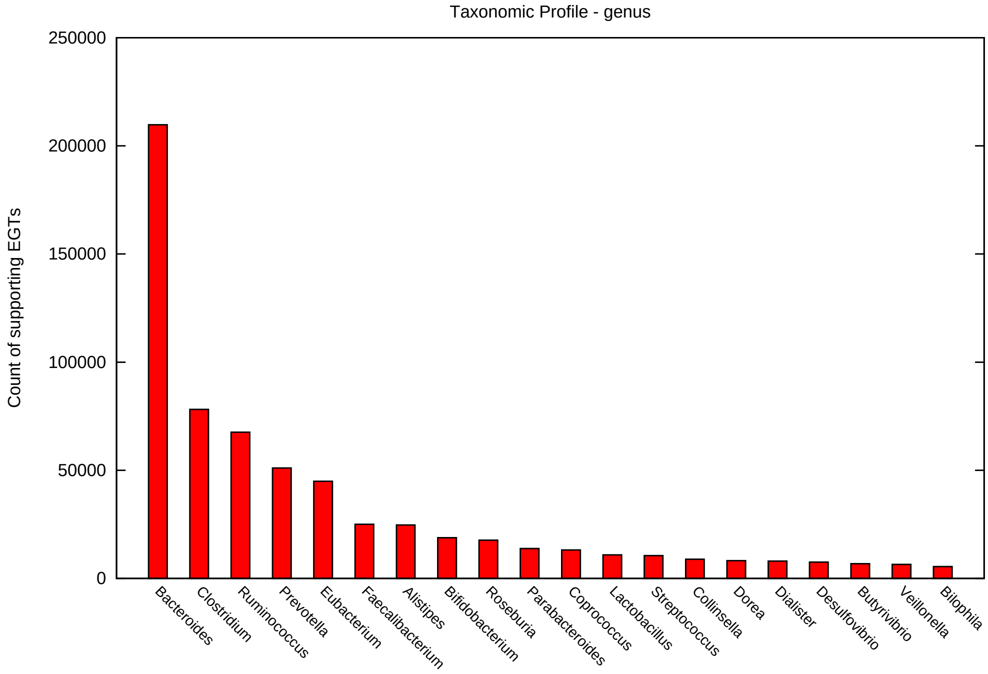


Figure 5.4: The 20 most abundant taxa of the human gut microbial gene catalogue at taxonomic rank genus.

sis [180]. Also the species *R. inulinivorans* and *R. intestinalis* of the genus *Roseburia*, which are found by CARMA3 but not by the 16S-rDNA analysis, are known to occur in human faeces [174, 39]. For the second most abundant phylum, the *Bacteroidetes*, the authors of the 16S-rDNA analysis report a high variability in the distribution of phylotypes in samples from different subjects. Nevertheless, all phylotypes reported by the authors of the 16S-rDNA analysis, *B. vulgatus*, *Prevotellaceae*, *B. thetaiotaomicron*, *B. caccae*, and *B. fragilis*, were among the 11 or, in case of *B. putredinis*, among the 22 most abundant taxa predicted by CARMA3 (Appendix Figures A.25 and A.26).

The comparison of the taxonomic predictions of the 16S-rDNA analysis and CARMA3 has revealed a high consistency in the results of both methods. This shows that CARMA3 can also be used for the taxonomic classification of amino acid sequences obtained from assembled Illumina reads.

5.9 Experiment 5: Applicability of Short Reads for Taxonomic Classification

A special challenge in metagenomics is the fact that the new sequencing techniques produce short (100-500 bp with 454) and ultra-short (35 bp with SOLiD, 35-100 bp with Illumina, 30-35 bp with Helicos [142]) reads. New bioinformatic tools have to be developed that can cope with both, the huge amount of data and the short read lengths. Especially the short read lengths have been considered the main bottleneck for the usage of ultra-short reads in metagenomics. Recent analyses, based on BLAST-searches, indicated a low prediction quality for short reads < 400 bp [217]. In contrast, Krause *et al.* [93] showed on a synthetic metagenome that even with reads as short as around 100 bp, high accuracy predictions with an average false positive rate of 0.1 to 2.5 percent are possible.

In this experiment we simulated short and ultra-short reads, as described in Section 5.2, to evaluate if reads, as short as 35 bp, can be used for taxonomic classification of a metagenome with CARMA2. It is known that some sequencing techniques exhibit correlations between read coverage and GC content [13, 38, 158]. By using simulated reads instead of real metagenomic reads we can be sure that any differences we see in the classification results between the data sets are only due to the different read lengths. If there is a bias in the 454 data, then we also have the same bias in the simulated data sets and our comparison should not be much affected by this.

First, we analyze the number and lengths of the EGTs obtained for each data set, then we compare the taxonomic classification results for the different read lengths.

As shown in Table 5.5, the number of reads in each data set decreases with increasing read length. This is because the 454 data set contains reads of different lengths and some of the reads are already too short to serve as a template for all simulated data sets. The relative amount of EGTs that is found in each data set increases with read length. Figure 5.5 shows the EGT length distribution in each data set as a function of read length. Shown are the minimum, 25% quantile, median, 75% quantile and the maximum.

Our results show that the median EGT length does not scale linearly with the read length. The length of Pfam families and domains poses an upper bound on the possible length of local alignments between translated reads and Pfam families. The longer a read is, the higher is the probability that parts of the reads lie outside of the matching gene and can not contribute to the EGT. In rare cases, it happens that an EGT is one amino acid longer than its read length divided by three. For example, in the set of EGTs produced from the 150 bp-reads data set, the longest EGTs are 51 amino acids long. This can occur when `BLASTx` predicts two frameshifts in one read.

For our analysis of the applicability of ultra-short reads with CARMA2, we considered seven different taxonomic ranks: superkingdom, phylum, class, order, family, genus and species. A first relative abundance for each taxon is obtained by dividing the absolute number of EGTs that predict this taxon by the total number of EGTs at the same taxonomic rank. The latter do not include EGTs that were assigned the taxonomic status “unknown”. We consider taxa with a relative abundance below the threshold 0.015 in all data sets, to be false positives. Therefore they are classified as “other”.

After applying the threshold we recompute the relative abundances for each taxon, this time subtracting both, “unknown” and “other” from the total number of EGTs at the same taxonomic rank. With this, we have normalized the relative abundances for the taxa such that they sum up to 1 and therefore ensured comparability between the data sets.

For scaling reasons, the fractions of “unknown” and “other” EGTs are not shown in the histograms (except “unknown” on superkingdom level). This data is given in Tables 5.6 and 5.7.

Table 5.5: Number of reads and EGT yield for each data set. Some metagenomic reads have matches to more than one Pfam family and therefore are translated into more than one EGT. The row “Unique” denotes the total number of EGTs where EGTs from the same read are counted only once. The row “Yield” denotes the fraction of (unique) EGTs that could be obtained from the corresponding data set.

Length	35 bp	40 bp	50 bp	60 bp	70 bp	100 bp	150 bp	200 bp	250 bp	original
Reads	616 069	616 031	613 943	606 760	598 811	584 168	550 945	492 305	297 852	616 072
EGTs	886	7 836	29 999	48 472	62 112	92 000	119 674	130 544	89 979	172 461
Unique	886	7 827	29 923	48 218	61 687	90 854	116 743	125 624	85 565	164 444
Yield	0.14 %	1.27 %	4.87 %	7.95 %	10.30 %	15.55 %	21.19 %	25.52 %	28.73 %	26.69 %
Yield(non-unique)	0.14 %	1.27 %	4.89 %	7.99 %	10.37 %	15.75 %	21.72 %	26.52 %	30.21 %	27.99 %

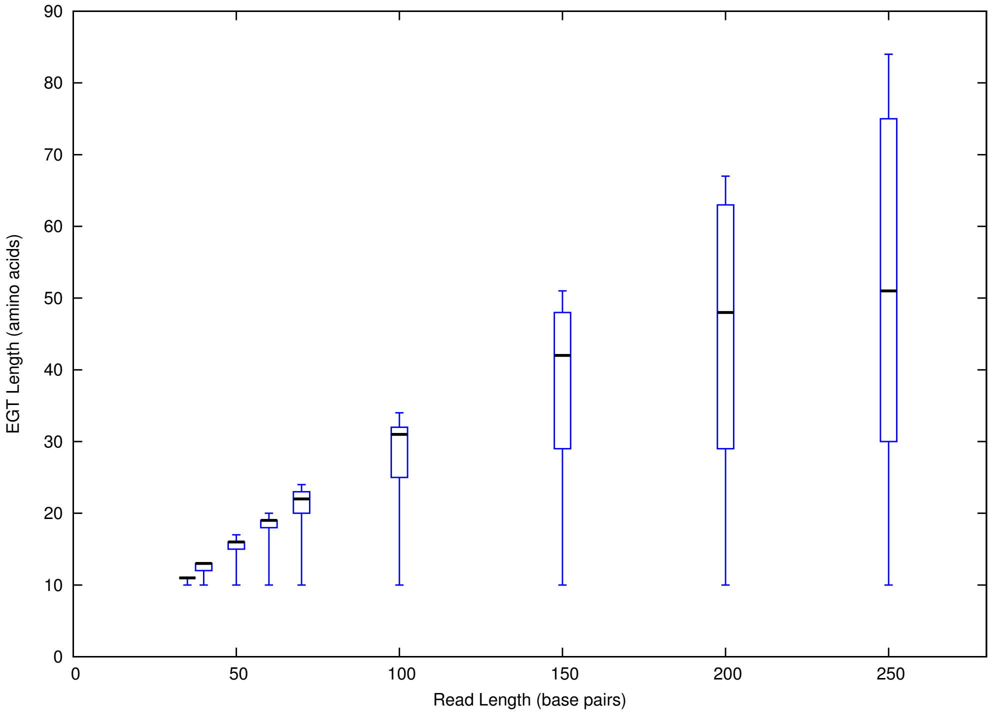


Figure 5.5: EGT length distribution in each data set as a function of read length. Shown are the minimum, 25% quantile, median, 75% quantile and maximum.

Table 5.6: Rate of "Unknown" EGTs that could not be classified further from the complete set of EGTs.

Read Length	Superkingdom	Phylum	Class	Order	Family	Genus	Species
35	0.09	0.31	0.38	0.45	0.52	0.53	0.59
40	0.09	0.26	0.37	0.43	0.51	0.52	0.57
50	0.09	0.27	0.38	0.43	0.51	0.52	0.58
60	0.09	0.28	0.39	0.45	0.53	0.54	0.61
70	0.09	0.29	0.4	0.46	0.54	0.56	0.63
100	0.1	0.32	0.43	0.49	0.58	0.6	0.68
150	0.11	0.33	0.44	0.52	0.6	0.62	0.71
200	0.11	0.34	0.45	0.52	0.61	0.63	0.73
250	0.11	0.32	0.44	0.51	0.6	0.63	0.73

Table 5.7: “Other” are EGT’s with a relative abundance below the threshold 0.015 and are not shown in the histograms. Here we show the rates of “Other” EGTs relative to the total number of classified EGTs for each taxonomic rank and data set.

Read Length	Superkingdom	Phylum	Class	Order	Family	Genus	Species
35	0.0011	0.0671	0.1651	0.2816	0.4057	0.4554	0.6776
40	0.0011	0.0678	0.1691	0.2901	0.4388	0.5056	0.7340
50	0.0019	0.0667	0.1609	0.2954	0.4591	0.5292	0.7606
60	0.0024	0.0619	0.1552	0.2864	0.4637	0.5302	0.7663
70	0.0023	0.0617	0.1554	0.2954	0.4535	0.5221	0.7505
100	0.0035	0.0655	0.1539	0.2891	0.4456	0.4978	0.7172
150	0.0071	0.0692	0.1565	0.2964	0.4555	0.5006	0.6756
200	0.0100	0.0651	0.1467	0.2938	0.4500	0.4954	0.6658
250	0.0137	0.0542	0.1377	0.2849	0.4317	0.4693	0.6364

Even though the taxonomic predictions on lower taxonomic ranks (order, family, genus and species) are known to be imprecise, we included them in our experiment in order to study the effect of using (ultra-)short reads compared to longer ones at all taxonomic ranks.

Figure 5.6 shows the results at taxonomic rank species. The complete set of figures for the evaluation at all taxonomic ranks can be found in Appendix Section A.10. The results show that CARMA2 predicts for all data sets and all taxonomic ranks the same taxa. For higher taxonomic ranks, even the relative abundance levels are similar between the different data sets. Deviations of 35 bp-reads for example can be seen on the level of order, where significantly more of *Thermotogales* and *Haemosporida*, and less of *Thermoanaerobacterales* are predicted. The 40 bp data set does not show these differences. Even more deviations can be found on lower taxonomic ranks, for example species.

Furthermore, as expected, the rate of EGTs that are not classified increases for lower taxonomic ranks for all data sets (Table 5.6). Interestingly, the rate of unclassified EGTs is smaller for shorter reads than for longer reads. This might be due to the circumstance that shorter EGTs need to have more sequence similarity to the Pfam families than longer EGTs, in order to achieve the same E-value threshold.

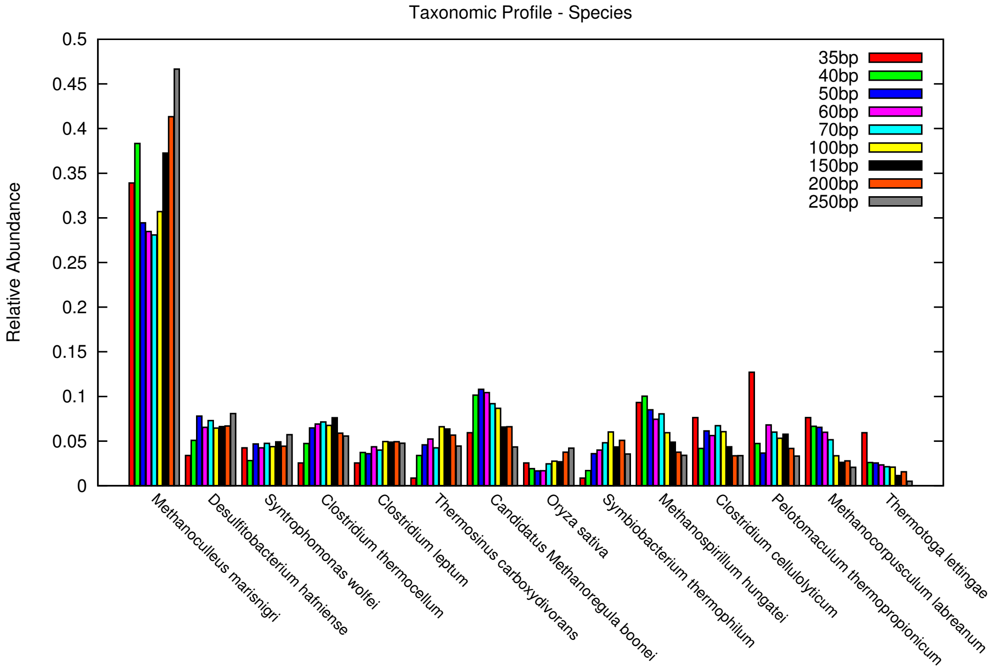


Figure 5.6: Taxonomic results on the level of species. Only taxa with an abundance of 0.015 or higher are shown.

Chapter 6

Conclusion and Outlook

Metagenomics is a relatively new field of research on natural microbial communities containing uncultured microbes. New methods that involve construction of metagenomic libraries and shotgun sequencing of whole metagenomes have helped to improve understanding the microbial world. Understanding microbes is not only important because they affect our health and can be used in industrial applications, microbes also have the potential to contribute to economically feasible and socially desirable alternatives for energy production and resource usage to solve the environmental and economical problems mankind faces.

In recent years, metagenomics has been spurred by the development of next-generation sequencing technologies. Despite its success in the analysis of microbial communities, there still remain many quality problems: DNA extraction, filtering, the sequencing protocol, and final sequencing produce various kinds of biases [38, 71, 113, 171]. Computational analyses that are based on comparisons with known sequences are intrinsically biased due to the fact that sequence databases contain mainly sequences from cultured species. The vast amount of data and short read lengths produced by sequencing technologies pose a challenge for the management and computational analysis.

In this thesis we have mainly been interested in computational methods to determine the taxonomic origin of metagenomic DNA sequences. Several existing methods have been reviewed in Chapter 2. The most critical issues of these methods are speed and accuracy of the taxonomic classification. Our main contribution in this field is the development of a novel method that is more accurate than other comparable methods while achieving competitive running times if the BLAST search is included in the time measurement.

With CARMA3 we have introduced a new method for the taxonomic classification of assembled as well as unassembled metagenomic sequences that can be used in combination with BLAST- and HMMER-based homology searches. Our method is able to classify protein-encoding DNA sequences, protein sequences and 16S-rDNA/RNA sequences. Except for the homology search and the fall-back scenario, our method is parameter-free. In addition, for the HMMER-based variant, our method also provides a functional classification of metagenomic sequences and therefore allows for the characterization of species composition and genetic potential of microbial samples.

Typically, a metagenomic sample contains many novel species that have not been sequenced before. We have simulated such a scenario with the order filtered database and have shown that in most cases CARMA3 not only performs better than existing BLAST-based methods, but most strikingly, it is better at avoiding false positive predictions on lower taxonomic ranks when only remote homologs are available for the classification of novel species.

As already pointed out in Section 3.3, we think that our method outperforms the other BLAST-based methods because reciprocal hits provide a reasonable estimation of the last common ancestor of the metagenomic sequence and its best hit in the sequence database. In contrast to the other methods, CARMA3 is not based on the LCA and therefore does not discard reciprocal hits that can provide valuable information for the taxonomic classification.

