# Development of multi-locus barcodes for identification of bacterial strains and species in environmental samples using next generation sequencing technologies

**Adeola Mujidat Rotimi**

# Development of multi-locus barcodes for identification of bacterial strains and species in environmental samples using next generation sequencing technologies

by

**ADEOLA MUJIDAT ROTIMI**

Submitted in partial fufilment of the requirements for the degree

***Philosophiae Doctor***

in the

Centre for Bioinformatics and Computational Biology

Department of Biochemistry, Genetics and Microbiology

Faculty of Natural and Agricultural Sciences

University of Pretoria

October, 2018

***In the beginning was the Word, and the Word was with God, and the Word was God.***

**John 1 vs 1**

## Dedication

***This work is dedicated to GOD Almighty and my late mother (Mrs Modupe Ayodele Helen Salawu); may your gentle soul continue to rest in perfect peace***

**SUBMISSION DECLARATION**

I, Adeola Mujidat Rotimi, declare that the thesis, which I hereby submit for the degree Philosophiae Doctor in the Department of Biochemistry, Genetics and Microbiology at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE: _____

DATE: 17th October, 2018

## PLAGIARISM STATEMENT

**UNIVERSITY OF PRETORIA**

**FACULTY OF NATURAL AND AGRICULTURAL SCIENCES**

**DEPARTMENT OF BIOCHEMISTRY, GENETICS AND MICROBIOLOGY**

Full name: Adeola Mujidat Rotimi

Student number: 11371855

Title of the work: Development of multi-locus barcodes for identification of bacterial strains and species in environmental samples using next generation sequencing technologies

**Declaration**

1. I understand what plagiarism entails and I am aware of the University's policy in this regard.

2. I declare that this thesis is my original work. Where someone else's work was used (whether from a printed source, the internet or any other source) due acknowledgment was given and reference was made according to departmental requirements.

3. I did not make use of another student's previous work to submit it as my own.

4. I did not allow and will not allow anyone to copy my work with the intention of presenting it as his or her work.

Signature _____ Date: 17th October, 2018

# ACKNOWLEDGEMENTS

**Firstly:** I would like to extend my greatest gratitude to the almighty GOD, the creator of heaven and earth. My GOD, I will forever be grateful to you

**Secondly: I would like to sincerely thank the following individuals:**

**My husband:** for believing in me; thanks for all your support, prayers, patience and understanding that made my PhD journey peaceful.

**Prof Oleg N. Reva (supervisor):** for accepting me as your PhD student, your humility, endless motivation, patience, support and professional supervision in the successful completion of this research project. Prof Reva, you are a great researcher and I hope to be like you some day. May GOD bless you and yours.

**Dr Rian Pierneef:** thank you, my good teacher, for always being willing to assist me. I really do appreciatiate your kind gestures.

**Prof Fourie Joubert (Head of Bioinformatics Centre):** thank you for accepting me into the Bioinformatics laboratory and all the assistance rendered to me when needed

**Mr Johann Swart (Systems Administrator):** I appreciate all the assistance given when needed

**My extended family (Colleagues): Lebirata, Shaza, Greg, Shawn, Dillon, Bianca, Ansie and all the others** for their moral support and understanding

**My parents:** Thank you for the great value you attach to education and all the sacrifices made for your kids to be educated.

**Siblings:** Ademola, thanks for always being willing to help and to all my other siblings; you are all amazing, thanks for your support always.

**My kids**: Opeyemi and Damilola, I love you both dearly.

**Friends:** Maltina Chukwu, Iqra and Shane, for your tremendous support and encouragement.

**South African National Research Foundation (NRF) and University of Pretoria:** for the financial support received for my PhD degree.

## SUMMARY

Metagenomic approaches have revealed the complexity of environmental microbiomes and the advancement in whole genome sequencing showed a significant level of genetic heterogeneity on species level. It has become clear that a superior pattern of bioactivity of bacteria applicable in biotechnology, as well as the enhanced virulence of pathogens, often requires distinguishing between closely related species or sub-species. Current methods for binning of metagenomic reads usually do not allow identification below the genus level and very often, stop at the level of families.

In this work, an attempt was made to improve metagenome binning resolution by creating genome-specific barcodes, based on the core and accessory gene sequences. This protocol was implemented in novel software tools available for use and download from http://bargene.bi.up.ac.za/. The most abundant barcode genes from the core genomes were found to encode for ribosomal proteins, some other central metabolic genes and ABC transporters. The performance of the created metabarcode sequences was evaluated using artificially generated and publicly available metagenomic datasets. Furthermore, a program, Barcoding 2.0, was developed to align reads against barcode sequences and calculate various parameters for scoring the alignment results and individual barcodes. Taxonomic units were identified in metagenomic samples by comparison of the calculated barcode scores to set cut-off values. In the study, it was found that varying sample sizes, i.e. the number of reads in a metagenome and metabarcode lengths had no significant effect on the sensitivity and specificity of the algorithm. Receiver operating characteristics curves were calculated for different taxonomic groups based on the results of identification of the corresponding genomes in artificial metagenomic datasets and the reliability of distinguishing between species of the same genus or family by the program was close to 100%.

The results showed that the novel online tool, BarcodeGenerator (http://bargene.bi.up.ac.za/), was an efficient approach to generating barcode sequences from a set of complete genomes provided by users. Another program, Barcoder 2.0, was made available from the same resource to enable efficient and practical use of metabarcodes for visualisation of distribution of organisms of interest in environmental and clinical samples.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF ABBREVIATIONS

**AUC**- Area under curve

**BCG**- Bacilli Caimette-Guérin

**BLAST**- Basic Local Alignment Search Tool

**BWA**- Burrows-Wheeler alignment

**ddNTPs**- dideoxynucleotides

**DNA**- Deoxyribonucleic acid

**dNTPs**- deoxyribonucleotides

**EDGAR**- Efficient Database frame work for Comparative Genome Analyses using BLAST Score Ratios

**eDNA**- Environmental deoxyribonucleic acid

**HMMs**- Hidden Markov models

**LSU**- Large Subunit

**MAC**- *Mycobacterium avium* complex

**MDR-TB**- Multidrug resistant *Mycobacterium tuberculosis*

**MEGAN**- Metagenome Analyser

**MLST**- Multi-locus sequence typing-

**MRSA**- Methicillin-resistant *Staphylococcus aureus*

**MUSCLE**- Multiple Sequence Comparison by Log-Expectation

**NCBI**-National Centre for Biotechnology Information

**NGS**- Next generation sequencing

**OTU**- Operational taxonomic units

**PCR**- Polymerase chain reaction

**PHP**- Hypertext Preprocessor

**RDP**- Ribosomal Database Project

**rMLST**- Ribosomal multi-locus sequence typing

**SBL**- Sequencing by ligation

**SBS**- Sequencing by synthesis

**SEN**- Sensitivity

**SMRT**- Single-molecule real-time

**SMS**- Single-molecule sequencing

**SNPs**- Single nucleotide polymorphisms

**SPE**- Specificity

**SRST**- Short read sequence typing

**wgMLST**- Whole genome multi-locus sequence typing

# CHAPTER 1:Literature Review

***Knowledge of sequences could contribute much to our understanding of living matter. Frederick Sanger***

## 1.1 Sequencing technologies and advance in genomic studies

In 1944, Oswald Theodore proved that deoxyribonucleic acid (DNA) was a genetic material. James D. Watson and Francis Crick demonstrated in 1953 that the double helical strand structure of DNA was made up of four bases, which led to the central dogma of molecular biology (Church and Gilbert, 1984; Liu *et al*., 2012). The arrangement of nucleic acids in polynucleotide chains contains the genetic information for heritable and biochemical properties of terrestrial life (Heather and Chain, 2016). Hence, knowing the order of sequences is of vital importance in a wide range of uses such as molecular cloning, breeding, finding pathogenic genes, comparative and evolution studies (Liu *et al*., 2012; Heather and Chain, 2016).

The chain termination or dideoxy sequencing method published by Federick Sanger in 1977 became the gold standard for sequencing for the next 30 years (Sanger *et al.,* 1977; McGinn and Gut, 2013). Dideoxy terminator DNA sequencing was initiated with the use of automated gel electrophoresis (slab gel) and fluorescent terminator chemistry (capillary gel-based systems) (McGinn and Gut, 2013). The introduction of next-generation sequencing (NGS) techniques in 2005 made the effects of the chain termination method more far-reaching, with a distinct increase in the amount of sequencing data produced per instrument (Mardis, 2017). Next-generation sequencing technologies do immense parallel sequencing, which generates millions of fragments of DNA from a single sample that is sequenced in unison (Grada and Weinbrecht, 2013). The massive parallel sequencing enables high-throughput sequencing, which allows a whole genome to be sequenced in less than a day (Grada and Weinbrecht, 2013). The main aim of these sequencing technologies is to decrease the time, effort and cost of whole genome sequencing (WGS) to a level where it can be done on a routine basis for research and clinical applications. At present, there are at least four generations of sequencing technologies that can be categorised by unique features (McGinn and Gut, 2013).

### 1.1.1 First-generation sequencing: Classical sequencing

In the early 1970s Sanger and his colleague, Coulson, developed enzymatic DNA sequencing, also known as Sanger sequencing, which uses DNA polymerase (Morey *et al.*, 2013). The sequencing method known as 'plus and minus' and the sequence of bacteriophage φX174 were also published in 1975 by Sanger and Coulson (Sanger and Coulson, 1975). In 1977, the same authors presented the 'chain termination' method, which was less tedious and more efficient than the plus and minus method (Sanger *et al.*, 1977; Morey *et al.*, 2013). The chain termination method makes use of chemical analogues of the deoxyribonucleotides (dNTPs), which are known monomers of DNA strands. The dideoxynucleotides (ddNTPs) cannot form a bond with the 5' phosphate of the next dNTP because it lacks the 3' hydroxyl group needed to form a bond (Heather and Chain, 2016). Fragment ladders of sequence are created by enzymatically extending a primer hybridised to a pool of template molecules and introducing specific T, C, G or A terminations along the template (Liu *et al.*, 2012; Heather and Chain, 2016). Series of improvements were made to the chain termination method in the following years, one of which was the non-enzymatic sequencing method developed by Maxam and Gilbert. The non-enzymatic method involves selective fragmentation of the region to be sequenced for each of the nitrogenous bases and the resulting fragment is loaded on a polyacrylamide gel. Radioactive labelling and exposure to film or the introduction of fluorescent dyes into the termination reaction and fluorescent imaging are mostly used for product revelation (Maxam and Gilbert, 1977; Heather and Chain, 2016).

The major disadvantages of first-generation sequencing are: (i) low throughput, due to template preparation; (ii) high background levels, variants that are present at low frequency, such as mosaics, are often difficult to detect; and (iii) in comparison to the new sequencing technologies available, the cost per base is still high (Mardis, 2011; Morey *et al.*, 2013).

### 1.1.2 Second-generation sequencing: Next-generation sequencing

Next-generation sequencing platforms were able to overcome the limitations associated with classical sequencing. The first commercially available NGS platform was the 454 Roche GS FLX system in 2004, followed by several others, as shown in Table 1 (Mardis, 2017). For assessment of any NGS platforms, the main factors to be considered are cost per base, read and depth, read accuracy, throughput and read length (Mardis, 2017).

**Table 1.1: NGS instruments and the year they were introduced; SBS, sequencing by synthesis; SMS, single-molecule sequencing; SBL, sequencing by ligation**

| YEAR | NEXT-GENERATION SEQUENCING INSTRUMENTS |
|---|---|
| **2004** | **454 (Roche) pyro-sequencing (SBS)** |
| **2006** | **Solexa 1G (Illumina) (SBS)** |
| **2007** | **ABI SOLiD (SBL)** |
| **2008** | **Helicos Helioscope (SMS)** |
| **2010** | **Ion Torrent (PGM)** |
| **2010** | **Pacific Biosciences SMRT (SMS)** |
| **2014** | **Oxford Nanopore MinION (SMS)** |
| **2015** | **Qiagen Gene Reader (SBS)** |

The two main processes involved in all NGS platforms are template preparation and sequencing. Template preparation is further divided into three steps, namely source nucleic acid extraction, library preparation and template amplification (Metzker, 2005, Ambardar *et al;* 2016). Two basic sequencing methods established so far are sequencing by synthesis (SBS) and sequencing by hybridisation and ligation (SBL), as shown in Figure 1.1 (Ambardar *et al.,* 2016). A polymerase and a signal (a fluorophore/change in ionic concentration) detect the integration of a nucleotide into the elongating strand in the SBS methods. Pyrosequencing, sequencing by reversible termination and sequencing by detection of hydrogen ions are the sequencing chemistries that fall under SBS. In SBL methods, a probe sequence that is bound to a fluorophore hybridises to a DNA fragment, which is then ligated to an adjacent oligonucleotide for imaging. The SBL is the basis of support for the oligonucleotide ligation detection (SOLiD) sequencing technique by Applied Biosystems (Ambardar *et al;* 2016). The major disadvantage associated with the NGS methods is the short read length needed to be assembled with the aid of a bioinformatics pipeline into the original length template and polymerase chain reaction (PCR) bias introduced by clonal amplification for discovery of the base incorporation signal.

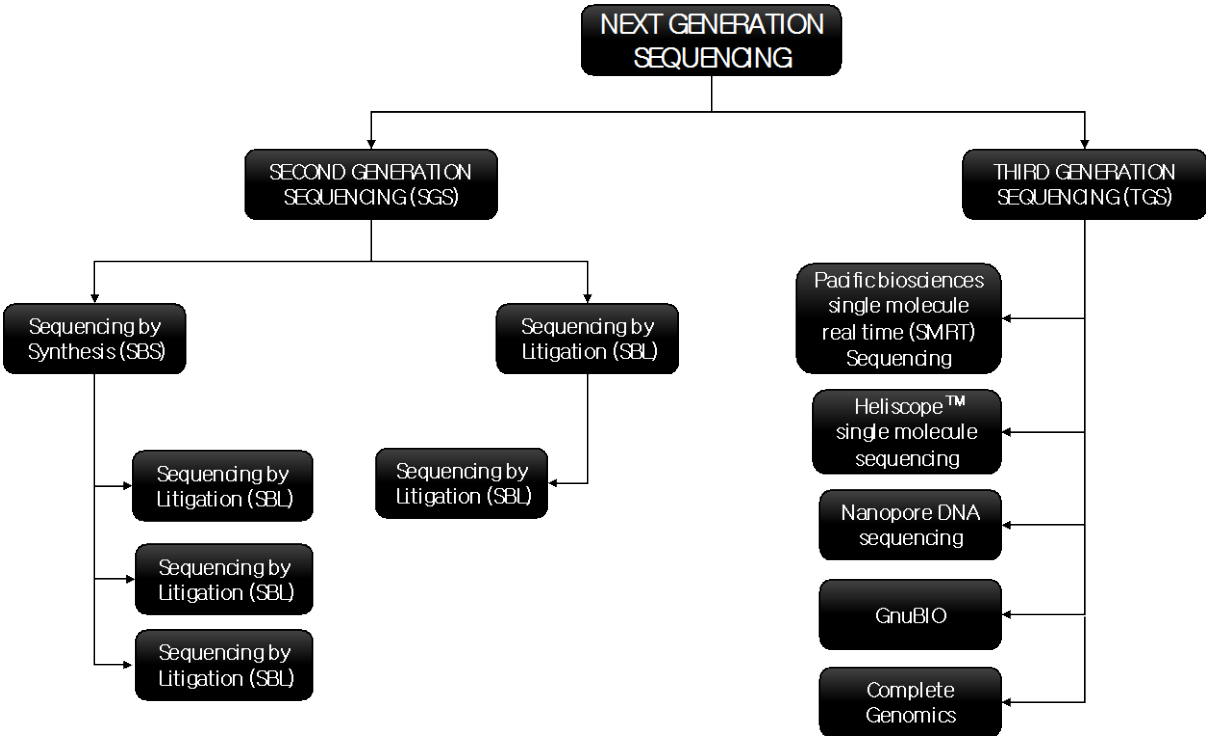



**Figure 1.1: Overview of the second- and third-generation sequencing technologies (Ambardar et al., 2016).**

### 1.1.3 Third-generation sequencing: Insight into the near future

Some researchers have argued that real-time sequencing, single-molecule sequencing (SMS) and divergence from prior technologies should define third-generation sequencing platforms (Schadt *et al.*, 2010; Niedringhaus *et al;* 2011; Pareek *et al;* 2011; Heather and Chain, 2016). Stephen Quake developed the first SMS technology, which was later commercialised by Helicos BioSciences. It works on the same principle as illumina, excluding the bridge amplification step (Braslavsky *et al.*, 2003; Harris *et al.*, 2008). The DNA template is attached to a planar surface and the property fluorescent reversible terminator dNTPs/virtual terminators are washed over one base at a time and imaged, before cleavage and cycling the next base over (Bowers *et al.*, 2009).

Pacific Biosciences developed the single-molecule real-time (SMRT) sequencing platform, which is the first third-generation sequencing technology to observe a single molecule of DNA polymerase directly as it synthesises a strand of DNA (Van Dijk *et al.*, 2014). Sequencing-by-synthesis is used by SMRT technology and it optically screens fluorescence marked nucleotides as they are merged into individual template molecules (Lee *et al.*, 2016). Read lengths of up to -100 000 bp with a throughput of > 8GB/day are produced by the PacBio RS II and the new release of PacBio can increase the throughput by as much as

seven-fold (Lee *et al.*, 2016).

The Oxford Nanopore MinIon, released in 2014, is the latest third-generation sequencing technology. It is a hand held device that sequences DNA electronically by measuring the minute disruptions to electric current as DNA molecules pass through a nanopore (Loman *et al*., 2015). The major limitation of this technology is that it is less accurate and its throughput is lower, which has limited its range to sequencing small genomes such as yeast (12 Mbp). However, by using error correction algorithms that are comparable to those available by PacBio reads, the per-nucleotide accuracy of genomes sequenced using the MinION has been measured to be >99.5 %. Because of its low cost and small size, it has been used extensively for studies in secluded settings, including the ebola outbreak in West Africa (Quick *et al*., 2016; Lee *et al.*, 2016).

### 1.1.4 Fourth-generation sequencing: *In situ* sequencing

The newly defined fourth-generation *in situ* sequencing technique makes use of second-generation NGS chemistry to read nucleic acid composition directly in fixed cells and tissues (Mignardi and Nilsson, 2014). Lee *et al.* (2014) demonstrated *in situ* sequencing of messenger ribonucleic acid (RNA) for the first time. They used a targeted method to sequence short nucleotide sequences in breast cancer tissue sections. Complementary DNA (cDNA) was first generated *in situ* and then padlock probes, which are approximately 70-base-long oligonucleotides, were used to encircle a short target sequence of four to six bases (Lee *et al.*, 2014).

Lee and colleagues further described a new technique to generate amplicons in a non-targeted approach in which random hexamers labelled with a sequencing adaptor are used to reverse transcribe RNA molecules *in situ*. The newly synthesised cDNA self-circularises and is amplified by rolling circle amplification. The amplicons are covalently linked to cellular proteins and can be produced in several different cell types, tissue sections and whole mount embryos.

Though *in situ* sequencing is still in its infancy, there is a possibility of it being used as a complementary tool to filter clinically important information from the huge amount of data produced by traditional NGS methods (Mignardi and Nilsson, 2014).

## 1.2 Metagenomics

Metagenomics can be defined as a culture-independent genomic investigation of microbial communities, which has emerged as a potent tool in the field of microbiology over the past two decades (Allan, 2014). Figure 1.2 shows a diagram of a typical metagenome project. The novelty of metagenomics lies in the fact that microbial DNA is isolated directly from the environmental sample, giving access to the whole microbial community, including the majority that has not been cultured in the laboratory. Metagenomics provides genetic information on possibly new biocatalyst, genomic linkages between function and phylogeny for uncultured organisms and evolutionary profiles of community function and structure. Metagenomics can also be supplemented with metatranscriptomic/metaproteomic techniques to define expressed activities (Wilmes and Bond, 2006; Gilbert *et al.*, 2008). Hence, metagenomics can be described as a potent tool for creating new hypotheses for microbial function such as the outstanding discoveries of proterorhodopsin-based photo-heterotrophy or ammonia-oxidising archaea (Beja *et al.*, 2000; Nicol and Schleper 2006).

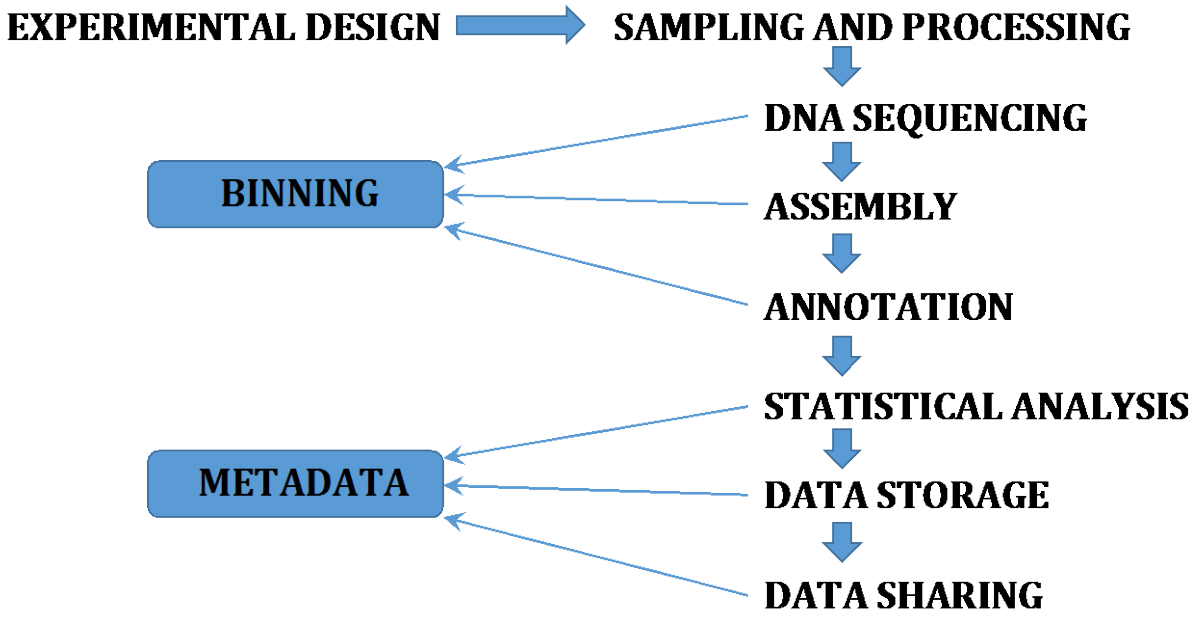



**Figure 1.1: Diagram of a typical metagenome project (Thomas et al., 2012).**

### 1.2.1. Methods and approaches of metagenomics

Two common methods are used for classification of taxonomic content of environmental samples. The first approach is sequencing of PCR amplified phylogenetic markers such as 16S ribosomal RNA (rRNA). This approach is referred to as amplicon analysis (marker gene metabarcoding). The second approach applies shotgun sequencing whereby all genomic DNA

in a community is sequenced (Handelsman, 2004; Peabody *et al.,* 2015).

Amplicon analysis or maker gene metabarcoding, also known as meta-genetics, is a rapid approach used to acquire a community diversity profile or to fingerprint a community using PCR amplification with universal primers, followed by sequencing of evolutionarily conserved genes such as the 16S rRNA gene (Tringe *et al.,* 2005; Oulas *et al.,* 2015). Amplicon sequencing has been applied in an extensive range of contexts that include among others bacterial metabarcoding, biomonitoring and community functioning analysis (Murray *et al.,* 2015). An environmental sample is collected and the total DNA is extracted from all cells in the sample. A taxonomically informative genomic maker that is common to virtually all organisms of interest is then targeted and amplified by PCR. The resulting amplicons are sequenced and bioinformatically characterised to identify which microbes exist in the sample and at what relative abundance. For bacteria and archaea, the amplicon approach commonly targets the small-subunit rRNA (16S rRNA) locus, which is both taxonomically and phylogenetically an informative marker (Pace *et al.,* 1986; Hugenholtz and Pace, 1996). Amplicon sequencing of the 16S locus has shown a remarkable quantity of microbial diversity on earth (Pace, 1997; Rappé and Giovannoni, 2003; Lozupone and Knight, 2007) and has been used to characterise the biodiversity of microbes from a variety of environments, comprising: (i) human microbiome (Human Microbiome Project Consortium, 2012a; Yatsunenko *et al.,* 2012); (ii) microbiota associated with *Arabidopsis thaliana* roots (Lundberg *et al.,* 2012); (iii) bacteria of the ocean thermal vent (McCliment *et al.,* 2006); (iv) bacterial communities of hot springs (Bowen De Leon *et al.,* 2013); (v) micro-flora of the Antarctic volcano mineral soils (Soo *et al.,* 2009) and many others. Associating 16S sequence profiles across samples explains how microbial diversity links with and scales across varieties of environmental habitats. Such observations have enabled insight into host-microbe relations and generated hypotheses about microbiota-based disease mechanisms (Turnbaugh *et al.,* 2009; Muegge *et al.,* 2011; Bulgarelli *et al.,* 2012; Smith *et al.,* 2013; Sharpton, 2014). Follow-up microbiota-manipulation research usually confirms these hypotheses (Smith *et al.,* 2013; David *et al.,* 2014; Sharpton, 2014). The most auspicious hypotheses and future planning of experiments tend to derive from the comparisons of microbiota associated with cohorts of hosts of specific genotypes or treatment conditions (Sharpton, 2014). Kuczynski and colleagues gave a thorough review on the use of 16S amplicon sequencing in a microbiota study in 2011.

The Ribosomal Database Project (RDP; http://rdp.cme msu.edu) is one of the tools

specifically created for processing high-volume amplicon sequence data (Cole *et al.,* 2013). The RDP 11.1, released in October 2013, comprises 2 809 406 aligned and annotated bacterial and archaeal small subunit rRNA gene sequences and 62 860 fungal large subunit rRNA gene sequences. The RDP tools provided by classifier and aligner pipelines have been upgraded to work with the fungal collection. Since the use of NGS platforms in characterising environmental microbial populations has increased rapidly in the past years, the sizes of environmental data sets have also increased. The RDP provides tools for browsing and searching the data collections, for taxonomic classification and nearest-neighbour (NN) search, for prime probe testing and for phylogenetic tree building. These new tools have been created with speed capability in mind. The recognised tools have been upgraded to accommodate the current changes of the sequencing technology. Many RDP tools are also made accessible as open-source stand-alone packages (Cole *et al.,* 2013).

Several limitations are associated with the amplicon sequencing approach, which include: (i) the fact that resolving a large portion of the diversity in a community is difficult, given several biases associated with PCR (Hong *et al.,* 2009; Sharpton *et al.,* 2011; Logares *et al.,* 2013; Sharpton, 2014); (ii) amplicon sequencing can produce widely variable estimates of diversity (Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012); (iii) sequencing errors and inaccurately assembled amplicons can produce artificial sequences that are usually difficult to identify (Wylie *et al.,* 2012); (iv) amplicon sequencing mostly only gives insight into the taxonomic composition of microbial community; it is usually difficult to resolve the biological functions linked with these taxa directly using this method; (Langille *et al.,* 2013); and (v) amplicon sequencing is limited to the analysis of taxa for which taxonomically informative genetic markers are known and can be amplified; new or highly diverged microbial species are difficult to study using this method (Acinas *et al.,* 2004; Sharpton, 2014).

The shotgun metagenomic sequencing technique, also known as WGS metagenomics, is an alternative method used to study uncultured microbiota that avoids many limitations of amplicon-based sequencing (Sharpton, 2014). WGS metagenomics has the ability to sequence the bulk of existing genomes within an environmental sample or community fully. This generates a community biodiversity profile that can be further linked with functional composition analysis of known and unknown organism lineages that are the genera or taxa (Tringe *et al.,* 2008; Oulas *et al.,* 2015). The DNA is extracted from cells in a community, but instead of aiming at a specific genomic locus for amplification, the entire DNA is then

clipped into tiny fragments that are autonomously sequenced. This results in DNA reads that correspond to specific genomic locations in numerous genomes available in the sample, including fungi and even multicellular organisms. Some of these reads will be sampled from taxonomically informative loci such as 16S, while others will be sampled from coding sequences that offer insight into the biological roles encoded in the genome (Sharpton, 2014). Shotgun metagenomics has advanced to address the following questions: (i) who is present in an environmental community; (ii) what those present are doing function-wise; and (iii) how these microorganisms interact to sustain a balanced ecological niche. It also offers unrestricted access to functional gene composition information derived from microbial communities residing in practical ecosystems (Oulas *et al.,* 2015). Hence, the present study aims at dealing with WGS metagenomics and the program (Barcoder software tools) created in this work is basically designed to work with WGS reads.

Notwithstanding the advantages linked with shotgun metagenomic sequencing, several limitations have been encountered. Since metagenomic data is relatively intricate and huge, this obfuscates its informatics analysis. It may be problematic to determine the genome from which a read was obtained. Moreover, most communities are so diverse that most genomes are far from being completely signified by the generated reads. Hence two reads from the same gene may not overlap and are thus difficult to compare correctly with a sequence alignment (Schloss and Handelsman, 2008; Sharpton *et al.,* 2011; Sharpton, 2014). When reads do overlap, it is not always obvious if they are from unique or repeated genomic fragments, which can challenge the sequence assembly (Mavromatis *et al.,* 2007; Mende *et al.,* 2012; Sharpton, 2014). Metagenomic analysis usually needs a huge volume of data to identify meaningful results because of the vast amount of genetic information being sampled. This need can cause computational challenges. Providentially bioinformatics software development is rapidly progressing and refining the ease and efficiency of metagenomics analysis (Sharpton, 2014).

Metagenomes can also contain unwanted DNA contaminations, for example the host DNA in samples generated from host-associated microbiota. In some scenarios, host DNA can so engulf the community DNA that complicated molecular approaches must be used to enhance the microbial DNA selectively before sequencing. Molecular and bioinformatics approaches needed to filter the host DNA from metagenomes either before or successive to sequencing of the data are being developed (Schmieder and Edwards, 2011b; Garcia-Garcerà *et al.,* 2013).

Though contamination is a general problem to all environmental sequencing studies, identification and removal of contaminants from metagenomics datasets are specifically challenging (Kunin *et al.,* 2008; Degnan and Ochman, 2012). It can become problematic to determine which reads were generated from which genome and chimeric assemblies are common. A metagenomic contaminant may reduce the coverage of genomes of interest, create chimeras and mislead the analysis of community function. Fortunately, software tools that identify and filter contaminants are made available (Schmieder and Edwards, 2011a). In 2011, Schmieder and Edwards developed DeconSeq, a rich framework for quick, automated identification and elimination of sequence contamination in longer read datasets > 150 bp mean read length. DeconSeq classifies likely contamination sequences, removes redundant hits with similarity to non-contaminants and offers graphical visualisations of the alignment results and classifications. DeconSeq allows scientists to automatically detect and proficiently remove unwanted sequence contamination from their datasets (Schmieder and Edwards, 2011). Other tools available include PRINSEQ (Schmieder and Edwards, 2011), Solexa QA (Cox *et al.,* 2010), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx-toolkit/) and FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/).

WGS metagenomics tends to be more expensive compared to amplicon-based metabarcoding, particularly when complex communities are sequenced or when the host DNA significantly outstrips the microbial DNA (Sharpton, 2014). The express reduction in the cost of sequencing has enhanced the popularity of WGS metagenomics. A rapid increase has been seen in the number of shotgun sequence datasets in the past years (Thomas *et al.,* 2012).

In recent years, WGS metagenomics has been used: (i) to identify novel viruses (Yozwiak *et al.,* 2012); (ii) to characterise the genomic diversity and function of uncultured bacteria (Wrighton *et al.,* 2012); (iii) to identify novel and industrially significant proteins (Godzik, 2011); (iv) to identify metabolic pathways controlled by gut microbiota, which were associated with human health and chronic disease development (Morgan *et al.,* 2012); and (v) to characterise plant and rhizosphere associated microbiota (Delmotte *et al.,* 2009; Bulgarelli *et al.,* 2013).

In 2016 Ranjan *et al*. performed a microbiome analysis where they compared the advantages of whole genome shotgun versus 16S amplicon sequencing. They studied the human faecal microbiome, accumulating a total of 194.1 x $10^6$ reads from a single sample by means of

multiple sequencing methods and platforms (Ranjan *et al.,* 2016). The 16S amplicon approach has been the technique used most frequently to analyse bacterial microbiomes because of providing several important advantages such as: (i) cost-effectiveness; (ii) data analysis being done by established pipelines; and (iii) the availability of a large body of archived data for reference. Nevertheless, the study by Rajan *et al*. showed substantial advantages of the WGS approach. WGS identified twice as many species as the 16S-based approach, with 32 x $10^6$ reads generated. Greater species diversity predicted by WGS was confirmed by calculating the Shannon and Simpson diversity coefficients. The WGS method also identified the presence of viruses, fungi and protozoa in the biotope that was obviously missed by the 16S rRNA approach. Another point being considered was the taxonomic resolution abilities of the 16S versus the WGS approaches (Ranjan *et al.,* 2016). Owing to the high level of conservation of 16S rRNA sequences, the RDP classifier often assigns the 16S amplicon reads maximum to the genus level and fails with the specie identifiation. In contrast, the WGS method can assertively bin the reads to the species level (Rajan *et al.,* 2016); however, this approach requires alignment of the reads against much larger reference databases.

In assessment of taxonomic diversity, marker gene analysis is one of the most computationally effective methods. This procedure involves: (i) aligning metagenomic reads to a database of taxonomically informative sequences such as genes 16S rRNA or internal transcribed spacer (ITS) regions; (ii) identifying those reads that show a reliable sequence similarity to the marker gene; and (iii) using the values of sequence similarity to perform an assignment or binning the read to the appropriate taxonomic unit (Sharpton, 2014).

One advantage of WGS sequencing is that this approach allows identification of multiple protein coding genes hosted by constituent microorganisms. Thus, the functionality of microbiomes can be predicted. Gene prediction can be done on assembled or unassembled metagenomic sequences. Three approaches by which genes are predicted in metagenomes are gene fragment recruitment, protein family classification and *de novo* gene prediction. However, because of the enormous diversity of bacterial genomes in natural environments, which significantly exceeds the capacity of available sequence databases, the majority of protein coding genes cannot be identified (Wu *et al.,* 2009). One of the most commonly used methods for detecting coding sequences in a metagenome is the use of the fragment recruitment approach to align metagenomic reads against a database of gene sequences. Metagenomic reads displaying a significant sequence similarity along their whole length to

respective gene sequences are considered representative subsequences of the gene. If the identified gene has a functional annotation, this approach to gene prediction can also concurrently offer a functional annotation to the recruited metagenomic sequences (Desai *et al.,* 2012). This approach has been useful in quantifying the genetic diversity of the gut microbiota (Qin *et al.,* 2010) and is commonly used for cataloguing specific genes present in metagenomes. This technique was designed as high-throughput likelihood-based gene identification. It relies on several read mapping algorithms that promptly evaluate to which extent a genomic fragment is similar to database sequence records. However, this comes at the cost of a possibility to return multiple various homolog sequences equally similar to a query read. Hence, this approach is not applicable to metagenomes obtained from communities comprising many unknown microorganisms, which sequences are scarce in public databases, particularly if the identification of new or highly divergent genes is an aim of the project (Sharpton, 2014).

A similar methodology involves the translation of each metagenomic read into six potential protein coding frames and matching each of the resulting peptides to a database of protein sequences by sequence alignment. Alignments can then be analysed to identify those metagenomic sequences that code translated peptides that show homology to proteins in the database. Translation tools such as transeq (Rice *et al.,* 2000) can be used to translate reads before conducting the protein sequence alignment using BLASTP or FASTA algorithms implemented in multiple online tools: USEARCH (Edgar, 2010), RAPsearch (Zhao *et al.,* 2012) and (iii) lastp (Kielbasa *et al.,* 2011). There are also implementations of these algorithms in stand-alone programs: blastx (Altschul *et al.,* 1997), USEARCH with ublast option or lastx (Kielbasa *et al.,* 2011). This gene prediction process is commonly used along with the functional annotation obtained from metadata records associated in databases with homologous protein sequences. Since this method depends on comparing metagenomic sequences to a reference database of known sequences, this approach is not appropriate for the identification of new types of protein. Only diverged homologs of known proteins can be predicted (Sharpton, 2014).

The *de novo* gene prediction approach can actually detect new genes. Gene prediction models that are capable of estimating diverse functions of microbial genes by analysing DNA sequence signals such as GC bias, codon usage, frequencies oligonucleotide and amino acids words and potential ORF length are used to evaluate the likelihood that a metagenomic read or contig contains a protein coding gene. This approach does not depend on the presence of

similar sequences in reference databases. Hence, these approaches can classify genes in a metagenome that share mutual functions with other microbial genes, but may be extremely diverged from any gene that has been revealed to date (Sharpton, 2014). Tools used for *de novo* gene prediction include: (i) MetaGene (Noguchi *et al.,* 2006), (ii) Glimmer-MG (Kelly *et al.,* 2012), (iii) MetaGeneMark (Zhu *et al.,* 2010), (iv) FragGeneScan (Rho *et al.,* 2010) and (v) Orphelia (Hoff *et al.,* 2009). Trimble *et al*. (2012) compared these methods by means of statistical simulations. Their concerts varied as a function of read properties such as length and sequencing error rate, with different approaches producing peak accuracies at diverse property thresholds, which makes it important for researchers to choose the right algorithm for their data carefully. For genome annotation, Yok and Rosen (2011) reported that gene predictions in metagenomes are improved when several approaches are used to the same data and then combined in a consensus approach (Yok and Rosen, 2011). Though these approaches need more time and means to envisage genes, they are usually more discerning than six-frame translation and may reduce the time consumed by the functional annotation of sequences, as fewer pairwise sequence comparisons may be needed (Trimble *et al.,* 2012). In case scenarios where the predicted gene is new relative to the database sequences, it can be challenging to decide whether the gene is an actual one or a false prediction (Sharpton, 2014).

### 1.2.2 NGS sequencing technology

Sample processing is the major and most important phase in any metagenomic research. Hence, DNA extracted should represent all cells present in the sample and an adequate quantity of high-quality DNA must be acquired for consequent library production and sequencing. Exact procedures are required for each sample type and different robust methods of DNA extraction are presented by different researchers (Venter *et al.,* 2004; Bruke *et al.,* 2009; Delmont *et al.,* 2011). Attempts have also been made to discover microbial diversity from different ecosystems using a single DNA extraction technology to ensure compatibility and a high level of precision.

Different sequencing technologies are now available, though Sanger sequencing is still considered the gold standard for sequencing, because of its low error rate, long read length (> 1,500 bp) and large insert sizes. These features will help improve assembly outcomes for shotgun data, which makes Sanger sequencing still appropriate for generating close to complete genomes in low-diversity environments (Goltsman *et al.,* 2009). Of all the NGS technologies, the Roche 454, Illumina and Ion Torrent systems have been used extensively in

metagenomic samples (Mardis, 2008; Metzker, 2010); however, the PacBio technology may replace them in the near future. Since the Roche 454 and Illumina technologies are mostly used in metagenomic research, it is of importance to describe their advantages and limitations in sequencing of metagenomics samples briefly (Oulas *et al.*, 2015)

The chemistry of the 454 pyrosequencer relies on immobilisation of DNA fragments on DNA-capture beads in a water-oil emulsion and then using PCR to amplify the fixed fragments. The beads are placed on a PicoTiterPlate. DNA polymerase is also packed in the plate and pyrosequencing takes place (Ronaghi *et al.*, 1998; Ronaghi, 2001). While Roche 454 pyrosequencing technology is considered highly reliable, it is associated with generation of several types of artefacts, which may affect the metagenomic data analysis and lead to biased results (Rosen *et al.*, 2012). One problem consists in generation of artificial replicates of the same read that may cause an overestimation of species abundance or functional gene abundance in a sample. Amplification errors in the form of single base pair mismatches and improper sequencing of mononucleotide stretches of DNA may cause frame shifts in protein-coding genes (Rothberg and Leamon, 2008). Chimera sequences generated by an undesired end joining of two or more true sequences can also affect the results of metabarcoding based on amplified 16S rRNA with respect to the species richness (Bordin *et al.*, 2013). The 454 pyrosequencing technology produces reads of up to 1 000 bp in length and >1 000 000 reads per run. The comparatively long reads length produced by this technology compared to other NGS technologies makes it more suitable for assembly genomes from shotgun metagenomic datasets and allows for better annotation accuracy (Wommack *et al.*, 2008; Thomas *et al.*, 2012)

Illumina dye sequencing by synthesis starts with the attachment of DNA molecules to primers on a glass slide, followed by amplification to generate local colonies of identical DNA fragments (Mardis, 2008). The production of DNA clusters is accompanied by an addition of fluorescently labelled adenine, cytosine, guanine and thymine terminator bases attached with a blocking group (Bentley *et al.*, 2008). These bases then compete for binding sites on the template DNA to be sequenced and unbound molecules are washed away. A laser is used to excite the dye after each synthesis cycle and a high-tenacity scan of the merged base is done. A chemical deblocking phase enables the removal of the 3' terminal blocking group together with the dye in a single step that generates a colour light impulse, which is recorded by the system. This procedure is repeated till the full DNA molecule is sequenced. Diverse Illumina sequencing instruments are dedicated to various uses. The

Hiseq2500 has large output of 1 000 GB per run but gives 125 bp reads. The MiSeq has an output of 15 GB and 25 million sequencing reads of 300 bp in length, of which clustered paired-end fragments can be sequenced from both ends, which can be combined so that 600 bp reads can be attained (Bantely *et al.*, 2008; Kircher *et al;* 2012; Oulas *et al;* 2015). Shorter read lengths generated by Illumina increase the chances of errors during assembly and then the annotation inaccuracies during shotgun metagenomics data analysis (Kircher *et al.*, 2012), while analysing 16S rRNA metabarcodes by Illumina obviates the need for time-consuming and inaccurate removal artifacts generated by Roche 454 pyrosequencing and makes this analysis less error-prone (Werner *et al.*, 2011). The greater coverage provided by Illumina enables a substantial reduction of systematic errors. This benefit and the low cost of sequencing are the defining reasons that have made Illumina the preferred NGS for metagenomics studies (Oulas *et al.*, 2015).

