Bioinformatics analysis of whole-genome shotgun metagenomic data

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

vorgelegt von
Master of Science
REWATI MUKUND TAPPU
aus Pune

Tübingen
2016

Tag der mündlichen Qualifikation	13.12.2016
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter	Prof. Dr. Daniel Huson
2. Berichterstatter	Prof. Dr. Kay Nieselt

Summary

Next-generation sequencing technologies, with their low costs and high throughputs, have benefited the field of microbial research to a great degree. The application of whole-genome shotgun sequencing to DNA extracted from an environmental sample enables avoiding the usually complex method of cultivation of pure cultures of microorganisms in the laboratory. This protocol is referred to as whole-genome shotgun metagenomic sequencing. The analysis of sequencing data mainly aims at the taxonomic and functional characterization of the microbial sample. Many algorithms and tools have been developed for the same. The design of the analysis pipeline is usually dictated by the specific project at hand.

In this thesis, we describe several aspects of analyzing whole-genome shotgun metagenomic data. Analysis usually begins with the quality check of raw sequencing data followed by its preprocessing to improve the read quality. When dealing with datasets containing several number of large samples, the preprocessing of the samples can take up considerable time and effort. However, if the binning of reads into different taxonomic and functional categories is the aim, a read with bad quality automatically gets filtered making the initial preprocessing unnecessary. Thus we first look into the effect of preprocessing on the ensuing analysis of the metagenomic samples. Next, we assess the correspondence between the different systems of functional classification typically used for metagenomic analyses. The reference proteins in databases like the NCBI-NR may have none or multiple identifiers belonging to a particular classification system. Consequently, a read aligning to such a reference may be placed into a functional group depending on the mapping of the reference to functional identifiers. We study the correspondence between the different classification systems using a few metagenomic samples.

Further, we describe the analysis of a dataset of human gut metagenomic samples obtained from obese patients undergoing a weight-loss diet-intervention. The obese patients were also detected positive for non-alcoholic fatty liver disease (NAFLD) and Metabolic Syndrome. The analysis is carried out using the popular metagenomic analysis tools DIAMOND and MEGAN. This study was carried out in order to track the effect of the diet-intervention on the gut flora composition and to relate the clinical parameters like weight-loss, NAFLD and metabolic syndrome to the microbiome.

A metagenomic sample could be subjected to analysis based directly on the reads or on an assembly. Both methods have their pros and cons. We explore the differences seen in the taxonomic and functional compositions between those two strategies and conclude that both

provide similar results with minor differences depending on the sample being assembled. At the end, we describe how a gene-centric assembly can be carried out with the tools DIAMOND and MEGAN and demonstrate the usefulness of such a gene-centric assembly in a metagenomic analysis pipeline by carrying out a gene-centric assembly across different gene families and metagenomic samples.

Zusammenfassung

Das Gebiet der mikrobiologischen Forschung hat in großem Maß von Next-Generation-Sequencing Technologien aufgrund ihrer niedrigen Kosten und ihrem hohen Durchsatz profitiert. Die Anwendung von Whole-Genome-Shotgun-Sequencing auf DNA, welche vorher aus Umweltproben extrahiert wurde, ermöglicht es die hufig aufwendige Kultivierung von Isolaten der Mikroorganismen im Labor zu umgehen. Diese Methode nennt man Whole-Genome-Shotgun Metagenomik. Die Analyse von Sequenzierdaten zielt hauptsächlich auf die taxonomische und funktionale Charakterisierung der mikrobiellen Probe ab. Dafür sind viele Algorithmen und Tools entwickelt worden. Bei einem solchen Tool wird normalerweise das Design der verwendeten Analysepipeline speziell auf ein spezifisches Projekt hin ausgerichtet.

In dieser Arbeit werden verschiedene Aspekte der Analyse von Daten aus der Whole-Genome-Shotgun Metagenomik beschrieben. Solch eine Analyse beginnt normalerweise mit der Qualitätskontrolle der Rohdaten, gefolgt von der Aufbereitung der Daten mit Ziel die Qualität der reads zu verbessern. Bei sehr groen Datensätzen, welche sehr viele groe Proben enthalten, können diese Vorarbeiten bereits betrchtliche Zeit und Mühe in Anspruch nehmen. Falls jedoch die Einteilung der reads in verschiedene taxonomische und funktionale Gruppen das Ziel ist, wird ein read mit schlechter Qualität in der Regel ohnehin automatisch herausgefiltert, was diese anfngliche Vorverarbeitung unnötig macht. In dieser Arbeit wird der Effekt der Vorbereitung auf die nachfolgende Analyse der metagenomischen Daten studiert. Auerdem wird der Zusammenhang verschiedener Klassifikationssysteme, welche typischerweise für metagenomische Analysen genutzt werden, untersucht. Die Referenzproteine in Datenbanken, wie zum Beispiel NCBI-NR, können keine oder auch mehrere Attribute eines speziellen Klassifikationssystems besitzen. Daher untersuchen wir basierend auf metagenomischen Datenstzen die Vergleichbarkeit verschiedener Klassifikationssysteme.

Zustzlich wird die Analyse eines Datensatzes von menschlichen Darmmikrobiomproben beschrieben, die von adipösen Patienten genommen wurden, welche bei einer Diätintervention zum Gewichtsverlust teilnahmen. Diese übergewichtigen Patienten litten zustzlich an nicht-alkoholischer Fettleber (NAFLD) und dem Metabolischen Syndrom. Die Analyse wurde mit den hufig verwendeten metagenomischen Softwaretools DIAMOND und MEGAN durchgeführt. Diese Studie wurde ausgeführt um den Effekt der Diät auf die Zusammensetzung des Darmmikrobioms zu untersuchen und klinische Parameter, wie Gewichtsverlust, NAFLD sowie metabolisches Syndrom, mit dem Mikrobiom in Verbindung zu bringen.

Eine metagenomische Probe kann dabei entweder direkt auf den jeweiligen reads oder aber auf einer Assemblierung untersucht werden. Sowohl read-basierte als auch auf Assemblierung basierende Analysen haben ihre Vor- und Nachteile. In dieser Arbeit untersuchen wir die Unterschiede welche sowohl in taxonomischer als auch funktionaler Zusammensetzung bei beiden Verfahren beobachtet werden und stellen fest, dass beide ähnliche Ergebnisse liefern, wobei kleine Unterschiede abhängig von der assemblierten Probe auftreten können. Abschließend wird beschrieben, wie eine gen-zentrische Assemblierung mit den tools DIAMOND und MEGAN ausgeführt werden kann und zeigen den Nutzen einer solchen gen-zentrischen Assemblierung innerhalb einer metagenomischen Analysepipeline auf, indem wir diese auf verschiedene Genfamilien und Proben anwenden.

Acknowledgments

I am very grateful to Prof. Dr. Daniel Huson, for supervising and guiding me through the course of my PhD. Working with him, has been a great privilege. I especially enjoyed the stress-free and friendly atmosphere that he brings to the group.

I express my heartfelt gratitude to my colleagues at the *Algorithms in Bioinformatics* group. I am thankful to Sina and Ania for the fruitful lunchtime discussions about work and life. I thank Hans, Mohamed, Nico, Benjamin and Monika for making the office a fun place. The talks over coffee with them were always refreshing and provoked many new ideas.

I would also like to thank Dr. Suparna Mitra for her kindest support. I thank Dr. Sandrine Louis for her useful collaboration.

I extend my thanks to everyone at the Center for Bioinformatics (ZBIT) for being friendly and helpful. I have learnt a lot from the professors, colleagues and the students here.

The multicultural environment and myriad interesting people, made life in Tübingen exciting and rewarding. I would like to express many thanks to my flatmates, friends and acquaintances.

I have some wonderful people as my closest friends who have been an important part of my journey so far. I would especially thank Priya and Priyanka for their friendship. I thank all my friends from the Bioinformatics Center, Pune University for the joyous memories created as classmates and for the innumerable things that I learnt from interactions with them.

I thank my siblings, Ketaki and Chaitanya, for the great times spent together. Lastly, but most importantly, I thank my beloved parents, Mr. Mukund Tappu and Mrs. Manjusha Tappu for always standing by me and helping me to achieve my dreams. They have always cheered me up whenever I felt low. Without their motivation and support, this would not have been possible.

I dedicate this thesis to my parents.

Rewati Mukund Tappu
Tübingen, September 2016

Dedicated to Aai & Baba

In accordance with the standard scientific protocol, I will use the personal pronoun *we* to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Chapter 1

Introduction

1.1 Metagenomics

Metagenomics is the sequencing and analysis of DNA extracted directly from environmental samples. Next-generation sequencing technologies offer high-throughput sequencing at reduced costs and this has provided a methodological base to carry out metagenomic and metatranscriptomic studies. Metagenomics has opened up several doors to decipher the biology of the various microbial communities present in different environmental conditions. This chapter gives a brief introduction to the field of metagenomics and discusses the analysis tools DIAMOND and MEGAN which can be used in combination to taxonomically and functionally characterize a metagenomic sample.

1.1.1 The importance of microorganisms

Microorganisms are ubiquitous and are present in almost all types of terrestrial, freshwater and marine habitats and contribute roughly about 60% of the earth's biomass [1, 2]. They have established symbiotic relationships with animal and plant hosts which can be mutualistic, parasitic or commensalistic in nature [3]. Microorganisms play an important role in the ecosystem that they thrive in, for example in the biogeochemical cycles on the earth, like the carbon and the nitrogen cycle [4]. The agriculture, food, pharmacy and biotechnology industries benefit from the numerous uses of commercial microbes. Specific microorganisms are capable of producing certain proteins and enzymes that are commercially important, e.g. lactic acid bacteria that are used in dairy products [5] or extremophiles that are used for performing catalysis at extremely high temperatures [6]. Microorganisms associated with a host play a major role in the health of the host, as is the case with the gut microbiota which participate in the development of the innate immune system [7]. However, some microorganisms also are pathogenic and cause infectious or life-threatening diseases, like *Yersinia pestis* [8]. Due to the vast spectrum of the different roles that microorganisms play, they exhibit tremendous genetic, metabolic and physiological diversity. The genomes of different microorganisms have evolved to suit the environment that they have created a niche in [9]. Moreover, several different species of microorganisms could be a part of a niche, in that they exist not in single, but as an assemblage.

Owing to their importance in several walks of life, large amounts of efforts have been put into understanding their biology. Traditional microbiology relied on pure cultures to study microbes. This method, although successful, cannot be used for microorganisms that grow in extreme conditions because of which they are difficult to obtain as pure cultures in

the laboratory. The Polymerase Chain Reaction (PCR) [10] was a breakthrough molecular technique that allowed the amplification of entire genes and therefore the community profiling with 16S rRNA sequences. With the advent of genomic technologies, sequencing became popular and led to several bacterial genome projects being completed. The first organism to have its complete genome sequenced was *Haemophilus influenzae* [11] followed by the sequencing of the genomes of several other model organisms like *Escherichia coli* [12]. A huge wealth of information has been gathered due to whole genome sequencing of several microorganisms. It has lead to a greater understanding of molecular biology and has provided us with a wealth of information about novel genes and biochemical pathways. Insight into the evolution of microbes has been obtained using comparative genomics. This has also helped in understanding pathogenesis, aiding vaccine design, crop improvement and others [2].

In addition, advances in the throughput and reduction in the cost of next-generation sequencing technologies have paved the way for an alternative approach to study microbes, involving the extraction and direct sequencing of DNA from an environmental sample. Next-generation sequencing technologies have made it possible for the new science of metagenomics to evolve [13, 14].

1.1.2 Next-generation sequencing technologies

After the elucidation of the structure of the deoxyribonucleic acid (DNA) molecule [15], efforts towards the sequencing of DNA molecules began. Sanger sequencing, developed by Sanger et al., [16] is one of the most widely used traditional methods of DNA sequencing which makes use of capillary electrophoresis to separate growing polypeptides from a template DNA molecule. The proportion of deoxynucleotides and dideoxynucleotides ensures the chain termination of the growing polypeptide resulting in identification of the base present at that position. Sanger sequencing is less prone to errors and generates long reads. But due to less throughput and intense manual laboratory work, it is not favorable for generating huge amounts of data. The next-generation sequencing technologies like the Illumina, 454-Ion Torrent, SOLiD, PacBio etc. offer very high throughput and low cost and hence are the best suited for huge genomic or metagenomic projects [17]. 454 sequencing involves the PCR amplification of DNA fragments that are attached to a bead and their washing over a PicoTiterPlate. Each bead falls into one well of the plate which contains enzymes for the polymerization of the DNA molecule. The addition of the correct base releases a pyrophosphate which is imaged. SOLiD is another platform which uses sequencing-by-ligation using

Table 1.1: Different sequencing technologies and their chemistry, run time, read length and run types.

Sequencing technology	**Clonal Amplification**	**Chemistry**	**Run types**	**Run time**	**Read length**
454	Emulsion PCR	Pyrosequencing (seq-by-synthesis)	Single end	23 hours	700
Illumina®	Bridge amplification	Reversible dye terminator (seq-by-synthesis)	Single and paired end	12 days	2*100
SOLiD®	Emulsion PCR	Oligonucleotide 8-mer chained ligation (seq-by-ligation)	Paired-end sequencing	6 days	75 bp
Ion Torrent	Emulsion PCR	Proton detection (seq-by-synthesis)	Bidirectional sequencing available	4 hours	400 bp
PacBio®	N/A (single molecule)	Phospholinked fluorescent nucleotides (seq-by-synthesis)	single molecule	2 days	8,500 bp
Nanopore (MinION)®	N/A (Single molecule)	Sequencing as DNA molecule passing through a nanopore under an applied electric field	–	–	5400 to 10,000 bp

*Information derived from Glenn 2011, [19], Hodkinson et al., [17] and Feng et al. [20]

a dibase incorporation system. Ion Torrent detects the change in the pH due to the release of a proton when a new base is added to the growing DNA polymer. PacBio and Nanopore sequencing involves the sequencing of individual DNA molecules and produce the longest reads.

Since most of the metagenomic datasets analyzed in the following thesis have been sequenced with the Illumina technology, we discuss that in more detail. Illumina uses the sequencing-by-synthesis (SBS) chemistry for sequencing and uses a flow cell with oligos attached. It uses reversible dye-terminators for the sequencing reaction. Sequencing begins with the hybridization of specific adapter sequences on the ends of the DNA fragments and then washing of the flow cell with these fragments. Polymerization occurs to produce replicates of the same fragment. Reversible dye-terminators are then used to wash the flow cell followed by washing of excess nucleotides. The flow cell is then imaged. Illumina HiSeq 2500 sequencing offers very high throughput producing 4 billion fragments in a paired-end fashion with 125 bases for each read in a single run. The MiSeq platform produces the longest reads, with 300 bases in length each. In the Table 1.1, the sequencing technologies and their corresponding read-lengths have been mentioned. Illumina sequencing has been used extensively in metagenomic studies due to its high-throughput [18].

1.1.3 Whole-genome shotgun metagenomic sequencing

Amplicon sequencing has long been used for the sequencing of microbial communities, which involved the PCR amplification of only a certain phylogenetically important gene to study the microbial diversity of environmental samples [21]. Different gene sequences can be targeted for amplification of bacteria (16S rRNA), eukaryotes (18S rRNA and ITS) and viruses (g23 and RdRp) [22]. Shotgun sequencing of a genome was an established protocol. The application of this protocol to microbial communities was first pioneered in a study by Venter et al. [4], where sampling was carried out from the Sargasso sea to characterize the microbial community. With this seminal work, the protocol of shotgun sequencing was applied to the total genomic DNA isolated from microbial communities. Since then, several important microbiome projects like the Human Microbiome Project Consortium [23], MetaHIT [24], Earth Microbiome Projects [25] etc. have been undertaken and these projects have given us insights into the composition of the microbial communities in the respective environments. Metatranscriptomics and metaproteomics data are also now being generated resulting in the understanding of the complexity of microbial communities, their genetic potential and their effect on the environment that they are part of. Co-occurrence and correlation between different species can be studied with metagenomics. WGS has given the opportunity to discover novel genomes [26]. The main questions to be answered when analyzing metagenomic datasets are:

1. Who is out there?

This mainly involves determining the taxonomic content of the sample. It deals with retrieving information about which species are present and in what abundance, and how diverse the sample is in terms of different taxonomic groups present.

2. What are they doing?

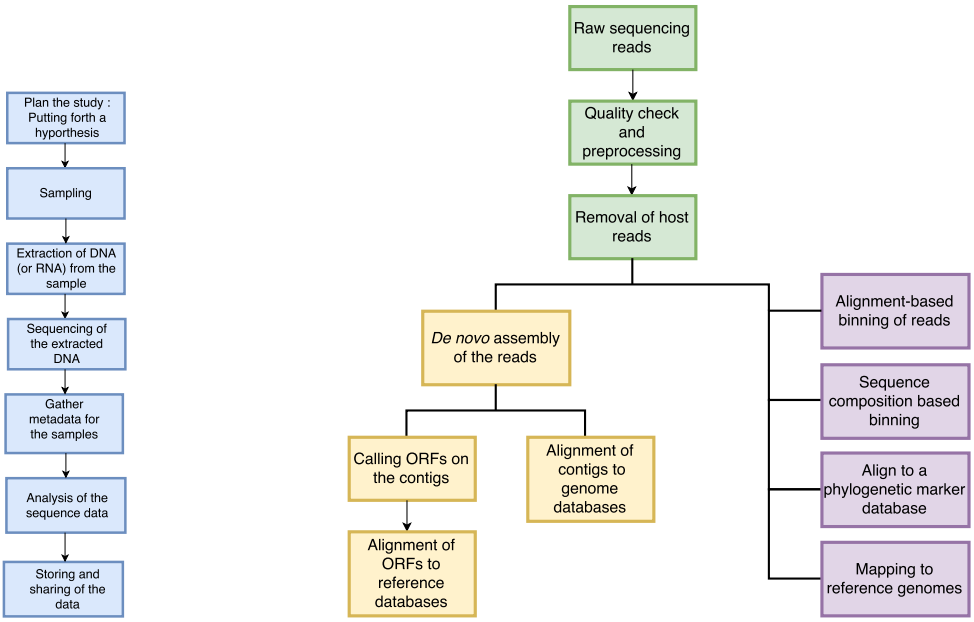
This mainly involves determining the functional content of the sample. In particular, it involves cataloging of the genes, modules and pathways present in the sample.

3. How do they compare?

This involves finding the similarities and differences between related samples. Different metagenomic samples can be compared with each other in terms of their taxonomic and functional content.

4. How do they relate to the environment they come from?

The correlations between the environmental factors and the taxonomic and functional content can be made to understand the biology of the microbial communities better.



(a) Steps involved in a metagenomic study. (b) Bioinformatic analysis of metagenomic samples.

Figure 1.1: The above schematic depicts a generic outline of a metagenomic study. It also lists the various options available for the analysis of the data.

A sufficient depth of sequencing from the DNA extracted is representative of the microbial community that it has been isolated from. Sequence analysis of this DNA allows the characterization of both the taxonomic and functional profiles of the communities. In a review by Thomas et al., [27] the different stages in a metagenomic workflow are described. Figure 1.1 shows a general workflow of the metagenomic pipeline and the bioinformatic analysis of the sequencing reads. Experiment design and planning is of utmost importance. The biological question to be answered using sequencing has to be well defined. The microbial community to be studied has to be sampled effectively before DNA can be extracted. When sampling, it is of advantage to consider biological or technical replicates for a later statistically sound analysis. Also, sampling along several time-points offers more insight into community dynamics [14].

DNA isolation should be carried out carefully with the appropriate laboratory protocol and contaminants should be avoided. The right sequencing technology and a good depth of coverage and total coverage is essential for obtaining a sufficient sampling and sequencing. In addition, recording the metadata concerning the samples is essential. This could involve physical characteristics of the sample like the treatment group of the sample, location, pH, or the time point that the sample was obtained from, or clinical parameters in the case

where the sample is from a host. Raw sequencing data can be stored along with the related metadata in public databases like the Sequence Read Archive, (SRA) [28].

1.1.4 Bioinformatics analysis of WGS metagenomic data

The huge amount of sequence data that metagenomic experiments produce necessitates the usage of advanced computational techniques for their analysis. Sequencing technologies are prone to error and hence the first step is the quality checking of the raw reads. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is a software commonly used to check the quality of the raw data. Based on the quality, the raw reads can be processed for further improvement. FASTX (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and PRINSEQ [29] offer functions to check read lengths, filter based on read quality scores, trim poly A/T tails, Ns etc. Specific tools for the removal of adapter sequences are also available and they include Cutadapt [30], Trimmomatic [31] and others. If filtering of host reads is necessary, then the reads are mapped using mapping tools like BWA or Bowtie to a genome of the host, if available or a closely related genome. The non-host reads are then analyzed. Several tools exist for the sequence analysis of raw reads and they can be classified into three main categories depending on the underlying algorithm they use [32]. The first one is sequence similarity searching or homology searching that involves the alignment of the reads to a database of reference sequences and this returns the reads and their matches to the reference sequences in the database. The most commonly used sequence alignment tool, BLAST [33], with its different flavors like BLASTN, BLASTX, TBLASTX and others are used in various scenarios. The National Center for Biotechnology Information (NCBI) holds several reference databases like NR, NT and GenBank. In some cases, custom databases are created if the aim is to analyze particular types of sequences. Several tools, CARMA [34], MG-RAST [35], Kraken [36] and MEGAN [37] carry out the binning of raw reads. The second kind of sequence analysis is composition based analysis where the k-mer compositions are determined and used for binning of the reads. PhyloPythia [38] and PhymmBL [39] are software tools that use sequence composition to classify sequences. The third type of analysis is extraction of phylogenetic marker genes and their subsequent comparison to a database of marker genes to determine the taxonomic composition of the metagenomic sample. Tools like MetaPhyler [40] and MetaPhylAn [41] use clade-specific marker genes for placing reads into a taxonomic group. Assembly of the metagenome to retrieve individual genomes can be performed using tools developed specifically for the same. SOAPdenovo [42], Ray [43], IDBA-UD [44] are some tools that are developed for this purpose. Statistical analysis and

visualization of multiple metagenomes is an important component of the pipeline. Tools like STAMP [45], MetaStats [46] and R packages like metagenomeSeq [47], and Vegan [48] contain functions for statistical tests to determine differences in the relative compositions of different species or gene families present in the sample.

1.2 The DIAMOND and MEGAN pipeline

Time efficient analysis of metagenomic data was previously a huge bottleneck. DIAMOND, written by Buchfink et al [49], is a very fast tool that alleviates this problem to a great extent. It has been shown to be upto 20,000 times faster than BLASTX when aligning short reads against the NCBI-nr database and with the same or similar sensitivity. DIAMOND is an open source software and it can be downloaded freely from the website - http://ab.inf.uni-tuebingen.de/software/diamond/. It uses a reduced alphabet, spaced seeds, increased seed length and double-indexing to speed up the database searching. The creation of a DIAMOND index for the NR database release of March 2016 took about 2.5 hours using 15 cores of a 32 core processor. A DIAMOND Alignment Archive output or DAA file is created as output consisting of the matches of a read against the sequences of a database. Options to set E-value and percent identity thresholds exist. The output DAA file can be opened in the popular metagenome analysis tool, MEGAN, which now has a new release, MEGAN 6 [37].

MEGAN is MEta Genome ANalysis tool that is used for the analysis of metagenomic or meta-transcriptomic data. A DAA file has to be "meganized" before it can be opened in MEGAN, using a tool called Meganizer, which is distributed as part of the software. Meganizer takes as input the DAA file and several mapping files, like GI to taxonomy, GI to KEGG, GI to COG, GI to InterPro2Go, GI to SEED and others. Based on the alignments obtained for a given read, it is placed in a taxonomic group using the Lowest Common Ancestor (LCA) algorithm. The LCA algorithm traverses through the hits obtained for a read and for all the hits that have the same bit score, it places the read in a taxonomic bin that is the lowest common ancestor of those hits. Alternatively, more specific assignments are made if the read scores a "best" match to a taxonomic group. Several parameters like the **MinScore**, **MinSupport**, **MinPercentIdentity**, **TopPercent** and others control the behavior of the LCA. The default taxonomy used is the hierarchical NCBI taxonomy containing the levels Domain, Kingdom, Phylum, Class, Order, Family, Genus, Species. The functional classification of a read is based on the KEGG, SEED, COG, PFAM and

InterPro2GO hierarchies. For outputs of a BLAST search or of other alignment programs, the "Import from BLAST" option is used. This dialog box allows the user to specify the mapping files required and as output produces the Read Match Archive (RMA) file which is an indexed file with reads and its matches, alignments, and classifications. Once a MEGAN file representing a metagenomic sample is created, it can be inspected and analyzed. Several plotting techniques like bar plots, pie-charts, bubble diagrams and heat maps are available in MEGAN to visualize samples. Multiple files corresponding to different metagenomic samples can be opened together and a comparative file can be created. PCoA analysis of multiple samples can be carried out using the Bray-Curtis, Jensen-Shannon diversity index and others. Alpha and beta diversity can be inspected using MEGAN. The number of reads assigned to each taxonomic or functional group is called the "read-count" data. This read count data can be exported as a tab-separated file. Also, tab-separated files containing information about the read names to the taxonomic or functional assignments and reads can be exported. MEGAN also allows the incorporation of metadata associated with a metagenomic file. In addition, the MEGANServer has been developed to access samples served on a public server. This makes it easy to use already analyzed samples without the need to download them.

1.3 Conclusions

This chapter introduces the field of metagenomics and describes the analysis strategies and software available for getting the most out of sequencing data. The DIAMOND and MEGAN pipeline is a fast, easy and straightforward solution to analyzing large whole genome shotgun metagenomic datasets.

Chapter 2

Preprocessing of raw metagenomic data for read quality improvement

2.1 Introduction

Sample preparation and sequencing using next-generation sequencing technologies is associated with certain error rates [50]. Different sequencing platforms have different error rates that depend on the underlying chemistry [51]. For example, the Illumina platform is known to produce single nucleotide substitution errors. The raw sequencing data could have a large number of "Ns" and low quality bases. The occurrence of poly-A/T tails is also very common. As mentioned in the previous section, one of the first steps in the analysis of metagenomic data is the processing of the raw reads for improvement of the read quality. FastQC is a popular software that is used for the initial assessment of the quality of the sequenced reads. It generates a quality report showing the per-base quality, per-base GC content, per-sequence quality and others. Accordingly, the raw reads need to be trimmed, or filtered. Many software tools exist for the same purpose, like FASTX toolkit, PRINSEQ etc. Tools like Cutadapt and Trimmomatic exist for the adapter removal from sequencing data. Samples obtained from a host warrant the removal of host-associated reads and mapping tools are used for carrying out the filtering of such reads. In metagenomic analyses, where binning of the reads is the goal, it may not be important for a low quality read to be filtered, because such a read may fail to obtain any alignment, automatically filtering it out. In this chapter, we assess the effect of preprocessing the raw sequence data for read quality on the ensuing metagenomic analysis.

2.2 Metagenomic samples

We selected five samples from the MetaHIT Project. We used the raw reads and ran a quality check on them using the FastQC software. Depending on the results, we designed a pipeline for the quality processing of the raw reads. We then compared the metagenomic samples before and after processing, with respect to their taxonomic and functional profiles. Table 2.1 shows the number of reads before and after preprocessing and the time required for the analysis.

2.3 Quality check and preprocessing

The FastQC quality report listed the per base sequence quality, the per sequence quality scores, and sequence length distributions as the parameters whose values were not satis-

Table 2.1: IDs of the MetaHIT samples used, the total number of reads before and after preprocessing and time required for analysis.

Samples	**Raw**				**Preprocessed**				
	Number of reads	DIAMOND Time (m)	Meganizer Time (s)	Total time (s)	Preprocessing Time (s)	Number of reads	DIAMOND Time (m)	Meganizer Time (s)	Total time (s)
bgi-MH0025	55,611,424	198	1,918	13,828	3,114	15,177,808	54	671	7,035
bgi-MH0031	55,568,300	198	1,210	13,110	3,082	9,843,691	35	295	5,485
bgi-MH0035	49,322,484	176	1,955	12,517	3,205	12,322,475	43	306	6,149
bgi-MH0072	49,297,218	175	2,356	12,913	3,293	2,835,957	10	99	3,999
bgi-MH0078	26,055,172	94	936	6,576	1,789	17,246,905	61	597	6,046

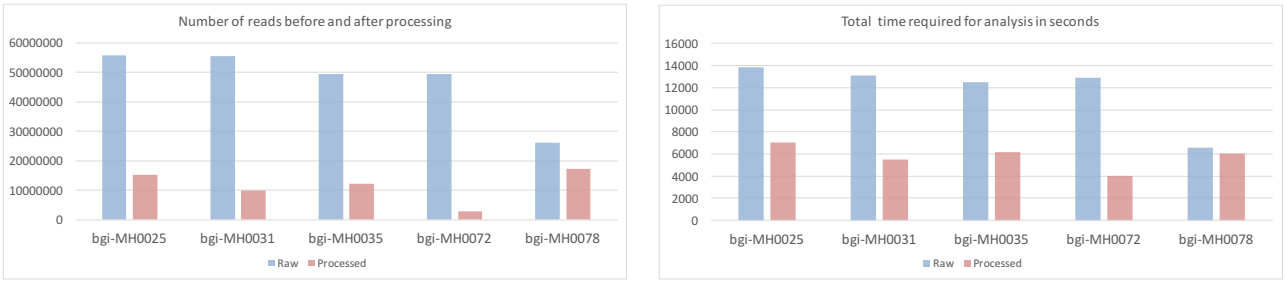


Figure 2.1: The number of reads after preprocessing decrease significantly as also the time required for their analysis.

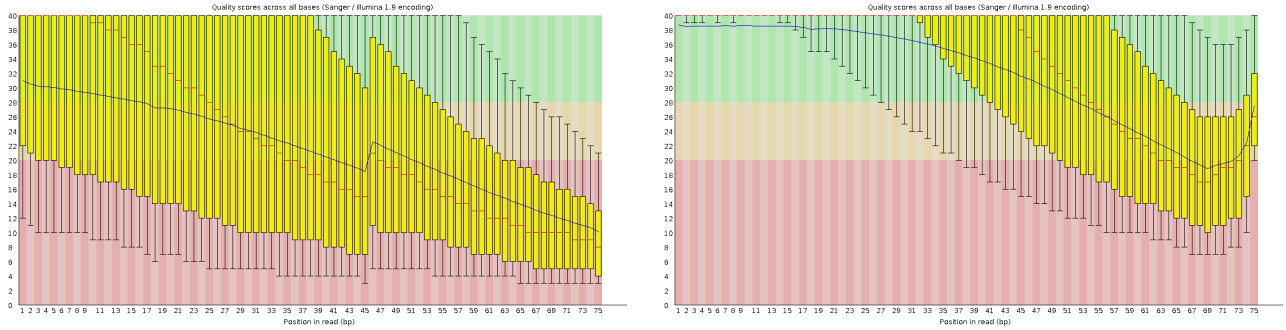
factory. The PRINSEQ parameters were chosen in accordance with the report. They can be summarized as follows: min_len (minimum length of the read) which was set to 70, min_qual_mean (average minimum quality of the bases in a read) which was set to 20, trim_qual_right (trimming of the right end of the read with quality below the specified quality) which was set to 20. For sample bgi-MH0078, since it had a huge number of duplicates, the derep option in PRINSEQ was used to remove exact duplicates. After the quality check, the number of reads got reduced significantly in all samples. The time required to run the analysis on all these samples was plot as a bar chart in Figure 2.1

2.4 Results

In order to check whether processing indeed affected the overall quality of the samples, FastQC was run again on the processed samples. Figure 2.2 (output of the FastQC software) shows the per base quality of the raw versus the processed samples.

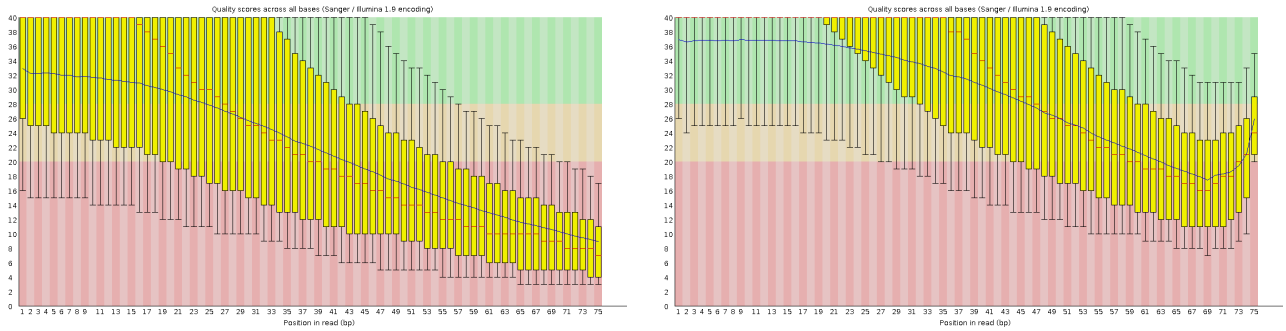
For parameters such as the sequence quality and duplication levels, there was improvement in the values after processing. Table 2.2 summarizes these values.