CARMA3 uses both, BLAST and HMMER3 for the taxonomic classification of metagenomic reads. One of the reasons we developed the HMMER3 variant was the idea that we could improve the speed of the reciprocal search by first finding the corresponding protein family with HMMER3 and then restricting the search of reciprocal hits to this smaller set of sequences from the same family. Indeed, for the future we plan to further increase the speed of CARMA3$_{\text{HMMER3}}$ by using BLASTp to search for the reciprocal hits within the protein family instead of computing the pairwise alignments for every Pfam family member. However, in nearly all cases the BLASTx-based variant classified significantly more reads than the HMMER3-based variant. In many cases it also had fewer false positives. Therefore we think that the BLASTx-based variant in our current setting is the preferable method for the taxonomic classification. The computational bottleneck of the CARMA3 pipeline is the homology search, in particular the BLAST search. In our evaluation the initial BLAST search accounted for over 98 % of the total running time. However, this is a problem shared with all BLAST-based approaches. Furthermore, we have shown in our evaluation that this problem can be dealt with by the use of data reduction strategies which include assembly and gene detection steps.

One of the reasons that the HMMER3-based variant does not perform as well as the BLASTx-based variant might be that the Pfam-A database contains less sequence information than the NR protein database. In our evaluation the NR protein database contained 3.55 Gaa (billion amino acids) while Pfam-A contained only 0.77 Gaa. The Pfam database also provides multiple alignments that have been created by aligning NCBI GenPept sequences [166] against Pfam-A. Since this additional sequence information might increase the classification accuracy we are planning to incorporate these alignments into the HMMER-based variant of CARMA3. Also, we are considering to include the Pfam-B database in the homology search as this should increase the fraction of metagenomic reads being classified.

Currently available biological sequence databases are known to be biased because they mainly contain sequences of species that are culturable. Although we have tried to minimize the effect of this bias on the results of our evaluation by creating the order filtered database, this bias has to be kept in mind when generalizing our evaluation results to metagenomic reads from uncultured species.

In our experiments we have shown that CARMA3 can also be used to taxonomically classify protein sequences derived from metagenomes like the human gut metagenome. We also have shown that the application of several data reduction strategies is a reasonable approach to handle the enormous amounts of data produced by recent sequencing technologies like Illumina.

In addition to CARMA3 we also have presented the web application WebCARMA, which makes metagenomic analyses of protein-encoding DNA sequences with CARMA3 easily accessible to the scientifc community. WebCARMA provides taxonomic and functional classifications in common data formats as well as basic visualizations of the profiles.

In an additional experiment we were able to show that ultra-short reads, as short as 35 bp, can be used for the taxonomic classification of a metagenome. The biogas data set we have used in the analysis is a low complexity data set with only a few prevalent species. Therefore, our results do not necessarily apply to data sets of higher complexity. Still, we think we have shown that ultra-short reads can indeed, in principle, be used for reliable taxonomic classification of a microbial community if the coverage is high enough.

6.1 Future Directions

Metagenomics with CARMA3 still leaves some room for improvements. Proper statistics to assess the significance of functional and taxonomic predictions based on short reads are still missing [90]. The abundance levels of the classification results have to be read with care. Species with larger genomes or more genes than other species will be overrepresented in the taxonomic profiles because more EGTs can be found. Therefore, more accurate results might be achieved by weighting EGTs using additional information like the genome size of the closest known relative.

CARMA3 has a better specificity than other methods which means that predictions of low abundant species are also more reliable. Still, CARMA3 makes many wrong predictions, many of which can be filtered away by using a simple abundance threshold. This threshold provides a trade-off between the ability to detect low abundant species and the ability to discard wrong predictions. A systematic evaluation that would allow for more conscious use of this threshold should be helpful.

Some predictions of CARMA3 are more reliable than others. For example, a taxonomic classification that is based on only one reference sequence should be considered less reliable than a classification based on a set of reference sequences representing a gene family. It also could be observed that classifications based on less significant BLAST hits are more likely to produce wrong predictions. Incorporating some kind of reliability measure therefore should help in interpreting the final taxonomic classification results.

CARMA3 was designed to detect protein-encoding DNA fragments. Longer fragments, like contigs or the complete genome, contain more than one gene. Currently, CARMA3 will only use the gene that obtains the highest BLAST bitscore to determine the taxonomic origin of the metagenomic fragment. Combining predictions based on each gene on such a fragment should lead to more robust and reliable predictions.

It is also possible to detect viral or plasmid genes using CARMA3. If such a gene is shared between a virus and a genome, or a plasmid and a genome in the reference database, it is likely that it will be assigned the taxonomic status “unknown” by CARMA3. The reason for this is that the structure of the NCBI taxonomy places entities like plasmids and viruses at the same taxonomic level like *Archaea*, *Bacteria* and *Eukaryota*. Any metagenomic fragment with two similarly significant BLAST hits to different entities will be assigned to the LCA of these entities, which is the root node of NCBI taxonomy tree. A future version of CARMA3 could account for this by restricting the reciprocal search to one entity and flagging the metagenomic sequence as putatively of viral or plasmid origin.

In Section 3.3 we discussed the problem that the best BLAST hit is often not necessarily the nearest neighbor. We have developed the BLAST variant of CARMA3 based on the assumption that the best BLAST hit is the nearest neighbor, or at least close to it. Our experience with the 16S-rDNA/RNA variant of CARMA3 confirmed that BLAST is not necessarily always the best choice to find the nearest neighbor. While noticing that 16S-rDNA sequences are a special case, because their sequence is highly conserved among different species, it still raises the question if there are better alternatives to BLAST also for our BLASTx-based variant of CARMA3. Nevertheless, a k-mer based approach like that of the RDP Classifier will not be able to provide the high sensitivity in homology detection of BLAST that is required for metagenomic protein-encoding DNA sequences.

Considering the speed of development of high-throughput sequencing technologies in the last years, it is likely that shotgun metagenomics will benefit from increasing read lengths and lower error rates. New developments like single-cell sequencing [219] even justify the hope that the ability to sequence near-complete or complete genomes of uncultured microbes will become a standard tool in metagenomics in the future. Such sequencing efforts are likely to help decreasing the current bias in sequence databases towards cultured species. Finally, we think that novel sequencing technologies and computational methods, like the one we have introduced in this work, will further help in shedding light on the still largely unknown microbial world.

Bibliography

[1] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res*, 12(5):281–290, 2005.

[2] S. G. Acinas, L. A. Marcelino, V. Klepac-Ceraj, and M. F. Polz. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol*, 186(9):2629–2635, May 2004.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Garland Science, 4 edition, 2002.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990.

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

[6] S. Anderson. Shotgun DNA sequencing using cloned dnase i-generated fragments. *Nucleic Acids Res*, 9(13):3015–3027, Jul 1981.

[7] F. E. Angly, B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer. The marine viromes of four oceanic regions. *PLoS Biol*, 4(11):e368, Nov 2006.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[9] G. C. Baker and D. A. Cowan. 16S rDNA primers and the unbiased assessment of thermophile diversity. *Biochem Soc Trans*, 32(Pt 2):218–221, Apr 2004.

[10] J. M. S. Bartlett and D. Stirling. A short history of the polymerase chain reaction. *Methods Mol Biol*, 226:3–6, 2003.

[11] P. J. L. Bell, A. Sunna, M. D. Gibbs, N. C. Curach, H. Nevalainen, and P. L. Bergquist. Prospecting for novel lipase genes using PCR. *Microbiology*, 148(Pt 8):2283–2291, Aug 2002.

[12] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res*, 37(Database issue):D26–D31, Jan 2009.

[13] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. R. Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008.

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

[15] S. A. Berger, D. Krompass, and A. Stamatakis. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol*, 60(3):291–302, May 2011.

[16] S. A. Berger and A. Stamatakis. Evolutionary placement of short sequence reads. Technical report, TU Munich, November 2009. Available: `http://arxiv.org/abs/0911.2852v1`.

[17] J. Besemer and M. Borodovsky. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res*, 27(19):3911–3920, Oct 1999.

[18] E. M. Bottos, W. F. Vincent, C. W. Greer, and L. G. Whyte. Prokaryotic diversity of arctic ice shelf microbial mats. *Environ Microbiol*, 10(4):950–966, Apr 2008.

[19] A. Brady and S. L. Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated markov models. *Nat Methods*, 6(9):673–676, Sep 2009.

[20] M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A*, 99(22):14250–14255, Oct 2002.

[21] M. Bunge, L. S. Søbjerg, A.-E. Rotaru, D. Gauthier, A. T. Lindhardt, G. Hause, K. Finster, P. Kingshott, T. Skrydstrup, and R. L. Meyer. Formation of palladium(0) nanoparticles at microbial surfaces. *Biotechnol Bioeng*, 107(2):206–215, Oct 2010.

[22] O. Béjà, M. T. Suzuki, E. V. Koonin, L. Aravind, A. Hadd, L. P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S. B. Jovanovich, R. A. Feldman, and E. F. DeLong. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol*, 2(5):516–529, Oct 2000.

[23] G. Campbell-Platt. Fermented foods – a world perspective. *Food Research International*, 27(3):253 – 257, 1994.

[24] T. R. Cech. Structural biology. the ribosome is a ribozyme. *Science*, 289(5481):878–879, Aug 2000.

[25] V. L. Chandler. Paramutation: from maize to mice. *Cell*, 128(4):641–645, Feb 2007.

[26] S. Chatterji, I. Yamazaki, Z. Bai, and J. Eisen. CompostBin: A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads. In M. Vingron and L. Wong, editors, *Research in Computational Molecular Biology*, volume 4955 of *Lecture Notes in Computer Science*, chapter 3, pages 17–28. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2008.

[27] A. C. Chinault and J. Carbon. Overlap hybridization screening: isolation and characterization of overlapping DNA fragments surrounding the leu2 gene on yeast chromosome iii. *Gene*, 5(2):111–126, Feb 1979.

[28] K. Chojnacka. Biosorption and bioaccumulation - the prospects for practical applications. *Environment International*, 36(3):299–307, 2010.

[29] N. Comfort. Essay reviews. [review of: Rabinow p. making PCR: a story of biotechnology. university of chicago press, 1996; and fujimura j. crafting science: a sociohistory of the quest for the genetics of cancer. harvard university press, 1996]. *Oral Hist Rev*, 26(2):181–186, 1999.

[30] D. Cowan, Q. Meyer, W. Stafford, S. Muyanga, R. Cameron, and P. Wittwer. Metagenomic gene discovery: past, present and future. *Trends Biotechnol*, 23(6):321–329, Jun 2005.

[31] A. I. Culley, A. S. Lang, and C. A. Suttle. Metagenomic analysis of coastal RNA virus communities. *Science*, 312(5781):1795–1798, Jun 2006.

[32] D. Dalevi, D. Dubhashi, and M. Hermansson. Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. *Bioinformatics*, 22(5):517–522, Mar 2006.

[33] D. Dalevi, N. N. Ivanova, K. Mavromatis, S. D. Hooper, E. Szeto, P. Hugenholtz, N. C. Kyrpides, and V. M. Markowitz. Annotation of metagenome short reads using proxygenes. *Bioinformatics*, 24(16):i7–13, Aug 2008.

[34] R. Daniel. The metagenomics of soil. *Nat Rev Microbiol*, 3(6):470–478, Jun 2005.

[35] E. F. DeLong, C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N.-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 311(5760):496–503, Jan 2006.

[36] A. L. Demain. Microbial biotechnology. *Trends in Biotechnology*, 18(1):26 – 31, 2000.

[37] N. N. Diaz, L. Krause, A. Goesmann, K. Niehaus, and T. W. Nattkemper. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, 10:56, 2009.

[38] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, 36(16):e105, Sep 2008.

[39] S. H. Duncan, G. L. Hold, A. Barcenilla, C. S. Stewart, and H. J. Flint. Roseburia intestinalis sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. *Int J Syst Evol Microbiol*, 52(Pt 5):1615–1620, Sep 2002.

[40] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*, pages – . Cambridge University Press, 2002.

[41] B. E. Dutilh, M. A. Huynen, and M. Strous. Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics*, Jun 2009. In press.

[42] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman. Diversity of the human intestinal microbial flora. *Science*, 308(5728):1635–1638, Jun 2005.

[43] S. R. Eddy. Profile hidden markov models (review). *Bioinformatics*, 14(9):755–763, 1998.

[44] R. A. Edwards, B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. M. Peterson, M. O. Saar, S. Alexander, E. C. Alexander, and F. Rohwer. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, 7:57, 2006.

[45] K. F. Ettwig, T. van Alen, K. T. van de Pas-Schoonen, M. S. M. Jetten, and M. Strous. Enrichment and molecular detection of denitrifying methanotrophic bacteria of the NC10 phylum. *Appl Environ Microbiol*, 75(11):3656–3662, Jun 2009.

[46] R. Feingersch and O. Béjà. Bias in assessments of marine SAR11 biodiversity in environmental fosmid and BAC libraries? *The ISME journal*, July 2009.

[47] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, 2003.

[48] N. Fierer and R. B. Jackson. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A*, 103(3):626–631, Jan 2006.

[49] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman. The pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–D222, Jan 2010.

[50] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, Jul 1995.

[51] K. U. Foerstner, C. von Mering, S. D. Hooper, and P. Bork. Environments shape the nucleotide composition of genomes. *EMBO Rep*, 6(12):1208–1213, Dec 2005.

[52] G. M. Gadd and C. White. Microbial treatment of metal pollution–a working biotechnology? *Trends Biotechnol*, 11(8):353–359, Aug 1993.

[53] W. Gerlach, S. Jünemann, F. Tille, A. Goesmann, and J. Stoye. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, 10(1):430, 2009.

[54] W. Gerlach and J. Stoye. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res*, 39(14):e91, May 2011.

[55] T. S. Ghosh, M. M. Haque, and S. S. Mande. DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics*, 11 Suppl 7:S14, 2010.

[56] D. E. Gillespie, S. F. Brady, A. D. Bettermann, N. P. Cianciotto, M. R. Liles, M. R. Rondon, J. Clardy, R. M. Goodman, and J. Handelsman. Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol*, 68(9):4301–4306, Sep 2002.

[57] W. Gish and D. J. States. Identification of protein coding regions by database similarity search. *Nat Genet*, 3(3):266–272, Mar 1993.

[58] G. González-Aguilar, J. Ayala-Zavala, G. Olivas, L. de la Rosa, and E. Álvarez-Parrilla. Preserving quality of fresh-cut products using safe technologies. *Journal für Verbraucherschutz und Lebensmittelsicherheit*, 5:65–72, 2010.

[59] H. C. Greenwell, L. M. L. Laurens, R. J. Shields, R. W. Lovitt, and K. J. Flynn. Placing microalgae on the biofuels priority list: a review of the technological challenges. *J R Soc Interface*, 7(46):703–726, May 2010.

[60] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, Oct 2003.

[61] D. Y. Guschin, B. K. Mobarry, D. Proudnikov, D. A. Stahl, B. E. Rittmann, and A. D. Mirzabekov. Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. *Appl Environ Microbiol*, 63(6):2397–2402, Jun 1997.

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

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

[64] M. Haque, T. S. Ghosh, N. K. Singh, and S. S. Mande. SPHINX–an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, 27(1):22–30, Jan 2011.

[65] M. M. Haque, T. S. Ghosh, D. Komanduri, and S. S. Mande. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25(14):1722–1730, July 2009.

[66] C. Heath, X. Hu, C. Cary, and D. Cowan. Isolation and characterisation of a novel, low-temperature-active alkaliphilic esterase from an antarctic desert soil metagenome. *Appl Environ Microbiol*, 75:4657–4659, 2009.

[67] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919, November 1992.

[68] A. Henne, R. Daniel, R. A. Schmitz, and G. Gottschalk. Construction of environmental DNA libraries in escherichia coli and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl Environ Microbiol*, 65(9):3901–3907, Sep 1999.

[69] A. Henne, R. A. Schmitz, M. Bömeke, G. Gottschalk, and R. Daniel. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on escherichia coli. *Appl Environ Microbiol*, 66(7):3113–3116, Jul 2000.

[70] M. Hess, A. Sczyrba, R. Egan, T.-W. W. Kim, H. Chokhawala, G. Schroth, S. Luo, D. S. Clark, F. Chen, T. Zhang, R. I. Mackie, L. A. Pennacchio, S. G. Tringe, A. Visel, T. Woyke, Z. Wang, and E. M. Rubin. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (New York, N.Y.)*, 331(6016):463–467, January 2011.

[71] K. J. Hoff. The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics*, 10:520, 2009.

[72] K. J. Hoff, T. Lingner, P. Meinicke, and M. Tech. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res*, 37(Web Server issue):W101–W105, Jul 2009.

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

[74] L.-N. Huang, H. Zhou, Y.-Q. Chen, S. Luo, C.-Y. Lan, and L.-H. Qu. Diversity and structure of the archaeal community in the leachate of a full-scale recirculating landfill as examined by direct 16S rRNA gene sequence retrieval. *FEMS Microbiol Lett*, 214(2):235–240, Sep 2002.

[75] J. A. Huber, D. B. M. Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D. A. Butterfield, and M. L. Sogin. Microbial population structures in the deep marine biosphere. *Science*, 318(5847):97–100, Oct 2007.

[76] P. Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biol*, 3(2):REVIEWS0003, 2002.

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

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

[79] P. H. Janssen, P. S. Yates, B. E. Grinton, P. M. Taylor, and M. Sait. Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions acidobacteria, actinobacteria, proteobacteria, and verrucomicrobia. *Appl Environ Microbiol*, 68(5):2391–2396, May 2002.

[80] P. L. Johnson and M. Slatkin. Inference of population genetic parameters in metagenomics: A clean look at messy data. *Genome Research*, 16(10):1320–1327, 2006.

[81] R. M. Johnson and R. E. Cameron. The physiology and distribution of bacteria in hot and cold deserts. *Journal of the Arizona Academy of Science*, 8:84–90, 1973.

[82] T. Joshi and D. Xu. Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics*, 8:222, 2007.

[83] J. Kennedy, B. Flemer, S. A. Jackson, D. P. H. Lejon, J. P. Morrissey, F. O'Gara, and A. D. W. Dobson. Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism. *Mar Drugs*, 8(3):608–628, 2010.

[84] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–664, Apr 2002.