PacBio offers longer read lengths of -10 000 bp compared to other sequencing technologies, therefore having the advantage of addressing issues of annotation and assembly for shotgun metagenomics (Metzker, 2010). The PacBio platform uses a process termed "storbing" to perform pair-end read sequencing. Notwithstanding the high read length of PacBio, this technology is limited by higher error rates and low coverage (Metzker, 2010; Oulas *et al.*, 2015). Ion Torrent provides higher quality than 454, particularly when sequencing homopolymers, but at a similar cost of about US$23 per Mb for the Ion Torrent PGM -314 chip. However, given that 454 will eventually stop being supported by life sciences, it is most likely that the former users of 454 pyrosequencing will switch to Ion Torrent sequencing chemistry in view of their similarities, such as the emulsion PCR step (Oulas *et al.*, 2015).

### 1.2.3 Assembly

The assembly procedure joins collinear metagenomic reads from the same genome into a single contiguous sequence and is suitable for creating longer sequences, which can simplify further bioinformatics analysis and genome comparison. Sometimes, complete or nearly completed genomes of non-cultured microorganisms can be assembled from metagenomic sets of reads (Iverson *et al.*, 2012; Wrighton *et al.*, 2012; Ruby *et al.*, 2013; Sharpton, 2014).

Two approaches are employed for assembly of metagenomic reads: reference-based assembly (co-assembly) and *de novo* assembly. Software packages used for reference-based assembly are MIRA or AMOS (http://sourceforge.net/projects/amos/) and Newbler (Roche) (Chevreux

*et al.,* 1999). The algorithms of these software packages are fast, memory-efficient and can be executed even on laptop computers in a few hours. Reference-based assembly performs better, if the metagenomic dataset sequences are closely related to one or several available reference genomes (Chevreux *et al.,* 1999).

The *de novo* assembly usually needs stronger computational resources. Hence, a whole class of assembly tools based on the de Bruijn graphs were specially created to handle large amounts of data (Pevzner *et al.,* 2001; Miller *et al.,* 2010). The machine requirements for the de Bruijn assemblers, such as Velvet and Soap, are still considerably higher than those for the reference-based assembly. Usually they require hundreds of gigabytes of memory in a single machine or a computer cluster, and the run time often takes days (Li *et al.,* 2009; Zerbino and Birney, 2008).

### 1.2.4. Binning and binning algorithms

Binning is defined as a process of assigning DNA sequences to taxon-specific groups, which may represent an individual genome or genomes of several closely related organisms (Thomas *et al.,* 2012). Usually, each sequence is either categorised into a taxonomic group (i.e operational taxonomic unit (OTU), genus, family) through comparison to some referential data, or clustered into groups of sequences that signify taxonomic groups based on shared characteristics such as sequence or nucleotide composition similarity. Binning plays a key role in metagenomics by (i) providing insight into the presence of groups of unknown organisms, which are difficult to separate and identify; (ii) providing insight into the distinct numbers and types of taxa in the community; and (iii) providing methods of reducing the complexity of data such as post-binning analyses, for example by read assembly, which can be performed independently on each set of the binned reads rather than on the whole population of data (Sharpton, 2014).

Different binning algorithms have been designed, which make use of different types of information in a given DNA sequence (Thomas *et al.,* 2012). The analysis of datasets obtained by shotgun sequencing includes characterisation of the taxonomic and functional diversity of specified environmental micro-flora by analysing DNA fragments originating from genomes of the inhabitant microbes (Mande *et al.,* 2012). Binning techniques available for these types of analyses can be grouped into two categories: taxonomy-dependent and taxonomy-independent (Mande *et al.,* 2012).

### 1.2.5 Taxonomy-dependent methods

Most of the methods for binning datasets from shotgun sequencing belong to the taxonomy-dependent category, which includes: (i) alignment-based methods, (ii) composition-based methods and (iii) hybrid-based methods (Mande *et al.,* 2012).

#### 1.2.5.1 Alignment-based methods

Most alignment-based methods work by aligning reads to sequences followed by some statistical procedures, such as hidden Markov models (HMMs) for example, which assign the sequences to known taxonomic groups. Algorithms such as BLAST, BLAT or read mapping approaches such as Burrows-Wheeler alignment (BWA) or BOWTIE are usually first used to align individual reads to nucleotide or protein sequences belonging to known genomes in the alignment-based methods (Altschul *et al.,* 1990; Kent *et al;* 2002; Langmead *et al.,* 2009; Li and Durbin, 2010; Mande *et al.,* 2012). Collections of reference sequences are usually obtained from public repositories such as Ensembl (http://www.ensembl.org/), DDBJ (http://www.ddbj.nig.ac.jp/), National Centre for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/blast/db/), PFAM (http://pfam.sanger.ac.uk), Uniprot (http://www.uniprot.org/), EMBL (http://www.ebi.ac.uk/embl/), NCBI Genbank (http://www.ncbi.ncbi.nlm.nih.gov/genbank/) and NCBI Refseq (http://www.ncbi.nlm.nih.gov/RefSeq). Analyses of the quality of alignment of the searched sequence against various reference sequences determine taxonomic groups to which the read can be assigned. The MG-RAST server and the CAMERA pipeline make use of this method, where reads are assigned to taxa of microorganisms corresponding to their respective best BLAST hits (Seshadri *et al.,* 2007; Meyer *et al.,* 2008). The major disadvantage of BLAST-based methods is the need for large computing power to align millions of reads against a large number of sequences constituting a reference database (Mande *et al.,* 2012). Since a large fraction of reads from datasets attained from the shotgun sequencing method originated from unidentified taxa, which might be of novel specie/genus/family, this fraction of sequences cannot be assigned to any taxonomic unit. Hence, the MG-RAST server and metagenome analyser (MEGAN) provide an option based on the lowest common ancestor (LCA) to infer taxonomic affiliation at the lowest possible level according to the sequence similarity of the top best hits (Huson *et al.,* 2007; Meyer *et al.,* 2008).

Huson *et al.* (2007) introduced a novel computational software tool called MEGAN, which is used for analysis of large datasets on personal computer (PC) machines. Basically, this

program was designed for a visualisation of local BLAST outputs, as first of all the set of metagenomic DNA reads should be aligned using BLAST against a local database of reference sequences. The MEGAN software then allows the taxonomic content of the datasets to be explored, using the NCBI taxonomy to summarise and order the result (Huson *et al.,* 2007). The MEGAN algorithm assigns each read to the LCA knob in the taxonomic tree that lies above all the species for which the reads have obtained significant alignment hit values. A rationale for doing this is that the reads that match widely conserved genes with similar hit values should be allocated to high-level taxa unifying all these equally matching groups of organisms. Reads that hit to a specific gene of a particular microorganism are allotted to lesser taxa. The number of reads confirming the presence of a specific taxonomic unit in the sample is also controlled. Hence, the reads can be binned by this approach across all taxonomic levels. The naïve LCA algorithm provides a rapid method for taxonomic binning, which runs at a rate of over 100 million reads and 2 billion alignments per hour on a PC machine (Huson *et al.,* 2007; Huson *et al.,* 2016). In 2016, a new release of the MEGAN software tool was produced and termed 'MEGAN Community Edition (CE)'. The MEGAN Community Edition (CE) allows interactive analysis and comparison of data, making it possible to explore hundreds of samples and billions of reads. All source code for MEGAN CE is made available at https://gitup.com/danielhuson/megan-ce. MEGAN CE also makes use of the naive LCA algorithm for taxonomic binning by default (Huson *et al.,* 2016).

MEGAN makes use of bit scores of individual BLAST hit as the main parameter for judging hit significance (Mande *et al.,* 2012). Research has shown that the single-parameter method adversely affects the accuracy of taxonomic assignments in diverse situations, especially taking into account that BLAST hits reflect the accuracy of local alignments (Monzoorul *et al.,* 2009; Ghosh *et al.,* 2010). Methods such as the Sort-ITEMS, DiScRIBinATE, ProViDE MetaPhyler and MARTA have been able to solve this limitation by using, apart from BLAST bit scores, several pre-computed thresholds for other alignment parameters such as percentages of identities of global sequence alignments, numbers of positives and gaps to determine the quality of alignments (Monzoorul *et al.,* 2009; Ghosh *et al.,* 2010; Horton *et al.,* 2010; Gosh *et al.,* 2011; Liu *et al.,* 2010).

The CARMA and AMPHORA tools also make use of HMM-based binning methods (Krause *et al.,* 2008; Wu and Eisen, 2008). Reads are first compared using BLASTx against protein sequences in the PFAM database with CARMA. The program subsequently creates a phylogenetic tree by comparing read alignments to different proteins using HMM-based

statistical parameters. A taxonomic dendrogram is eventually inferred for the analysed reads (Krause *et al.,* 2008). For the AMPHORA, reads are first compared using an HMM algorithm against reference sequences representing 31 phylogenetic marker gene families. Subsequently, a phylogenetic tree is created embodying all the reads and the sequences belonging to the best scoring HMM hits. Taxonomic assignments are then attained in a mode similar to CARMA (Wu *et al.,* 2008). Other approaches that make use of HMMs/reference trees for the assignment procedures include: (i) ML TreeMap; (ii) Treephyler; (iii) pplacer; and (iv) papara (Masten *et al.,* 2010; Schreiber *et al.,* 2010; Stark *et al.,* 2010; Berger and Stamatakis, 2011). These approaches also make use of either the Bayesian or maximum-likehood algorithms to compute confidence scores. Treephyler uses the PFAM database in its workflow, while MLTreeMap compares query sequences using the HMM approach against protein sequences of 40 marker gene families (Schreiber *et al.,* 2010; Stark *et al.,* 2010). The pplacer and papara procedures offer a comprehensive scope of algorithms, which can be used for placing reads into the best scoring insertion edge on a user-specified reference phylogenetic tree (Masten *et al.,* 2010; Berger *et al.,* 2011).

#### 1.2.5.2 Composition-based methods

Compositional binning makes use of the fact that the genomes have conserved nucleotide composition such as a certain GC or abundance distribution of k-mers, which will be reflected in the sequence fragments of the genomes, codon usage and oligonucleotide usage patterns for comparing reads to sequences or models present in reference databases (Mande *et al.,* 2012; Thomas *et al.,* 2012). Composition-based methods differ in the way they characterise, measure and compare compositional properties. Most tools involve an initial preparation phase during which one or more compositional properties of known genomes are used for creating genome-specific reference models or classifiers (Mande *et al.,* 2012). The Phylopythia and NBC classifiers create genome- or clade-specific classifiers with support vector machines (SVMs) and naïve-Bayesian methods to capture and represent oligonucleotide usage forms seen in known taxonomic clades (McHardy *et al.,* 2007; Rosen *et al.,* 2010).

TACOA first creates genome-specific models by analysing tetra- and penta-nucleotide usage forms. A kernelised-NN (k-NN) method is then used to decipher taxonomic assignments of individual reads (Diaz *et al.,* 2009). The Phymm tool represents oligonucleotide usage forms of reference genomes as interpolated Markov models. Reads are usually scored against these

models and a Bayesian method is thereafter used to draw inferences (Brady *et al.*, 2009).

The ClaMS tool creates training models using de Bruijin graphs and Markovian chain algorithms. During the classification stage, a similar procedure is used for creation and comparison of signatures of query reads against pre-computed signatures of the training sequences (Pati *et al.*, 2011). In 2011, Nalbantoglu *et al.* developed a semi-supervised approach, which performs read binning by creating an index referred to as the relative abundance index, which shows the under-abundance patterns of k-mers in sequences belonging to different known taxonomic clades. The index is then used as a measure to associate a given taxon to a query sequence (Nalbantoglu *et al.*, 2011).

All the approaches described above adopt the idea that each genome can be represented by a single DNA compositional model of genome-signature k-mer frequencies. Some authors criticise this assumption by pointing out a significant level of compositional heterogeneity in genomic loci of several organisms (Cole *et al.*, 1998). From these observations, it was concluded that the representation of each genome by one single composition model may be inappropriate (Mohammed *et al.*, 2011). The INDUS algorithm disregards this hypothesis and characterises each genome in the form of multiple vectors. Each vector captures the form of tetranucleotide frequencies of individual 1-kb segments created by dicing the particular genome. In the assignment process, INDUS makes use of compositional distances between the query read and the closest known set of reference segments for determining an appropriate taxonomic level of assignment for the query. Hence, the final assignment is made to a consensus taxon that matches the next reference segments at or above the known taxonomic level (Mohammed *et al.*, 2011).

#### 1.2.5.3 Hybrid methods

Hybrid binning approaches make use of both alignment- and composition-based strategies for taxonomic classification. SPHINX and PhymmBL are two examples of the hybrid approach (Brady *et al.*, 2009; Mohammed *et al.*, 2011). SPHINX makes use of a two-phase binning algorithm. In the first stage, it compares the composition of a specified read with those of reference sequences. Hence, it speedily finds a subset of clusters of reference sequences that are next in composition to a given read. In the second stage, the taxonomic classification of the query read is inferred by first aligning the query read to reference sequences in the closet cluster and then engaging a similarity-based method such as Sort-ITEMS (Mohammed *et al.*, 2011). PhymmBL combines the composition-based method of Phymm with the alignment-

based stage that involves BLAST to enhance the confidence of taxonomic assignments (Brady *et al.,* 2009).

Taxonomy-dependent binning approaches are usually used to classify sequences constituting metagenomic datasets. However, lengths of metagenomic read, which are mostly dependent on the used sequencing techniques, are observed to be the main factor that determines selection of the binning approach. The alignment-based and composition-based approaches are most suitable for relatively longer lengths, though the composition-based approach is of more benefit in respect of the speed of execution and low memory requirement (Mande *et al.,* 2012). Lykidis *et al.* (2011) performed a taxonomic characterisation of the terephthalate (TA) wastewater metagenome using a compositional approach (Phylopythia), given that the input was of adequate length (Lykidis *et al.,* 2011). The assembled sequence data contained 37 818 and 14 526 contiguous fragments of intermediate length, the largest fragment being approximately 240 kb and 45 fragments between 24 and 167 kb. Phylopthia helped in identifying specific microbial species that played an important role not only in the degradation of TA but also in maintaining the stability of this distinctive microbial community. However, for lower length sequences providing very weak compositional signals, alignment-based approaches yielded much better performance, as seen in the studies conducted by Gupta *et al.* (2011) and Belda-Ferre *et al.* (2012). Sequences with a length of 200-400 bp in both malnourished gut metagenome (Gupta *et al.,* 2011) and oral metagemones (Belda-Ferre *et al.,* 2012) respectively, the hybrid (SPHINX and PhymmBL algorithms) and alignment-based (MEGAN) binning approaches resulted in the best binning outputs.

For most ultra-short sequences, a pre-assembly phase is necessary before performing taxonomic binning. In a comparative study conducted by Qin *et al.* (2010) on reads from human gut metagenomes, reads of length of < 75 bp generated by the Illumina sequencing technique were first assembled into contigs and then classified using MEGAN.

### 1.2.6 Taxonomy-independent or read clustering approaches

Approaches under this category include: (i) TETRA; (ii) CompostBin; (iii) AbundanceBin; (iv) variants of Self Organizing Maps (SOMs); and (v) MetaCluster (Teeling *et al.,* 2004; Ultsch and Moerchen, 2005; Chan *et al.,* 2008; Chatterji *et al.,* 2008). The simplest methodology is used by TETRA. In a known sequence dataset, TETRA computes pairwise correlations between tetra-nucleotide usage patterns of all reads. This information is then used for segregation of reads into unique bins (Teeling *et al.,* 2004). The SOMs program

applies a neural network-based method, which involves multidimensional clustering of data points. The results of clustering are then plotted onto a two-dimensional map. Both methods use 4-mer frequencies in their algorithms, as it has been demonstrated in papers that 4-mers provide programs with much better discrimination power compared to any other k-mer patterns (Pride *et al.*, 2003). The CompostBin approach uses frequencies of k-mers of different lengths and then applies a weighted PCA-based approach to lessen the dimensionality of the output plot (Chatterji *et al.*, 2008).

One limitation reported for TETRA was that in a metagenome sample with uneven species distribution there was a tendency to group reads originating from the genome of the most abundant organism into several clusters. However, this limitation has been dealth with using the AbundanceBin approach (Wu and Ye, 2011). The main goal of the AbundanceBin is to apply different clustering parameters for reads with different abundance levels. Though the AbundanceBin works proficiently with samples having extremely diverse abundance levels, this approach created artificial bins when the species distribution in a sample was even. However, in practice environmental samples with even specie distributions are very unlikely to occur (Wu and Ye, 2011). The program MetaCluster, developed by Leung and colleagues, attempts to address problems with even species distribution samples by making use of a two-phase approach. Reads are segregated into taxonomically homogenous clusters in the first stage. The second stage, performed by MetaCluster, involves the merging of diverse clusters by generating probabilistic models based on a GC-content analysis of fragments constituting these clusters (Leung *et al.*, 2011).

Most of the taxonomy-independent approaches are more relevant for metagenomes where the numbers of taxonomically classifiable species are very low. Results obtained using taxonomy-independent approaches can also help in downstream processes such as assembly (Mande *et al.*, 2012).

### 1.2.7 Strategies of validation of binning results

To validate taxonomy-dependent binning approaches, validation should be done using simulated metagenomics datasets and databases. The reads in these datasets should simulate the lengths as well as the sequencing errors characteristic of different sequencing technologies. To ensure confidence of validation, multiple datasets of varying sizes should be tried (Mande *et al.*, 2012). To evaluate the performance of different metagenomics tools, Fidelity of Analysis of Metagenomic Samples (FAMeS) is one of the datasets providing

access to multiple simulated datasets and at present is used as a gold standard (Mavromatis *et al.,* 2007). Datasets of diverse taxonomic complexity usually contain about 100 000 reads, having lengths ranging from 650 to 1000 bp. These reads are sampled from 112 real genome sequencing projects, which are populated with typical sequencing errors associated with the Sanger sequencing technology. It should be noted that this database has no data sets simulating the typical errors of the current popular NGS technologies (Mavromatis *et al.,* 2007; Mande *et al.,* 2012). To simulate NGS reads, software tools such as Metasim and ART have to be used (Richter *et al.,* 2008; Huang *et al.,* 2011).

MetaSim, which is a sequencing simulator of genomic and metagenomic data, can be used to generate a collection of artificial reads from provided genomes that mimic typical errors of Roche 454 or Illumina sequencing technologies. Based on a collection of real genome sequences, the program constructs a metagenome allowing unequal representation of the initial genomes in the resulting data set. By representing various levels of taxonomic nodes of the NCBI taxonomy, binning of different programs can be compared in terms of sensitivity and specificity (Richter *et al.,* 2008).

Haung *et al.* (2011) developed the program ARTS, which is a set of simulation tools to generate artificial next-generation sequencing reads. ART simulates both single-end and paired-end reads of the three most popular next-generation sequencing technologies: Roche 454, Illumina and SOLiD. This functionality is most important for testing tools developed for processing and analysis of next-generation sequencing data, for example read alignment, *de novo* assembly and genetic variation detection. ART produces simulated sequencing reads by imitating the sequencing process with built-in technology-specific read error models and base quality value profiles parameterised empirically from large sequencing datasets (Haung *et al.,* 2011).

Simulation of reference databases is another important question to be considered during the assessment of taxonomic-dependent approaches. Simulated databases should be populated by artificial reference sequences showing some level of similarity at different taxonomic levels to the sequences used as input reads. Hence, the simulation is usually done using a leave one scheme, where species are removed from reference databases and validation is performed using reads from this species (Huson *et al.,* 2007; Monzoorul *et al.,* 2009). However, in real metagenomics cases, query reads may also originate from completely unknown taxonomic clades, which cannot be binned at all. So, to incorporate these cases in

the evaluation process, it has been recommended to use the omit one clade strategies, where sequences belonging to an entire clade, such as genus, family, order, class, phylum and above, are removed from the reference database. Sort-ITEMS and DiScRIBinATE make use of this validation strategy (Monzoorul *et al.,* 2009; Gosh *et al.,* 2010).

Four parameters, accuracy, specificity, execution time and required compute power, are used to quantify the binning efficiency of the taxonomy-dependent approaches. The assignment of a read is said to be precise if it is assigned to any taxon that lies in the taxonomic lineage of the source organism read and when assignment specificity is well-defined in terms of the taxonomic levels such as the strain, species, genus, family, order, class, phylum and superkingdom to which the read is assigned. Assignment at the strain level is said to be most precise in cases where the reads originate from an identified strain, sequences of which are present in the reference database, though in most metagenomic case scenarios, the taxon corresponding to the source organism of a read is absent from the reference database (Mande *et al.,* 2012). Different taxonomy-dependent approaches make use of various measures to quantify accuracy and specificity. Methods such as MEGAN and Sort-ITEMS compute the percentage of properly assigned reads at different taxonomic levels and use this information as a measure of sensitivity and specificity (Huson *et al.,* 2007; Monzoorul *et al.,* 2009).

It is of importance to note that a balance exists between the sensitivity and specificity of a method and the requirements for the time of computing. Though composition-based approaches have been shown to outscore alignment-based methods in terms of execution time and computing power, the comparatively lower sensitivity and specificity of these methods compared to alignment-based methods and their limited use with metagenomic datasets containing short reads are still a challenge. However, hybrid binning methods make use of both alignment- and composition-based approaches in order to exploit the relative benefits of both (Teeling *et al.,* 2004; Mohammed *et al.,* 2011).

The efficiency of binning of reads by taxonomy-independent methods is assessed by the following parameters: taxonomic homogeneity of the resultant bins and the number as well as size of bins generated. Hence, in an ideal situation an efficient method should form *n* number of taxonomically homogenous bins where *n* is the number of species constituting the validation dataset. However, in scenarios where multiple homogenous bins arise from the segregation of reads originating from the same species, such parameters as the normalised mutual information and F-score have to be used (Mande *et al.*, 2012). Sun *et al.* (2012) gave

a detailed description of the application of these parameters. Different algorithms used by microbiologists were surveyed in this paper and the authors also compared in a large benchmark study seven representative approaches, which address different issues of concern. A novel protocol was introduced, which allowed different algorithms to be compared using the same platform and different criteria to enable a qualitative assessment of the clustering performance of each algorithm. The newly developed program, ESPRIT-TREE, was found to be one of the best algorithms available in terms of computational efficiency and clustering accuracy (Sun *et al.*, 2012).

Well-established tools, such as the primer-E package, allow for various multivariate analyses, which include generation of multivariate and multidimensional scaling plots, similarity analysis (ANOSIM) and identification of species or functions that contribute to differentiation between two samples (SIMPER) (Clarke, 1993). Multivariate statistics are also incorporated in web-based tools called Metastats, which show high-level discriminatory power to distinguish between replicated metagenome datasets originating from the gut microbiota of lean and obese mice (Turnbaugh *et al.,* 2009; White *et al* 2009). The ShotgunFunctionalizeR package is also known to provide different statistical procedures for assessing functional differences between samples, both for individual genes and for entire pathways using the R statistical package (Kristiansson *et al.,* 2009).

### 1.2.8 Annotation of metagenomic reads

Two different initial schemes can be used for the annotation of metagenomes. The first approach is applied when assembly of genomes from metagenomic reads is one of the main objectives of the study and the assembly has produced large contigs. In this case it is preferable to use existing pipelines for genome annotation, such as RAST and IMG, hence the minimum length of contigs required for this method is 30 000 bp or longer. The second approach is applied for annotation of individual reads or short contigs. Tools specifically developed for metagenomic annotation should be used to perform this task (Aziz *et al.,* 2008; Markowitz *et al.,* 2014)

Metagenomic sequence data annotation can be executed in two steps: features of interest such as coding and non-coding genes are identified (feature prediction step) and then putative gene functions can be assigned by homology search in taxonomic neighbours (functional annotation step). The feature prediction step is the process of labelling sequences as genes or genomic elements. Tools such as FragGeneSan, MetaGeneMark, MetaGeneAnnotator and

Orphelia were specifically designed to handle prediction of CDS in metagenomic reads (McHardy *et al.,* 2007; Noguchi *et al.,* 2006; Hoff *et al.,* 2009; Rho, 2010; Yok *et al.,* 2011). At present, estimates show that only 20-50% of metagenomic sequences can be annotated, leaving the question of the importance of the remaining genes unanswered (Gilbert *et al.,* 2010). The annotation is rarely performed *de novo*. On the contrary, mapping of reads to databases of genes with known functions is preferable. Sequences that cannot be mapped to any known sequence are termed orphan. Orphans constitute the never-ending genetic novelty in microbial metagenomics (Yooseph *et al.,* 2007).

Many databases are available and can be used to annotate metagenomic reads functionally. They commonly come in two varieties: sequence and HMM databases. Searching metagenomic reads through a database of sequences has a tendency to be comparatively quick and may generate more specific hits for the reads that are closely related to sequences in the database, whereas comparing metagenomic reads to an HMM database tends to detect more vaguely related and diverged members of the family, though the precision of identification of very short sequences is not well established. Commonly used sequence databases consist of the SEED annotation system, which is used by MG-RAST and links precise family level functions to higher-order functional subsystems (Overbeek *et al.,* 2014). The KEGG orthology class has also been demonstrated to be particularly valuable, as it maps suitably to KEGG metabolic pathway modules (Kanehisa *et al.,* 2014). The MetaCyc class is alike in that the families are mapped to extensively curated and well-defined metabolic pathways (Caspi *et al.,* 2014). The EggNOG is a database of non-supervised groups of orthologous proteins that incline to be improved frequently, so as to include a huge amount of sequence diversity (Powell *et al.,* 2014). HMM databases suitable for querying metagenomic reads are limited by the Pfam, which uses HMMs to model protein domains (Finn *et al.,* 2013). At present, generation of HMM databases of the full-length and phylogenetically varied protein family is under process. These new databases may be exemplified by Phylofacts (Afrasiabi *et al.,* 2013) and SiftingFamilies, which are also regularly upgraded, like EggNOG (Sharpton *et al.,* 2012; Sharpton, 2014).

### 1.2.9 Sharing and storage of metagenomic data

Tools such as IMG/MER, CAMERA, MG-RAST and EBI metagenomics, which also incorporate QIIME, offer an integrated environment for: (i) analysis, (ii) management, (iii) storage and (iv) sharing of metagenome projects (Oulas *et al.,* 2015). This requires a

constant, generally accepted annotation scheme to be considered to allow for efficient data exchange, integration, sharing and visualisation between different platforms. This will decrease the necessity for reprocessing of metagenomic datasets further, an assignment that is very costly computationally (Oulas *et al.,* 2015).

The Genomic Standards Consortium is currently investing heavily in a generally acknowledged language that shares ontologies and nomenclatures, thus providing a common standard for exchange of data resulting from the analysis of metagenomic projects. Hence, Minimum Information about Metagenome Sequence (Yilmaz *et al.,* 2011) and Minimum Information about a MARKer Sequence (Field *et al.,* 2011) have been devised, which offer a scheme of standard languages for metadata annotation (Thomas *et al.,* 2012; Oulas *et al.,* 2015).

### 1.3 Barcoding

A broad range of genetic data about microorganisms of importance has been made accessible with the advancement seen in different NGS platforms. However, it remains a challenge for many researchers to process this enormous quantity of genetic data to resolve practical demands. Genetic barcoding of microorganisms is the first major area where NGS has met the need of applied microbiology (Reva *et al.,* 2014). Kress and Erickson (2008), defined barcodes as 400-800 bp DNA fragments, which serve as explicit specie identifiers. Deoxyribonucleic acid barcoding is an advanced technique for rapid specie identification based on a standard fragment of DNA sequence (Albu *et al.,* 2011). A DNA barcode gene region should fulfil three major conditions: it should (i) have substantial species-level genetic variability and divergence; (ii) have conserved flanking sites for developing universal PCR primers for varied taxonomic use (however, the present NGS technologies have made this requirement obsolete as there is currently no need to care about PCR-based amplification); and (iii) require a short sequence length to enable the present capabilities of DNA extraction and amplification (Kress and Erickson, 2008) (Table 1.2).

In bacteriology, DNA barcoding was initiated using 16S rRNA as taxonomic markers. This was then followed by the use of many other housekeeping gene sequences as possible barcodes (Weisburg *et al.,* 1991; Case *et al.,* 1997). The 16S rRNA is one of the most sequenced DNA fragments used for specie identification, since it is well conserved in eubacteria and Archaea. This permits the creation of universal primers that enclose different informative variable regions (Coenye and Vandamme, 2003). Some of the limitations

associated with barcoding with 16S rRNA include: (i) the gene is too conserved for discrimination of closely related species; and (ii) the possession of various copies of variable copies of 16S rRNA by an organism also causes a problem (Kunst and Devine, 1991). However, a significant factor that made the 16S rRNA gene suitable for phylogenetic inferences was its resistance to horizontal exchange (Woese and Fox, 1977). The ITS region of nuclear ribosomal DNA (rDNA) was recommended as genetic makers for eukaryotes (fungi), while the mitochondrial gene cytochrome *c* oxidase I (COI) was recognised as a universal DNA barcode for animals (Hebert *et al*., 2003; Nilsson *et al*., 2008).

The main barcoding bodies and resources are Consortium for the Barcode of Life (CBOL), Quarantine Barcoding of Life (QBOL) and Barcode of Life Datasystem (BOLD) (Salvolainen *et al*., 2005; Ali *et al.*, 2014). The CBOL was established in 2004 and has more than 170 member groups from 50 countries to endorse DNA barcoding as the universal standard of identification of biological specimens (Miller, 2005). The QBOL aims to obtain DNA barcode data of significant species of bacteria and other organisms to establish an analytical tool for quarantine (Bonants *et al.,* 2010). The BOLD workbench supports the possession, analysis, storage and publication of DNA barcode records. It enables bioinformatics opening by collecting morphological, molecular and distributional data. The BOLD could be described as the global beginning of the identification of species, which allows users to refer to a specialised database such as disease vector species, threatened species and pathogenic strains (Ball and Armstrong, 2006; Lebonah *et al*., 2014). It is freely accessible and enables researchers to carry out neighbour-joining clustering, identification of taxa using a restructured sequence library and storage of information on the different groups studied (Amanda and Luciane, 2010).

### 1.3.1 Advantages of DNA barcoding

Generally, DNA barcoding is designed to benefit population genetics and systemic studies of habitat-specific bacterial consortia. Barcoding approaches are used for inventory of biodiversity, routine identification of species of interest in environmental samples and flagging of atypical specimens for detailed taxonomic research, since there are almost 1.7 million named species and probably another 10 million (excluding bacteria and archaea) that have not been counted. By contrast to phylogenetics and taxonomy, which aim at inferring relationships of common ancestry, the objective of molecular barcoding is the identification of the presence or absence of taxonomic units of interest in selected environmental samples

or habitats. In population genetics studies, DNA barcoding can provide a first signal of the extent and nature of population divergences; comparative studies of the population diversity in many species and detection and tracing down of microorganisms of interest in the environment (Stoeckle 2003; Hajibabaei *et al.*, 2007).

### 1.3.2 Challenges of DNA barcoding

Since DNA-based species identification depends on differentiating intraspecific from interspecific genetic variation, the ranges of these types of differences are unidentified and may vary between groups. Attempting to resolve newly diverged species may be difficult, since there is no universal DNA barcode gene, no single gene that is conserved in all domains of life and exhibits enough sequence divergence for species discrimination (Kress and Erickson, 2008). Two dynamics that may also strongly affect how well barcode markers work at species identification and discovery are database design and sequence search strategies. The exact method or algorithm to be used in searching a barcode database has not been thoroughly investigated or debated, particularly as regards a multi-locus DNA barcode (Kress and Erickson, 2008).

Advances in sequencing technologies allowed scaling up the barcoding approaches to study environmental populations of microorganisms by using metagenomic technologies. One of the most popular technologies is metabarcoding, which is built upon massive parallel sequencing of species-specific marker genes (i.e. 16S rRNA) from environmental DNA samples by means of universal primers. Metabarcoding should be clearly distinguished from both metagenomics and phylogenetics, which are common sources of confusion. Metagenomics, in contrast to metabarcoding, relies on WGS approaches, omitting either amplification or cloning of DNA fragments. By contrast to phylogenetics, the aim of metabarcoding is to identify known species in the sample, leaving out questions of phylogenetic relations between them and/or new species proclamation.

In the presented work, an idea of multi-locus barcoding is introduced for data mining of metagenomic datasets. It resembles to some extent multi-locus sequence typing (MLST), but in contrast to MLST and metabarcoding it relies on the WGS approach instead of amplification of marker loci. It allows an increase in the number of marker loci to be considered, as there is no need to develop primers for every locus and amplify them separately. In contrast to metagenomics, multi-locus barcoding is not instrumental either for genomic contig assembly or for a functional analysis of bacterial populations. The aim of

multi-locus barcoding remains the same as for other barcoding approaches: species identification in environmental samples and tracing down the strains of interest.

At-a-glance comparison of multi-locus barcoding to metabarcoding is presented in Table 1.2. In subsequent sections of this chapter, a more detailed analysis of multi-locus barcoding is given.

**Table 1.1: Criteria for barcode evaluation**

| Criteria for Evaluation | Multi-locus Barcoding | Metabarcoding |
|---|---|---|
| Barcode sequence length | Usually longer than 10 000 bp, up to 200 000 bp. | Length < 400 bp is usually appropriate; however, some authors recommend sequencing the complete 16S rRNA that is around 1500 bp. |
| Sensitivity and taxonomic resolution. | Adjustable to different taxonomic levels | 16S rRNA is believed to be species-specific, but bacterial genera are usually identified with appropriate statistical reliability. |
| Specificity and false positive rate | Specificity increases with the number of individual marker loci in the barcode sequence. | Generation of chimeric 16S rRNA sequences is a common problem for all sequencing technologies based on the massive parallel sequencing by synthesis technologies (Roche 454, Illumina and Ion Torrent). |
| Sufficient coverage requirement and biased predictions | This technology assumes further reduction of the sequencing price, as high coverage above 100 is a necessity for this approach. | This approach allows a significant enrichment of sequences of interest by amplification. Coverage above 15 is sufficient for species identification; however, the amplification process based on universal primers is rather biased. |
| Applicability for phylogenetic inferences | Not applicable | Not applicable |
| Dependence on availability of comprehensive databases | Barcodes are developed “in fly” for user-provided sets of genomes, but the barcodes are applicable to identification of these organisms only. | Metabarcoding allows identification of any organism deposited in the reference database. |

### 1.3.3 Multi-locus barcoding and metabarcoding

Evolving molecular methods show fast advancement with tools used for specie identification in environmental samples (Pochon *et al*., 2013; Wood *et al*., 2013). Of interest is DNA barcoding and metabarcoding, which can possibly offer more precise and standardised, high-resolution taxonomic data (Hajibabei *et al*., 2007; Ji *et al.,* 2013; Zaiko *et al*., 2015). Metabarcoding is defined as a quick technique of high-throughput, DNA-based identification of species of interest from a complex and degraded sample of environmental DNA or from a bulk group of specimens. Metabarcoding techniques are faced with several limitations, which can obstruct their ability to yield robust, comparable biodiversity estimates (Cristescu, 2014). Limitations associated with metabarcoding are: (i) generation of amplification biases that contribute to errors that can influence biodiversity estimates; this is usually caused by dependency on the intermediate PCR step, which enriches the DNA templates extracted from a bulk sample; (ii) low taxonomic resolution due to a high level of conservation of these sequences; and (iii) generation of chimeric sequences that artificially increase the species richness of the samples (Bik *et al.,* 2012; Ratnasingham and Hebert, 2013)

Though metabarcoding has been labelled as a promising method rendering conventional DNA barcoding irrelevant (conventional DNA barcoding involves the use of a single gene to identify a given species through the comparison of nucleotide sequences in the DNA to that of same gene in other species) (Taylor and Harris, 2012), DNA multi-locus barcoding and metabarcoding are complementary methods and impending biodiversity research would benefit from harmonising the two methods. The two techniques are similar in that they both use DNA-based identification of species, but they have differing assets that are determined by their unique sequencing technologies and precise aims. The multi-locus barcoding method involves aligning DNA reads generated by WGS metagenomics against multiple taxon-specific loci, while metabarcoding involves enormous targeted parallel sequencing of one or several marker genes. Both methods make use of the massive advantages associated with second-generation technology, which enables the production of loads of sequences at a run. However, the ability to interpret the results is highly dependent on both the skill of the experimenter and accessibility of sophisticated computer tools and well-populated databases providing access to reference sequences of interest (Janzen *et al.,* 2005; Ji *et al.,* 2013).

Several international projects have been launched with the aim to barcode multiple live organisms of the earth or specific environments. The DNA barcoding approaches have been

designed to build a link between molecular ecologists and morphological taxonomists by generating reference databases based on verified and curated specimens (Barrett and Hebert 2005; Hebert and Gregory 2005). This alliance can also be extended to metabarcoding techniques by designing metabarcodes within the standardised barcodes. The taxonomic data associated with barcode sequences enables researchers to place OTUs in important evolutionary, physiological and ecological perspectives. With further progressing of barcoding initiatives and better curated specimens of museum assortments, more reference sequences will be provided to link researchers to useful biological information (Janzen *et al.,* 2005; Ji *et al.,* 2013).

### 1.3.4 DNA barcoding of bacteria

Barcoding of bacterial communities is of enormous importance to resolve several health care, ecologically and epidemiologically associated problems (Reva *et al.,* 2014). The importance of bacterial community barcoding for health care problems has been demonstrated through fingerprinting of both gut micro-flora and the Human Microbiome Project (Eckburg *et al.,* 2005; www.hmpdacc.org/index.php). It was reported that the micro-flora of every individual are distinct owing to the effect of external factors such as dietary specificity, lifestyle, medication and genetic specificity (Zoetendal *et al.,* 1998; Suau *et al.,* 1999; Hayashi *et al.,* 2002; Lay *et al.,* 2005; Mueller *et al.,* 2006; Dicksved *et al.,* 2007; Jernberg *et al.,* 2007; Dethlefsen *et al.,* 2008). Different studies also reported that the micro-flora of an individual may cause predilection to obesity, as well as several other immune and inflammatory diseases, such as diabetes (Larsen *et al.,* 2010; Hullar and Lampe, 2012; Kelly and Mulder, 2012; Shanahan, 2012). The management of an individual course of disease, treatment with specific drugs and selection of the most appropriate therapeutic regimens can also be supervised by barcoding of microbial communities. Profiling of complex communities of potentially pathogenic microorganisms in airways of cystic fibrosis patients may be linked to the inception of disease (Zemanick *et al.,* 2011).

Studies on bacterial pathogens give a representation of how populations of bacteria act as a group, but with inadequate resolution to know how microorganisms act as individuals. Hence, it is important to produce markers, which will allow differentiation between lineages and genetic variants in mixed populations during an extended infection. Signature marked mutagenesis based on insertion of transposon sequences serving as barcodes was presented to be instrumental in the study of the dynamics of bacterial populations. This approach allowed

tracing the fate of every individual mutant in the population (Hensel *et al.,* 1995; Chang and Mekalanos, 1998; Darwin and Miller, 1999; Edelstein *et al.,* 1999; Mecsas *et al.,* 2001).

Several success stories of the application of barcoding and metabarcoding are considered below.

Makarova *et al*. (2012) established a universal DNA barcode based on the elongation factor Tu (*tuf*) gene for phytoplasma identification. They also designed a set of primers, which amplified a 420-444 bp fragment of *tuf* from all 91 strains of phytoplasma (16S rRNA, groups –I through –VII, -IX through –XII, -XV and –XX). Assessment of the neighbour-joining trees constructed from the *tuf* barcode showed that the *tuf* tree was congruent to those based on 16S rRNA but provided greater inter- and intra-group divergence. Hence, they demonstrated that *tuf* sequences can be applicable to the barcoding of phytoplasmas. The *tuf* barcodes performed much better than the 1.2 kbp fragment of the 16S rRNA genes and offered an easy-to-use technique for phytoplasma identification (Makarova *et al*., 2012).

In another paper, the effect of fire on microbial communities in chaparral soils was tested by means of DNA barcoding. This technique allowed a comparison of microorganisms found in burnt and unburnt soil samples. DNA barcoding was based on analysis of 16S rRNA genes. Two sets of primers were used for PCR, one for bacteria and another for archaea. Purified DNA was then sub-cloned into TA plasmids. Sequencing of 62 plasmids generated an array of DNA data, which was then used to search the GenBank database with the BLASTNprogram. The generated data revealed that the most abundant microbes found in the unburnt samples were less visible in the burnt samples. Larger diversity of microbes was also observed in the burnt soil samples, meaning that most of the soil archaea microbes might have moved in quickly after the fire; after the community stabilised, these microbes could be displaced (Natalie, 2013).

In oceans, microbial life is responsible for almost all production and mediation of all biogeochemical processes (Sogin *et al.,* 2006). However, the dimension of taxonomic and genetic versatility of these microbiota is poorly understood. Advances in community genomics and metagenomic techniques are leading to valuable insight into the prokaryotic diversity and processes of molecular evolution of ocean-inhabiting bacterial communities (DeLong, 2004; Tyson *et al.,* 2004; Venter *et al.,* 2004; Tringe *et al.,* 2005; DeLong *et al.,* 2006; Leclerc *et al.,* 2007).