A MEGAN comparison file was then created from the 10 samples - (5 raw samples and 5 processed samples). A principle coordinate analysis (Figure 2.3) using the JensenShannon



bgi-MH0025 - Before processing

After processing



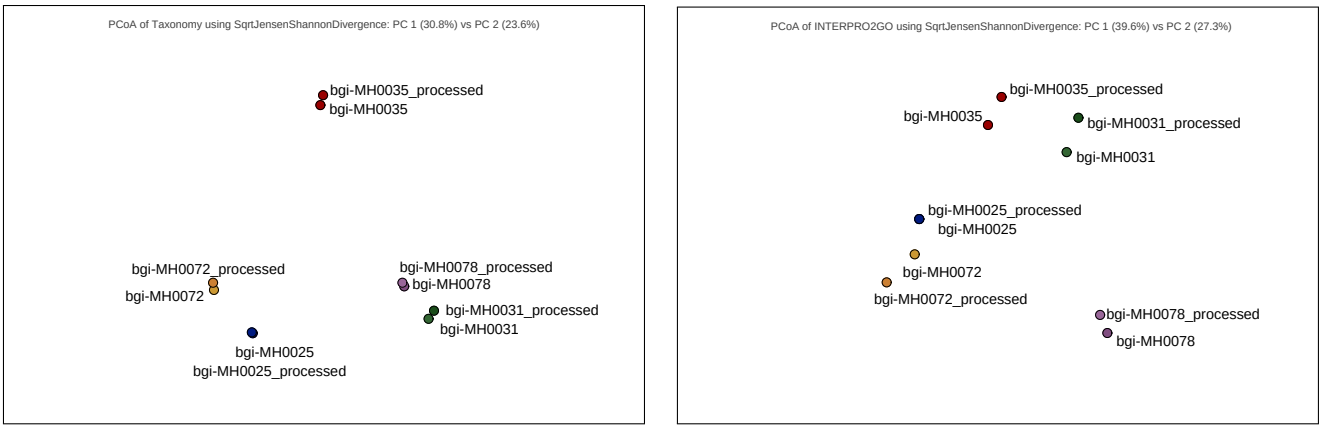
bgi-MH0035 - Before processing

After processing

Figure 2.2: The FastQC reports for the quality scores before and after preprocessing, for two samples.

Table 2.2: Table showing the values of the sequence quality before and after processing.

Samples	Raw : Sequence quality	Processed : Sequence Quality	Raw : Duplication levels	Processed : Duplication levels
bgi-MH0025	31	34	4.3%	2.66%
bgi-MH0031	16	31	8.56%	2.74%
bgi-MH0035	17	32	9.96%	3.61%
bgi-MH0072	15	39	8.8%	1.1%
bgi-MH0078	39	39	43.76%	0%



PCoA on species classifications PCoA on InterPro classifications

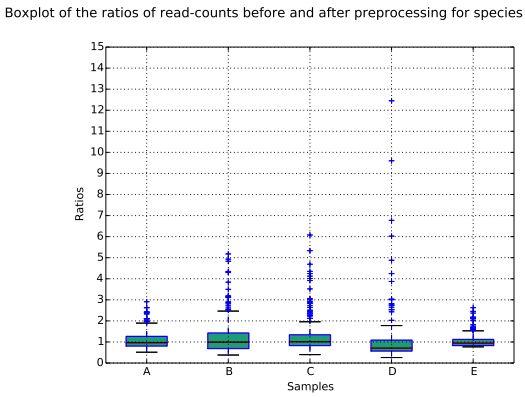
Figure 2.3: PCoA in MEGAN for the ten samples at the level of taxonomy and function (InterPro classifications.)

divergence index on the taxonomy and functional classification (InterPro2GO) of the samples shows that each pair of raw and preprocessed samples clusters together, indicating that raw and preprocessed samples do not differ to a great extent.

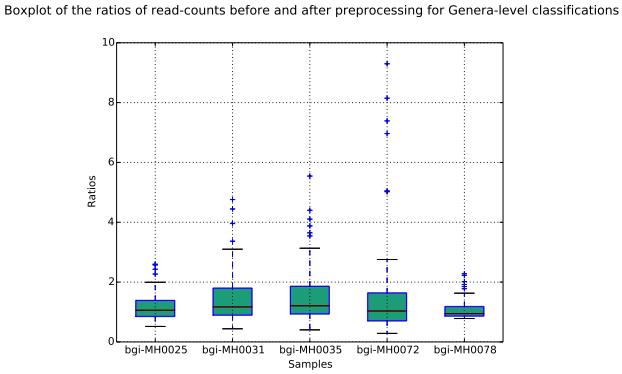
Further, read-count data was extracted from the comparison file and the difference in the read counts for the raw and processed samples was assessed. For each feature in the taxonomy (species and genus level), KEGG, eggNOG and InterPro classifications, the ratio of the read counts before and after processing the samples was determined. A ratio of 1.0 indicates that the read count did not change and greater than 1.0 or less than 1.0 indicates decrease and increase in the read count pointing to the fact that processing changed the analysis results. The boxplots in Figure 2.4 are the ratios for each features in all the classifications considered. The scatterplot shows the average ratios.

For all the classifications, a majority of the features have a ratio close to 1, with there being several features having a high value for the ratios. This indicates that after preprocessing, changes at the read-count level do occur. As can be seen, for bgi-MH0072, the preprocessing filtered a lot of reads and subsequently, this sample has a high average ratio for all the classifications tested.

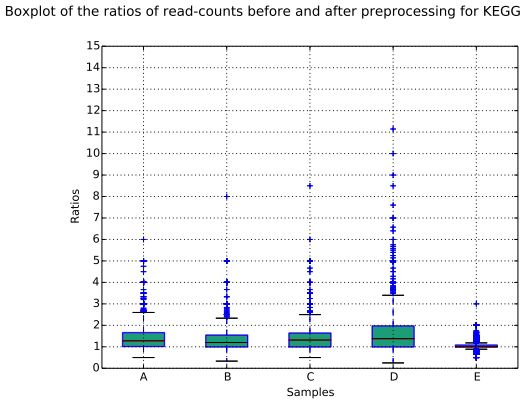
To determine whether the bad quality filtered reads are assigned to a taxonomic or functional classification in the raw unprocessed samples, the read-name to taxonomic assignments for the raw samples was extracted. These assignments were checked for the presence of any bad quality read. In all 5 samples, the bad quality reads were not a part of the taxonomic profile of the raw samples. Thus the low-quality reads are excluded during the read assignment process.



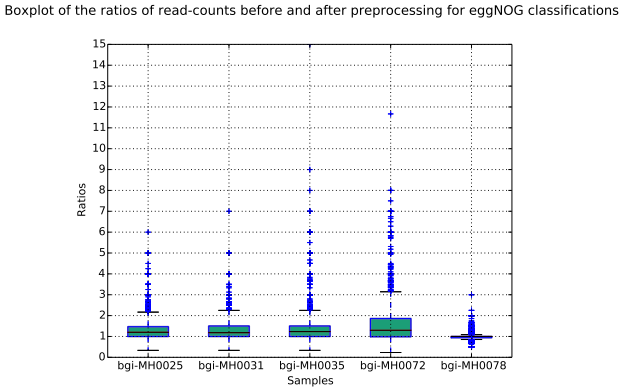
(a) Species



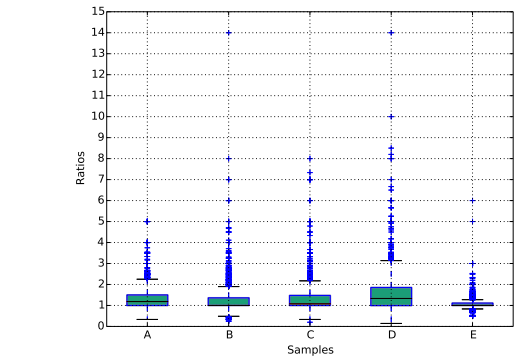
(b) Genera



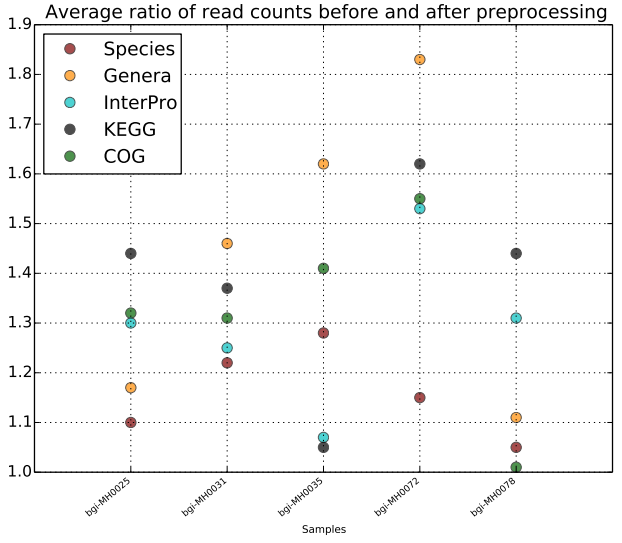
(c) KEGG



(d) eggNOG



(e) InterPro



(f) Average ratios for all features.

Figure 2.4: The ratios of the read-counts for the individual features at all levels of classification before and after preprocessing tend to be close to 1. In comparison to the species level classifications, the genera level classifications tend to have higher ratios, since it is a higher level of classification.

2.5 Conclusions

Metagenomic analysis starts with a quality check of the samples in consideration. Although quality checking is essential, it might not always lead to a significant change in the taxonomic and functional profiles obtained for metagenomic analysis. Preprocessing however, does improve the quality of the samples as determined by experiments on the five samples considered. However, a considerable amount of time has to be devoted to checking the quality and to decide what parameters are suitable for the preprocessing of the data. This time required is not accounted for, but with a large number of samples, it is certain that this could take up a considerable amount of time. At the read-count level, the processed and raw samples show some differences. While an improvement in the read quality of the samples is certain, it might not be a significant improvement. Thus, it is important to process the samples and check sample quality, but in some cases it might be skipped because it does not affect the final results to a great extent. It is recommended in the case where the dataset is small.

Chapter 3

Correspondence between different systems of functional classification

3.1 Introduction

After alignment of newly sequenced genomic or metagenomic data to reference sequences, different systems for functional annotation can be used to put the reads into a biological context like an orthologus group, pathway or a module. Several systems of functional classification exist, like the KEGG [52], SEED [53], InterPro [54] and eggNOG [55]. These systems are employed for functional classification in MEGAN. Depending on the alignments of a read to reference proteins, the reads are binned into a functional classification system based on the membership of the reference protein to any orthologous group in the classification system. In the study described in the chapter, we aimed at comparing three different classification systems, the KEGG, InterPro and eggNOG (COG).

3.2 Mapping the classifications for reference proteins from the NCBI-NR database

The NCBI-nr database released in June 2016 was downloaded from the NCBI FTP which has a total of 89,362,690 sequences. The number of unique GI (Gene Identification) numbers from the database were 2,80,103,394. This was compared with the mappings from GI to KEGG, GI to COG and GI to InterPro. The comparisons yielded the percentage of the total number of GI's that have all three (KEGG, COG and InterPro) assignment, the percentage that have only one of the three and the percentage that has any two assignments. The Venn diagram in Figure 3.1 is annotated with these values. Only about 51% of the total number of GIs have at least one identifier associated which leaves about 49% of the total number of GIs in the NR database without any identifier.

3.3 Mapping the classifications for reads from metagenomic samples

Next, we wanted to determine the correspondence between the classification systems by tracking the assignment of each read in a metagenomic sample to the different systems. For a given read, we determine which KEGG, COG and InterPro group it bins into. Three metagenomic datasets analyzed with the DIAMOND and MEGAN pipeline were used for this purpose. One is a mock metagenome consisting of 64 microbial species [56], one is a MetaHIT

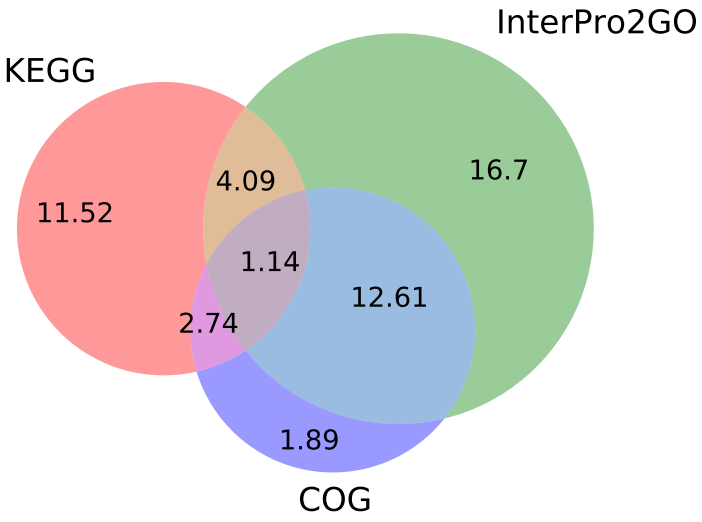
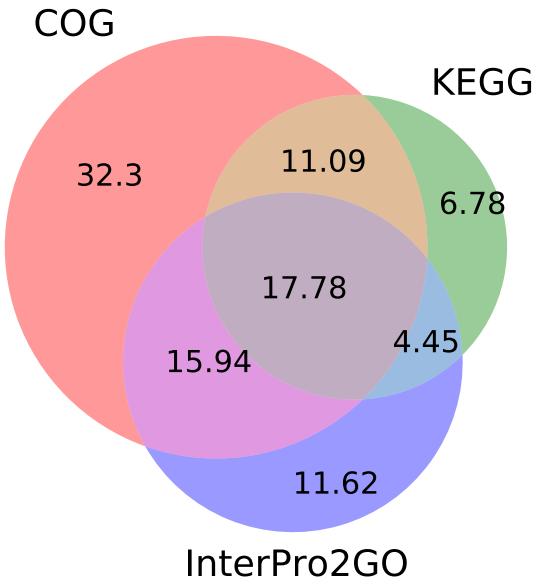


Figure 3.1: Percentage of the total number of GIs in the NCBI-nr database that have all, two or one assignment. The percentage of the GIs that have all identifiers for all three classifications is very low as compared to the percentage of the GIs that have only a KEGG or InterPro identifier associated.

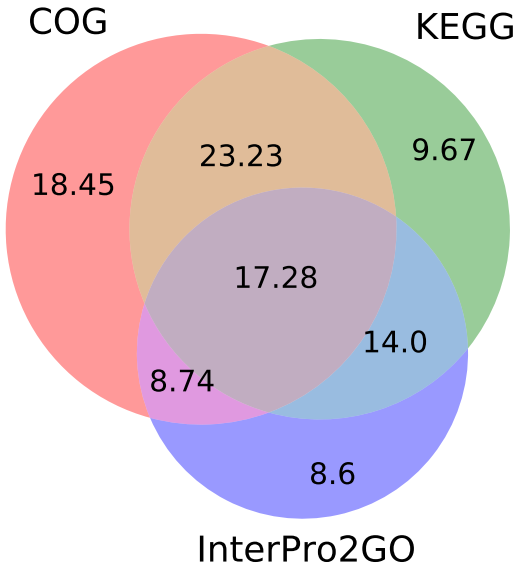
sample previously used for the preprocessing study and the third is a human gut microbial sample from the study by Louis et al. [57]. For the three classification systems, CSV files with the read name and the identifier of the category that it is binned into were downloaded. Starting with this information, the number of reads that got assigned exclusively to an eggNOG category, or a KEGG group or an InterPro group was calculated. The number of reads that got assigned by all the three categories and the number of reads that got assigned to two systems but not the third one was calculated. The Venn diagrams in Figure 3.2 depict the percentages of the total reads that each category amounts to.

From these analyses we conclude that the assignment rate of the COG classification system is the highest in the three samples studied. There is a consistency between the results for the three samples, as is evidenced from the values. One identifier from any functional classification system is expected to map to multiple identifiers from the other classification systems. For each identifier from each classification system, the number of different KEGG, eggNOG and InterPro identifiers it maps to was determined. An average of these values was calculated, shown in Table 3.1

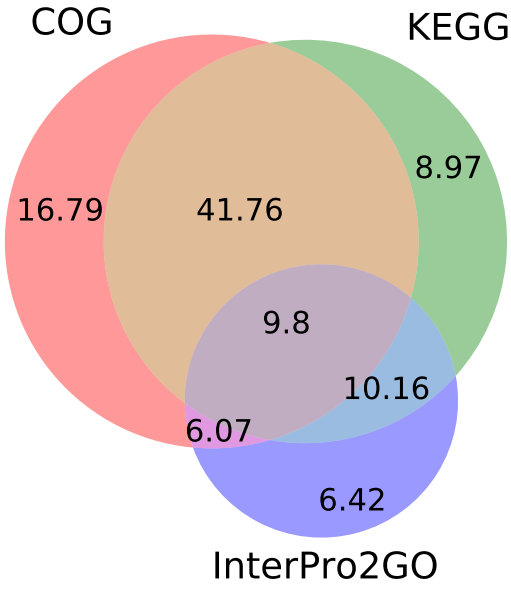
That the scores tend to have a value greater than 1, reveals that each identifier may map to multiple other identifiers in other classification systems. The highest values are for the InterPro identifiers. This may be because InterPro mostly characterizes protein domains,



(a)



(b)



(c)

Figure 3.2: Venn diagrams depicting the overlaps between the classification systems. (a)-(c) : Percentage overlap shared by the reads in the samples, (b) : Mock metagenome, (c) : AS50_0 (d) : bgi-MH0025

Table 3.1: Average number of the different orthologous groups mapping to a given identifier.

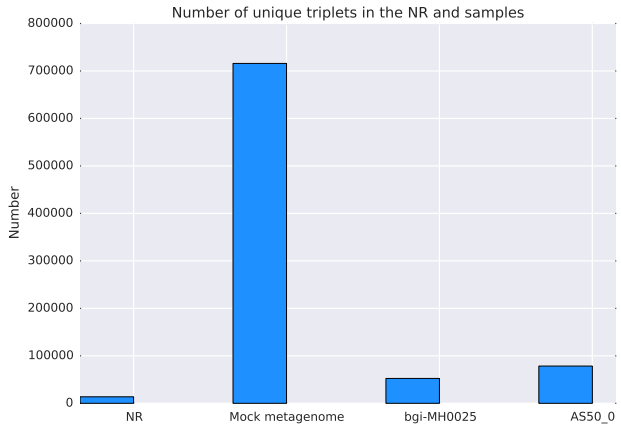
Classification system	**KEGG**	**InterPro**	**COG**
KEGG	–	1.52	1.47
InterPro	1.88	–	2.37
COG	1.71	1.69	–

Table 3.2: Table of the shared and non-shared triplets.

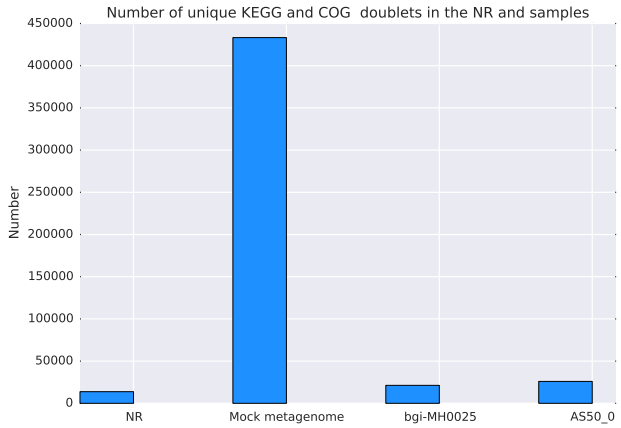
Type	Sample	Unique in NR	Total in read-mappings	Present	Absent
Triplets	Total	13,734			
	Mock metagenome		715,902	1,578	714,324
	bgi-MH0025		52,415	1,332	51,083
	AS50_0		78,474	1,256	77,218
Doublets InterPro and KEGG	Total	18,218			
	Mock metagenome		150,588	1,925	148,663
	bgi-MH0025		42,000	1,522	40,478
	AS50_0		30,170	1,297	28,873
Doublets KEGG and COG	Total	13,779			
	Mock metagenome		433,286	2,373	430,913
	bgi-MH0025		21,287	1,694	19,593
	AS50_0		25,945	1,509	24,436
Doublets InterPro and COG	Total	24,429			
	Mock metagenome		595,986	2,784	593,202
	bgi-MH0025		58,591	1,836	56,755
	AS50_0		35,722	1,565	34,157

and a single protein may have multiple domains.

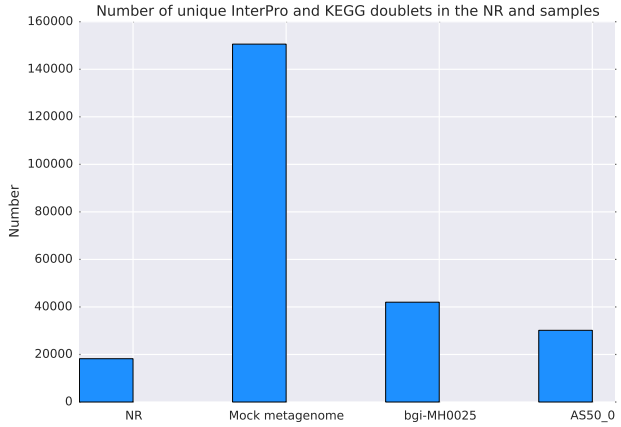
We expect that the mappings derived from the publicly available mapping files will be the same as the mappings obtained from the metagenomic samples. However, an analysis of the triplets, that is read-mappings of the InterPro-KEGG-COG revealed that from the metagenomic datasets, triplets could be found that do not occur in the NR mappings. However, some of them are common. The unique GI-KEGG-InterPro-COG mappings were considered. Then from the three datasets, unique KEGG-InterPro-COG triplets were extracted. The Table 3.2 contains the number of unique triplets and doublets and the number of triplets or doublets that appear in both the NR mappings and read mappings. Figure 3.3 shows the number of unique triplets and doublets encountered in the NR and the samples.



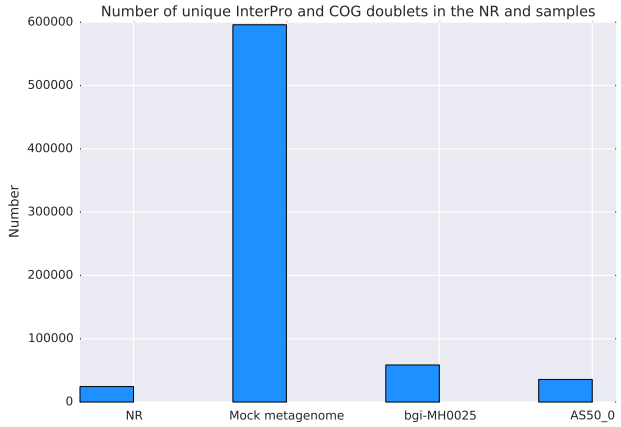
(a) Unique triplets in NR and samples



(b) Unique KEGG and COG doublets



(c) Unique InterPro and KEGG doublets



(d) Unique InterPro and COG doublets

Figure 3.3: The number of unique triplets and doublets as derived from the NR and the metagenomic samples. The number of triplets and doublets is very high in the mock metagenomic sample.

The unique number of triplets and doublets in the read-mappings is much more than what is present in the GI to functional category mappings. This is because for a given read, it may align to multiple reference sequences and the sequence that the read is finally assigned to may not be associated with an identifier from all the classification systems. This difference is especially stark in the mock metagenomic sample, where the number of reads is very high.

3.4 Conclusions

The comparison of the different systems for functional classifications reveal that the orthologous groups as defined by a given classification system may correspond to one or more orthologous groups from other classification systems. Different systems have different rates of assignment. One orthologous group in a given system of classification in effect corresponds to multiple different groups in another system of classification. It is recommended that multiple different systems of classification be used when analyzing metagenomic samples. Comparing the annotations for a given read as per each classification system is also informative. In the future, efforts could be concentrated towards creating a meta-viewer which combines the annotations and simultaneously classifies a read into multiple classification systems.

Chapter 4

Analysis of a dataset of gut metagenome samples from obese patients

Metagenomics has proven to be of great utility in studying the different microbes present on and inside the human body, particularly the human gut microbiota. The intestinal flora play a vital role in the health of its human host and their dysbiosis has been proved to be instrumental in shaping several diseases like obesity, metabolic syndrome and non-alcoholic fatty liver disease. Studies have suggested that strong connections exist between gut microbiota, diet and obesity. In this chapter we discuss the analysis of the metagenomic samples obtained from shotgun sequencing of fecal samples from 16 obese patients undergoing a formula diet treatment for weight-loss. We discuss the key results and observations from the analysis.

4.1 The human gut microbiota

The human body is inhabited by millions of commensal microorganisms that help in the normal functioning and the maintenance of good health of the human body [58, 59]. The number of microbial cells outnumber the somatic cells [23] and contribute to the biomass. Tremendous diversity of bacteria, archaea, viruses and fungi are present as part of the microbiome. In recent years the advancement of genomic technologies have made it possible to characterize the human microbiome [60]. For example the Human Microbiome Project aimed at sequencing samples from different individuals and various body sites like the oral tract, gastrointestinal tract, the skin and others. These studies have revealed that microbes have created specific niches on the human body [61].

One of the most complex and intriguing microbial ecosystems in the human body is the gut microbial system. The gut microbiota possess genes that are important in the metabolism and hence they are extremely important and are deemed a vital organ of the human body [62]. Colonization of the intestine begins at birth and rapidly changes and attains an adult stage, with major changes occurring again at old age. They are composed of trillions of cells representing several different species [7]. Species from seven phyla are present in the gut namely, Bacteroidetes, Firmicutes, Actinobacteria, Verrucomicrobiota, Tenericutes, and Fusobacteria. Bacteroidetes and Firmicutes are the groups that are most predominant. They harvest nutrients from the diet that are otherwise inaccessible to the host. Many species especially from the Bacteroidetes phyla possess polysaccharide degrading enzymes that their mammalian hosts do not have. The gut microbiota take part in fat storage. They carry out the metabolism of drugs and other xenobiotics and protect the host from colonization by pathogens. Development of the innate immune system also depends on the gut microbiota.

They also play a role in the production of vitamins [63]. In spite of similarities at high taxonomic levels like the phylum, the gut microbial composition varies substantially between individuals across geographical locations and age [64]. In the study by Arumugam et al., [65] three enterotypes corresponding to three dominating species were described. Apart from the enterotypes, many factors contribute to the variability of the gut microbiome such as the age and lifestyle. The gut microbiome is also known to vary over time. In addition, several disease states have been associated with the increase or decrease in relative abundance of specific species. Dysbiosis of the the gut flora has been linked to inflammatory bowel disease [66], obesity [67], colorectal cancer [68] and other medical conditions.

4.2 Hohenheim Obesity Project

4.2.1 Gut microbiota, diet and obesity

Obesity is an important public health issue in recent times and is associated with co-morbidities like metabolic syndrome and non-alcoholic fatty liver diseases. Obese people are prone to high blood pressure. Many factors such as lifestyle, eating habits and genetics contribute to it. Proper treatment and care is of huge importance if the patient is known to be clinically obese, as it can even cause life threatening diseases like cardiovascular diseases [69].

The gut microbiota have been shown to play a role in obesity in several mouse and human experiments. Dysbiosis of the gut microbiota has been seen in obese versus lean patients [70]. By fermenting the indigestible polysaccharides, the gut microbiota contribute to the energy harvest from diet. This has mostly been linked to the relative abundance of members of the Firmicutes and Bacteroidetes phyla [71].

Diet is known to have a rapid and reversible impact on the gut microbiota. Diet interventions are a common method of treating obesity and they primarily have an effect on the composition of the gut microbiota [72]. Studies that look at the effect of diet on the gut microbiota for a long period of time are scarce. In the work described in this chapter, we were interested in examining the gut microbiome of obese patients who underwent a diet intervention over a period of two years.

The gut microbiome of 16 obese patients who participated in a weight-loss intervention was characterized. Formula diet was used as intervention and DNA extracted from fecal samples was sequenced. Weight-loss, non-alcoholic fatty liver disease (NAFLD), and Metabolic Syndrome were monitored. This dataset will be further referred to as the “Ho-

henheim Dataset" and the study "Hohenheim Obesity Project", described in Louis et al., [57]. [1]

4.2.2 The study

Selection of subjects

16 obese patients from a multi-center clinical trial and research project "Obesity and the gastrointestinal tract" (ClinicalTrials.gov identifier: NCT01344525) were selected. At the start of the study, written and informed consent was obtained from all the patients. The patients were excluded for chronic or current gastrointestinal disease, severe eating disorders, and treatment with anti-, pre- or probiotics within 3 months before collection of the samples. All patients belong to the Bacteroides-enterotype as determined through the sequencing of the first sample. The patients with a similar BMI and a similar age at the start of the study were considered. The final cohort consisted of 16 subjects (9 women) with a mean BMI of 43 ± 7 kg*m^{-2} and age of 40 ± 8 years at the start of the study. A defined multidisciplinary weight-loss program was then carried out for 12 months followed up for another 12 months. During this period, the participants were examined at six time points, at the start, (T0) and at 3, 6, 12, 18 and 24 months (T3-T24). At T0, all participants underwent a thorough medical examination. A set of clinical tests were performed at all time points, explained in a later section.

Weight-loss intervention

The multidisciplinary weight-loss program (OPTIFAST® 52, Nestlé Inc.) has a positive effect on the weight-loss of the patients, explained at a great length in [73]. The 52 week program involves lifestyle modification based on psychology, medicine, dietetics and exercise. It also consists of consumption of a low-carbohydrate, inulin containing formula diet for 3 months. Inulin is the only source of energy during this period. The three months of diet intervention are followed by 8 weeks of reintroduction to normal food. This is again followed by a maintenance phase where the patients slowly begin the intake of normal food, while keeping their weight stable. Compliance to the program was assessed through participation in the meetings. Patients were also given dietary diaries to fill before the start of the program.

[1] The metagenomic analysis and its results in the publication was carried out using MALT, however with the availability of DIAMOND, the samples were re-analyzed and the results are described in this thesis. The main results with MALT and DIAMOND software tools do not differ.

Clinical parameters

At every time-point, several clinical parameters like the weight, height, blood and liver parameters were measured. Liver sonography was performed by a trained physician. Blood serum was analyzed for alanine aminotransferase (ALT), gamma-glutamyl-transferase (GGT), C-reactive protein (CRP), leukocytes, fasting glucose, insulin, HbA1c, total-, LDL- and HDL-cholesterol, and triglycerides in a certified medical laboratory (Laborärzte Sindelfingen, Germany). The Homeostasis Model Assessment-Insulin Resistance (HOMA-IR) index was calculated to estimate insulin sensitivity. The Fatty Liver Index (FLI) is a validated marker of risk for fatty liver disease and this was calculated. We used the definition of the International Diabetes Foundation (http://www.idf.org/metabolic-syndrome) for determining the metabolic syndrome state of study participants at different time points. IDFmetSynd and NCEmetSynd are the parameters measured for metabolic syndrome. The parameters were later correlated with the microbiome composition to inspect correlations between them.

Analysis of gut microbiota by shotgun sequencing

Collection of fecal samples was carried out at the six time-points. DNA extraction from the stool samples was performed using the "PSP-Spin-Stool-DNA-Plus Kit with lyses enhancer according to the manufacturers instruction (Stratec Molecular, Berlin, Germany). Whole-metagenome shotgun sequencing was done by the company CeGat, Tuebingen, Germany. Illumina HiSeq 2500 sequencer was used for producing 2x100 (paired-end) reads. On an average, the sequencing led to 2.1 GB per sample, with a sequencing depth of 10.9 million reads per paired-end sequencing (s=6.3 million). Out of the 96 samples expected to be obtained (16x6), 4 samples were missing.

4.2.3 Weight-loss, NAFLD and Metabolic Syndrome

Towards the end of the diet intervention, that is at the end of three months (T3), all patients showed a significant decrease in the weight (i.e. a high value of relative weight-loss). This relative weight-loss did not remain constant throughout the reintroduction and maintenance phases, with there being many fluctuations. At the end of the diet intervention, some patients were successful in losing more than 10% of their initial weight (a total of 9 patients) and the rest of the patients lost less than 10% of their initial weight (a total of 7 patients). This led us to divide the patients into two groups, the successful intervention group and the non-successful intervention group. The relative weight-loss in shown in Figure 4.1.