[85] A. Kislyuk, S. Bhatnagar, J. Dushoff, and J. S. Weitz. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics*, 10:316, 2009.

[86] A. Knietsch, S. Bowien, G. Whited, G. Gottschalk, and R. Daniel. Identification and characterization of coenzyme B12-dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. *Appl Environ Microbiol*, 69(6):3048–3060, Jun 2003.

[87] A. Knietsch, T. Waschkowitz, S. Bowien, A. Henne, and R. Daniel. Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on escherichia coli. *Appl Environ Microbiol*, 69(3):1408–1416, Mar 2003.

[88] R. Knippers. *Molekulare Genetik*. G. Thieme, 2006.

[89] L. B. Koski and G. B. Golding. The closest blast hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–542, Jun 2001.

[90] A. Kowalczyk, T. Conway, B. Beresford-Smith, S. Choudhury, S. Sukumar, K. Polyak, and I. Haviv. Significance tests for short read concentrations. *in preparation*, 2009.

[91] L. Krause, N. N. Diaz, D. Bartels, R. A. Edwards, A. Pühler, F. Rohwer, F. Meyer, and J. Stoye. Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics*, 22(14):e281–e289, Jul 2006.

[92] L. Krause, N. N. Diaz, R. A. Edwards, K.-H. Gartemann, H. Krömeke, H. Neuweger, A. Pühler, K. J. Runte, A. Schlüter, J. Stoye, R. Szczepanowski, A. Tauch, and A. Goesmann. Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. *J Biotechnol*, 136(1-2):91–101, Aug 2008.

[93] L. Krause, N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res*, 36(7):2230–2239, April 2008.

[94] M. Kröber, T. Bekel, N. N. Diaz, A. Goesmann, S. Jaenicke, L. Krause, D. Miller, K. J. Runte, P. Viehöver, A. Pühler, and A. Schlüter. Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. *J Biotechnol*, 142(1):38–49, Jun 2009.

[95] V. Kunin, A. Engelbrektson, H. Ochman, and P. Hugenholtz. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol*, 12(1):118–123, Jan 2010.

[96] D. J. Lane, B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*, 82(20):6955–6959, Oct 1985.

[97] J. Laserson, V. Jojic, and D. Koller. Genovo: de novo assembly for metagenomes. *J Comput Biol*, 18(3):429–443, Mar 2011.

[98] G. R. LeCleir, A. Buchan, and J. T. Hollibaugh. Chitinase gene sequences retrieved from diverse aquatic habitats reveal environment-specific distributions. *Appl Environ Microbiol*, 70(12):6977–6983, Dec 2004.

[99] H. C. M. Leung, S. M. Yiu, B. Yang, Y. Peng, Y. Wang, Z. Liu, J. Chen, J. Qin, R. Li, and F. Y. L. Chin. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics*, 27(11):1489–1495, Jun 2011.

[100] R. E. Ley. Obesity and the human microbiome. *Curr Opin Gastroenterol*, 26(1):5–11, Jan 2010.

[101] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–272, Feb 2010.

[102] Y. Lin and S. Tanaka. Ethanol fermentation from biomass resources: current state and prospects. *Appl Microbiol Biotechnol*, 69(6):627–642, Feb 2006.

[103] D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, Mar 1985.

[104] B. Liu, T. Gibbons, M. Ghodsi, and M. Pop. MetaPhyler: Taxonomic profiling for metagenomic sequences. In *Proceedings of 2010 IEEE Bioinformatics and Biomedicine*, pages 95–100, Dec 2010.

[105] Y. G. Liu and R. F. Whittier. Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. *Genomics*, 25(3):674–681, Feb 1995.

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

[107] P. Lorenz and J. Eck. Metagenomics and industrial applications. *Nat Rev Microbiol*, 3(6):510–516, Jun 2005.

[108] W. Ludwig and H.-P. Klenk. Overview: A phylogenetic backbone and taxonomic framework for prokaryotic systamatics. *Bergey's Manual of Systematic Bacteriology*, 1:49–65, 2001.

[109] I. A. MacNeil, C. L. Tiong, C. Minor, P. R. August, T. H. Grossman, K. A. Loiacono, B. A. Lynch, T. Phillips, S. Narula, R. Sundaramoorthi, A. Tyler, T. Aldredge, H. Long, M. Gilman, D. Holt, and M. S. Osburne. Expression and isolation of antimicrobial small molecules from soil DNA libraries. *J Mol Microbiol Biotechnol*, 3(2):301–308, Apr 2001.

[110] A. Magi, M. Benelli, A. Gozzini, F. Girolami, F. Torricelli, and M. L. Brandi. Bioinformatics for next generation sequencing data. *Genes*, 1(2):294–307, 2010.

[111] S. Mahmood, T. E. Freitag, and J. I. Prosser. Comparison of pcr primer-based strategies for characterization of ammonia oxidizer communities in environmental samples. *FEMS Microbiol Ecol*, 56(3):482–493, Jun 2006.

[112] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. Mcdade, M. P. Mckenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, July 2005.

[113] F. Martin-Laurent, L. Philippot, S. Hallet, R. Chaussod, J. C. Germon, G. Soulas, and G. Catroux. DNA extraction from soils: old bias for new microbial diversity analysis methods. *Appl Environ Microbiol*, 67(5):2354–2359, May 2001.

[114] F. A. Matsen, R. B. Kodner, and E. V. Armbrust. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11:538, 2010.

[115] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 4(6):495–500, Jun 2007.

[116] J. McEntyre and J. Ostell, editors. *The NCBI Handbook*. National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda MD, 20894 USA, 2002.

[117] A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*, 4(1):63–72, Jan 2007.

[118] M. D. Megonigal, E. F. Rappaport, R. B. Wilson, D. H. Jones, J. A. Whitlock, J. A. Ortega, D. J. Slater, P. C. Nowell, and C. A. Felix. Panhandle PCR for cDNA: a rapid method for isolation of MLL fusion transcripts involving unknown partner genes. *Proc Natl Acad Sci U S A*, 97(17):9597–9602, Aug 2000.

[119] P. Meinicke, K. P. Asshauer, and T. Lingner. Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics*, 27(12):1618–1624, Jun 2011.

[120] A. Melis and T. Happe. Hydrogen production. Green algae as a source of energy. *Plant Physiol*, 127(3):740–748, Nov 2001.

[121] M. L. Metzker. Emerging technologies in DNA sequencing. *Genome Res*, 15(12):1767–1776, Dec 2005.

[122] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386, 2008.

[123] R. Milanowski, B. Zakryś, and J. Kwiatowski. Phylogenetic analysis of chloroplast small-subunit rRNA genes of the genus euglena ehrenberg. *Int J Syst Evol Microbiol*, 51(Pt 3):773–781, May 2001.

[124] M. Moran. Metatranscriptomics: Eavesdropping on complex microbial communities. *Microbe*, 4(7):329–335, 2009.

[125] D. J. Munroe and T. J. R. Harris. Third-generation sequencing fireworks at marco island. *Nat Biotechnol*, 28(5):426–428, May 2010.

[126] K. V. Myrick and W. M. Gelbart. Universal fast walking for direct and versatile determination of flanking sequence. *Gene*, 284(1-2):125–131, Feb 2002.

[127] O. U. Nalbantoglu, S. F. Way, S. H. Hinrichs, and K. Sayood. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics*, 12:41, 2011.

[128] NCBI. Ncbi taxonomy statistics.

[129] H. Noguchi, J. Park, and T. Takagi. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*, 34(19):5623–5630, 2006.

[130] H. Noguchi, T. Taniguchi, and T. Itoh. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res*, 15(6):387–396, Dec 2008.

[131] H. Ochman, A. S. Gerber, and D. L. Hartl. Genetic applications of an inverse polymerase chain reaction. *Genetics*, 120(3):621–3, 1988.

[132] A. M. O'Hara and F. Shanahan. The gut flora as a forgotten organ. *EMBO Rep*, 7(7):688–693, Jul 2006.

[133] N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen. Analyzing natural microbial populations by rRNA sequences. *ASM News*, 51:4–12, 1985.

[134] N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen. The analysis of natural microbial populations by ribosomal RNA sequences. *Adv. Microb*, 9:1–55, 1986.

[135] S.-J. Park, C.-H. Kang, J.-C. Chae, and S.-K. Rhee. Metagenome microarray for screening of fosmid clones containing specific genes. *FEMS Microbiol Lett*, 284(1):28–34, Jul 2008.

[136] K. R. Patil, P. Haider, P. B. Pope, P. J. Turnbaugh, M. Morrison, T. Scheffer, and A. C. McHardy. Taxonomic metagenome sequence assignment with structured output models. *Nat Methods*, 8(3):191–192, Mar 2011.

[137] E. Pelletier, A. Kreimeyer, S. Bocs, Z. Rouy, G. Gyapay, R. Chouari, D. Rivière, A. Ganesan, P. Daegelen, A. Sghir, G. N. Cohen, C. Médigue, J. Weissenbach, and D. L. Paslier. "candidatus cloacamonas acidaminovorans": genome sequence reconstruction provides a first glimpse of a new bacterial division. *J Bacteriol*, 190(7):2572–2579, Apr 2008.

[138] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. Meta-IDBA: a de novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94–i101, jul 2011.

[139] P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A*, 98(17):9748–9753, Aug 2001.

[140] M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*, 26(7):1641–1650, Jul 2009.

[141] D. T. Pride and T. Schoenfeld. Genome signature analysis of thermal virus metagenomes reveals archaea and thermophilic signatures. *BMC Genomics*, 9:420, 2008.

[142] D. Pushkarev, N. F. Neff, and S. R. Quake. Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, 27:847–850, August 2009.

[143] P. L. Pushpam, T. Rajesh, and P. Gunasekaran. Identification and characterization of alkaline serine protease from goat skin surface metagenome. *AMB Express*, 1(3), March 2011. [License: CC BY 2.0, `http://creativecommons.org/licenses/by/2.0`].

[144] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. L. Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, M. I. T. Consortium, P. Bork, S. D. Ehrlich, and J. Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, Mar 2010.

[145] C. Quince, A. Lanzén, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods*, 6(9):639–641, Sep 2009.

[146] J. Raes, K. U. Foerstner, and P. Bork. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol*, 10(5):490–498, Oct 2007.

[147] M. S. Rappé and S. J. Giovannoni. The uncultured microbial majority. *Annu Rev Microbiol*, 57:369–394, 2003.

[148] D. E. Rawlings. Heavy metal mining using microbes. *Annu Rev Microbiol*, 56:65–91, 2002.

[149] H. Rees, S. Grant, B. E. Jones, W. D. Grant, and S. Heaphy. Detecting cellulase and esterase enzyme activities encoded by novel genes present in environmental DNA libraries. *Extremeophiles*, 7:415–421, 2003.

[150] A. L. Reysenbach, L. J. Giver, G. S. Wickham, and N. R. Pace. Differential amplification of rRNA genes by polymerase chain reaction. *Appl Environ Microbiol*, 58(10):3417–3418, Oct 1992.

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

[152] C. S. Riesenfeld, R. M. Goodman, and J. Handelsman. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol*, 6(9):981–989, Sep 2004.

[153] C. S. Riesenfeld, P. D. Schloss, and J. Handelsman. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*, 38:525–552, 2004.

[154] I. B. Rogozin, K. S. Makarova, D. A. Natale, A. N. Spiridonov, R. L. Tatusov, Y. I. Wolf, J. Yin, and E. V. Koonin. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res*, 30(19):4264–4271, Oct 2002.

[155] M. R. Rondon, P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol*, 66(6):2541–2547, Jun 2000.

[156] G. Rosen, E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj. Metagenome fragment classification using N-mer frequency profiles. *Adv Bioinformatics*, 2008:205969, 2008.

[157] G. L. Rosen, E. R. Reichenberger, and A. M. Rosenfeld. NBC: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, 27(1):127–129, Jan 2010.

[158] R. Rosenkranz, T. Borodina, H. Lehrach, and H. Himmelbauer. Characterizing the mouse ES cell transcriptome with illumina sequencing. *Genomics*, 92(4):187–194, Oct 2008.

[159] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987.

[160] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated markov models. *Nucleic Acids Res*, 26(2):544–548, Jan 1998.

[161] S. L. Salzberg and J. A. Yorke. Beware of mis-assembled genomes. *Bioinformatics*, 21(24):4320–4321, Dec 2005.

[162] R. Sandberg, G. Winberg, C. I. Bränden, A. Kaske, I. Ernberg, and J. Cöster. Capturing whole-genome characteristics in short sequences using a naïve bayesian classifier. *Genome Res*, 11(8):1404–1409, Aug 2001.

[163] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695, Feb 1977.

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

[165] D. C. Savage. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol*, 31:107–133, 1977.

[166] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej,

D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 38(Database issue):D5–16, Jan 2010.

[167] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 37(Database issue):D5–D15, Jan 2009.

[168] P. D. Schloss and J. Handelsman. Status of the microbial census. *Microbiol Mol Biol Rev*, 68(4):686–691, Dec 2004.

[169] A. Schlüter, T. Bekel, N. N. Diaz, M. Dondrup, R. Eichenlaub, K.-H. Gartemann, I. Krahn, L. Krause, H. Krömeke, O. Kruse, J. H. Mussgnug, H. Neuweger, K. Niehaus, A. Pühler, K. J. Runte, R. Szczepanowski, A. Tauch, A. Tilker, P. Viehöver, and A. Goesmann. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J Biotechnol*, 136(1-2):77–90, Aug 2008.

[170] T. M. Schmidt, E. F. DeLong, and N. R. Pace. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol*, 173(14):4371–4378, Jul 1991.

[171] R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, Mar 2011.

[172] J. W. Schopf, A. B. Kudryavtsev, D. G. Agresti, T. J. Wdowiak, and A. D. Czaja. Laser–raman imagery of earth's earliest fossils. *Nature*, 416(6876):73–76, Mar 2002.

[173] F. Schreiber, P. Gumrich, R. Daniel, and P. Meinicke. Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, 26(7):960–961, Apr 2010.

[174] K. P. Scott, J. C. Martin, C. Chassard, M. Clerget, J. Potrykus, G. Campbell, C.-D. Mayer, P. Young, G. Rucklidge, A. G. Ramsay, and H. J. Flint. Microbes and health sackler colloquium: Substrate-driven gene expression in roseburia inulinivorans: Importance of inducible enzymes in the utilization of inulin and starch. *Proc Natl Acad Sci U S A*, Aug 2010.

[175] J. L. Sebat, F. S. Colwell, and R. L. Crawford. Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl Environ Microbiol*, 69(8):4927–4934, Aug 2003.

[176] B. Shahbaba and R. M. Neal. Gene function classification using bayesian models with hierarchy-based priors. *BMC Bioinformatics*, 7:448, 2006.

[177] C. Simon and R. Daniel. Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biotechnol*, 85(2):265–276, Nov 2009.

[178] C. Simon, J. Herath, S. Rockstroh, and R. Daniel. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Appl Environ Microbiol*, 75(9):2964–2968, May 2009.

[179] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, 103(32):12115–12120, Aug 2006.

[180] Y. L. Song, C. X. Liu, M. McTeague, P. Summanen, and S. M. Finegold. Clostridium bartlettii sp. nov., isolated from human faeces. *Anaerobe*, 10(3):179–184, Jun 2004.

[181] R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res*, 6(7):2601–2610, Jun 1979.

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

[183] A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, Nov 2006.

[184] M. Stark, S. A. Berger, A. Stamatakis, and C. von Mering. MLTreeMap–accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics*, 11:461, 2010.

[185] J. L. Stein, T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol*, 178(3):591–599, Feb 1996.

[186] H. W. Stokes, A. J. Holmes, B. S. Nield, M. P. Holley, K. M. Nevalainen, B. C. Mabbutt, and M. R. Gillings. Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl Environ Microbiol*, 67(11):5240–5246, Nov 2001.

[187] T. Strachan and A. P. Read. Human molecular genetics, 2nd edition. In *In Alberto Apostolico, Maxime Crochemore, Zvi Galil, and Udi Manber, editors, Combinatorial Pattern Matching, 4th Annual Symposium, volume 684 of Lecture Notes in Computer Science*, pages 228–242. Wiley, 1999.

[188] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239):719–724, Apr 2009.

[189] W. R. Streit and R. A. Schmitz. Metagenomics–the key to the uncultured microbes. *Curr Opin Microbiol*, 7(5):492–498, Oct 2004.

[190] M. Ströck. DNA Overview, 2006. [License: CC BY-SA 3.0, `http://creativecommons.org/licenses/by-sa/3.0/legalcode`] Source: `http://commons.wikimedia.org/wiki/File:DNA_Overview.png`, 13 July 2011.

[191] M. T. Suzuki and S. J. Giovannoni. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol*, 62(2):625–630, Feb 1996.

[192] U. Szewzyk, R. Szewzyk, and T. A. Stenström. Thermophilic, anaerobic bacteria isolated from a deep borehole in granite in sweden. *Proc Natl Acad Sci U S A*, 91(5):1810–1813, Mar 1994.

[193] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol*, 6(9):938–947, Sep 2004.

[194] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner. 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, Oct 2004.

[195] B. Temperton, D. Field, A. Oliver, B. Tiwari, M. Muhling, I. Joint, and J. A. Gilbert. Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. *The ISME Journal*, 3(7):792–796, April 2009.

[196] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994.

[197] A. H. Treusch, A. Kletzin, G. Raddatz, T. Ochsenreiter, A. Quaiser, G. Meurer, S. C. Schuster, and C. Schleper. Characterization of large-insert DNA libraries from soil for environmental genomic studies of archaea. *Environ Microbiol*, 6(9):970–980, Sep 2004.

[198] S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, Apr 2005.

[199] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, Oct 2007.

[200] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–1031, Dec 2006.

[201] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, Mar 2004.

[202] T. Uchiyama, T. Abe, T. Ikemura, and K. Watanabe. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol*, 23(1):88–93, Jan 2005.

[203] United Nations. *World Population Prospects: The 2010 Revision*. Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, New York, 2010.