Studies exploring the community dynamics of microbes depend heavily on genecentric metagenomic profiling using 16S rRNA and 60 kDa chaperonin protein (cpn60) as marker genes. Links *et al.* (2012) assessed DNA barcoding techniques based on amplification of 16S rRNA genes and the protein coding *cpn60* genes. The *cpn60* gene reported as a universal target that outperformed the traditionally used 16S rRNA as a barcode sequence. These authors suggested *cpn60* as an ideal barcode for species-level characterisation of bacterial communities. Assembling consensus sequences for barcodes was also reported to be a good method of tracking and identification of new microbes by metagenomics (Links *et al.*, 2012).

Liu *et al*. (2013) verified whether *Mollitrichosiphum,* an aphid genus with life cycles on subtropical woody host plants, and *Buchnera,* the main endosymbiont of aphids, evolved in parallel. *Buchnera* belongs to the γ subdivision of proteobacteria and is commonly assumed to be present in all aphid species, where it exists in specific cells termed bacteriocytes (Buchner, 1965; Lebonah *et al.,* 2014). The following aphid genes, mitochondrial COI, cytochrome oxidase subunit I and *Cytb,* cytochrome b: nuclear *EF1α,* translation elongation factor 1 apha and two *Buchnera* genes, namely 16S rDNA and *gnd* for gluconate-6-phosphate dehydrogenase, were used to reconstitute the phylogenies of these species. The phylogenetic trees of aphids and *Buchnera* were then compared. It was reported that phylogenetic evidence for the parallel evolution of *Millitrichosiphum* and *Buchnera* at the intraspecific as well as interspecific levels supported the prospect of using endosymbiont genes to analyse the evolutionary history and biogeographical distribution of host organisms. These authors also explored the possibility of the *Buchnera* gene *gnd* being used as a barcode marker for aphid identification. This study showed that *Buchnera* gene *gnd* was also suitable for barcoding as a marker for aphids, just like the traditional COI barcode (Liu *et al.,* 2013).

### 1.3.5 Barcoding and multi-locus sequence typing

Since the single-gene technique of DNA barcoding fails to differentiate between closely related organisms on the level of species and sub-species, there is a dire need for the development of more sensitive DNA markers in both medical and biotechnological microbiology (van Belkum *et al.,* 2001; Urwin and Maiden, 2003). Hence, it was hypothesised that the comparison of strains by various gene sequences would provide better resolution to distinguish between closely related organisms. Multi-locus sequence typing was then introduced, which uses PCR amplification and sequencing of small fragments of multiple genes comprising diagnostic signals instead of one barcode gene or whole-gene

sequence data (Maiden *et al.,* 1998). However, the challenge with MLST is that different housekeeping genes and a varying number of polymorphic sites are used in diagnostic protocols of different groups of organisms, which impedes cross-platform evaluation (Urwin and Maiden, 2003). To aid DNA barcoding, numerous potent laboratory information management systems have been presented, including the BIGSdb database, which is now integrated into the PubMLST website (Jolley and Maiden, 2010). The BIGSdb database permits cross-referencing among diverse MLST datasets and makes use of data for genome functionality, epidemiology and evolutionary predictions (Joelly *et al.,* 2012a).

Joelly *et al.* (2012b) proposed an MLST typing technique called ribosomal MLST (rMLST), which indexes variations in 53 genes encoding bacterial ribosome protein subunits (rps genes) as a way of incorporating microbial genealogy and typing. Grouping provided by rMLST was consistent with the present nomenclature systems independently of the clustering algorithm being used (Joelly *et al.,* 2012b). Moreover, by increasing the analytic sets of polymorphic sites to a larger number of housekeeping genes using high-throughput sequencing techniques, MLST datasets might become universal and useable in several microbial research projects (Reva *et al.,* 2014). A variant MLST technique called short read sequence typing (SRST) was introduced by Inouye *et al.* (2012). The SRST technique maps IIIumina DNA reads against target sequences, which are then spontaneously recovered from the MLST database (http://pubmlst.org). These short reads are initially mapped by the BWA tool and then processed by Samtools (Li *et al.*, 2009; Li and Durbin, 2010). Genometa is another program for effective barcoding of bacterial communities and populations using short IIIumina DNA read (Davenport *et al.*, 2012). BOWTIE is used to map reads instead of BLASTN in this program. An extended version of the open source Integrated Genome Browser browser is used to view the alignment (Davenport *et al.*, 2012).

### 1.4 Research aim and objectives

#### Aim

The aim of this study was to create a computer system for multi-locus genetic barcoding suitable for identification and tracking down of biotechnological strains in the environment. These software tools provide access to databases of genetic barcodes for program testing and application in biotechnology.

## Objectives

- To create and test a novel software tool allowing automatic creation of diagnostic multi-locus metabarcodes (concatenated sequences of marker genes) for user-provided groups of microorganisms. For case studies, the following groups of organisms representing different levels of taxonomic relatedness were used: *Bacillus cereus, Escherichia/Shigella, Lactobacillus, Mycobacteria, Streptococcus, Salmonella* and *Prochlorococcus.* The composition of each group will be explained in detail in the next chapter.
- To evaluate the performance of the designed metabarcodes in terms of sensitivity and specificity by using publicly available metagenomic datasets and artificially created metagenomic datasets.
- To analyse functional categories of marker genes selected by the program for diagnostic metabarcodes designed for different groups of microorganisms.
- To develop a standardised web-based pipeline (BarcodeGenerator) for metabarcode development for any given group of microorganism.
- To develop a standardised barcoding pipeline implemented in the form of a stand-alone Python program available for download from the BarcodeGenerator Web-site to perform taxonomic binning of metagenomics reads against designed multi-locus barcodes.

**References**

Acinas SG, Marcelino LA, Klepac-Ceraj V and Polz MF (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *Journal of Bacteriology,* 186, pp. 2629-2635

Afrasiabi C, Samad B, Dineen D, Meacham C and Sjölander K (2013). The PhyloFacts FAT-CAT web server: orthology identification and function prediction using fast approximate tree classification. *Nucleic Acids Research,* 41, pp. W242-W248

Albu M, Nikbakht H, Hajibabaei M and Hickey DA (2011). The DNA Barcode Linker. *Molecular Ecology Resources,* 11, pp. 84-88

Ali MA, Gyulai G, Hidvégi N, Kerti B, Al Hemaid FMA, Pandey AK and Lee J (2014). The changing epitome of species identification-DNA barcoding. *Saudi Journal of Biological Sciences,* 21, pp. 204-231

Allan E (2014). Metagenomics unrestricted access to microbial communities. *Virulence,* 5, pp. 397-398.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). Basic local alignment search tool. *Journal of Molecular Biology,* 215, pp. 403-410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25, pp. 3389-3402

Amanda CP and Luciane M (2010). DNA barcoding and traditional taxonomy unified through integrative taxonomy: a view that challenges the debate questioning both methodologies. *Biota Neotropica,* 10, pp. 30-33

Ambarda S, Gupta R, Trakoo D, Lal R and Vakhlu J (2016). High Throughput Sequencing: An Overview sequencing Chemistry. Indian Journal of Microbiology, 56(4), pp. 394-404

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Oslon R, et al., (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics,* 9:75

Ball SL and Armstrong KF (2006). DNA barcodes for insect pest identification: a test case

with tussock moths (Lepidoptera: *Lymantriidae*). *Canadian Journal of Forest Research.* 36, pp. 337-350

Barrett RDH and Hebert PDN (2005). Identifying spiders through DNA Barcodes. *Canadian Journal of Zoology,* 83, pp. 481-491

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science,* 289, pp. 1902-1906

Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simón-Soro A, Pignatelli M and Mira A (2012). The oral metagenome in health and disease. *The ISME Journal,* 6, pp. 45-46

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, et al., (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456, pp. 53-59

Berger SA and Stamatakis A (2011). Aligning short reads to reference alignments and trees. *Bioinformatics,* 27, pp. 2068-2075

Bik HM, Porazinska DL, Creer S, Caporaso G, Knight R and Thomas WK (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution,* 27, pp. 233-243

Bonants P, Groenewald E, Rasplus JY, Maes M, De Vos P, Frey J, Boonham N, Nicolaisen M, Bertacini A, Robert V, Barker I, Kox L, et al., (2010). QBOL: a new EU project focusing on DNA barcoding of quarantine organisms. *EPPO Bulletin,* 40, pp. 30-33

Bowen De León K, Gerlach R, Peyton BM and Fields MW (2013). Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geyser Basin, Yellowstone National Park. *Frontiers in Microbiology,* 4(330), doi:10.3389/fmicb.2013.00330

Bowers J, Mitcheel J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D, Lowman GM, Marappan S, et al., (2009). Virtual terminator nucleotides for next generation DNA sequencing. *Nature Methods,* 6, pp.593-595

Brady A and Salzberg SL (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated markov models. *Nature Methods,* 6, 673-676

Braslavsky I, Hebert B, Kartalov E and Quake SR (2003). Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 100, pp. 3960-3964

Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B and Albert J (2013). PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PLos One*, 8:e70388

Bruke C, Kjelleberg S and Thomas T (2009). Selective extraction of bacterial DNA from the surfaces of macroalgae. *Applied and Environmental Microbiology*, 75, pp. 252-256

Buchner P (1965). Endosymbiosis of animals with plant microorganisms. *Interscience*, New York, NY, USA pp. 909.

Bulgarelli D, Rott M, Schaeppi K, Ver Loren van Themaat E, Ahmadine N, Assenza F, Rauf P, Huettel B, Reinhardt R, Schmelzer E, Peplies J, Gloeckner FO, Amann R, Eickhorst T and Schulze-Lefert P (2012). Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature*, 488, pp.91-95

Bulgarelli D, Schlaeppi K, Spaepen S, Ver Loren van Themaat E and Schulze-Lefert P (2013). Structure and functions of the bacterial microbiota of plants. *Annual Review of Plant Biology*, 64, pp. 807-838

Case RJ, Bouchner Y, Dallöf I, Holmström C, Doolittle WF and Kjelleberg S (1997). Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, 73, pp. 278-288

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Krummenacker AKM, Latendresse M, Mueller LA, et al., (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome database. *Nucleic Acids Research*, 42, pp. D459-D471

Chan CK, Hsu AL, Halgamuge SK and Tang SL (2008). Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*, 9:215

Chang SL and Mekalanos (1998). Use of signature-tagged transposon mutagenesis to identify *Vibrio cholera* genes critical for colonization. *Molecular Microbiology*, 27, pp.797-805

Chatterji S, Yamazaki I, Bai Z and Eisen JA (2008). CompostBin: a DNA composition-based

algorithm for binning environmental shotgun reads. *Research in Computational Molecular Biology*, 4955, pp. 17-28

Chevreux B, Wetter T and Suhai S (1999). Genome Sequence Assembly Using Trace Signals and Additional Sequence Information Computer Science and Biology. *Proceedings of the German Conference on Bioinformatics*, 99, pp. 45-46

Church GM and Gilbert W (1984). Genomic sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 81, pp. 1991-1995

Coenye T and Vandamme P (2003). Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters*, 28, pp. 45-49

Clarke KR (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18, pp. 117-143

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR and Tiedje JM (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42, pp. D633-D642

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry III CE, Tekaia F, Badcock K, Basham D, Brown D, et al.,(1998). Deciphering the biology of *Mycobacterium tuberculosis* from complete genome sequence. *Nature*, 393, pp. 537-544

Cox MP, Peterson DA and Biggs PJ (2010). SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, 11, doi: 10.1186/1471-2105-11-485

Cristescu ME (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology and Evolution*, 29, pp. 566-571

Darwin AJ and Miller VL (1999). Identification of *Yersinia enterocolitica* genes affecting survival in an animal host using signature-tagged transposon mutagenesis. *Molecular Microbiology*, 32, pp. 51-62

Davenport CF, Neugebauer J, Beckmann N, Friedrich B, Kameri B, Kokott S, Paetow M,

Siekmann B, Wieding-Drewes M, Wienhöfer M, Wolf S, Tümmeler B, et al., (2012). Genometa- a fast and accurate classifier for short metagenomic shotgun reads. *PLoS ONE,* 7:e41224

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ and Turnbaugh PJ (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature,* 505, pp. 559-563

Degnan PH and Ochman H (2012). Illumina-based analysis of microbial community diversity. *The ISME Journal,* 6, pp. 183-194

Delmont TO, Robe P, Clark I, Simonet P and Vogel TM (2011). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *Journal of Microbiological Methods,* 86, pp. 397-400

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, von Mering C and Vorholt JA (2009). Community proteogenomics reveals insights into the physiology of phyllospere bacteria. *Proceedings of the National Academy of Sciences of the United States of America,* 106, pp. 16428-16433

DeLong EF (2004). Microbial population genomics and ecology: the road ahead. *Environmental Microbiology,* 6, pp. 875-878

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW and Karl DM (2006). Community genomics among stratified microbial assemblages in the oceans interior. *Science,* 311, pp. 496-503

Desai N, Antonopoulous D, Gilbert JA, Glass EM and Meyer F (2012). From genomicsto metagenomics. *Current Opinion in Biotechnology,* 23, pp. 72-76

Dethlefsen L, Huse S, Sogin ML and Relman DA (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biology,* 6:e280

Diaz NN, Krause L, Goesmann A, Niehaus K and Nattkemper TW (2009). TACOA: taxonomic classification of environmental genomic fragments using a kermelized nearest neighbour approach. *BMC Bioinformatics,* 10:56

Dicksved J, Filstrup H, Bergstrm A, Rosenquist M, Pershagen G, Scheynius A, Roos S, Alm JS, Engstrand L, Braun-Fahrländer C, von Mutius E and Jansson JK (2007). Molecular fingerprinting of the fecal microbiota of children raised according to different lifestyles. *Applied and Environmental Microbiology,* 73, pp. 2284-2289

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE and Relman DA (2005). Diversity of human intestinal microbial flora. *Science,* 308, pp. 1635-1638

Edelstein PH, Edelstein MAC, Higa F and Falkow S (1999). Discovery of virulence genes of Legionella pneumophila by using signature tagged mutagenesis in a guinea pig pneumonia model. *Proceedings of the National Academy of Sciences of United States of America,* 96, pp. 8190-8195

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26, pp. 2460-2461

Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi I, Klenk HP, Knight R, Kottmann R, et al., (2011). The Genomic Standards Consortium. *PLoS Biology,* 9:e1001088

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J and Punta M (2013). Pfam: the protein families database. *Nucleic Acids Research,* 42, pp.D222-D230

Garcia-Garcerà M, Garcia-Etxebarria K, Coscollà M, Latorre A and Calafell F (2013). A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin. *PLoS ONE,* 8(9):e74914

Ghosh TS, Mohammed MH, Komanduri D and Mande SS (2011). ProViDE: a software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation,* 6, pp. 91-94

Ghosh TS, Monzoorul Haque M and Mande SS (2010). DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenome sequences. *BMC Bioinformatics,* 11:S14

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P and Joint I (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial

communities. *PLoS One*, 3:e3042

Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, Huse S, Hughes M, Joint I, Somerfield PJ and Mühling M (2010). *PLoS One*, 5:e15545

Godzik A (2011). Metagenomics and the protein universe. *Current Opinion in Structural Biology*, 21, pp. 398-403

Goltsman DSA, Denef VJ, Singer SW, Verberkmoes NC, Lesfsrud M, Mueller RS, Dick SJ, Sun CL, Wheeler KE, Zemla A, Baker BJ, Hauser L, Land M, Shah MB, et al., (2009). Community genomic and proteomic analysis of chemoautotrophic iron-oxidizing "*Leptospirillum rubarum"* (Group II) and *"Leptospirillum ferrodiazotrophum"* (Group II) bacteria in acid mine drainage biofilms. *Applied and Environmental Microbiology*, 75, pp. 4599-4615

Grada A and Weinbrecht K (2013). Next-Generation Sequencing: Methodology and Application. Journal of Investigative Dermatology, 133:e11

Gupta SS, Mohammed MH, Ghosh TS, Kanungo S, Nair GB and Mande SS (2011). Metagenome of the gut of a malnourished child. *Gut Pathogen*, 3:7

Hajibabaei M, Singer GAC, Hebert PDN and Hickey DA (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, 23, pp. 167-172

Handelsman J (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reveiws*, 68, pp. 669-685

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Bralavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, et al., (2008). Single-molecule DNA sequencing of a viral genome. *Science*, 320, pp. 106-109

Haung W, Li L, Myers JR and Marth GT (2011). ART: a next-generation sequencing read simulator. *Bioinformatics*, 28, pp. 593-594

Hayashi H, Sakamoto M and Benno Y (2002). Fecal microbial diversity in a strict vegetarian as determined by molecular analysis and cultivation. *Microbiology and Immunology*, 46, pp. 819-831

Heather JM and Chain B (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107, pp. 1-8

Hebert PDN, Cywinska A, Ball SL and DeWaard JR (2003). Biological identifications through DNA barcodes. *Prcoceedings of the Royal Society B Biological Sciences,* 270, pp. 313-321

Hebert PDN and Gregory TR (2005). The promise of DNA barcoding for taxonomy. *Systematic Biology,* 54, pp. 852-859

Hensel M, Shea JE, Gleeson C, Jones MD, Dalton E and Holden DW (1995). Simultaneous identification of bacterial virulence by genes by negative selection. *Science,* 269, pp. 400-403

Hoff KJ, Lingner T, Meinicke P and Tech M (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Research,* 37, pp. W101-W105

Hong S, Bunje J, Leslin C, Jeon S and Epstein SS (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *Multidisciplinary Journal of Microbial Ecology,* 3, pp. 1365-1373

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

Hugenholtz P and Pace NR (1996). Identifying microbial diversity in natural environment: a molecular phylogenetic approach. *Trends in Biotechnology,* 14, pp. 190-197

Hullar MA and Lampe JW (2012). The gut microbiome and obesity. *73$^{rd}$ Nestle Nutrition Institute Workshop,* pp. 67-79

Human Microbiome Project Consortium (2012a). Structure, function and diversity of the healthy human microbiome. *Nature,* 486, pp. 215-221

Huson DH, Auch AF, Qi J and Schuster SC (2007). MEGAN analysis of metagenomic data. *Genome Research,* 17, pp. 377-386

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ and Tappu R (2016). MEGAN Community-Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology,* 12:e1004957

Inouye M, Conway TC, Zobel J, and Holt KE (2012). Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC Genomics,* 13:338

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL and Armbust EV (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science,* 335, pp .587-590

Janzen DH, Hajibabaei M, Burns JM, Hallwachs W, Remigio ED and Hebert PDN (2005). Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society B Biological Sciences,* 360, pp. 1835-1845

Janzen DH, Hallwachs W, Blandin P, Burns JM, Cadiou JM, Chacon I, Dapkey T, Deans AR, Epstein ME, Espinoza B, Franclemont JC, Haber WA et al (2009). Integration of DNA barcoding into angoing inventory of complex tropical biodiversity. *Molecular Ecology Resources,* 9, pp. 1-26

Jenberg C, Lofmark S, Edlund C and Jansson JK (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME Journal,* 1, pp. 56-66

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, et al., (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters.* 16, pp. 1245-1257

Joelley KA and Maiden MC (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics,* 11:595

Jolley KA, Hell DMC, Bratcher HB, Harrison OB, Feavers IM, Parkhill J and Maiden MCJ (2012a). Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid Web-based analysis methods. *Journal of Clinical Microbiology,* 50, pp. 3046-3053

Joelly KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarattttthna H, Harrison OB, Sheppard SK, Cody AJ and Maiden MCJ (2012b). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology,* 158, pp. 1005-1015

Jumpstart Consortium Human Microbiome Project Data Generation Working Group (2012).

Evaluation of 16S Rdna-based community profiling for human microbiome research. *PLoS ONE,* 7:e39315

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M and Tanabe M (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research,* 42, pp. D199-D205

Kelley DR, Liu B, Delcher AL, Pop M and Salzberg SL (2012). Gene prediction with glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research,* 40:e9

Kelly D and Mulder IE (2012). Microbiome and immunological interactions. *Nutrition Reveiew,* 70, pp. S18-S30

Kent WJ (2002). BLAT-the BLAST –like alignment tool. *Genome Research,* 12, pp. 656-664

Kielbasa SM, Wan R, Sato K, Horton P and Firth MC (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research,* 21, pp. 487-493

Kircher M, Sawyer S and Meyer M (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the illumine platform. *Nucleic Acids Research,* 40:e3

Krause L, Diazz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA and Stoye J (2008). Phylogenetic classification of short environmental DNA fragments. *Nucliec Acids Research,* 36, pp.2230-2239

Kress WJ and Erickson DL (2008). DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America,* 105, pp. 2761-2762

Kristiansson E, Hugenholtz P and Dalevi D (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics,* 25, pp. 2737-2738

Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D and Knight R (2011). Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics,* 13, pp. 47-58

Kunin V, Copeland A, Lapidus A, Mavromatis K and Hugenholtz P (2008). A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reveiws,*

72, pp. 557-578

Kunst F and Devine K (1991). The project of sequencing the entire *Bacillus subtilis* genome. *Research Microbiology,* 142, pp. 905-912

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Thurber RLV, Knight R, Beiko RG and Huttenhower C (2013). Predictive functional profiling of microbial communities using 16S rDNA marker gene sequences. *Nature Biotechnology*, 31, pp. 814-821

Langmead B, Trapnell C, Pop M and Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology,* 10:R25

Larsen N, Vogensen FK, van den Berg FW, Nielsen DS, Andreasean AS, Pedersen BK, Al-sound WA, Søresen SJ, Hansen LH and Jakobsen M (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE,* 5:e280

Lay C, Rigottier-Gois L, Holmstrøm K, Rajilic M, Vaughan EE, de Vos WM, Collins MD, Thiel R, Namsolleck P, Blaut M and Doré J (2005). Colonic microbiota signatures across five northern European countries. *Applied and Environmental Microbiology,* 71, pp. 4153-4155

Lebonah DE, Dileep A, Chandrasekhar K, Sreevani S, Sreedevi B, and Pramoda Kumari J (2014). DNA Barcoding on Bacteria: A Reveiew. *Advances in Biology,* pp. 1-9

Leclerc M, Juste C, Marteau P, Nalin R, Blottiere H and Dore J (2007). Intestinal metagenomics and nutrition. *Annuals of Nutrition and Metabolism,* 51(supplement 1, no.13)

Lee JH, Daugharty ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, et al., (2014). Highly Multiplexed Subcellular RNA Sequencing in Situ. Science, 343, pp. 1360-1363

Lee H, Gurtowski J, Yoo S, Nattestad M, Marcus S, Goodwin S, McCombie WR and Schatz (2016). Third-generation sequencing and the future of genomics. *bioRxiv The preprint server for Biology*

Leung HCM, Yiu SM, Yang B, Peng Y, Wang Y, Liu Z, Chen J, Qin J, Li R and Chin FYL (2011). A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics,* 27(11), pp. 1489-1495

Li H and Durbin R (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* 26, pp. 589-595

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25, pp. 2078-2079

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L and Law M (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012:11

Links MG, Dumonceaux TJ, Hemmingsen and Hill JE (2012). The Chaperonin-60 Universal Targert Is a Barcode for Bacteria That Enables *De Novo* Assembly of Metagenomic Sequence Data. *PLoS ONE,* 7:e49755

Liu B, Gibbons T, Ghodsi M and Pop M (2010). MetaPhyler: taxanomic profiling for metagenomic sequences. *Proceedings of IEEE Bioinformatics Biomed*, pp. 95-100

Liu L, Huang X, Zhang R, Jiang L and Qiao G (2013). Phylogenetic congruence between *Mollitrichosiphum* (Aphididae:Greenideinae) and Buchnera indicates insect-bacteria parallel evolution. *Systematic Entomology*, 38, pp. 81-92

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, et al., (2013). Metagenomic 16S rDNA illumine tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology,* 16, pp. 2659-2671

Loman NJ, Quick J and Simpson JT (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12, pp. 733-735

Lozupone CA and Knight R (2007). Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 104, pp. 11436-11440

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S, Tremblay J, Engelbrektson A, Kunin V, Glavina del Rio T, Edgar RC, Eickhorst T, Ley RE, et al., (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature*, 488, pp. 86-90

Lykidis A, Chen CL, Tringe SG, McHardy AC, Copeland A, Kyriprides NC, Hugenholtz P, Macarie H, Olmos A, Monroy O and Liu WT (2011). Multiple syntrophic interactions in a

terephthate-degrading methanogenic consortium. *The ISME Journal*, 5, pp. 122-130

Maiden MC, Bygraves JA, Feil E, Mrelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M and Spratt BG (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Preceedings of the National Academy of Sciences of the United States of America*, 95, pp. 3140-3145

Makarova O, Contaldo N, Paltrinieri S, Kawube G, Bertaccini A and Niclaisen M (2012). DNA barcoding for identification of "*Candidatus phytoplasmas"* using a fragment of the elongation factor Tu gene. *PLoS ONE*, 17:e52092

Mande SS, Mohammed MH and Ghosh TS (2012). Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics*, 13, pp. 669-681

Mardis ER (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9, pp. 387-402

Mardis ER (2011). A decade's perspective on DNA sequencing technology. *Nature* 470, pp. 198-203

Mardis ER (2017). PERSPECTIVE DNA sequencing technologies: 2006-2016. *Nature Protocols*, 12, pp. 213-218

Matsen AFA, Kodner RB and Armbrust EV (2010). Pplacer: linear time maximum likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11:538

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, et al., (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4, pp. 495-500

Maxam AM and Gilbert W (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74, pp. 560-564

McCliment EA, Voglesonger KM, O'Day PA, Dunn EE, Holloway JR and Cary SC (2006). Colonization of nascent, deep-sea hydrothermal vents by novel Archaeal and Nanoarcheal assemblage. *Enviromental Microbiology*, 8, pp. 114-125

McGinn S and Gut IG (2013). DNA sequencing-spanning the generations. *New Biotechnology*, 30, pp. 366-372

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P and Rigoutsos I (2007). Accurate phylogenetic classification of variable length DNA fragments. *Nature Methods*, 4, pp. 63-72

Mecas J, Bilis I and Falkow S (2001). Identification of attenuated *Yersina pseudotuberculosis* strains and characterization of an orogastric infection in BALB/c mice on day 5 postinfection by signature-tagged mutagenesis. *Infection and Immunity*, 69, pp. 2779-2787

Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, Raes J and Bork P (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE*, 7:e31386

Metzker ML (2005). Emerging technologies in DNA Sequencing. *Genome Research*, 15, pp.1767-1776

Metzker ML (2010). Sequencing technologies-the next generation. *Nature Review Genetics*, 11, pp.31-46

Meyer F, Paarmann D, D'Souza M, Oslon R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J and Edwards RA (2008). The Metagenomics RAST server - A public resource for automatic phylogenetics and functional analysis of Metagenomes. *BMC Bioinformatics*, 9:386

Mignardi M and Nilsson M (2014). Fourth-generation sequencing in the cell and the clinic. *Genome Medicine*, 6(31), (http://genomemedicine.com/content/6/4/31)

Miller JR, Koren S and Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95, pp. 315-327

Miller SE (2005). Proposed standards for BARCODE records in INSDC (BRIs), in Request documentation for continuation of support by the Alfred P. Sloan Foundation submitted by the Smithsonian Institution on behalf of Consrtium for the barcode of Life: 22 January 2006 Robert H, pp. 36-38

Mohammed MH, Ghosh TS, Reddy RM, Reddy CVSK, Singh NK and Mande SS (2011). INDUS-a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC Genomics*, 12:S4

Mohammed MH, Ghosh TS, Singh NK and Mande SS (2011). SPHINX-an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, 27, pp. 22-30

Monzoorul Haque M, Ghosh TS, Komanduri D and Mande SS (2009). Sort-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25, pp. 1722-1730

Morey M, Fernàndez-Maimiesse A, Castiñeiras D, Fraga JM and Cocho JA (2013). A glimpse into past, present and future DNA sequencing. *Molecular Genetics and Metabolism*, 110, pp. 3-24

Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, Leleiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, et al., (2012). Dysfunction of the intestine microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13:R79

Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzàlez A, Fontana L, Bernard H, Knight R and Gordon J (2011). Diet drivers convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, 332, pp. 970-974

Mueller S, Saunier K, Hanisch C, Norin E, Alm L, Midtvedt T, Cresci A, Silvi S, Orpianesi C, Verdenelli MC, Clavel T, Koebnickc C, Zunft HJF, et al., (2006). Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. *Applied and Environmental Microbiology*, 72, pp. 1027-1033

Murray DC, Coghlan ML and Bunce M (2015). From Benchtop to Desktop: Important Considerations when Designing Amplicon Sequencing Workflows. *PLoS ONE*, 10:e0124671

Nalbantoglu OU, Way SF, Hinrichs SH and Sayood K (2011). RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics*, 12:41

Natalie JWW (2013). Determining the microbial diversity in chaparral soils before and after wildfires through DNA barcoding. Project Number-S1119, California State Science Fair, Project Summary

Nicol GW and Schleper C (2006). Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle?. *Trends in Microbiolgy*, 14, pp. 207-212

Niedringhaus TP, Milanova D, Kerby MB, Snyder MP and Barron AE (2011). Landscape of Next-Generation Sequencing Technologies. *Analytical Chemistry*, 83, pp. 4327-4341

Noguchi H, Park J and Takagi T (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research*, 34, pp. 5623-5630

Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvantidis C and lliopoulas I (2015). Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. *Bioinformatics and Biology Insights*, 9, pp. 75-88

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N and Larsson KH (2008). Intraspecific ITS variability in the Kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics*, 4, pp. 193-201

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Vonstein SHV, Wattam AR, Xia F and Stevens R (2014). The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Research*, 42, pp. D206-D214

Pace NR (1997). A molecular view of microbial diversity and the biosphere. *Science* 276, pp. 734-740

Pace NR, Stahl DA, Lane DJ and Oslen GJ (1986). The analysis of natural microbial populations by ribosomal RNA sequences. *Advances in Microbial Ecology*, 9, pp. 1-55

Pareek CS, Smoczynski R and Tretyn A (2011). Sequencing technologies and genome sequencing. Journal of Applied Genetics, 52, pp. 413-435

Pati A, Heath LS, Kyrpides NC and Ivanova N (2011). ClaMS: a classifier for metagenomic sequences. *Standards in Genomic Sciences*, 5, pp. 248-253

Peabody MA, Van Rossum T, Lo R and Brinkman FSL (2015). Evaluation of shotgun metagenomics sequence classification methods using *in silico* and *in vitro* simulated communities. *BMC Bioinformatics*, 16(363), doi: 1186/s12859-015-0788-5

Pochon X, Bott NJ, Smith KF and Wood SA (2013). Evaluating detection limits of Next-Generation sequencing for the surveillance and monitoring of international marine pests.

*PLoS ONE*, 8:e73935

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, Jensen LJ, von Mering C and Bork P (2014). Eggnog v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, 42, pp. D231-D239

Pride DT, Meinersmann RJ, Wassenaar TM and Blaser MJ (2003). Evoluntionary implications of microbial genome tetranucleotide frequency biases. *Genome Research,* 13, pp. 145-158

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, et al., (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature,* 464, pp. 59-65

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, et al., (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, pp. 228-232

Ranjan R, Rani A, Metwally A and McGee HS (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications,* 469, pp. 967-977

Rappé MS and Giovannoni SJ (2003). The uncultured microbial majority. *Annual Review of Microbiology*, 57, pp. 369-394

Ratnasingham S and Hebert PDN (2013). A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS ONE,* 8:e66213

Reva ON, Chan WY, Bezuidt OKI, Lapa SV, Safronova LA, Avdeeva LV, Borriss R. Genetic Barcoding of Bacteria and its Microbiology and Biotechnology Applications. In Bioinformatics and Data Analysis in Microbiology, Ed. O. Tastan Bishop, Caister Academic Press. 2014. pp. 230-243

Rho M, Tang H and Ye Y (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research,* 38:e191

Rice P, Longden I and Bleasby A (2000). EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, 16, pp. 276-277

Richter DC, Ott F, Auch AF, Schmid R and Huson DH (2008). MetaSim-a sequencing simulator for genomics and metagenomics. *PLoS One*, 3:e3373

Ronaghi M (2001). Pyrosequencing sheds light on DNA sequencing. *Genome Research*, 11, pp. 3-11

Ronaghi M, Uhlén M and Nyrén P (1998). A sequencing method based on realtime pyrophosphate. *Science*, 281, pp. 363-365

Rosen GL, Reichenberger ER and Rosenfeld AM (2010). NBC: the naïve bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, 27, pp. 127-129

Rosen MJ, Callan BJ, Fisher DS and Homles SP (2012). Denoising PCR-amplified metagenome data. *BMC Bioinformatics*, 13:283

Rothberg JM and Leamon JH (2008). The development and impact of 454 sequencing. *Nature Biotechnology*, 26, pp. 1117-1124

Ruby JG, Bellare P and Derisi JL (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3: Genes, Genome, Genetics*, 8, pp. 865-880

Sanger F, Coulson AR (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology, 94, pp. 441-446

Sanger F, Nicklen S and Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 74, pp. 5463-5467

Schadt EE, Turner S and Kasarskis A (2010). A window into third-generation sequencing. Human Molecular Genetics, 19, pp. R227-R240

Schloss PD and Handelsman J (2008). A statistical toolbox for metagenomics assessing functional diversity in microbial communities. *BMC Bioinformatics*, 9(34), doi: 10.1186/1471-2105-9-34

Schmieder R and Edwards R (2011a). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE*, 6:e17288

Schmieder R and Edwards R (2011b). Quality control and preprocessing of metagenomic

datasets. *Bioinformatics,* 27, pp. 863-864

Schreiber F, Gumrich P, Daniel R and Meinicke P (2010). Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics,* 26, pp. 960-961

Seshadri R, Kravitz SA, Smarr L, Gilna P and Frazier M (2007). CAMERA-a community resource for metagenomics. *PLoS Biology,* 5:e75

Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, Eisen JA and Pollard KS (2011). PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Computational Biology,* 7:e1001061

Sharpton TJ, Jospin G, Wu D, Langille MG, Pollard KS and Eisen JA (2012). Sifting through genomes with iterative-sequence clustering produces a large, phylogenetically diverse protein-family source. *BMC Bioinformatics,* 13:64

Shanahan F (2012). The microbiota in inflammatory bowel disease: friend, bystander, and sometime-villian. *Nutrition Reveiws,* 70, pp. S31-S37

Sharpton TJ (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science,* 5:209

Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, Kau AL, Rich SS, Concannon P, Mychaleckyj JC, Liu J, Houpt E, Li JV, et al., (2013). Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science,* 339, pp. 548-554

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM and Herndl GJ (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Preceedings of the National Academy of Sciences of the United States of America,* 103, pp. 12115-12120

Soo RM, Wood SA, Grzymski JJ, McDonald IR and Cary SC (2009). Microbial biodiversity of thermophilic communities in hot mineral soils of Tramway Ridge, Mount Erebus, Antarctica. *Environmental Microbiology,* 11, pp. 715-728

Stark M, Berger SA, Stamatakis A and von Mering C (2010). MLTreeMap-accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics,* 11:461

Stoeckle M (2003). Taxonomy, DNA, and the Barcode of life. *Bioscience,* 53, pp. 796-797

Suau A, Bonnet R, Sutren M, Gordon JJ, Gibson GR, Collins MD and Doré J (1999). Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and Environmental Microbiology,* 65, pp. 4799-4807

Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X and Mai V (2012). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Breifings in Bioinformatics,* 13, pp. 107-121N

Taylor HR and Harris WE (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources,* 21, pp. 377-388

Teeling H, Waldmann J, Lombardot T, Bauer M and Glöckner FO (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usuage patterns in DNA sequences. *BMC Bioinformatics,* 5:163

Thomas T, Gilbert J and Meyer F (2012). Metagenomics-a guide from sampling to data analysis. *Microbial Informatics and Experimentation,* 2:3

Trimble WL, Keegan KP, D'Sousa M, Wilke A, Wilkening J, Gilbert J and Meyer F (2012). Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC Bioinformatics,* 13(183), doi: 10.1186/1471-2105-13-83

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P and Rubin EM (2005). Comparative metagenomics of microbial communities. *Science,* 308, pp. 554-557

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, HeatH AC, et al., (2009). A core gut microbiome in obese and lean twins. *Nature,* 457, pp. 480-484

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS and Branfield JF (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature,* 428, pp. 37-43

Ultsch A and Moerchen F (2005). ESOM-Maps: tools for clustering visualization and classification with Emergent SOM Technical Report, Vol 46. Germany: Department of

Mathematics and Computer Science, University of Marburg, 2005.

Urwin R and Maiden MCJ (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology,* 11, pp. 479-487

van Belkum A, Struelens M, de Visser A, Verbrugh H and Tibayrenc M (2001). Role of genomic typing in taxonomy, evoluntionary genetics, and microbial epidemiology. *Clinical Microbiology Reveiws,* 14, pp. 547-560

Van Dijk EL, Auger H, Jasczczyszyn Y and Thermes C (2014). Ten years of next-generation sequencing technology. Trends in Genetics, 30, pp. 418-426

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, et al., (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science,* 304, pp. 66-74

Werner JJ, Zhou D, Caporaso JG, Knight R and Angenent LT (2011). Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *The ISME Journal,* 6, pp. 1273-1276

Weisburg WG, Barns SM, Pelletier DA and Lane DJ (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology,* 173, pp .697-703

White RA, Blainey PC, Fan HC, Quake SR (2009). Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics,* 10:116

Wilmes P and Bond PL (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends in Microbiology,* 14, pp. 92-97

Woese CR and Fox GE (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America.* 74, pp. 5088-5090

Womack KE, Bhavsar J and Ravel J (2008). Metagenomics: read length matters. *Applied and Environmental Microbiology,* 74, pp. 1453-1463

Wood SA, Smith KF, Banks JC, Tremblay LA, Rhodes L, Mountfort D, Cary SC, Pochon X (2013). Molecular genetic tools for environmental monitoring of New Zealand's aquatic habitats, past, present and the future. *New Zealand Journal of Marine and Freshwater*

*Research*, 47, pp. 90-119

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE and Banfield JF (2012). Fermentation, hydrogen, and sulphur metabolism in multiple uncultivated bacteria phyla. *Science*, 337, pp. 1661-1665

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A et al., (2009). A phylogeny-driven genomic encyclopaedia of bacteria and archae. *Nature*, 462, pp. 1056-1060

Wu M and Eisen JA (2008). A simple, fast and accurate method of phylogenomic inference. *Genome Biology*, 9:R151

Wu YW and Ye Y (2011). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, 18, pp. 523-534

Wylie KM, Truty RM, SharptonTJ, Mihindukulasuriya KA, Zhou Y, Gao H, Sodergren E, Weinstock GM and Pollard KS (2012). Novel bacteria taxa in the human microbiome. *PLoS ONE*, 7:e35294

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Doominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, et al., (2012). Human gut microbiome viewed across age and geography. *Nature*, 486, pp. 222-227

Yilmaz P, Kottman R, Field D, Knight R, Cole JR, Amaral-Zetter L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, *et al.*, (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) sepcifications. *Nature Biotechnology*, 29, pp. 415-420

Yok NG and Rosen GL (2011). Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics*, 12, doi: 10.1186/1471-2105-12-20

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, et al., (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biology*, 5:e16

Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E and DeRisi JL (2012). Virus identification in an unknown tropical febrile illness cases using deep sequences. *PLoS Neglected Tropical Diseases*, 6:e1485

Zaiko A, Samuiloviene A, Ardura A and Garcia-Vazquez E (2015). Metabarcoding approach for nonindigenous species surveillance in marine coastal waters. *Marine Pollution Bulletin,* 100, pp. 53-59

Zerbino DR and Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research,* 18, pp. 821-829

Zemanick ET, Sagel SD and Harris JK (2011). The airway microbiome in cystic fibrosis and implications for treatment. *Current Opinion in Pediatrics,* 23, pp. 319-324

Zhao Y, Tang H and Ye Y (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics,* 28, pp. 125-126

Zhu W, Lomsade A and Borodovsky M (2010). *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Research,* 38:e132

Zoetendal EG, Akkermans AD and De Vos WM (1998). Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Applied and Environmental Microbiology,* 64, pp. 3854-3859

# CHAPTER 2 Selection of reference genomes of microorganisms for case studies and design of the BarcodeGenerator: A novel software tool for generation of diagnostic barcode sequences


## Abstract

Different microorganisms are of important use in medicine and biotechnology. Despite the diverse bioactivity of these microorganisms, they are almost indistinguishable by phenotype and by the 16S rRNA, which makes it challenging to trace and identify them in nature. In this chapter, a novel software online tool, BarcodeGenerator, was used for the creation of barcode sequences. This allows identification of most suitable marker sequences for DNA-based multi-local barcoding for identification and monitoring of medical and biotechnological strains.


## 2.1 Introduction

Microorganisms are found everywhere in nature. Different communities of microbes flourish in various environments, ranging from the human gut, the rhizosphere and unreceptive habitat such as acid runoff to geothermal hotsprings (Simmons *et al.,* 2008; Walter and Ley, 2011; Philippot *et al.,* 2013; Sharp *et al.,* 2014). Different studies of cultured microbes have shown that they are important constituents of these environments and offer vital ecosystem services (Arrigo, 2005; van der Heijden *et al.,* 2008). Microbiota are referred to as microbes that associate with a macroscopic host organism (Hooper *et al.,* 2012). Full understanding of a macroscopic organism's composition needs investigation of its microbiota. Regrettably, some microbes are extremely difficult to culture in the laboratory (Sharpton *et al.,* 2014).

Metagenomics is the study of the entire genomic content of a microbial community that bonds the three domains of life:: archea, bacteria and eukaryotes (Galagan *et al.,* 2005; Wylie *et al.,* 2012; Chun and Rainey, 2014; Kwong *et al.,* 2015; Land *et al.,* 2015). All DNA and RNA samples may be isolated from a microbial populace, skipping the step of cultivation. Then DNA reads can be obtained by massive parallel sequencing and identified by searching through databases of known reference for binning to known species or assigning them to a cluster of reads of unknown origin (Kulski, 2016). Microbial communities are isolated and studied from different environments, such as the aquatic and terrestrial environments, host-associated ecosystems and various human-engineered systems, such as those involved with food, water, waste production and agriculture (Bashir *et al.,* 2014; Kwong *et al.,* 2015).