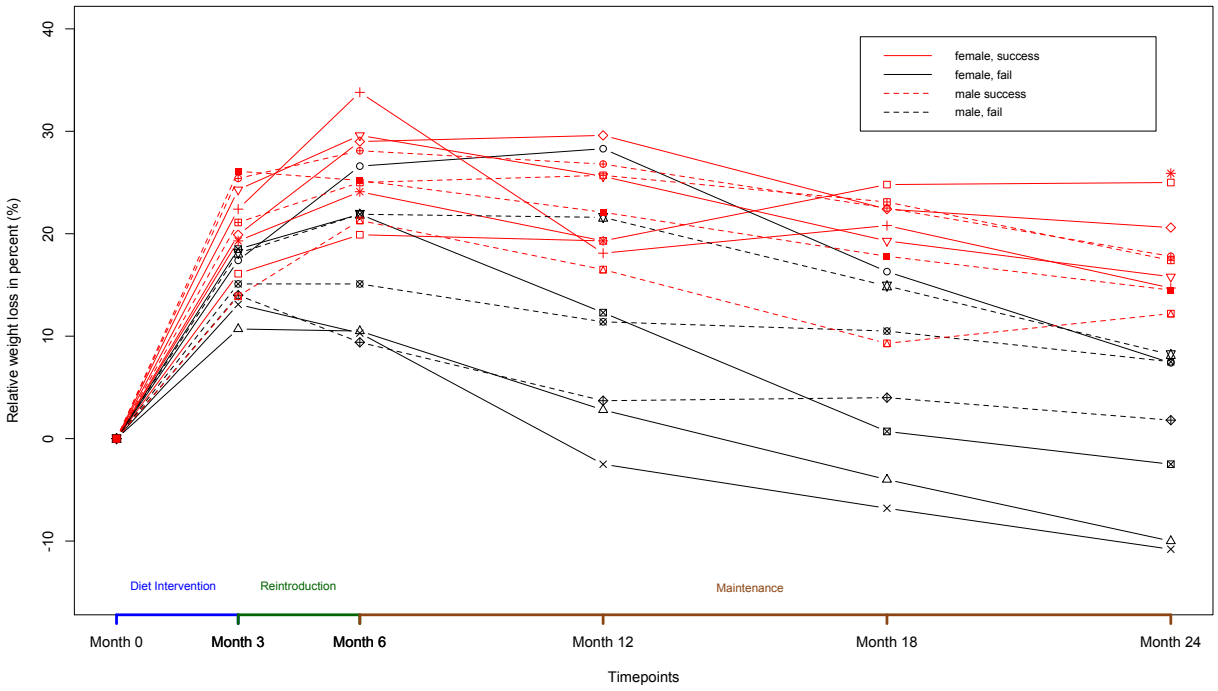


Figure 4.1: Relative weight-loss over the period of the study for the 16 patients. Black : positive intervention group. Gray : negative intervention group. The solid lines are used to depict female participants and the dotted lines to depict male participants.[2]

Both the grade of NAFLD and also the values of the Metabolic Syndrome parameter reduced for most patients after the diet intervention.

4.2.4 Objectives of the study

The main focus of the study was the characterization of the changes in the gut microbiome with respect to taxonomical and functional composition during the participation of the patients in the diet intervention. We wanted to know which bacterial groups, if any, fluctuate most in their abundance values during this time period. The second goal was to relate external measurements of weight, NAFLD and Metabolic Syndrome to the microbiome and assess whether significant differences are present in the gut microbiota composition between patients who have significantly different values for the assessed clinical parameters. Another aim was to determine whether the success or failure of the diet treatment is dictated by the initial composition of the gut microbiota, that is at T0.

4.3 Sequencing and bioinformatics analysis

4.3.1 Characterization of changes in the gut microbiome composition over the diet-intervention

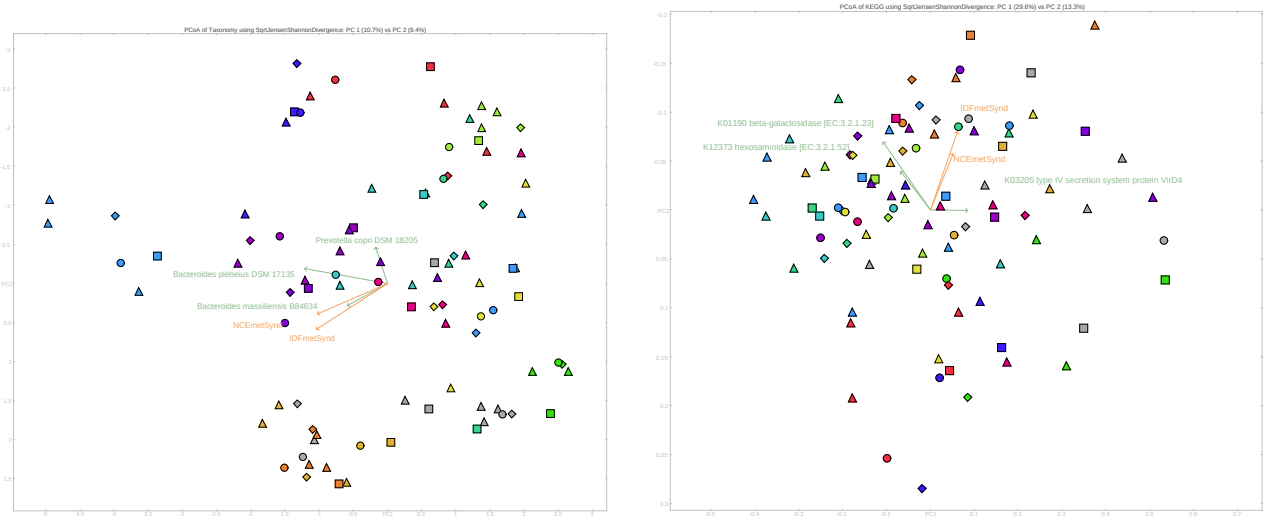
Sequencing of the DNA extracted from the fecal samples was carried out with the Illumina HiSeq 2500 technology. Paired-end reads with 100 base-pairs each were generated. A quality check was performed using the FastQC software. Processing of the raw sequence data was carried out with the PRINSEQ software. The reads were filtered for Ns and mean quality. The resulting reads were subjected to a DIAMOND analysis against the NCBI-NR (February-2015) version. The so obtained DAA files were meganized and imported in MEGAN. Metagenome data is available at the NCBI Server under Bioproject ID PRJNA290729. The DAA files are available on the MeganServer in the LouisEtAl2016 folder. A MEGAN comparison file was created by normalizing the read counts obtained. It also indicates the number of reads that got assigned to the Taxonomy, KEGG, COG and InterPro2GO classifications. Considering all 92 samples together, the most abundant phyla were Bacteroidetes (67.35%), Firmicutes (27.29%), Proteobacteria (2.23%), Actinobacteria (1.41%) and Verrucomicrobia (0.99%). The most abundant KEGG pathways were the

[2]We wanted to investigate whether the gender of the patient is an important factor in weight-loss, but the analysis showed no correlation between gender and dietintervention.

Biosynthesis of amino acids, Metabolism of cofactors and vitamins, Carbon metabolism, Purine metabolism and Pyrimidine metabolism. The predominant InterPro terms in all samples considered together were the IPR026892 Glycoside hydrolase family 3, IPR005094 Endonuclease relaxase, MobA/VirD2, IPR004764 Hydrophobe/amphiphile efflux-1 HAE1, IPR025705 Beta-hexosaminidase, IPR000322 Glycoside hydrolase family 31, IPR027256 P-type ATPase, subfamily IB, IPR006275 Carbamoyl-phosphate synthase, large subunit, IPR012754 DNA-directed RNA polymerase, subunit beta-prime, IPR004602 UvrABC system subunit A, IPR011895 Pyruvate-flavodoxin oxidoreductase, IPR001668 Plasmid recombination enzyme and IPR000743 Glycoside hydrolase, family 28. There can be seen a huge abundance of carbohydrate fermenting proteins.

The 92 samples part of the dataset were visualized as a PCoA (Figure 4.2) using the Jenson-Shannon index at the level of species and KEGG classifications in MEGAN. Each point in the PCoA corresponds to a sample and is colored and shaped according to the patient and the time-point that it belongs to. For the PCoA using the species counts, no strong separation is seen between the samples, although a general observation is that the samples from the same patient but at different time points tend to cluster. For the KEGG classification, no specific clusters are seen. The green and orange arrows depict the biplots and the triplots. At the species level, the Bacteroides species, namely, *Bacteroides massiliensis*, *Bacteroides plebeius DSM 17135* and *Prevotella copri DSM 18205* are the species responsible for the most difference between samples. For the KEGG classifications, K12373 hexosaminidase [EC:3.2.1.52], K01190 beta-galactosidase [EC:3.2.1.23], K03205 type IV secretion system protein VirD4 are responsible for most separation between samples. In both cases, the clinical parameters IDFmetSynd and NCEmetSynd are the ones responsible for most difference between samples. This is in line with a common observation that related gut microbiota samples tend to exhibit very similar composition at the level of function but in fact show many differences at the level of species and taxonomic composition. To gain a better understanding, we inspected the taxonomic composition at the level of genera. The Figure 4.3 depicts the abundance at the level of genera with the legend in Figure 4.4

In order to statistically test the differences between the composition along the time-points, we carried out Wilcoxon test at all taxonomic levels as well as functional levels. The Wilcoxon test was carried out between T0 - T3, T3 - T6, T6 - T12, T12 - T18 and T18 - T24. Additionally, we carried the tests at T0 - T6, T0 - T12, T0 - 18 and T0 - T24. This was done at all phylogenetic levels like Phylum, Class, Order, Family, Genus, Species



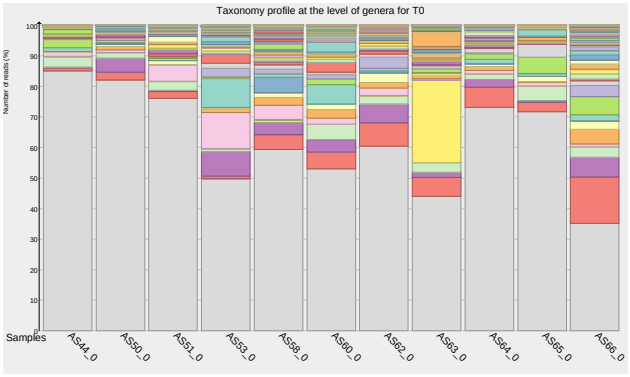
(a) Species (b) KEGG

Figure 4.2: PCoA of the 92 samples at the level of species shows clustering of the samples from the same patient at different time-points. At the level of functional classification, no specific clustering is seen.

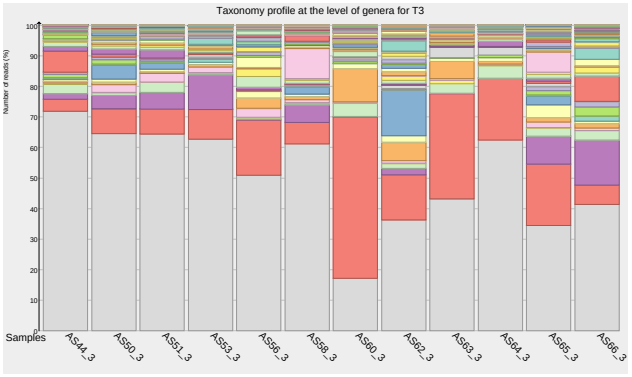
as well as the functional level, InterPro2GO, KEGG and COG. A p value of less than 0.05 was considered significant. The Wilcoxon tests revealed that the most changes occur before and after the actual diet intervention, that is from time point T0 and T3 and T3 and T6. This is also the period corresponding to most weight-loss. At higher phylogenetic levels the changes in relative abundance are less conspicuous, but at the level of genera and the species, there are many groups with significant changes in relative abundance. This also points to the idea that different individuals show a greater degree of dissimilarity when lower levels of taxonomy are considered than when higher taxonomic levels are considered. The strongest changes are between the relative abundances of many species of Alistipes and Roseburia. The Alistipes species showed a significant increase in the abundance from T0 to T3 and a significant decrease from T3 to T6. Roseburia showed a decrease from T0 to T3 but an increase from T3 to T6. Figure 4.5 and Figure 4.6 depict the species with the most abundant changes as per the Wilcoxon tests are shown as a tree in MEGAN.

The number of changes at the level of species decreases for the later time points. The changes from T6 to T12, T12 to T18 and T18 and T24 are shown in Figure 4.7 as a tree. From T12 to T18, a general decrease in Bacteroidetes species and an increase in the Firmicutes species is observed.

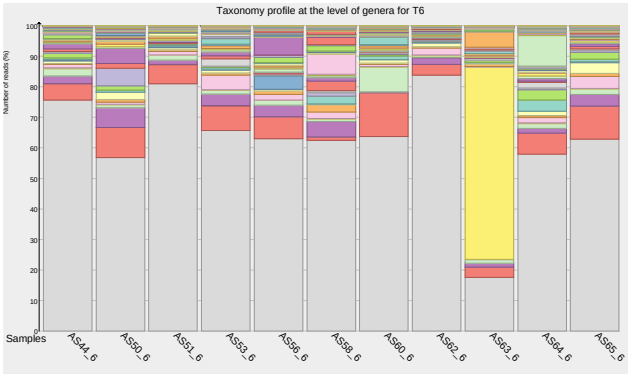
For the functional levels the most changes were observed from T0 to T3 and T3 to



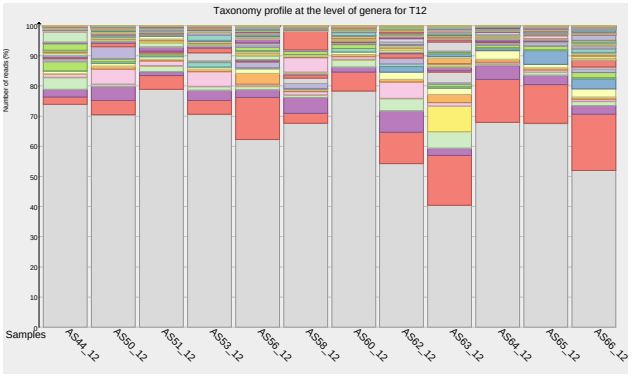
(c) T0 (Start of intervention)



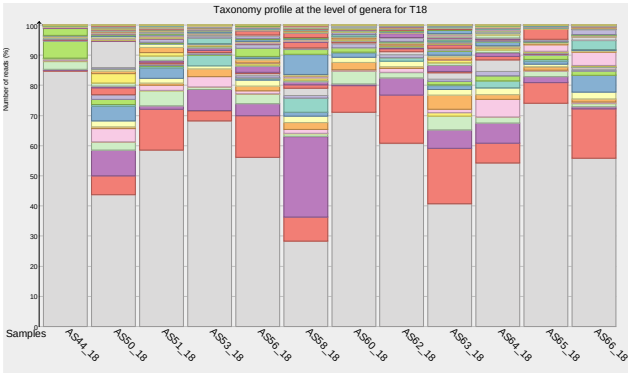
(d) T3 (End of intervention)



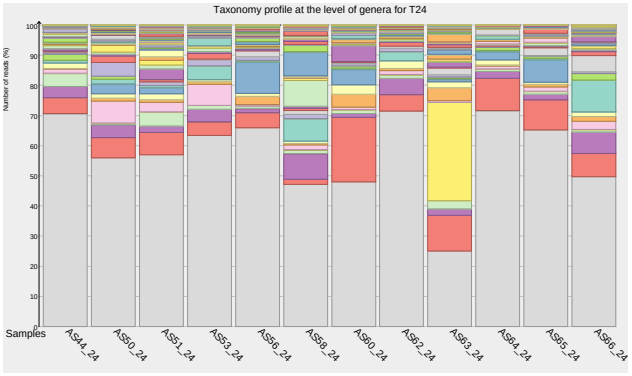
(c) T6



(d) T12



(c) T18



(d) T24

Figure 4.3: Stacked bar-plots of the read-count data at the level of genera shows the increase in the abundance of the Alistipes genera at the end of the diet-intervention.

Figure 4.4: Legend detailing the different genera.

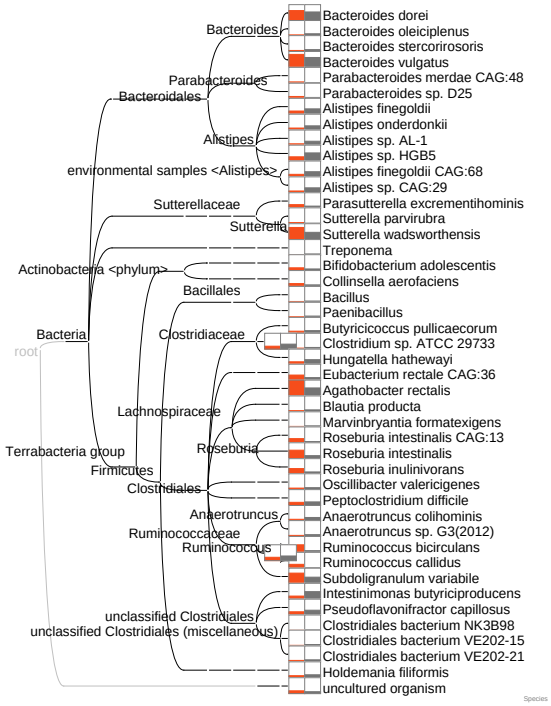


Figure 4.5: The species that significantly change in relative abundance over T0 and T3 are shown. Orange : The mean read-count of the species for T0 over all patients. Gray : The mean read-count of the species for T3 over all patients. As compared to T0, the relative abundance of several species of Alistipes increase in abundance. Species belonging to other genera, notably the Roseburia, decrease in relative abundance.

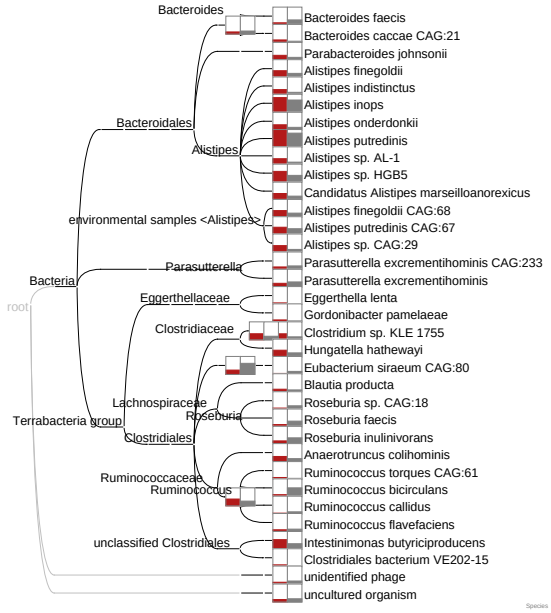
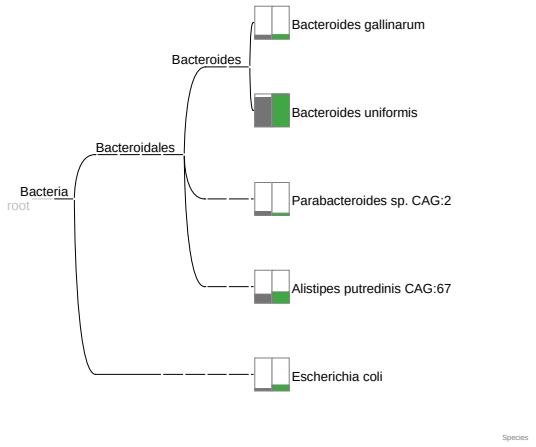
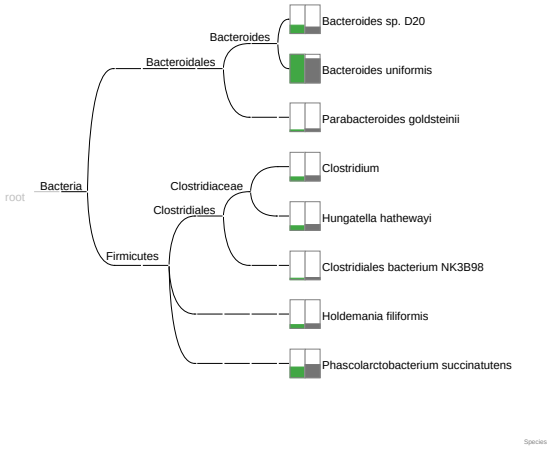


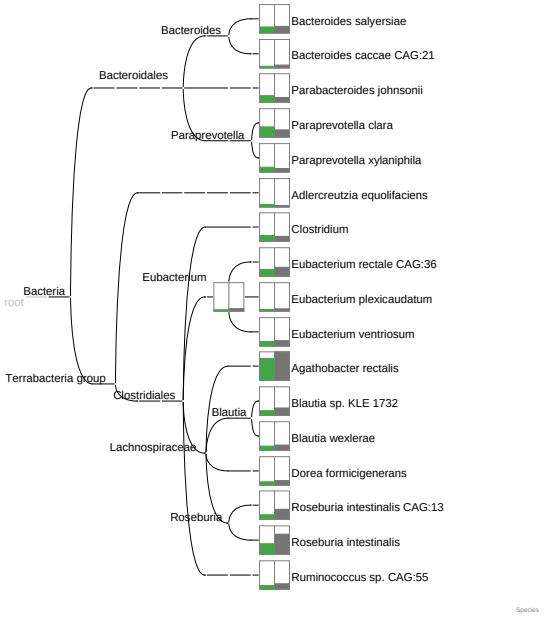
Figure 4.6: Red : The mean read-count of the species for T3 over all patients. Gray : The mean read-count of the species for T6 over all patients. From the end of diet-intervention (T3) to the start of the maintenance phase, (T6), a reversal of the previous changes occurs, with decrease in the relative abundance of Alistipes species and increase in the relative abundance of the Roseburia species.



(a) T6 and T12



(b) T12 and T18



(c) T18 and T24

Figure 4.7: The tree represents the genera that significantly change in relative abundance for the maintenance phase. Fewer significant changes are observed, pointing to the resilience of the gut microbiota, where Green : previous time-point and Gray : later time-point.

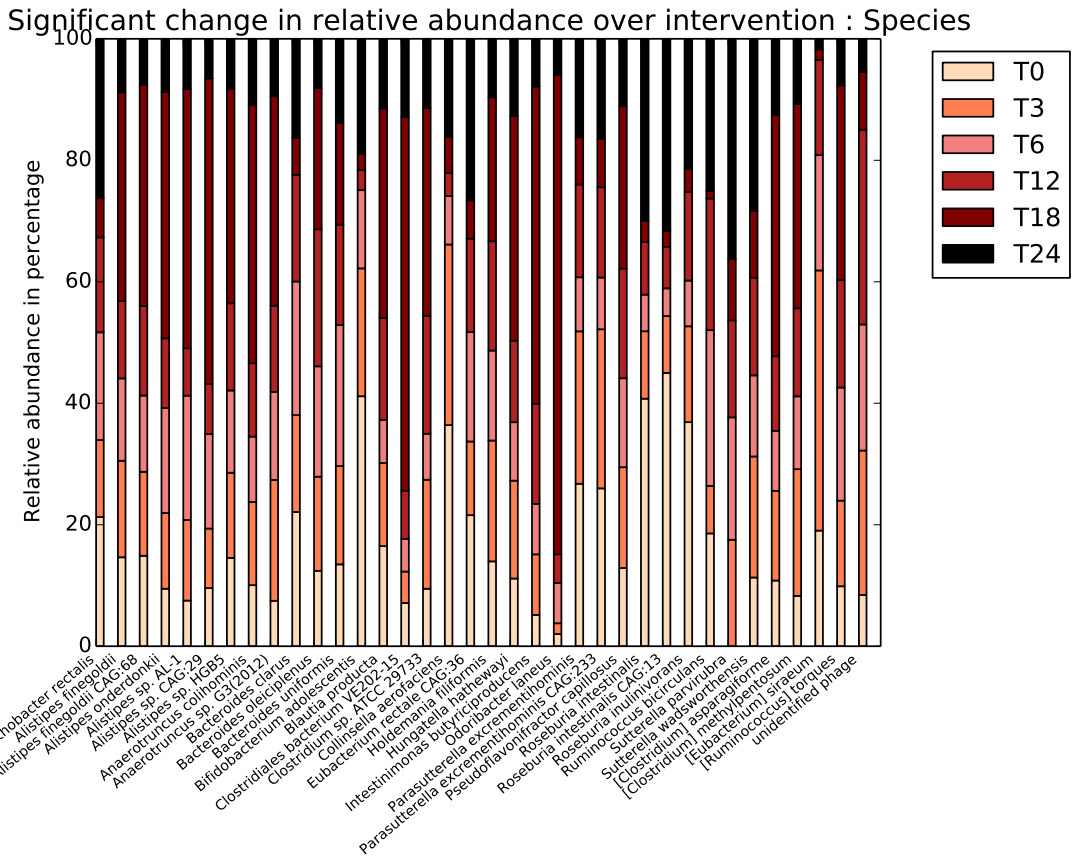


Figure 4.8: The species most fluctuating in relative abundance as determined by the Friedman's test are shown as a stacked bar-plot. Each value in the bar is the mean read-count for the particular time-point for each of the species, over all patients.

T6. The most abundant and significantly changing KO groups belonged to carbohydrate metabolism enzymes like beta-galactosidase, alpha-glucosidase and others. These enzymes usually decreased from T0 to T3. In the changes from T3 to T6, the KO groups belonged to a range of pathways and not only the carbohydrate pathways. A Friedman's test was carried out to check the overall changes in the gut microbial composition. The species most fluctuating in relative abundance and the KEGG groups most fluctuating in relative abundance are shown in Figure 4.8 and Figure 4.9. The mean abundance of all patients of a given species (or KO group) at a given time-point is plot.

Table 4.1 describes the KO groups that are abundant in the samples and significantly change in relative abundance. As can be seen, for the KO groups, the values over the time-points are constant, unlike the case with the species composition. It is a common observation that the abundance of functional categories are similar across related samples, and are not

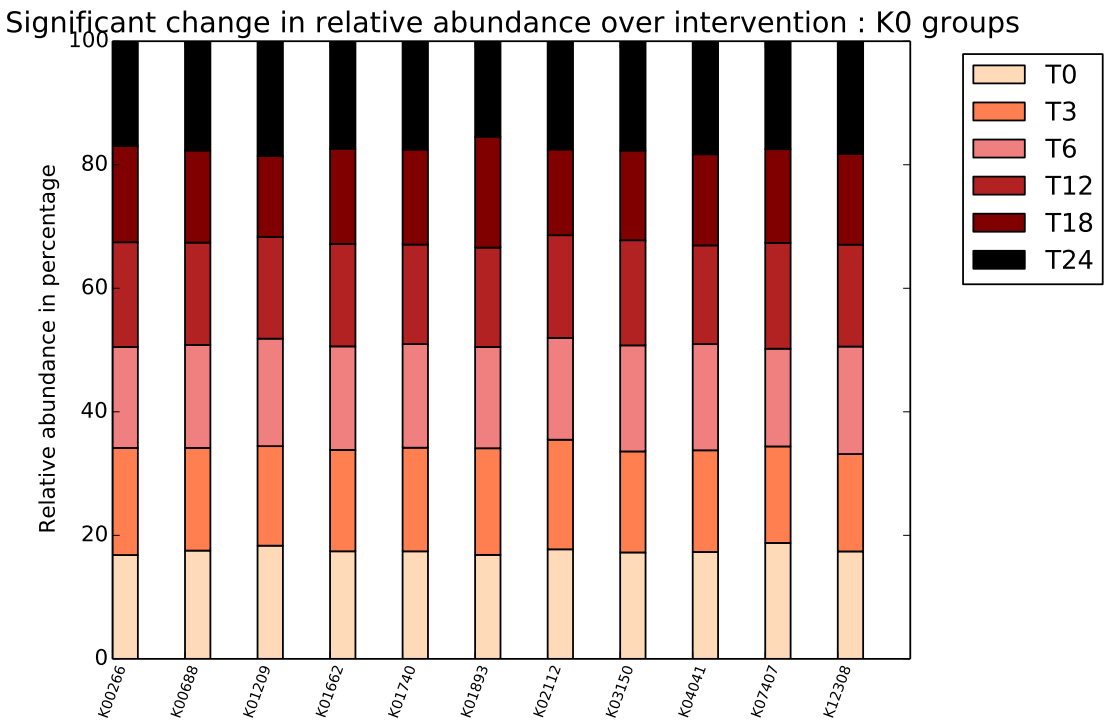


Figure 4.9: The Friedman's test for the KEGG functional classifications

Table 4.1: Description of the KO groups determined to be significantly changing in their relative abundance over the intervention.

KO	Orthologous group	Pathway
K00266	glutamate synthase (NADPH/NADH) small chain	Biosynthesis of amino acids
K00688	starch phosphorylase	Carbohydrate metabolism
K01209	alpha-N-arabinofuranosidase	Carbohydrate metabolism
K01662	1-deoxy-D-xylulose-5-phosphate synthase	Metabolism of cofactors and vitamins
K01740	O-acetylhomoserine (thiol)-lyase	Amino acid metabolism
K01893	asparaginyl-tRNA synthetase	Translation
K02112	F-type H+-transporting ATPase subunit beta	Energy metabolism
K03150	2-iminoacetate synthase	Metabolism of cofactors and vitamins
K04041	fructose-1,6-bisphosphatase III	Glycolysis / Gluconeogenesis
K07407	alpha-galactosidase	Galactose metabolism
K12308	beta-galactosidase	Galactose metabolism

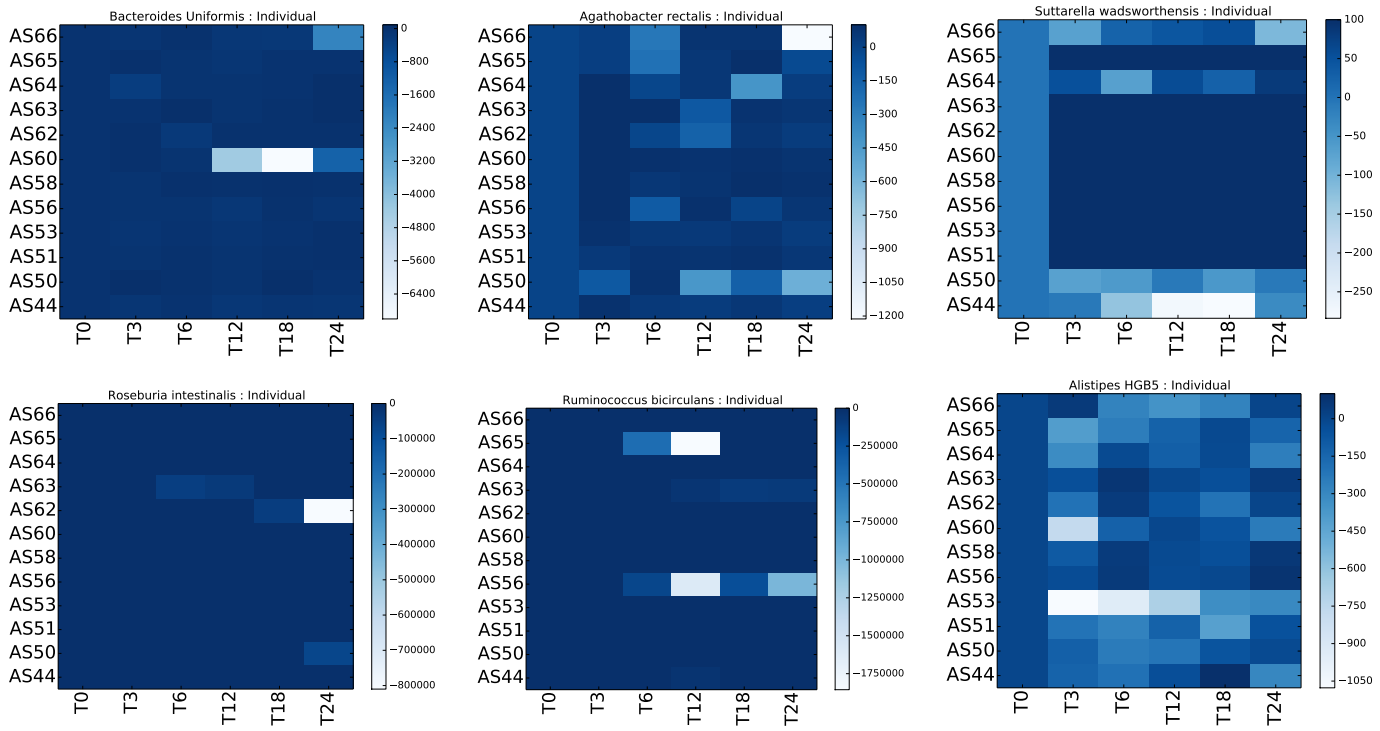


Figure 4.10: The colors are reflective of whether the read-count abundance increased or decreased with respect to T0 along the rest of the time-points. For every individual, this trend is different.