[204] A. C. van der Kuyl, C. L. Kuiken, J. T. Dekker, and J. Goudsmit. Phylogeny of african monkeys based upon mitochondrial 12S rRNA sequences. *J Mol Evol*, 40(2):173–180, Feb 1995.

[205] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74, Apr 2004.

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

[207] G. C. Wang and Y. Wang. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology*, 142 (Pt 5):1107–1114, May 1996.

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

[209] R. A. Waterland and R. L. Jirtle. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol*, 23(15):5293–5300, Aug 2003.

[210] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

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

[212] T. Wicker, E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, 7:275, 2006.

[213] T. Williams, C. Kelley, and many others. Gnuplot 4.4: an interactive plotting program. `http://gnuplot.sourceforge.net/`, March 2010.

[214] L. L. Williamson, B. R. Borlee, P. D. Schloss, C. Guan, H. K. Allen, and J. Handelsman. Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Appl Environ Microbiol*, 71(10):6335–6344, Oct 2005.

[215] C. R. Woese. Bacterial evolution. *Microbiol Rev*, 51(2):221–271, Jun 1987.

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

[217] E. K. Wommack, J. Bhavsar, and J. Ravel. Metagenomics: Read length matters. *Appl. Environ. Microbiol.*, 74(5):1453–1463, January 2008.

[218] J. C. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLoS Comput Biol*, 6(2):e1000667, Feb 2010.

[219] T. Woyke, D. Tighe, K. Mavromatis, A. Clum, A. Copeland, W. Schackwitz, A. Lapidus, D. Wu, J. P. McCutcheon, B. R. McDonald, N. A. Moran, J. Bristow, and J.-F. Cheng. One bacterial cell, one complete genome. *PLoS One*, 5(4):e10314, 2010.

[220] L. Wu, D. K. Thompson, G. Li, R. A. Hurt, J. M. Tiedje, and J. Zhou. Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl Environ Microbiol*, 67(12):5780–5790, Dec 2001.

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

[222] Y.-W. Wu and Y. Ye. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol*, 18(3):523–534, Mar 2011.

[223] K. Yamada, T. Terahara, S. Kurata, T. Yokomaku, S. Tsuneda, and S. Harayama. Retrieval of entire genes from environmental DNA by inverse PCR with pre-amplification of target genes using primers containing locked nucleic acids. *Environ Microbiol*, 10(4):978–987, Apr 2008.

[224] E. Yergeau, S. A. Schoondermark-Stolk, E. L. Brodie, S. Déjean, T. Z. DeSantis, O. Gonçalves, Y. M. Piceno, G. L. Andersen, and G. A. Kowalchuk. Environmental microarray analyses of antarctic soil microbial communities. *ISME J*, 3(3):340–351, Mar 2009.

[225] F. Yu, Y. Sun, L. Liu, and W. Farmerie. GSTaxClassifier: a genomic signature based taxonomic classifier for metagenomic data analysis. *Bioinformation*, 4(1):46–49, 2010.

[226] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–829, May 2008.

[227] F. Zhou and Y. Xu. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, 26(16):2051–2052, Aug 2010.

[228] W. Zhu, A. Lomsadze, and M. Borodovsky. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res*, 38(12):e132, Jul 2010.

Appendix A

Appendix

A.1 WebCARMA Data Formats

In the following we describe the input and output file formats of WebCARMA.

A.1.1 Input Requirements

WebCARMA accepts as input a FASTA file containing metagenomic DNA sequences. Optionally, this file can be uploaded as a compressed file, either as `zip`, `gzip` or `tgz` archive. An additional requirement is that the FASTA description lines contain unique names.

The taxonomic classification of protein or 16S-rDNA sequences is not yet supported, although this functionality is now implemented in CARMA3. A new version of WebCARMA, which is currently under development, will support the upload and analysis of this kind of sequence data.

Users should note that WebCARMA performs no quality check on the sequences, e.g., duplicates or low-quality sequences are not removed. CARMA was designed to analyze bacterial and archaeal DNA, but it is also possible to analyse eukaryotic DNA. These sequences consist mainly of non-coding DNA, to which BLASTx and HMMER3 cannot find homologous protein references. In our experience, the homology search often still reports many low quality matches to the reference protein sequences. Therefore, the result should be interpreted with care in this case.

A.1.2 Description of WebCARMA Output Files

The output of WebCARMA is a `tgz`-compressed archive file that contains the following files:

```
result.egt                        superkingdom.pdf
blastx_result.tax                 phylum.pdf
hmm_result.tax                    class.pdf
functional_profile.tsv            order.pdf
taxonomic_profile.tsv             family.pdf
functional_profile.pdf            genus.pdf
                                  species.pdf
```

Text lines in the examples below that were too long to fit on the page, were broken and marked with the "↦"-symbol at each line break.

EGTs – `result.egt`

This file contains the environmental gene tags (EGTs) that were obtained by aligning the translated reads against their Pfam family in the HMMER variant of CARMA3. The FASTA description line contains information about matching Pfam family, read identifier, HMMER3 E-value, and a list of Gene Ontology identifiers with individual fields being separated by the sequence "=+=". This format was introduced by CARMA1 and for compatibility reasons we adopted it in the subsequent versions of CARMA. The box below shows one exemplary FASTA entry consisting of a description line and an EGT protein sequence.

```
>PF04961.4=+=HWI-EAS217_1_2013P:1:1:383:736  ↦
  =+=3.2e-07=+={GO:0044237,GO:0003824}
LPKKTDEEKAARKAAI
```

Taxonomic Classifications – `blastx_result.tax` and `hmm_result.tax`

Below is an exemplaric line of a `hmm_result.tax` file. Each line consists of a tab separated list of values, where columns contain information about read identifier, Pfam family, list of Gene Ontology identifiers (GO-Id), NCBI taxon identifier, prettyprint taxon name and E-value. The format of `blastx_result.tax` differs from `hmm_result.tax` in that the entry of Pfam family is not used. In case of `hmm_result.tax`, the E-value refers to the HMMER3 E-value, whereas in case of `blastx_result.tax`, it refers to the E-value of the best BLAST hit.

```
072343_1987_0335<tab>PF01312<tab>  ↦
  {GO:0016020,GO:0009306}<tab>68295<tab> ↦
  Bacteria(superkingdom)!Firmicutes(phylum)!Clost...  ↦
  <tab>7e-30
```

For further processing of these data, we recommend to use the NCBI taxon identifier instead of the prettyprint taxon name.

Functional Profile – `functional_profile.tsv`

The example below shows three lines from a `functional_profile.tsv` file. Each GO-Id is represented by one line that shows the corresponding Gene Ontology term with its category in round brackets, as well as number of EGTs in `result.egt` that have this GO-Id assigned.

```
GO:0051287<tab>"NAD or NADH binding (molecular_function)"↦
  <tab>molecular_function<tab>105
GO:0005737<tab>"cytoplasm (cellular_component)"<tab>  ↦
  cellular_component<tab>103
GO:0046168<tab>"glycerol-3-phosphate catabolic process  ↦
  (biological_process)"<tab>biological_process<tab>101
```

Taxonomic Profile – `taxonomic_profile.tsv`

CARMA does not directly create taxonomic profiles, it just tells for each EGT, which gene it encodes and from which taxon it most likely originates. To get a better overview of the metagenomic content of a sample, one needs to have a histogram that states for each taxon how many supporting metagenomic sequences have been found in the sample. This information is provided by the file `taxonomic_profile.tsv`. It contains for each taxon the number of metagenomic sequences that have been assigned to this taxon by the BLAST variant of CARMA3. The first column shows the taxonomic rank of the taxon.

```
order<tab>"Poales"<tab>165
order<tab>"Clostridiales"<tab>39
class<tab>"Liliopsida"<tab>180
class<tab>"Clostridia"<tab>40
```

Functional Profile – `functional_profile.pdf`

The file `functional_profile.pdf` provides a visualization of the functional profile that is given by `functional_profile.tsv`. It shows for the 40 most abundant GO-terms the numbers of metagenomic sequences that support the corresponding GO-terms. An example of such a functional profile of a complete metagenomic 454 data set from a biogas plant microbial community produced with WebCARMA is depicted in Figure A.1. The underlying data set is described in Section 5.2.

Taxonomic Profiles – `{superkingdom.pdf, ..., species.pdf}`

The files `superkingdom.pdf`, `phylum.pdf`, `class.pdf`, `order.pdf`, `family.pdf`, `genus.pdf`, and `species.pdf` are visualizations of the file

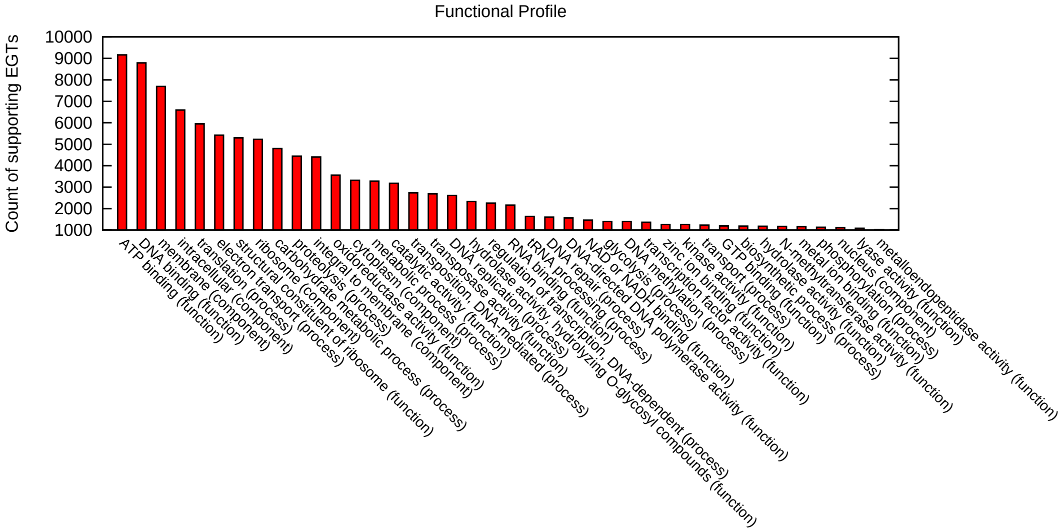


Figure A.1: Example of a functional profile: 40 most abundant GO-terms in the metagenome of an agricultural biogas reactor.

taxonomic_profile.tsv. For each taxonomic rank, only the 40 most abundant taxa are shown. In addition, taxa with a relative abundance below 0.01 have been discarded. Examples of these files can be found in Appendix A.9.

A.1.3 Tools

The profiles WebCARMA provides by default have been created using certain parameters which a user might want to change, like the cut-off threshold for the relative abundance of taxa. Therefore, we provide Perl scripts for download that can easily be used as templates for own data processing pipelines. In the following, we give a short overview of these scripts. A manual with more detailed explanations can be found on the WebCARMA site.

getFunctionalProfile.pl

This script takes result.egt as input and creates taxonomic_profile.tsv.

getTaxonomicProfile.pl

Taking blastx_result.tax as input, the file taxonomic_profile.tsv is created. This script can alternatively be used to generate a taxonomic profile from hmm_result.tax.

`getComparativeTaxonomicProfile.pl`

To compare two or more metagenomic data sets, that have been analyzed with CARMA, it is possible to create a comparative taxonomic profile with this script. Examples of visualizations of such a comparative profile can be found in Appendix Sections A.8 and A.10.

A.2 Simulated Metagenome

Table A.1: The 25 genomes from NCBI used to simulate the metagenomic reads. The number of reads refers to the simulated metagenome with average read length 265 bp.

Genome	taxonomy id	# reads
Aliivibrio salmonicida LFI1238	316275	1205
Bdellovibrio bacteriovorus HD100	264462	967
Brucella melitensis biovar Abortus 2308	359391	856
Burkholderia mallei SAVP1	320388	1297
Burkholderia multivorans ATCC 17616	395019	1789
Chlamydia trachomatis D/UW-3/CX	272561	253
Clostridium phytofermentans ISDg	357809	1276
Colwellia psychrerythraea 34H	167879	1377
Cyanothece sp. ATCC 51142	43989	1307
Escherichia coli B str. REL606	413997	1123
Haemophilus parasuis SH0165	557723	598
Helicobacter pylori B38	592205	404
Mycobacterium abscessus ATCC 19977	561007	1305
Orientia tsutsugamushi str. Ikeda	334380	499
Pseudomonas aeruginosa PA7	381754	1714
Rhodopseudomonas palustris BisB5	316057	1224
Shigella boydii CDC 3083-94	344609	1202
Sinorhizobium meliloti 1021	266834	973
Staphylococcus aureus subsp. aureus Mu50	158878	736
Staphylococcus epidermidis ATCC 12228	176280	684
Streptococcus pneumoniae G54	512566	534
Sulfurovum sp. NBC37-1	387093	681
Synechococcus sp. CC9605	110662	650
Vibrio cholerae M66-2	579112	1020
Vibrio parahaemolyticus RIMD 2210633	223926	1326
Total		25000

A.3 Evaluation of CARMA3 on Reads of Different Lengths

Table A.2: Reads simulated using the MetaSim 454 error model with 400 bp read length.

	order filtered		species filtered		all	
	TP	FP	TP	FP	TP	FP
superkingdom	14190	1175	21238	151	23010	19
phylum	9905	1153	20034	201	22919	28
class	4307	1289	16723	269	20826	30
order	–	2082	15851	246	21456	31
family	–	959	11696	213	18805	25
genus	–	151	7078	452	16357	99
species	–	2	–	49	417	17

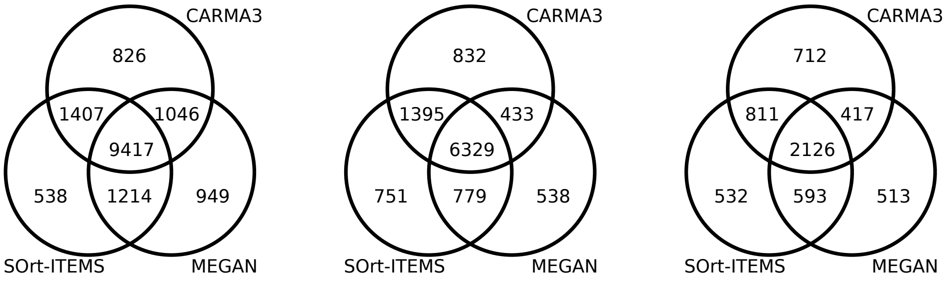
Table A.3: Reads simulated using the MetaSim 454 error model with 80 bp read length.

	order filtered		species filtered		all	
	TP	FP	TP	FP	TP	FP
superkingdom	3174	174	8558	39	12411	14
phylum	2208	531	8107	146	12218	41
class	884	564	6517	174	10745	47
order	–	1114	6469	251	11062	76
family	–	665	4916	299	9526	73
genus	–	190	2854	354	8597	90
species	–	42	–	485	2724	44

Table A.4: Reads simulated using the MetaSim Illumina error model with 80 bp read length.

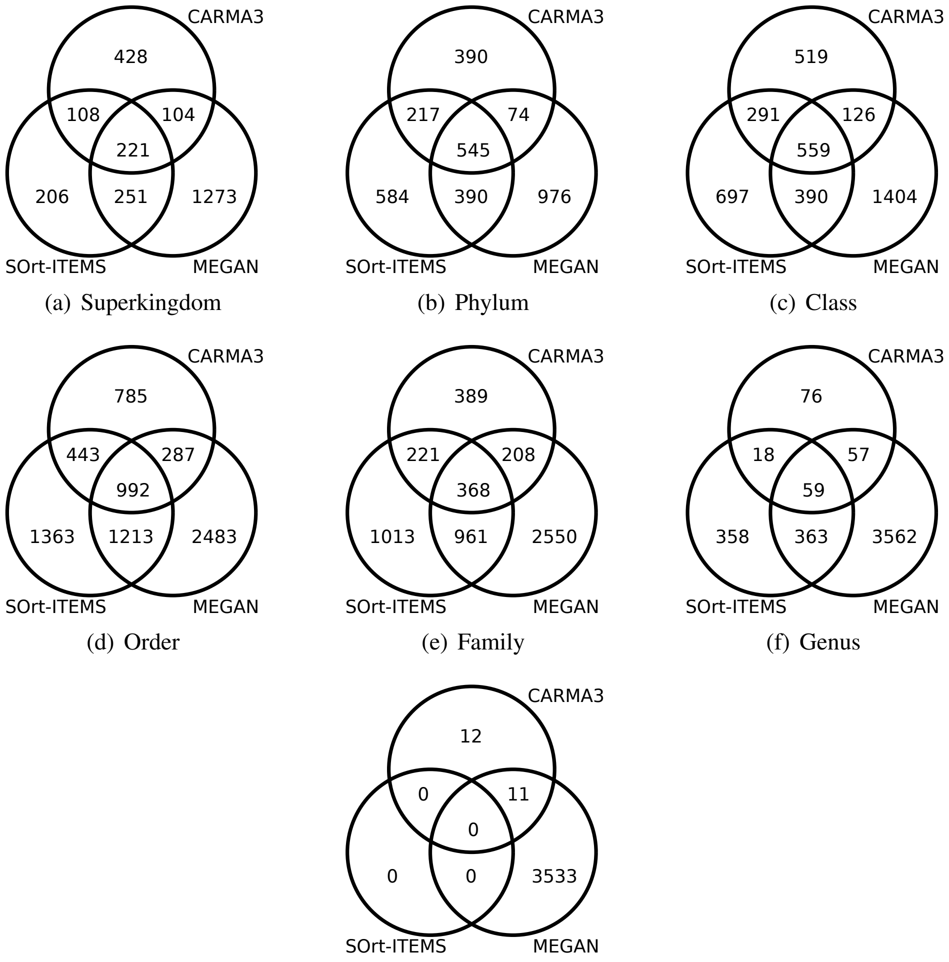
	order filtered		species filtered		all	
	TP	FP	TP	FP	TP	FP
superkingdom	7112	335	15692	34	20305	6
phylum	5139	904	15012	182	20157	22
class	2223	1021	12483	226	18331	24
order	–	2143	12291	342	18809	37
family	–	1320	9769	439	16736	38
genus	–	363	6420	672	15759	82
species	–	99	–	1190	6773	114

A.4 Overlap of Classifications for Order-Filtered Data Set



(a) Superkingdom (b) Phylum (c) Class

Figure A.2: True positives overlap of CARMA3, SOrt-ITEMS and MEGAN for the order-filtered data set at taxonomic ranks superkingdom to class.