Hospitals are no exception; they are a source of pathogenic microorganisms, of which those of most interest develop resistance to commonly used medical antibiotics, such as methicillin-resistant *Staphylococcus aureus*, multidrug-resistant *Mycobacterium tuberculosis* and others. Hence, NGS is seen as a vital growing application for epidemiological studies of various pathogens (Grad and Lipsitch, 2014).

Next-generation sequencing and its application in metabarcoding and metagenomics enable researchers to study a complex microbial population without isolation and cultivation of strains. The most common method is the amplification of fragments of 16S rRNA from whole DNA samples using universal primers, followed by massive parallel sequencing of the amplified fragments (Wang and Qian, 2009). Genes encoding 16S rRNA were the earliest metagenomic targets used for identification of different species in environments before the first NGS microbial studies were performed using Roche 454 pyrosequencing and Illumina platforms. New opportunities and the benefits of application of NGS in metabarcoding were demonstrated in numerous publications following the introduction of these technologies: metabarcoding of mining sites and the surface waters of the gulfs, seas, and oceans are but a few projects benefitting from NGS (Gilbert and Dunpont, 2011; Wylie *et al.*, 2012). However, a limitation of classical 16S rRNA metabarcoding consists in the fact that the many microorganisms are almost indistinguishable by 16S rRNA (Safronova *et al.*, 2012). In general, the level of sensitivity of 16S rRNA metabarcoding is the taxonomic level of genera; above that sensitivity is insufficient to distinguish between microorganisms showing different activities in terms of biotechnology or medical practice.

Kress and Erickson (2008) defined DNA barcoding as a fast and robust technique for specie identification based on marker nucleotide sequences (Kress and Erickson, 2008). However, since the single-gene technique of DNA barcoding does not differentiate between closely related species and sub-species, it is of limited importance to develop diagnostic sets of marker sequences for biotechnological and medical microbiology (van Belkum *et al.*, 2001; Urwin and Maiden, 2003; Reva *et al.*, 2014). Hence, it was hypothesised that the comparison of bacterial strains by using multiple gene sequences would give better resolution of their core relationships than a single gene (Maiden *et al.*, 1998). The MLST technique was introduced, which made use of DNA sequences of internal fragments of multiple housekeeping genes for definitive identification of microorganisms (Maiden *et al.*, 1998; Urwin and Maiden, 2003). Various researchers have developed different techniques for MLST, some of which include rMLST, multi-locus sequence analysis (MLSA) and whole

genome MLST (wgMLST) (Jolley *et al.*, 2012; Glaser and Kämpfer, 2015; Katz *et al.*, 2017). The rMLST technique indexes variations in 53 genes encoding bacterial ribosome protein subunits (*rps* genes) as a way of incorporating microbial genealogy and typing. Grouping provided by rMLST was consistent with the present nomenclature systems independently of the clustering algorithm being used (Jolley *et al.*, 2012). The MLSA technique is used to obtain better differentiation of species within a genus. Partial sequences of genes coding for housekeeping genes are used to create phylogenetic trees, but can also be used as taxonomic markers in MLSA research. The MLSA technique has also been suggested as a replacement for DNA-DNA hybridisation in species delineation (Glaser and Kämpfer, 2015). The two basic techniques used for species delineation by WGS are wgMLST and single nucleotide polymorphisms (SNPs). In particular, these approaches were used to survey outbreaks of pathogens (Glaser and Kämpfer, 2015). As with the traditional MLST, alleles in wgMLST are either the same as in reference or different, which implies that any nucleotide substitution, insertion or deletion is equivalent to one allele change. In wgMLST, several thousand loci can be matched and estimated distances between them then are used either for species or strain delineation, or to infer phylogenetic relationships by clustering algorithms. For the SNP technique, counts of single nucleotide substitutions are used to deduce phylogenetic relatedness or genetic typing. SNP protocols have been implemented in various software packages (Katz *et al.*, 2017).

Multi-locus sequence typing approaches were also promoted by the advance in NGS. Different software applications have been developed, using various techniques to calculate the sequence types (STs) from the NGS data. However, not all MLST calling applications function as required. Challenges encountered with these programs include (i) computationally inefficient methods; (ii) false positive results; (iii) out-of-date databases; (iv) inability to call alleles from low coverage reads; and (v) variable performance of mixed samples. Hence, there is room for improvement (Page *et al.*, 2017).

The work hypothesis of this study was that several limitations of the traditional metabarcoding and MLST/wgMLST approaches can be overcome by creating a standardised computational approach (BarcodeGenerator) for a dynamic selection of marker genes for multi-locus barcoding depending on research tasks and the taxonomic level of separation specified by users. It is furthermore of interest to investigate functional categories of genes showing the best performance in metabarcoding of different groups of microorganisms. To

evaluate the performance of selected metabarcode sequences, publicly available metagenomic datasets from NCBI and MG-RAST were used.

### 2.2 Selection of microorganisms for case studies

This project aimed at developing a new software tool for an automated selection of marker genes to distinguish between various microorganisms at different levels of taxonomic relatedness. Different microorganisms were used for case studies in this work*: Bacillus cereus, Escherichia* and *Shigella, Lactobacillus, Mycobacteria, Prochlorococcus, Salmonella Shewanella* and *Streptococcus.* The strains used represent different species and subspecies, including pathogenic and biotechnological strains.

#### 2.2.1 Bacillus

The genus *Bacillus* comprises rod-shaped, endospore-forming bacteria that belong to the phylum *Firmicute* (Rooney *et al.*, 2009). The species of this genus have an abundant spread in nature and are found in practically every environment (Rooney *et al.*, 2009; Alina *et al.*, 2015). They are actively involved in carbon and nitrogen cycling, while species such as *Bacillus anthracis* and *Bacillus cereus* are known human and livestock pathogens. Since most of these *Bacillus* species are non-pathogenic, they have frequently been used in biotechnological and industrial applications (Bischoff *et al.*, 2006; Price *et al.*, 2007; Rooney *et al.*, 2009). *Bacillus cereus* is an opportunistic pathogen that causes food poisoning, expressed by diarrhoeal or emetic syndromes. *Bacillus cereus* is closely related to *Bacillus anthracis* and the insect pathogen *Bacillus thuringiensis* (Ivanova *et al;* 2003). *Bacillus anthracis* is used as a biological weapon, while *Bacillus thuringienis* is used as pesticide. *Bacillus anthracis* and *Bacilllus thuringienis* do have plasmid-borne specific toxins and this fact is usually used to differentiate them from *Bacillus cereus* (Ivanova *et al.*, 2003). Table 2.1 shows the different strains and species of *Bacillus* used in this study.

**Table 2.1: Strains of *Bacillus* used in this study**

| STRAINS | NCBI ID |
|---|---|
| *Bacillus cereus* ATCC 14579 | NC_004722 |
| *Bacillus anthracis* str.A0248 | NC_012659 |
| *Bacillus cereus* F837/76 | NC_016779 |
| *Bacillus thuringiensis serovar chinensis* | NC_017208 |
| *Bacillus thuringiensis serovar konkukian* str. 97-27 | NC_005957 |
| *Bacillus cereus* B4264 | NC_011725 |
| *Bacillus thuringiensis* str. Al Hakam | NC_008600 |
| *Bacillus anthracis* str. Ames Ancestor | NC_007530 |

| | |
|---|---|
| *Bacillus cereus* AH820 | NC_011773 |
| *Bacillus weihenstephanensis* KBAB4 | NC_010184 |
| *Bacillus cereus* NC7401 | NC_016771 |
| *Bacillus cereus biovar anthracis* str. CI chromosome | NC_014335 |
| *Bacillus cereus* ATCC 10987 | NC_003909 |
| *Bacillus cereus* G9842 chromosome | NC_011772 |
| *Bacillus anthracis* str. Sterne chromosome | NC_005945 |
| *Bacillus thuringiensis* BMB171 | NC_014171 |
| *Bacillus cereus* AH187 | NC_011658 |
| *Bacillus cereus* E33L | NC_006274 |
| *Bacillus cereus* 03BB102 | NC_012472 |
| *Bacillus anthracis* str. Ames chromosome | NC_003997 |
| *Bacillus cereus* Q1 | NC_011969 |
| *Bacillus anthracis* str. CDC 684 | NC_012581 |
| *Bacillus thuringiensis serovar finitimus* YBT-020 | NC_017200 |

### 2.2.2 Escherichia coli and Shigella

*Escherichia coli* belongs to the *Enterobacteriaceae* family; it is gram-negative, rod-shaped and oxidase-negative. *E. coli* can be either non-motile or motile, with a petritirichious flagella. It can grow anaerobically or aerobically, preferably at $37^{o}$ C. It is readily isolated from faecal samples (Croxen *et al.,* 2013), has been described as a known commensal of the gastrointestinal tract in warm-blooded animals and is used as the everyday laboratory mainstay. However, pathogenic *E. coli* has also been reported, which causes human diseases ranging from those affecting the gastrointestinal tract to extra-intestinal sites such as the urinary tract, bloodstream and the central nervous system (Kaper *et al.,* 2004; Croxen and Finlay 2010; Croxen *et al.,* 2013). Though various aetiological agents have been reported as the cause of diarrhoea, pathogenic *E. coli* stands out among others as a major cause. A case control study aimed at understanding the burden of paediatric diarrhoeal disease in Sub-Saharan Africa and South Asia (Levine *et al.,* 2012) reported by the Global Enteric Multi-Centre study showed that entrotoxigenic *E. coli* and *Shigella* are two of the four causative agents for moderate to severe diarrhoea among children in these areas (Kotloff *et al.,* 2013). The main diarrhoeagenic *E. coli* phenotypes are: (i) enteropathogeni *E. coli*; (ii) Shiga toxin producing *E. coli*; (iii) *Shigella*/entero-invasive *E. coli* ; (iv) enteroaggregative *E. coli*; (v) diffusely adherent *E. coli*; (vi) enterotoxigenic *E. coli;* and (vii) adherent invasive *E. coli* (Croxen *et al.,* 2013).

Entero-invasive *E.coli/Shigella* spp are described as facultative intracellular pathogens and the aetiological agents of bacillary dysentery, also known as shigellosis. *Bacillus dysenteriae,* also called *Shigella,* was first identified in 1897 by Kiyoshi Shiga during an epidemic in

Japan, where it infected more than 91 000 people, causing a mortality rate of more than 20% (Trofa *et al.,* 1999). Fifty years later, EIEC was identified as having similar biochemical, genetic and pathogenic functions as *Shigella* (Lan *et al.,* 2004). *Shigella* is a non-motile, lysine decarboxylase negative microorganism that does not ferment lactose, with the exception of *S. sonnei,* which is a slow lactose fermenter (Scheutz and Strockbine, 2005; Strockbine and Maurelli, 2005). Conventionally, *Shigella* is classified based on biochemical, serological and clinical phenotypes and not on the phylogenetic relationship (Ewing 1949; Strockbine and Maurelli, 2005, Kalluri *et al.,* 2004). This comprises 49 sero- and subserotypes that are further clustered into four species: (i) *S. dysenteriae* (sero A, 15 serotypes); (ii) *S. flexneri* (serogroup B, 14 sero- and subserotypes); (iii) *S. boydii* (serogroup C, 19 serotypes) and (iv) *S. sonnei* (serogroup D, one serotype). *S. boydii* was formerly subdivided into 20 serotypes, but the phylogenetic analysis showed that *S. boydii* 13 fits into the *E. albertii* lineage, which is quite distinct from the typical *Shigella* (Croxen *et al.,* 2013). Table 2.2 shows the strains and species of *Escherichia coli* and *Shigella* used in this study.

**Table 2.2: Strains of *Escherichia coli* and *Shigella* used in this study**

| Strains | NCBI: ID |
|---|---|
| *Shigella boydii* CDC 3083-94 | NC_010658 |
| *Escherichia coli* CFT073 | NC_004431 |
| *Escherichia coli* O157:H7 | NC_002655 |
| *Escherichia coli* O111:H- str | NC_013364 |
| *Shigella flexneri* 2a str | NC_004741 |
| *Escherichia coli* UT189 | NC_007946 |
| *Shigella boydii* Sb227 | NC_007613 |
| *Escherichia fergusonii* ATCC | NC_011740 |
| *Escherichia coli* O26:H11 | NC_013361 |
| *Escherichia coli* O157:H7 | NC_011353 |
| *Escherichia coli* BL21-Gold | NC_012947 |
| *Escherichia coli* APEC O1 | NC_008563 |
| *Escherichia coli* SE11 | NC_011415 |
| *Escherichia coli* 55989 | NC_011748 |
| *Escherichia coli* IAI1 | NC_011741 |
| *Esherichia coli* E24377A | NC_009801 |
| *Escherichia coli* O157:H7 | NC_013008 |
| *Escherichia coli* 536 | NC_008253 |
| *Escherichia coli* UMNO26 | NC_011751 |
| *Escherichia coli* ED1a | NC_011745 |
| *Escherichia coli* 055:H7 | NC_013941 |
| *Escherichia coli* SMS-3-5 | NC_010498 |
| *Escherichia coli* O103:H2 | NC_013353 |
| *Escherichia coli* IA139 | NC_011750 |
| *Escherichia coli* HS | NC_009800 |
| *Escherichia coli* B str | NC_012967 |
| *Escherichia coli* str.K -12 substr. | NC_010473 |
| *Escherichia coli* O157:H7 Str. Sakai | NC_002695 |

| | |
|---|---|
| *Shigella dysenteriae* Sd197 | NC_007606 |
| *Escherichia coli* S88 | NC_011742 |
| *Escherichia coli* O127:H6 str.E2348/69 | NC_011601 |
| *Shigella flexneri* 2a str. 301 | NC_004337 |
| *Escherichia coli* BW2952 | NC_012759 |
| *Shigella sonnei* 53G | NC_016822 |
| *Shigella sonnei* Ss046 | NC_007384 |
| *Escherichia coli* ATCC 8739 | NC_010468 |

### 2.2.3 Lactobacillus

*Lactobacillus* is described as the largest genus of the lactic acid bacteria group, which comprises 50 species in total. These are found in the oral, vaginal and intestinal regions of most animals. *Lactobacillus* is used in the production of cheese (Blaiotta *et al.,* 2001), yogurt (Omogbai *et al.,* 2005), bacteriocin (Vuyst and Leroy, 2007) and other products because it produces lactic acid, which prevents the growth of other organisms as well as dropping the pH of food products (Thavasi *et al.,* 2011). *Lactobacillus* has also been reported to be used as probiotics, prebiotics (Teitelbaum and Walker, 2002) and biotherapeutics (Buddington, 2009). Lactic acid bacteria are mostly seen in various natural environments and are characterised by precise *lactobacilli* compositions such as *L. acidophilus* and *L. delbrueckii* spp. *bulgaricus. Lactobacillus helveticus* are the classic representatives of the micro-flora of fermented milk products such as yoghurt and kefir, while the *L. casei* group comprising *L. casei, L. paracasei* and *L. rhamnosus* can be found in various types of cheese (Bouton *et al.,* 2002; Markiewicz *et al.,* 2010). *Latobacillus delbrueckii* has also been illustrated as a strain producing biosurfactants and crude oil biodegrading compounds (Thavasi *et al.,* 2006). *Lactobacillus* is furthermore known to help prevent infections of the urogenital and intestinal tract. The dominance of *Lactobacillus* in the vagina is linked with a reduced risk of bacterial vaginosis and urinary tract infections. Hence, the instillation of *Lactobacillus* GR-1 and B-54 or RC-14 strains into the vagina has been reported to reduce the risk of urinary tract infections and improve the maintenance of the normal flora (Reid and Burton, 2002). Table 2.3 shows the *Lactobacillus* strains and species used in this study.

**Table 2.3: Strains of *Lactobacillus* used in this study**

| STRAINS | NCBI:ID |
|---|---|
| *Lactobacillus casei* str. Zhang | NC_014334 |
| *Lactobacillus acidiphilus* NCFM | NC_006814 |
| *Lactobacillus rhamnosus* Lc 705 | NC_013199 |
| *Lactobacillus reuteri* DSM 20016 | NC_009513 |
| *Lactobacillus delbrueckii* subsp. bulgaricus ATCC 11842 | NC_008054 |

| | |
|---|---|
| *Lactobacillus fermentum* IFO 3956 | NC_010610 |
| *Lactobacillus acidophilus* 30SC | NC_015214 |
| *Lactobacillus salivarius* UCC118 | NC_007929 |
| *Lactobacillus casei* BL23 | NC_010999 |
| *Lactobacillus sanfranciscensis* TMW | NC_015978 |
| *Lactobacillus reuteri* SD2112 | NC_015697 |
| *Lactobacillus kefiranofaciens* ZW3 | NC_015602 |
| *Lactobacillus crispatus* ST1 | NC_014106 |
| *Lactobacillus rhamnosus* GG | NC_013198 |
| *Lactobacillus johnsonii* NCC 533 | NC_005362 |
| *Lactobacillus gasseri* ATCC 33323 | NC_008530 |
| *Lactobacillus brevis* ATCC 367 | NC_008497 |
| *Lactobacillus delbruecki* subsp. bulgaricus ND02 | NC_014727 |
| *Lactobacillus reuteri* JCM 1112 | NC_010609 |
| *Lactobacillus buchneri* NRRL B-30929 | NC_015428 |
| *Lactobacillus helveticus* DPC 4571 | NC_010080 |
| *Lactobacillus delbruecki* subsp. bulgaricus ATCC BAA-365 | NC_008529 |
| *Lactobacillus amylovorus* GRL 1112 | NC_014724 |
| *Lactobacillus sakei* subsp. sakei 23k chromosome | NC_007576 |
| *Lactobacillus plantarum* WCFS1 | NC_004567 |
| *Lactobacillus ruminis* ATCC 27782 | NC_015975 |
| *Lactobacillus plantarum* subsp. plantarum ST-III | NC_014554 |
| *Lactobacillus johnsonii* F19785 | NC_013504 |

### 2.2.4 Mycobacteria

The family *Mycobacteriacea* is made up of only one genus, *Mycobacterium. Mycobacterium* species are thin, slightly curved to straight bacilli, forming no spores and non-motile. Cells of mycobacteria are 0.2 to 6 µm x 1.0 to 10 µm (Eisenstadt, 1995). The genus is made up of more than 50 species (Wayne *et al;* 1992). Most mycobacterial species are ubiquitous and can be found in water, soil, food and vegetation. *M. bovis* infection has been developed by consuming unpasteurised milk. Bacilli Calmette-Guérin, which is a strain of *M. bovis*, is widely used for immunisation against tuberculosis. It is also administered as an immunotherapeutic agent for the treatment of superficial bladder carcinoma or melanoma. *Mycobacterium fortuitum* has been described as a normal flora of the skin (Eisenstadt, 1995). Pathogenic isolates of *Mycobacterium* include (i) *M. tuberculosis* ─ the causative agent of human tuberculosis; (ii) *M. bovine* ─ the causative agent of bovine tuberculosis; (iii) *M. leprae* ─ the causative agent of leprosy; (iv) *M. ulcerans,* which causes Buruli ulcers and is the third most common form of mycobacterial disease in humans; and (v) *M. marium* ─ the causative agent of fish tank granuloma in humans and granulomatous lesions similar to those of *M. tuberculosis* in zebra fish (Demangel *et al.,* 2009; Rahman *et al.,* 2014). The non-pathogenic groups are *M. gilvum*, *M. vanbaalenii* and *M. smegmatis* (Raham *et al.,* 2014). Opportunistic pulmonary infections are mostly caused by members of the *Mycobacterium*

*avium* complex (MAC) that includes *M. avium* and *M. avium-M. intracellulare*, while Crohn's disease in humans is suspected to be caused by the third member of the MAC group, *Mycobacterium avium* subsp. *paratuberculosis* (Cook, 2010; Chiodini *et al.*, 2012).

Seven strains of *M. leprae* have been well characterised, namely India2, Thai53, TN, Africa, NHDP63, NHDP98 and Br4923. India2, Thai53 and TN are of SNP type 1, Africa clade is of SNP type 2, NHDP63 and NHDP98 are of SNP type 3 while Br4923 is of SNP type 4 (Monot *et al.*, 2009; Akinola *et al.*, 2013). *Mycobacterium smegmatis* is an aerobic fast-growing non-pathogenic *Mycobacterium*, which has similar features with pathogenic mycobacteria (Cordone *et al.*, 2011). It can adjust to micro-aerobiosis by changing from the active growth to dormant or latent stages. *Mycobacterium smegmatis* is mainly valuable in understanding the cellular processes that are significant to pathogenic mycobacteria such as *M. leprae*, *M. tuberculosis* and *M. avium* subsp. *paratuberculosis* (He and De Buck, 2010; Akinola *et al.*, 2013). Table 2.4 shows the *Mycobacteria* strains and species used in this study.

**Table 2.0.4: Strains of *Mycobacteria* used in this study**

| STRAINS | NCBI:ID |
|---|---|
| *Mycobacterium leprae* | NC_002677 |
| *Mycobacterium marinum* M | NC_010612 |
| *Mycobacterium gilvum* PYR-GCK | NC_009338 |
| *Mycobacterium* sp. KMS | NC_008705 |
| *Mycobacterium ulcerans* Agy99 | NC_008611 |
| *Mycobacterium sp.* JDM601 | NC_015576 |
| *Mycobacterium smegmatis* str. MC2 155 | NC_008596 |
| *Mycobacterium* sp. MOTT36Y | NC_017904 |
| *Mycobacterium avium* 104 | NC_008595 |
| *Mycobacterium abscessus* | NC_010397 |
| *Mycobacterium tuberculosis* H37Rv | NC_000962 |
| *Mycobacterium intracellulare* MOTT-02 | NC_016947 |
| *Mycobacterium sp.* JLS | NC_009077 |
| *Mycobacterium* sp. MCS | NC_008146 |
| *Mycobacterium avium* subsp. paratuberculosis k-10 | NC_002944 |
| *Mycobacterium vanbaalenii* PYR-1 | NC_008726 |

### *2.2.5 Prochlorococcus*

*Prochlorococcus* is described as the minutest (< 1 µm) in diameter and most abundant (3 x $10^{27}$ cells) photosynthetic microorganism on planet earth. *Prochlorococcus* is a unicellular marine cyanobacterium, which is found throughout the euphotic zone of open ocean between $45^0$ N and $40^0$ S, where it carries out a significant portion of global photosynthesis (Partensky *et al.*, 1999; Flombauum *et al.*, 2013; Biller *et al.*, 2014). The genome of *Prochlorococcus* is

the smallest of any known free-living photosynthetic cells, ranging from 1.6 to 2.7 Mbp (Kettler *et al.,* 2007). Though the core set of genes present is shared by all strains, a notable diversity in gene content was reported among isolates. The *Prochlorococcus* group has an open pan-genome such that each newly sequenced genome typically contains various novel genes never identified before (Kettler *et al.,* 2007). Research on the genomic and metagenomic features has provided wide understanding of the features of ocean ecosystems (Rocap *et al.,* 2003; Martiny *et al.,* 2009; Coleman and Chisholm, 2010), microbial evolution (Zhaxybayeva *et al.,* 2009; Baumdicker *et al.,* 2012) and the intrinsic relationships between genotype, phenotypic and ecological variations in marine populations (Moore *et al.,* 1998; Zinser *et al.,* 2007; Kashtan *et al.,* 2014). Using *Prochlorococcus* genomes as reference sequences has also been tremendously valuable for interpreting marine metagenomic and metatranscriptomic datasets (Venter *et al.,* 2004; Friaz-Lopex *et al.,* 2008; Poretsky *et al.,* 2009; Rusch *et al.,* 2010; Biller *et al.,* 2014).

*Prochlorococcus* isolates can be subdivided into taxonomically and ecologically distinguishable lineages based on their adaptation to general environmental settings such as high-light (HL) versus low-light (LL) (Moore *et al.,* 1998) or temperate regimes (Johnson *et al.,* 2006). Across the globe, warm surface water is mostly dominated by a particular clade of *Prochlorococcus*, namely eMIT9312. The eMIT9312 clade has a high rate of sequence divergence while upholding notable conservation both in gene content and synteny, notwithstanding worldwide distribution that spans huge gradients in the bioavailable concentrations of the macronutrients nitrogen (N) and phosphorous (P). Table 2.5 shows the strains of *Prochlorococcus marinus* used in this study.

**Table 2.5: Strains of *Prochlorococcus* used in this study**

| STRAINS | NCBI:ID |
|---|---|
| *Prochlorococcus marinus* str.MIT 9301 | NC_009091 |
| *Prochlorococcus marinus* str. MIT 9312 | NC_007577 |
| *Prochlorococcus marinus* str. NATL1A | NC_008819 |
| *Prochlorococcus marinus* str. MIT 9313 | NC_005071 |
| *Prochlorococcus marinus* str. MIT 9215 | NC_009840 |
| *Prochlorococcus marinus* str. AS9601 | NC_008816 |
| *Prochlorococcus marinus* str. MIT 9303 | NC_008820 |
| *Prochlorococcus marinus* subsp. pastoris str. CCMP1986 | NC_005072 |
| *Prochlorococcus marinus* str. NATL2A | NC_007335 |
| *Prochlorococcus marinus* str. MIT 9211 | NC_009976 |
| *Prochlorococcus marinus* subsp. marinus str. CCMP1375 | NC_005042 |
| *Prochlorococcus marinus* str. MIT 9515 | NC_008817 |

### 2.2.6 Salmonella

The genus *Salmonella* belongs to the *Enterobacteriaceace* family. *Salmonella* are rod-shaped facultative anaerobes. The genus *Salmonella* is divided into typhoidal serotypes *Salmonella enterica* var. Typhi (*S. typhi*); and *Salmonella enterica* var. Paratyphi (*S. paratyphi A*), as well as multiple non-typhoidal *Salmonella* serotypes usually called the NTS serotypes (Feasey *et al.,* 2012). In developed countries, non-typhoidal *Salmonella* are mostly the causative agents of self-limiting diarrhoea in people, while bloodstream or focal infections are usually uncommon (Laupland *et al.,* 2010) and mostly occur in individuals with particular risk factors (Gordon, 2008). However, in sub-Saharan Africa, non-typhoidal *Salmonella* are the most common bloodstream isolates in children and adults presenting with fever (Gilks *et al.,* 1990; Gordon *et al.,* 2008; Reddy *et al.,* 2010) and are usually linked to a mortality rate of 20-25% (Feasey *et al.,* 2012).

Whole-genome sequencing of pathogens, immunological trials and characterisation of bacteria-host interactions at the cellular, humoral and mucosal level helped to generate a comprehensive view on the evolution and emergence of this pathogen (Feasey *et al.,* 2012). *Salmonella typhimurium* or *Salmonella enterica* var Enteritidis (*S. enteritidis*), which are non-typhoidal *Salmonella*, have been reported to be the major cause of disease across Africa (Feasey *et al.,* 2010). Researchers have also reported disease outbreaks associated with the following serotypes: (i) *Salmonella enterica* var lsangi (*S. isangi*) in South Africa (Wadula *et al.,* 2006); (ii) *Salmonella enterica* var concord (*S. concord*) in Ethopia (Beyene *et al.,* 2011); and (iii) *Salmonella enterica* var Stanleyville (*S. stanleyville*) and *Salmonella enterica* var Dublin (*S. dublin*) in Mali (Tennant *et al.,* 2010). Non-typhoidal *Salmonella* has been established as a major HIV-related pathogen in sub-Saharan African adults (Gilks *et al.,* 1990). While the non-typhoidal *Salmonella* have a broad range of hosts among humans and animals, the typhoidal serotypes *S. typhi* and *S. paratyphi A* are totally host-constrained to people, causing invasive disease in immune-competent individuals (Feasey *et al.,* 2012). Table 2.6 shows strains and species of *Salmonella* used in this study.

**Table 2.6: Strains of *Salmonella* used in this study**

| STRAINS | NCBI:1D |
|---|---|
| *Salmonella enterica* subsp.enterica serovar Gallinarium str. 287/91 | NC_011274 |
| *Salmonella enterica* subsp.enterica serovar Typhimurium str. LT2 | NC_003197 |
| *Salmonella enterica* subsp.enterica serovar Choleraesuis str.SC-B67 | NC_006905 |
| *Salmonella enterica* subsp.enterica serovar Schwarzengrund str.CVM19633 | NC_011094 |
| *Salmonella enterica* subsp.enterica serovar Typhimurium str.14028S | NC_016856 |

| | |
|---|---|
| *Salmonella bongori* NCTC 12419 | NC_015761 |
| *Salmonella enterica* subsp.enterica serovar Newport str. SL254 | NC_011080 |
| *Salmonella enterica* subsp.enterica serovar Dublin str. CT_02021853 | NC_011205 |
| *Salmonella enterica* subsp.enterica serovar Typhi str. CT18 | NC_003198 |
| *Salmonella enterica* subsp.enterica serovar Paratyphi A str. ATCC 9150 | NC_003198 |
| *Salmonella enterica* subsp.enterica serovar Typhimurium str. SL1344 | NC_016810 |
| *Salmonella enterica* subsp.enterica serovar Typhi str.P-stx-12 | NC_016810 |
| *Salmonella enterica* subsp.enterica serovar Gallinarum/pullorum str. RKS5078 | NC_016831 |
| *Salmonella enterica* subsp.enterica serovar Typhimurium str.T000240 | NC_016860 |
| *Salmonella enterica* subsp.enterica serovar Typhimurium str.UK-1 | NC_0166863 |
| *Salmonella enterica* subsp.enterica serovar Typhimurium str. ST4/74 | NC_016857 |
| *Salmonella enterica* subsp.enterica serovar Heidelberg str. SL476 | NC_011083 |
| *Salmonella enterica* subsp.enterica serovar Paratyphi A str. AKU_12601 | NC_011147 |
| *Salmonella enterica* subsp.enterica serovar Typhi str. Ty2 | NC_004631 |
| *Salmonella enterica* subsp.enterica serovar Agona str. SL483 | NC_011149 |
| *Salmonella enterica* subsp.enterica serovar Paratyphi C strain RKS4594 | NC_012125 |
| *Salmonella enterica* subsp.enterica serovar Typhimurium str. 798 | NC_017046 |

### 2.2.7 Shewanella

*Shewanella* are facultative anaerobes, Gram-negative gamma-proteobacteria found in varied kinds of environments, but mostly in marine sediments and frequently in association with fish (Ivanova *et al.,* 2004; Dikow 2011; Wright *et al.,* 2016). *Shewanella* species signify a vital family of dissimilatory metal-reducing bacteria, which enables the transmission of metabolically produced electrons from a cell interior to external electron acceptors such as solid metal oxides during anaerobic respiration (Fredrickson *et al.,* 2008; Wang *et al.,* 2011). The metal-reducing capability of *Shewanella* has been credited mostly to a set of surface-linked Mtr/Omc proteins, such as three outer membrane decaheme *c*-type ctytochromes, MtrC, MtrF and OmcA; two periplasmic decaheme cytochromes, Mtr and MtrD; and two outer membrane non-cytochrome proteins, MtrB and MtrE (Meyers *et al.,* 1997; Beliaev *et al.,* 1998; Pitts *et al.,* 2003; Wang *et al.,* 2011). These proteins are known to work jointly to transmit electrons towards the outside of the cell outer membrane and are highly conserved across the *Shewanella* genus. At the cell interface, at least two unique extracellular electron transfer (EET) pathways have been identified, namely the direct charge transfer from the cell surface and the use of self-secreted soluble redox mediators (Beliaev *et al.,* 2001; Pitts *et al.,* 2003). The unique EET systems of *Shewanella* have been comprehensively studied in various biotechnological applications such as bioremediation (Hau and Gralnick, 2007), heavy metal detoxification (Liu *et al.,* 2002; Hau and Gralnick, 2007) and electricity production in microbial fuel cells (Bretscher *et al.,* 2007; Wang *et al.,* 2011).

*Shewanella oneidensis* MR-1 is a recognised dissimilatory metal-reducing bacterium with a distinct respiration design. It has modular electron transport pathways and a huge number of terminal reductases to reduce ferric oxides, manganese oxides, nitrate fumarate, sulphur oxyanions, dimethyl sulphoxide and trimetlylamine oxide (Heidelberg *et al;* 2002; Fredrickson *et al.,* 2008; Li *et al.,* 2013). The study of *S. oneidensis* genome suggests that there is a vastly varied electron transport system comprising 42 putative *c*-type cytochromes essential in the reduction of chromate, cobalt (iii), vanadium (v) and uranium (vi) salts and oxides (Meyers *et al.,* 2004; Hau *et al.,* 2008; Belchik *et al.,* 2011; Li *et al.,* 2013). Flavin mononucleotide and riboflavin secreted by *S. oneidensis* MR-1 have also been reported to increase the bioreduction of extracellular electron acceptors (Canstein *et al.,* 2008; Marsili *et al.,* 2008; Liu *et al.,* 2013).

Wright and colleagues (2016) tested a number of *Shewanella* strains for their manganese oxidising capacity in aerobic conditions (Wright *et al.,* 2016). *Shewanella loihica* strain PV-4 was reported as the strongest oxidiser, producing oxides at a rate of 20.3 mg/litre/day and oxidising Mn(II) in concentrations of up to 9 mM. Analysis of compounds produced by *S. loihica* PV-4 and another strong oxidiser, *S. putrefaciens* CN-32, identified finely grained nanosize oxide particles with an identical Mn oxidation state of 3.86. By contrast, the strain *S. oneidensis* MR-1 was the weakest oxidiser of all tested *Shewanella* (Wright *et al.,* 2016). Table 2.7 shows the strains and species of *Shewanella* used in this study.

**Table 2.7: Shows strains of *Shewanella* used in this study**

| STRAINS | NCBI: ID |
|---|---|
| *Shewanella* sp. MR-7 | NC_008322 |
| *Shewanella* sp MR-4 | NC_008321 |
| *Shewanella oneidensis* MR-1 | NC_004347 |
| *Shewanella baltica* OS155 | NC_009052 |
| *Shewanella woodyi* ATCC 51908 | NC_010506 |
| *Shewanella pealeana* ATCC 700345 | NC_009901 |
| *Shewanella baltica* OS223 | NC_011663 |
| *Shewanella frigidimarina* NCIMB 400 | NC_008345 |
| *Shewanella violacea* DSS12 | NC_014012 |
| *Shewanella halifaxensis* HAW-EB4 | NC_010334 |
| *Shewanella dentrificans* OS217 | NC_007954 |
| *Shewanella baltica* OS678 | NC_016901 |
| *Shewanella piezotelerans* WP3 | NC_011566 |
| *Shewanella sediminis* HAW-EB3 | NC_009831 |
| *Shewanella amazonensis* SB2B | NC_008700 |
| *Shewanella* sp W3-18-1 | NC_008750 |
| *Shewanella baltica* OS195 | NC_009997 |
| *Shewanella putrifaciens* CN-32 | NC_009438 |
| *Shewanella* sp. ANA-3 | NC_008577 |

| | |
|---|---|
| *Shewanella loihica* PV-4 | NC_009092 |
| *Shewanella baltica* OS185 | NC_009665 |

### 2.2.8 Streptococcus

The name *Streptococcus* originated from the Greek word "strepto", which means twisted, and "coccus", which means spherical. Over 100 species of *Streptococcus* have now been identified. According to the Lancefield system, streptococci were grouped by the carbohydrate composition of cell walls (Nobbs *et al.,* 2009). These substances are group-specific antigens, which belong to three different classes of chemical compounds: polysaccharides determining the groups A, B, C, E, F and G of Staphylococci, teichoic acids determining the groups D and N and lipoteichoic acid characteristic for the group H (Nobbs *et al.,* 2009). Streptococci can also be classified based on the 16S rRNA gene sequences (Kilian, 2005). The pyogenic group is made up of *S. pyogenes*, which is Lancefield group A; *S. agalactiae* and *S. uberis*, which are group B; *S. dysgalactiae* (group C, G or L) and *S. equi* (group C) (Nobbs *et al.,* 2009). Several groups are common among isolates from human oral and nasopharynx cavities, which include *S. oralis, S. mitis, S. gordon* and *S. pneumonia.* Because of intensive horizontal gene exchange between these organisms, phylogenetic relations between these species remain unclear (Humtsoe *et al.,* 2005; Bergmann and Hammerschmidt, 2006). This has led to the introduction of the anginosus and salivarius groups, which comprise mostly human and animal oral cavity isolates, and also the bovis group. Many species isolated from oral cavities of humans and animals remain unclassified: *S. mutans* and *S. soborinus* (human isolates); *S. downei* from macaques; *S. ratei* from rats and *S. criceti* from hamsters (Nobbs *et al.,* 2009). It is known that these bacteria are linked to the development of dental caries. Oral cavity microbes are usually referred to as viridans streptococci because of the greenish pigmentation produced by these bacteria when grown on blood agar. This reaction is often termed alpha-haemolysis and is suggestive of the presence of hydrogen peroxide production (Nobbs *et al.,* 2009).

*Streptococcus pneumonia* is a Gram-positive coccus and a member of the lactic acid bacteria, which has been described as a foremost source of morbidity and mortality worldwide. The World Health Organisation (WHO) reported that approximately 1 million children die of pneumococcal disease every year in third-world countries (Hoskin *et al.,* 2001; WHO/UNICEF, 2005; WHO, 2007). Pneumococcal infections have been reported to be the foremost cause of death from vaccine-preventable illnesses in children younger than five

years (CDC, 2006). Invasive diseases caused by pneumococci include meningitis and pneumonia associated with bacteraemia and emphysema. The risk factors for developing invasive pneumococcal disease (IPD) include age, with the highest risk of incidence among young children less than two years old and also elderly people older than 65 years; ethnicity and geographic location with the ability to attend care centres being the main factor, as well as associated chronic sickness (Fletcher *et al.,* 2006; WHO, 2007; Isaacman *et al.,* 2010).

*Streptococcus pyogens,* otherwise known as group A streptococcus (GAS) can cause minor human infections such as pharyngitis and impetigo, and also serious systemic infections such as necrotising fasciitis and streptococcal toxic shock syndrome. Furthermore, recurrent GAS infections can activate auto-immune diseases such as acute poststreptococcal glomerulonephritis, acute rheumatic fever and rheumatic heart diseases (Walker *et al.,* 2014).

*Streptococcus agalactiae,* also known as Lancefield's group B streptococcus, is a Gram-positive coccus that causes septicaemia and meningoencephalitis in various species of marine and freshwater fish globally (Eldar *et al.,* 1995; Evans *et al.,* 2002; Barony *et al.,* 2017). *Streptococcus agalactiae* is also known to cause septicaemia and meningitis in the newborn (Bohnsack *et al.,* 2008) and has been described in other animals, such as guinea pigs, camels, cats, dolphin, horses and frogs as well (Johri *et al.,* 2006). This disease has been described as a key hindrance to the growth of Brazillian aquaculture because it causes high occurrences of disease in Nile tilapia, which is the most frequently farmed fish in Brazil (Mian *et al.,* 2009; Barony *et al.,* 2017).

*Streptococcus suis* comprises an intricate population made up of heterogenous strains (Feng *et al.,* 2009), which can be classified into 35 serotypes based on the composition of capsule antigens (Wertheim *et al.,* 2009). *Streptococcus suis* signifies a health problem in the swine industry globally. It has been described as an evolving zoonotic pathogen that causes severe human infections, clinically presenting with different diseases or syndromes such as meningitis, septicaemia and arthritis (Feng *et al.,* 2014). Table 2.8 shows the strains and species of *Streptococcus* used in this study.