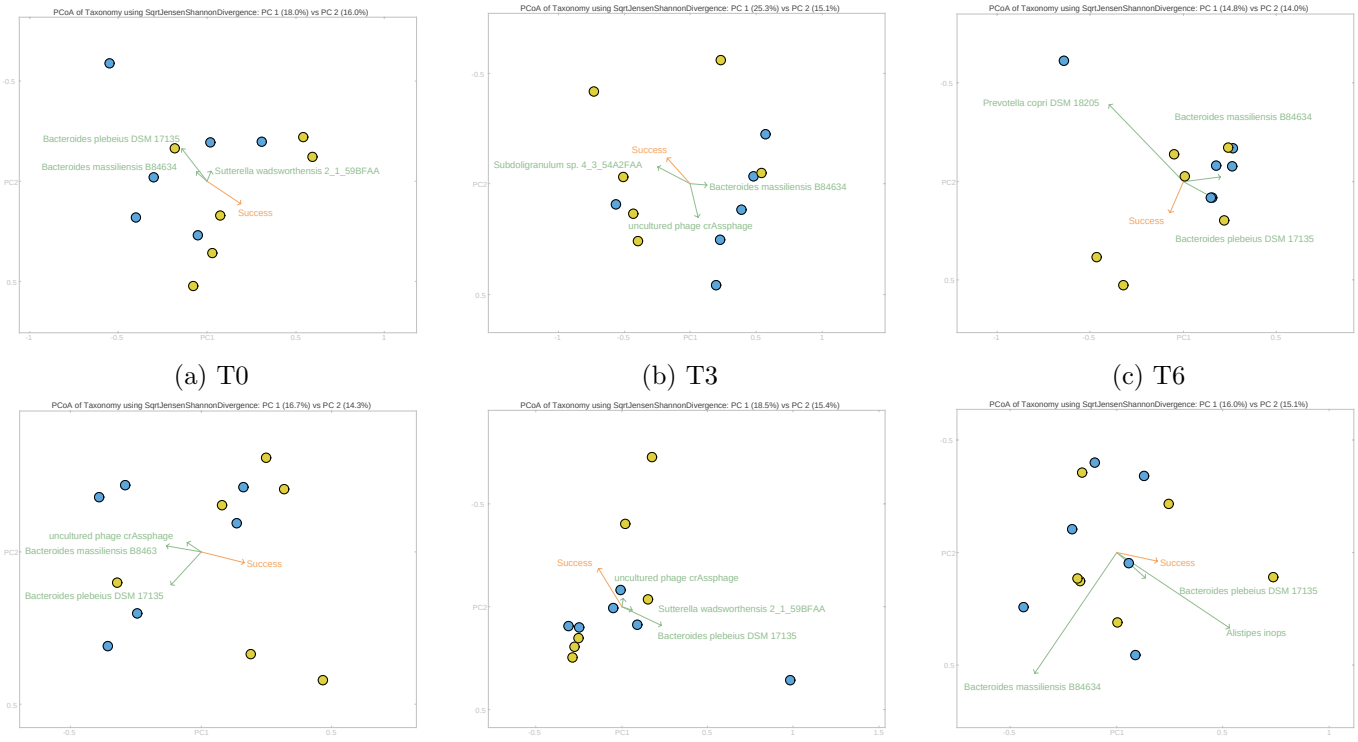
variable as is the case with the species.

The statistical tests helped in determining the taxonomic groups that are most fluctuating, however, every patient reacts differently and hence it is important to consider the individual trends. For the species determined to be significantly changing, the read-count abundance relative to T0 was calculated and plot as heatmaps (Figure 4.10).

We therefore conclude that inspecting the general trends in a gut microbiome dataset is important, but it is also of interest to study the dynamics of the gut microbiome of a single patient over the time-points. This would lead to more understanding of the individualistic nature of the gut microbiome and help in personalized treatment of obesity.

4.3.2 Differences between the gut microbiota of patients with successful (PI) and non-successful (NI) diet-intervention

Depending on the relative weight-loss at the end of the diet-intervention, the patients were grouped into two categories – the successful intervention group consisting of patients



(a) T0 (b) T3 (c) T6

(d) T12 (e) T18 (f) T24

Figure 4.11: PCoA plots at the species level showing the clustering between the different samples, blue : non-successful diet intervention, yellow : successful diet intervention.

that lost at least 10 % or more of their weight with respect to the weight at the start of the intervention (T0) and non-successful intervention group, consisting of patients that lost less than 10 % of their weight. A Student's t-test was performed to find whether significant differences exist between the two groups. The main difference between the groups was in the Firmicutes to Bacteroidetes ratio at the start and end of the intervention, Figure 4.11

At T3, the patients with more than 10% weight-loss at T24 have a significantly high amount of Firmicutes to Bacteroidetes as compared to the patients with less than 10% weight-loss at T24. Thus a greater relative abundance of the members of the Firmicutes species are indicative of success in the intervention (Figure 4.12).

This also points to the idea that the success of the diet intervention may depend on the initial composition of the gut microbiome and also on the individual patients. Obesity needs a personalized treatment and studies such as these could provide with essential clues regarding the design of the treatment.

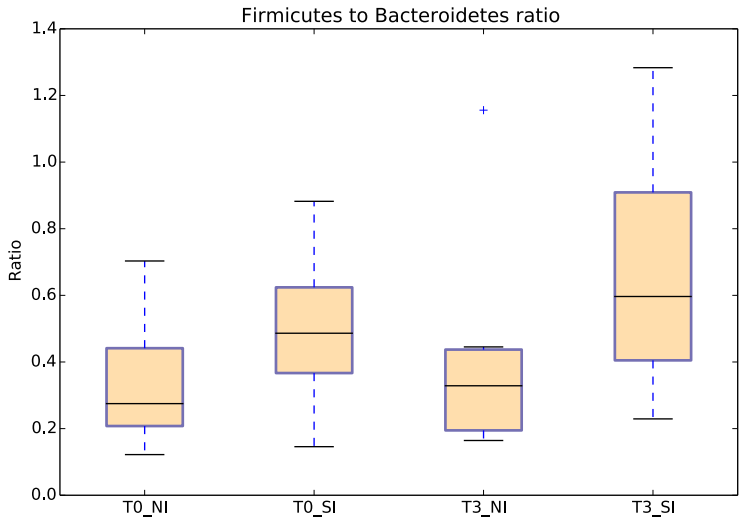


Figure 4.12: Boxplot of the Firmicutes and Bacteroidetes ratio. NI - Non-successful Intervention, SI - Successful Intervention

4.3.3 Relating the clinical parameters for NAFLD and Metabolic Syndrome

Fatty Liver Index (FLI) and HOMA-IR were used for the diagnosis of NAFLD. The diet intervention had a positive effect on the grade of NAFLD. At T3, most patients had a reduced grade or negative result for presence of NAFLD. At T3, a total of 8 patients were not positive for NAFLD as measured by the two parameters. At T24, 8 patients were not positive for NAFLD. At time points T3 and T24, the differences in the relative abundances of species between the patients with or without NAFLD were tested.

Tests performed for finding significant differences revealed that those patients with NAFLD had much lower counts of the bacterium *Subdoligranulum variabile* as compared to the patients that did not have NAFLD. This agrees with the observation by Bajaj et al., [74] who found out that cirrhotic patients had reduced relative abundance of this bacterium.

For Metabolic Syndrome, the bacterium *Faecalibacterium prausnitzii* and *Eubacterium ventriosum* was found to be significantly higher in patients without any metabolic syndrome.

4.3.4 Time-series clustering of the patients.

The above study was carried out over a period of 2 years and produced time-series data. The time-series is not a long one, since only few time-points are recorded. This is often a

problem with huge metagenomic projects, especially whole-genome shotgun projects where obtaining data for several time points, although possible, is limited due to the finances involved. Many methods specifically focus on the statistical analysis of time-series metagenomic data as reviewed in [75]. However, many of the methods are statistical-model based and it is not possible to create statistical models with few data-points. Thus, there is always a need of novel statistical techniques to analyze time-series metagenomic data. Many studies on time-series metatranscriptomic data are available. In the case of the Hohenheim Project, one of the questions we aimed at answering was whether any difference existed between the microbiomes of successful and non-successful patients. For this it was also important to find the "overall" microbiome composition along the time-points for each patient. In order to determine how the patients cluster overall during the entire time-period, we binned the patients in clusters along each time-point, and then determined which samples have the propensity to be in a given group.

Specifically, read-count data at the level of genera was extracted and normalized for the 16 patients at each time-point. The Vegan veg-dist package was used for creating a distance matrix according to Bray-Curtis metric. The distance matrix was then subjected to a Partitioning Around Medoids (PAM) clustering. The output of the PAM clustering for each time-point was used to determine, for each pair of patients, the number of times they are together in a cluster. Based on this, a distance matrix was created, and to visualize it, a dendrogram was created (Figure 4.13).

This offers some additional insight into how the patients/samples themselves are related, however, these clusters do not explain a particular phenotype, for example weight-loss or NAFLD.

4.4 Conclusions and discussion

Whole genome shotgun sequencing of the fecal samples to study the gut microbiota is now commonplace and has given us the opportunity to study the gut flora at the taxonomic and functional level. For the Hohenheim Project, the metagenomic sequencing coupled with the correlation of clinical parameters contributes towards understanding how the gut microbiota are affected after a weight-loss formula diet treatment. Longitudinal analysis proved to be beneficial in tracking the gut microbial changes long after intervention. However, a deeper depth of sequencing and more time-points could have contributed to the robustness of the analysis.

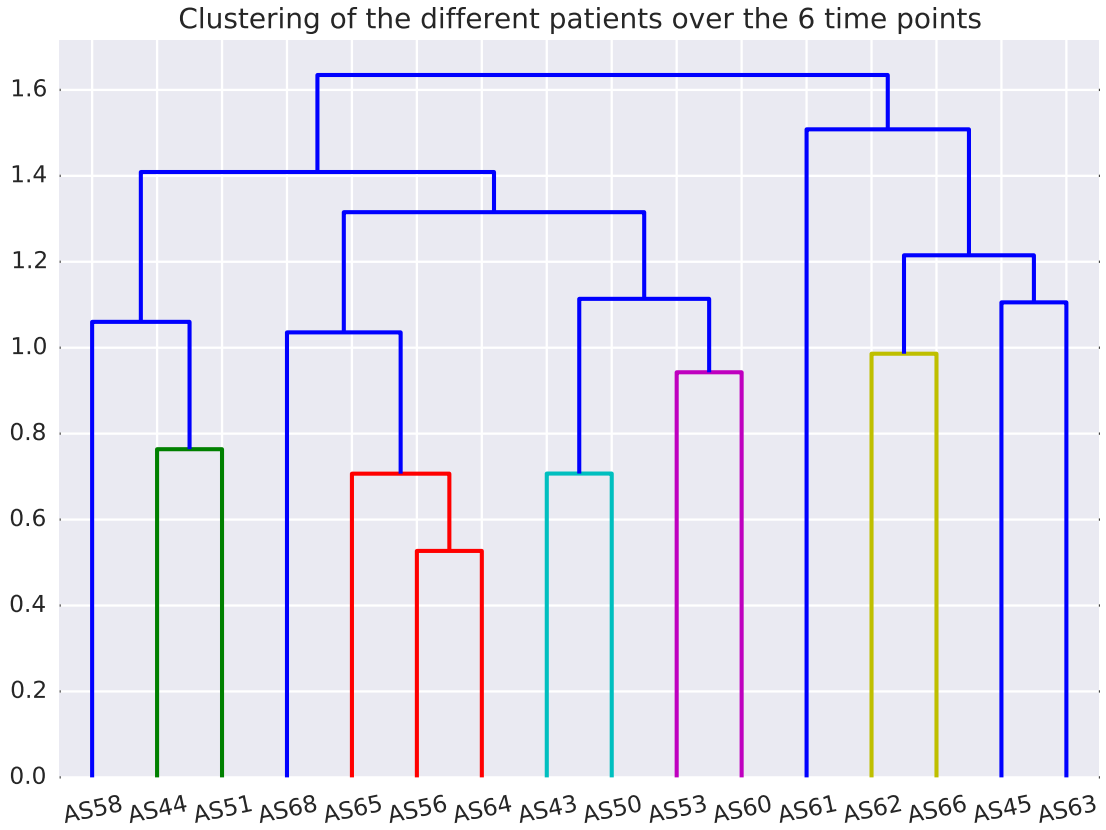


Figure 4.13: Clustering of the patients according to similar taxonomic profiles over the diet-intervention.

Formula diet treatment results in significant loss in weight accompanied by change in the microbial composition of the gut microbiota, but this change is transient and does not persist in the maintenance phase (for a longer period of time). The greatest impact was seen at month 3, at the end of the diet intervention. Samples from the same patient, but at different time points are more similar than samples from other patients. The diet intervention was characterized by significant changes in many groups of bacteria, both from the Bacteroidetes and Firmicutes phyla. But the Firmicutes to Bacteroidetes ratio was significantly high in the patients with a successful diet intervention treatment. This indicates that there is a possibility that the success of diet intervention could be detected at an early stage.

Several different groups of species like the Alistipes, Akkermansia, and the butyrate producing Roseburia are all important species in the diet intervention. Differences in the taxonomic composition are easily observable, but it is difficult to do so at the level of function. However, studies such as these are bring us closer in understanding the interplay between diet, obesity and the gut flora and designing personalized care for obese patients.

Chapter 5

To assemble or not to assemble

Metagenomic assembly is a complex computational problem due to the size of the datasets and the presence of reads from multiple different organisms in different coverages in the datasets. The specific goals of the metagenomic project dictate whether assembly is required or could be skipped. This chapter compares the taxonomic and functional profiles obtained by assembly and analysis of the contigs, with the profiles obtained by raw read analysis.

5.1 Challenges involved in the *de novo* assembly of metagenomic datasets

The size and complexity of WGS metagenomic datasets render their assembly a complex task. However, the ease of assembly of such datasets depends on the taxonomic diversity of the microbial sample being assembled and also the depth and coverage of the sequencing [76]. In the case of a microbial sample with a low taxonomic diversity, for example a mock metagenomic community consisting of a few organisms, the sequencing can be designed to have a sufficient depth and coverage. In this case, the sample will consist of sufficient number of reads originating from all the organisms present and with a satisfactory uniform coverage across all genomes. Here assembly is a relatively easy task. In the case where there is huge taxonomic diversity, for example a soil microbial sample, the different organisms present in a community are not uniformly abundant. The different organisms have varying genome sizes. As a result, the depth of sequencing may not be uniform across all the genomes present in the sample. The sequenced reads often may not represent the complete complement of the organisms present in the sample, making assembly harder. In any case, the repeats in the genomes pose a challenge. *De novo* assembly is generally a time consuming and computationally intensive task requiring much hands-on work.

5.2 Tools used for the assembly of metagenomic datasets

In spite of the challenges, many tools carrying out *de novo* assembly of metagenomic samples exist. Ray is an assembler that makes use of a de Bruijn graph for determining seeds from the coverages, and these seeds are extended based on overlaps. The process stops when the seed cannot be extended and the contigs are returned. Ray-Meta offers options for using multiple cores and this distributed computing helps to make the software scalable. In

addition to assembly, it also offers taxonomic profiling. The IDBA-UD assembler also uses de Bruijn graphs, but iteratively produces scaffolds by first starting from a minimum value of k, and increasing this value in the next iteration. Also, the threshold for removing low depth contigs is increased. Meta-Velvet [77], MetAMOS [78] and other assemblers also use de Bruijn graphs. Some tools can combine short-read data with long read data, for example the software Cerculean [79].

5.3 *De novo* assembly as compared to a raw read based analysis

Both read-based and assembly-based approaches have been successful in profiling a microbial community using metagenomic data. The problem with short reads is that they may align with equivalent scores to multiple reference sequences and hence for short reads from highly conserved regions of the genome, their organism of origin cannot be easily determined. Various algorithms and software tools are being continuously developed for short read assignment. Depending on the goal of the project, the analysis pipeline can be designed for either the sequence analysis of the reads leading to their binning into taxonomic and functional categories or their assembly into contigs. The contigs can then be analyzed with the appropriate tools. These two approaches can be used separately or in combination. As already discussed, assembly is a complex problem and even though both read-based and assembly-based analyses are known to be equally good approaches, an assembly still remains important. In the following chapter we describe the results obtained from comparison of the taxonomic and functional profiles using short reads with the profiles using long contigs. The raw reads and the long contigs warrant a different approach for their analysis. However a comparison can be made between the taxonomic and functional profile that is generated as a result of the analysis. Short reads are binned based on their alignments to a database and the contigs are subjected to ORF calling and annotation of the obtained ORFs. A pictorial representation of the possible analysis strategies of the read-based and contig-based approaches is shown in Figure 5.1

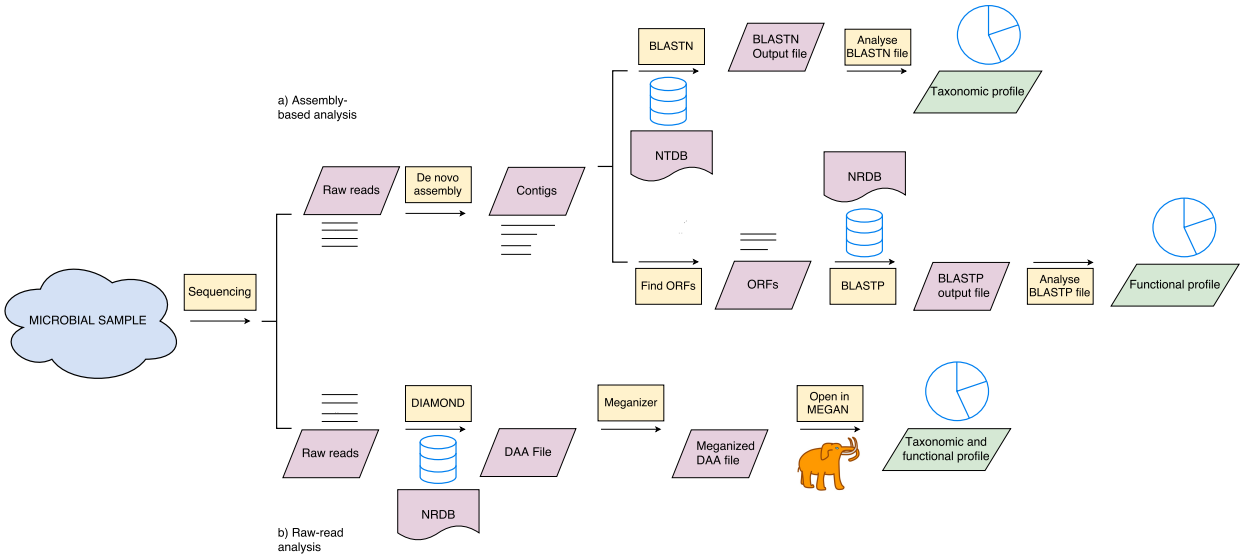


Figure 5.1: a. A generic approach for assembly-based analysis. b) A generic approach for raw-read based analysis.

We considered three metagenomic samples for the study. The first sample is a mock metagenomic sample from the study by Shakya et al. [56]. This mock community consists of 48 bacterial and 16 archael organisms. Illumina sequencing data generated from this mock community with 101 million reads and a read length of 101 bases was used for the study. The second sample was taken from the Hohenheim Obesity Project, sample AS53_6 (SRS1028280). This sample was chosen because of its average size as compared to other samples in the dataset. The third metagenomic sample is a simulated metagenomic sample created from a total of 10 bacterial species commonly found in the human gut. The 10 bacterial species and their abundances are summarized in the Table 5.1. The simulation tool ART [80] was used for creating 101 base-pair reads with a 20-fold coverage. The raw reads were subjected to the DIAMOND and MEGAN pipeline and the read count data was extracted, both at the level of taxonomy and functional classification. For both the datasets, the raw-read pipeline and the contig pipeline as described in the Figure 5.1 was used. The Table 5.2 summarizes the number of reads present in the samples, time required for the analysis, number of reads assigned to taxonomic and functional categories and the number of contigs generated after assembly.

Table 5.1: Organisms used for creating a simulated metagenome and their abundance (number of genomes used).

Organism	**NCBI Accession**	**Genome count**
Bacteroides xylanisolvens XB1A	FP929033	1
Eubacterium siraeum 70/3	FP929044	1
Odoribacter splanchnicus DSM	NC_015160	1
Bifidobacterium longum subsp	NC_004307	1
Akkermansia muciniphila ATCC	NC_010655	2
Parabacteroides distasonis ATCC	NC_009615	3
Faecalibacterium prausnitzii L2-6	NZ_GG697149	4
Alistipes finegoldii DSM	NC_018011	5
Butyrate-producing bacterium SS3/4	FP929062	5
Bifidobacterium adolescentis ATCC	NC_008618	9

Table 5.2: Number of reads per sample, number of reads assigned and number of contigs.

Sample	Number of reads	MEGAN + DIAMOND Time (s)	Assigned Taxa	Assigned Function (KEGG)	Assembly Time	Number of contigs
Mock metagenome	101,000,000	1,07,460	76,119,959 (∼ 75 %)	23,377,731 (∼23 %)	7 days	71,550
AS53_6	17,892,974	22,903	10,984,455 (∼ 61%)	2,326,503 (∼13%)	13 hours, 25 minutes	318,478
Simulated metagenome	20,480,388	11,163	16,436,900 (∼80%)	1,898,704 (∼ 9%)	4 hours	11,701

The mock sample is huge in size with 101 million reads but produced 71,550 contigs which were few as compared to the human gut sample (AS53_6) which has 17,892,974 reads but produced 3,18,478 contigs. The simulated metagenome produced the least number of contigs (11,701). The human gut sample is the most diverse and hence a huge number of contigs was expected. There are fewer contigs in samples with a less diverse taxonomic composition.

5.4 Comparison of the taxonomic profiles

For the mock metagenome, out of the 76,119,959 reads assigned to a taxonomic group, 20,831,992 are assigned to the species level, 7,296,701 reads are assigned at the strain level, 11,170,251 at the species level, and 2,365,040 are false positives. The taxonomic profile of the raw reads has 24 different phyla, 38 classes, 51 orders, 62 families, 77 genera and 132

species. Out of 132 species, 23 were correct at level of strain, 50 at the level of species and 59 were false positives. A high number of false positives is expected in a read-based analysis. The false positive organisms share a phylogenetic closeness with the species part of the mock community, since most of them belong to the same genera or family. The profile does not contain any false negatives, in that all the 64 organisms part of the profile have at least a few reads aligning to them.

The contigs produced for the mock metagenome were subjected to BLASTN against the NCBI-NT database. Analysis of the BLASTN output file of the contigs revealed that out of 71,550 contigs, a total of 66,652 contigs found an alignment (4,898 contigs did not align to any reference). Out of 66,652 contigs, 64,431 belonged to true positive references and 2,221 in total belonged to false positive references. All 64 organisms are part of the profile and no false negatives exist.

For the human gut microbiome sample AS53_6, the taxonomic profile for the raw reads has 216 different bacterial species. The taxonomic profile for the contigs contained 1,008 different species. Out of the 3,18,478 contigs, 54,523 obtained an alignment. It was unexpected that the profile for the raw reads contained less contigs than the profile for the contigs.

For the simulated metagenome, the taxonomic profile for the reads contained a vast number of false positives. The most abundant organisms were the ones that belonged to the 10 species part of the simulated metagenome. However, the read counts were not representative of the abundance of the different species in the metagenome. For example, the most reads should belong to the species *Bifidobacterium adolescentis ATCC* since it is the most abundant, however, this was not the case. For the contigs, the taxonomic profile consisted of all the 10 species.

For a more direct comparison, the reads from each sample were mapped to the respective contigs. The read to taxonomic group assignment was compared with the contig to taxonomic group assignment. We expect that, if a read is placed in a particular taxonomic group, the contig that the read is part of, will probably be placed in the same taxonomic group. We then calculate the percentage of reads, that are part of a contig, and map to the same taxonomic group, to the resolution of strain or species. The Figure 5.2 is a bar plot depicting the percentage of reads inspected that have the same taxonomic annotation as the contig it was mapped into.

For the simulated metagenome and the mock metagenome, about 50% of the reads that map to a contig are assigned to the same strain as the contig. Thus, for a simple and less diverse sample, short-read assignment is easier than for a more diverse sample such as the

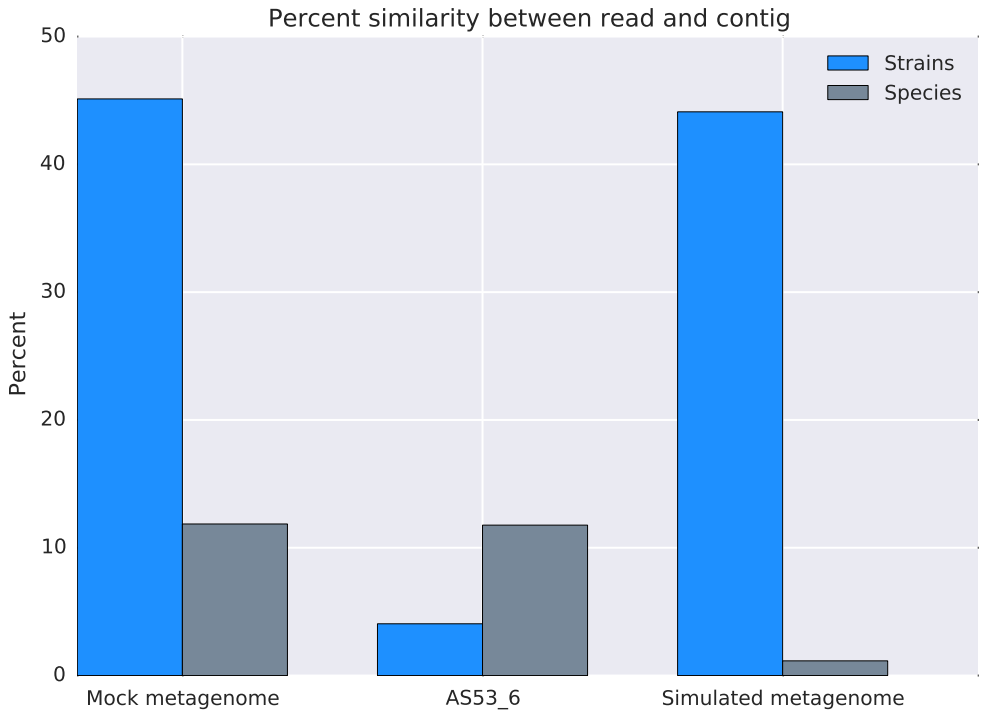


Figure 5.2: Barplot showing the percentage of reads that are assigned to the same strain, or the same species as the contig that it maps to.

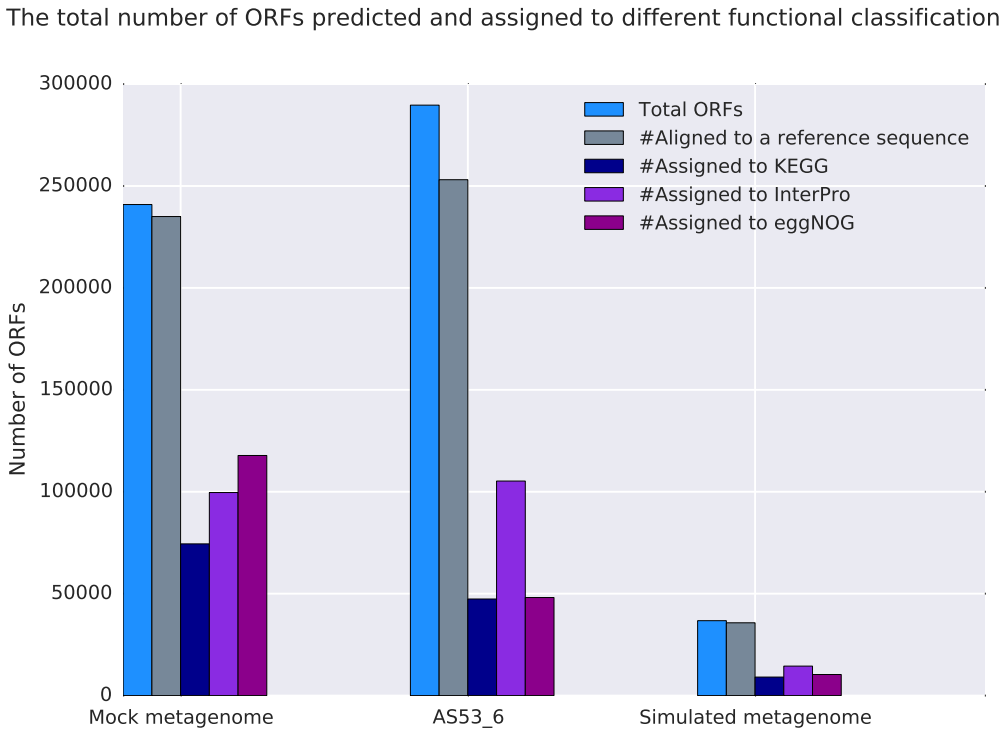


Figure 5.3: The number of ORFs predicted by MetaGeneMark for each sample and the number of ORFs that are placed in the three functional classification systems. Very few ORFs out of the total are placed into a functional classification system. For the gut microbial sample, this difference is the highest.

gut microbial sample.

5.5 Comparison of the functional profiles

The read based functional profile were extracted from the MEGAN files. For the functional profiling of the contigs, open reading frames were called on the contigs using the tool MetaGeneMark [81]. The ORFs were output as amino-acid sequences and DIAMOND BLASTP analysis was carried out against the NR database for the assignment of functions to contigs. The meganized DIAMOND output file was imported in MEGAN and the functional profiles were examined. The assignment rate of the contigs was very low for the ORFs. The number of ORFs for the mock, AS53_6 and the simulated metagenome are shown in Figure 5.3

For the functional profiles across the metagenomes, there was a strong correlation between the number of reads and the number of contigs assigned to the orthologous groups. However,

Table 5.3: InterPro profile of AS53_6 for the 10 most abundant orthologous groups. The 10 most abundant InterPro groups as derived from the analysis of the reads is the same as the 10 most abundant InterPro groups as derived from the analysis of the contigs.

InterPro ID	Contig Abundance	Read Abundance
IPR002528 Multi antimicrobial extrusion protein	1,636	56,004
IPR020449 Transcription regulator HTH, AraC- type	1,008	39,773
IPR026892 Glycoside hydrolase family 3	890	43,638
IPR027256 P-type ATPase, subfamily IB	606	18,854
IPR025705 Beta-hexosaminidase	422	31,713
IPR010327 FldB/FldC dehydratase alpha/beta subunit	416	11,271
IPR004805 DNA polymerase III, alpha subunit	360	13,931
IPR000322 Glycoside hydrolase family 31	337	23,535
IPR006275 Carbamoyl-phosphate synthase, large subunit	333	19,148
IPR003373 Ferrous iron transport protein B	331	10,534
IPR015937 Aconitase/isopropylmalate dehydratase	329	14,316
IPR003688 Type IV secretion system protein TraG/VirD4	319	29,669
IPR005094 Endonuclease relaxase, MobA/VirD2	310	41,860
IPR005936 Peptidase, FtsH	296	10,407
IPR018044 Peptidase S11, D-alanyl-D-alanine carboxypeptidase A	293	5,566
IPR004576 Transcription-repair coupling factor	292	12,643
IPR001463 Sodium:alanine symporter	287	8,816

the abundance as defined simply by the number of reads or the number of ORFs present in the specific bin, is not always concordant. For example, for the analysis of contigs (ORFs), the four most abundant KO groups K03088 (RNA polymerase sigma-70 factor), K06147 (ATP-binding cassette), K02003 (putative ABC transport system ATP-binding protein) and K02004 (putative ABC transport system permease protein) have no contigs assigned to them, whereas these KO groups fetch a significant amount of reads. For the InterPro2GO analysis, the abundance profile for both the reads and the contigs was similar. There was a strong correlation between the number of reads and the number of ORFs that bin into that group. For the eggNOG assignments as well, there was a strong correlation.

Thus, the functional composition by the read analysis or by the contig analysis both result in fairly similar results with small discrepancies. In this analysis, the KEGG classification system showed the most deviations. It must be noted that change in mapping files due to newer releases of the databases might result in differences in the functional assignments. As an example of the correspondence between the profile obtained for reads and contigs, Table 5.3 lists the most abundant InterPro identifiers.

5.6 Conclusions

An attempt was made at comparing the results obtained for a read-based and an assembly-based analysis of metagenomic samples. The assembly of the samples took significantly more time as compared to a read-based analysis. But the time required depends on the size and

complexity of the sample (when using a particular assembler). When analyzing multiple metagenomes, this could be a bottleneck. In two out of the three cases, the raw read based taxonomic profile yielded a lot of false positives as compared to the contig taxonomic profile of the metagenome. Since short reads align with less specificity, this was expected. Taxonomic or functional assignment depends on the alignments and since short reads align with less specificity, their analysis might yield slightly different results as compared to the contigs. As revealed by the above study, some differences exist between read profiles and contig profiles, though they may not affect the conclusions about taxonomic diversity that are made with them. The two analysis strategies could be useful in different situations. Alternatively, they could be used in consolidation, usually by mapping the reads back to the contigs for improved assignment.

In conclusion, metagenomic assembly could be carried out provided sufficient computational power and time is at disposal. Moreover, a read-based analysis and an assembly together can also be useful. Mapping the reads back to the assembled contigs can better determine the taxonomic origin of the read.

Chapter 6

Gene-centric assembly of orthologous gene families in microbiome sequencing data using MEGAN 6

A gene-centric assembly involves the assembly of all reads in a microbiome sample that have been assigned to a specific gene family, after the sequence analysis of the reads. As opposed to a *de novo* assembly of an entire microbiome sequencing sample, a gene-centric assembly could be computationally less intensive. A new algorithm based on the overlap-layout-consensus paradigm has been developed in MEGAN 6 which can be used to assemble reads belonging to one or more orthologous gene families. This chapter demonstrates how a gene-centric assembly is performed with the DIAMOND and MEGAN 6 pipeline and evaluates the performance of MEGAN as compared to other well-known assemblers like SOAPdenovo, Ray and IDBA-UD. Examples of how a gene-centric assembly could be included as part of a metagenome analysis pipeline is provided by running MEGAN's assembler on different kinds of gene families and metagenomic samples.