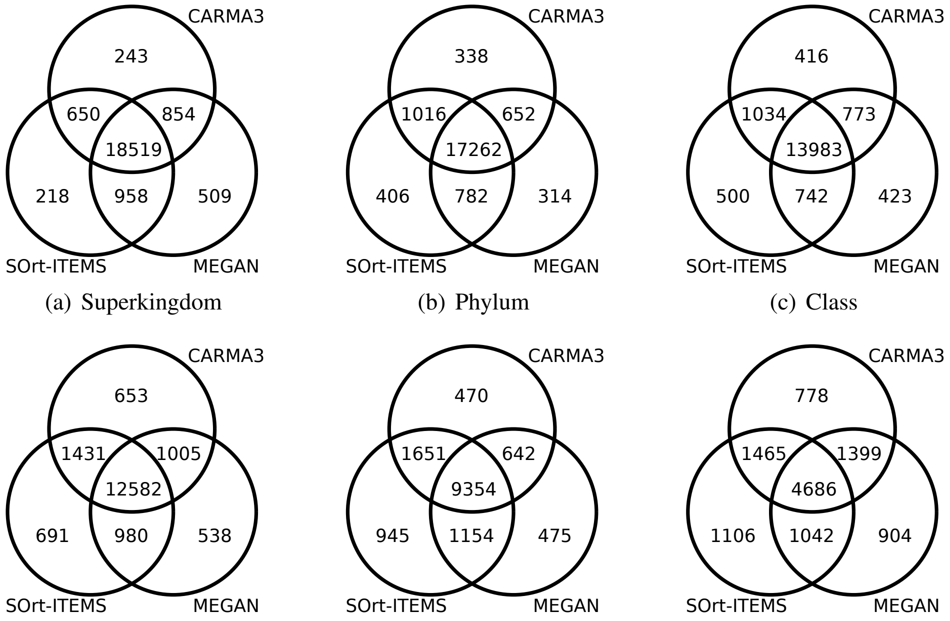
(a) Superkingdom (b) Phylum (c) Class

(d) Order (e) Family (f) Genus

(g) Species

Figure A.3: False positives overlap of CARMA3, SOrt-ITEMS and MEGAN for the order-filtered data set at taxonomic ranks superkingdom to species.

A.5 Overlap of Classifications for Species-Filtered Data Set



(a) Superkingdom

(b) Phylum

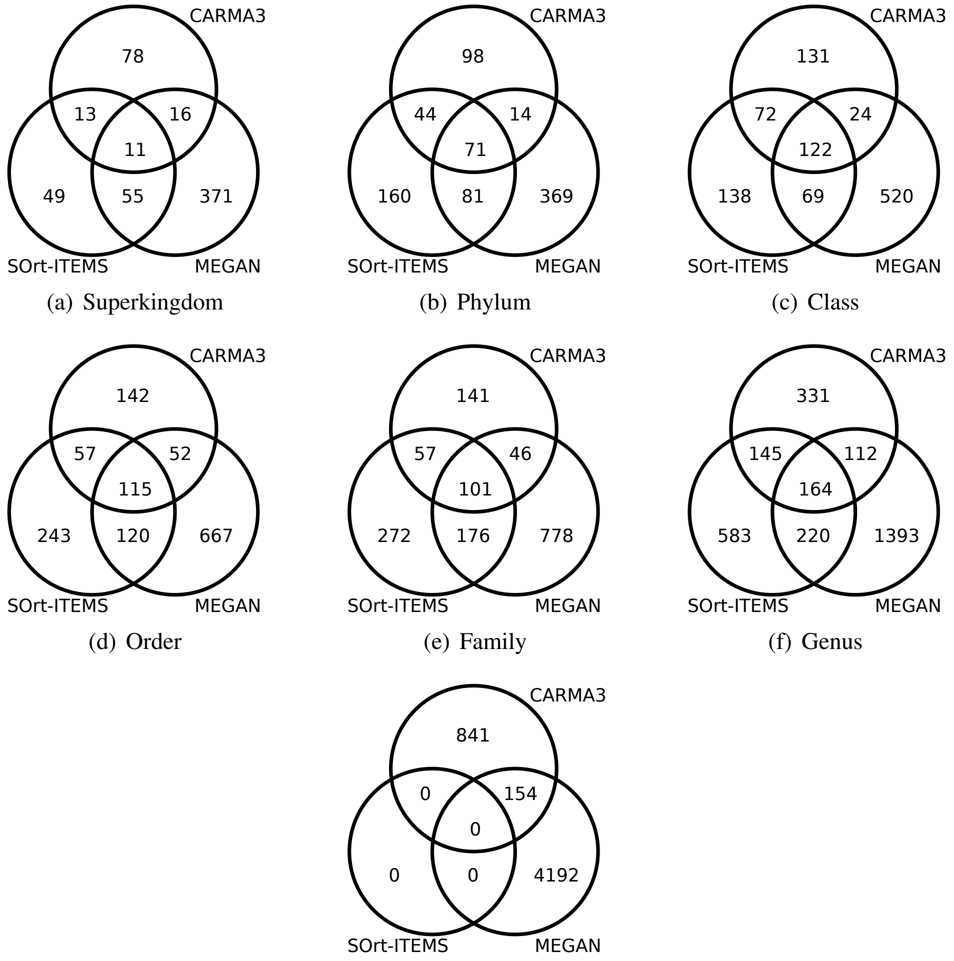
(c) Class

(d) Order

(e) Family

(f) Genus

Figure A.4: True positives overlap of CARMA3, SOrt-ITEMS and MEGAN for the species-filtered data set at taxonomic ranks superkingdom to genus.



(a) Superkingdom

(b) Phylum

(c) Class

(d) Order

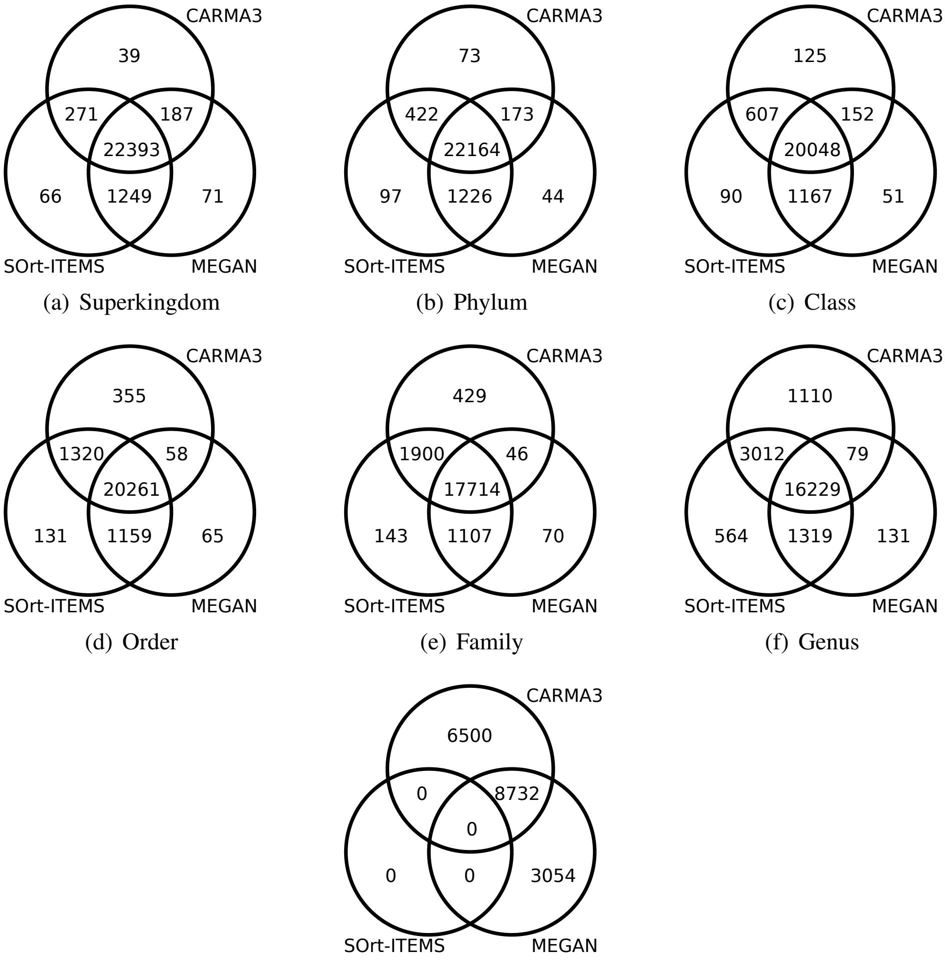
(e) Family

(f) Genus

(g) Species

Figure A.5: False positives overlap of CARMA3, SOrt-ITEMS and MEGAN for the species-filtered data set at taxonomic ranks superkingdom to species.

A.6 Overlap of Classifications for Unfiltered Data Set



(a) Superkingdom

(b) Phylum

(c) Class

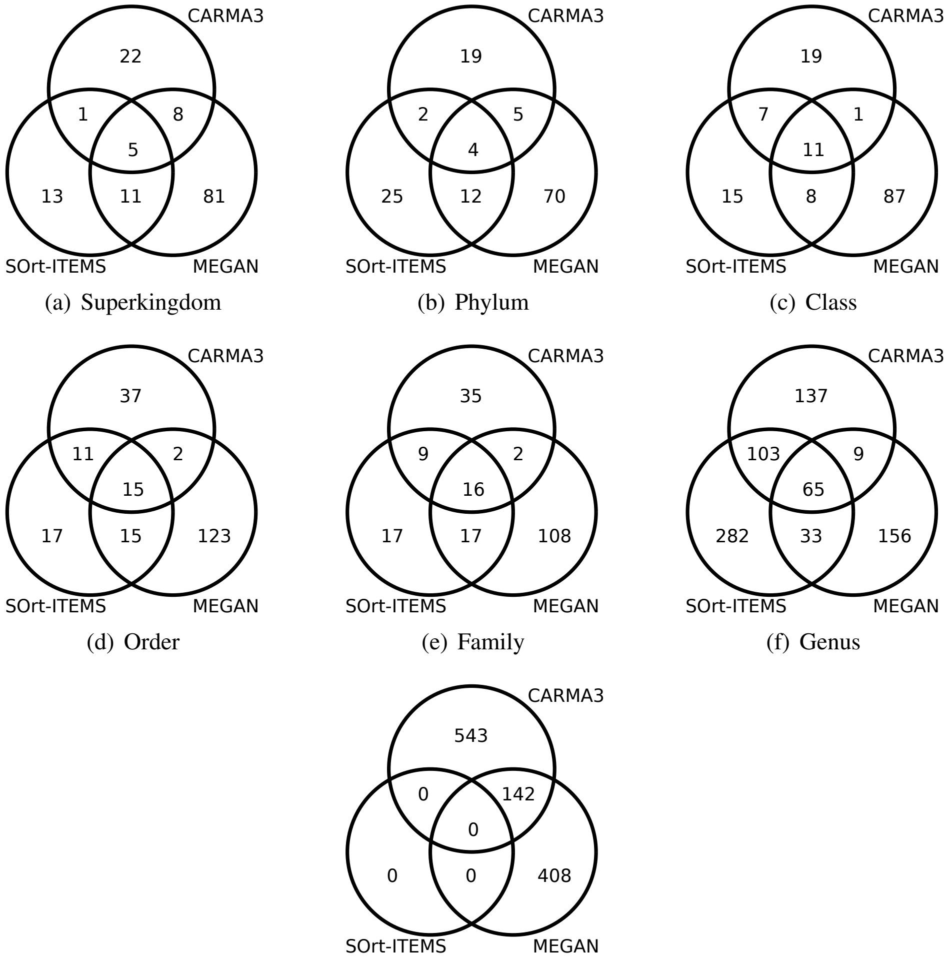
(d) Order

(e) Family

(f) Genus

(g) Species

Figure A.6: True positives overlap of CARMA3, SOrt-ITEMS and MEGAN for the unfiltered data set at taxonomic ranks superkingdom to species.



(a) Superkingdom

(b) Phylum

(c) Class

(d) Order

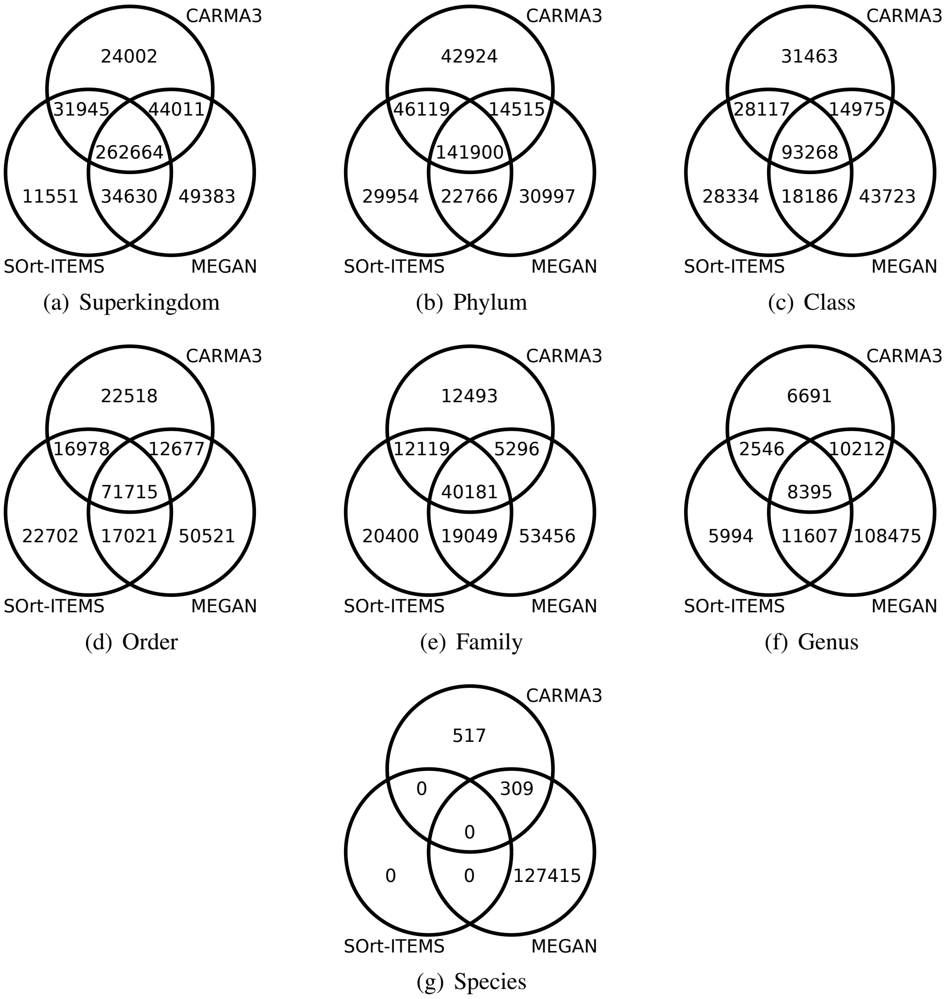
(e) Family

(f) Genus

(g) Species

Figure A.7: False positives overlap of CARMA3, SOrt-ITEMS and MEGAN for the unfiltered data set at taxonomic ranks superkingdom to species.

A.7 Biogas Plant Microbial Community – Overlap



(a) Superkingdom

(b) Phylum

(c) Class

(d) Order

(e) Family

(f) Genus

(g) Species

Figure A.8: Number of reads classified by each method at the corresponding taxonomic rank.

A.8 Biogas Plant Microbial Community – Comparative Taxonomic Profile

Table A.5: Fraction of reads that have been discarded from the comparative taxonomic profile due to the cut-off threshold 0.01.

	CARMA3	SOrt-ITEMS	MEGAN
superkingdom	0	0	0
phylum	0.045	0.054	0.073
class	0.036	0.054	0.120
order	0.065	0.106	0.277
family	0.087	0.142	0.456
genus	0.091	0.090	0.495
species	0.267	–	0.689

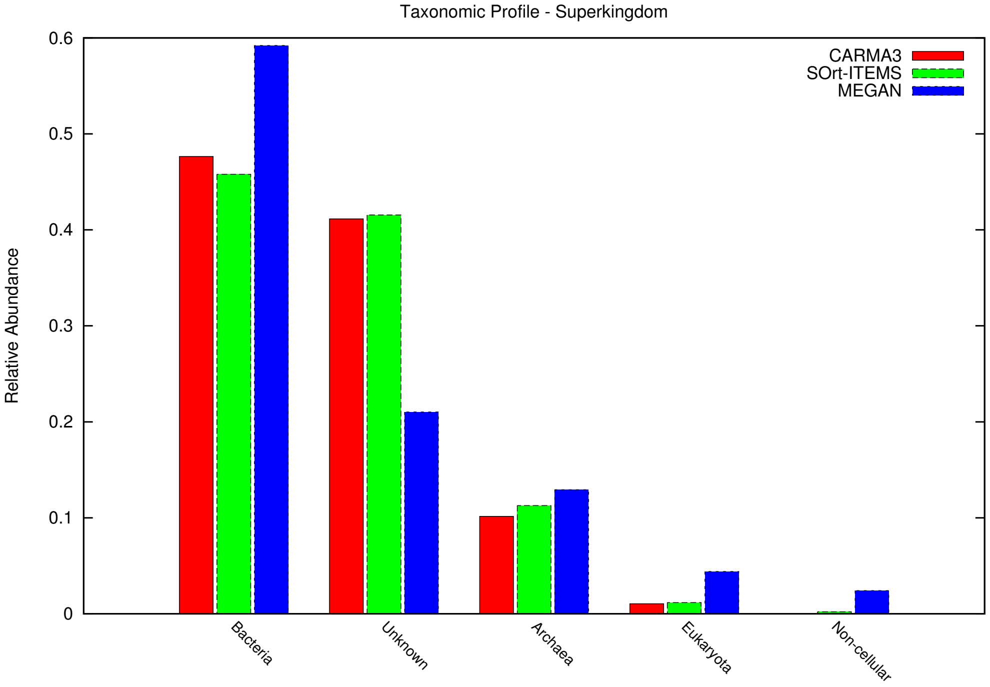


Figure A.9: Comparative taxonomic profile of a biogas plant microbial community at taxonomic rank superkingdom.