**Table 2.8: Strains of *Streptococcus* used in this study**

| Strains | NCBI: ID |
|---|---|
| *Streptococcus agalactiae* A909 | NC_007432 |
| *Streptococococcus pneumoniae* TCH8431/19A | NC_014251 |
| *Streptococococcus pneumoniae* CGSP14 | NC_010582 |
| *Streptococococcus suis* SC84 | NC_012924 |

| | |
|---|---|
| *Streptococcus pneumoniae* D39 | NC_008533 |
| *Streptococcus pyogens* MGAS2096 | NC_008023 |
| *Streptococcus uberis* 0140J | NC_012004 |
| *Streptococcus gallolyticus* subsp.gallolyticus ATCC BAA-2069 | NC_015215 |
| *Streptococcus suis* ST3 | NC_015433 |
| *Streptococcus pneumoniae* 670-6B | NC_014498 |
| *Streptococcus suis* P1/7 | NC_012925 |
| *Streptococcus agalactiae* NEM316 | NC_004368 |
| *Streptococcus suis* 05ZYH33 | NC_009442 |
| *Streptococcus pyogenes* SSI-1 | NC_004606 |
| *Streptococcus equi* subsp. zooepidemicus MGCS10565 | NC_011134 |
| *Streptococcus pyogenes* MGAS6180 | NC_007296 |
| *Streptococococcus parasanguinis* ATCC 15912 | NC_015678 |
| *Streptococcus pneumonia* AP200 | NC_014494 |
| *Streptococcus gallolyticus* UCN34 | NC_013798 |
| *Streptococcus pneumonia* G54 | NC_011072 |
| *Streptococcus pyogenes* str. Manfredo | NC_009332 |
| *Streptococcus pneumonia* 70585 | NC-012468 |
| *Streptococcus pyogenes* MGAS315 | NC_004070 |
| *Streptococcus salivarius* CCHSS3 | NC_015760 |
| *Streptococus pyogenes* MGAS9429 | NC_008021 |
| *Streptococcus dysgalactiae* subsp.equisimilis GGS_124 | NC_012891 |
| *Streptococcus pneumonia* Taiwan 19F-14 | NC_012469 |
| *Streptococcus pneumonia* P1031 | NC_012467 |
| *Streptococcus infantarius* subsp. infantarius CJ18 | NC_016826 |
| *Streptococcus sanguinis* SK36 | NC_009009 |
| *Streptococcus pyogenes* MGAS10750 | NC_008024 |
| *Streptococcus mutans* NN2025 | NC_013928 |
| *Streptococcus pyogenes* NZ131 | NC_011375 |
| *Streptococcus gordonii* Str. Challis substr. CH1 | NC_009785 |
| *Streptococcus pneumoniae* TIGR4 | NC_003028 |
| *Streptococcus equi* subsp. zooepidemicus | NC_012470 |
| *Streptococcus mutans* UA159 | NC_004350 |
| *Streptococcus pseudopneumonia* IS7493 | NC_015875 |
| *Streptococcus pyogenes* MGAS10394 | NC-006086 |
| *Streptococcus pneumonia* R6 | NC_003098 |
| *Streptococcus equi* subsp.equi 4047 | NC_012471 |
| *Streptocccccus pneumonia* Hungary 19A-6 | NC_010380 |
| *Streptococcus suis* | NC_009443 |
| *Streptococcus pneumoniae* JJA | NC_012466 |
| *Streptococcus pyogenes* MGAS8232 | NC_003485 |
| *Streptococcus parauberis* KCTC 11537 | NC_015558 |
| *Streptococcus thermophiles* LMD-9 | NC_008532 |
| *Streptococcus thermophiles* CNRZ1066 | NC_006449 |
| *Streptococcus pyogenes* MGAS5005 | NC_007297 |
| *Streptococcus agalactiae* 2603V/R | NC_004116 |
| *Streptococcus pyogenes* MGAS15252 | NC_017040 |
| *Streptococcus pyogenes* MGAS10270 | NC_008022 |
| *Streptococcus suis* BM407 | NC_012926 |
| *Streptococcus macedonicus* ACA-DC | NC_016749 |
| *Streptococcus oralis Uo5* | NC_015291 |
| *Streptococcus pneumonia* ATCC 700669 | NC-011900 |
| *Streptococcus pasteurians* ATCC 43144 | NC_015600 |
| *Streptococcus thermophiles* LMG 18311 | NC_006448 |
| *Streptococcus mitis* B6 | NC_013853 |

## 2.3 Design and implementation of a computer algorithm for the generation of diagnostic barcode sequences

The basic principles of selection of barcode sequences were explained in detail in a previous publication by Reva *et al.* (2014) and developed further in this work. The sequence alignment was performed by the MUSCLE algorithm (Edgar *et al.*, 2014). Orthology prediction was done by reciprocal BLASTP implemented in an in-house Python 2.7 script. All the programs for this work were written on Python 2.5 and made accessible at the website http://bargene.bi.up.ac.za/ through a PHP framework.

### 2.3.1 Input data

To generate a set of barcode sequences, the user should upload corresponding genome sequences in GenBank or FASTA format in a single archived file. A minimum of three genomes is required. The maximum size of the file to be uploaded should be < 500 MB. The proportion of accessory genes required should be selected alongside the desired length needed for the barcode sequences to be created. Then the program algorithm identifies taxa-specific genes and generates diagnostic barcodes as explained below. Schematically, the program algorithm is shown in Figure 2.1. The algorithm was implemented as a Python 2.5 program integrated into a PHP framework creating a Web-based user interface (see Chapter 4).

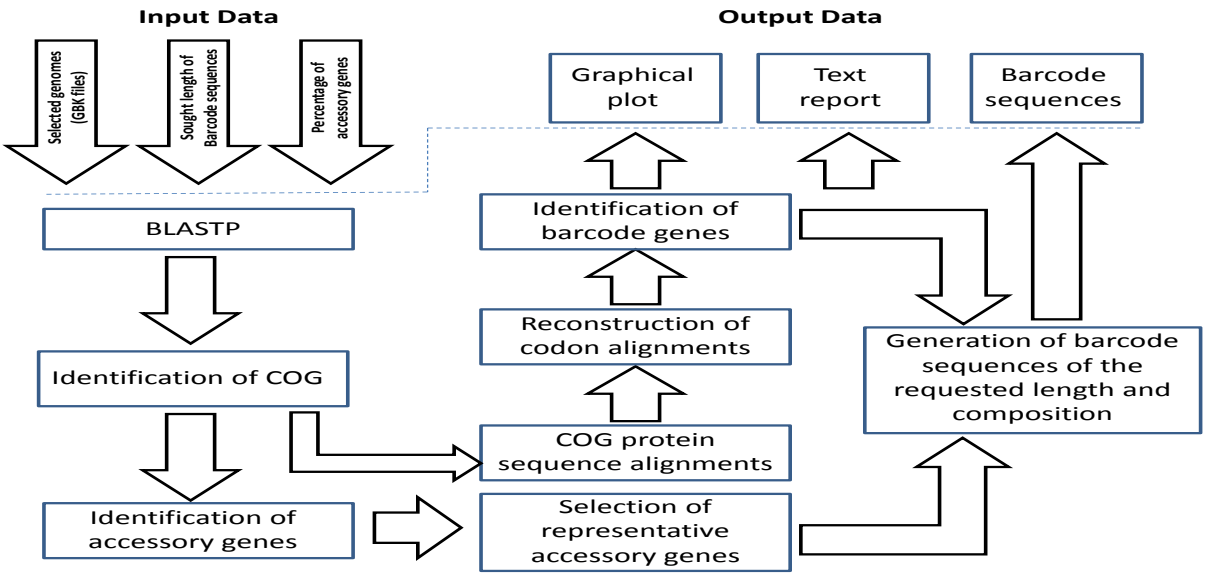



**Figure 2.1: Shows how barcode sequences are generated from the BarcodeGenerator**

### 2.3.2 Orthology prediction

Orthology prediction is performed by a reciprocal BLASTP alignment of all coding sequences predicted in all user-submitted genomes against one another. Two genes were considered as orthologs, if they produced reciprocally the best hit alignments against each other with e-values $\leq 0.0001$ and BLAST score $\geq 75$. A special group of accessory genes found exclusively in one given genome was marked as unique genes.

### 2.3.3 Identification of barcode genes

Core and non-unique accessory genes were grouped into clusters of orthologous genes (COG) and further processed by sequence alignment using the MUSCLE algorithm (Edgar *et al.*, 2010). First, protein sequences were aligned and these alignments were used to reconstruct accurate codon alignments of DNA sequences of corresponding protein-coding genes. Analysis of frequencies of nucleotide and amino acid substitutions in the alignments allows assigning COGs to different categories representing different dynamics and involvement of these genes in evolutionary processes. The following statistical parameters were calculated by the program:

- Percentage of amino acid substitutions (sense mutations) of a total number of nucleotide substitutions in a given pair of orthologous genes;
- Diversity between protein sequences that was calculated as 1 ─ identity;
- Prevalence of positive matches in a pair-wise protein alignment, which includes identical matches and conserved amino acid substitutions, over identities. The value was calculated as (positives – identities)/identities. Conserved pairs of amino acids were identified using the PAM250 matrix of frequencies of amino acid substitutions; and
- The frequency of nucleotide substitutions per 100 bp stretches of DNA, calculated for each pair of aligned DNA (codon) sequences.

The rationale of this analysis was that the different rates of accumulation of nucleotide and amino acid substitutions may depict different categories of genes, i.e. evolutionary conserved genes, genes under strong positive selection and highly variable genes. The distribution of identified COG in a 3D plot is automatically calculated by the program BarcodeGenerator for every pair of submitted genomes and then these values are summarised for every COG and

returned as a graphical output. Figure 2.2 shows this graphical output of the program BarcodeGenerator designed for the automatic creation of diagnostic barcodes. In Figure 2, the individual COG are depicted by dots, which are projected along the axes presenting the percentage of sense mutations in DNA alignments, percentage of mismatches in protein alignments and normalised difference between positives and identities in aligned protein sequences calculated as (positives – identities)/identities. The X axis is the percentage of sense mutations over the total number of nucleotide substitutions, Y is the difference between protein sequences (1 – percentage of identities) and Z (vertical axis) is the ratio (positives-identities)/identities.

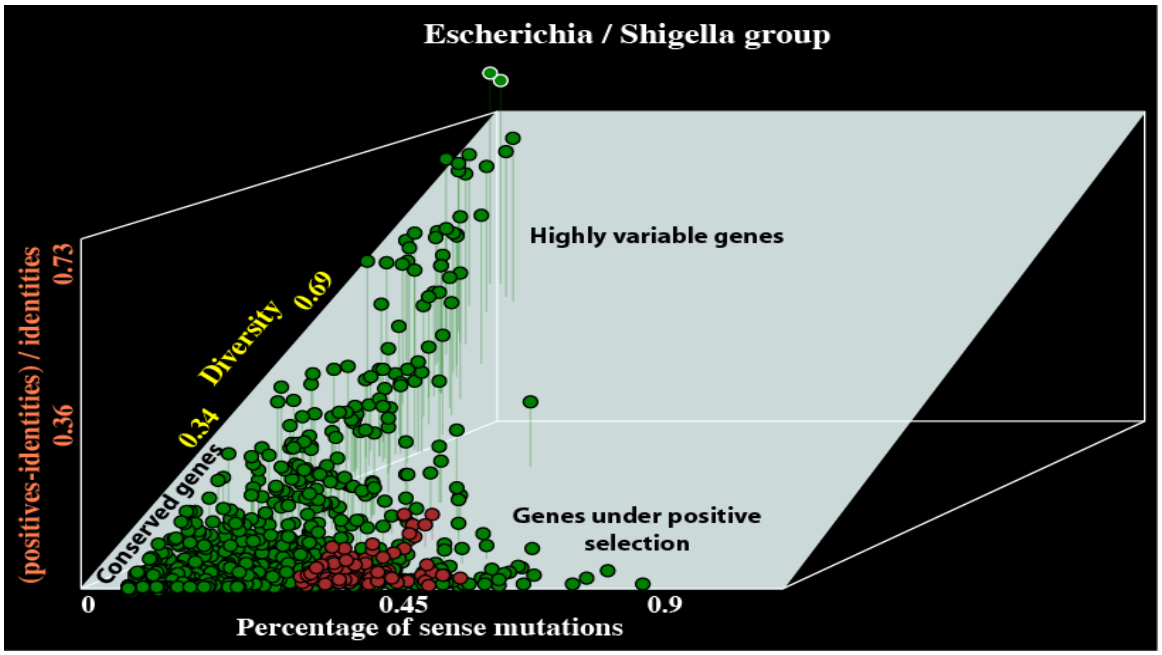



**Figure 2.2: Individual orthologous gene pairs are depicted by dots projected into 3D space, X axis is the percentage of sense mutations over the total number of nucleotide substitutions; Y is the difference between protein sequences (1 – percentage of identities); and Z (vertical axis) is the ratio (positives-identities)/identities. Core genes suitable for barcoding were highlighted in brown**

The COGs from the analysis can be grouped into several categories: conserved; positively selected and highly variable genes. The conserved genes under moderate positive selection (highlighted in Figure 2.2 in brown) were proved to be suitable for barcoding (Reva *et al.*, 2014). Appropriateness of COGs for barcoding was scored as $X \times (1 - X) \times (1 - Y) / (Z + 1)$, where X, Y and Z are values of the respective axes in Figure 2.2. Clusters of COG are ordered by these scores from large to small and then nucleotide sequences of the genes from highly scored COGs are concatenated into barcode

sequences until the requested length of barcodes is achieved. The chosen barcode genes provide a sufficient number of nucleotide substitutions to distinguish between the organisms of interest, yet they are sufficiently conserved to ensure correct orthology prediction.

Users may request the addition of a portion of accessory genes to barcode sequences to improve the sensitivity of the barcodes. In this case, the program first identifies two genomes in the dataset provided, which share the smallest number of accessory genes with each other, but share these genes with other genomes in the group. Then the program selects the accessory genes from these two genomes giving preference to longer genes shared by a bigger number of genomes of the group. An example of the selection of 50 accessory genes from 10 genomes of the genus *Shewanella* is shown in Figure 2.3. First, it was identified that the genomes NC_004347 (*S. oneidensis*) and NC_010334 (*S. halifaxensis*) possess a great number of accessory genes but do not share them with each other. Orthologous genes, depicted by blue bars in Figure 2.3, were found in other genomes: NC_008345 (*S. frigidmarina*), NC_008700 (*S. amazonensis*), NC_009052 (*S. baltica* OS155), NC_009997 (*S. baltica* OS195), NC_011663 (*S. baltica* OS223), NC_009438 (*S. putrefaciens*) and NC_010506 (*S. woodyi*). Genes from this selection were used to replace the core genes in the created barcodes to fit the requested length and core/accessory gene proportion in the resulting barcode sequences. To avoid overrepresentation of accessory genes in the two reference genomes (in Figure 2.3 these genomes are NC_004347 and NC_010334, which accessory genes are depicted by red bars), these genes in barcodes were partly replaced with unique genes identified in these genomes, provided that the length of these genes was above 300 bp. Unique genes were used also to fill in the barcode sequence of the genome NC_008750 (*Shewanella* sp. W3-18-1), which did not share accessory genes with any other genomes in this group. If there are no accessory genes suitable for barcode sequences, barcodes made of the core genes will be returned despite the user request.

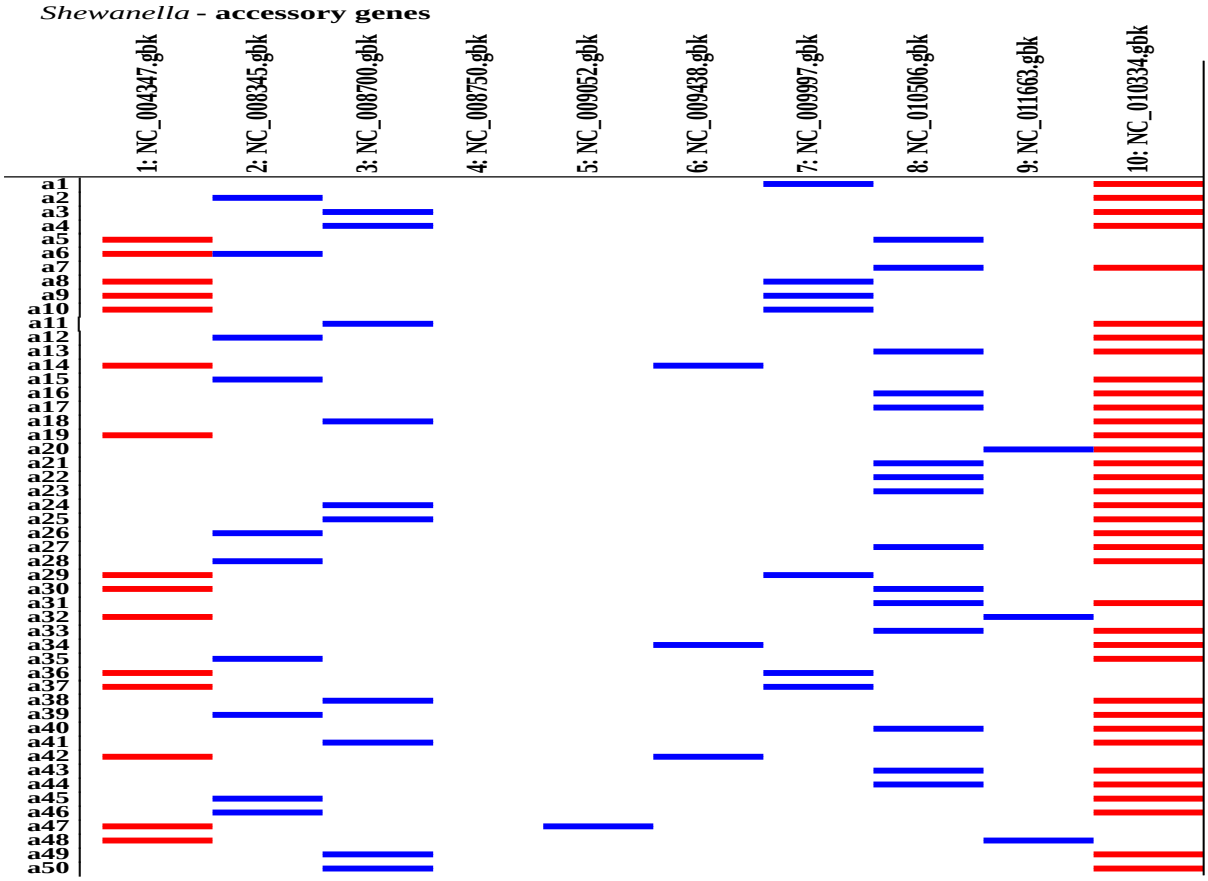



**Figure 2.3: Selection of 50 accessory genes for barcodes to distinguish between *Shewanella* genomes. Sharing of accessory genes is depicted by red and blue bars.**

### 2.3.4 Output data

BarcodeGenerator generates three output files: (i) the core gene plot graphical output, which is a scalable vector grapics (SVG) file shown in Figure 2.2, (ii) the barcode information, which is a text report and (iii) the barcode sequences generated in FASTA format. This information is sent to the e-mail address of the user entered on the website. All groups of microorganisms introduced at the beginning of this chapter were used for case studies, which involved generation of barcode sequences of different lengths and with different proportions of core and accessory genes. All these barcodes and supporting information were made available through the project website at http://seqword.bi.up.ac.za/barcoder_help_download/barcodes/index.html. This page is shown in Figure 2.4.

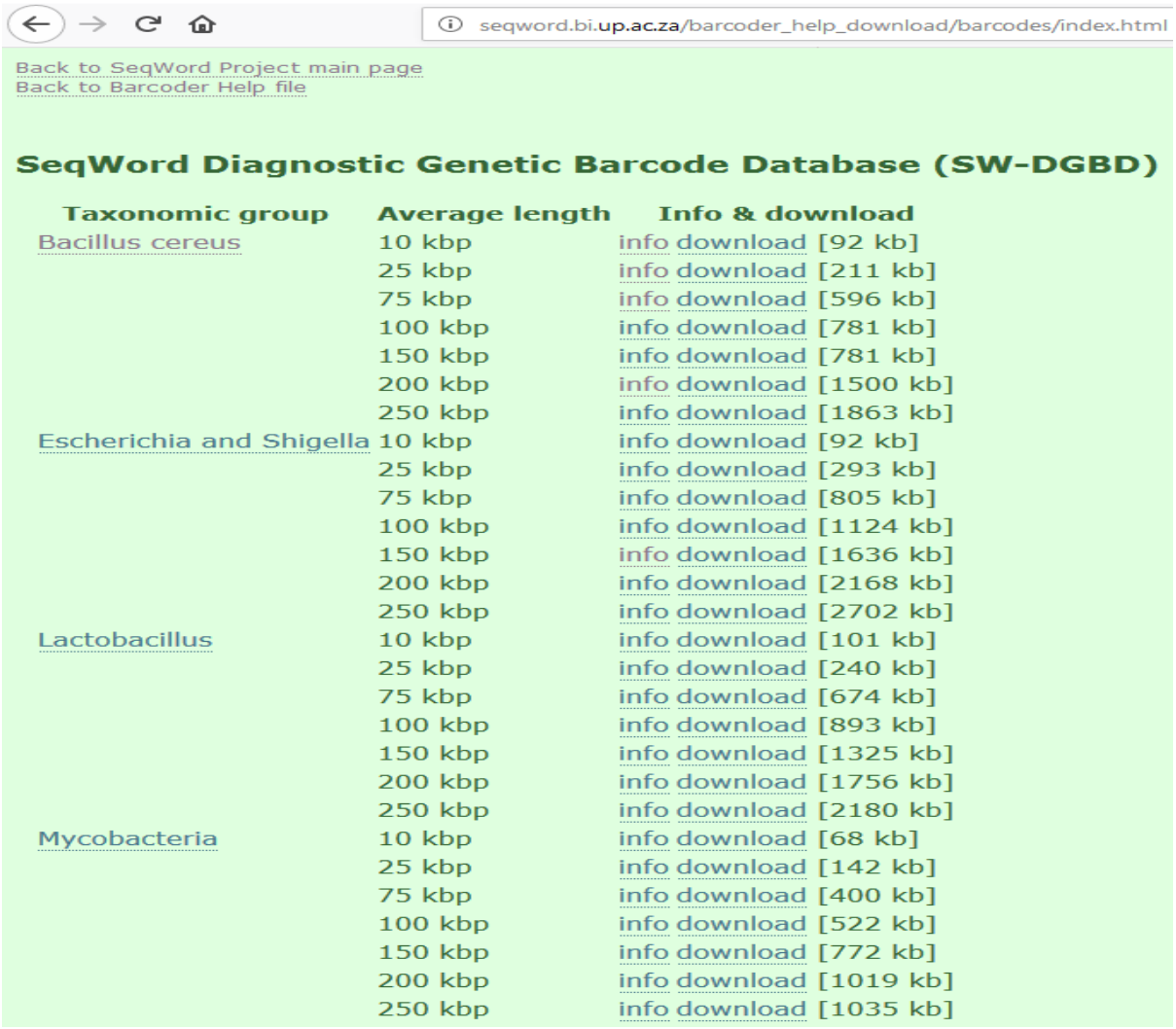



**Figure 2.4: Web-page with the list of barcode sequences generated for this project for testing and evaluation of the developed software tools.**

Additional information about every barcode sequence, including the plot of COG distribution and the list of genes selected for each barcode sequence, is accessible by clicking the links “info” next to each barcode in the list shown in Figure 2.4. The information page for a selected barcode is shown in Figure 2.5.

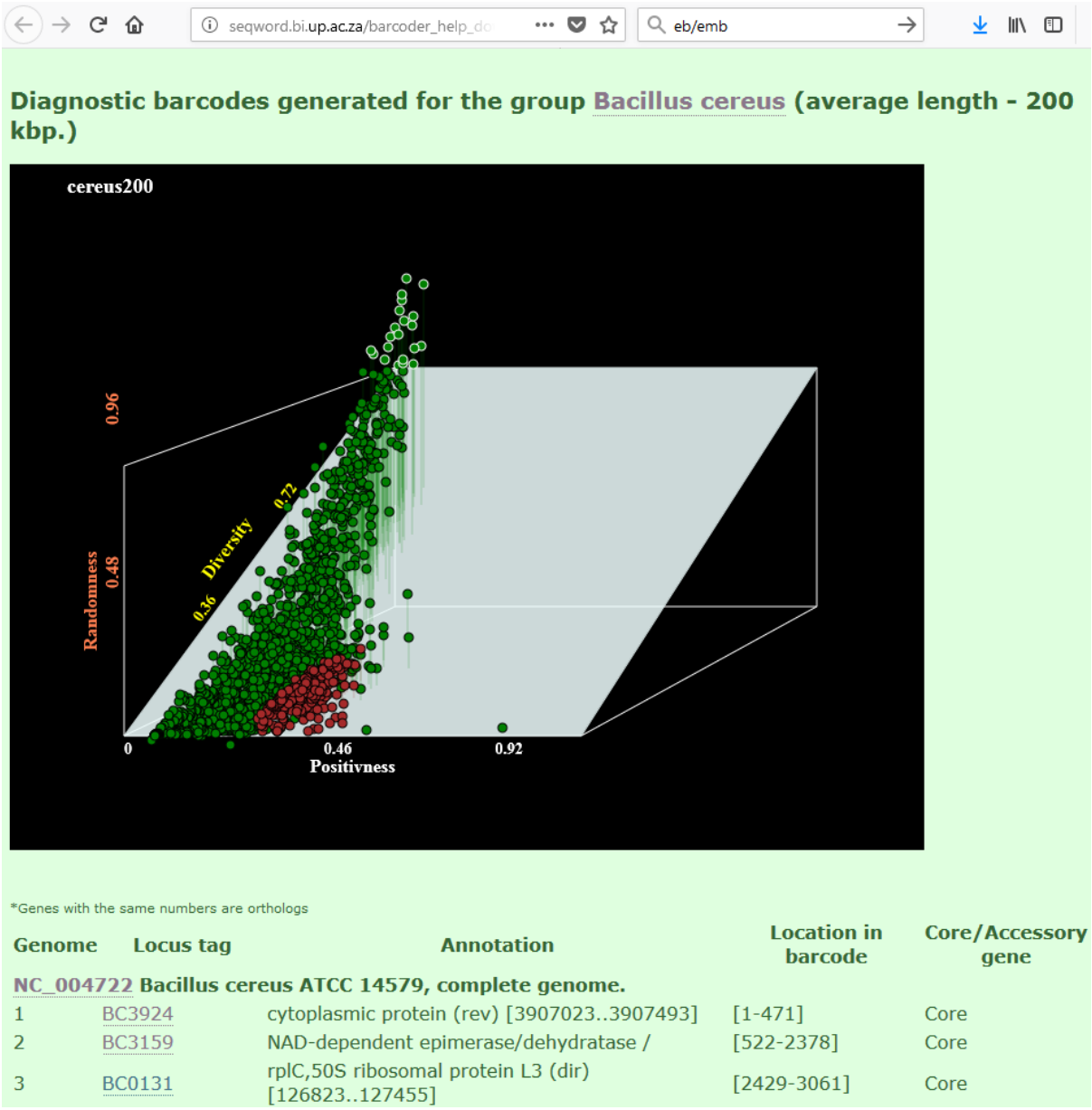



**Figure 2.5: Information page about barcode sequences generated for the *Bacillus cereus* group.**

On all pages of the project web portal all bacterial and gene names are hyperlinked to the corresponding web resources of the NCBI database to allow access to all possible additional information about the subject.

### 2.3.5 Identification of categories of core and accessory genes automatically selected by the BarcodeGenerator for diagnostic barcodes

It was of interest to investigate which categories of genes were selected for barcodes in different groups of organisms by the above-mentioned algorithm. Among the core genes selected for barcodes, the most abundant group was the genes encoding for ribosomal

proteins (Table 2.9 and Figure 2.6). Ribosomal proteins are considered to facilitate the folding of the rRNA and the maintenance of an ideal configuration, which both speed up protein synthesis and accuracy (Wool, 1996). They are also known to have extraribosomal functions involved in replication, translation, RNA processing, DNA repair and autogenous regulation of translation (Wool, 1996). Most importantly, the ribosomal proteins are regarded as the best markers for studying the phylogenetic relationship because they are universal and are made up of highly conserved as well as variable domains (Patwardhan *et al.*, 2014). This finding is in agreement with many publications reporting ribosomal proteins as the most suitable taxonomic and phylogenetic markers used in rMLST (Jolley *et al.*, 2012b; Glaeser and Kämpfer, 2015). Ribosomal proteins comprised up to 15% of the sequences selected for barcodes by the program BarcodeGenerator. Other genes belonged to purine and pyrimidine biosynthetic pathways, ATP-binding cassette (ABC) transporters, tRNA synthetases and amido-transferases, various oxidoreductases, acyl carrier proteins and several other functional categories.

**Table 2.9: Functional categories of genes selected from the core part of genomes for barcode sequences in different groups of microorganisms used as case studies.**

| TYPES OF GENES | NUMBER OF GENES |
|---|---|
| Ribosomal protein | 3137 |
| Purine biosynthesis | 450 |
| ABC transporter ATP-binding protein | 242 |
| Aspartyl/Glutamyl-tRNA amido-transferase | 155 |
| Oxidoreductase | 142 |
| Cysteinyl-tRNA synthetase | 100 |
| Acyl carrier protein | 98 |
| Thymidine kinase | 97 |
| Cytochrome c-type biogenesis protein | 96 |
| Triosephosphate isomerase | 94 |
| others | 17126 |

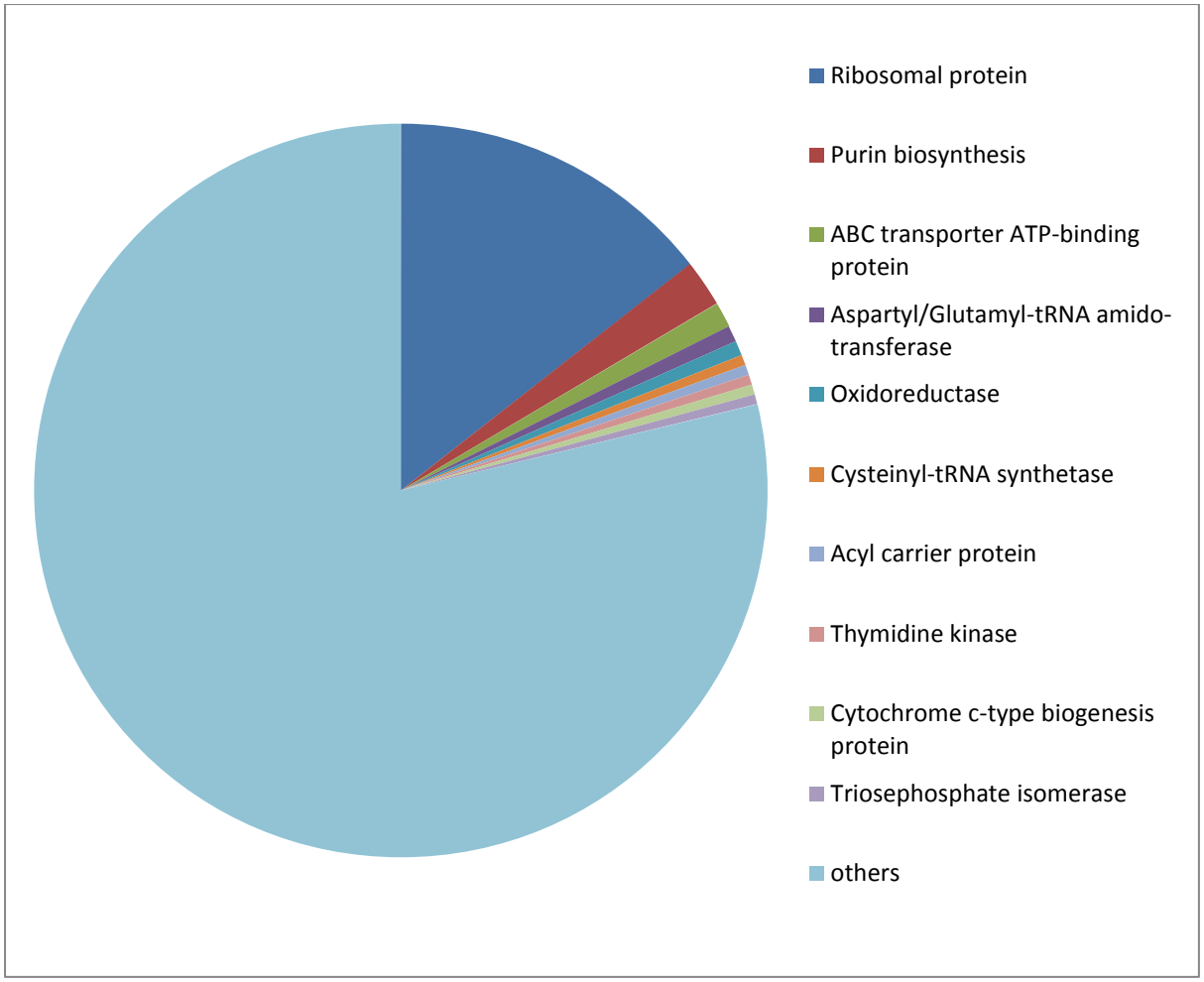



**Figure 2.6: Pie chart showing the different functional categories of genes selected from the core genes**

Synthesis of purine rings plays a principal metabolic role in all cells. The products that are AMP and GMP offer bases for DNA and RNA, as well as for a quantity of important coenzymes such as NAD, NADP, FAD and signalling molecules such as cAMP (Smith and Atkins, 2002). Purine and pyrimidine nucleotides are synthesised *in vivo* at an amount constant with physiological requirements. The intracellular mechanism senses and controls the pool amounts of nucleotide triphosphates, which increase during growth and tissue regeneration when cells are speedily dividing. Three processes contribute to purine nucleotide biosynthesis: (i) synthesis from an amphibolic intermediate, which is *de novo* synthesis; (ii) phosphoribosylation of purines and (iii) phosphorylation of purine nucleosides (Rodwell, 2003).

The ABC transporters are essential membrane proteins that effectively transport all necessary molecules across the lipid membrane against the concentration gradient, using the energy

obtained from the hydrolysis of ATP to ADP. The ABC transporters are found in almost all living organisms and are responsible for a large variety of processes. The ABC domain is also seen in proteins that may couple ATP hydrolysis to function other than transport, such as DNA repair (Doolittle *et al.*, 1986; Linton and Higgins, 1998; Moussatova *et al.*, 2008). Specialised ABC transporters also transport various choices of substrates such as ions, sugars or amino acids to larger compounds, like antibiotics, drugs, lipids and oligopeptides. They also take part in the uptake of nutrients or secretion of toxins in bacteria, as well as conferring multidrug resistance on bacterial cells by pumping diverse drugs and antibiotics into extracellular spaces (Moussatova *et al.*, 2008).

Aminoacyl-tRNA synthetases (ARSs) are made up of an ancient family of enzymes that is found in virtually all cells from the three main kingdoms of life. They are known to catalyse the esterification reactions that bond amino acids with cognate tRNAs bearing the right anticodon triplet to confirm the precise transfer of information directed by the genetic code (Schimmel, 1987; Yao and Fox, 2012). The aminocylation reaction takes place in a two-phase process in which amino acids are first activated by ATP, forming an intermediate aminoacyl adenylate, and then transferred to the 3'-end of the tRNA to form the aminoacyl-tRNA end product (Ibba and Soll, 2000; Yao and Fox, 2012). All ARSs are made up of catalytic and anticodon recognition domains to catalyse aminoacylation reactions exactly for their cognate amino acids. Hence, to enable translational fidelity and sustain usual cellular function, various ARSs have developed editing activities to hydrolyse the inactivated amino acids or mischarged tRNAs and avoid insertion of incorrect amino acids during protein synthesis (Schimmel, 2008).

Oxidoreductases are made up of a huge group of enzymes, which catalyses biological oxidation-reduction reactions (May, 1999). Oxidoreductases make use of the integration of various cofactors such as haeme, flavin and metal ions to catalyse redox reactions. During these reactions, they make use of various electron acceptors and a huge amount of electron-donating substrates generating several products of industrial interest (Gygli and van Berkel, 2015). Several bacteria are made up of proton-translocating membrane-bound nicotinamide adenine dinucleotide (NADH)-quinone oxidoreductases, which show important genetic, spectral and kinetic resemblance with their mitochondrial equivalents (Sled *et al.*, 1993). The mitochondrial NADH:ubiqinone oxidoreductase (complex I, coupling site I) plays an important role in the oxidation of NADH, the reducing product of cellular metabolism, by the

respiratory chain. However, complex I remains the most complex and least understood energy-transducing proto-motive device of the respiratory chain (Sled *et al.*, 1993).

Amidotransferase is a class of enzymes that makes use of the ammonia obtained from the hydrolysis of glutamine for consequent chemical reactions catalysed by the same enzyme. The ammonia intermediate does not dissociate into solution during the chemical alteration (Raushel *et al.*, 1999). A detailed example of the structure and mechanism shown by this group of enzymes is provided by carbamoyl phosphate synthetase. Carbamoyl phosphate synthetase is isolated from *E. coli* as a heterodimeric protein. The smaller of the two subunits is used to catalyse the hydrolysis of glutamine to glutamate and ammonia, while the larger subunit catalyses the formation of carbamoyl phosphate using 2 mol of ATP, bicarbonate and ammonia. Kinetic research has led to a suggested chemical mechanism for this enzyme that needs carboxy phosphate, ammonia and carbamate as kinetically competent reaction intermediates. The amidotransferase part of the CPS best illustrates how protein synthesis can stimulate the capture and use of ammonia obtained from the hydrolysis of glutamine (Raushel *et al.*, 1999).

Acyl carrier proteins (ACPs) perform a major role in *de novo* fatty acid synthesis. Fatty acid synthases (FASs) can be grouped into two unique classes: (i) type I is made up of a single large multifunction polypeptide, which is mostly seen in mammals, fungi and some bacteria; and (ii) type II FASs are mostly seen in archaea, bacteria and plants and are usually categorised by the involvement of distinct mono-functional enzymes for fatty acid synthesis. Acyl carrier proteins exist as a separate domain (Jenke-Kodama *et al.*, 2005; Hung *et al.*, 2017). The *de novo* biosynthesis of FAs takes place through a conserved set of reactions, which are carried out during the elongation cycle (Smith and Sherman, 2008). Acyl carrier proteins are major constituents, which covalently bind all fatty acyl intermediates. During the initial phase, ACPs attach a phosphopantetheine group from CoA on a serine residue of ACP in a conserved Asp-Ser-Leu motif to form holo-ACP (Mofid *et al.*, 2002; Hung *et al*., 2017). The first substrate of FASs, malonyl-CoA, is moved to ACP and the acetyl-CoA unit; the C2 is expanded to a butyrl group, the C4. The synthetic cycle is then reiterated multiple times depending on when the saturated C16 or C18 acyl-ACPs are generated for utilisation in membrane biosynthesis (Chan and Vogel, 2010).

Thymidine kinase (TK), is one of the major enzymes in the pyrimidine salvage pathway, which catalyses the phosphorylation of thymidine to thymidine 5'-monophosphate. The physiological importance of TK is shown by its extraordinary action in cells, which is involved in active DNA synthesis by the evolution of an elaborate feedback structure to control it (Saito and Tomioka, 1984).

Cytochromes (cysts) are pervasive haemoproteins that are major constituents of the energy transduction pathways and important for cellular processes ranging from chemical energy (ATP) production to planned cell death, also known as apoptosis (Moore and Pethigrew, 1990; Jiang and Wang; 2004; Bertini *et al.,* 2006; Verissimo and Daldal, 2014). Cyctochrome c biogenesis is a complex process taking place in almost all organisms and enables the covalent ligation of haeme to an apocyt c. It depends on key cellular functions such as protein translocation followed by post-translational modification extracytoplasmic protein folding and degradation, redox homeostasis, metal cofactor acquisition and insertion into target proteins. Different maturation processes, the systems I to IV sharing similar characteristics, were recognised (Kranz *et al.,* 2009; Sanders *et al.,* 2010; Stevens *et al.,* 2011). In the first system, all apocyts c are synthesised in the cytoplasm and translocated through the sec pathway (Natale *et al.,* 2008; Facey and Kuhn, 2010) across a lipid bilayer into a cellular section where they mature and function. This section is usually on the positive (*p*) side of an energy-transducing membrane such as the bacterial periplasmic space with the exception of the cyst $b_6f$ complex cyst c, also termed $C_x$ or $C_n$, which is formed on the negative (*n)* side of the thylakoid membranes (de Vitry, 2011). In the second system, biosynthesis and transport of haeme and translocation of apocysts occur through a unique and autonomous process, which is coordinated spatially and temporarily to minimise the cytotoxic effects of haeme and proteolytic degradation of apocyts c (Goldman *et al.,* 1996; Moore and Helmann 2005). For the third system both the haeme iron atom and the apocyt c haeme-binding motif Cys thiol groups need to be reduced for thioether bond formation (Kranz *et al.,* 1998; Sanders *et al.,* 2010). For the fourth system, devoted chaperons and enzymes are needed for ligation of haeme to the apocyts c in a stereo-specific configuration. Mature cyst c are assembled into their respective cyst c complexes following their biogenesis (Verissimo and Daldal, 2014).

Analysis of the functions of genes selected by the program for diagnostic barcodes demonstrated that all these genes are involved in central indispensable metabolic processes of

all microorganisms. It guarantees that all these genes always constitute the core part of bacterial genomes; they are sufficiently conserved to be unambiguously identified in metagenomic reads but owing to accumulation of random and positively selected mutations, these genes provide sufficient signals to distinguish between species and sub-species of microorganisms. This analysis confirmed that the program can properly identify and select the genes that are suitable for diagnostic barcodes.

Among accessory genes, the most frequently selected were IS1 and IS2 transposases, membrane proteins, transcriptional regulators and capsular polysaccharide biosynthesis proteins (Table 2.10 and Figure 2.7).

Membrane proteins play a major role in identifying and transmitting outside signals into cells, thereby enabling them to network and respond to their environment in a detailed way. There are two major groups of membrane proteins: (i) those that span the membrane through secondary structures and (ii) those that span it as β barrels. The β barrels groups are usually found in the outer membranes of bacteria, mitochondria and chloroplasts, whereas the helical group is usually found in every other place, making them the most abundant group. The *in vivo* folding of the two classes of proteins is totally different (Bowie, 2004) and usually embroils a discrete cellular mechanism to catalyse the process (Fleming, 2014; Cymer *et al.*, 2015).

Aziz *et al.* (2010) reported that transposases are the most profuse genes in both completely sequenced genomes and environmental metagenomes and are also the most abundant in metagenomes. Transposase genes are known to encrypt DNA binding enzymes, mostly members of the polynucleotidyl transferase superfamily, which catalyses the cut and paste reactions, thereby enhancing the movement of DNA segments to new sites (Rice and Baker, 2001; Aziz *et al.*, 2010). These move double-stranded DNA (dsDNA) directly by excision and insertion and may be linked with insertion sequences (ISs), but most frequently they catalyse their own mobilisation (Crucio and Derbyshire, 2003; Aziz *et al.*, 2010). Insertion sequences make up a significant part of most bacterial genomes. More than 500 different ISs have been identified and many are still being revealed (Mahillon and Chandler, 1998). The DNA ISs IS1, IS2 and IS3 are natural components of *E. coli* and K12 chromosome, where they are available in several duplicates (Brahma *et al.*, 1982). The IS1 was one of the first bacterial ISs to be isolated and identified (Mahillon and Chandler, 1998). The original

examples were obtained from F'*lac*-proB plasmid (IS*I*k) and the multiple drug resistance plasmid R100 (IS*I*R) (Ohtsubo and Ohtsubo, 1978; Mahillon and Chandler, 1998).