6.1 Introduction to gene-centric assembly

Next-generation sequencing technologies that are used for the sequencing of microbiome samples produce short reads. They can either be subjected to a *de novo* assembly or can be analyzed using homology searching against a reference database, or both [82]. Homology searches for short reads, for example using the DIAMOND and MEGAN pipeline, ultimately lead to the binning of the reads into different taxonomic and functional categories. The targeted assembly of all reads belonging to a specific orthologous group of genes (that is belonging to a particular bin), is referred to as a gene-centric assembly. A new algorithm based on the overlap-layout-consensus paradigm has been implemented in MEGAN 6 that makes use of protein alignments, that is alignment of reads to protein references, for assembly of the reads. In this protein-alignment-guided assembly, perfect overlaps of reads aligning to the same protein reference are used for the creation of an overlap graph. One or more nodes belonging to one or more functional classification systems like KEGG [52], InterPro2GO [54], SEED [53] can be assembled using MEGAN. The different parameters for the assembly algorithm are the **minOverlap**, which is the minimum number of bases that the two reads should overlap by, **minReads**, the minimum number of reads required for a contig, **minLength**, minimum length for a contig, **minAvCoverage** the minimum average coverage for a contig and **maxPercentIdentity**, the maximum percent identity that two contigs are allowed to have. Other software that have specifically been designed for gene-centric assembly are Xander [83] and SAT assembler [84]. These software align reads to HMMs and assemble those reads that find a significant alignment to the HMMs. A raw

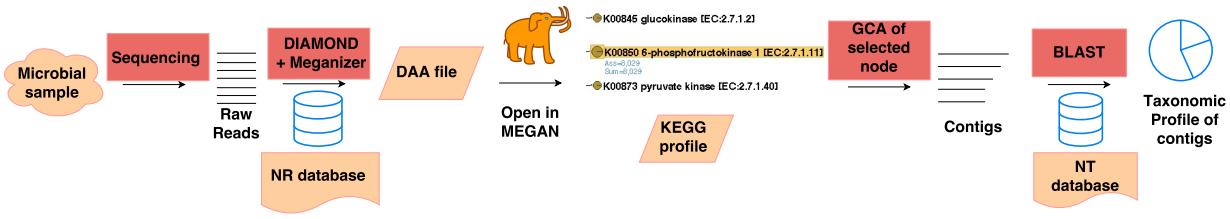
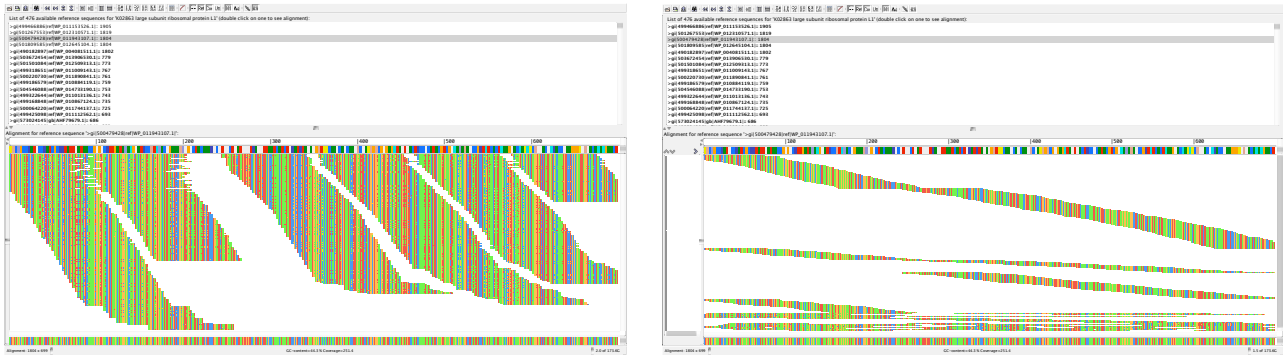


Figure 6.1: Overview of a gene-centric assembly approach.



(a) Alignment of reads

(b) Alignment of contigs

Figure 6.2: The alignment viewer in MEGAN. (a) Alignment of reads against a protein reference sequence. (b) Alignment of the reads as part of the contigs against a protein reference sequence.

metagenomic sample can be analyzed with DIAMOND, meganized with the mapping files for the desired classifications and imported in MEGAN. One or more orthologous groups can be selected and assembled. The Figure 6.1 is a pictorial representation of an overview of a gene-centric assembly workflow using MEGAN.

MEGAN also provides an alignment viewer for visualizing the alignment of reads to the protein references, and also the alignment of the contigs produced as a result. The alignment viewer is shown in Figure 6.2. In addition, the overlap graph can be exported as an image file.

6.2 Evaluation of the MEGAN assembler using a set of gene families

The evaluation of the MEGAN assembler was carried out in order to achieve two goals - one, as proof-of-principle that using gene-centric assembly is convenient and easily feasible using the DIAMOND and MEGAN pipeline and second, to compare the performance of

MEGAN's assembler with other established metagenomic assemblers. We analyzed a mock metagenomic community described in the study by Shakya et al., [56], consisting of 64 organisms. Exactly 38 single-copy phylogenetic marker genes from the study by [85] and 3 multiple-copy genes (*cheA*, *ftsZ* and *atoB*) were assembled using the gene-centric assembler. We expect that if 64 organisms are part of a community, then for each single-copy gene, 64 contigs should be produced. Each of the 64 contigs should ideally cover 100 percent of the span of the gene. A sequence analysis of these contigs (for e.g. carrying out BLASTN, or BLASTX) should result in the alignment of each contig to one of the 64 organisms, with no false positives or false negatives.

First, DIAMOND was run on the raw reads. The DIAMOND output file was imported in MEGAN and a gene-centric assembly of the marker genes was carried out. The reads aligning to the gene families were used as an input to other assembly software like SOAPdenovo [42], Ray [43] and IDBA-UD [44]. The contigs thus produced for all the marker genes by all the assembly algorithms were further evaluated for their assembly quality.

6.2.1 Mock metagenomic dataset

A mock metagenomic dataset described in the publication from [56], composed of 48 bacterial and 16 archael genomes was chosen for the study (SRA run SRR606249). The synthetic mock community was created from known amounts of purified gDNA of the 64 organisms and was used for metagenomic sequencing using 454 and Illumina sequencing technologies. The Illumina dataset consists of 101 base-pair long paired end reads with about 54 million reads per paired end file. This raw sequence data was obtained and used for the study. DIAMOND analysis of the raw reads of the mock metagenome against the NR protein database (version - February 2015) was carried out. This resulted in more than 1 billion alignments, involving 87 million reads. The DIAMOND output file was meganized with default LCA parameters and imported in MEGAN 6.

6.2.2 Genes used for evaluation

The phylogenetic marker genes for bacteria and archaea as described in a study by Wu et al. [85] were used for the evaluation. The KEGG Orthologous (KO) groups corresponding to the marker genes were inspected in the dataset. Out of the 40 marker genes that are part of the list, 35 genes had corresponding KO groups in the dataset. In addition, the well known marker gene *rpoB* and archael *rpoB'* and *rpoB"* were included, along with the three

mentioned multiple-copy genes, resulting in a total of 41 gene families. The protein sequences from all 64 genomes were downloaded. BLASTX was used to retrieve the protein sequences for the gene families used for evaluation. From these, reference lengths of the proteins sequences were calculated. Further, the complete genome sequences of the 64 organisms was downloaded. Table 6.1 contains the KO groups with their descriptions.

6.2.3 Gene-centric assembly in MEGAN

MEGAN's gene-centric assembler was run on the 41 gene-families with default parameters and a 98 percent similarity cut-off. The reads binned into the KO groups corresponding to the 41 genes were extracted. These reads were used as input to SOAPdenovo, Ray and IDBA-UD assemblers and contigs were thus obtained for the genes from each of the 4 assemblers using default parameters. The minimum length of the contigs was set to 200 base-pairs. The contigs were further evaluated for their coverage of the gene and the taxonomic profile (the number of different true positive references detected).

6.2.4 Evaluation of the contigs produced

The contigs produced for each of the 38 genes with the different assemblers were evaluated in the following three ways:

1. Contig Statistics

Basic contig statistics like the number of contigs generated by each assembler for each marker gene, the lengths of the contigs produced for each gene was determined using custom scripts.

2. Validation of the contigs by sequence analysis

The contigs produced by the assemblers were subjected to a BLASTN analysis against the NCBI-NT database. Our assumption is that the quality of the alignments of the contigs produced for each gene to the nucleotide sequences available in the database will point to the authenticity of the contigs. Also, alignment of a contig to any one of the reference organism present in the mock community will further validate that contig. The BLASTN output files were imported in MEGAN and were used for a first inspection of the BLASTN results. The BLASTN text output files were parsed using a custom script for a detailed analysis of the results. Each BLASTN output file, corresponding to the alignments of the contigs produced for one gene, was parsed separately. For each contig, out of the total number of hits that it

Table 6.1: KO groups and their descriptions : Single-copy gene families

KO group	**Description**
K00626	acetyl-CoA C-acetyltransferase
K03044	archael *rpoB1*
K03045	archael *rpoB2*
K03043	bacterial *rpoB*
K03531	cell division protein
K01889	phenylalanyl-tRNA synthetase alpha subunit
K01890	phenylalanyl-tRNA synthetase beta subunit
K01933	phosphoribosylformylglycinamidine cyclo ligase
K03470	ribonuclease HII
K02863	ribosomal protein L1
K02864	ribosomal protein L10
K02867	ribosomal protein L11
K02871	ribosomal protein L13
K02874	ribosomal protein L14
K02876	ribosomal protein L15
K02878	ribosomal protein L16
K02881	ribosomal protein L18
K02886	ribosomal protein L2
K02890	ribosomal protein L22
K02895	ribosomal protein L24
K02897	ribosomal protein L25
K02904	ribosomal protein L29
K02906	ribosomal protein L3
K02926	ribosomal protein L4
K02931	ribosomal protein L5
K02933	ribosomal protein L6
K02946	ribosomal protein S10
K02948	ribosomal protein S11
K02950	ribosomal protein S12
K02952	ribosomal protein S13
K02956	ribosomal protein S15
K02961	ribosomal protein S17
K02965	ribosomal protein S19
K02967	ribosomal protein S2
K02982	ribosomal protein S3
K02988	ribosomal protein S5
K02992	ribosomal protein S7
K02994	ribosomal protein S8
K02996	ribosomal protein S9
K03110	signal recognition particle protein
K03407	two-component system

obtained, the hit with the best bit score was considered. The hits with at least 90 percent of the score of the best hit were evaluated. If any of the hits thus evaluated belonged to any one of the 64 organisms, then this contig was considered a true positive. If none of these hits correspond to any one organism from the mock community, then the contig was deemed false positive. If the contig did not fetch any alignment, it was a "no-hit". Thus a contig was annotated as either true positive, false positive or "no-hit". For each gene, the total complement of the reference organisms that had an aligned contig is referred to as the taxonomic profile for the gene. For a reference organism, it is either a true positive, if it is part of the 64 organisms present in the mock community or false positive if it is not a part of the mock community. From the 64 organisms, if an organism is absent from the organisms detected to be present, then it is a false negative reference. Thus a reference is annotated as either true positive, false positive or false negative.

3. Mapping against the 64 reference genomes

The contigs were mapped against a database of the full genome sequences of the 64 organisms using the BWA [86] mapping tool. In case of BWA, a contig was annotated either a true positive when it mapped to any reference, or a no-hit when it did not map to any reference. A reference is either a true positive if at least one contig mapped to it or a false negative if no contig mapped to it. For each contig of each gene, the BLASTN and the BWA output was compared. The output SAM files were used as input to SAMtools [87] to produce MD tags for the SAM files. The MD tags were used to determine the coverage of the longest contig. The coverage is defined as the total number of identities in the alignment, that is the length of the alignment subtracted by the number of bases that are mismatches. The percent coverage is defined as the total coverage divided by the length of the reference. The coverage was plot as heatmap with the 64 organisms on the X-axis and the genes on the Y-axis. Sensitivity of the assembler is determined by calculating the number of true positive references that are covered at least 50% by the longest contig aligning to it, divided by the total number of references expected to be present for the gene-family.

4. Validation using DNA-protein alignment

The contigs that fail to obtain an alignment using BLASTN were subjected to a DNA-protein alignment (BLASTX). This was done in order to evaluate whether such contigs were truly invalid, or whether they fetch an alignment in a DNA-protein alignment. The BLASTX output files were parsed in order to determine whether the contig aligns to the same gene

that it is produced for. In this case, only the best hit was considered. All contigs align to a valid reference in either a DNA-DNA or DNA-protein alignment.

6.2.5 Results

Contig statistics

The number of contigs that were produced for each marker gene by each assembler had a positive correlation with the number of reads that got recruited per marker gene. For example, the *rpoB* gene had the highest number of reads assigned to it, and also the highest number of contigs produced for it, i.e. 184 (by MEGAN's assembler). All assemblers produce a comparable number of contigs, as shown in Figure 6.3. IDBA-UD tends to produce the least number of contigs per gene. All the contigs produced by the assemblers get validated, either through a DNA-DNA (BLASTN) or a DNA-protein (BLASTX) alignment. For all assemblers, on an average $\sim$ 98% of the contigs aligned with $\sim$ 98 % identity to a true positive reference.

Sensitivity of determining a true positive reference

In the ideal case, for each single-copy gene, one contig for each of the 64 organisms is expected to be produced. But the taxonomic profile obtained for each gene does not contain all 64 organisms. We describe a reference to be "detected" if at least 50% of it is covered by the longest contig produced for it. Figure 6.4 shows the sensitivity of the different assemblers for the genes. MEGAN shows a very high sensitivity in most cases as compared to the other assemblers.

No assembler produces contigs that represent the complete complement of 64 organisms that are part of the mock community. In some cases, for example the archael *rpoB*, it is expected that some of the bacterial reference organisms will not appear in the total number of true positives. The *rpoB* gene codes for the beta-subunit of the bacterial RNA polymerase and is an important single-copy gene involved in transcription. In archael genomes, *rpoB'* and *rpoB"* code for the beta-subunit of the gene. For *rpoB*, the total number of true positive organisms was thus 48 and for both *rpoB'* and *rpoB"* it is 16. For ribosomal protein L29, the number of reads assigned is the lowest leading to very few contigs being generated for the same. Consequently, few organisms are part of the profile generated for it. There are some references however, that are not part of the profile for any gene. For example, *Sulfurihydrogenibium yellowstonense SS-5* is not a part of any profile. This bacterium has

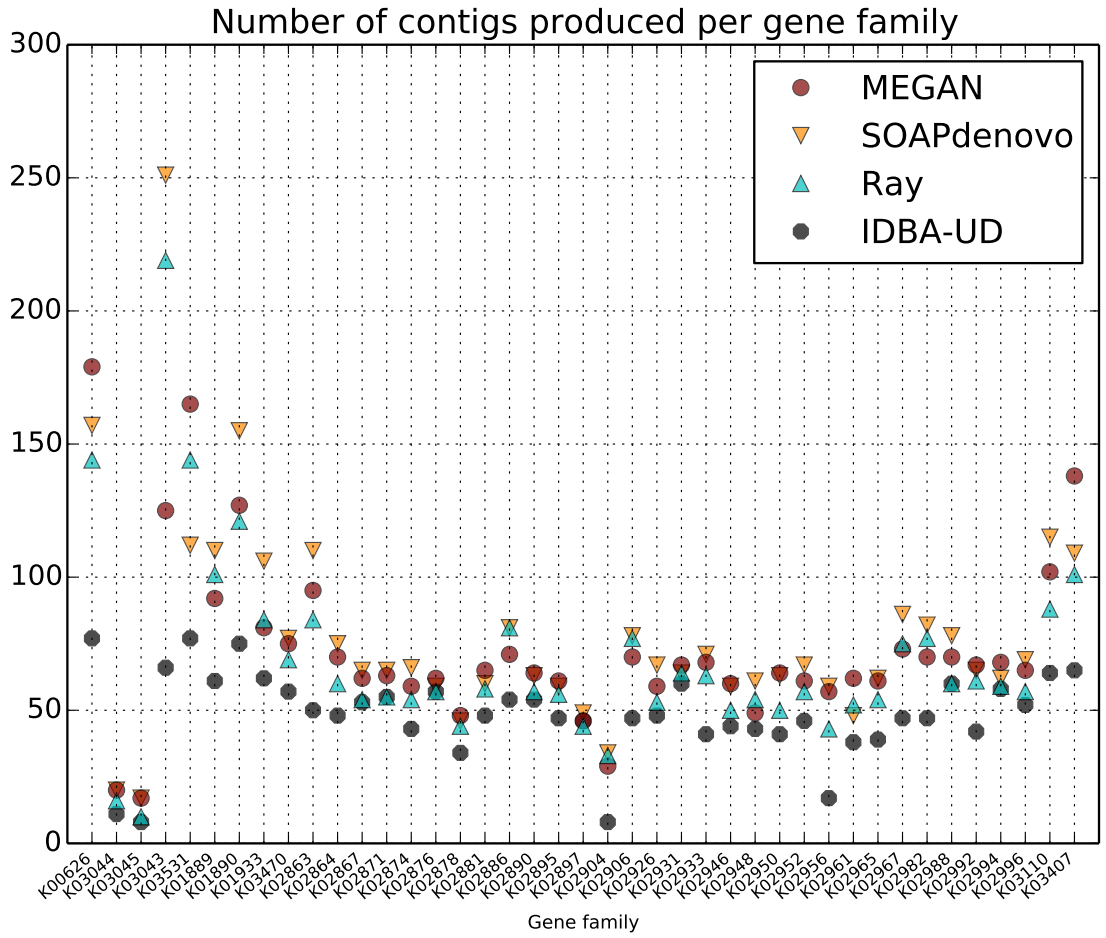


Figure 6.3: The number of contigs produced and validated by all assemblers is shown as a scatter plot.

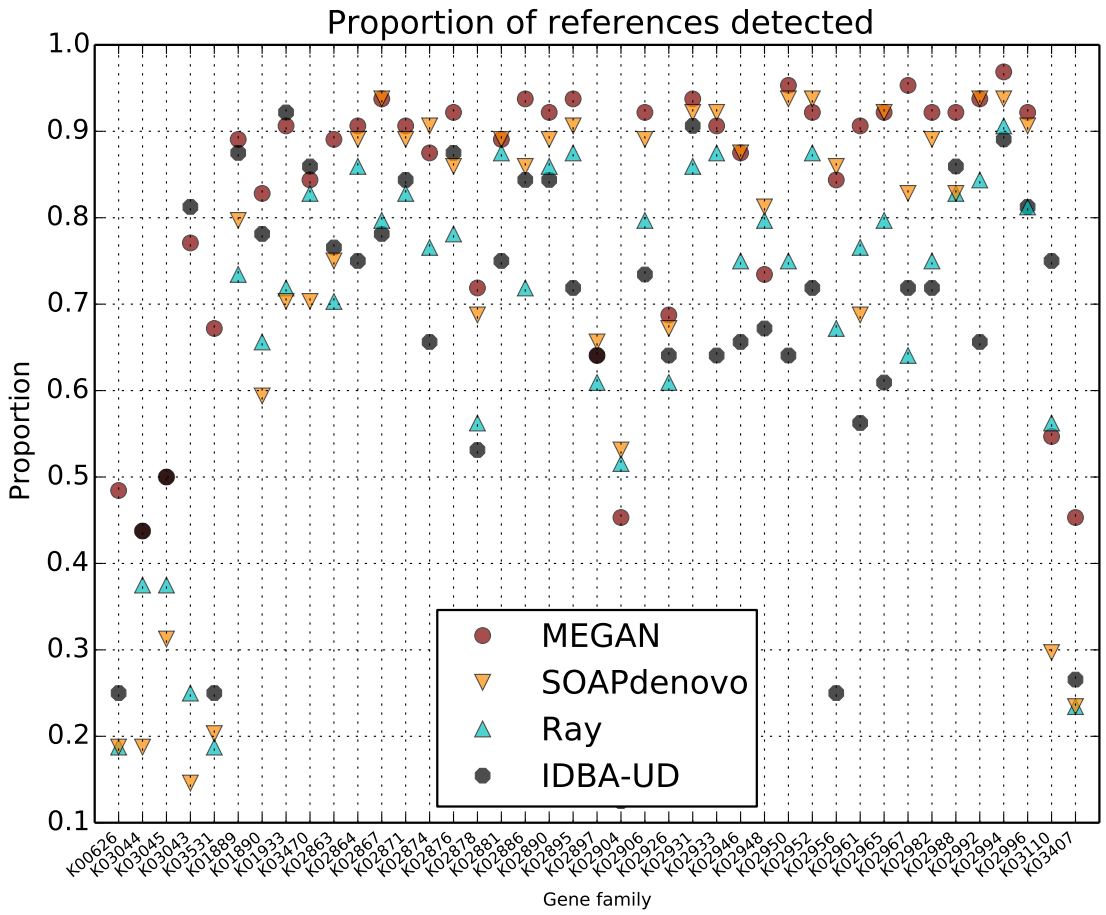


Figure 6.4: The sensitivity of the assemblers in detecting the reference organisms for different genes.

a low number of reads assigned to it as compared to the other organisms. An Illumina read simulation of the genome of this bacterium using the ART simulator [80] was carried out and the resulting reads were analyzed using DIAMOND. However very few KEGG orthologous groups are detected to be present for this bacterium. This explains why the bacterium is a false negative for all the genes.

Coverage of the genes by the longest contig

For a given gene family, the coverage of the longest gene to its reference was calculated. The Figures 6.5, 6.6, 6.7, 6.8 are the percent coverages plot as heatmaps, darker the color, higher is the percent. The value -1 indicates that no contig maps to the reference. The coverage plot also shows the true positive references obtained per gene. Overall, the percent coverage by the contigs produced by MEGAN is high. SOAPdenovo and Ray produce long contigs for most genes, but do not produce contigs for several reference organisms.

The taxonomic profile created for a specific gene is similar across the assemblers. For example, in the case of phenylalanyl-tRNA synthetase alpha subunit and phenylalanyl-tRNA synthetase beta subunit, the profiles cluster with respect to the gene and not with respect to the assembler. As shown in Figure 6.9, the profiles generated for alpha subunit by the assemblers are more similar to each other than the profiles generated for the beta subunit. All assemblers perform uniformly over the genes, serving as a sanity check and pointing to the fact that a targeted assembly is a useful analysis strategy.

6.3 Gene-centric assembly across samples and gene-families

In the following section, we describe the analysis of contigs assembled with MEGAN's assembler from reads belonging to a wide range of orthologous gene families and from both real and simulated datasets.

6.3.1 Mock metagenome, multiple-copy genes

In order to examine the contigs produced from a gene-centric assembly of gene families that are involved in metabolism and are not usually single-copy genes, we considered a wider range of gene families. We run the analysis on all genes belonging to the Glycolysis pathway. The KO groups and their descriptions are listed in Table 6.2.