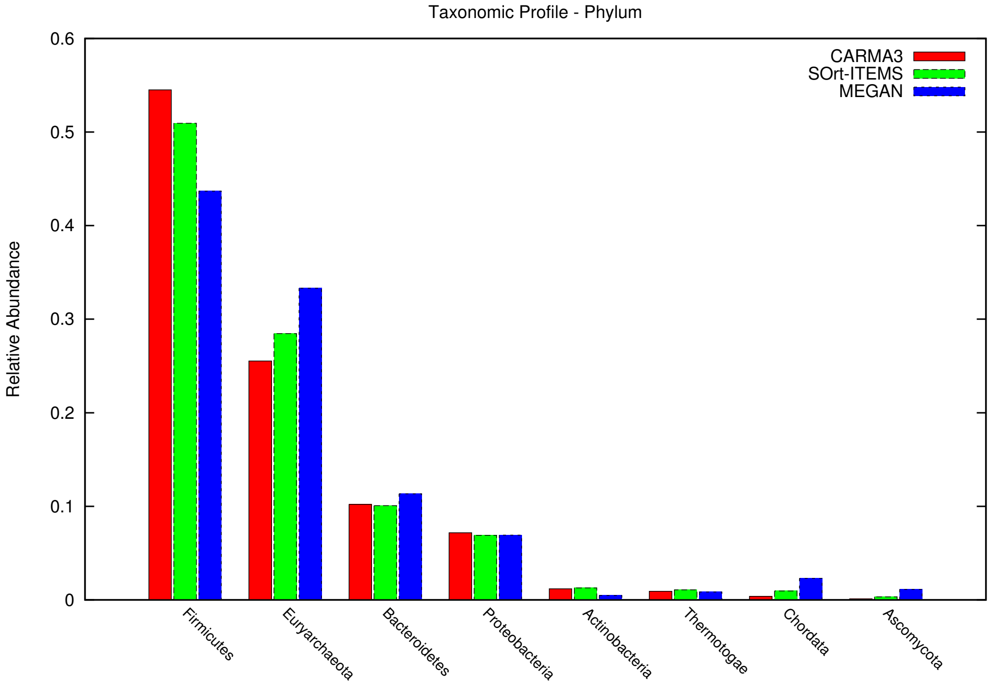


Figure A.10: Comparative taxonomic profile of a biogas plant microbial community at taxonomic rank phylum.

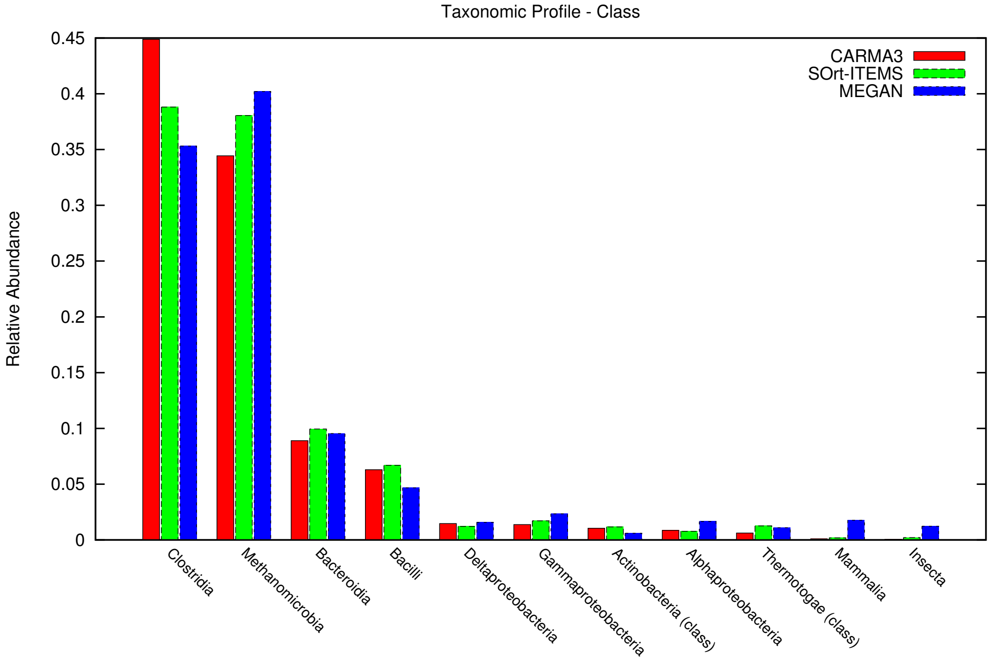


Figure A.11: Comparative taxonomic profile of a biogas plant microbial community at taxonomic rank class.

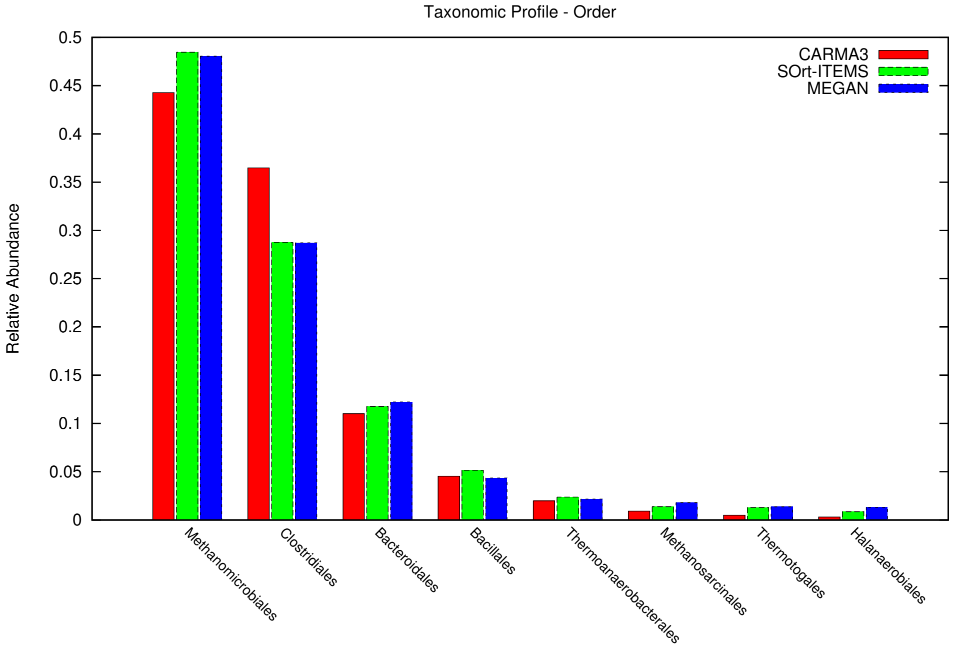


Figure A.12: Comparative taxonomic profile of a biogas plant microbial community at taxonomic rank order.

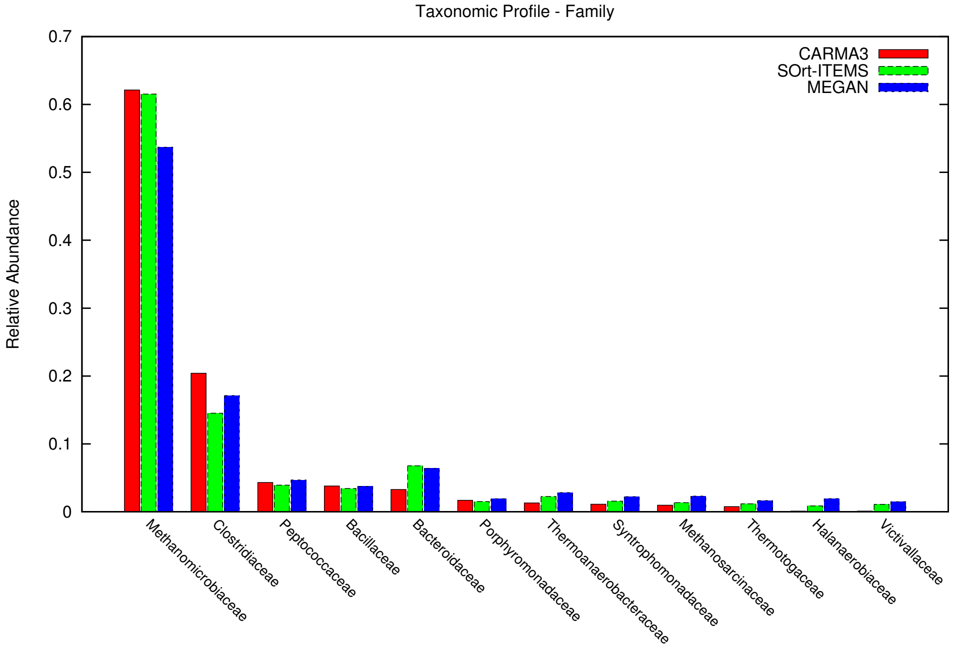


Figure A.13: Comparative taxonomic profile of a biogas plant microbial community at taxonomic rank family.

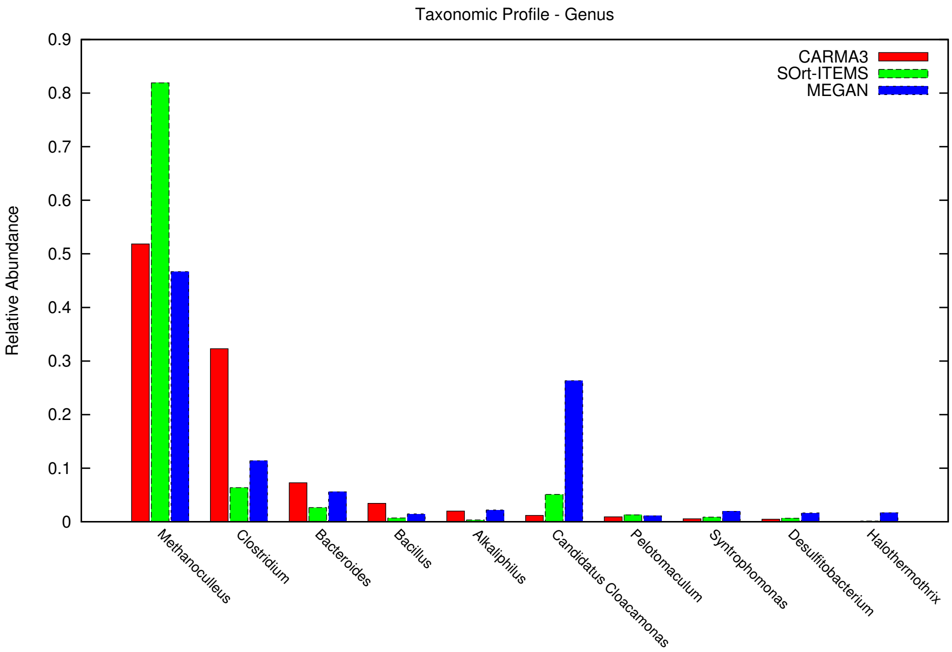


Figure A.14: Comparative taxonomic profile of a biogas plant microbial community at taxonomic rank genus.

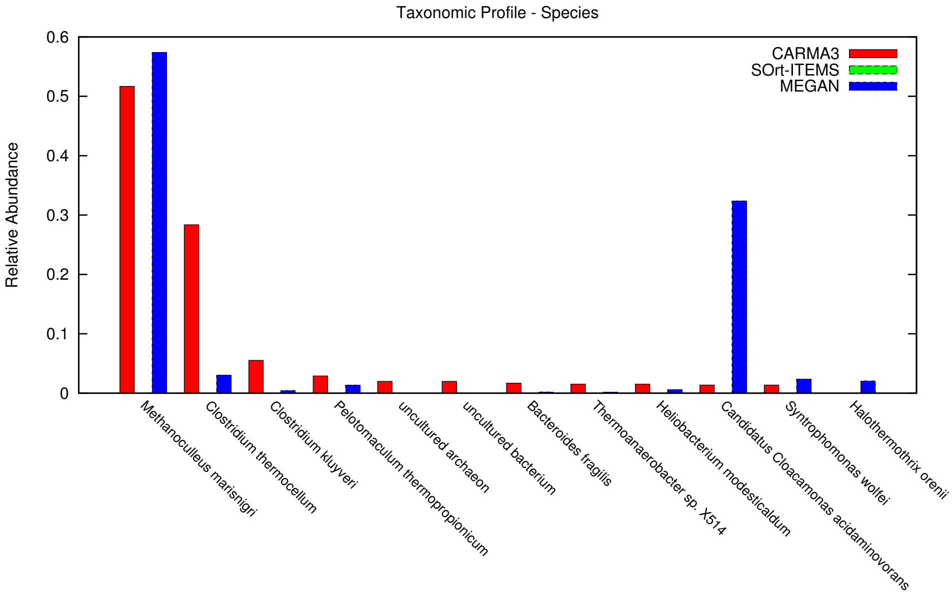


Figure A.15: Comparative taxonomic profile of a biogas plant microbial community at taxonomic rank species. Note that SOrt-ITEMS does not make predictions at taxonomic rank species.

A.9 Taxonomic Profile of the Human Gut Microbial Gene Catalogue

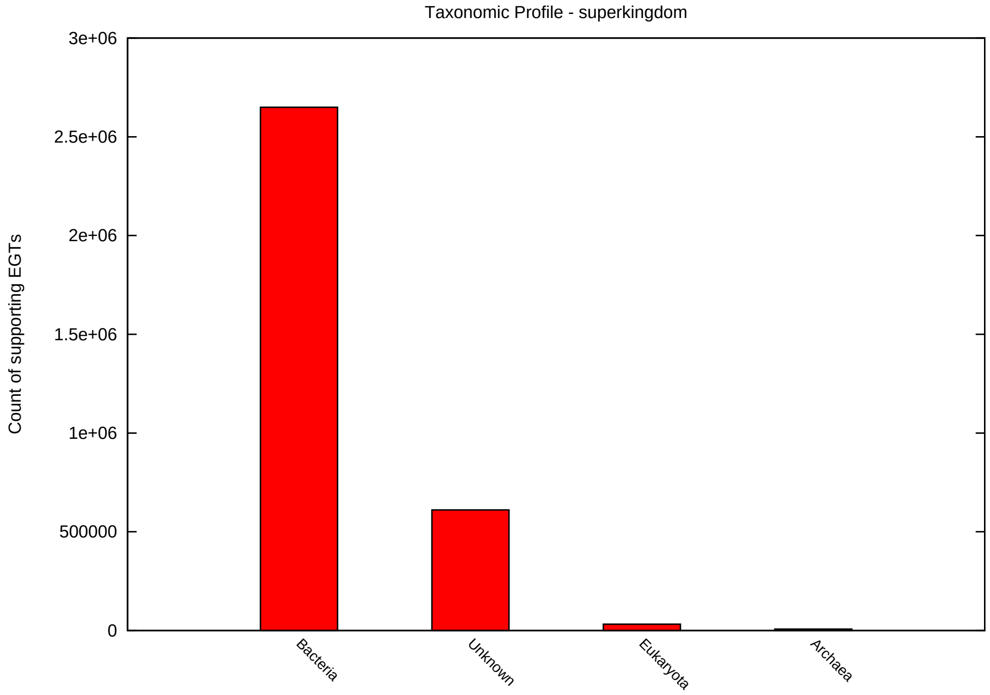


Figure A.16: All taxa of the human gut microbial gene catalogue at taxonomic rank superkingdom.

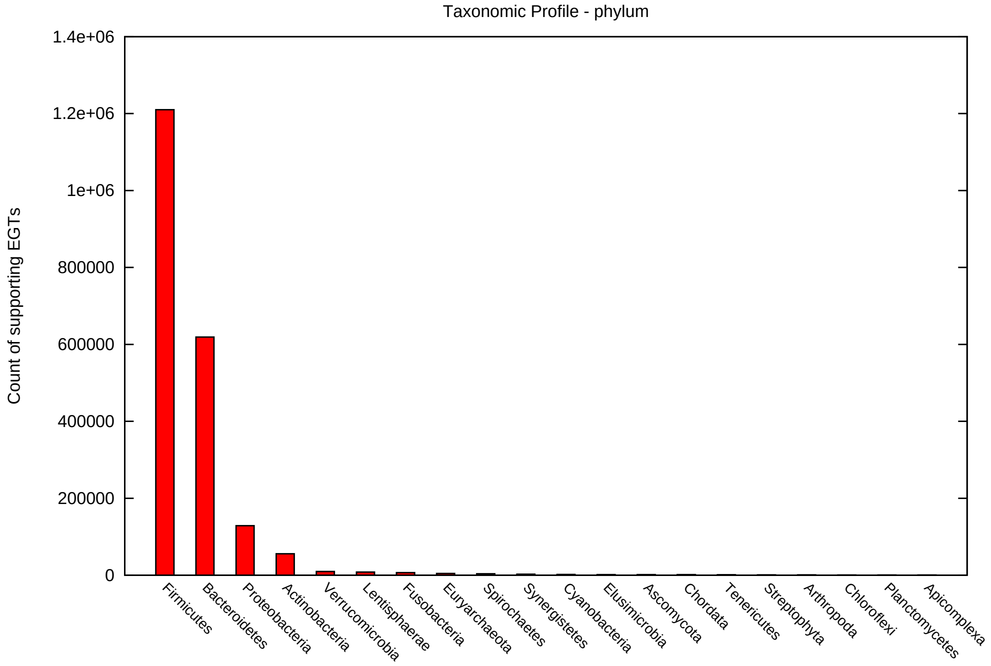


Figure A.17: The 20 most abundant taxa of the human gut microbial gene catalogue at taxonomic rank phylum.