The regulation of gene expression is mainly facilitated by proteins termed 'transcription factors (TFs), which identify and bind precise nucleotide sequences and affect the transcription of neighbouring genes (Chalancon and Babu, 2013). Transcription factors are described as DNA binding proteins that bind to precise regions and the *cis*-regulatory elements in the promoter regions of certain genes and finally have an impact on gene expression. In addition to a DNA binding domain that identifies the DNA, most TFs also contain extra-regulatory domains such as small molecule-binding domains and enzyme domains that respond to a signal such as a small molecule (Chalancon and Babu, 2003).

Capsular polysaccharides are usually found on the outermost surface of a varied array of bacteria and are sometimes associated to the cell surface through covalent attachments to phospholipids or a lipid A molecule. Capsular polysaccharides are well hydrated and are normally made up of more than 95% water. They have repeating single monosacchride units that are linked by glycosidic linkages (Taylor and Roberts, 2005). The ability of *S. pneumonia* to regulate CPS production might be a significant factor responsible for its survival in various host environments. The utmost expression of CPS is important for systemic virulence, because of its antiphagocytic properties. Intrusive illnesses are consistently followed by asymptomatic colonisation of the nasopharynx and the thickness of the capsule may have an impact on the degree of exposure of additional significant pneumococcal surface structures, such as adhesins that are needed during this early colonisation stage (Morona *et al.*, 2004).

The gene *gal*T of *E. coli* codes for the enzyme galactose-1-phosphate uridyltransfersae. The *gal*T gene is involved in the metabolism of galactose and it catalyses the reversible conversion of UDP-glucose and galactose-1-phosphate to UDP-galactose and glucose-1-phosphate through an uridylated enzyme intermediate (McCorvie and Timson, 2011; McCorvie *et al.*, 2013).

Permeases are defined as membrane proteins that transduce free energy stored in electrochemical ion gradient into a concentration gradient (Abramson *et al.*, 2003). The *E. coli* lactose permease is one of the most studied members of the main superfamily of

transporters. The molecule is made up of N- and C-terminal domains, with each having six transmembrane helices, symmetrically structured within the permease. A huge internal hydrophilic cavity open to the cytoplasmic side shows the innermost conformation of the transporter. The structure with a bound lactose homolog β-D- galactopyranosyl-1-thio-β-D-galactose shows the sugar binding site of the cavity and residues that play key roles in substrate identification and proton translocation are recognised (Abramson *et al.,* 2003).

Diacetyl is an essential aroma compound and plays a key role in the flavour of dairy products. Usually *L. lactis* undergoes homolactic fermentation and most of the dominant intermediate pyruvate is converted to lactate, a reaction catalysed by lactate dehydrogenase (LDH) with the oxidation of NADH to $NAD^+$ for maintaining a redox balance (Neves *et al.,* 2005). In aerobic situations the activities of α-acetolactate synthase, i.e ALS and NADH oxidase (NOX), are highly increased (Bassit *et al.,* 1993; Guo *et al.,* 2012). Alpha-acetolactate synthase catalyses the pyruvate to acetolactate. After a decarboxylation process, α acetolactate is then converted to acetoin and diacetyl. Reoxidation of NADH by NOX usually replaces the role of LDH in the regeneration of $NAD^+$, leaving room for the accumulation of the two aroma compounds (Lopez de Felipe, 2000; Guo *et al.,* 2012). Hence, in the presence of Oxygen, *L. lactis* shows the metabolic shift from homolactic to mixed-acid product formation comprising lactate, acetate and $CO_2$, which makes diacetyl accumulation restricted (Guo *et al.,* 2012). Hence, various methods to enhance diacetyl production in *L. lactis* have been established, such as the overexpression of *als* and *nox*-2 and the inactivation of the *ldh* and *α*-acetolate decarboxylase (*aldB*) genes. Therefore, excessive pyruvate was channelled to diacetyl through ALS while the flux pyruvate to lactate was almost eradicated (Guo *et al.,* 2012).

Hence, the accessory genes selected by the program for barcode sequences belonged to two categories: selfish mobile genetic elements infecting bacteria and functional genes, which provided bacteria with biosynthetic capacities important in specific habitats, or for a molecular redress of surface compounds to avoid the immune response of host organisms.

**Table 2.10: Shows the different types of genes selected among the accessory genes for barcode sequences in the different groups of microorganisms used as case studies**

| Types of genes | Number of genes |
|---|---|
| Membrane protein | 159 |
| Insertion element IS2 transposase | 145 |
| Transcriptional regulator | 123 |
| Capsular polysaccharide biosynthesis | 60 |
| Transposase ORF A | 55 |
| Galactose-1-phosphate uridylyltranseferase | 50 |
| Premease | 48 |
| Cytochrome c-type biogenesis protein | 48 |
| Alpha-acetolactate decarboxylase | 46 |
| Others | 3583 |

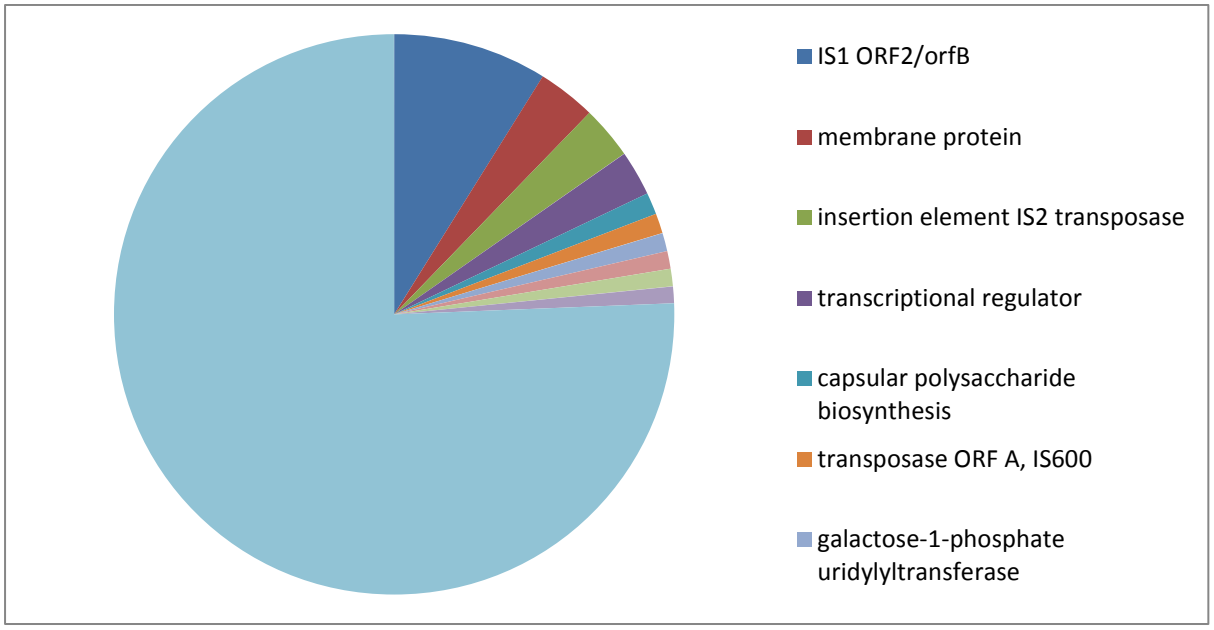



**Figure 2.7: Pie chart showing the different classes of genes selected for the accessory genes**

## 2.4 Conclusion

Advances in present-day sequencing technologies have made it affordable to sequence and compare whole genomes of related microorganisms in the infancy of clonal segregation and speciation. Hence, there is a need for new computation techniques for mining an enormous

quantity of data generated by next-generation sequencing technologies. It is also of importance to identify and highlight marker sequences most suitable for the strains of interest and their biological activity. Multi-locus barcoding is a promising method for dependable identification of strains of closely related bacteria in environmental samples.

The aim of this work was to create an interactive computational service for the identification of the most suitable marker sequence for DNA-based multi-barcoding. The BrcodeGenerator is a novel software tool available for use at http:/bargene.bi.up.ac.za/. The program BarcodeGenerator creates a specific barcode sequence based on the core and accessory gene provided by the user. The program then returns a link with the generated barcode sequences in FASTA format, information on the genes selected for barcodes and a graphical file in SVG format. The researcher also investigated which categories of genes were selected for barcodes in different groups of organisms by the above-mentioned algorithm. Among the core genes selected for barcodes, the most abundant group was the genes encoding for ribosomal proteins. The next question to address is how efficient the developed barcodes are in binning metagenomic reads to distinguish between closely related organisms. Development of the program for binning DNA reads against multi-locus barcodes and statistical validation of the results of the binning will be covered in the following chapters.

## References


Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR and Iwata S (2003). Structure and mechanism of the lactose premease of *Escherichia coli. Science,* 5633, pp. 610-615

Akinola RO, Mazandu GK and Mulder NJ (2013). A systems level comparison of *Mycobacterium tuberculosis, Mycobacterium leprae* and *Mycobacterium smegmatis* based on functional interaction network analysis. *Bacteriology and Parasitology,* 4:173

Alina SO, Constantinscu F and Petruta CC (2015). Biodiversity of *Bacillus subtilis* group and beneficial traits of *Bacillus* species useful in plant protection. *Romanian Biotechnological Letters,* 20, pp. 10737-10750

Arrigo KR (2005). Marine microorganisms and global nutrient cycles. *Nature,* 437, pp. 349-355

Aziz RK, Breitbart M and Edwards RA (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acid Research,* 38, pp. 4207-4217

Barony GM, Tavares GC, Pereira FL, Carvalho AF, Dorella FA, Leal CAG and Figueiredo HCP (2017). Large-scale genomic analyses reveal the population structure and evolutionary trends of *Streptococcus agalactiae* strains in Brazillian fish farms. *Scientific Reports,* 7:13538

Bashir Y, Singh SP and Konwar BK (2014). Metagenomics: An application based perspective. *Chinese Journal of Biology,* pp. 1-7

Bassit N, Boquien CY, Picque D and Corrieu G (1993). Effect of initial oxygen concentration on diacetyl and acetoin production by *Lactococcus lactis* subsp *lactis,* biovar diacetyl lactis. *Applied and Environmental Microbiology,* 59, pp. 1893-1897

Baumdicker F, Hess WR and Pfaffelhuber P (2012). The infinitely many genes model for the distributed genome of bacteria. *Genome Biology and Evolution,* 4, pp. 443-456

Belchik SM, Kennedy DW, Dohnalkova AC, Wang Y, Sevinc PC, Wu H, Lin Y, Lu HP, Fredrickson JK and Shi L (2011). Extracellular reduction of hexavalent chromium by cytochromes mtrC and omcA of *Shewanella oneidensis* MR-1. *Applied and Environmental Microbiology,* 77, pp. 4035-4041

Beliaev AS and Saffarini DA (2001). *Shewanella putrefaciens mtrB* encodes an outer membrane protein required for Fe (III) and Mn reduction (1998). *Journal of Bacteriology,* 180, pp. 6292-6297

Bergmann and Hammerschmidt S (2006). Versatility of pneumococcal surface proteins. *Microbiology,* 152, pp. 295-303

Bertini I, Cavallaro G and Rosato A (2006). Cytochrome *c*: occurrence and functions. *Chemical Reviews,* 106, pp. 90-115

Beyene G, Nair S, Asrat D, Mengistu Y, Engers H and Wain J (2011). Multidrug resistant Salmonella concord is a major cause of salmonellosis in children in Ethopia. *The Journal of Infection in Developing Countries,* 5, pp. 23-33

Biller SJ, Berube PM, Berta-Thompson JW, Kelly L, Roggendack SE, Awad L, Roache-Johnson KH, Ding H, Giovannoni SJ, Rocap G, Moore LR and Chisholm SW (2014). Genomes of diverse isolates of marine cyanobacterium *Prochlorococcus. Scientific Data,* 1:140034

Bischoff KM, Rooney AP, Li XL, Liu S and Hughes SR (2006). Purification and characterization of a family 5 endoglucanase from a moderately thermophilic strain of *Bacillus licheniformis. Biotetechnology Letters,* 28, pp. 1761-1765

Blaiotta G, Moschetti G, Simeoli E, Andolfi R, Villani F and Coppola S (2001). Monitoring lactic acid bacteria strains during 'Cacioricotta' cheese production by restriction endonuclease analysis and pulsed-field gel electrophoresis. *Journal of Dairy Research,* 68, pp. 139-144

Bohnsack JF, Whiting A, Gotlschalk M, Dunn DM, Weiss R, Azimi PH, Philips JB III, Welsman LE, Rhoads GG and Lin C FY (2008). Population structure of invasive and colonizing strains of *Streptococcus agalactiae* from neonates of six U.S academic centers from 1995-1999. *Journal of Clinical Microbiology,* 46, pp. 1285-1291

Bouton Y, Guyot P, Beuvier E, Tailliez P and Grappin R (2002). Use of PCR-based methods and PFGE for typing and monitoring homofermentative *Lactobacilli* during Comte' cheese ripening. *International Journal of Food Microbiology,* 76, pp. 27-28

Bowie JE (2004). Membrane proteins: A new method enters the fold. *Proceedings of the National Academy of Sciences of the United States of America,* 101, pp. 3995-3996

Brahma N, Schumacher A, Cullum J and Saedler H (1982). Distribution of the *Escherichia coli* K12 insertion sequences IS1, 1S2 and IS3 among other bacterial species. *Journal of General Microbiology,* 128, pp. 2229-2234

Bretschger O, Obraztsova A, Sturm CA, Chang IS, Gorby YA, Reed SB, Culley DE, Reardson CL, Barua S, Romine MF, Zhou J, Beliaev AS, et al.,. An exploration of current production and metal oxide reduction of *Shewanella oneidensis* MR-1 wild type and mutants. *Applied and Environmental Microbiology,* 73, pp. 7003-7012

Buddington R (2009). Using probiotics and prebiotics to manage the gastrointestinal tract ecosystem. In: Charalampopoulos D, Rastall RA (Eds), Prebiotics and Probiotics Science and Technology. Springer Science + Business Media, New York, pp. 1-32

Canstein HV, Ogawa J, Shimizu S and Lloyd JR (2008). Secretion of flavins by *Shewanella* species and their role in extracellular electron transfer. *Applied and Environmental Microbiology,* 74, pp. 615-623

Centers for Disease Control and Prevention (2006). Vaccine preventable deaths and the global immunization vision and strategy. *Morbidity and Mortality Weekly Report,* 55, pp. 511-515

Chalancon G and Babu MM (2013). Structure and evolution of transcriptional regulatory networks. In *Bacteria Gene Regulation and Transcriptional Networks, Cambridge, United Kingdom,* Chapter 8, pp. 121-126

Chan DI and Vogel HJ (2010). Current understanding of fatty acid the acyl carrier protein. *Biochemical Journal,* 430, pp. 1-19

Chiodini RJ, Chamberlin WM, Sarosiek J and McCallum RW (2012). Crohn's disease and the mycobacterioses: A quarter century later. Causation or simple association? *Critical Reviews in Microbiology,* 38, pp. 52-93

Chun J and Rainey FA (2014). Integrating genomics into the taxonomy and systematics of bacteria and archaea. *International Journal of Systematic and Evolutionary Microbiology,* 64, pp. 316-324

Coleman ML and Chisholm SW (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences of the United States of America,* 107, pp. 18634-18639

Cook JL (2010). Nontuberculous mycobacteria: opportunistic environmental pathogens for predisposed hosts. *British Medical Bulletin,* 96, pp. 45-59

Cordone A, Audrain B, Calabrese I, Euphrasie D and Reyrat JM (2011). Characterization of a *Mycobacterium smegmatis uvrA* mutant impaired in dormancy induced hypoxia and low carbon concentration. *BMC Microbiology,* 11:231

Croxen MA and Finlay BB (2010). Molecular mechanisms of *Escherichia coli* pathogenicity. *Nature Reviews Microbiology,* 8, pp. 26-28

Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M and Finlay BB (2013). Recent advances in understanding enteric pathogenic *Escherichia coli. Clinical Microbiology Reveiws,* 26, pp. 822-880

Crucio MJ and Derbyshire KM (2003). The outs and ins of transposition: from mu to kangaroo. *Nature Reviews Molecular Cell Biology,* 4, pp. 865-877

Cymer F, von Heijne G and White SH (2015). Mechanisms of integral membrane protein insertion and folding. *Journal of Molecular Biology,* 427, pp. 999-1021

Damangel C, Stinear TP and Cole ST (2009). Buruli ulcer: reductive evolution enhances pathogenicity of *Mycobacterium ulcerans. Nature Reviews Microbiology,* 7, pp. 50-60

de Vitry C (2011). Cytochrome *c* maturation system on the negative side of bioenergetics membranes: CCB or system IV. FEBS Journal, 278, pp. 4189-4197

Dikow RB (2011). Genome-level homology and phylogeny of *Shewanella* (Gammaproteobacteria: Iteromonadales: Shewanellaceae). *BMC Genomics,* 12:237

Doolittle RF, Johnson MS, Husain I, Van Houten B, Thomas DC and Sancar A (1986). Dominal evolution of a prokaryotic DNA repair protein and its relationship to active-transport proteins. *Nature,* 323, pp. 451-453

Eisenstadt and Hall DGS (1995). Microbiology and classification of *Mycobacteria. Clinics in Dermatology,* 13, pp. 197-206

Eldar A, Bejerano Y, Livoff A, Horovitcz A and Bercovier H (1995). Experimental streptococcal meningo-encephalitis in cultured fish. *Veterinary Microbiology,* 43, pp. 33-40

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, pp. 2460-2461

Evans JJ, Klesius PH, Gilbert PM, Shoemaker CA, Al Sarawi MA, Landsberg J, Duremdez R, Al Marzouk A, Al Zenki S (2002). Characterization of β-haemolytic Group B *Streptococcus agalactiae* in cultured seabream, *Sparus auratus L.*, and wild mullet, *Liza klunzinger* (Day), in Kuwait. *Journal of Fish Diseases,* 25, pp. 505-513

Ewing WH (1949). *Shigella* nomenclature. *Journal of Bacteriology,* 57, pp. 633-638

Facey SJ and Kuhn A (2010). Biogenesis of bacterial inner-membrane proteins. *Cellular and Molecular Life Sciences,* 67, pp. 2343-2362

Feasey NA, Archer BN, Heyderman RS, Sooka A, Dennis B, Gordon MA and Keddy KH (2010). Typhoid fever and invasive nontyphoid Salmonellosis, Malawi and South Africa. *Emerging Infectious Disease,* 16, pp. 1448-1451

Feasey NA, Dougan G, Kingsley RA, Heyderman RS and Gordon MA (2012). Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa. *Lancet,* 379, pp.2489-2499

Feng Y, Zhang H, Wu Z, Wang S, Cao M, Hu D and Wang C (2014). *Streptococcus suis* Infection, an emerging/reemerging challenge of bacterial infectious diseases? *Virulence,* 5, pp. 477-497

Fleming KG (2014). Energetics of membrane protein folding. *Annual Review of Biophysics,* 43, pp. 233-255

Fletcher MA, Laufer DS, Mclntosh EDG, Cimino C and Malinoski FJ (2006). Controlling invasive pneumococcal disease: is vaccination of at-risk groups sufficient. *The International Journal of Clinical Practice,* 60, pp. 450-456

Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jia ON, Karl DM, Li WKW, Lomas MW, Veneziano D, Vera CS, Vrugt JA and Martiny AC (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus.*

*Proceedings of the National Academy of Sciences of the United States of America,* 110, pp. 9824-9829

Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, Gardner TS, Nealson KH, Osterman AL, Pinchuk G, Reed JL, Rodionov DA, et al (2008). Towards environmental systems biology of *Shewanella. Nature Reviews Microbiology,* 6, pp. 592-603

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW and DeLong EF (2008). Microbial community gene expression in Ocean surface waters. *Proceedings of the National Academy of Sciences of the United States of America,* 105, pp. 3805-3810

Galagan JE, Henn MR, Ma L-J, Cuomo CA and Birren B (2005). Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Research,* 15, pp. 1620-1631

Gilbert JA and Dunpont CL (2011). Microbial metagenomics beyond the genome. *Annual Review of Marine Science,* 3, pp. 347-371

Gilks CF, Brindle RJ, Newnham RS, Watkins WM, Waiyaki PG, Were JBO, Otieno LS, Simani PM, Bhatt SM, Lule GN, Okelo GBA, Brindle RJ, Newnham RS, Gilks CF and Warrell DA (1990). Life-threatening bacteremia in HIV-1 seropositive adults admitted to hospital in Nairobi, Kenya. *The Lancet,* 336, pp. 545-549

Glaser PS and Kämfer P (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonmy. *Systematic and Applied Microbiology,* 38, pp. 23

Goldman BS, Gabbert KK and Kranz RG (1996). Use of heme reporters for studies of cytochrome biosynthesis and heme transport. *Journal of Bacteriology,* 178, pp. 6338-6347

Gordon MA (2008). *Salmonella* infections in immunocompromised adults. *Journal of Infection,* 56, pp. 413-422

Grad YH and Lipsitch M (2014). Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biology,* 5:538

Guo T, Kong J, Zhang L, Zhang C and Hu S (2012). Fine tuning of the lactate and diacetyl production through promoter engineering in *Lactococcus lactis. PLoS ONE,* 7:e36296

Gygli G and van Berkel WJH (2015). Oxizymes for biotechnology. *Current Biotechnology,* 4, pp. 100-110

Hau HH and Gralnick JA (2007). Ecology and biotechnology of the genus *Shewanella. Annual Review Microbiology,* 61, pp. 237-258

Hau HH, Gilbert A, Courselle D and Gralnick JA (2008). Mechanism and consequences of anaerobic respiration of cobalt by *Shewanella oneidensis* strain MR-1. *Applied and Environmental,* 74, pp. 6880-6886

Haung J, Xue C, Wang H, Schmidt W, Shen R and Lan P (2017). Genes of ACYL carrier protein family show different expression profiles and overexpression of ACYL carrier protein 5 modulates fatty acid composition and enhances salt stress tolerance in *Arabidopsis. Frontiers in Plant Science,* 8:987

He and De Buck J (2010). Cell wall proteome analysis of *Mycobacterium smegmatis* strain MC2 155. *BMC Microbiology,* 10:121

Heidelberg JF, Paulsen IT, Nelson KE, Gaidos EJ, Nelson WC, Read TD, Eisen JA, Seshadri R, Ward N, Methe B, Clauton RA, Meyer T, Tsapin A, et al., (2002). Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis. Nature Biotechnology,* 20, pp. 1118-1123

Hooper LV, Littman DR and Macpherson AJ (2012). Interactions between microbiota and the immune system. *Science,* 336, pp. 1268-1273

Hoskins J, Alborn JR WE, Arnold J, Blaszczak LC, Burgett S, Dehoff BS, Estrem ST, Fritz L, Fu DJ, Fuller W, Geringer C, Gilmour R, Glass JS et al., (2001). Genome of the bacterium *Streptococcus pneumoniae* strain R6. *Journal of Bacteriology,* 183, pp. 5709-5717

Humtsoe JO, Kim JK, Xu Y, Keene DR, Hook M, Lukomski S and Wary KK (2005). A streptococcal collagen-like protein interacts with α2β1 integrin and induces intracellular signalling. *The Journal of Biological Chemistry,* 280, pp. 13848-13857

Ibba M and Soll D (2000). Aminoacyl-tRNA synthesis. *Annual Review of Biochemistry,* 69, pp. 617-650

Isaacman DJ, Mclntosh D and Reinert RR (2010). Burden of lnvasive pneumococcal disease and serotype distribution among *Streptococcus pneumoniae* isolates in young children in Europe: impact of the 7-valent pneumococcal conjugate vaccine and considerations for future conjugate vaccines. *International Journal of Infectious Diseases,* 14, pp. e197-e209

Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V, Bhattacharyya A, Reznik G, Mikharlova N, Lapidus A, Chu L, Mazur M, Goltsman E, et al., (2003). Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis. Letters to nature,* 423, pp. 87-91

Ivanova EP, Gorshkova NM, Bowman JP, Lysenko AM, Zhukova NV, Sergeev AF, Mikhailov VV and Nicolau DV (2004). *Shewanella pacific asp*. Nov., a polyunsaturated fatty acid-producing bacterium isolated from sea water. *International Journal of Systematic and Evolutionary Microbiolgy,* 54, pp. 1083-1087

Jenke-Kodama H, Sandmann A, Muller R and Dittmann E (2005). Evoluntionary implications of bacterial polyketide synthases. *Molecular Biology and Evolution,* 22, pp. 2027-2039

Jiang X and Wang X (2004). Cytochrome *c* mediated apoptosis. *Annual Review of Biochemistry*, 73, pp. 87-106

Joelly KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarattttthna H, Harrison OB, Sheppard SK, Cody AJ and Maiden MCJ (2012b). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology,* 15, pp. 1005-1015

Johnson ZI, Zinser ER, Coe A McNulty NP, Woodward EMS and Chisholm SW (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science,* 311, pp. 1737-1740

Johri AK, Paoletti LC, Glser P, Dua M, Sharma PK, Grandi G and Rappuoli R (2006). Group B *Streptococcus:* global incidence and vaccine development. *Nature Reviews Microbiology,* 4, pp. 932-942

Kalluri P, Cummings KC, Abott S, Malcolm GB, Hutcheson K, Beall A, Joyce K, Polyak C, Woodward D, Caldeira R, Rodgers F, Mintz ED and Strockbine N (2004). Epidemiological features of a newly described serotype of *Shigella boydii. Epidemiology and Infection,* 132, pp. 579-583

Kaper JB, Nataro JP and Mobley HL (2004). Pathogenic *Escherichia coli. Nature reviews Microbiology,* 2, pp. 123-140

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus. Science*, 344, pp. 416-420

Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, Van Domselaar G, Deng X, Carleton HA (2017). A comparative analysis of the Lysve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Frointiers in Microbiology*, 8:375

Kettler GC, Martiny AC, Haung K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P and Chisholm SW (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus. PLoS Genetics*, 3:e231

Kilian M (2005). *Streptococcus* and *Lactobacillus*, pp. 833-881. *In* Borriello P, Murray PR and Funke G (ed.), Topley and Wilson's Microbiology and Microbial Infections. Hodder Arnold, London, United Kingdom

Kotloff KL, Nataro JP, Blackwelder W, Nasrin D and Farag T (2013). Burden and aetiology of diarrhoea disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet*, 382, pp. 209-222

Kranz R, Lill R, Goldman B, Bonnard G and Merchant S (1998). Molecular mechanisms of cytochrome *c* biogenesis: three distinct systems. *Molecular Microbiology*, 29, pp. 383-396

Kranz RG, Richard-Fogal C, Taylor JS and Frawley ER (2009). Cyctochrome *c* biogenesis: mechanisms for covalent modifications and trafficking of heme and for heme-iron redox control. *Microbiology and Molecular Biology Reviews*, 73, 510-528

Kress WJ and Erickson DL (2008). DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America*, 105, pp. 2761-2762

Kulski JK (2016). An overview of the history, tools and "omics" applications and challenges. Kulski JK (Ed.). InTech, Rijeka, Croatia, pp. 3-60

Kwong JC, McCallum N, Sintchenko V and Howden BP (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology*, 47, pp. 199-210

Lan R, Alles MC, Donohoe K, Martinez MB and Reeves PR (2004). Molecular evoluntionary relationships of enteroinvasive *Escherichia coli* and *Shigella* spp. *Infection* and *Immunity*, 72, pp. 5080-5088

Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar J, Poudel S and Ussery DW (2015). Insights from 20 yrs of bacterial genome sequencing. *Functional Integrative Genomics*, 15, pp. 141-161

Laupland KB. Schønheyder HC, Kennedy KJ, Lyytikäinen O, Valiquette L, Galbraith J, Collignon P and the International Bacteremia Surveillance Collaborative (2010). *Salmonella enterica* bacteraemia*:* a multinational population-based cohort study. *BMC Infectious Diseases*, 10:95

Levine MM, Kotloff KL, Nataro JP and Muhsen K(2012). The Global Enteric Multicenter Study (GEMS):impetus, rationale and genesis. *Clinical Infectious Diseases*, 55, pp. S215-S224

Li DB,Cheng YY, Wu C, Li WW, Li Na, Yang ZC, Tong ZH and Yu HQ (2013). Selenite reduction by *Shewanella oneidensis* MR-1 is mediated by fumarate reductase in periplasm. *Scientific Reports*, 4:3735

Linton KJ and Higgins CF (1998). The *Escherichia coli* ATP-binding cassette (ABC) proteins. *Molecular Microbiology*, 28, pp. 5-13

Liu C, Gorby YA, Zachara JM, Fredrickson JK and Brown CF (2002). Reduction kinetics of Fe (III), Co(III), U (VI), Cr(VI) and TC (VII) in cultures of dissimilatory metal-reducing bacteria. *Biotechnology and Bioengineering*, 80, pp. 637-649

Liu G, Zhou J, Meng X, Fu SQ, Wang J, Jin R and Hong LV (2013). Decolorization of azo dyes by marine *Shewanella* strains under saline conditions. *Applied Microbiology and Biotechnology*, 97, pp. 4187-4197

Lopez de Felipe F, Kleerebezem M, de Vos WM and Hugenholtz J (2000). Lactic acid bacteria as a cell factory: rerouting ofcarbon metabolism in *Lactococcus latis* by metabolic engineering. *Enzyme and microbial technology*, 26, pp. 840-848

Mahillon J and Chandler M (1998). Insertion sequences. *Microbiology and Molecular Biology Reviews*, 3, pp. 725-774

Maiden MC, Bygraves JA, Feil E, Mrelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M and Spratt BG (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Preceedings of the National Academy of Sciences of the United States of America,* 95, pp.3140-3145

Markiewicz LH, Biedrzycka E, Wasilewska E and Bielecka M (2010). Rapid molecular identification and characteristics of *Lactobacillus* strains. *Folia Microbiologica,* 55, pp.481-488

Marsili E, Baron DB, Shikhare ID, Coursolle D, Gralnick JA and Bond DR (2008). *Shewanella* secrets flavins that mediate extracellular electron transfer. *Proceedings of the National Academy of Sscience of the United States of America,* 105, pp. 3968-3973

Martiny AC, Huang Y and Li N (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environmental Microbiology,* 11, pp. 1340-1347

May SW (1999). Applications of oxidoreductases. *Current Opinion in Biotechnology,* 4, pp. 370-375

McCorvie TJ and Timson DJ (2011). The structural and molecular biology of type I galactosemia: enzymology of galactose 1-phosphate uridylytransferase. *IIUBMB Life Journal,* 63, pp. 694-700

McCorvie TJ, Gleason TJ, Fridovich-Keil JL and Timson DJ (2013). Misfolding of galactose 1-phosphate uridylyltransferase can result in type I galactosemia. *Biochimica et Biophysica Acta,* 1832, pp. 1279-1293

Mian GF, Godoy DT, Leal CAG, Yuhara TY, Costa GM and Figueiredo HCP (2009). Aspects of the natural history and virulence of *S. agalactiae* infection in Nile tilapia. *Veterinary Microbiology,* 136, pp. 180-183

Mofid MR, Finking R and Marahiel MA (2002). Recognition of hybrid peptidyl carrier proteins/acyl carrier proteins in nonribosomal peptide synthetase modules by the 4'-

phosphopantetheinyl transferases AcpS and Sfp. *The Journal of Biological Chemistry*, 277, pp. 17023-17031

Monot M, Honoré N, Garnier T, Zidane N, Sherafi D, Paniz-Mondolfi A, Matsuoka M, Taylor GM, Donoghue HD, Bouwman A, Mays S, et al (2009). Comparative genomic and phylogenetic analysis of *Mycobacterium leprae. Nature Genetics*, 41, pp. 1282-1289

Moore GR and Pettigrew GW (1990). Cytochromes *c* evolutionary, structural and physicochemical aspects. Springer-Verlag; New York

Moore LR, Rocap G and Chisholm SW (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature,* 393, pp. 464-467

Moore CM and Helman JD (2005). Metal ion Homeostasis in *Bacillus subtilis. Current Opinion in Microbiology*, 8, pp. 188-195

Morona JK, Miller DC, Morona R and Paton JC (2004). The effect that mutations in the conserved capsular polysaccharide biosynthesis genes *cpsA, cpsB* and *cpsD* have on virulence of *Streptococcus pnuemoniae. The Journal of Infectious Diseases,* 189, pp. 1905-1913

Moussatova A, Kandt C, O'Mara ML and Tieleman DP (2008). ATP-binding cassette transporters in *Escherichia coli. Biochimica et Biophysica Acta,* 1778, pp. 1757-1771

Myers CR and Myer JM (1997). Cloning and sequence of CymA, a gene encoding a tetraheme cytochrome c required for reduction of iron (III), fumarate and nitrate by *Shewanella putrefaciens* MR-1. *Journal of Bacteriology,* 179, pp. 1143-1152

Myers JM, Antholine WE and Myers CR (2004). Vanadium (v) reduction by *Shewanella oneidensis* MR-1 requires menaquinone and cytochromes from the cytoplasmic and outer membranes. *Applied and Environmental Microbiology*, 70, pp. 1405-1412

Natale P, Bruser T and Driessen AJ (2008). Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—distinct translocases and mechanisms. *Biochim Biophys Acta*, 1778, pp. 1735-1756

Neves AR, Pool WA, Kok J, Kuipers OP and Santos H (2005). Overview on sugar metabolism and its control in *Lactococcus lactis*-the input from in vivo NMR. *FEMS microbiology reviews,* 29, pp. 531-554

Nobbs AH, Lamont RJ and Jenkinson HF (2009). *Streptococcus* adherence and colonization. *Microbiology and Molecular Biology Reviews,* 73, pp. 93-104

Ohtsubo H and Ohtsubo E (1978). Nucleotide sequence of an insertion element ISI. *Proceedings of the National Academy of Science of the United States of America,* 75, pp. 615-619

Omogbai BA, Ikenebomeh MJ and Ojeaburu SI (2005). Microbial utilization of stachyose in soymilk yogurt production. *African Journal of Biotechnology,* 4, pp. 905-908

Page AJ, Alikhan N-F, Carleton HA, Seemann T, Keane JA, Katz LS (2017). Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microbial Genomics,* 3:8

Partensky F, Hess WR and Vaulot D (1999). *Prochlorococcus,* a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews,* pp. 106-127

Patwardhan A, Ray S and Roy A (2014). Molecular markers in phylogenetic studies-A Review. *Phylogenetics and Evolutionary Biology,* 2:2

Philippot L, Raaijmakers JM, Lemanceau P and van der Putten WH (2013). Going back to the rhizosphere. *Nature Reviews Microbiology,* 11, pp. 789-799

Pitts KE, Dobbin PS, Reyes-Ramirez F, Thomson AJ, Richardson DJ and Seward HE (2003). Characterization of *Shewanella oneidensis* MR-1 decaheme cytochrome mtrA: expression in *E.coli* confers ability to reduce soluble Fe (III) chelates. *Journal of Biological Chemistry,* 278, pp. 27758-27765

Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP and Moran MA (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North pacific subtropical gyre. *Environmental Microbiology,* 11, pp. 1358-1375

Price NPJ, Rooney AP, Swezey JL, Perry E and Cohan FM (2007). Mass spectroscopic analysis of lipopeptide for *Bacillus* strains isolated from diverse geographical locations. *FEMS Microbiology Letters,* 271, pp. 83-89

Rahman SA, Singh Y, Kohli S, Ahmad J, Ehtesham NZ, Tyagi AK Hasnain SE and Lee SY (2014). Comparative analyses of non-pathogenic, opportunistic, and totally pathogenic

mycobacteria reveal genomic and biochemical variabilities and highlight the survival attributes of *Mycobacterium tuberculosis. mBio,* 5:e02020

Raushel FM, Thoden JB and Holden HM (1999). The Amidotransferase family of enzymes: molecular machines for the production and delivery of ammonia. *Biochemistry,* 38, pp. 7892-7899

Reddy EA, Shaw AV and Crump JA (2010). Community-acquired bloodstream infections in Africa: a systematic review and meta-analysis. *Lancet Infectious Diseases,* 10, pp. 417-432

Reid G and Burton J (2002). Use of *Lactobacillus* to prevent infection by pathogenic bacteria. *Microbes and Infection,* 4, pp. 319-324

Reva ON, Chan WY, Bezuidt OKI, Lapa SV, Safronova LA, Avdeeva LV, Borriss R. Genetic Barcoding of Bacteria and its Microbiology and Biotechnology Applications. In Bioinformatics and Data Analysis in Microbiology, Ed. O. Tastan Bishop, Caister Academic Press. 2014. pp. 230-243

Rice PA and Baker TA (2001). Comparative architecture of transposase and integrase complexes. *Nature Structural and Molecular Biology,* 8, pp. 302-307

Rocap G, Distel DL, Waterbury JB and Chisholm SW (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Applied and Environmental Microbiology,* 68, pp. 1180-1191

Rodwell VW. Metabolism of purine and pyrimidine nucleotides Chapter 34. In Murrary-Medicine Harper's Illustrated Biochemistry 26th Edition 2003. Lange Medical Book/McGraw-Hill Medical Publishing Division

Rooney AP, Price NPJ, Ehrhardt C, Swezey JL and Bannan JD (2009). Phylogeny and molecular taxonomy of *Bacillus subtilis* species complex and description of *Bacillus subtilis* subsp. *inaquosorum* subsp.nov. *International Journal of Systemic and Evoluntionary Biology,* 59, pp. 2429-2436

Rusch DB, Matiny AC, Dupont CL, Halpern AL and Venter JC (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proceedings of the National Academy of Sciences of the United States of America,* 107, pp. 16184-16189

Safronova LA, Zelena LB, Klochko VV and Reva ON (2012). Does the applicability of *Bacillus* strains in probiotis rely on taxonomy? *Canadian Journal of Microbiology*, 58, pp. 212-219

Saito H and Tomioka H (1984). Thymidine kinase of bacteria: activity of the enzyme in actinomycetes and related organisms. *Journal of General Microbiology*, 130, pp. 1863-1870

Sanders C, Turkarslan S, Lee DW and Daldal F (2010). Cyctochrome *c* biogenesis: the Ccm system. *Trends in Microbiology*, 18, pp. 266-274

Scheutz F and Strockbine NA (2005). Genus I. *Escherichia* pp.607-624. In Garrity GM et al (ed), Bergey's manual of systemic bacteriology. Springer Publishing Company, New York, NJ

Schimmel P (1987). Aminoacyl tRNA synthetases: general scheme of structure-function relationships in the polypeptides and recognition of transfer RNAS. *Annual Review of Biochemistry*, 56, pp. 125-158

Schimmel P (2008). Development of tRNA synthetases and connection to genetic code and disease. *Protein Science*, 17, pp. 1643-1652

Sharp CE, Brady AL, Sharp GH, Grasby SE, Stott MB and Dunfield PF (2014). Humboldt's spa: microbial diversity is controlled by temperature in geothermal environments. *International Society for Microbial Ecology*, 8, pp. 1166-1174

Sharpton TJ (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5:209

Sled VD, Friedrich T, Leif H, Weiss H, Meinhardt SW, Fukumori Y, Calhoun MW, Gennis RB and Ohnishi T (1993). Bacterial NADH-Quinone oxidoreductases: iron-sulfur clusters and related problems. *Journal of Bioenergetics and Biomembranes*, 22, pp. 347-356

Simmons SL, DiBartolo G, Denef VJ, Goltsman DSA, Thelen MP and Banfield JF (2008). Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biology*, 6:e177

Smith JL and Sherman DH (2008). Biochemistry, an enzyme assembly line. *Science*, 321, pp. 1304-1305

Smith PMC and Atkins CA (2002). Purine biosynthesis. Big in cell division, even bigger in nitrogen assimilation. *Plant Physiology*, 28, pp. 793-802

Steven JM, Mavridou DA, Hamer R, Kritsiligkou P, Goddard AD and Ferguson SJ (2011). Cytochrome *c* biogenesis system I. *FEBS Journal*, 278, pp. 4170-4178

Strockbine NA and Maurelli AT (2005). Genus XXXV. *Shigella*, pp.811-823. In Garrity GM et al (ed), Bergey's manual of systemic bacteriology. Springer Publishing Company, New York, NJS

Taylor CM and Roberts IS (2005). Capsular polysaccharides and their role in virulence. Rusesell W, Herwald H (Eds). In *Concepts in Bacterial Virulence*, 12, pp. 55-66

Teitelbaum JE and Walker WA (2002). Nutritional impact of pre- and probiotics as protective gastrointestinal organisms. *Annual Review of Nutrition*, 22, pp. 107-138

Tennant SM, Diallo S, Levy H, Livio S, Sow SO, Tapia M, Fields PI, Mikoleit M, Tamboura B, Kotloff KL, Nataro JP, Galen JE and Levine MM (2010). Identification by PCR of non-typhoidal *Salmonella enterica* serovars associated with invasive infections among febrile patients in Mali. *PLoS Necglected Tropical Diseases*, 4:e621

Thavasi R (2006). Biosurfactants from marine hydrocarbonoclastic bacteria and their application in marine oil pollution abatement. Ph.D Thesis, Annamali University, India p.162

Thavasi R, Jayalakshmi S and Banat IM (2011). Application of biosurfactant produced from peanut oil cake by *Lactobacillus delbrueckii* in biodegradation of crude oil. *Bioresource Technology*, 102, pp. 3366-3372

Trofa AF, Ueno-Oslen H, Oiwa R and Yoshikawa M (1999). Dr. Kiyoshi Shiga: discoverer of the dysentery *bacillus. Clinical Infectious Diseases*, 29, pp. 1303-1306

Urwin R and Maiden MCJ (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology*, 11(10), pp. 479-487

van Belkum A, Struelens M, de Visser A, Verbrugh H and Tibayrenc M (2001). Role of genomic typing in taxonomy, evoluntionary genetics, and microbial epidemiology. *Clinical Microbiology Reveiws*, 14, pp. 547-560

Van Der Heijden, Bargett RD and Vanstralen NM (2008). The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecology Letters,* 11, pp. 296-310

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Roggers YH and Smith HO (2004). Environmental genome shotgun sequencing of the *Sargasso* Sea. *Science,* 304, pp. 66-74

Verissimo AF and Daldal F (2014). Cytochrome *c* biogenesis system 1: an intricate process catalysed by a maturase supercomplex? *Biochim Biophys Acta*, 1837, pp. 989-998

Vuyst LD and Leroy F (2007). Bacteriocins from lactic acid bacteria: production, purification, and food applications. *Journal of Molecular Microbiology and Biotechnology,* 13, pp. 194-199

Wadula J, von Gottberg A, Kilner D, de Jong G, Cohen C, Khoosal M, Keddy K and Crewe-Brown H (2006). Nosocomial outbreak of extended spectrum β-lactamase-producing *Salmonella isangi* in pediatric wards. *The Pediatric Infectious Disease Journal,* 25, pp. 843-844

Walker MJ, Barnett TC, McArthur JD, Cole JN, Gillen CM, Henningham A, Sriprakash KS, Sanderson-Smith ML and Nizet V (2014). Disease manifestations and pathogenic mechanisms of Group A *Streptococcus. Clinical Microbiology Reviews,* 27, pp. 264-301

Walter J and Ley R (2011). The human gut microbiome: ecology and recent evolutionary changes. *Annual Review of Microbiology,* 65, pp. 411-429

Wang G, Qian F, Saltikov CW, Jiao Y and Li Y (2011). Microbial reduction of Graphene oxide by *Shewanella. Nano Research,* 4, pp. 563-570

Wang and Qian (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE,* 4:e7401

Wayne LG and Sramek HA (1992). Agents of newly recognized or infrequently encountered mycobacterial diseases. *Clinical Microbiology Review,* 5, pp. 1-25

Wertheim HFL, Nghia HDT, Taylor W and Schultsz C (2009). *Streptococcus suis*: An emerging human pathogen. *Clinical Infectious Diseases*, 48, pp. 617-625

Wool IG (1996). Extraribosomal functions of ribosomal proteins. *Trends in Biochemical Sciences,* 21, pp. 164-165

World Health Organization (2007). Pneumococcal conjugate vaccine for childhood immunization: WHO position paper. *Weekly Epidemiological Record,* 82, pp. 93-104

World Health Organization/United Nations Children's Fund (WHO/UNICEF) (2005). Global immunization vision and strategy. Washington DC: World Health Organization

Wright MH, Farooqui SM, White AR, and Greene AC (2016). Production of manganese oxide nanoparticles by *Shewanella* species. *Applied and Environmental Microbiology,* 85, pp. 5402-5409

Wylie KM, Truty RM, Sharpton TJ, Mihindukulasuriya KA, Zhou Y, Gao H, Sodegren E, Weinstock GM and Pollard KS (2012). Novel bacteria taxa in the human microbiome. *PLoS ONE*, 7:e35294

Yao P and Fox PL (2012). Aminoacyl-tRNA synthetases in medicine and disease. *EMBO Molecular Medicine* 5, pp. 332-343

Zhaxybayeva O, Doolittle WF, Papke RT and Gogarten JP (2009). Interwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus* marinus. *Genome Biology and Evolution,* 1, pp. 325-339

Zinser ER, Johnson ZI, Coe A, Karaca E, Veneziano D, Chisholm SW (2007). Influence of light and temperature on *Prochloroccus* ecotype distributions in the Atlantic ocean. *Limnology and Oceanography*, 52, pp. 2205-2220

# CHAPTER 3 Program implemention for Barcoding 2.0


## Abstract

Metagenomic approaches have revealed the complexity of environmental microbiomes and advancement in WGS has led to a significant level of genetic heterogeneity on the species level. It has become clear that a superior pattern of bioactivity of bacteria applicable in biotechnology, as well as the enhanced virulence of pathogens, often requires researchers to distinguish between closely related species or sub-species. Current methods for binning of metagenomic reads usually do not allow identification below the genus level and very often stop at the level of families. In this chapter, an attempt was made to improve metagenome binning resolution using the Barcoding 2.0 program to align reads against barcode sequences and calculate various parameters for scoring the alignment results and individual barcodes. Taxonomic units were identified in metagenomic samples by comparison of the calculated barcode scores to set cut-off values.