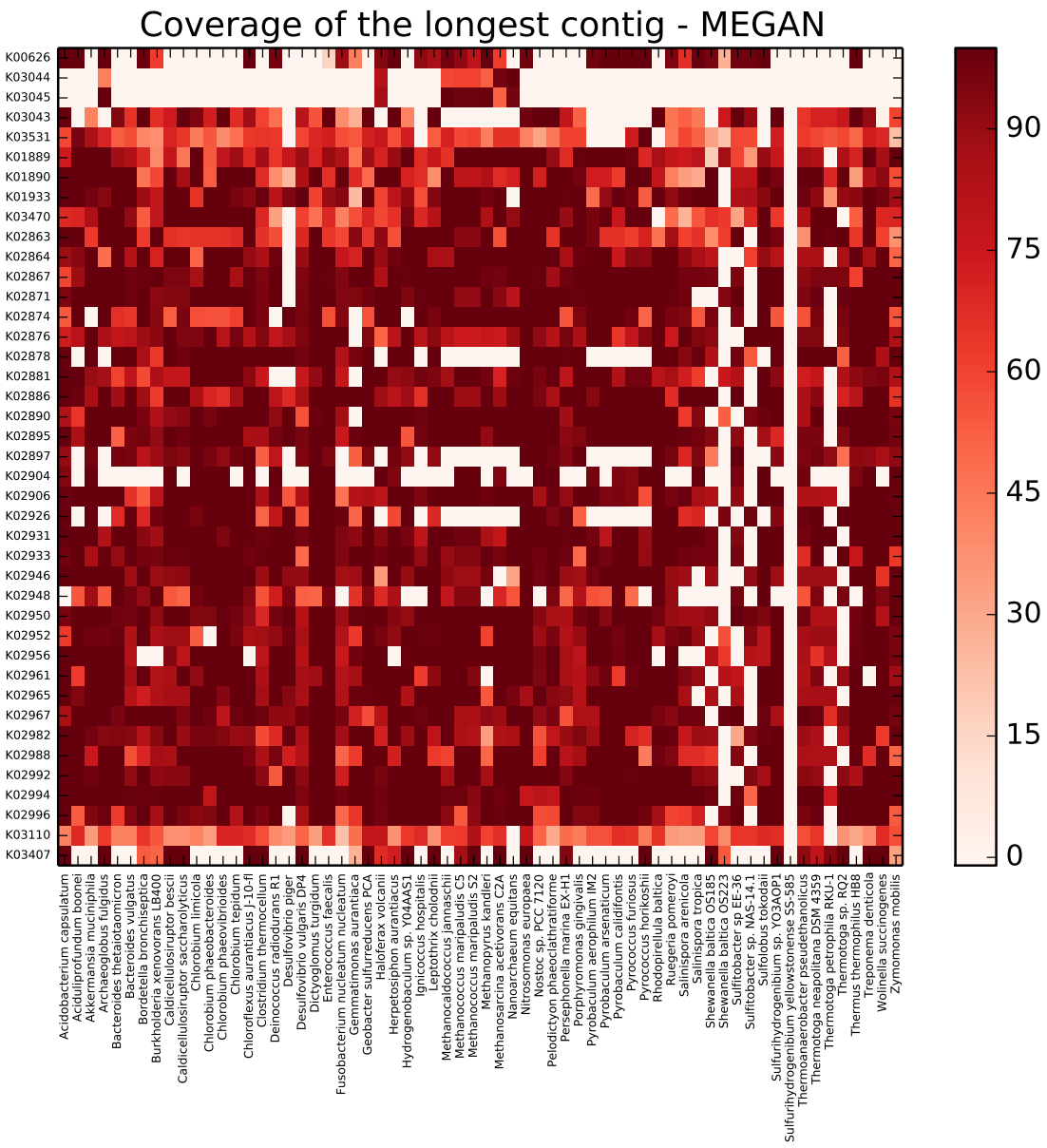


Figure 6.5: Percentage coverage of the longest contig - MEGAN

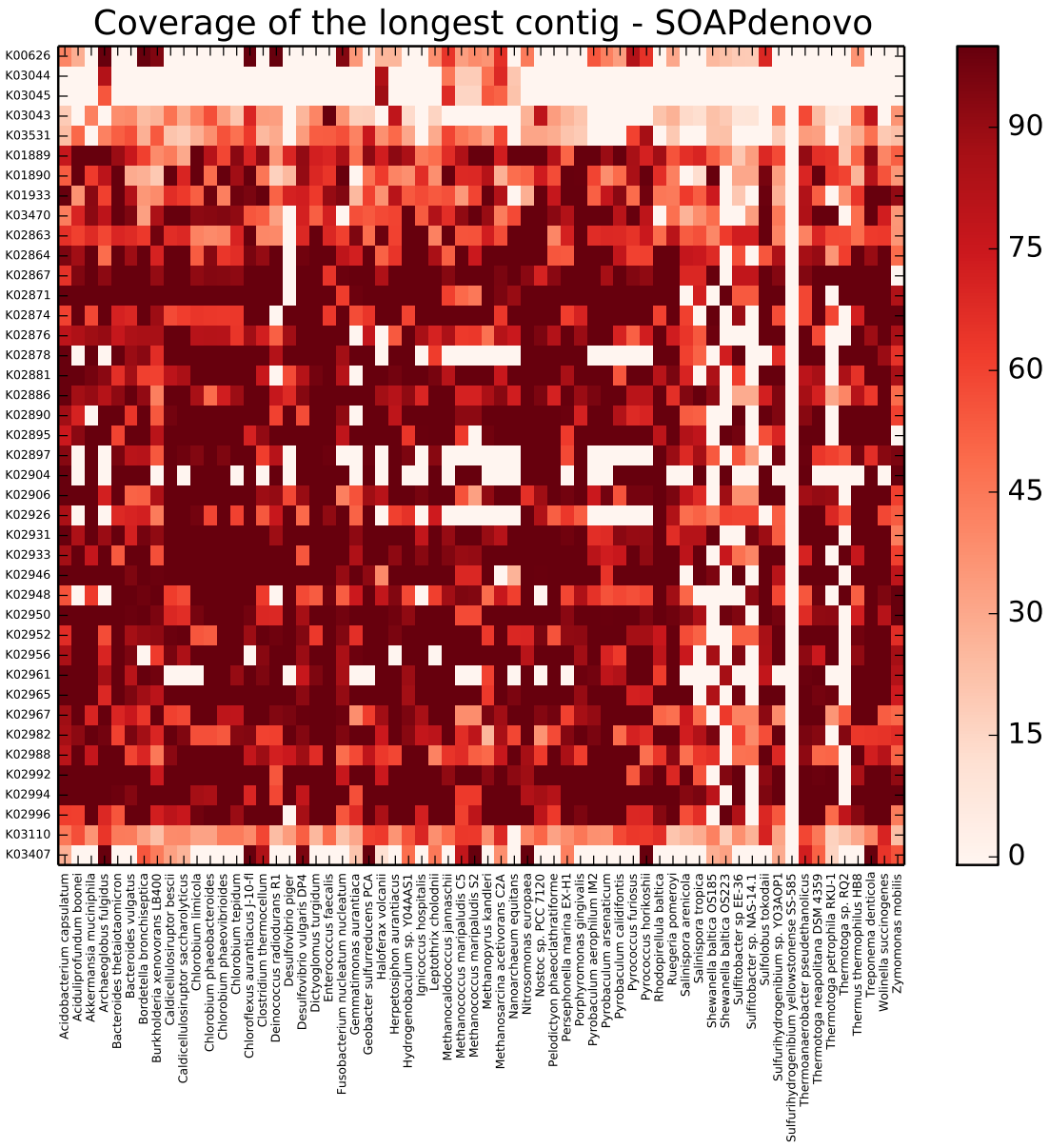


Figure 6.6: Percentage coverage of the longest contig - SOAPdenovo

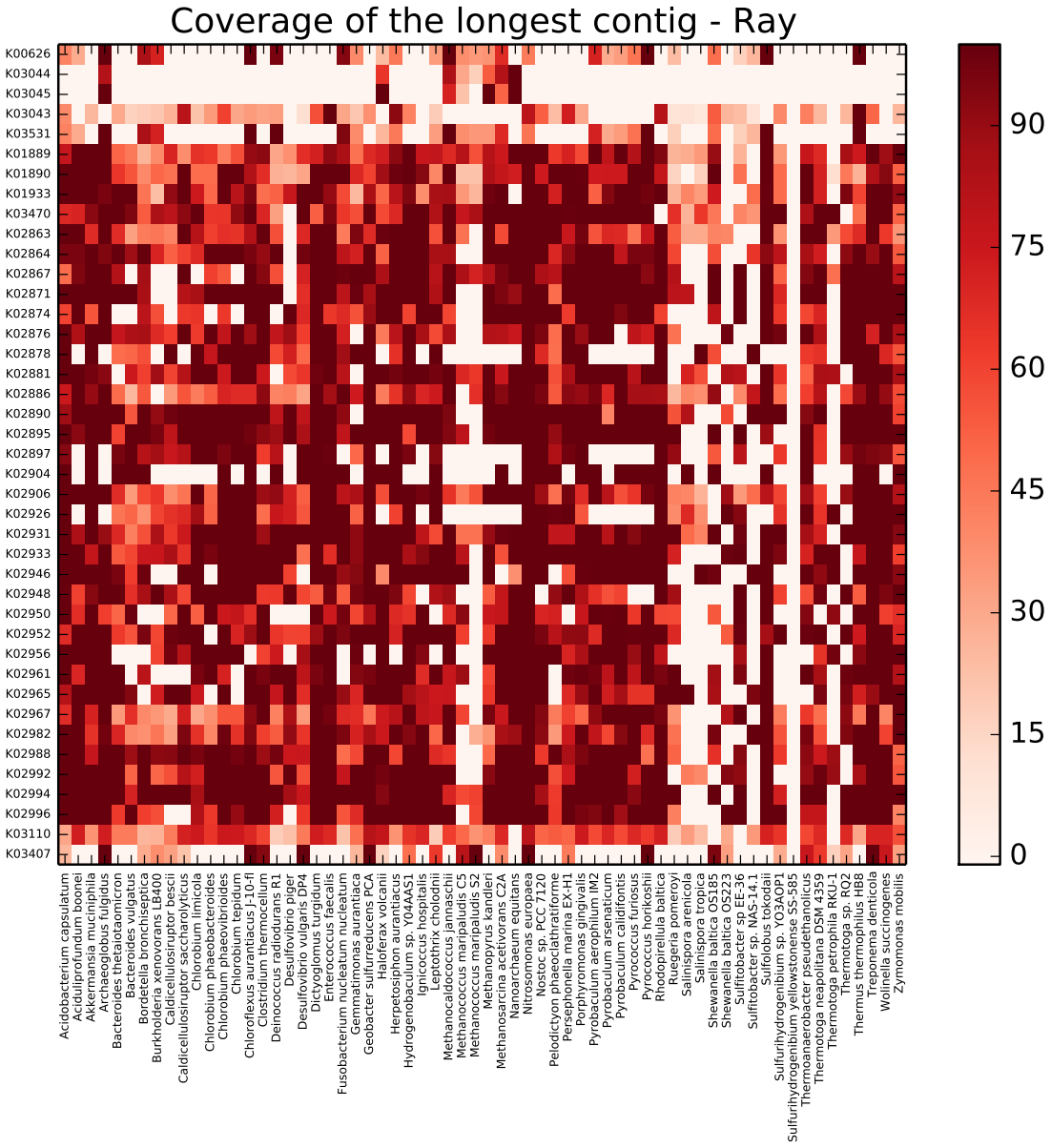


Figure 6.7: Percentage coverage of the longest contig - Ray

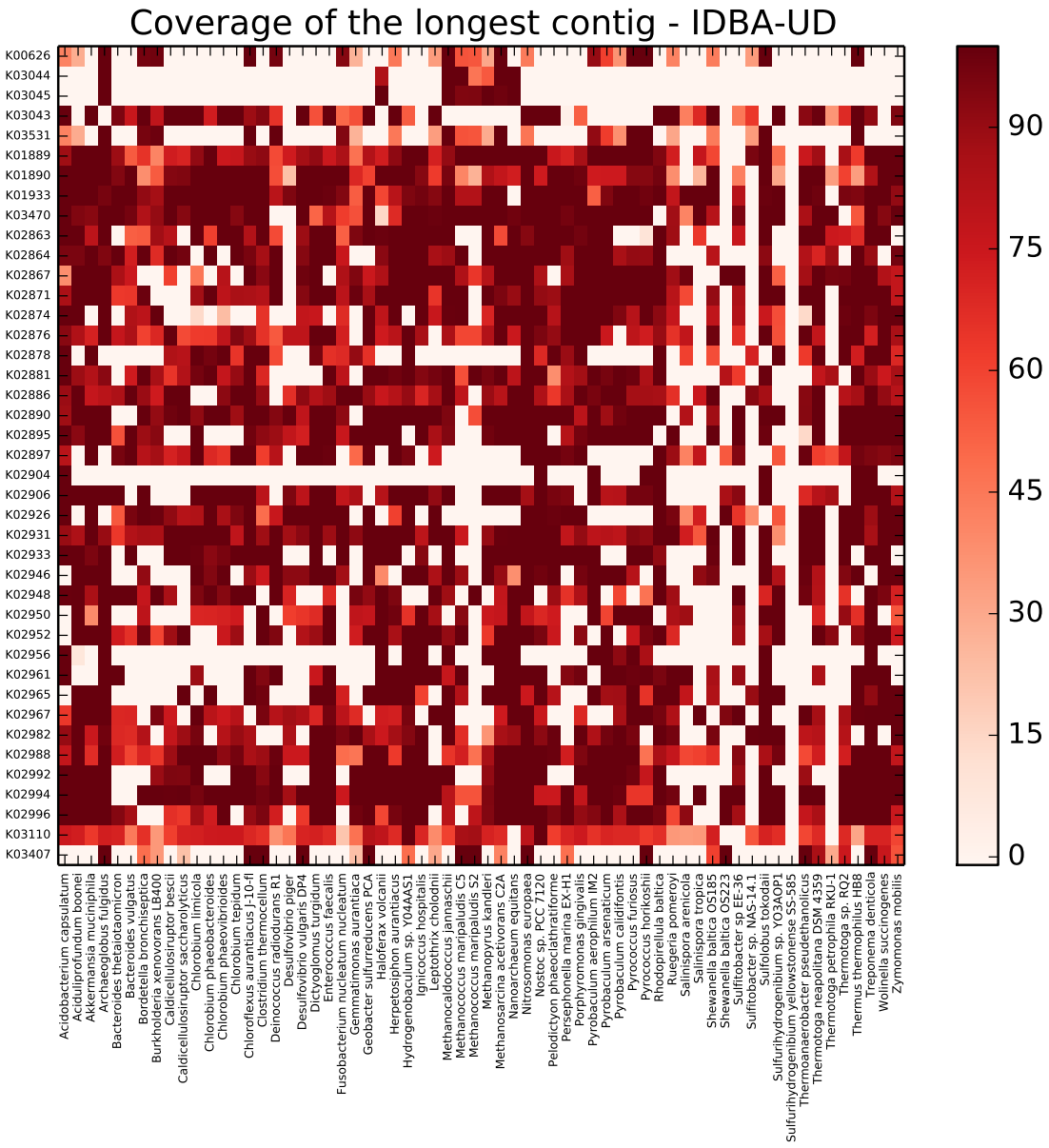


Figure 6.8: Percentage coverage of the longest contig - IDBA-UD

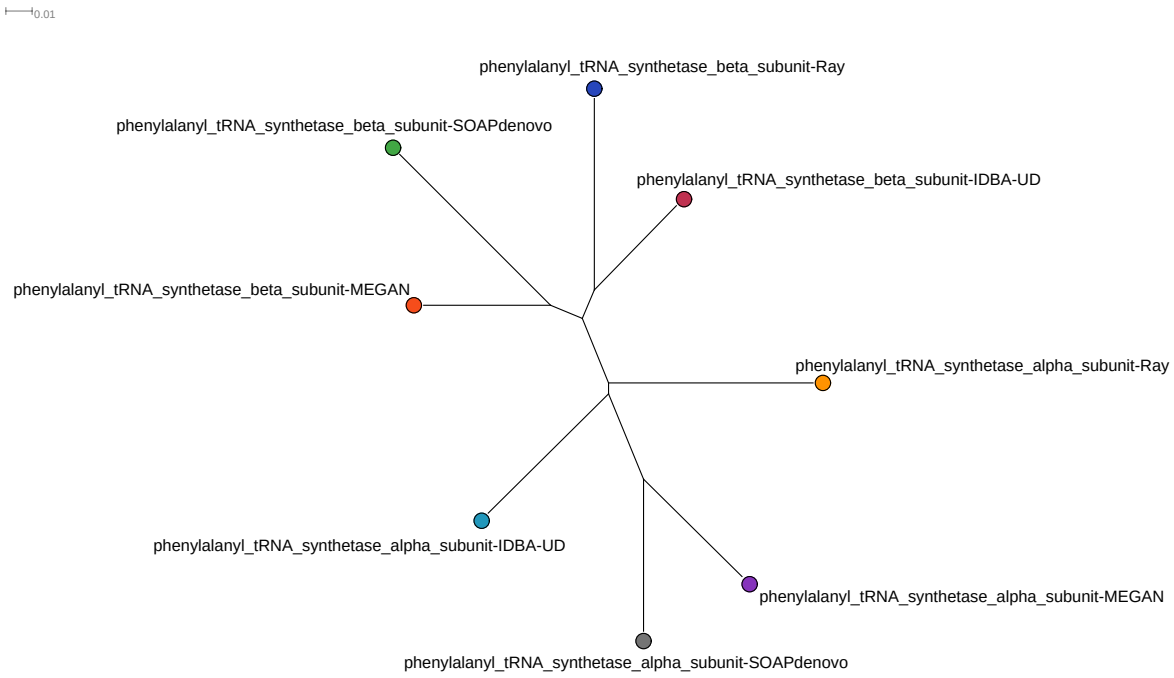


Figure 6.9: Clustering of the taxonomic profiles generated by the contigs for the phenylalanyl-tRNA alpha and beta genes shows that the profiles for alpha and beta subunits from all the assemblers cluster together. The clustering was performed in MEGAN using the Jensen-Shannon Divergence method.

Table 6.2: KO groups and their descriptions : Glycolysis pathway gene families

KO group	Description
K01895	acetyl-CoA synthetase
K00873	pyruvate kinase
K00382	dihydrolipoamide dehydrogenase
K01803	triosephosphate isomerase (TIM)
K00927	phosphoglycerate kinase
K01689	enolase
K01834	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase
K01624	fructose-bisphosphate aldolase, class II
K00845	glucokinase
K00850	6-phosphofructokinase 1
K00134	glyceraldehyde 3-phosphate dehydrogenase
K01810	glucose-6-phosphate isomerase
K00001	alcohol dehydrogenase
K01596	phosphoenolpyruvate carboxykinase (GTP)
K03738	aldehyde:ferredoxin oxidoreductase
K00171	pyruvate ferredoxin oxidoreductase delta subunit
K00172	pyruvate ferredoxin oxidoreductase gamma subunit
K00169	pyruvate ferredoxin oxidoreductase alpha subunit
K00170	pyruvate ferredoxin oxidoreductase beta subunit
K03841	fructose-1,6-bisphosphatase I
K00128	aldehyde dehydrogenase (NAD+)
K00627	pyruvate dehydrogenase E2 component (dihydrolipoamide acetyltransferase)
K00161	pyruvate dehydrogenase E1 component alpha subunit
K00162	pyruvate dehydrogenase E1 component beta subunit
K00016	L-lactate dehydrogenase
K13953	alcohol dehydrogenase, propanol-preferring
K01785	aldose 1-epimerase
K01610	phosphoenolpyruvate carboxykinase (ATP)
K01835	phosphoglucomutase
K00121	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase
K00163	pyruvate dehydrogenase E1 component
K02446	fructose-1,6-bisphosphatase II
K01623	fructose-bisphosphate aldolase, class I
K13954	alcohol dehydrogenase
K01792	glucose-6-phosphate 1-epimerase
K01222	6-phospho-beta-glucosidase
K04041	fructose-1,6-bisphosphatase III
K00886	polyphosphate glucokinase
K02779	PTS system, glucose-specific IIC component
K04072	acetaldehyde dehydrogenase / alcohol dehydrogenase
K02777	PTS system, sugar-specific IIA component
K01223	6-phospho-beta-glucosidase
K02791	PTS system, maltose/glucose-specific IIC component
K11532	fructose-1,6-bisphosphatase II / sedoheptulose-1,7-bisphosphatase
K00844	hexokinase
K00129	aldehyde dehydrogenase (NAD(P)+)
K00114	alcohol dehydrogenase (cytochrome c)
K14028	methanol dehydrogenase (cytochrome c) subunit 1
K00002	alcohol dehydrogenase (NADP+)
K13810	transaldolase / glucose-6-phosphate isomerase
K00131	glyceraldehyde-3-phosphate dehydrogenase (NADP+)
K06859	glucose-6-phosphate isomerase, archaeal
K00150	glyceraldehyde-3-phosphate dehydrogenase (NAD(P))
K00918	ADP-dependent phosphofructokinase/glucokinase
K01905	acetyl-CoA synthetase (ADP-forming)
K11389	glyceraldehyde-3-phosphate dehydrogenase (ferredoxin)

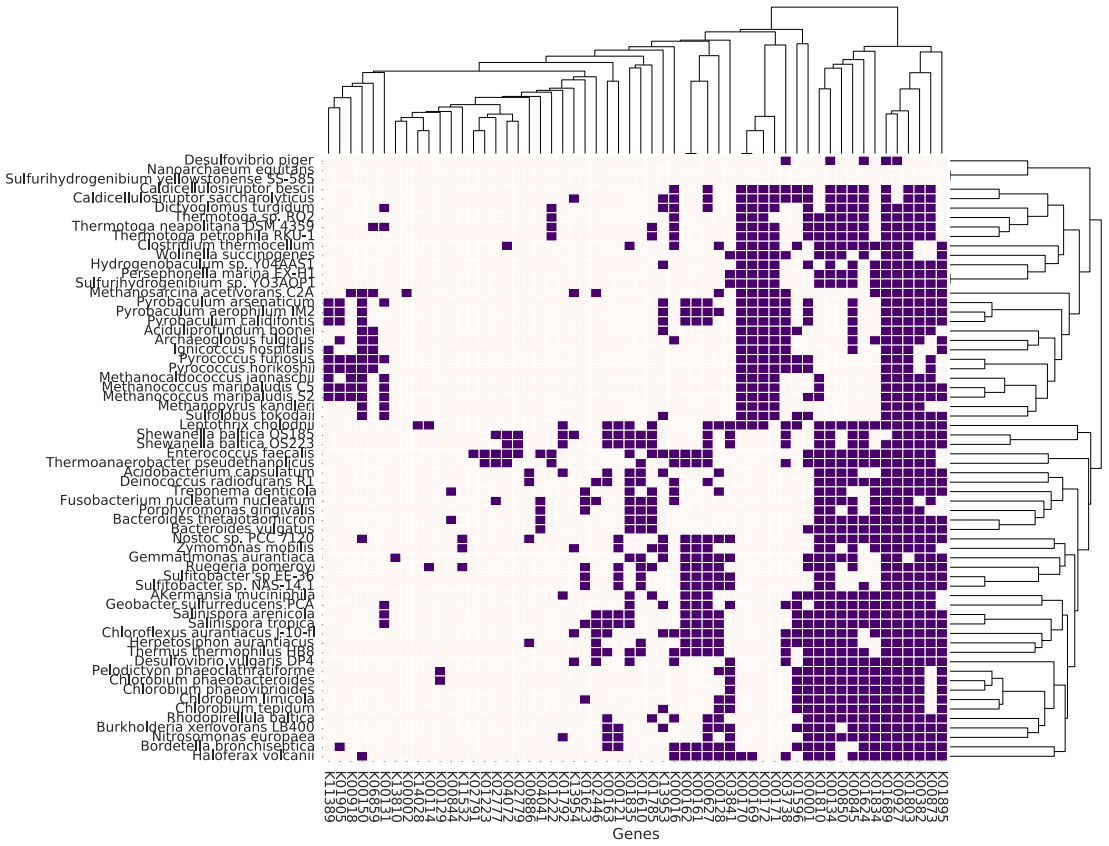


Figure 6.10: A two-way hierarchical clustering shows the organisms grouping according to the presence or absence of contigs belonging to the gene families studied.

Since this pathway is one of the predominant pathways in most metagenomic datasets and consists of many different genes, it was chosen for the analysis. As described previously, a gene-centric assembly was carried out in MEGAN with default parameters and the resulting contigs were mapped against the genome database of 64 reference genomes (with the BWA tool). Depending on whether at least one contig per gene mapped to a reference organism from the mock community, a matrix was created, with the values 1 and 0. The value 1 indicates that at least one contig mapped to the reference and the value 0 indicates that no contig mapped to the reference. The matrix was subjected to a two-way hierarchical clustering using the Python Scipy [88] library and was plot as a heatmap, Figure 6.10.

The clusters so formed are related genes that have similar taxonomic profiles and related organisms that have similar gene profiles. Most reference organisms have at least one contig that belongs to the following KO groups - K01895, K00873, K00382, K01803, K00927 and

Table 6.3: Lysine pathway gene families

KO group	Description
K01929	UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase
K01928	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate–2 6-diaminopimelate ligase
K00290	saccharopine dehydrogenase NAD L-lysine forming
K05825	2-aminoadipate transaminase
K03340	diaminopimelate dehydrogenase
K01586	diaminopimelate decarboxylase
K10206	LL-diaminopimelate aminotransferase
K01778	diaminopimelate epimerase
K01439	succinyl-diaminopimelate desuccinylase
K00821	acetylornithine N-succinyldiaminopimelate aminotransferase
K00674	2 3 4 5-tetrahydropyridine-2-carboxylate N-succinyltransferase
K00215	4-hydroxy-tetrahydrodipicolinate reductase
K01714	4-hydroxy-tetrahydrodipicolinate synthase
K00133	aspartate-semialdehyde dehydrogenase
K12524	bifunctional aspartokinase homoserine dehydrogenase 1
K00928	aspartate kinase
K00003	homoserine dehydrogenase

K01689. These KO groups correspond to the most important enzymes in the glycolysis pathway and their ubiquity is expected. Similarly, the KO groups K00169, K00170, K00171 and K00172 which are the pyruvate ferredoxin oxidoreductase alpha, beta, delta and gamma subunits, cluster according to their presence in a subset of the organisms. The grouping of the 14 archaeal organisms is based on the presence of the contigs for the gene-families K00131, K06859 , K00150, K00918, K01905, K11389 which play a glycolytic role in the hyperthermophilic archaea [89]. Contigs for these genes are missing in the profile for the other organisms. In this case, the contigs obtained after assembly could be segregated depending on their taxonomic origin and studied further.

6.3.2 Real metagenome, single-copy and multiple-copy genes

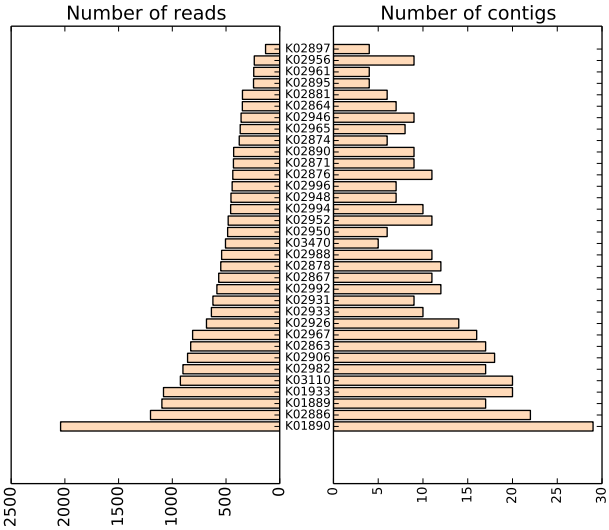
A gut microbiome sample from the Hohenheim Obesity Project was chosen for the analysis of the performance of gene-centric assembly on samples from a real metagenomic sequencing experiment. Both single (38 marker genes) and multiple-copy genes were assembled. Out of all different pathways, the lysine biosynthesis pathway attracts an average amount of reads and therefore was chosen for the analysis. The list of those genes is provided in the Table 6.3. Ideally, we would expect the taxonomic profile as analyzed using the reads to be similar to the taxonomic profile obtained by the analysis of the contigs produced for each gene. However, since the sequencing depth may not always be high, one or more contigs per gene for the rare taxonomic groups is highly unfeasible. To analyze the contigs, we carry out BLASTN against a genome database and compare the taxonomic profile thus obtained to the taxonomic profile obtained with raw reads.

The sample chosen for this was AS66_6 (SRA run SRR2155395), also because of its average size. The DAA file is available at the MeganServer database in the folder LouisEtAl2016. The 38 single copy genes and the lysine biosynthesis pathway genes were assembled with default parameters. They were analyzed using the usual DIAMOND and MEGAN pipeline. Out of the 38 genes, 34 were assembled, since no reads were assigned to 4 genes. Both BLASTN and BLASTX were carried out. In addition to the gene-centric assembly, the reads belonging to the orthologous groups studied were extracted and DIAMOND was run on them. The meganized DAA file was imported in MEGAN 6.

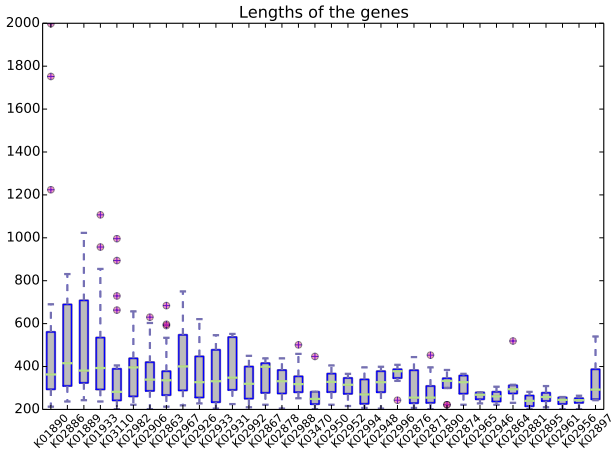
The number of reads and the contigs for single and multiple copy genes are shown in the Figure 6.11.

The number of reads and the number of contigs have a positive correlation. The reads taxonomic profile contains many more species than the contigs taxonomic profile. For the genes in this case, the MEGAN read count data consisting of the organism and the number of contigs that are binned into the group were considered as input to clustering. The genus and species level ranks were selected. A taxonomic group was considered only if at least one contig aligns to it. As can be seen from the heatmap for the single-copy genes, most species do not have a contig for all the 34 genes. The taxonomic groups that are known to be abundant in the dataset obtain at least one contig for almost all the genes. The different genera abundant in the sample are the Alistipes, Clostridium, Bacteroides, Faecalibacterium, Eubacterium, Suttarella, Roseburia etc. and each of these genera have one or more contigs aligned to them, that belong to different gene-families. Thus a gene-centric assembly in this case may produce contigs for the organisms that are abundant in the sample and therefore have a sufficient number of reads belonging to them. This explains the positive correlation between the number of reads and the number of contigs. The Figure 6.12 is a two-way hierarchical clustering of the data.

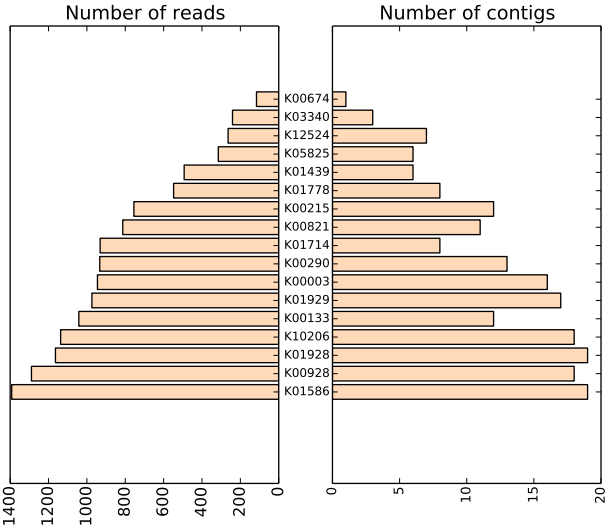
Same is the case for mulitple-copy genes where the contigs are created only for the most abundant organisms. However, as compared to the single-copy genes, this profile is still less diverse. One explanation could be that for the single-copy genes, they are ubiquitous, and are expected to be most present in most species, and therefore more number of species are expected to be a part of it. In the case of these genes, the number of "no-hits" is very high. The contigs belonging to these nodes were subjected to a BLASTX against the NCBI-NR database. The contigs align to the reference proteins that they are expected to be from. One reason for the presence of such contigs could be due to the presence of uncharacterized species that are part of the gut microbiota.



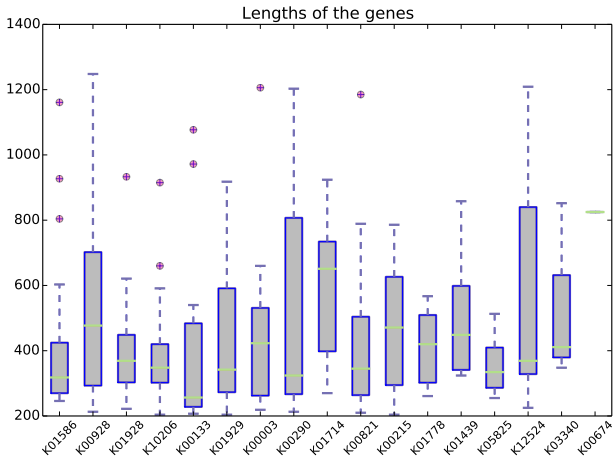
(a) Number of reads and contigs - Single-copy genes



(b) Lengths of the contigs - Single-copy genes



(a) Number of reads and contigs - Lysine pathway.



(b) Lengths of the contigs - Lysine pathway.

Figure 6.11: Number of reads, number of contigs and length of contigs for AS66_6

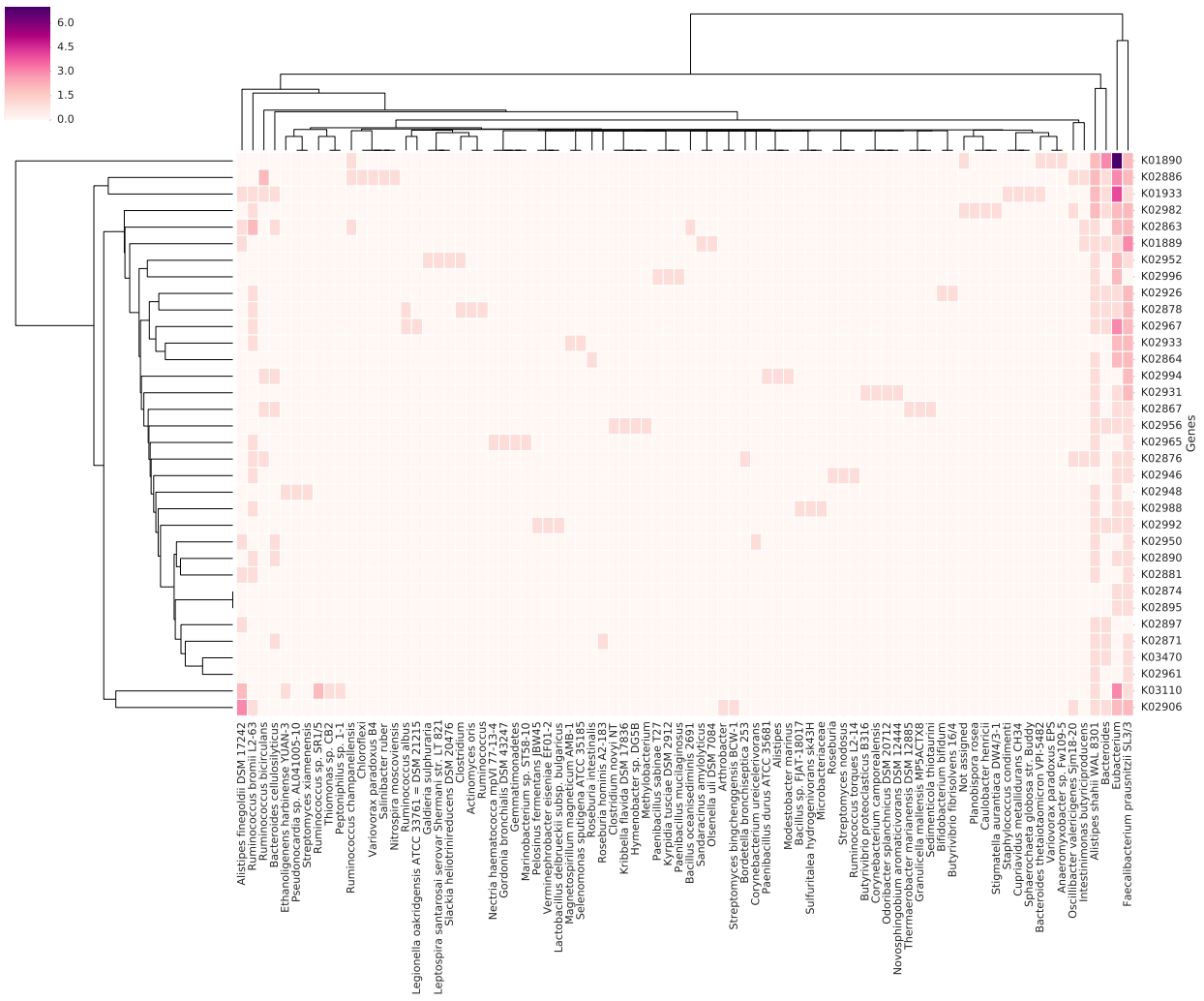


Figure 6.12: A two-way hierarchical clustering of the profile for the AS66_6 single copy genes. The abundant species in the sample have at least one contig for most gene families belonging to them.

The Figure 6.13 is a two-way hierarchical clustering of the data.

6.3.3 Simulated metagenome, single-copy and multiple-copy genes

For the simulated metagenome used in the assembly study, a gene-centric assembly was run on the reads assigned to the 38 single-copy genes and the genes from the lysine pathway. Figure 6.14 shows the number of reads, and the number of contigs and a boxplot of the length of the contigs.

The simulated metagenome was used to determine whether perfect contigs for all the genes studied are obtained, since this is an easy sample with deep enough sequencing. Also in the case of the simulated metagenome, the taxonomic profile for the reads and the contigs differs, with the profile for the contigs being less diverse. For single-copy genes, most reference organism had at least one contig aligned to it. But for multiple-copy genes like the lysine pathway gene-families, only a few contigs corresponding to all reference organisms is produced.

It is expected that for single-copy genes, for 10 organisms in the community, 10 contigs will be produced. In the case of the *rpoB* gene, 73 contigs were produced where multiple contigs find an alignment to each reference organism and at least one contig for each of the 10 organisms is present. For most other genes, not all organisms had a contig aligning to them, with the profile for each gene being different. The profile for related genes is, however similar, and phenylalanyl-tRNA synthetase is a good example. For example, the alpha and beta subunit of the gene have a similar taxonomic profile, barring just one organism (Figure 6.15).

For aspartate kinase, 10 contigs corresponding to the 10 genomes were produced (Figure 6.16). With a simple, less diverse and very well sequenced metagenome, contigs for every species present is relatively easy to obtain, however this depends on the gene being assembled, in this case aspartate kinase.

Figure 6.17 is the two-way hierarchical clustering heatmap for the taxonomic profile of the contigs belonging to single-copy genes and Figure 6.18 for the Lysine pathway genes. Even for a simple, less diverse and deeply sequenced sample like the simulated metagenome, there are cases where contigs for all reference sequences are not obtained. This further confirms that the gene family being assembled affects the results of the assembly.