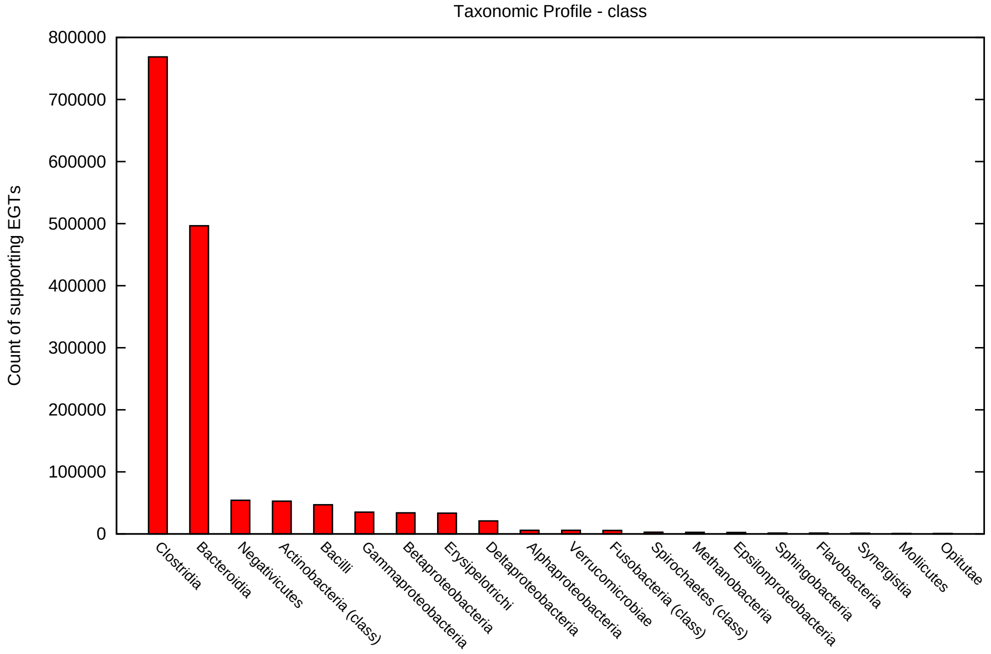


Figure A.18: The 20 most abundant taxa of the human gut microbial gene catalogue at taxonomic rank class.

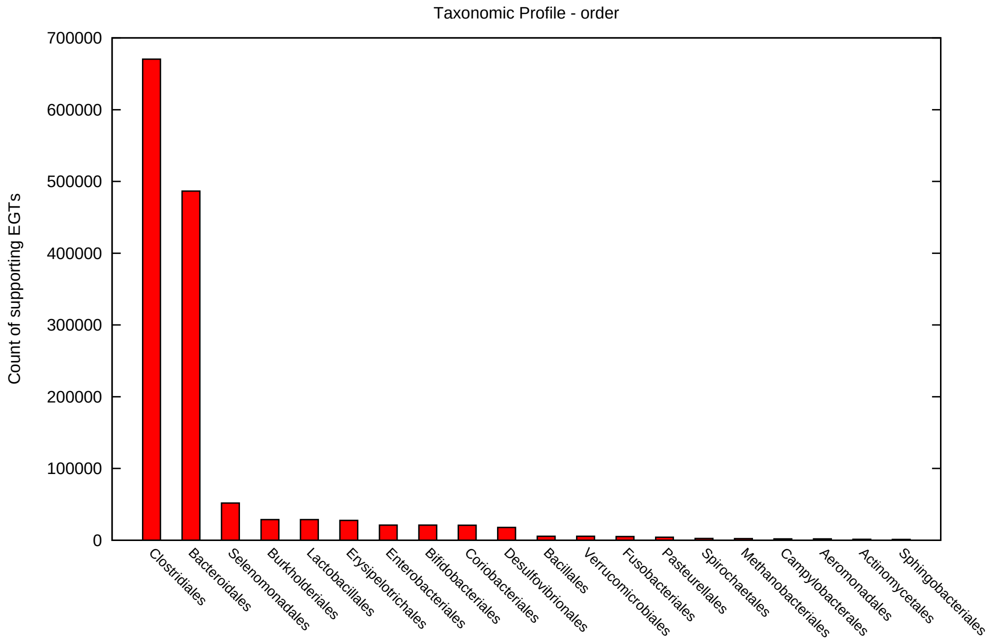


Figure A.19: The 20 most abundant taxa of the human gut microbial gene catalogue at taxonomic rank order.

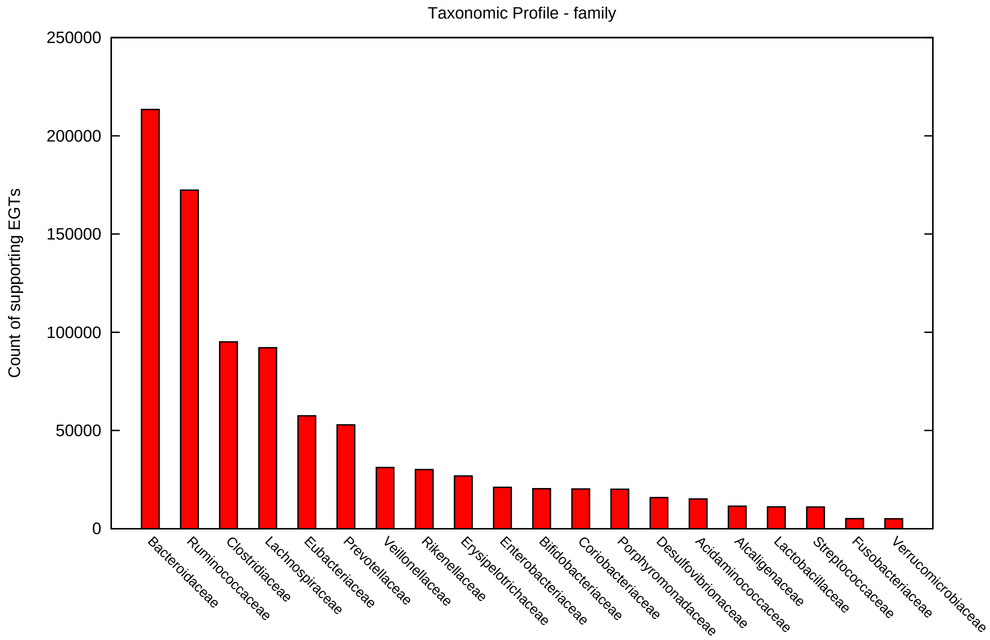


Figure A.20: The 20 most abundant taxa of the human gut microbial gene catalogue at taxonomic rank family.

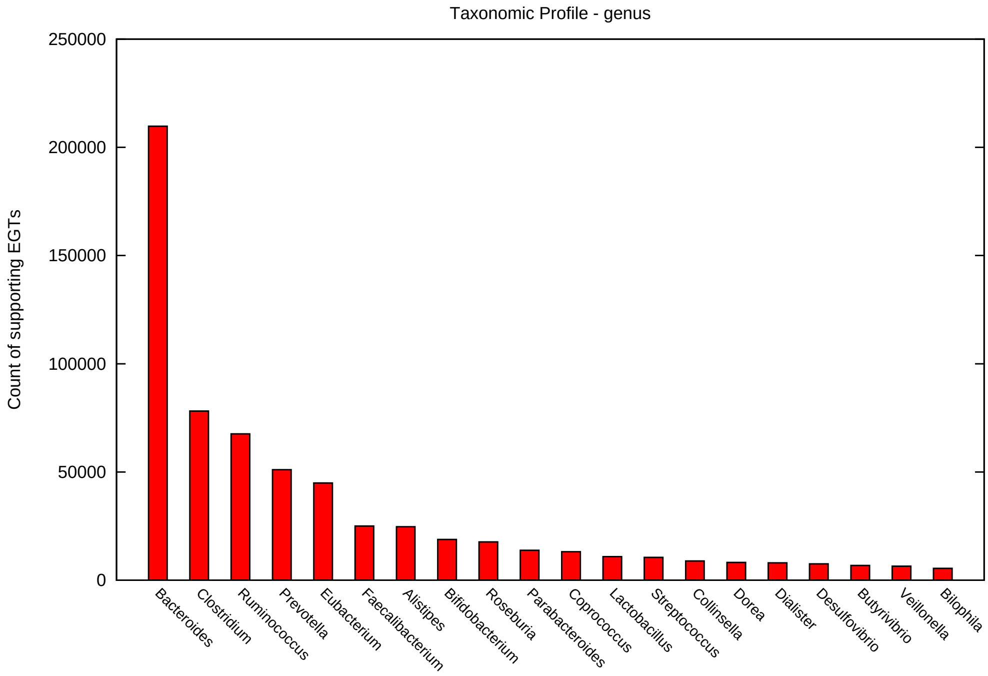


Figure A.21: The 20 most abundant taxa of the human gut microbial gene catalogue at taxonomic rank genus.

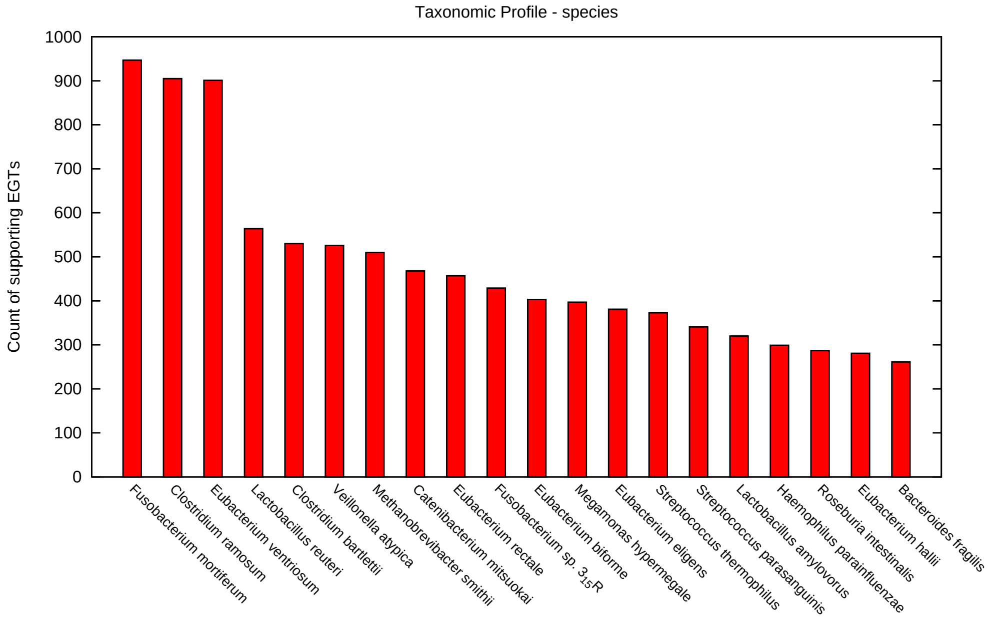


Figure A.22: The 20 most abundant taxa of the human gut microbial gene catalogue at taxonomic rank species.

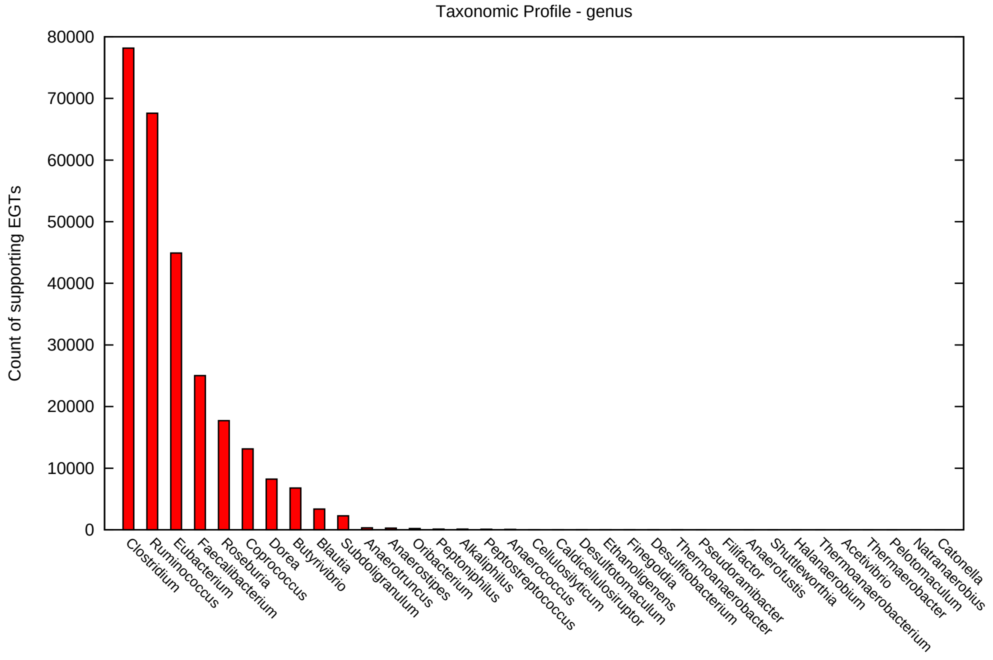


Figure A.23: All genera from the class *Clostridia* in the human gut microbial gene catalogue.

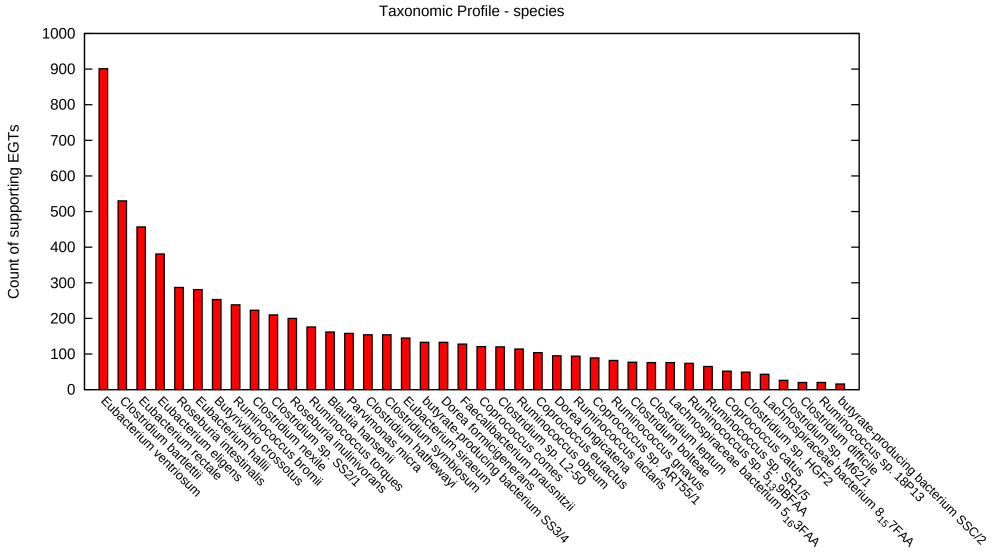


Figure A.24: The 40 most abundant species from the class *Clostridia* in the human gut microbial gene catalogue.

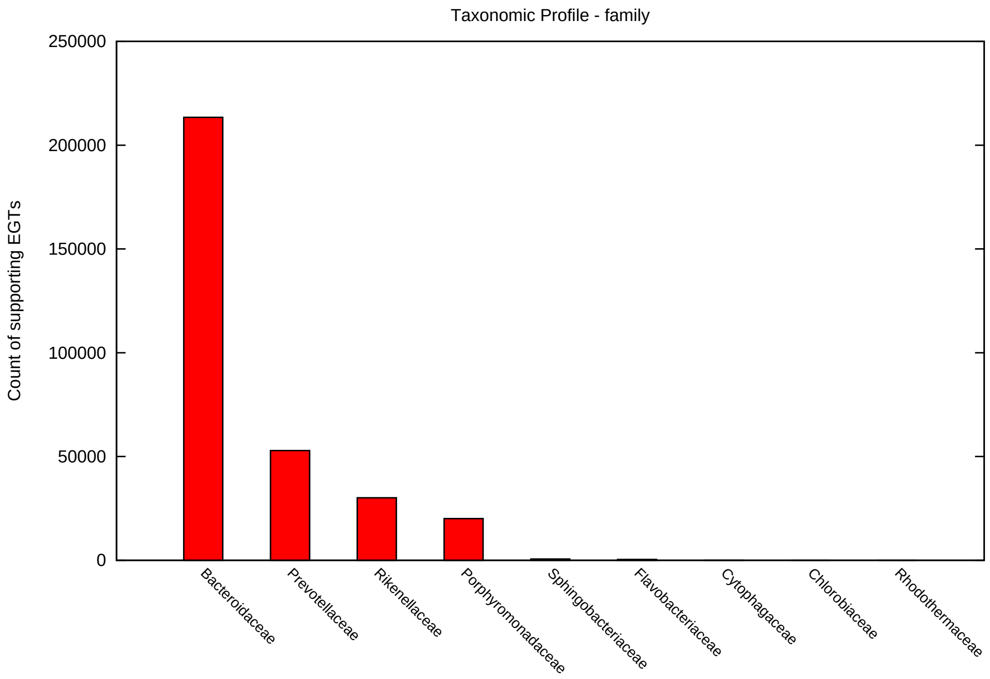


Figure A.25: All families from the phylum *Bacteroidetes* in the human gut microbial gene catalogue.