## 3.1 Introduction

Metagenomics can be defined as a technique used for the direct investigation of genomes that contribute to an environmental sample (Handlesman *et al.,* 1998; Thomas *et al.,* 2012). Over the years, the field of metagenomics has transformed from sequencing of cloned DNA fragments using Sanger technology to direct sequencing of DNA without heterologous cloning (Tyson *et al.,* 2004; Gilbert *et al.,* 2008; Desai *et al.,* 2012). Metagenomics offers access to the functional gene composition of microbial communities, which enables a wider depiction than phylogenetic surveys, and a strong tool for creating new hypotheses of microbial functions, such as the discovery of proteorhodophsin (Beja *et al.,* 2000; Gilbert *et al.,* 2008; Desai *et al.,* 2012).

Advances in sequencing technologies have provided researchers with the ability to describe the microbial composition of environmental or clinical samples with exceptional resolutions promptly. A wealth of genetic data has become available owing to these approaches, providing new understanding of environmental and human microbial ecology (Hong *et al.,* 2012). The reduction in the cost of sequencing has also rapidly enhanced the development of

sequence-based metagenomics. The number of metagenome sequence datasets has increased dramatically in the past few years (Thomas *et al.,* 2012). Hence, metagenomics researchers have to analyse huge short-read datasets using tools designed for long reads and more specifically for clonal datasets (Desai *et al.,* 2012).

Binning is generally referred as a method used for grouping reads or contigs and assigning them to OTUs. Normally, each sequence is either classified into a taxonomic group such as OTU, genus or family through association to some referential data, or clustered into groups of sequences that denote taxonomic groups centred on common characteristics such as the GC content (Sharpton, 2014). Binning plays a key part in the analysis of metagenomes, such as: (i) depending on the approach used, binning can give understanding of the presence of new genomes that are challenging to identify; (ii) it can be used to provide better insight into the unique numbers and kinds of taxa in a given community; and (iii) binning can decrease the intricacy of data, as used in post-binning analysis in assembly that can be carried out autonomously on each set of the binned reads rather than on the whole population of data (Sharpton, 2014). There are three common types of binning algorithms, namely sequence composition, sequence similarity and fragment recruitment. Sequence compositional binning uses metagenome sequence characteristics such as tetramer frequency to cluster or classify sequences into taxonomic groups. Some of these approaches, like PhyloPithia, analyse whole genome sequences ahead of time to train classifiers that stratify sequences into taxonomic groups (McHardy *et al.,* 2007; Patil *et al.,* 2011), while other approaches, such as emergent self-organising maps, use sequence characteristics to cluster metagenomic reads into unique classes without demanding a reference database and can be used to classify earlier unidentified organisms (Dick *et al.,* 2009; Sharpton, 2014). Unlike composition-based methods, sequence similarity approaches need larger computational resources, as every read is normally aligned to a big volume of sequences. Sequence similarity-based approaches give better annotation accuracy and resolution compared to compositional binning. The MEGAN tool is one of the most commonly used sequence similarity methods usings BLAST to compare reads to a database of sequences that are annotated with NCBI taxonomy (Huson *et al.,* 2011). The fragment recruitment method identifies reads that show almost matching alignments to genome sequences, such as mapping and screen reads based on genomes to which they map. However, there are at present few tools that can handle both mapping of reads to a database of genomes and the calculation of genome abundance. One such tool is

Genometa, which provides users with a graphical user interface (Davenport *et al.*, 2012; Sharpton, 2014).

However, most methods used for binning of metagenomic reads do not allow identification below the genus level and very often stop on the level of bacterial families (Thomas *et al.*, 2012). In this work, an attempt was made to improve the metagenome binning resolution by using the novel Barcoding 2.0 program, which is available from http://bargene.bi.up.ac.za/. The program Barcoding 2.0 is a command-line program on Python 2.5/2.7 designed to align metagenome reads (Roche 454 and Illumina) against taxon-specific barcode sequences generated by the online program BarcodeGenerator (chapter 2).

### 3.2 Methods and research design

Command-line program Barcoding 2.0 is available for download from the Barcoder web portal. To validate the program, MetaSim software was used to generate collections of artificial reads simulating metagenome data sets (Richter *et al.*, 2008). Sequence alignment was performed by MUSCLE algorithm (Edgar *et al.*, 2004). Orthology prediction was done by reciprocal BLASTP implemented by an in-house Python 2.5/2.7 script. For data visualisation, *matplotlib* 1.5.1 Python module (https://matplotlib.org/1.5.1/index.html) was used. All the programs for this work were written on Python 2.5 (compatible with Python 2.7) and made accessible at the website http://bargene.bi.up.ac.za/ through a PHP framework.

### 3.3 Program implementation

Barcode sequences generated by BarcodeGenerator can be used for identification of species of interest in environmental metagenome samples sequenced by Roche 454 or Illumina technologies. Barcoding 2.0 is an application written in Python 2.5 (compatible with Python 2.7) with a command-line user interface made available for downloading from the BarcodeGenerator website (http://bargene.bi.up.ac.za/). The program uses BLASTN to align reads against the generated barcode sequences and then calculates several parameters for scoring the results of the BLASTN alignment and individual barcodes. First, read alignment records with BLASTN scores below an estimated *S'* score cut-off value are filtered out. The cut-off *S'* is calculated by equation 1:

$$S' = S + \frac{L-S}{1+e^{\frac{3(L-S)}{S \times lg(N)}}} - 10 \times \left( ln\left(\frac{2S+100}{L+100}\right) - 1 \right) \qquad (1)$$

where $S$ – an average BLASTN score of all aligned reads; $L$ – an average length of reads; and $N$ – number of aligned reads.

The program then calculates the alignment specificity ($a_{specificity}$) of read alignments (equation 2) by estimating the number of metagenomics reads ($N_{aligned\_reads}$) that were successfully aligned against the given number of barcode sequences ($N_{barcodes}$) and the total number of BLASTN matches ($N_{matches}$):

$$a_{specificity} = 1 - \frac{N_{matches} - N_{aligned\_reads}}{N_{aligned\_reads} \times (N_{barcodes} - 1)}. \qquad (2)$$

Values of specificity vary in the range from 0 to 1. The value of 0 indicates no specificity, i.e. every read in a given metagenome was aligned against every barcode sequence in the set. The value of 1 reports no overlap between reads aligned to different barcodes – maximal specificity.

Thereafter the program calculates the specificity of every read ($r_{specificity}$):

$$r_{specificity} = \frac{(N_{barcodes} - N_{read\_aligned\_barcodes})}{(N_{barcodes} - 1)}. \qquad (3)$$

It can be seen from equation 3 that if one read was aligned against all barcodes, its specificity is 0; and if the read was aligned only against one barcode, its specificity is 1.

Then the program calculates two scores, ReadScore1 and ReadScore2, for every aligned read per barcode by equations 4 and 5, respectively:

$$ReadScore1 = \frac{BLASTN\_score}{read\_length} \times \frac{r_{specificity} + EXP(r_{specpficity} \times r_{vicinity}) + 1}{r_{specificity} + EXP(r_{vicinity}) + 1} \qquad (4)$$

$$ReadScore2 = a_{specificity} \frac{N_{reads|barcode}}{N_{reads}} \times \frac{BLASTN\_score}{read\_length} \times \frac{r_{specificity} + 1.5^{(r_{specificity} \times r_{vicinity})} + 1}{r_{specificity} + 1.5^{(r_{vicinity})} + 1}. \qquad (5$$

It should be emphasised that *ReadScore2* is barcode-specific, i.e. reads aligned to several barcodes will have different *ReadScore2* values but the same value of *ReadScore1*. In

equations 4 and 5, the coefficient $r_{vicinity}$ was calculated for every read to avoid downgrading those reads, which were aligned to several barcodes of closely related organisms. First, a matrix of Jaccard distances is calculated for the set of barcodes, where the distance between two barcodes is *1 – number_of_common_reads / total_number_of_reads*. If one read was aligned to several barcodes, the parameter $r_{vicinity}$ for this read is calculated as *10 × max_barcode_subset_distance / max_matrix_distance*. Values of $r_{vicinity}$ are in the range of 0 to 10. If the read is specifically aligned against only one barcode, its $r_{vicinity}$ is 0. If the read is aligned against several barcodes of closely related organisms, the parameters $r_{vicinity}$ will be small and the read will be scored high. However, if the read is promiscuously aligned against many unrelated barcodes, the parameters $r_{vicinity}$ will be high and the read will be scored low.

After scoring all the aligned reads, the program calculates scores for every barcode to identify corresponding species in the metagenome sample. Scores *BarcodeScore1* and *BarcodeScore2* (equation 6) are calculated from *ReadScore1* (equation 4) and *ReadScore2* (equations 5) respectively and are independent of the lengths of barcode sequences.

$$BarcodeScore_i = \frac{1+\sum_i ReadScore}{1+\frac{3\times BarcodeLength_i \times \sum BLASTN\_score}{4\times \sum_i^N BarcodeLength}} - 1 \qquad (6)$$

After downloading the program archive file from the project website, the file has to be unzipped to a local directory. The structure of internal folders of the program is shown in Fig. 3.1. Diagnostic barcode sequences generated by the program BarcodeGenerator (chapter 2) should be copied to the folder *input* as FASTA files. In the example in Fig. 3.1 this is the file *Lactobacillus_barcodes.fasta*. Metagenome reads should be stored in FASTA files in a subfolder within the folder *input*. In the given example, the folder metadata was created, which contains multiple metagenome files. All these metagenome datasets will be analysed for the presence of species of interest in a single program run with individual reports for each metagenome file. Optionally, a phylogenetic tree file may be provided as an input (file *Lactobacillus.tre* in Fig. 3.1). Optimally, this tree file should contain information about all organisms indicating which barcodes are stored in the the barcode sequence file; however, the phylogenetic tree may comprise only part of these organisms and contain more organisms, which were not barcoded. Obviously, names of barcode sequences must correspond to the names in the phylogenetic tree. The tree dendrogram, if provided, will be included in the

graphical report file. The program can be run on computers with Python 2.5/2.7 installed. To run the program, the user has to double-click the file *run.py* in the top folder of the program. A command line window will appear, as shown in Figure 3.1.

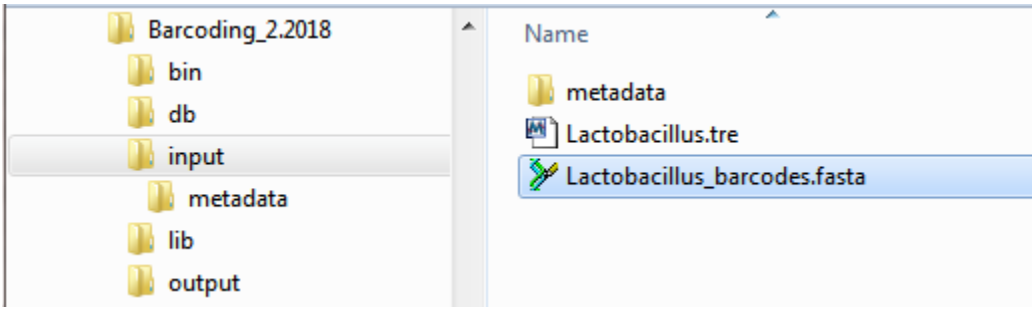



**Figure 3.1: Folders of the program Barcoder unzipped to a local directory.**

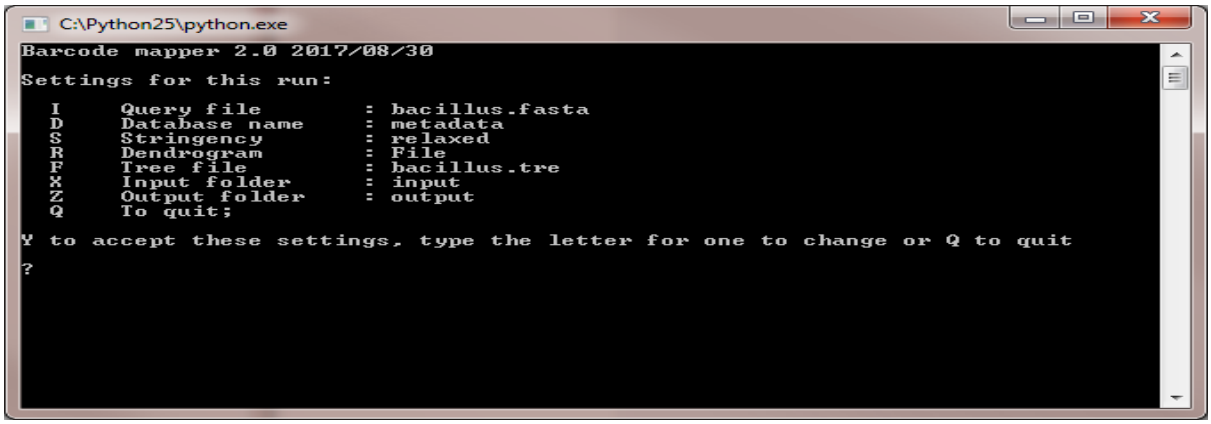



**Figure 3.2: An initial command-line window of the program Barcoding 2.**

By default the barcode file *bacillus.fasta* is set in the command-line window (Figure. 3.2). This file is provided as an example with the program download. To change the input file, use the option <I+Enter> and enter the name of the barcode file to use. Use the option <F+Enter> to change the name of the input tree file, if the file is available. If there is no such file, use <R+Enter> to clear this option. The folder with metagenome files may be changed using the option <D+Enter+*new_folder_name*+Enter>. The option <S> allows setting of species identification stringency that will be explained below in this chapter. The command-line window ready to run with the example files is shown in Figure. 3.3.

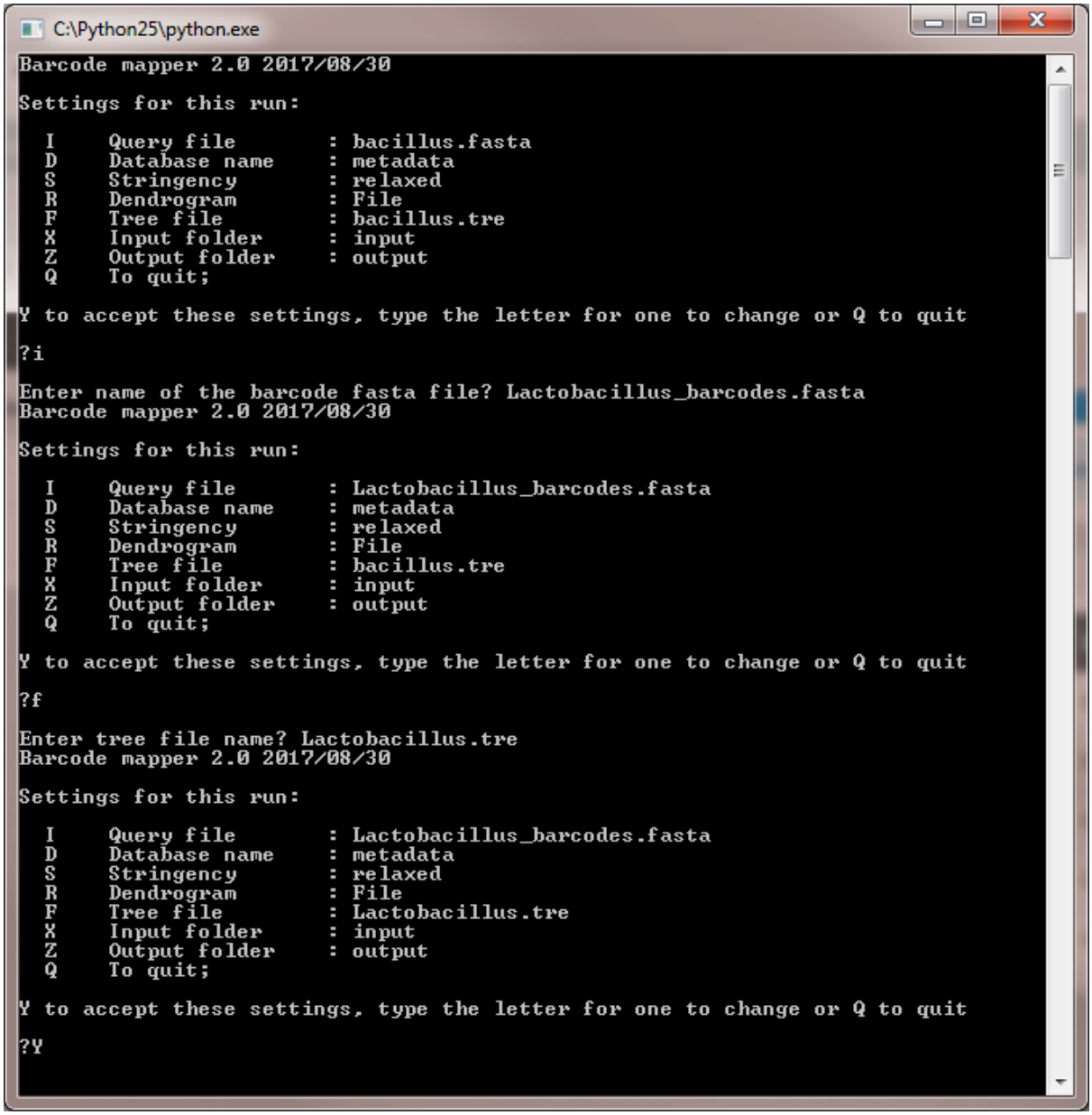



**Figure 3.3: Setting of program run options in the command line program interface.**

Keyboard combination <Y-Enter> will run the program for execution with the set parameters. Advanced users can run this program with the same settings from the command prompt line shown in Figure. 3.4.

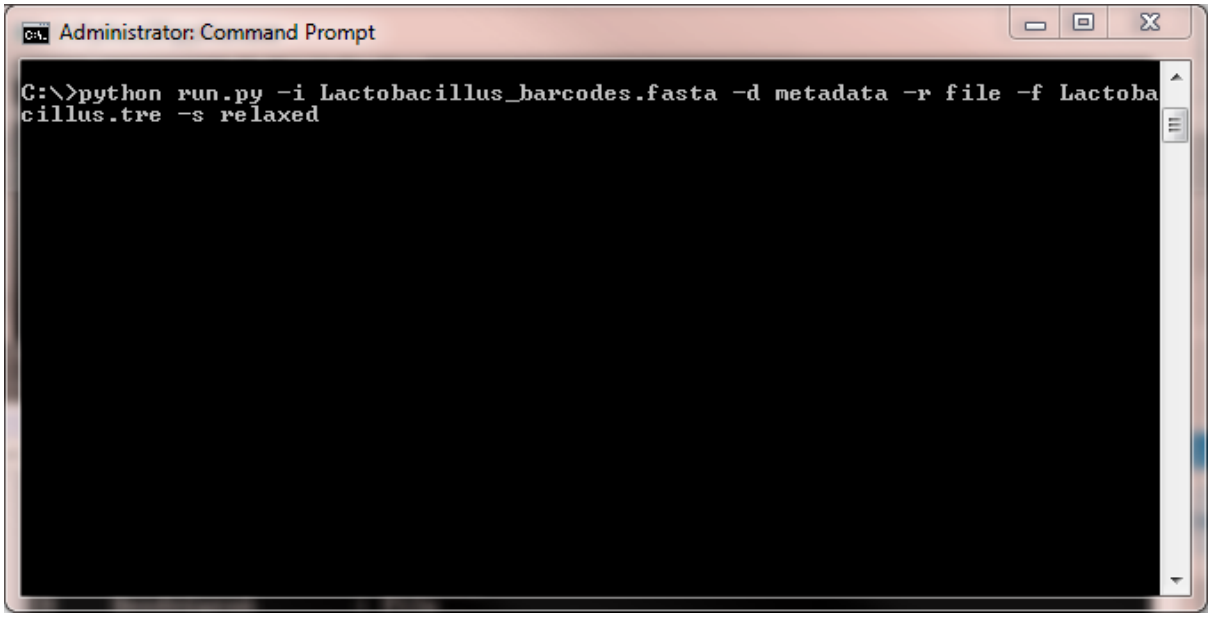



**Figure 3.4: Command prompt run of the program.**

### 3.3.1 Barcoding program workflow and identification of optimal program run parameters

An overview of the program workflow is shown in Figure. 3.5.

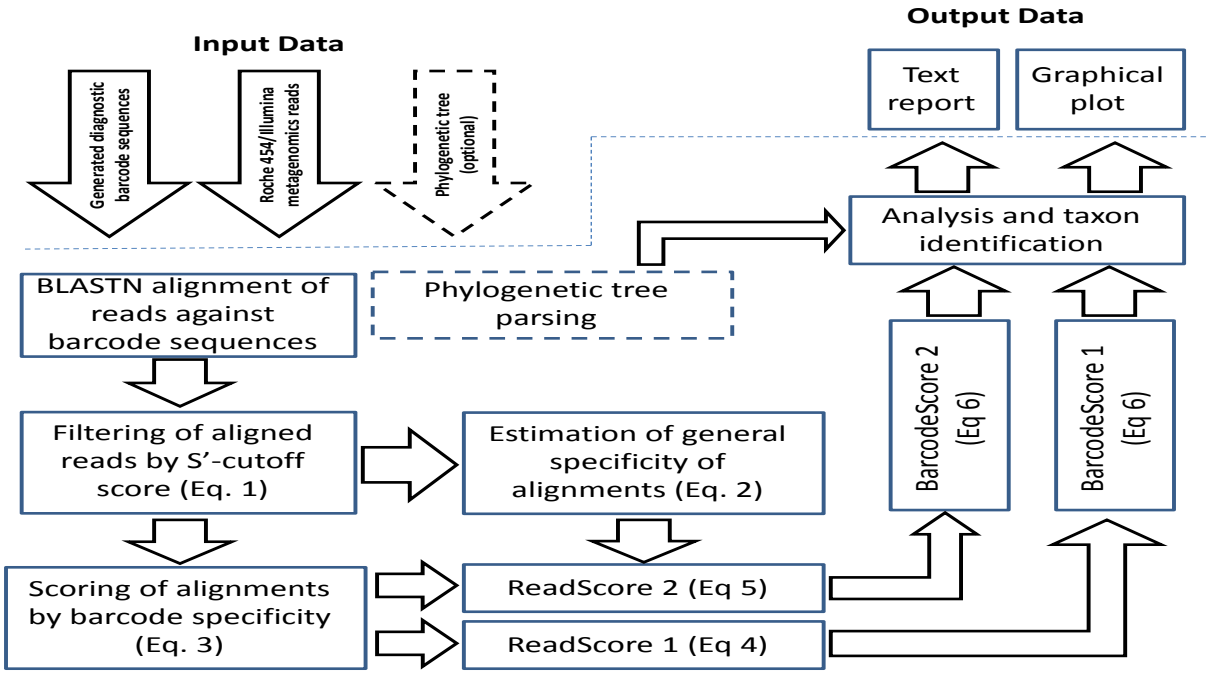



**Figure 3.5: An overview of how Barcoding 2.0 program works.**

The program aligns metagenome reads against barcode sequences and then performs statistical analysis as explained above by equations 1-6. In the next step, the program generates graphical and text output files. The text output file contains a list of barcoded genomes with assigned scores. The graphical SVG file presents these scores in the form of a

histogram. If a phylogenetic tree is provided, barcode bars in this graph are distributed along the corresponding nodes of the phylogenetic tree.

The program uses cut-off values of the barcode scores (see equation 6) to evaluate the results of identification of barcoded organisms in metagenomic samples. To validate the program and identify optimal settings of cut-off values, an artificial metagenome was created comprising DNA reads generated by the program MetaSim from several reference genomes (Table 3.1).

**Table 3.1: Composition of the artificial metagenomic dataset generated by MetaSim from reference *Shewanella, Escherichia, Shigella, Lacobacillus* and *Mycobacterium* genomes.**

| **Reference genomes** | | | **Number of reads (200-500bp)** | | | | |
|---|---|---|---|---|---|---|---|
| | | | **1** | **2** | **3** | **4** | **5** |
| ***Shewanella*** | | | **10,000** | **50,000** | **100,000** | **300, 000** | **500,00** |
| *S. amazonensis* SB2B | NC_008700 | 5% | 5 | 5 | 5 | 5 | 5 |
| *S. frigidmarina* NCIMB 400 | NC_008345 | 10% | 10 | 10 | 10 | 10 | 10 |
| *Shewanella sp.* MR-4 | NC_008321 | 15% | 15 | 15 | 15 | 15 | 15 |
| ***Escherichia/Shigella*** | | | | | | | |
| *E. coli* ATCC8739 | NC_010468 | 5% | 5 | 5 | 5 | 5 | 5 |
| *E. coli* BL21 | NC_012947 | 10% | 10 | 10 | 10 | 10 | 10 |
| *Shigella dysenteriae* Sd197 | NC_007606 | 15% | 15 | 15 | 15 | 15 | 15 |
| ***Lactobacillus*** | | | | | | | |
| *L. sanfranciscensis* TMW1 | NC_015978 | 5% | 5 | 5 | 5 | 5 | 5 |
| *L. plantarum* WCFS1 | NC_004567 | 10% | 10 | 10 | 10 | 10 | 10 |
| *L. fermentum* IFO3956 | NC_010610 | 15% | 15 | 15 | 15 | 15 | 15 |
| ***Mycobacterium*** | | | | | | | |
| *M. avium* Env77 | NC_008595 | 5% | 5 | 5 | 5 | 5 | 5 |
| *M. abscessus* ATCC 19977 | NC_010397 | 5% | 5 | 5 | 5 | 5 | 5 |

Values for *BarcodeScore1* and *BarcodeScore2*, which are dependent on the percentage of reads in a metagenome, are shown in Fig. 3.6A and B, respectively. *BarcodeScore1* was more sensitive to the presence of specific reads in metagenomes. It may be appropriate for a quantitative identification of taxa, while *BarcodeScore2* reflects the abundance of specific reads in metagenomes better.

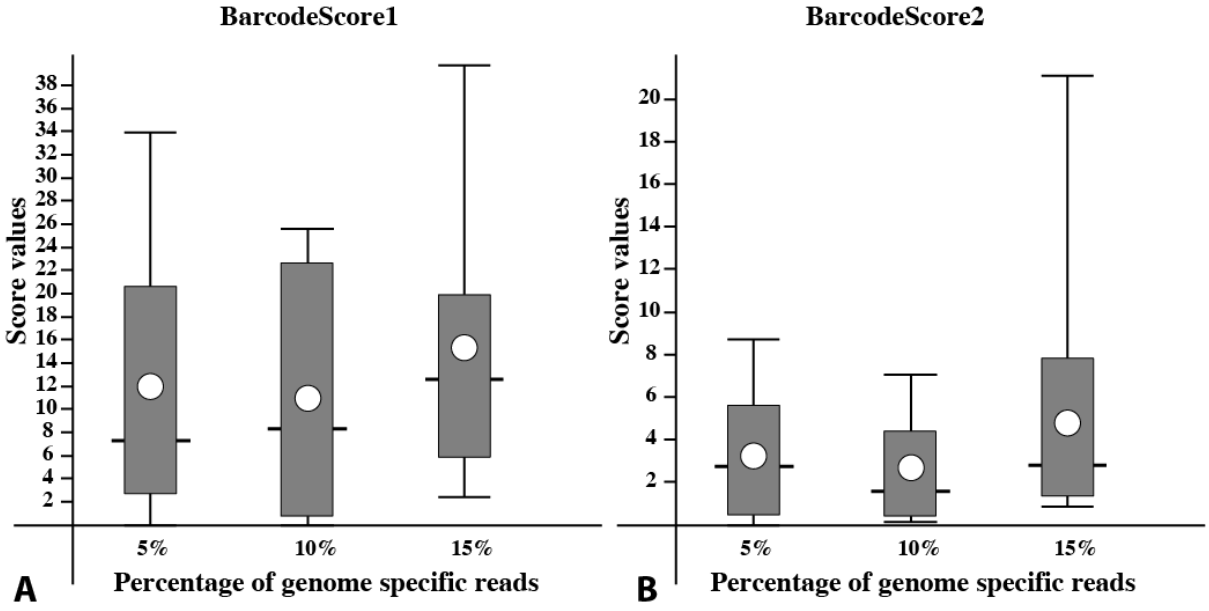



**Figure 3.6: Distribution of calculated values for A) BarcodeScore1 and B) BarcodeScore2 based on the percentage of genome specific reads in artificial metagenomes. Whisker lines depict the minimal, maximal and median values; grey bars show middle quartiles and the open cycles indicate the average values.**

Taxonomic units are identified in metagenomic samples by comparison of the calculated barcode scores to pre-computed cut-off values. True positives (TP) would be the genomes that were used for preparation of the artificial metagenomes and correctly identified by the program. Those genomes were false negative (FN), which the program failed to identify. False identification of other genomes represented in a set of barcodes leads to false positives (FP), but if excluded from the program output, they are true negatives (TN). To evaluate the barcoding performance with different cut-off values, parameters of sensitivity, specificity and the ratio of TPs over false predictions TP/(FP + FN), were calculated.

The distribution of values for TP/(FP + FN) calculated for a matrix of combinations of *BarcodeScore1* and *BarcodeScore2* cut-offs is shown in Table 3.2 and Figure 3.7. The highest proportion of TPs over false predictions was achieved for the pair of cut-offs *BarcodeScore1* = 2.5 and *BarcodeScore2* = 1. However, in the program the cut-off values

*BarcodeScore1* = 2.3 and *BarcodeScore2* = 0.5 were set by default as the *relaxed* mode to allow for higher sensitivity in case of an increase in the number of false positives. The setting *BarcodeScore1* = 2.5 and *BarcodeScore2* = 1.0 is available as the *stringent* mode. Switching between the *relaxed* and *stringent* modes is performed by using the option <S> in the command line interface (see Fig. 3.3).

**Table 3.2: TP / (FP + FN) values calculated for a matrix of combinations of *BarcodeScore1* and *BarcodeScore2* cut-offs. Combinations of pairs of score cut-off values for the relaxed and stringent operation modes are highlighted.**

| BarcodeScore 2 cutoff values | BarcodeScore 1 cutoff values | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 2.1 | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 | 2.9 |
| 0.5 | 0.59 | 0.59 | 0.62 | 0.62 | 0.62 | 0.62 | 0.63 | 0.63 | 0.64 | 0.65 |
| 0.6 | 0.64 | 0.64 | 0.66 | 0.66 | 0.66 | 0.66 | 0.66 | 0.65 | 0.65 | 0.66 |
| 0.7 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 | 0.69 | 0.69 | 0.68 | 0.69 |
| 0.8 | 0.74 | 0.74 | 0.75 | 0.75 | 0.74 | 0.75 | 0.74 | 0.73 | 0.72 | 0.72 |
| 0.9 | 0.81 | 0.81 | 0.82 | 0.82 | 0.81 | 0.82 | 0.81 | 0.81 | 0.79 | 0.79 |
| 1 | 0.84 | 0.84 | 0.85 | 0.85 | 0.85 | **0.85** | 0.85 | 0.84 | 0.83 | 0.83 |
| 1.1 | 0.79 | 0.79 | 0.79 | 0.79 | 0.79 | 0.79 | 0.78 | 0.78 | 0.77 | 0.77 |
| 1.2 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.75 | 0.75 |
| 1.3 | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.78 | 0.77 | 0.77 |
| 1.4 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.74 | 0.74 |

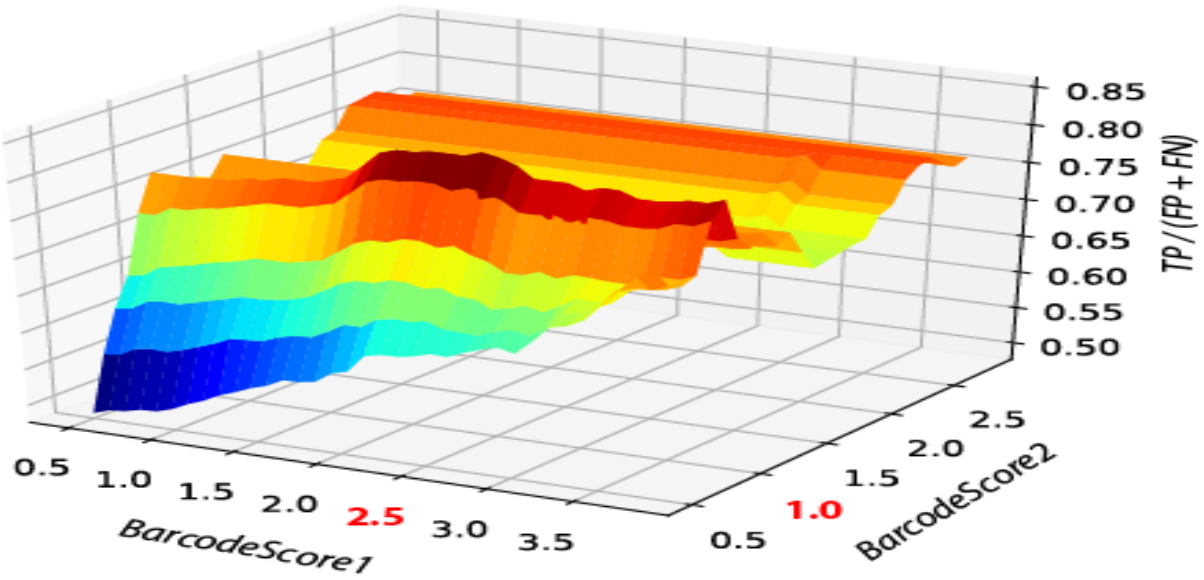



**Figure 3.7 Surface plotting of the distribution of TP / (FP + FN) values calculated for different pairs of cut-off values of the *BarcodeScore* 1 and 2.**

The barcoding program in the relaxed and stringent modes was used for processing of artificial metagenomes of different sample sizes on the generated barcodes of different lengths. It was found that the sample size (number of reads in a metagenome) had no effect on the sensitivity and specificity of the algorithm in the range from 10 000 to 500 000 (Table 3.3, Figure 3.8A and calculated receiver operating characteristics (ROC) diagrams are shown in Appendix 1. In all these experiments, metagenomic datasets of different size were aligned against barcodes of the same sequence length of 50 000 bp. However, in this range of values, the percentage of TPs grew with the sample size proportionally with the number of FPs. This is illustrated in a series of output files calculated for the artificial metagenomes aligned against *Shewanella* barcode sequences, as displayed in Figure 3.10. This series of figures demonstrates an increasing number of identified genomes (green bars in the figures), both TPs and FPs, with the increase in the sample size. The artificial metagenome comprised fragments generated from three *Shewanella* genomes: *Shewanella* sp. MR-4 [NC_008321] ─ 15% of reads of the metagenome), *S. frigidmarina* NCIMB 400 [NC_008345] ─ 10% and *S. amazonensis* SB2B [NC_008700] ─ 5%. In Figure 3.10A, the size of the metagenome was 10 000 reads. The strains NC_008321 and NC_008345 were reliably identified (green bars), while the minority strain NC_008700 was putatively detected (orange bar), together with two other strains, NC_008577 and NC_008700, which were not present in the metagenome (FPs). Figure 3.10B shows the results of barcoding, when the size of the metagenome was increased to 50 000 reads. In this case, all three strains of *Shewanella* included in the metagenome were identified, together with two FPs. In Figures 3.10C and D the size of the metagenome was progressively increased, which caused an increase in identification scores for both TP and FP predictions. False positive predictions may result from close phylogenetic relatedness between barcoded strains, which will be discussed below in this chapter and in chapter 5.

The ratio TP / (FP + FN) was generally higher in smaller metagenomes (see Table 3.3, Figure 3.8A).