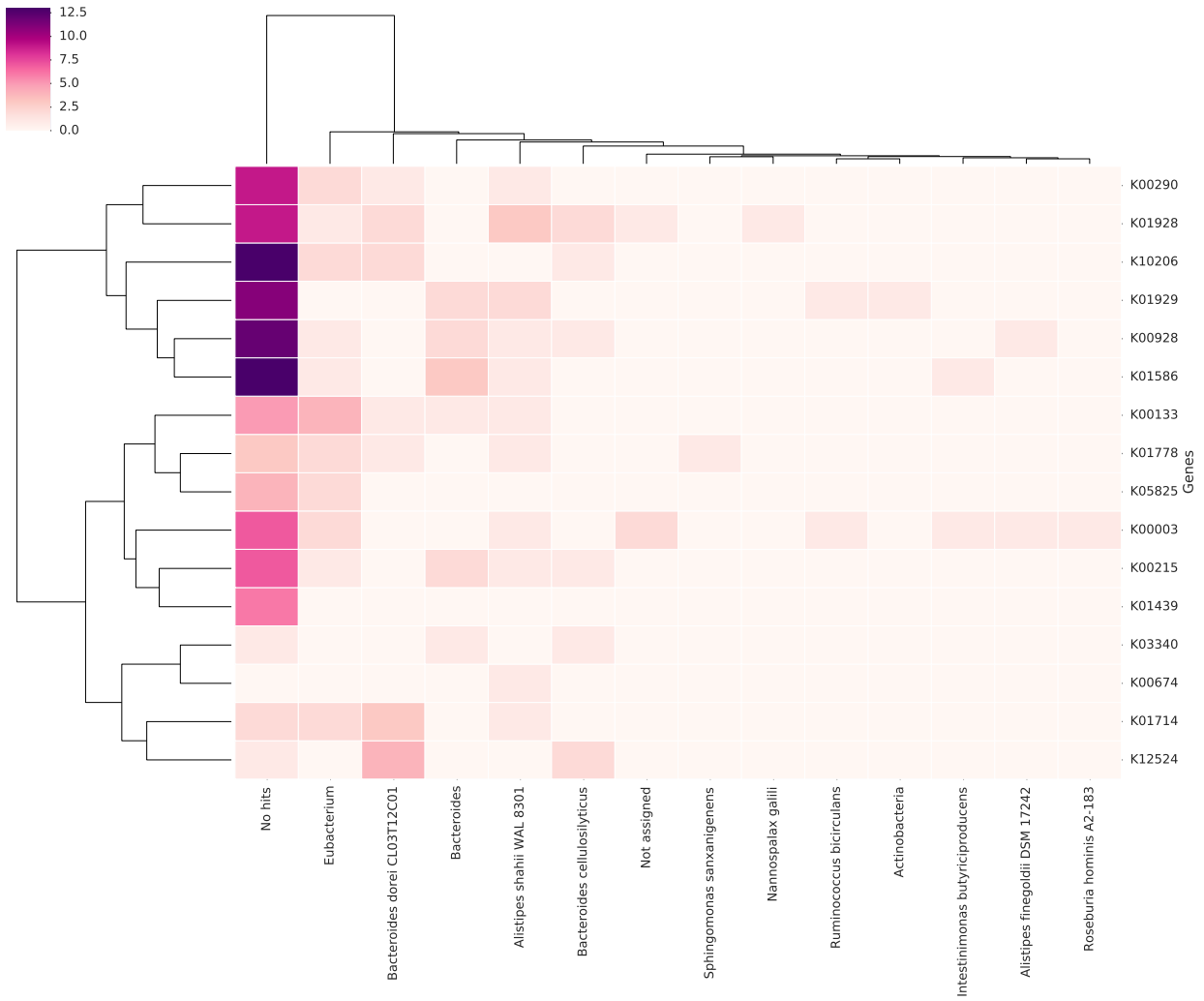
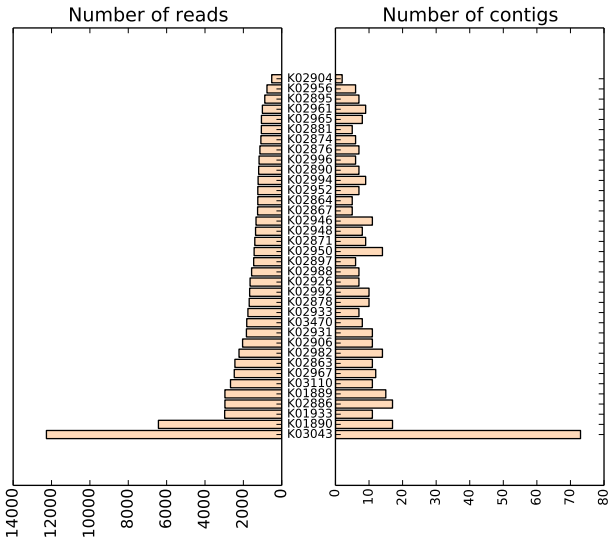
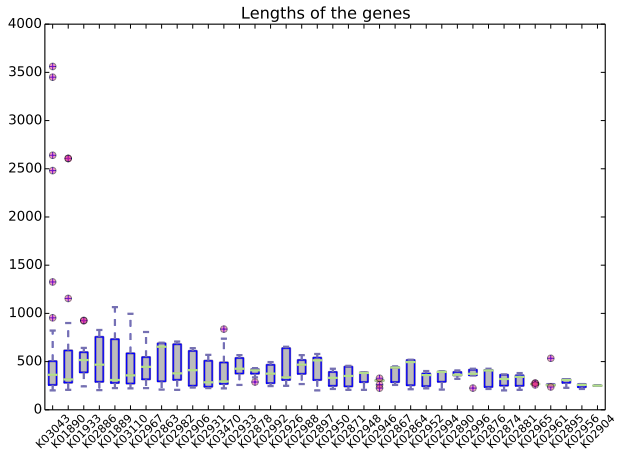


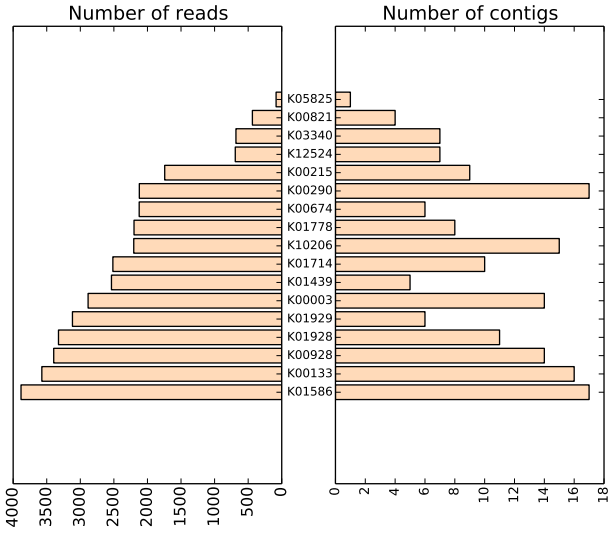
Figure 6.13: A two-way hierarchical clustering of the profile for the AS66_6 lysine pathway genes shows that very few species have at least one contig for the gene families part of the pathway. A high number of "no-hits" are present.



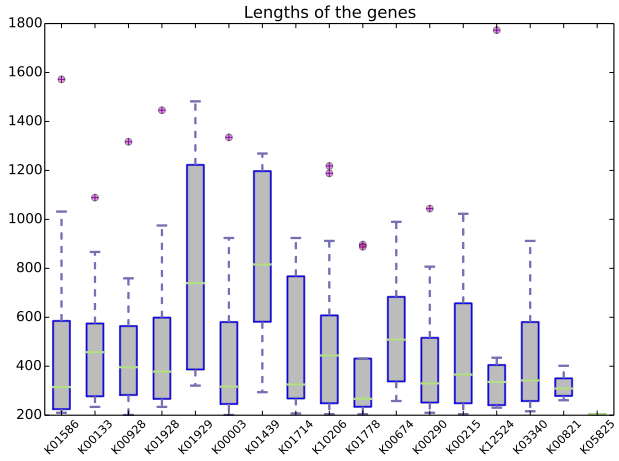
(a) Number of reads and number of contigs.



(b) Lengths of the contigs.



(a) Number of reads and number of contigs.



(b) Lengths of the contigs.

Figure 6.14: Number of reads, number of contigs and length of contigs

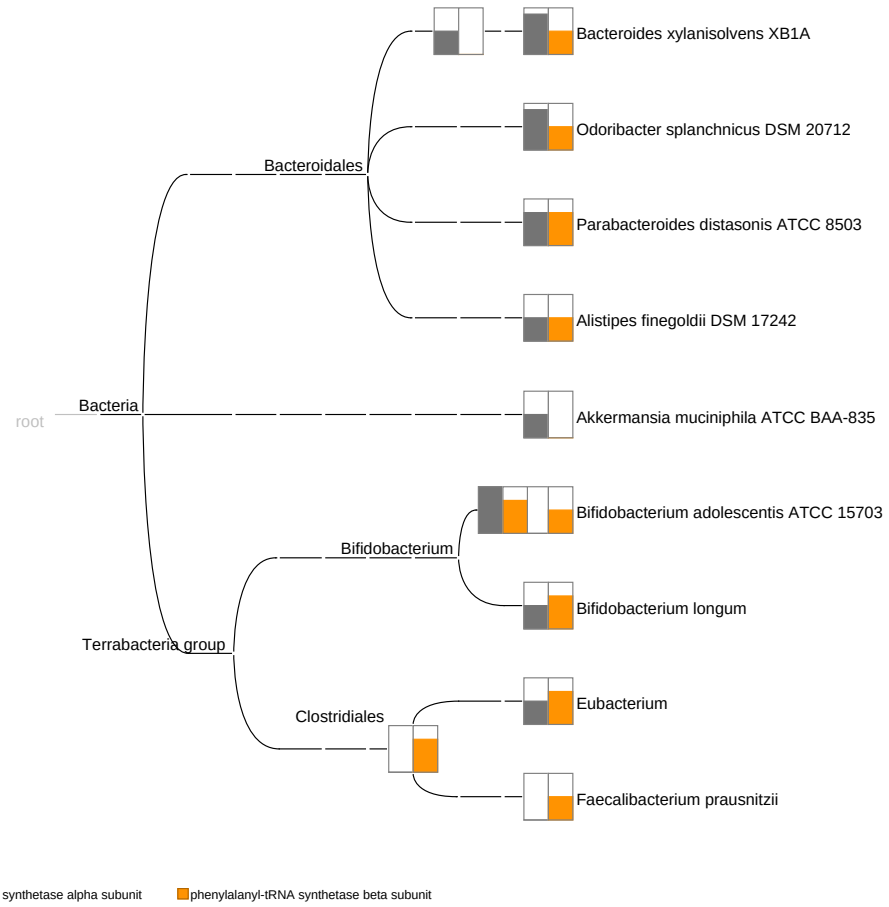


Figure 6.15: The taxonomic profile for the phenylalanine-tRNA alpha and beta subunits are highly similar to each other as compared to the other genes.

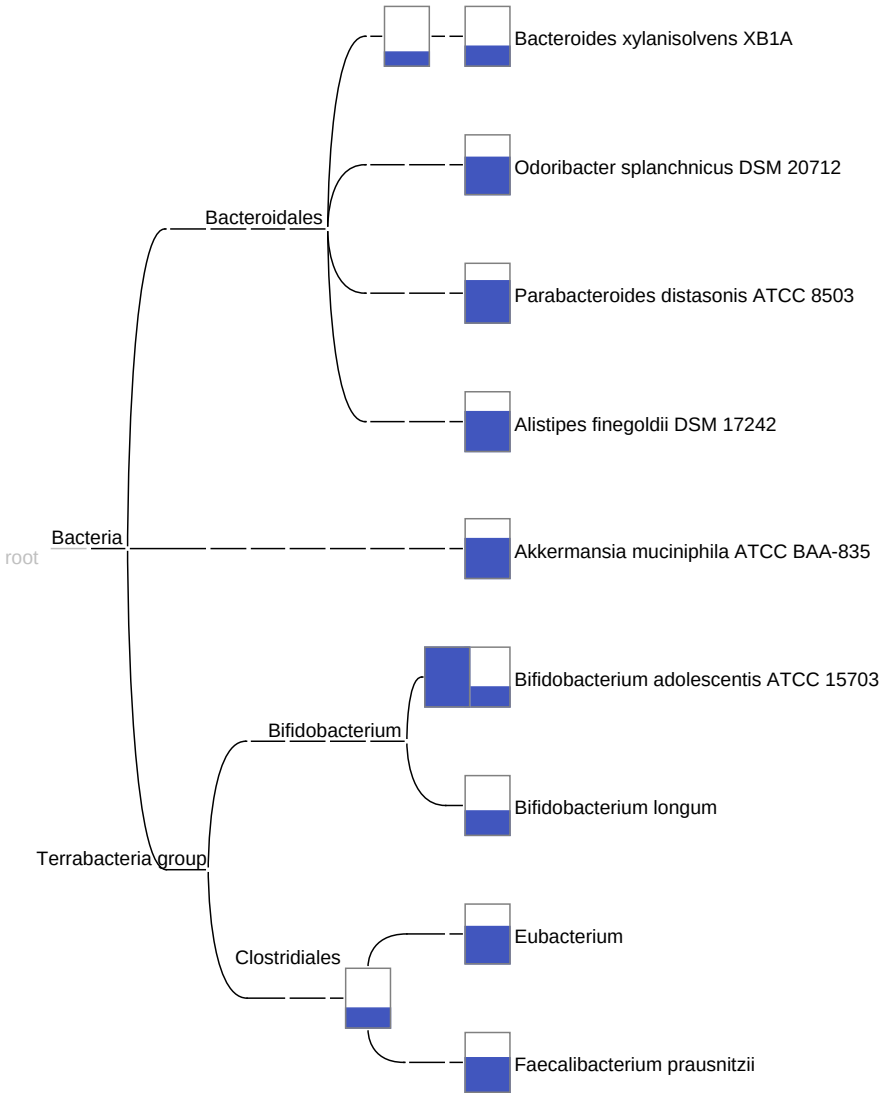


Figure 6.16: The aspartate kinase gene with all 10 contigs corresponding to 10 genomes of the simulated metagenome.

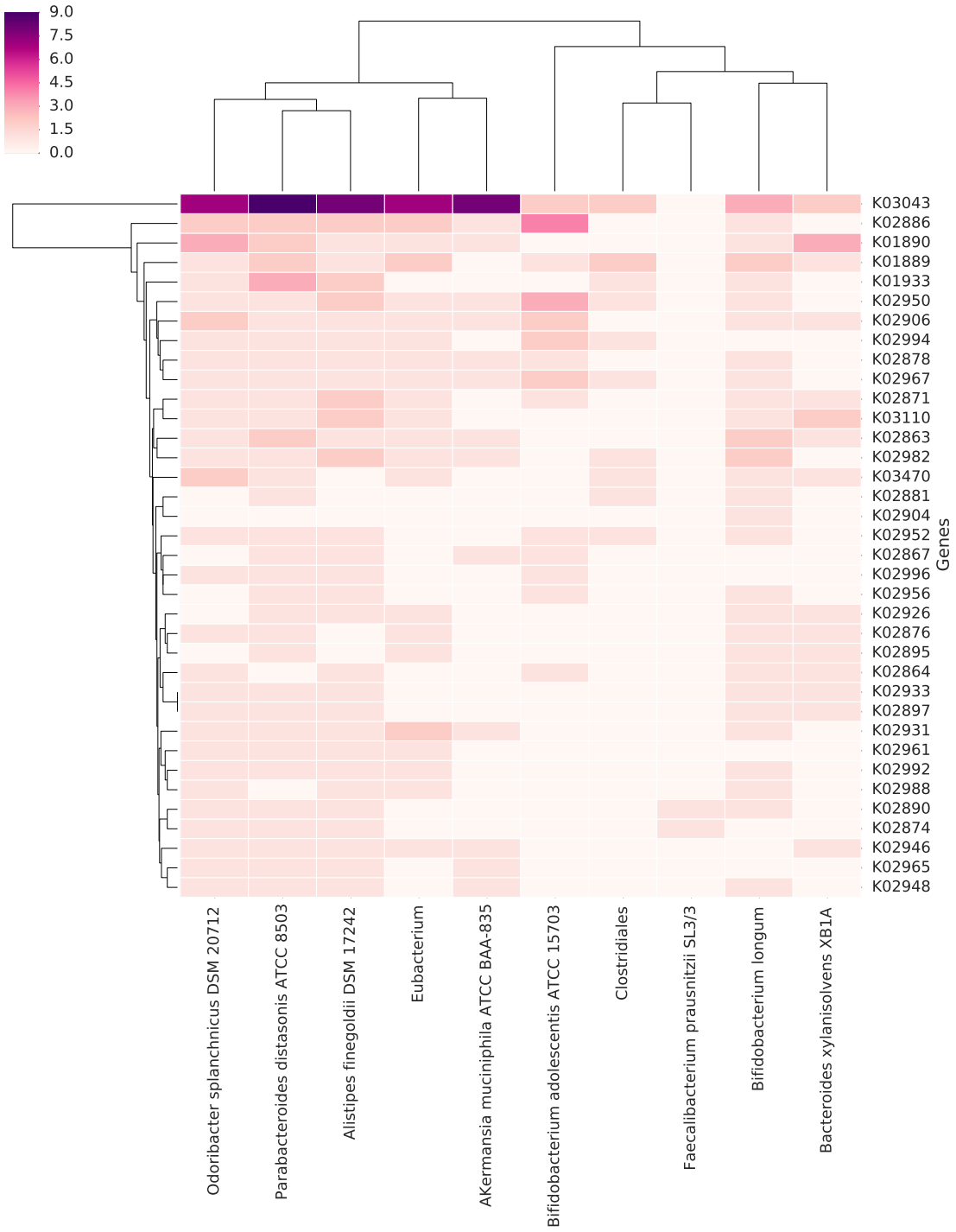


Figure 6.17: A two-way hierarchical clustering of the profile for the simulated metagenome single-copy genes. For most gene families, at least one contig per reference sequence is obtained.

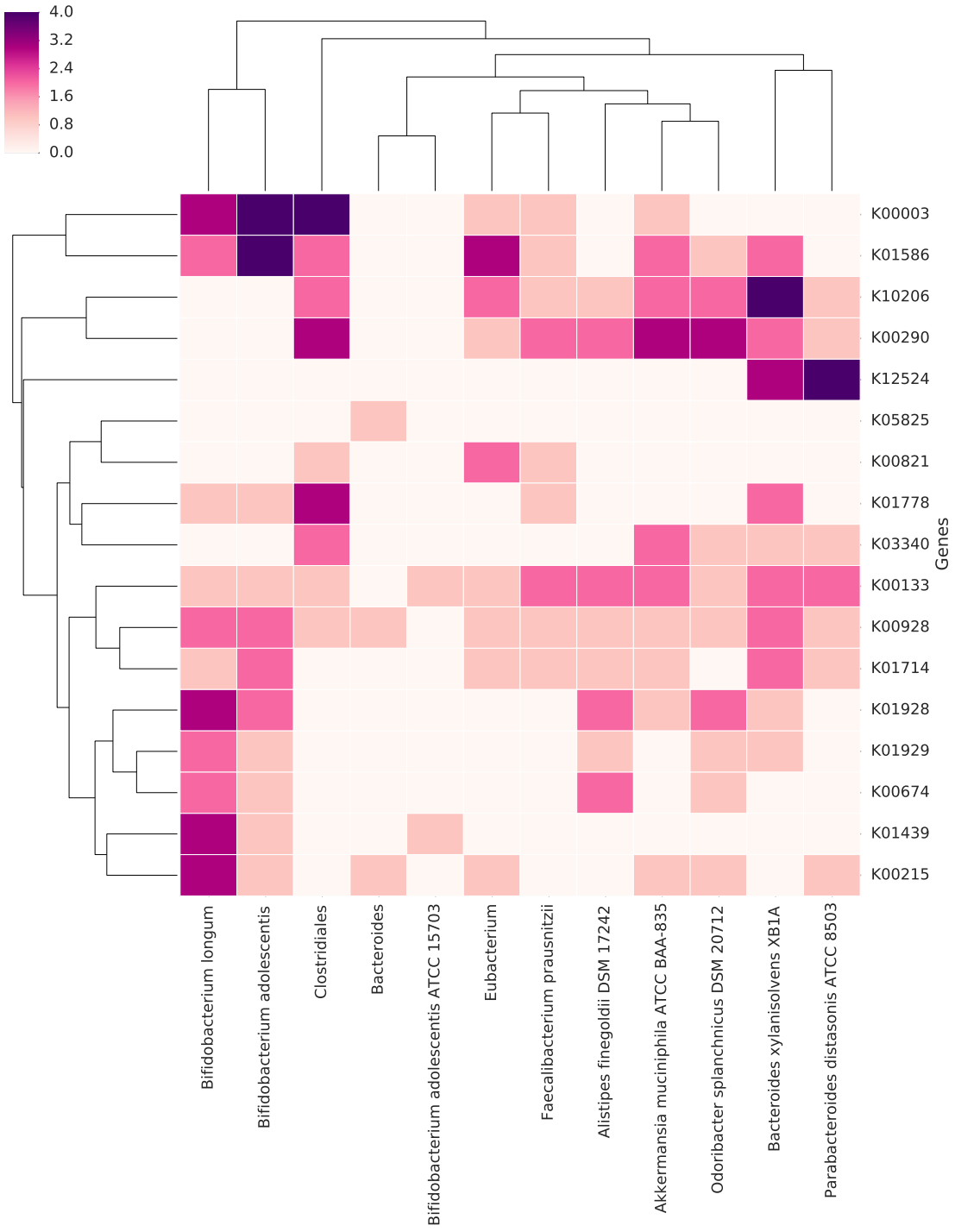


Figure 6.18: A two-way hierarchical clustering of the profile for the simulated metagenome Lysine pathway genes.

Table 6.4: The orthologous gene families and the number of reads assigned to them

Orthologous group	InterPro	#Reads	KEGG	#Reads	COG	#Reads
Ornithine carbamoyltransferase	IPR024904	30,994	K00611	25,720	COG0078	26,682
Orotate phosphoribosyltransferase	IPR023031	18,446	K00762	18,800	COG0461	14,684
Sulphite reductase beta subunit	IPR011808	10,664	K11181	9338	COG2221	23,892
Inorganic pyrophosphatase	IPR008162	11,056	K01507	8541	COG0221	7308
DNA polymerase III, delta subunit	IPR005790	18,022	K02340	20,522	COG1466	15,923

6.3.4 Gene-centric assembly of related orthologous groups from different classification systems.

Different functional classification systems like the KEGG, COG, InterPro2GO and SEED may have different number of reads assigned to related orthologous groups. In order to study the results obtained from carrying out a gene-centric assembly from the same orthologous group but from different classification systems, the following genes were chosen from the mock community and assembled. The orthologous groups are shown in Table 6.4:

The scatterplot depicts the number of contigs resulting from it (Figure 6.19).

The number of contigs produced may differ in a few cases. We therefore conclude that choice of functional classification system may affect the result of a gene-centric assembly. It may be useful to produce assemblies using different classification systems and compare the contigs produced.

6.4 Conclusions

In the case where specific genes are to be studied, a gene-centric assembly of metagenomic data could prove advantageous as shown by this study. A gene-centric assembly adds to the repertoire of the different analyses that are part of the WGS metagenome analysis pipeline. Targeted assembly is also less computationally intensive and requires comparatively less amount of time.

MEGAN's assembler does as well as and in some cases better than other assemblers, for example in the "detection" of reference organisms. SOAPdenovo and Ray tend to be similar in performance. IDBA-UD produces very few contigs, is highly specific but performs worse than MEGAN in the "detection" of the references. The number of contigs produced and the performances of all assemblers is similar on a per-gene basis, for example as shown in the case of the phenylalanyl t-RNA alpha and beta genes. This shows that gene-centric assembly, as an approach itself, is robust and useful. The advantage of MEGAN over the other assemblers is perhaps the easy-to-use GUI and less amount of computation required

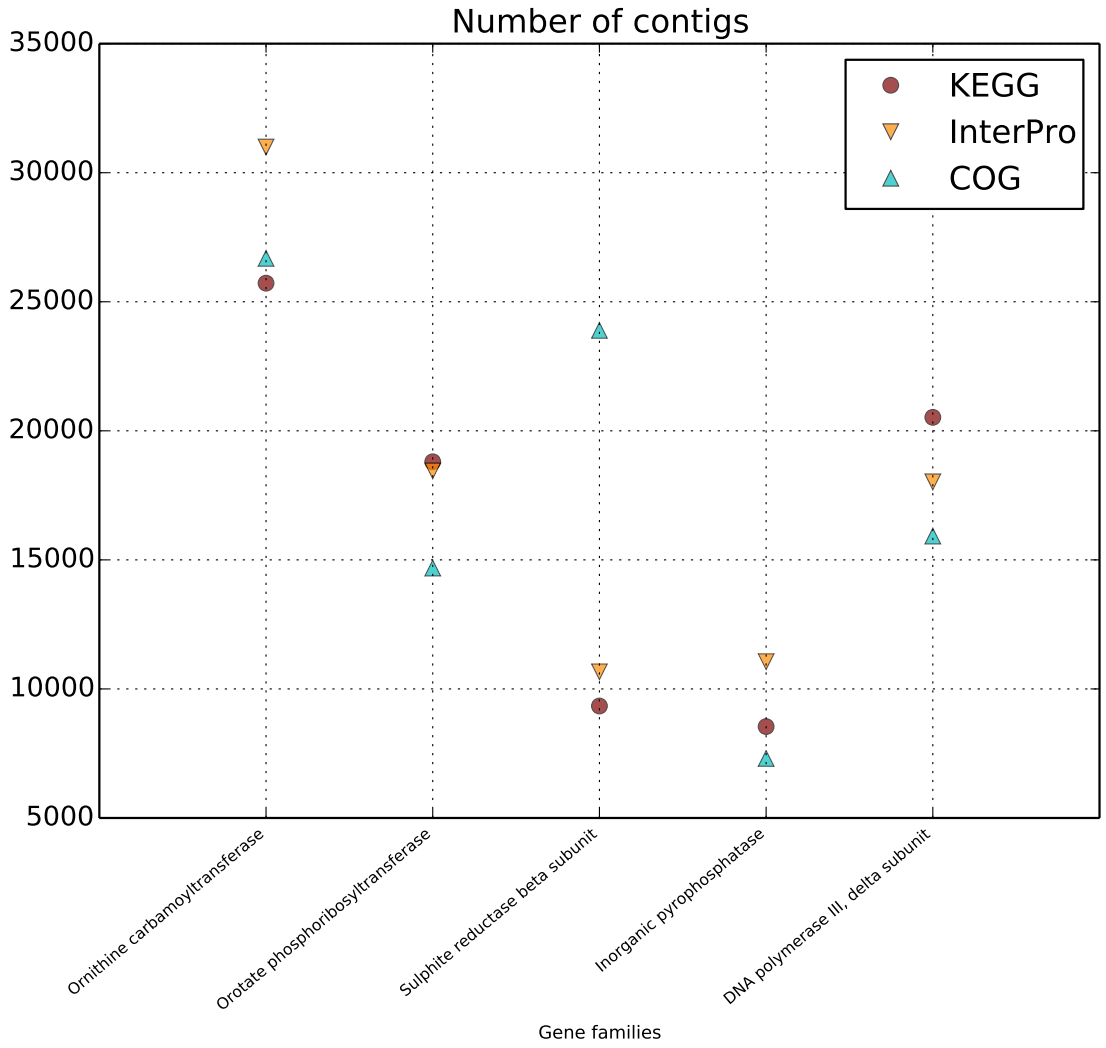


Figure 6.19: Results for different functional classification systems may differ as a result of the difference in the number of reads and therefore contigs produced after assembly.

to get the assembler running. MEGAN 6 also provides a good visualization of the alignment of the reads to the reference sequences, and enables the export of the overlap-layout graph produced for the reads belonging to the node being assembled. At the algorithmic level, even though a gene-centric assembly strategy is not novel for metagenomic data analysis, using BLASTX for the recruitment of the reads is novel.

The gene-centric assembly of the 38 genes yielded contigs, most of which could be mapped or aligned back to the organisms present in the mock community. In each case, there were some reference organisms for which no contig mapped to it. This could be due to absence of that specific gene in the genome or it could be because of there being no coverage of sequencing in that region of the genome.

The quality of the contigs produced after a gene-centric assembly depends on the specific gene that is to be studied and differs from gene to gene. The single-copy genes that are universal, are much better assembled than multiple-copy genes. In the evaluation study for example, contigs produced for most of the ribosomal protein coding genes are true positive and with a few number of false negative references. The taxonomic profile generated for a gene may not represent the complete taxonomic diversity of the sample. This could be due to the absence of that particular gene in some genomes that are part of the community. Most of the contigs detected to be false positives through the BLASTN search align to reference organisms that are closely related to the reference organisms actually present in the mock community. Appearance of false positives or false negatives is not always due to error in assembling, but due to the inherent nature of microbial genomes which are prone to horizontal gene transfer or presence of highly conserved genes. Some genes might not get assembled and usually these are the gene-families that have very few reads assigned to them. Another expected but important observation is that there is a strong correlation between the number of reads assigned and the number of contigs generated. Some genes have a wider taxonomic representation than others. The assembled contigs may vary slightly when using different classification systems, depending again on the number of reads that it has obtained. A gene-centric assembly can be incorporated in many different ways in a typical whole genome shotgun metagenomic analysis pipeline.

Chapter 7

Conclusions and Outlook

The work presented in this thesis was concentrated on the bioinformatics analysis of whole-genome shotgun metagenomic sequencing data and examining in detail a few of the several aspects involved. The results obtained could serve as a guideline when analyzing metagenomic datasets.

7.1 Potential of whole-genome shotgun metagenomic studies

Culture-independent methods have made it possible to study the otherwise unknown microbes in different environments, study their genetic makeup and to correlate this information with the physical properties of the source of the sample. WGS metagenomic data are one step ahead of amplicon sequencing data as they enable the profiling a microbial community both at a taxonomic and functional level. A species-level resolution of the taxonomic profile can be obtained with WGS sequencing as opposed to generally only a genera-level resolution with 16S ribosomal RNA (rRNA) sequencing. Discovery of novel genes, genomes and pathways is feasible with WGS sequencing. Metadata accompanying the sequencing data aids in putting forth a context for its analysis. For example, in the case of the Hohenheim Obesity Project, different clinical parameters that were measured enabled the grouping of the patients into distinct categories. Then the analysis of whether significant differences exist in the gut microbial compositions between the groups could be carried out, leading to interesting observations.

However, even though WGS provides a catalogue of the species and genes present, they do not help with understanding the mechanisms of microbial function in the community. Metagenomic studies need to be coupled with metatranscriptomics, metaproteomics and metabolomics to better understand the workings of microbial communities.

7.2 Designing a WGS metagenomic experiment

Collaborative efforts between biologists and bioinformaticians is very important for metagenomic studies. In order to glean the most information from a WGS metagenomic dataset, it is important to design the sequencing experiment optimally. The biological question(s) that are aimed to be answered with the sequencing should be formulated concretely. As a first, observations related to the community should be translated to a specific hypothesis. For example, in the case of the Hohenheim Obesity Project, the observation that the patients lose a

significant amount of weight after going through the weight-loss treatment was accompanied with a hypothesis that the microbiota undergo several changes along the diet treatment. Sequencing at different time-points help in understanding community dynamics. Technical and biological replicates are important since they enable the consolidation of observations with statistical and biological significance. Unfortunately this depends on the finances available. Good sequencing data invariably guarantees that the analysis will result in biologically important results. Detailed metadata should be collected and used appropriately.

7.3 Designing the analysis pipeline

The time available and the computational power at disposal are important factors to consider when designing the analysis pipeline for a metagenomic dataset. Specifics of the analysis pipeline, like whether the preprocessing is essential, which functional classification systems to use and whether assembly of raw reads is required, depends on the goals of the metagenomic project.

The analysis results depend heavily on the samples being analyzed. Deeply sequenced samples are prone to give better results. The microbial diversity in the sample also has an effect on the analysis results. For example, a mock metagenome with a relatively low diversity is an easy dataset to analyze as compared to a more diverse sample with a huge number of uncharacterized microbial species, like a gut sample or a soil sample.

Development of various tools for the analysis of metagenomic data is necessary. The problem of binning short reads and read-assignment is still a challenging one and novel ideas are needed to solve this problem. Metagenomic datasets coupled with metadata need good statistical and visualization techniques for determining patterns in the datasets. For example, novel ideas for the statistical analysis and visualization of time-series metagenomic data are crucial.

7.4 Gene-centric assembly as an analysis strategy

Targeted assembly of metagenomic data is a useful analysis strategy. The DIAMOND and MEGAN 6 pipeline that can be used to carry out a gene-centric assembly is fast and easy to use. The evaluation reveals that it performs as well as or better than some other assemblers. However, the quality and quantity of the contigs produced depends on the gene being assembled and the sample it is being assembled from. Mock metagenomic datasets

that have a low diversity and are deeply sequenced, produce a good number of contigs for most gene families. When used on real sequencing samples, a gene-centric assembly could be helpful for revealing novel patterns in the sample.

7.5 Future perspectives

A more comprehensive analysis of the correspondence between the different systems of functional classification could be carried out. This could be achieved by comparing at the level of pathways and analyzing more metagenomic datasets. The gene-centric assembly approach could be used for the analysis of more datasets, and studied in different contexts, for example metatranscrptiome data. Visualization and statistical approaches for analyzing time-series metagenomic data could be developed. There is a possibility for combining the ecological information with the time-series information to decipher community dynamics from the time-series metagenomic data.

Chapter 8

APPENDIX

8.1 Contributions

(a) Preprocessing

Daniel Huson (DH) put forth the idea, Rewati Tappu (RT) designed the analysis pipeline.

(b) Corresponding between the functional classification systems

Daniel Huson (DH) and Rewati Tappu (RT) put forth the idea. RT ran the analysis.

(c) Hohenheim Obesity Project

Laboratory work and experiment design was done by the Stephan C. Bischoff, Sandrine Louis and Antje Damms-Machado of the Institute of Clinical Nutrition, University of Hohenheim, Stuttgart, Germany. RT analyzed the dataset using the tools DIAMOND developed by Benjamin Buchfink (BB) and MEGAN developed by DH. Statistical analysis was carried out by RT and SL.

(d) Gene-centric assembly

DH put forth the idea and implemented the algorithm in MEGAN. The evaluation of the contigs was designed by RT, Dr. Adam Bazinet (AB), Prof. Dr. Michael Cummings, Dr. Rohan Williams and Prof. Dr. Kay Nieselt. RT tested the algorithm on different samples and gene-families.

8.2 Publications

Louis S, Tappu RM, Damms-Machado A, Huson DH, Bischoff SC. Characterization of the Gut Microbial Community of Obese Patients Following a Weight-Loss Intervention Using Whole Metagenome Shotgun Sequencing. PLoS One. 2016 Feb PMID: 26919743

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 2016 Jun PubMed PMID: 27327495

Huson DH, Tappu R, Bazinet AL, Xie C, Cummings MP, Nieselt K, Williams R. Fast and simple protein-alignment-guided assembly of orthologous gene families from microbiome sequencing reads. Microbiome. 2017 Jan 25 PMID: 28122610.