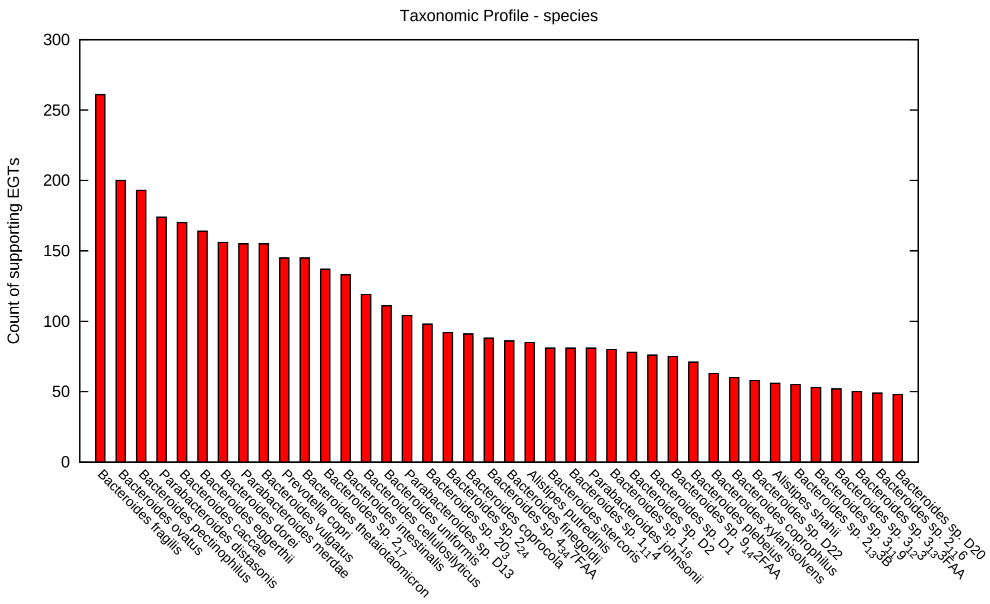


Figure A.26: The 40 most abundant species from the phylum *Bacteroidetes* in the human gut microbial gene catalogue.

A.10 Applicability of Short Reads for Taxonomic Classification

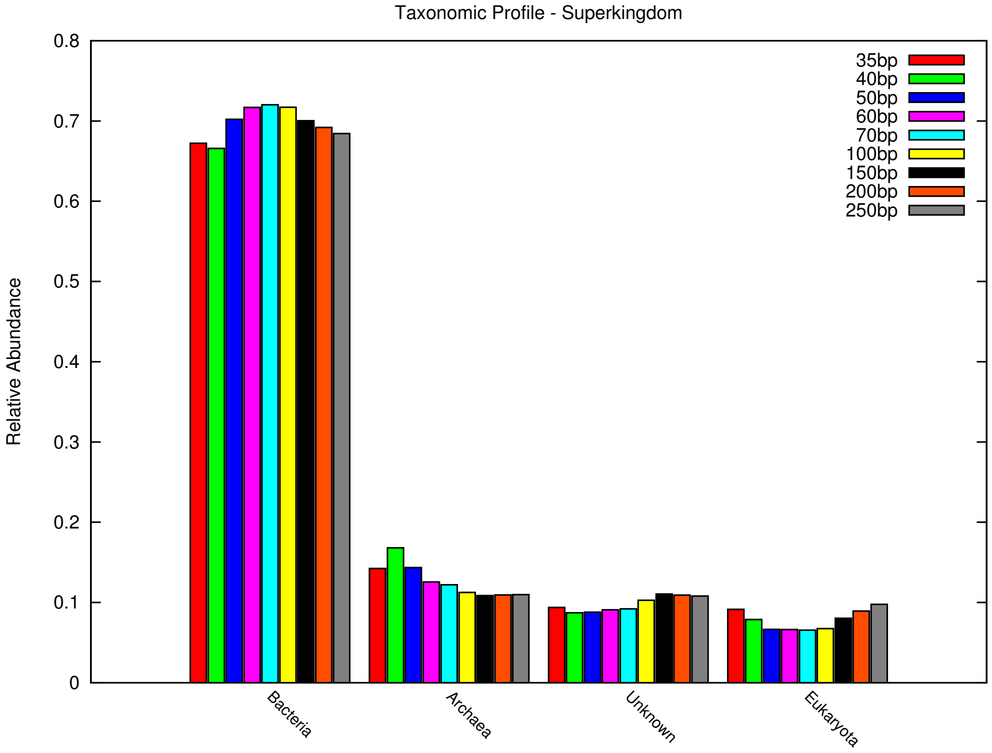


Figure A.27: Taxonomic results on the level of superkingdom.

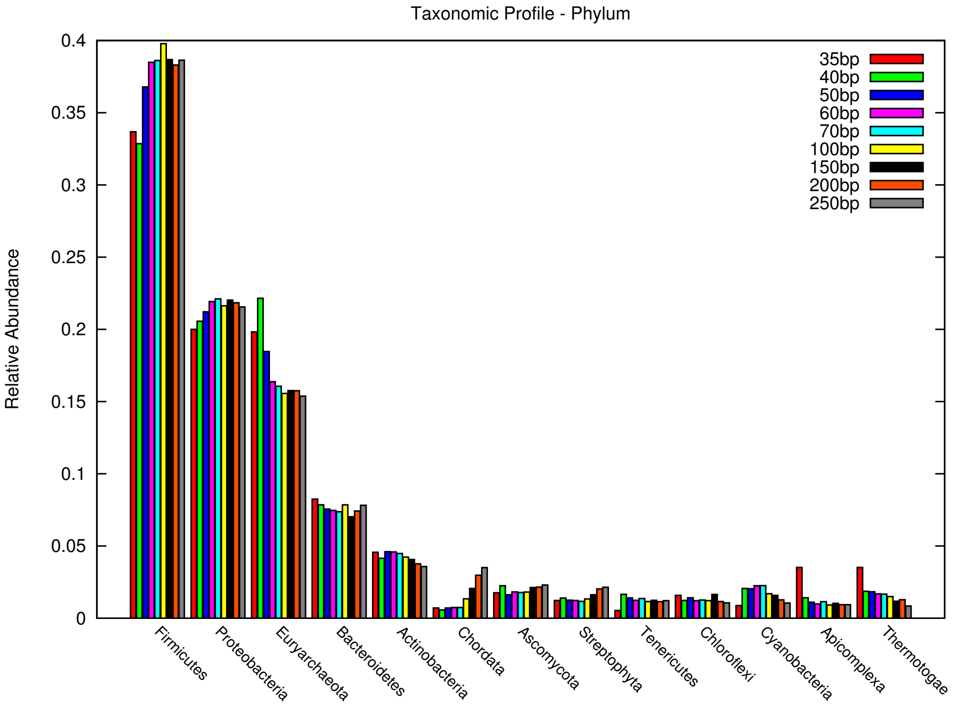


Figure A.28: Taxonomic results on the level of phylum. Only taxa with an abundance of 0.015 or higher are shown.

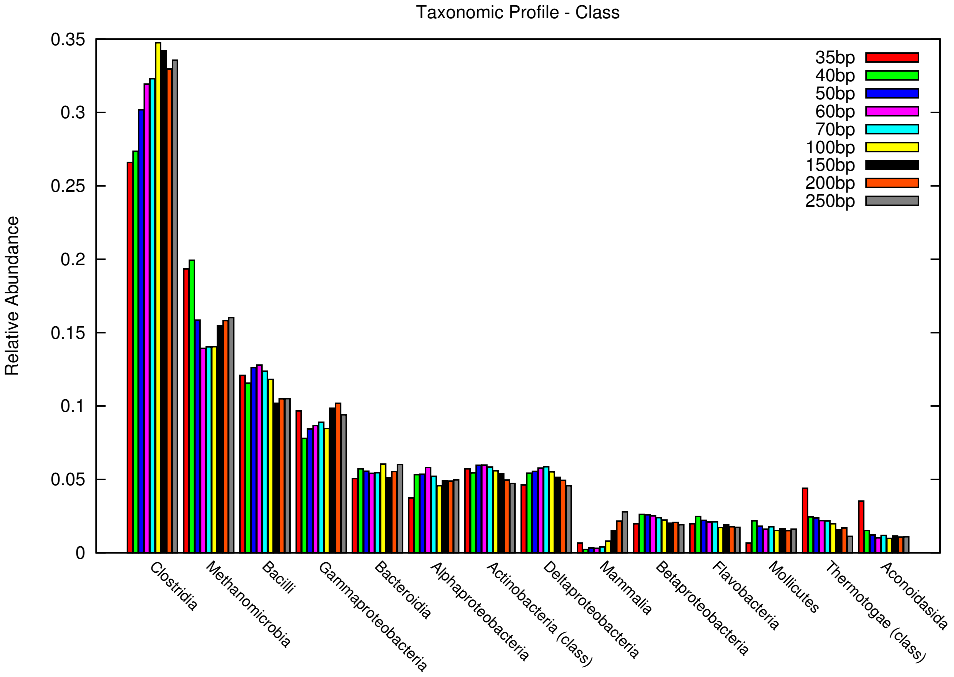


Figure A.29: Taxonomic results on the level of class. Only taxa with an abundance of 0.015 or higher are shown.

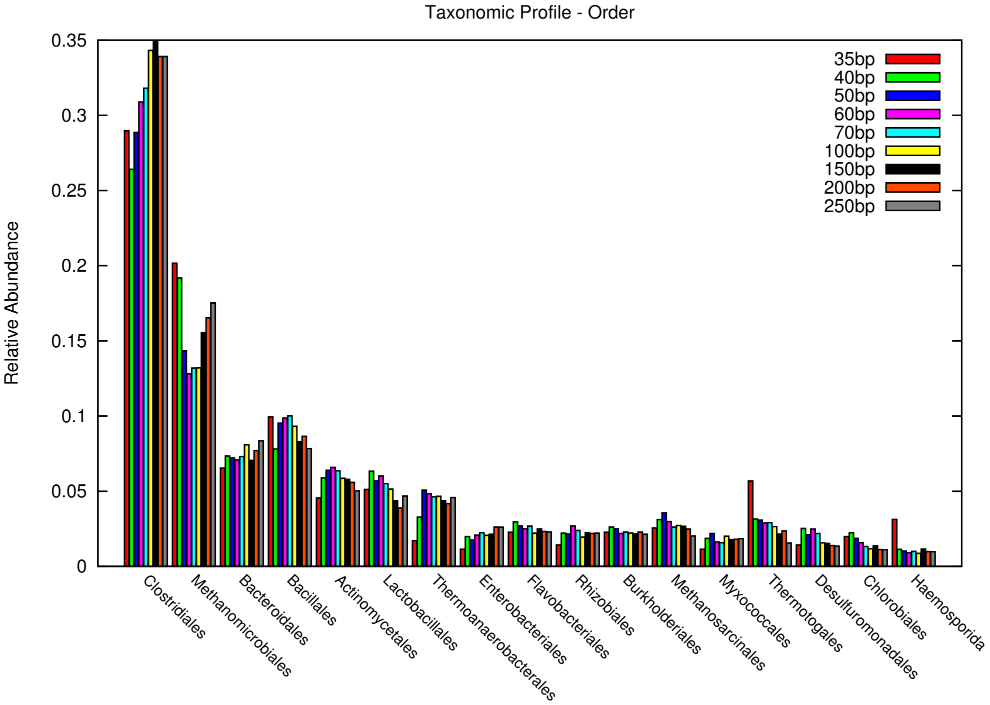


Figure A.30: Taxonomic results on the level of order. Only taxa with an abundance of 0.015 or higher are shown.

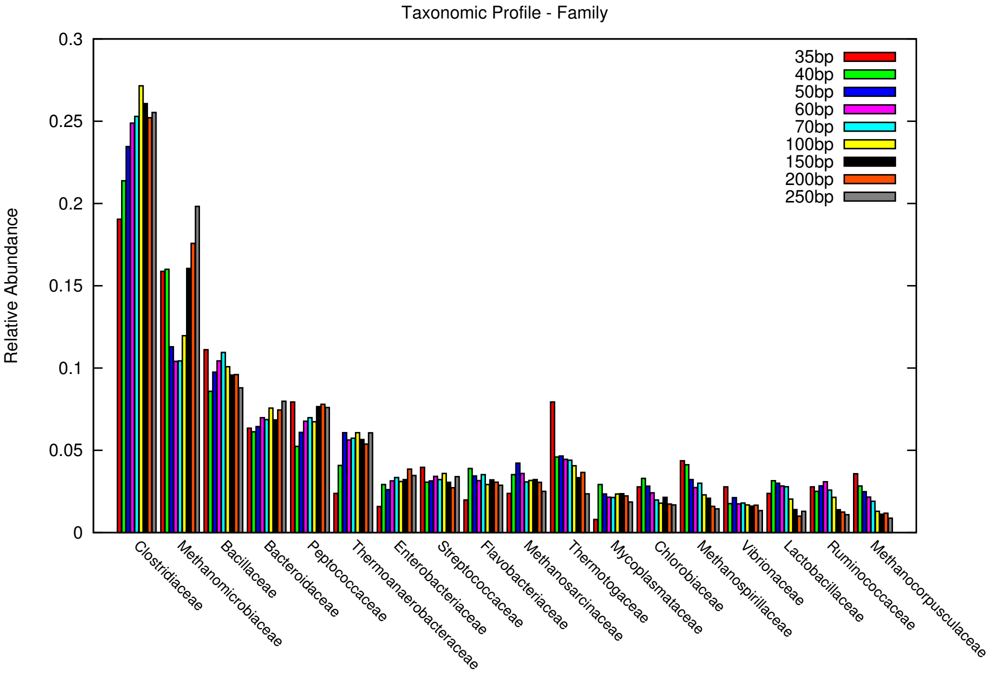


Figure A.31: Taxonomic results on the level of family. Only taxa with an abundance of 0.015 or higher are shown.

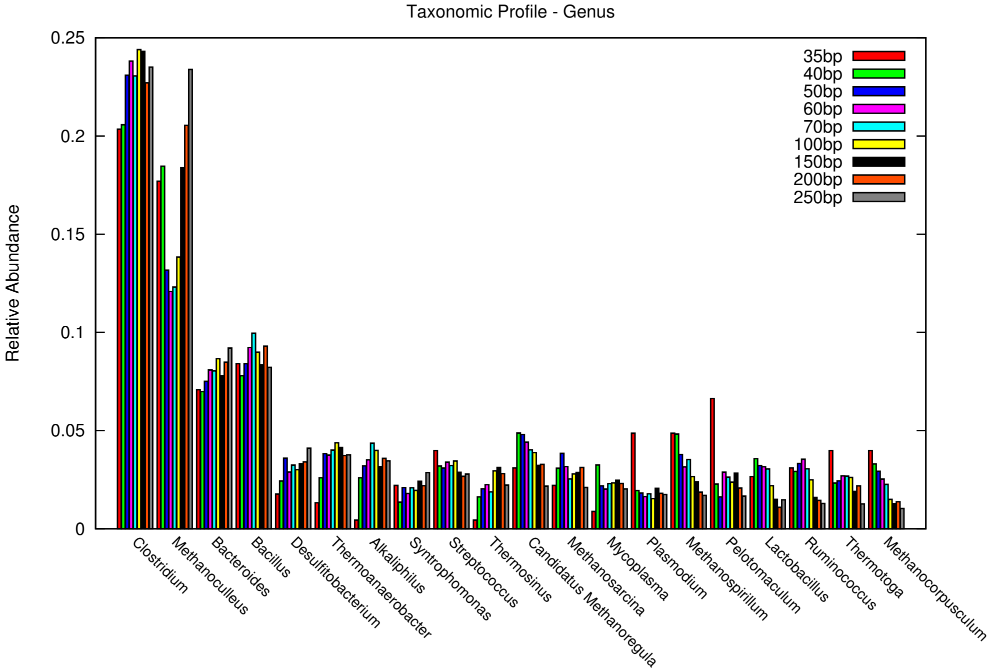


Figure A.32: Taxonomic results on the level of genus. Only taxa with an abundance of 0.015 or higher are shown.

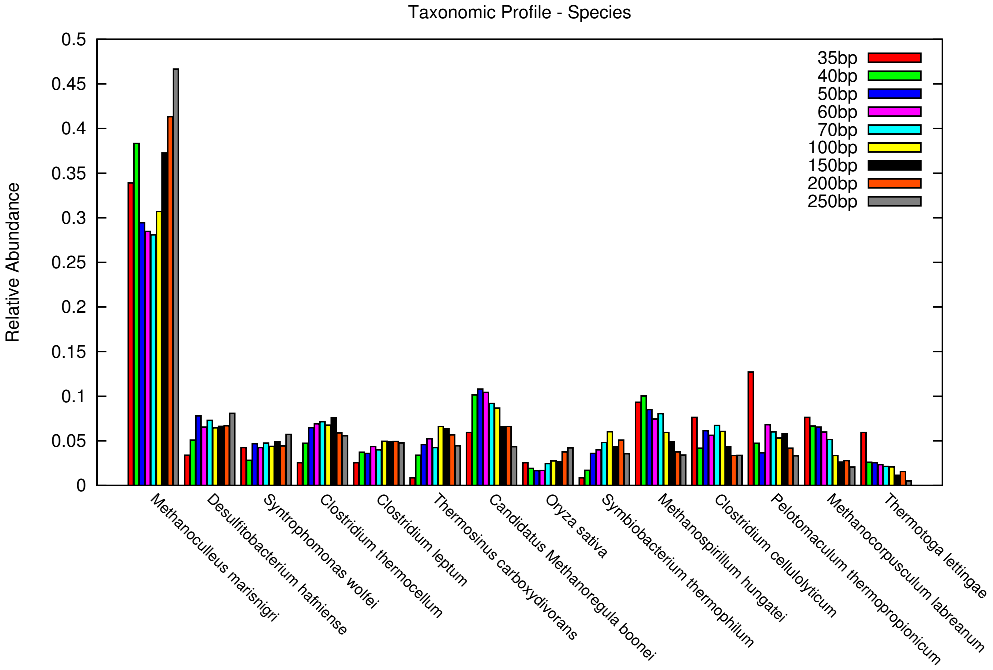


Figure A.33: Taxonomic results on the level of species. Only taxa with an abundance of 0.015 or higher are shown.