8.3 Bachelors and Masters thesis supervised

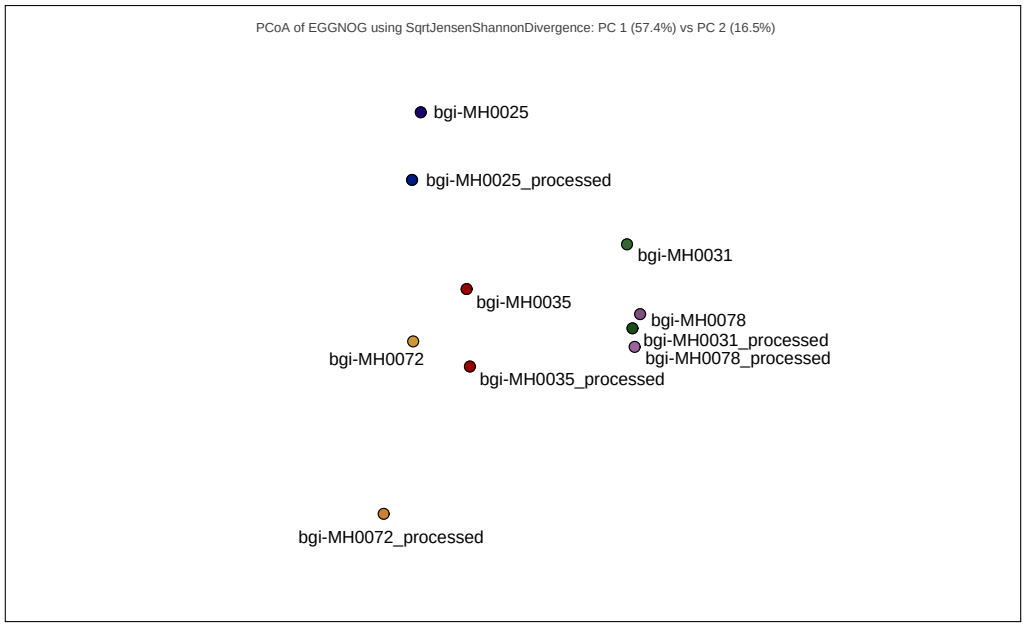
(a) Fast comparison of metagenomic samples - Marc Uwe Engelhardt

(b) Analysis of microbiome data in the context of Adipositas - Sanja Köhler

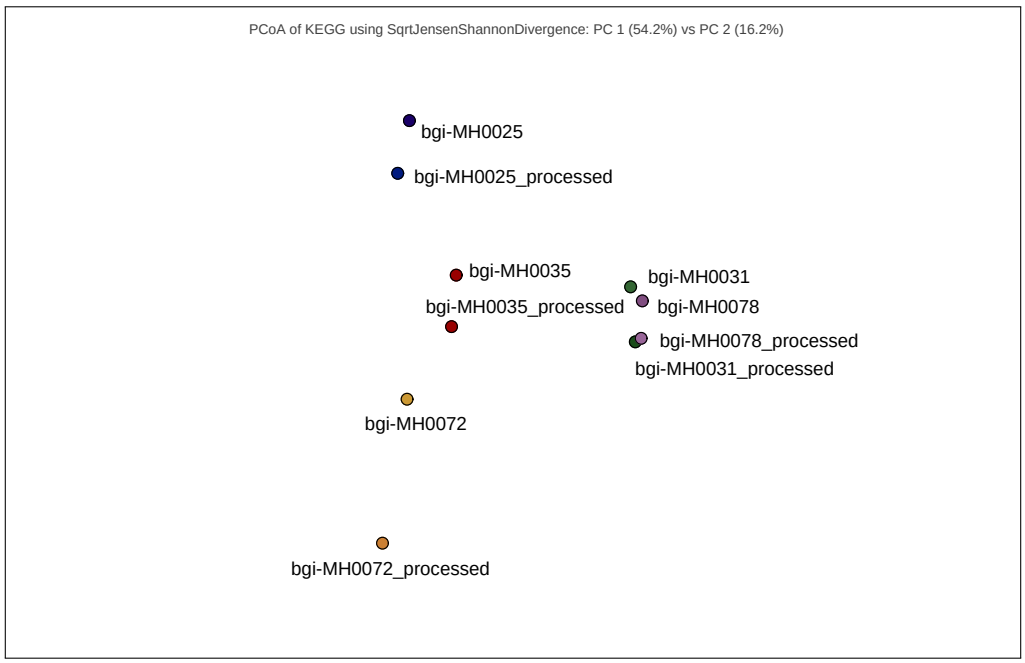
(c) Comparison of the KEGG and InterPro functional classification systems - Baiyu Lin

8.4 Supplementary materials

8.4.1 Preprocessing of metagenomic reads



(a) eggNOG

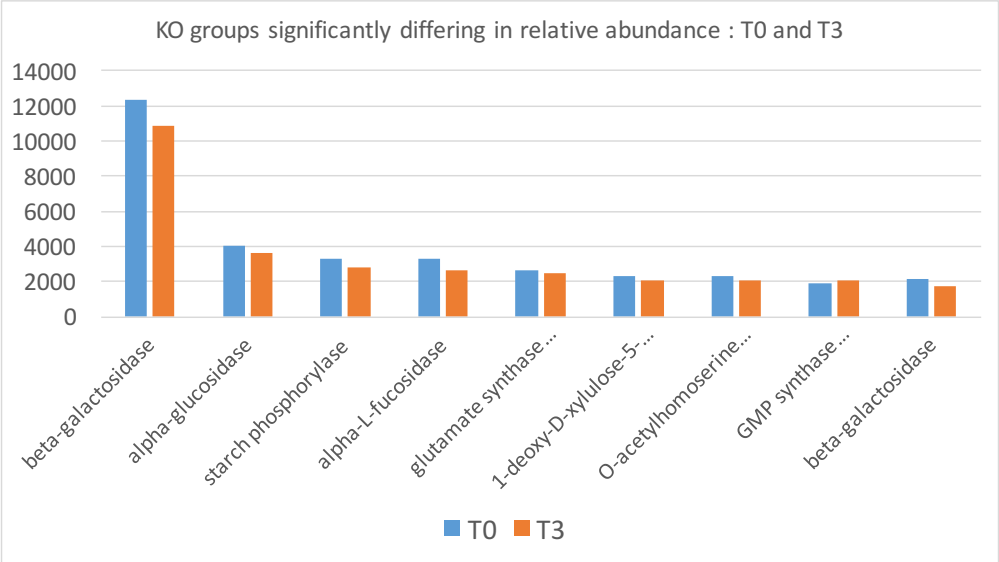


(b) KEGG

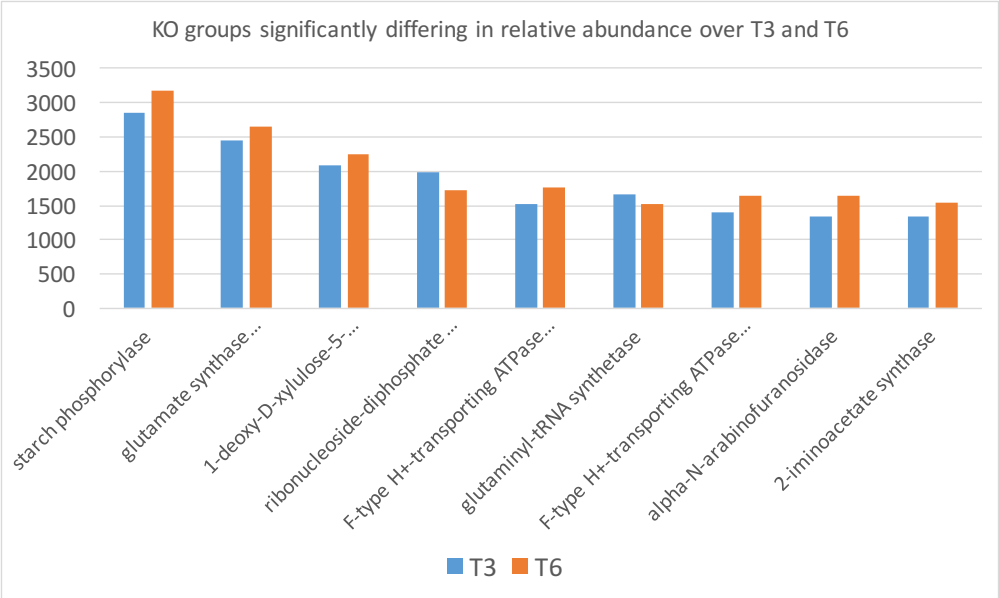
Figure 8.1: The PCoA plots of the MetaHIT samples before and after preprocessing. (a) : For the eggNOG classifications, (b): For the KEGG classifications.

8.4.2 Hohenheim Obesity Project

(a) The KO groups significantly shown to be changing in relative abundance are plot as the mean of the read-count values for the time-point.



(a) T0 to T3



(b) T3 to T6

Figure 8.2: Important KO groups that are affected with the diet-intervention are plot as their mean value over all patients for the specific time-point.

8.4.3 Gene-centric assembly

(a) The average-coverage of a given gene family for all the 64 reference-organisms is plot as a scatterplot.

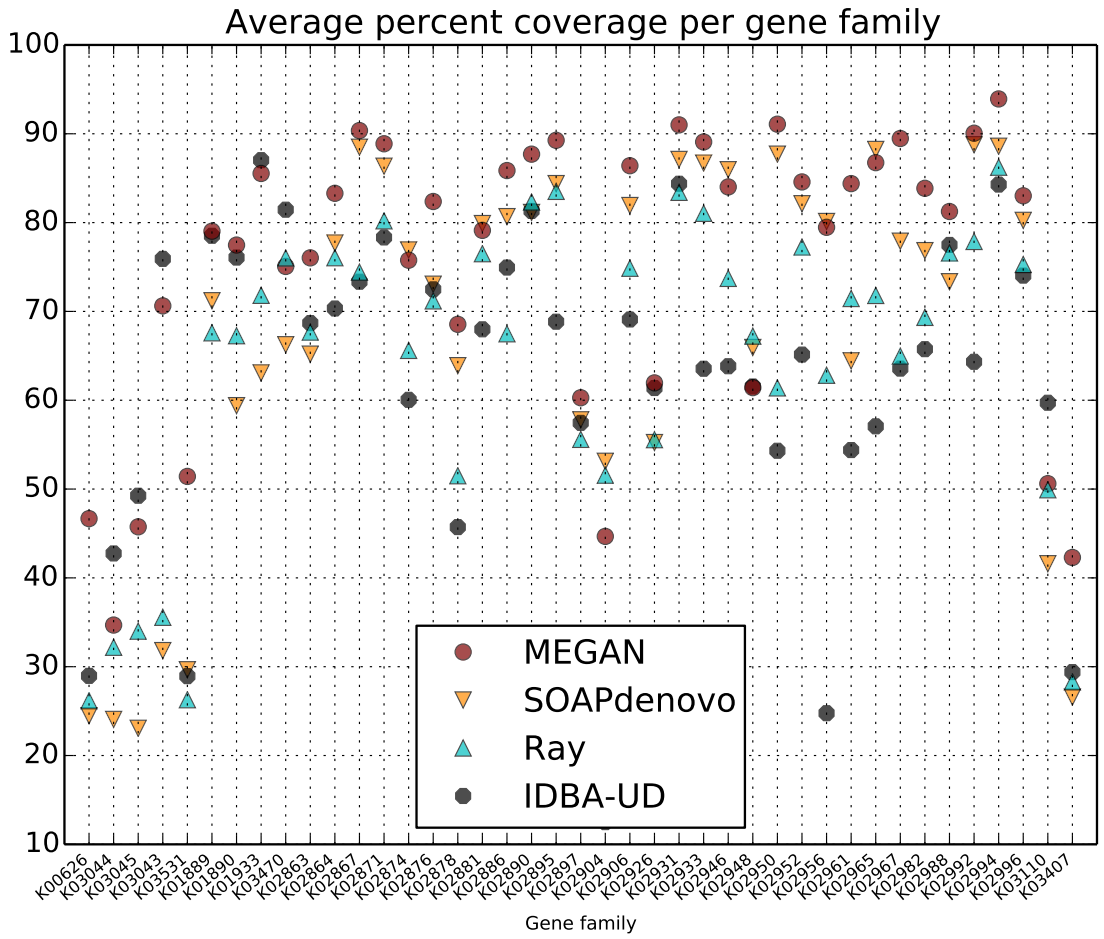


Figure 8.3: The average-coverage over 64 references for the gene families.

Bibliography

[1] John C. Wooley, Adam Godzik, and Iddo Friedberg. A primer on metagenomics. *PLoS Computational Biology*, 6(2), 2010.

[2] Claire M Fraser, Jonathan a Eisen, and Steven L Salzberg. Microbial Genome Sequencing. *Nature*, 406(6797):799–803, 2000.

[3] Aleksandar D Kostic, Michael R Howitt, and Wendy S Garrett. Exploring host microbiota interactions in animal models and humans. *Genes & Development*, 27:701–718, 2013.

[4] J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan a Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, Derrick E Fouts, Samuel Levy, Anthony H Knap, Michael W Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)*, 304(5667):66–74, 2004.

[5] Frédéric Leroy and Luc De Vuyst. Lactic acid bacteria as functional starter cultures for the food fermentation industry. *Trends in Food Science and Technology*, 15(2):67–78, 2004.

[6] David C. Demirjian, Francisco Morís-Varas, and Constance S. Cassidy. Enzymes from extremophiles. *Current Opinion in Chemical Biology*, 5(2):144–151, 2001.

[7] Michelle G. Rooks and Wendy S. Garrett. Gut microbiota, metabolites and host immunity. *Nature Reviews Immunology*, 16(6):341–352, 2016.

[8] J Parkhill, B W Wren, N R Thomson, R W Titball, M T Holden, M B Prentice, M Sebaihia, K D James, C Churcher, K L Mungall, S Baker, D Basham, S D Bentley, K Brooks, A M Cerdeno-Tarraga, T Chillingworth, A Cronin, R M Davies, P Davis,

G Dougan, T Feltwell, N Hamlin, S Holroyd, K Jagels, A V Karlyshev, S Leather, S Moule, P C Oyston, M Quail, K Rutherford, M Simmonds, J Skelton, K Stevens, S Whitehead, and B G Barrell. Genome sequence of Yersinia pestis, the causative agent of plague. *Nature*, 413(6855):523–527, 2001.

[9] Otto G Berg and C G Kurland. Evolution of microbial genomes: sequence acquisition and loss. *Molecular Biology and Evolution*, 19(12):2265–76, 2002.

[10] RK Saiki, DH Gelfand, S Stoffel, SJ Scharf, R Higuchi, GT Horn, KB Mullis, and HA Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491, 1988.

[11] RD Fleischmann, MD Adams, O White, RA Clayton, EF Kirkness, AR Kerlavage, CJ Bult, JF Tomb, BA Dougherty, JM Merrick, and et al. Whole-genome random sequencing and assembly of Haemophilus Influenzae Rd. *Science*, 269(5223):496–512, 1995.

[12] V Burland, G Plunkett, H J Sofia, D L Daniels, and F R Blattner. Analysis of the Escherichia coli genome VI: DNA sequence of the region from 92.8 through 100 minutes. *Nucleic Acids Research*, 23(12):2105–19, 1995.

[13] Alan W. Walker, Sylvia H. Duncan, Petra Louis, and Harry J. Flint. Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends in Microbiology*, 22(5):267–274, 2014.

[14] Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews : MMBR*, 72(4):557–78, Table of Contents, 2008.

[15] J. D. WATSON and F. H. CRICK. The structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, 18:123–131, 1953.

[16] F Sanger, S Nicklen, and a R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, 1977.

[17] Brendan P Hodkinson and Elizabeth A Grice. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in Wound Care*, 4(1):50–58, 2015.

[18] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

[19] Travis C. Glenn. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, 11(5):759–769, 2011.

[20] Nanopore-based fourth-generation {DNA} sequencing technology. *Genomics, Proteomics and Bioinformatics*, 13.

[21] NR Pace, DA Stahl, DL Lane, and GJ Olsen. The analysis of natural microbial populations by rrna sequences. *Advances in Microbial Ecology*, 9(5):1–55, 1986-01-01 00:00:00.0.

[22] Miguel I. Uyaguari-Diaz, Michael Chan, Bonnie L. Chaban, Matthew A. Croxen, Jan F. Finke, Janet E. Hill, Michael A. Peabody, Thea Van Rossum, Curtis A. Suttle, Fiona S. L. Brinkman, Judith Isaac-Renton, Natalie A. Prystajecky, and Patrick Tang. A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome*, 4(1):20, 2016.

[23] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire Fraser-liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804–810, 2007.

[24] Junhua Li, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, Jens Roat Kultima, Edi Prifti, Trine Nielsen, Agnieszka Sierakowska Juncker, Chaysavanh Manichanh, Bing Chen, Wenwei Zhang, Florence Levenez, Juan Jun Jian Wang, Xun Xu, Liang Xiao, Suisha Liang, Dongya Zhang, Zhaoxi Zhang, Weineng Chen, Hailong Zhao, Jumana Yousuf Al-Aama, Sherif Edris, Huanming Yang, Juan Jun Jian Wang, Torben Hansen, Henrik Bjorn Bjørn Nielsen, Soren Søren Brunak, Karsten Kristiansen, Francisco Guarner, Oluf Pedersen, Joel Doré, S Dusko Ehrlich, Peer Bork, Juan Jun Jian Wang, Joel Dore, S Dusko Ehrlich, MetaHIT Consortium, Peer Bork, and Juan Jun Jian Wang. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotech*, advance on(8):834–41, 2014.

[25] Jack A Gilbert, Janet K Jansson, and Rob Knight. The Earth Microbiome project: successes and aspirations. *BMC Biology*, 12(1):69, 2014.

[26] Natalya Yutin, Sofiya Shevchenko, Vladimir Kapitonov, Mart Krupovic, and Eugene V Koonin. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biology*, 13(1):95, 2015.

[27] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):3, 2012.

[28] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1):2010–2012, 2011.

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

[30] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, 2011.

[31] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[32] Adam L Bazinet and Michael P Cummings. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13(1):92, 2012.

[33] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–10, 1990.

[34] Wolfgang Gerlach, Sebastian Jünemann, Felix Tille, Alexander Goesmann, and Jens Stoye. Webcarma: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, 10(1):1–10, 2009.

[35] Folker Meyer, D Paarmann, M D'Souza, and Etal. The metagenomics RAST servera public resource for the automatic phylo- genetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386, 2008.

[36] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

[37] Daniel H. Huson, Sina Beier, Isabell Flade, Anna Gorska, Mohamed El-Hadidi, Suparna Mitra, Hans-Joachim Ruscheweyh, and Rewati Tappu. Megan community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol*, 12(6):1–12, 06 2016.

[38] Alice Carolyn McHardy, Héctor García Martín, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4(1):63–72, 2007.

[39] Arthur Brady and Steven L Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, 6(9):673–6, 2009.

[40] Bo Liu, Theodore Gibbons, Mohammad Ghodsi, Todd Treangen, and Mihai Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, 12 Suppl 2(Suppl 2):S4, 2011.

[41] Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–4, 2012.

[42] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Yong Yunjie Yong Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yunjie Yong Yunjie Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jun Jian Wang, Tak-Wah Lam, and Jun Jian Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, 2012.

[43] Sébastien Boisvert, François Laviolette, and Jacques Corbeil. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, 17(11):1519–1533, 2010.

[44] Yu Peng, Henry C M Leung, S. M. Yiu, and Francis Y L Chin. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, 2012.

[45] Donovan Parks and Robert Beiko. *STAMP: Statistical Analysis of Metagenomic Profiles*, pages 1–6. Springer New York, New York, NY, 2013.

[46] James Robert White, Niranjan Nagarajan, and Mihai Pop. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Computational Biology*, 5(4), 2009.

[47] Jn Paulson. metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. *Bioconductor.Jp*, pages 1–20, 2014.

[48] Jari Oksanen, Roeland Kindt, Pierre Legendre, Bob O'Hara, M Henry H Stevens, Maintainer Jari Oksanen, and MASS Suggests. The Vegan package. *Community Ecology Package*, 10, 2007.

[49] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2014.

[50] Edward J Fox, Kate S Reid-Bayliss, Mary J Emond, and Lawrence a Loeb. Accuracy of Next Generation Sequencing Platforms. *Next generation, sequencing & applications*, 1(1):1–4, 2014.

[51] Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16), 2008.

[52] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[53] Ross Overbeek, Robert Olson, Gordon D. Pusch, Gary J. Olsen, James J. Davis, Terry Disz, Robert A. Edwards, Svetlana Gerdes, Bruce Parrello, Maulik Shukla, Veronika Vonstein, Alice R. Wattam, Fangfang Xia, and Rick Stevens. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42(D1):206–214, 2014.

[54] Alex Mitchell, Hsin Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, Amaia Sangrador-Vegas, Maxim Scheremetjew, Claudia Rato, Siew Yit Yong, Alex Bateman, Marco Punta, Teresa K. Attwood, Christian J A Sigrist, Nicole Redaschi, Catherine Rivoire, Ioannis Xenarios, Daniel Kahn, Dominique Guyot, Peer Bork, Ivica Letunic, Julian Gough, Matt Oates, Daniel Haft, Hongzhan Huang, Darren A. Natale, Cathy H. Wu, Christine Orengo, Ian Sillitoe, Huaiyu Mi, Paul D. Thomas, and Robert D. Finn. The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Research*, 43(D1):D213–D221, 2015.

[55] Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian von Mering, and Peer Bork. eggnog 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44(D1):D286–D293, 2016.

[56] Migun Shakya, Christopher Quince, James H. Campbell, Zamin K. Yang, Christopher W. Schadt, and Mircea Podar. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental Microbiology*, 15(6):1882–1899, 2013.

[57] Sandrine Louis, Rewati Mukund Tappu, Antje Damms-Machado, Daniel H. Huson, and Stephan C. Bischoff. Characterization of the gut microbial community of obese patients following a weight-loss intervention using whole metagenome shotgun sequencing. *PLoS ONE*, 11(2):1–18, 2016.

[58] Ruth E. Ley, Daniel A. Peterson, and Jeffrey I. Gordon. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4):837–848, 2006.

[59] L Dethlefsen, M McFall-Ngai, and D A Relman. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, 449(7164):811–818, 2007.

[60] Simone MacCaferri, Elena Biagi, and Patrizia Brigidi. Metagenomics: Key to human gut microbiota. *Digestive Diseases*, 29(6):525–530, 2011.

[61] Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G. FitzGerald, Robert S. Fulton, Michelle G. Giglio, Kymberlie Hallsworth-Pepin, Elizabeth A. Lobos, Ramana Madupu, Vincent Magrini, John C. Martin, Makedonka Mitreva, Donna M. Muzny, Erica J. Sodergren, James Versalovic, Aye M. Wollam, Kim C. Worley, Jennifer R. Wortman, Sarah K. Young, Qiandong Zeng, Kjersti M. Aagaard, Olukemi O. Abolude, Emma Allen-Vercoe, Eric J. Alm, Lucia Alvarado, Gary L. Andersen, Scott Anderson, Elizabeth Appelbaum, Harindra M. Arachchi, Gary Armitage, Cesar A. Arze, Tulin Ayvaz, Carl C. Baker, Lisa Begg, Tsegahiwot Belachew, Veena Bhonagiri, Monika Bihan, Martin J. Blaser, Toby Bloom, Vivien Bonazzi, J. Paul Brooks, Gregory A. Buck, Christian J. Buhay, Dana A. Busam, Joseph L. Campbell, Shane R. Canon, Brandi L. Cantarel, Patrick S. G. Chain, I-Min A. Chen, Lei Chen, Shaila Chhibba,

Ken Chu, Dawn M. Ciulla, Jose C. Clemente, Sandra W. Clifton, Sean Conlan, Jonathan Crabtree, Mary A. Cutting, Noam J. Davidovics, Catherine C. Davis, Todd Z. DeSantis, Carolyn Deal, Kimberley D. Delehaunty, Floyd E. Dewhirst, Elena Deych, Yan Ding, David J. Dooling, Shannon P. Dugan, Wm Michael Dunne, A. Scott Durkin, Robert C. Edgar, Rachel L. Erlich, Candace N. Farmer, Ruth M. Farrell, Karoline Faust, Michael Feldgarden, Victor M. Felix, Sheila Fisher, Anthony A. Fodor, Larry J. Forney, Leslie Foster, Valentina Di Francesco, Jonathan Friedman, Dennis C. Friedrich, Catrina C. Fronick, Lucinda L. Fulton, Hongyu Gao, Nathalia Garcia, Georgia Giannoukos, Christina Giblin, Maria Y. Giovanni, Jonathan M. Goldberg, Johannes Goll, Antonio Gonzalez, Allison Griggs, Sharvari Gujja, Susan Kinder Haake, Brian J. Haas, Holli A. Hamilton, Emily L. Harris, Theresa A. Hepburn, Brandi Herter, Diane E. Hoffmann, Michael E. Holder, Clinton Howarth, Katherine H. Huang, Susan M. Huse, Jacques Izard, Janet K. Jansson, Huaiyang Jiang, Catherine Jordan, Vandita Joshi, James A. Katancik, Wendy A. Keitel, Scott T. Kelley, Cristyn Kells, Nicholas B. King, Dan Knights, Heidi H. Kong, Omry Koren, Sergey Koren, Karthik C. Kota, Christie L. Kovar, Nikos C. Kyrpides, Patricio S. La Rosa, Sandra L. Lee, Katherine P. Lemon, Niall Lennon, Cecil M. Lewis, Lora Lewis, Ruth E. Ley, Kelvin Li, Konstantinos Liolios, Bo Liu, Yue Liu, Chien-Chi Lo, Catherine A. Lozupone, R. Dwayne Lunsford, Tessa Madden, Anup A. Mahurkar, Peter J. Mannon, Elaine R. Mardis, Victor M. Markowitz, Konstantinos Mavromatis, Jamison M. McCorrison, Daniel McDonald, Jean McEwen, Amy L. McGuire, Pamela McInnes, Teena Mehta, Kathie A. Mihindukulasuriya, Jason R. Miller, Patrick J. Minx, Irene Newsham, Chad Nusbaum, Michelle O'Laughlin, Joshua Orvis, Ioanna Pagani, Krishna Palaniappan, Shital M. Patel, Matthew Pearson, Jane Peterson, Mircea Podar, Craig Pohl, Katherine S. Pollard, Mihai Pop, Margaret E. Priest, Lita M. Proctor, Xiang Qin, Jeroen Raes, Jacques Ravel, Jeffrey G. Reid, Mina Rho, Rosamond Rhodes, Kevin P. Riehle, Maria C. Rivera, Beltran Rodriguez-Mueller, Yu-Hui Rogers, Matthew C. Ross, Carsten Russ, Ravi K. Sanka, Pamela Sankar, J. Fah Sathirapongsasuti, Jeffery A. Schloss, Patrick D. Schloss, Thomas M. Schmidt, Matthew Scholz, Lynn Schriml, Alyxandria M. Schubert, Nicola Segata, Julia A. Segre, William D. Shannon, Richard R. Sharp, Thomas J. Sharpton, Narmada Shenoy, Nihar U. Sheth, Gina A. Simone, Indresh Singh, Christopher S. Smillie, Jack D. Sobel, Daniel D. Sommer, Paul Spicer, Granger G. Sutton, Sean M. Sykes, Diana G. Tabbaa, Mathangi Thiagarajan, Chad M. Tomlinson, Manolito Torralba, Todd J. Treangen, Rebecca M. Truty, Tatiana A. Vishnivetskaya, Jason Walker, Lu Wang, Zhengyuan Wang,

Doyle V. Ward, Wesley Warren, Mark A. Watson, Christopher Wellington, Kris A. Wetterstrand, James R. White, Katarzyna Wilczek-Boney, YuanQing Wu, Kristine M. Wylie, Todd Wylie, Chandri Yandava, Liang Ye, Yuzhen Ye, Shibu Yooseph, Bonnie P. Youmans, Lan Zhang, Yanjiao Zhou, Yiming Zhu, Laurie Zoloth, Jeremy D. Zucker, Bruce W. Birren, Richard A. Gibbs, Sarah K. Highlander, Barbara A. Methé, Karen E. Nelson, Joseph F. Petrosino, George M. Weinstock, Richard K. Wilson, and Owen White. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[62] Les Dethlefsen, Paul B. Eckburg, Elisabeth M. Bik, and David A. Relman. Assembly of the human intestinal microbiota. *Trends in Ecology and Evolution*, 21(9):517–523, 2006.

[63] James M Kinross, Ara W Darzi, and Jeremy K Nicholson. Gut microbiome-host interactions in health and disease. *Genome Medicine*, 3(3):14, 2011.

[64] T Yatsunenko, F E Rey, M J Manary, I Trehan, M G Dominguez-Bello, M Contreras, M Magris, G Hidalgo, R N Baldassano, A P Anokhin, A C Heath, B Warner, J Reeder, J Kuczynski, J G Caporaso, C A Lozupone, C Lauber, J C Clemente, D Knights, R Knight, and J I Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012.

[65] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H. Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin, Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G. Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M. de Vos, Søren Brunak, Joel Doré, María Antolín, François Artiguenave, Hervé M. Blottiere, Mathieu Almeida, Christian Brechot, Carlos Cara, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Konrad U. Foerstner, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Wolfgang Huber, Johan van Hylckama-Vlieg, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Jan Knol, Omar Lakhdari, Severine Layec, Karine Le Roux, Emmanuelle Maguin, Alexandre Mérieux, Raquel Melo Minardi, Christine M'rini, Jean Muller, Raish Oozeer, Julian Parkhill, Pierre Renault,

Maria Rescigno, Nicolas Sanchez, Shinichi Sunagawa, Antonio Torrejon, Keith Turner, Gaetana Vandemeulebrouck, Encarna Varela, Yohanan Winogradsky, Georg Zeller, Jean Weissenbach, S. Dusko Ehrlich, and Peer Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[66] C Manichanh, L Rigottier-Gois, E Bonnaud, K Gloux, E Pelletier, L Frangeul, R Nalin, C Jarrin, P Chardon, P Marteau, J Roca, and J Dore. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, 55(2):205–11, 2006.

[67] Harry J. Flint. Obesity and the Gut Microbiota. *Journal of Clinical Gastroenterology*, 45(December):S128–S132, 2011.

[68] Jiyoung Ahn, Rashmi Sinha, Zhiheng Pei, Christine Dominianni, Jing Wu, Jianxin Shi, James J. Goedert, Richard B. Hayes, and Liying Yang. Human gut microbiome and risk for colorectal cancer. *Journal of the National Cancer Institute*, 105(24):1907–1911, 2013.

[69] Isabelle Lemieux. Abdominal obesity and metabolic syndrome. *Nature*, 444(December):881–887, 2006.

[70] Ruth E Ley, Fredrik Bäckhed, Peter Turnbaugh, Catherine A Lozupone, Robin D Knight, and Jeffrey I Gordon. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11070–5, 2005.

[71] N M Delzenne and P D Cani. Interaction between obesity and the gut microbiota: relevance in nutrition. *Annu Rev Nutr*, 31:15–31, 2011.

[72] Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, Sudha B Biddinger, Rachel J Dutton, and Peter J Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–63, 2014.

[73] S C Bischoff, a Damms-Machado, C Betz, S Herpertz, T Legenbauer, T Löw, J G Wechsler, G Bischoff, a Austel, and T Ellrott. Multicenter evaluation of an interdisciplinary 52-week weight loss program for obesity with regard to body weight, comorbidities and quality of lifea prospective study. *International Journal of Obesity*, 36(4):614–624, 2012.

[74] J. S. Bajaj, P. B. Hylemon, J. M. Ridlon, D. M. Heuman, K. Daita, M. B. White, P. Monteith, N. a. Noble, M. Sikaroodi, and P. M. Gillevet. Colonic mucosal microbiome differs from stool microbiome in cirrhosis and hepatic encephalopathy and is linked to cognition and inflammation. *AJP: Gastrointestinal and Liver Physiology*, 303(6):G675–G685, 2012.

[75] Karoline Faust, Leo Lahti, Didier Gonze, Willem M de Vos, and Jeroen Raes. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology*, 25:56 – 66, 2015. Environmental microbiology Extremophiles.

[76] Adina Howe and Patrick S G Chain. Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Frontiers in Microbiology*, 6(JUL):10–13, 2015.

[77] Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 2012.

[78] Todd J. Treangen, Sergey Koren, Daniel D. Sommer, Bo Liu, Irina Astrovskaya, Brian Ondov, Aaron E. Darling, Adam M. Phillippy, and Mihai Pop. Metamos: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biology*, 14(1):1–20, 2013.

[79] Viraj Deshpande, Eric D K Fung, Son Pham, and Vineet Bafna. Cerulean: A hybrid assembly using high throughput short and long reads. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8126 LNBI:349–363, 2013.

[80] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. Art: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2012.

[81] Wenhan Zhu, Alexandre Lomsadze, and Mark Borodovsky. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research*, 38(12):e132, 2010.

[82] Thomas J Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5(June):209, 2014.

[83] Qiong Wang, Jordan A Fish, Mariah Gilman, Yanni Sun, C Titus Brown, James M Tiedje, and James R Cole. Xander : employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome*, 3:32, 2015.

[84] Yuan Zhang, Yanni Sun, and James R. Cole. A scalable and accurate targeted gene assembly tool (SAT-Assembler) for next-generation sequencing data. *PLoS Computational Biology*, 10(8):e1003737, 2014.

[85] Dongying Wu, Guillaume Jospin, and Jonathan A. Eisen. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS ONE*, 8(10), 2013.

[86] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[87] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[88] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2016-09-07].

[89] Corné H Verhees, Servé W M Kengen, Judith E Tuininga, Gerrit J Schut, Michael W W Adams, Willem M De Vos, and John Van Der Oost. The unique features of glycolytic pathways in Archaea. *The Biochemical Journal*, 375(Pt 2):231–246, 2003.