**Table 3.3: Influence of metagenome sample size on the program performance**

| Sample size | Operation mode | AUC | Sensitivity | Specificity | TP / (FP + FN) |
|---|---|---|---|---|---|
| 10000 | relaxed | 0.95 | 0.72 | 0.98 | 1.60 |
| 50000 | relaxed | 0.97 | 0.81 | 0.96 | 1.50 |
| 100000 | relaxed | 0.97 | 0.81 | 0.92 | 0.90 |
| 300000 | relaxed | 0.98 | 0.91 | 0.86 | 0.66 |
| 500000 | relaxed | 0.99 | 0.91 | 0.81 | 0.47 |
| 10000 | stringent | 0.94 | 0.54 | 0.97 | 0.85 |
| 50000 | stringent | 0.97 | 0.82 | 0.96 | 1.50 |
| 100000 | stringent | 0.96 | 0.73 | 0.95 | 1.0 |
| 300000 | stringent | 0.97 | 0.82 | 0.94 | 1.13 |
| 500000 | stringent | 0.98 | 0.82 | 0.94 | 1.13 |

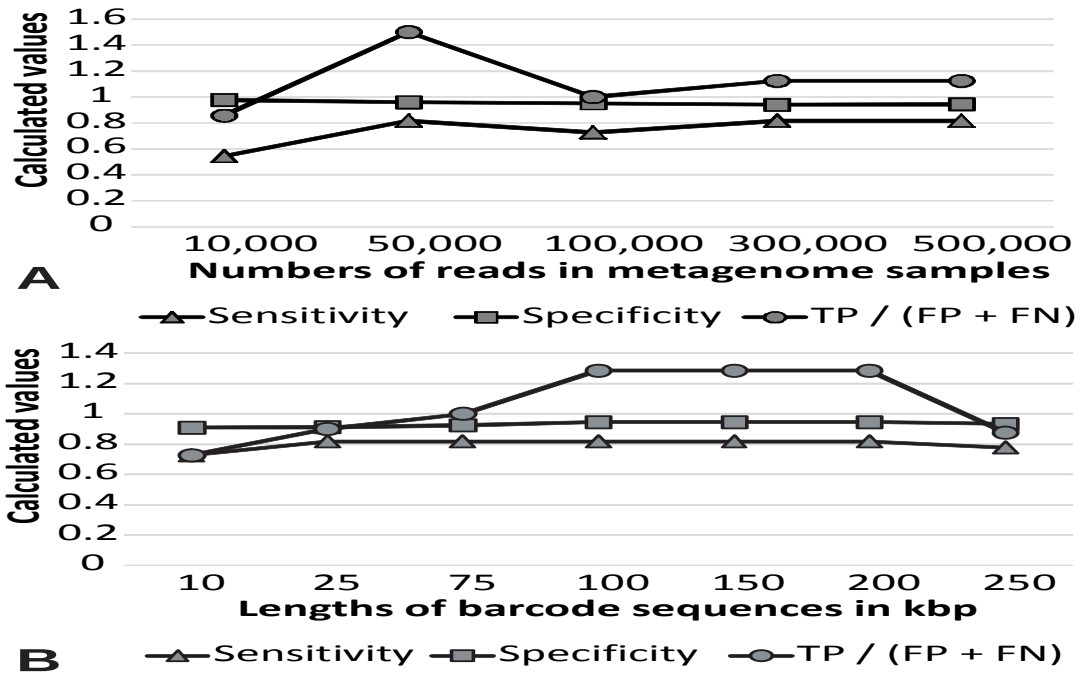



**Figure 3.8: Influence of the A) metagenome sample size and B) length of barcode sequence on the program performance.**

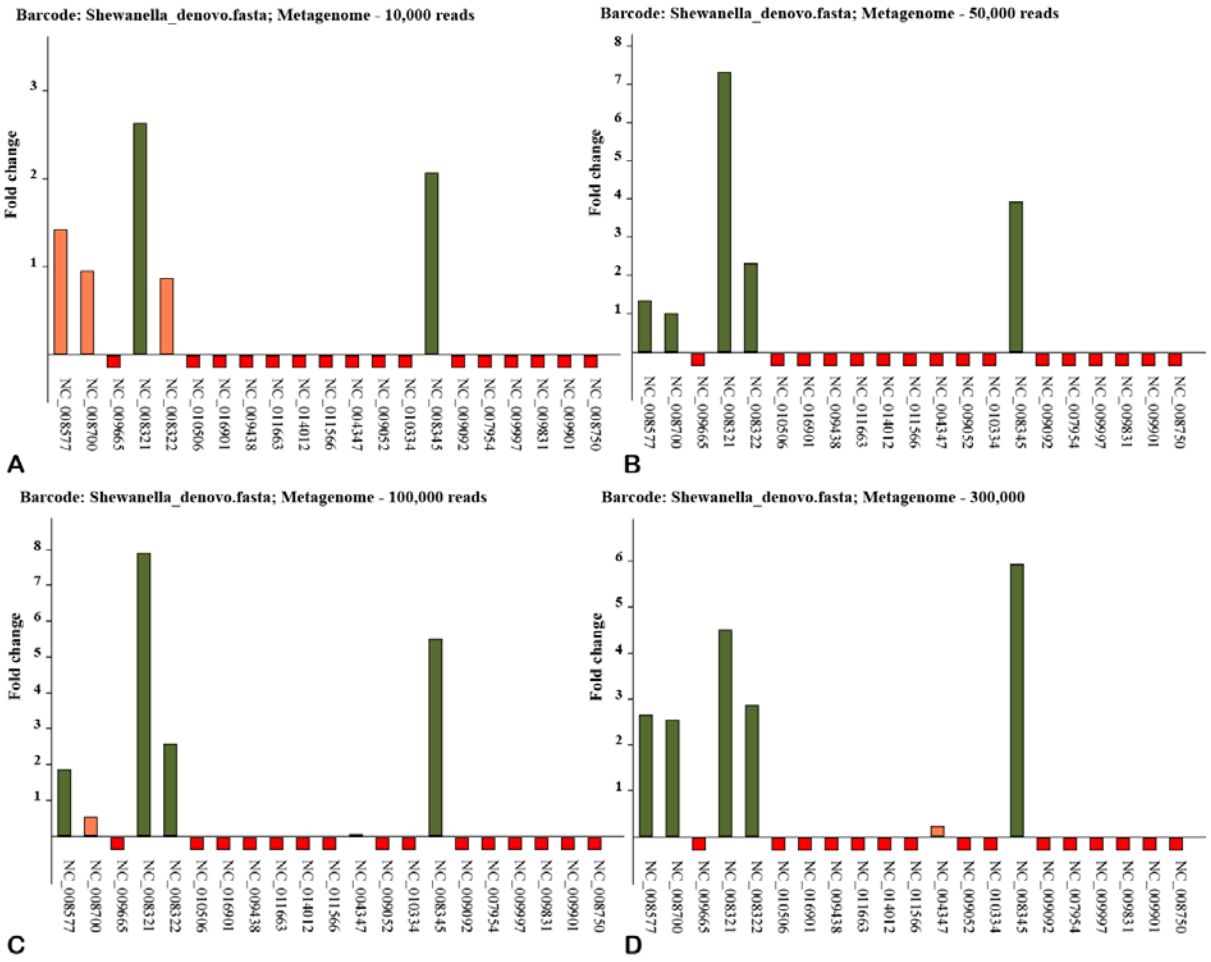



**Figure 3.9: Results of *Shewanella* strain identification in artificial metagenomes of different sizes: A) 10,000; B) 50,000; C) 100,000 and D) 300,000 reads. The strains *Shewanella* sp. MR-4 [NC_008321], *S. frigidmarina* NCIMB 400 [NC_008345] and *S. amazonensis* SB2B [NC_008700] comprised 15%, 10% and 5% of the total number of reads, respectively. Identification of other barcoded strains was considered as false-positives**

In another set of experiments, the influence of the barcode sequence lengths on barcoding accuracy was examined. It was found that the specificity and sensitivity were constant and independent on lengths of barcode sequences in the range from 10 to 250 kbp (Table 3.4, Figure 3.8B and see also calculated ROC diagrams in Appendix 2). However, the ratio TP/(FP + FN) was optimal when the barcode sequences were in the range from 100 to 200 kbp. Shorter barcodes reduced the number of TPs, as many reads remained unidentified and longer barcodes increased the number of FP predictions. The influence of the barcode sequence length on the program performance was tested on artificial metagenomic datasets with 500 000 randomly generated reads.

**Table 3.4: The influence of the length of barcode sequence on the program performance**

| Barcode length (kbp) | Operation mode | AUC | Sensitivity | Specificity | TP / (FP + FN) |
|---|---|---|---|---|---|
| 10 | relaxed | 0.89 | 0.73 | 0.88 | 0.57 |
| 25 | relaxed | 0.94 | 0.91 | 0.86 | 0.71 |
| 75 | relaxed | 0.93 | 0.82 | 0.89 | 0.75 |
| 100 | relaxed | 0.94 | 0.82 | 0.89 | 0.75 |
| 150 | relaxed | 0.93 | 0.82 | 0.89 | 0.75 |
| 200 | relaxed | 0.89 | 0.82 | 0.90 | 0.82 |
| 250 | relaxed | 0.92 | 0.78 | 0.9 | 0.64 |
| 10 | stringent | 0.89 | 0.73 | 0.91 | 0.73 |
| 25 | stringent | 0.93 | 0.82 | 0.91 | 0.90 |
| 75 | stringent | 0.93 | 0.82 | 0.92 | 1.0 |
| 100 | stringent | 0.93 | 0.82 | 0.95 | 1.28 |
| 150 | stringent | 0.93 | 0.82 | 0.95 | 1.29 |
| 200 | stringent | 0.88 | 0.82 | 0.95 | 1.29 |
| 250 | stringent | 0.91 | 0.78 | 0.93 | 0.87 |

### 3.3.2 Program performance on different groups of microorganisms

Program performance was affected by the level of taxonomic relatedness between barcoded organisms. Receiver operating characteristics curves were calculated for different taxonomic groups based on the results of identification of corresponding genomes in artificial metagenomic datasets (Table 3.5 and Figure 3.12). In addition to sensitivity and specificity parameters, the area under curve was calculated, which is considered a performance measure of diagnostic tools. Distinguishing between species of the same genus or family by the program was close to optimal. However, it was problematic for the program to differentiate between representatives of different lineages of *Escherichia* and *Shigella* (Figure 3.13). It was assumed that the addition of accessory genes in barcodes may improve the diagnostic performance. Comparison of identification results when the barcodes of the

*Escherichia/Shigella* group of the same length (150 000 bp) with different proportions of core and accessory genes were used is shown in Figure 3.14. It was found that an increase of accessory genes in barcodes hampered distinguishing between closely related organisms even more, compared to barcodes based solely on core genes. It may be explained by related organisms exchanging frequently mobile elements in a random fashion, which impedes proper differentiation between them. However, including species-specific accessory genes may improve identification on higher taxonomic levels.

**Table 3.5: shows the ROC result calculated for different taxonomic groups**

| Group of micro-organisms | Operation mode | AUC | Sensitivity | Specificity | TP / (FP + FN) |
|---|---|---|---|---|---|
| Ecol_Shig | relaxed | 0.74 | 0.47 | 0.87 | 0.19 |
| Lactobacillus | relaxed | 1 | 1 | 0.96 | 3 |
| Mycobacteria | relaxed | 1 | 0.9 | 0.95 | 1.8 |
| Shewanella | relaxed | 1 | 1 | 0.89 | 1.5 |
| Ecol_Shig | stringent | 0.7 | 0.33 | 0.98 | 0.33 |
| Lactobacillus | stringent | 1 | 0.93 | 0.96 | 2.33 |
| Mycobacteria | stringent | 1 | 0.9 | 0.94 | 1.8 |
| Shewanella | stringent | 1 | 0.87 | 0.9 | 1.18 |

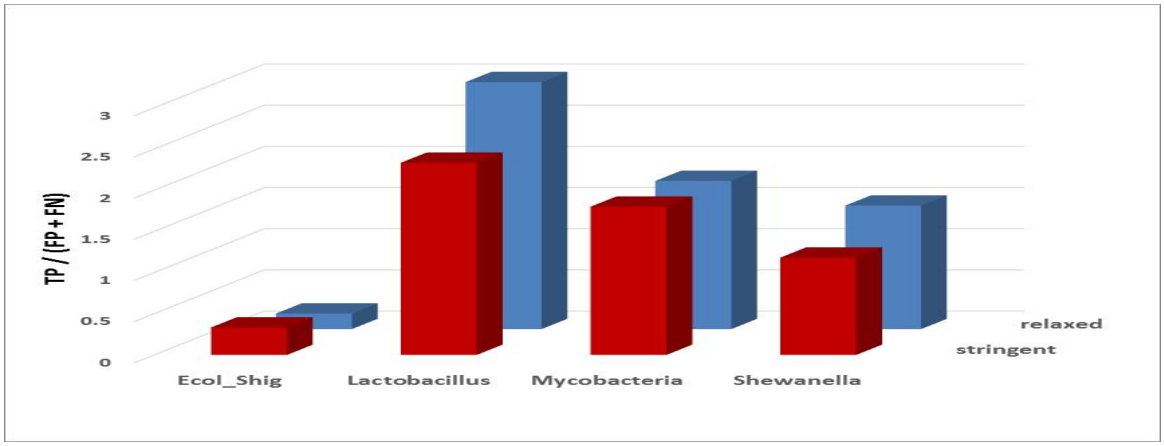



**Figure 3.10: Histogram for the taxonomic relatedness between organisms used as case study**

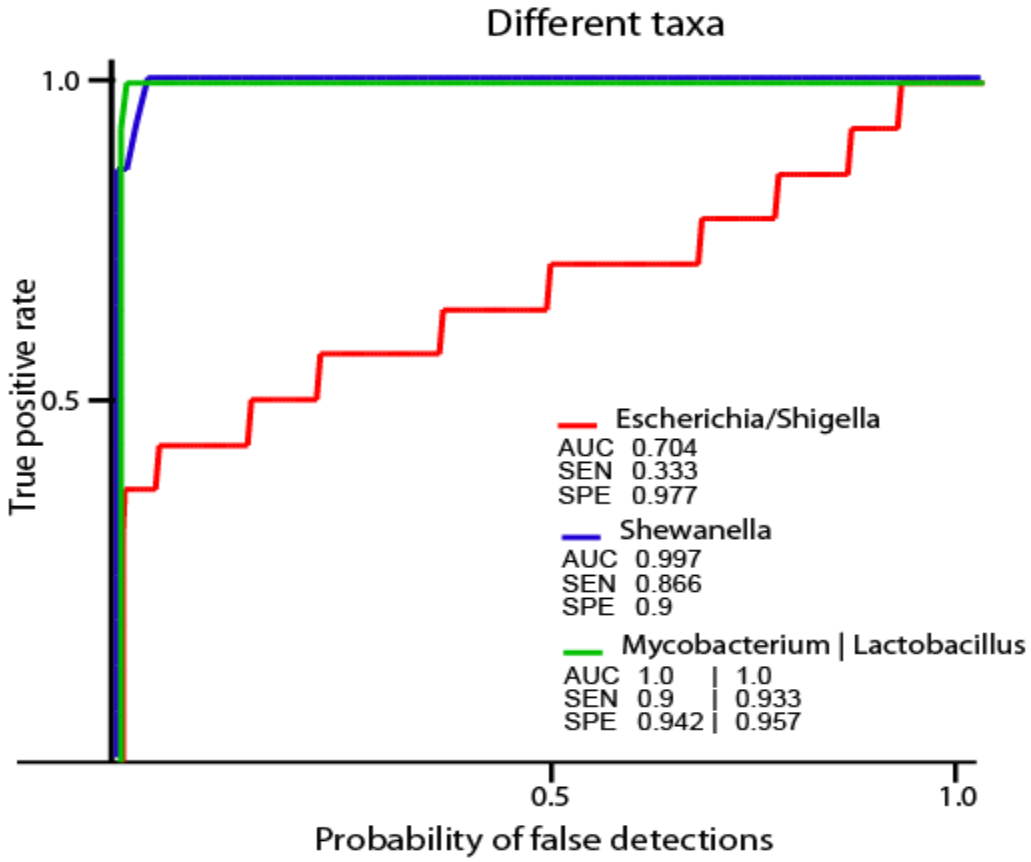



**Figure 3.11: ROC diagrams of barcoding of genomes on different taxonomic levels. The following parameters were calculated: AUC – area under the curve; SEN – sensitivity; and SPE – specificity**

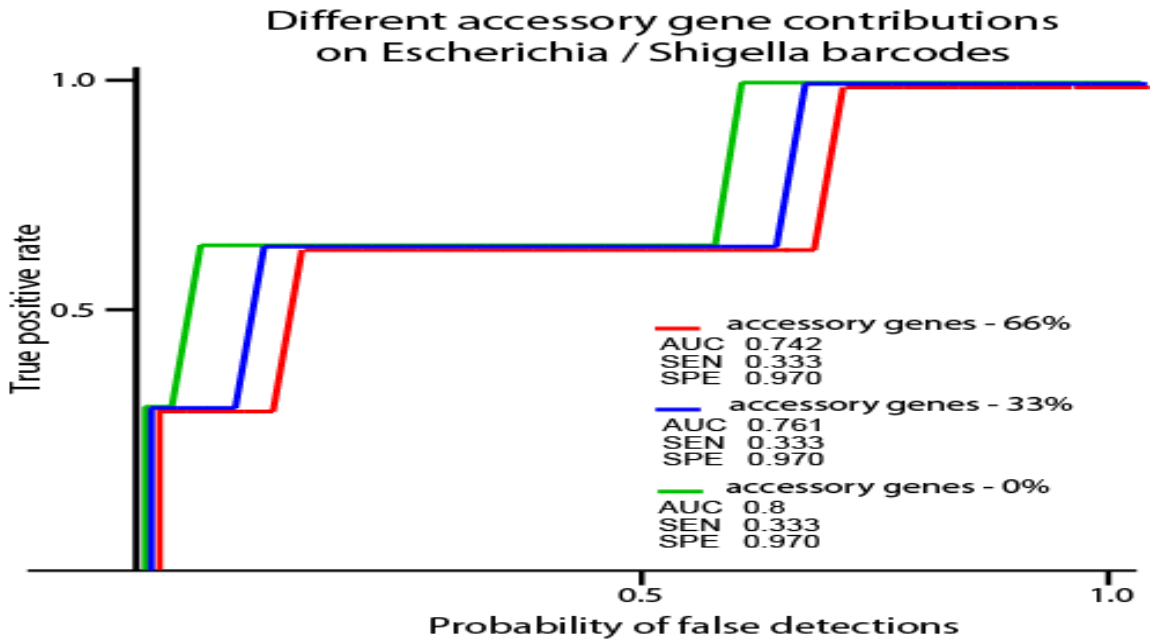



**Figure 3.12: ROC diagrams of barcoding of genomes of the *Escherichia / Shigella* group by barcodes with different contribution of accessory genes. The following parameters were calculated: AUC – area under the curve; SEN – sensitivity; and SPE – specificity.**

## 3.4 CONCLUSION

In this chapter, a novel command-line program, Barcoding 2.0, was used for binning of metagenomic reads against barcode sequences generated with the BarcodeGenerator. The program MetaSim, which is a sequencing simulator, was used to generate collections of DNA reads from chosen bacterial genomes to design artificial metagenomic datasets with known species composition and specie abundance. Metagenomes of different sample sizes (of 10 000, 50 000, 100 000, 300 000 and 500 000 bp) were generated by random selection of DNA fragments of a specified range of length from the selected reference organisms to simulate next-generation sequencing.

The program (Barcoding 2.0) uses BLASTN to align reads against barcode sequences and then calculates scores for the BLASTN alignment and individual barcodes. After scoring all the aligned reads, the program calculates scores for every barcode to identify organisms present in metagenome samples. Taxonomic units are identified by comparison of calculated barcode scores to standard cut-off values set by default.

The researcher also performed two experiments using varying metagenomes of different sample size and barcode sequences of different lengths. In the first experiment, metagenomic datasets of varying sizes of 10 000 to 500 000 reads were aligned against barcodes of the same length (50 kbp). It was found that the sample size (the number of reads in a metagenome) has no effect on the sensitivity or specificity of the algorithm. In this range of values, the percentage of TPs increased with the sample size proportionally to the number of false positives. The ratio of TPs over false prediction was higher in smaller metagenomes. Furthermore, when varying lengths of barcode sequences (10 to 250 kbp) were used for aligning an artificial metagenomic dataset of 500 000 bp, the sensitivity and specificity remained unchangeable. However, the ratio of the TPs over FNs was optimal when the barcode sequences were in the range from 100 to 200 kbp.

Receiver operating characteristic curves of the algorithm performance were calculated for all experiments with artificial metagenomics datasets. Distinguishing between species of the same genus or family by the program was close to perfect, but in distinguishing between strains of *Escherichia coli* and *Shigella* the program fared worse. Closely related organisms could be identified better when barcodes were based solely on core genes.

Hence, Barcoding 2.0 enables efficient and practical use of metabarcodes for visualisation of distribution of organisms of interest in environmental and clinical samples. Barcoding 2.0 is available for download from the same source as the BarcodeGenerator (http://bargene.bi.up.ac.za/).

**References**

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN and DeLong EF (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science,* 289, pp. 1902-1906

Davenport CF, Neugebauer J, Beckmann N, Friedrich B, Kameri B, Kokott S, Paetow M, Siekmann B, Wieding-Drewes M, Wienhöfer M, Wolf S, Tümmeler B, Ahlers V and Sprengel F (2012). Genometa- a fast and accurate classifier for short metagenomic shotgun reads. *PLoS ONE,* 7:e41224

Desai N, Antonopoulous D, Gilbert JA, Glass EM and Meyer F (2012). From genomicsto metagenomics. *Current Opinion in Biotechnology,* 23, pp. 72-76

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP and Banfield JF (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, 10:R85

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26, pp. 2460-2461

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P and Joint I (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One,* 3:e3042

Handelsman J, Rondon MR, Brady SF, Clardy J and Goodman RM (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry and Biology,* 5, pp. R245-R249

Hong S, Bunge J, Leslin C, Jeon S and Epstein SS (2009). Polymerase chain reaction primer miss half of rRNA microbial diversity. *The ISME Journal,* 3, pp. 1365-1373

Huson DH, Mitra S, Ruscheweyh HJ, Weber N and Schuster SC (2011). Integrative analysis of environmental sequences using MEGAN$_4$. *Genome Research,* 21, pp. 1552-1560

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P and Rigoutsos I (2007). Accurate phylogenetic classification of variable length DNA fragments. *Nature Methods,* 4, pp. 63-72

Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T and McHardy AC (2011). Taxonomic metagenome sequence assignment with structured output models. *Nature Methods,* 8, pp. 191-192

Richter DC, Ott F, Auch AF, Schmid R and Huson DH (2008). MetaSim-a sequencing simulator for genomics and metagenomics. *PLoS One,* 3:e3373

Sharpton TJ (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science,* 5:209

Thomas T, Gilbert J and Meyer F (2012). Metagenomics-a guide from sampling to data analysis. *Microbial Informatics and Experimentation,* 2:3

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS and Branfield JF (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature,* 428, pp. 37-43

# CHAPTER 4: Barcoder web interface and case study of barcode-guided species detection


## Abstract

In this chapter the Barcoder software tools are discussed and a case study of barcode-guided species is provided. A detailed explanation is provided on using the Barcoder software tools. The web page/web application, Barcoder, is available at http://bargene.bi.up.ac.za.


## 4.1 Introduction

The Barcoder web application serves as an interactive computational service for identification of the most suitable marker sequences for DNA for multi-local barcoding. In this chapter the researcher discusses in detail how the Barcorder web application works. For framework data visualisation, *matplotlib* 1.5.1 Python module (https://matplotlib.org/1.5.1/index.html) was used. This web application was made accessible at the webpage http://bargene.bi.up.ac.za/ through a PHP

## 4.2 BarcodeGenerator

This webpage provides users with online access to the program BarcodeGenerator, which creates diagnostic barcodes based on the genome sequences of species of interest submitted by users (Figure 4.1). The computational algorithm implemented in this program was described in detail in Chapter 2. The program BarcodeGenerator allows for the creation of barcode sequences based on a given set of genomes. It compares all pairs of genomes and selects barcodes (DNA sequences) from core and accessory genes, depending on the program run parameters. The program allows addition of accessory genes, which are believed to be genome-specific and may improve the sensitivity of the barcode sequences. However, in Chapter 3 it was demonstrated that the addition of accessory genes to barcode sequences of closely related organisms (the *Escherichia-Shigella* group was considered) may worsen the sensitivity of barcoding owing to random sharing of horizontally transferred genes by these organisms. The researcher may suggest the use of accessory genes to distinguish between closely related species, but not sub-species or lineages of the same species.

To generate a set of barcode sequences, the user has to upload corresponding genome sequences in GenBank format in a single archived file. The archived file has to have a minimum of three sequence files, but the maximum file size has to be below 500 MB.

Uploading of the input file is performed by using a corresponding key in the web interface (Figure 4.1). The user may then change the proportion of accessory genes in the generated barcode sequences and request the approximate length of barcode sequences. The project name is entered alongside the e-mail address, which will be used to provide the user with links to output files with generated barcode sequences and other supporting information.

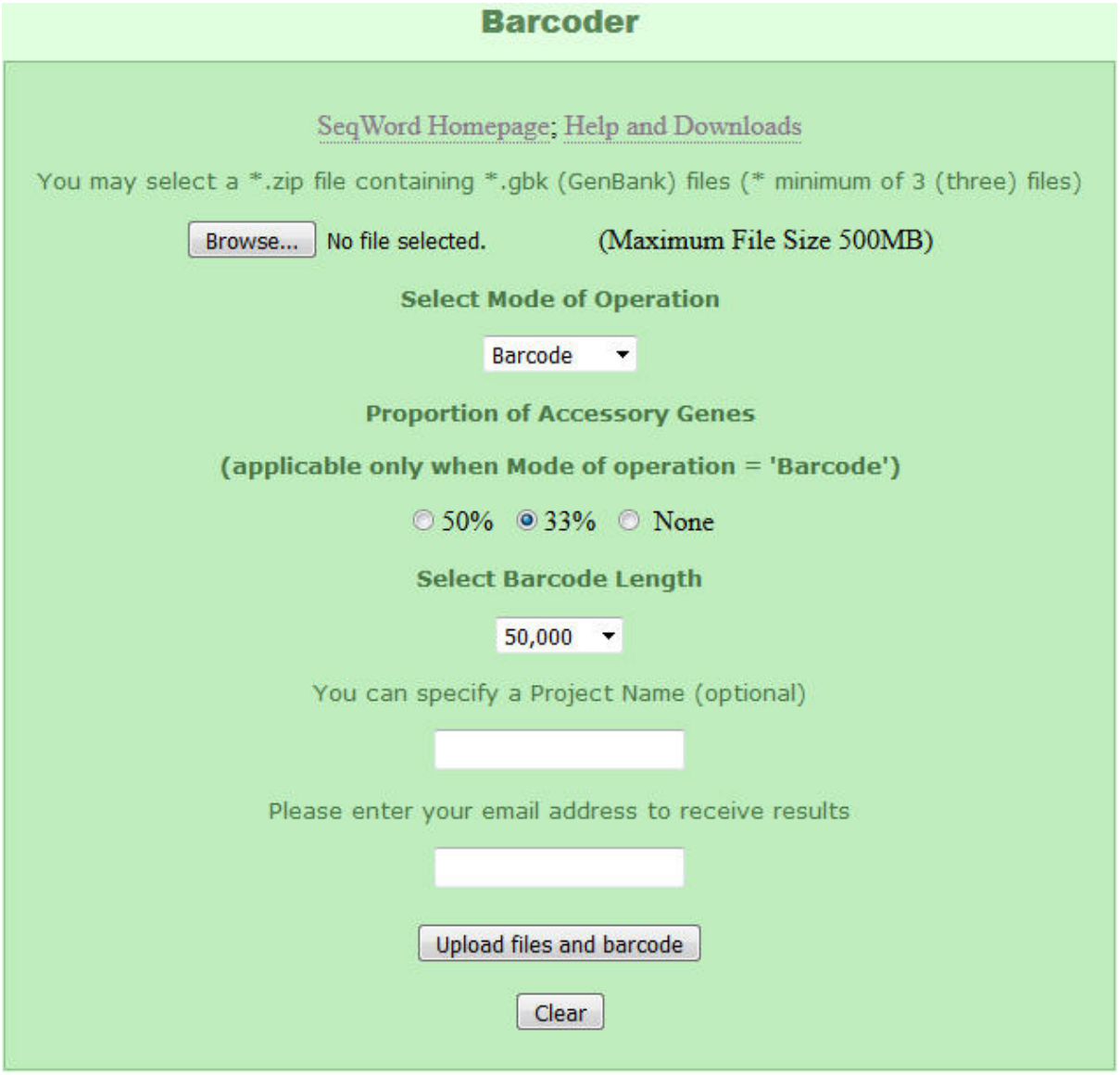



**Figure 4.1: The screenshot of the BarcodeGenerator Web-interface**

By default, the program will look in provided genomes for genes most suitable for identification (barcoding) of these organisms. Several alternative algorithms were implemented and may be chosen from the drop-box 'Select Mode of Operation':

- Darwinian – select genes under highest pressure of the positive Darwinian selection;
- Conserved – select the most conserved genes in the given genomes;

- Hotspotted – select the most variable genes in the given genomes.

Figure 4.2 shows examples of gene selections by these different algorithms from the same input set of Thermotoga genomes.

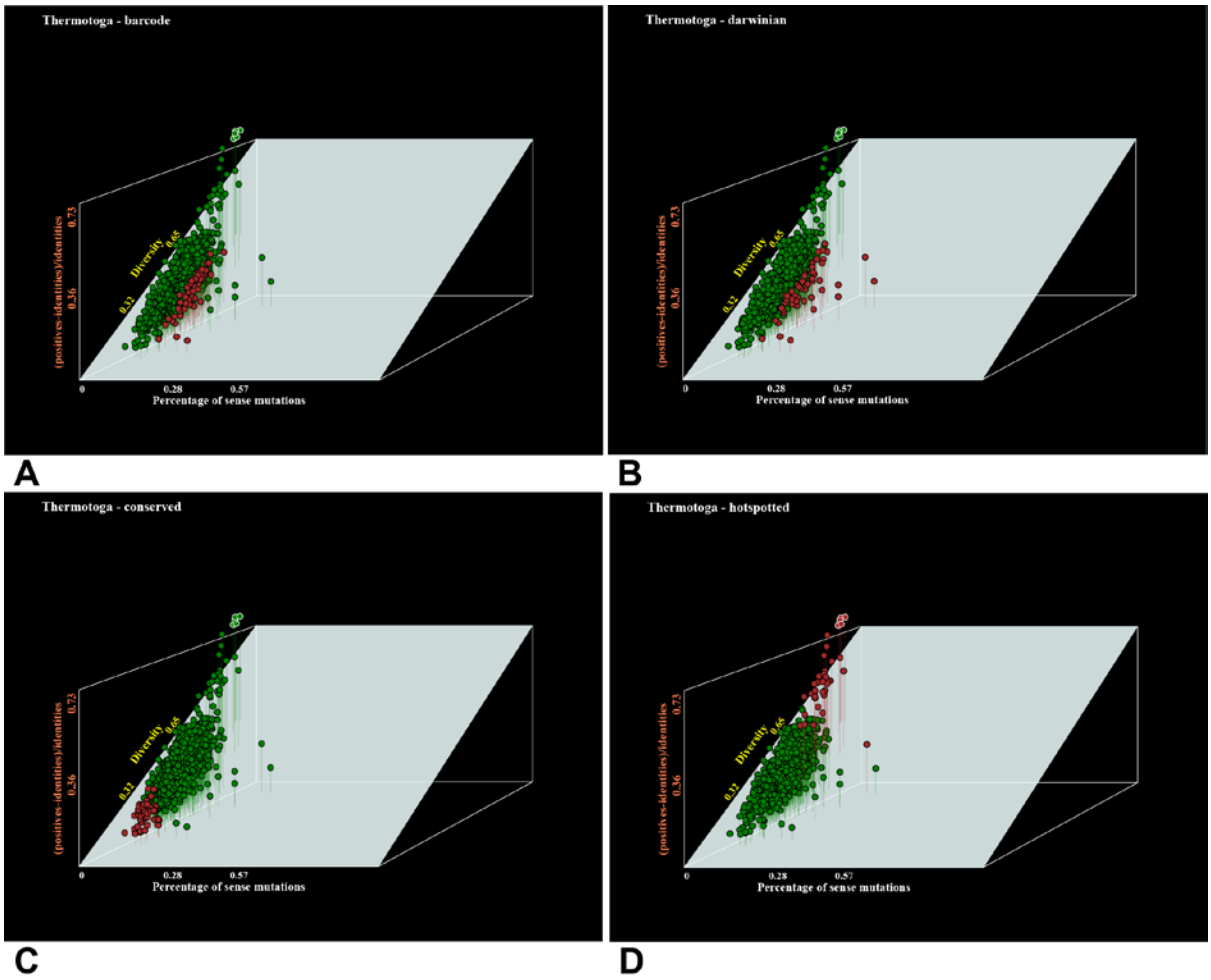



**Figure 4.2: Graphical outputs of the program BarcodeGenerator generated for A) Barcode; B) Darwinian; C) Conserved and D) hotspotted algorithms. Selected clusters of orthologous genes are shown in brown colour.**

The difference between these algorithms consists only in the way the score is calculated for different clusters of orthologous genes:

- Barcode

$$Score = {X(1-X)(1-Y)}/{(Z+1)} \qquad \text{(Eq. 4.1)}$$

- Darwinian

$$Score = \begin{bmatrix} if\ Z \leq 0.3 \rightarrow & Score = {X}/{(Z+1)} \\ else \rightarrow & Score = 0 \end{bmatrix} \qquad \text{(Eq. 4.2)}$$

- Conserved

$$Score = {}^{1}/_{(X+1)(Y+1)(Z+1)} \quad \text{(Eq. 4.3)}$$

- Hotspotted

$$Score = (X+1)(Y+1)(Z+1). \quad \text{(Eq. 4.4)}$$

In all these equations, X, Y and Z are the values of the corresponding axes in Figure 4.2; i.e. X is the percentage of sense mutations in alignments of sequences of orthologous genes; Y is 1 ─ identities of alignments; Z is (positives ─ identities)/identities.

In this work, only the barcode algorithm was considered in detail.

### 4.2.1 Local version of BarcodeGenerator

The local version of the program BarcodeGenerator is a command-line program that can be run on Python 2.7 or Python 2.5. It was designed to select the most appropriate genes for genetic barcoding and generate barcode sequences that can be used for the analysis and visualisation of metagenomic datasets by using another program, Barcoding 2.0, provided from the same web page.

To use the local version of BarcodeGenerator, the user has to download an archived ZIP file of BarcodeGenerator to the local computer and unzip it. When all the archive content has been extracted, the following folders will appear in the computer, as shown in Figure 4.3: *bin*, *gbk_examples*, *input*, *lib* and *output*. The desired GenBank files of the organisms needed to be barcoded are then copied into the *input* folder. To run the program, the user double-clicks on the file *run.py* in the top-level folder of the program. The command-line interface of the BarcodeGenerator is shown in Figure 4.4.

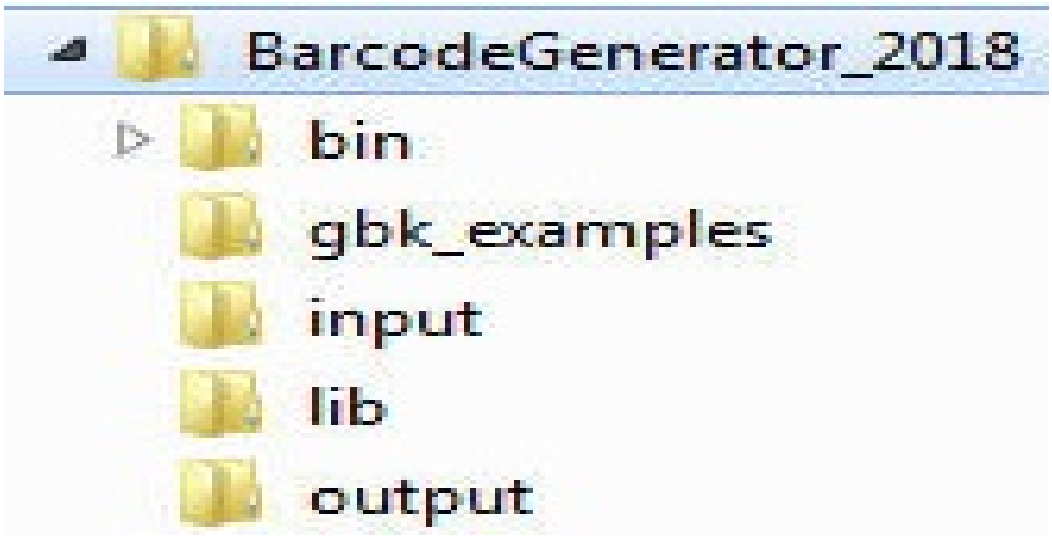



**Figure 4.3: Unzipped folder structure of the local version of BarcodeGenerator.**

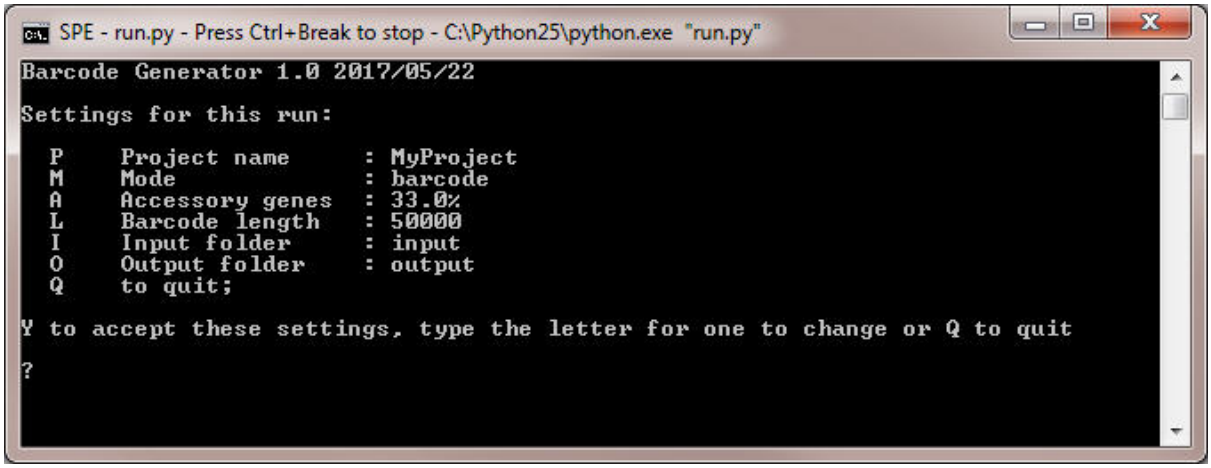



**Figure 4.4: Command line interface of BarcodeGenerator**

Option **P** is used to change the project name. A new folder title by the project name will be created in the folder *output* where all the resulting output files of the current program run will be stored. The operation mode (option **M**) by default is *barcode* to generate barcode sequences. The program may also allow selection of genes of the following categories: (i) Darwinian (orthologous genes under positive selection); and (ii) conserved (the most conserved genes and hotspotted (orthologous genes with the highest number of random mutations) as discussed above (see Figure 4.2). Option **A** is used to specify the percentage of accessory genes in the barcode sequences. This option is available only with the barcode mode of operation. Option **L** is used to set an average length of generated barcode sequences in bp. When **Y** is typed and the *Enter* key pressed, the program starts showing the progress bar. The program generates three output files and stores them to the folder with the project

name in *output* folder. These files are: (i) graphical core gene plot SVG file, as those shown in Figure 4.2; (ii) barcode info-file in text format; and (iii) generated barcode sequences in FASTA format.

### 4.3 Barcoding 2.0 command line interface for metagenome analysis and visualisation

Barcoding 2.0, provided with a command-line user interface, is available for download from the Barcoder web page. The command-line program Barcoding 2.0 can be used for binning of reads of WGS metagenomes (Figure 4.5). The program Barcording 2.0 is a command-line program in Python 2.5/2.7 designed to align metagenomic reads of Roche 454 and Illumina against taxon-specific barcode sequences generated by the online program BarcodeGenerator. The program performs a BLASTN alignment of reads against the barcode sequences and scores every barcode in the set, as explained in Chapter 3.

The user needs to download the zip file Barcoding 2.0 to a computer and unzip it. The unzipped file is made up of the following folders :(i) bin, (ii) db, (iii) input, (iv) lib and (v) output (Figure 4.6). The user then copies the following files into the input folder: (i) FASTA file with barcode sequences and named as *barcodes.fas*; and (ii) FASTA files of metagenomic reads of one or several metagenomes stored in a new folder in the folder *input*. Optionally, the user can copy to the *input* folder to a phylogenetic tree in phylip/Newick format to align barcoded taxa against the phylogenetic tree. Taxonomic units in the tree file MUST have the same names in the barcode file. A minimum of three barcoded taxa should be present in the tree file; however, the total number of barcodes may differ from the total number of taxonomic units in the tree file. Taxonomic units not found among barcodes will be ignored and the barcode sequences not represented in the tree file will be grouped outside the tree in the graphical output file. The user can then run the program by clicking on the file run.py in the top-level folder of the program.


**Figure 4.5: Command line interface of Barcoding 2.0 with the argument setting by default.**

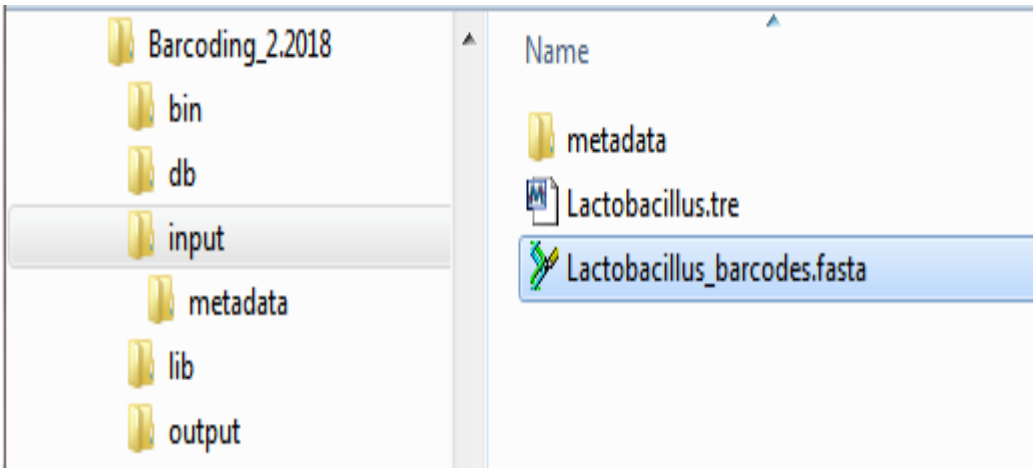



**Figure 4.6: Folders of the program Barcoding 2.0 unzipped to a local directory.**

The name given in the *Query file* option (Figure 4.5) is the name of the barcode file and the option *Database name* indicates the name of the subfolder containing FASTA files of metagenomic reads. These settings should correspond to real names in the *input* folder, otherwise these names can be changed using keys **I** and **D**, respectively. The option **C** is used to choose the stringency parameter between *relaxed* and *stringent*. The stringent mode corresponds to cut-off values 2.5 and 1 of BarcodeScore1 and BarcodeScore2, while in the relaxed mode these values are 2.3 and 0.5 (see discussion in Chapter 3, Table 3.3 and Figure 3.16 for detail). If the user has a phylogenetic tree file *(phylotree.tre* for example), it should

be copied and placed in the input folder. The user can type **F** and press the Enter key; the program will then allow the user to enter the name of the tree file, as shown in Figure 4.7.

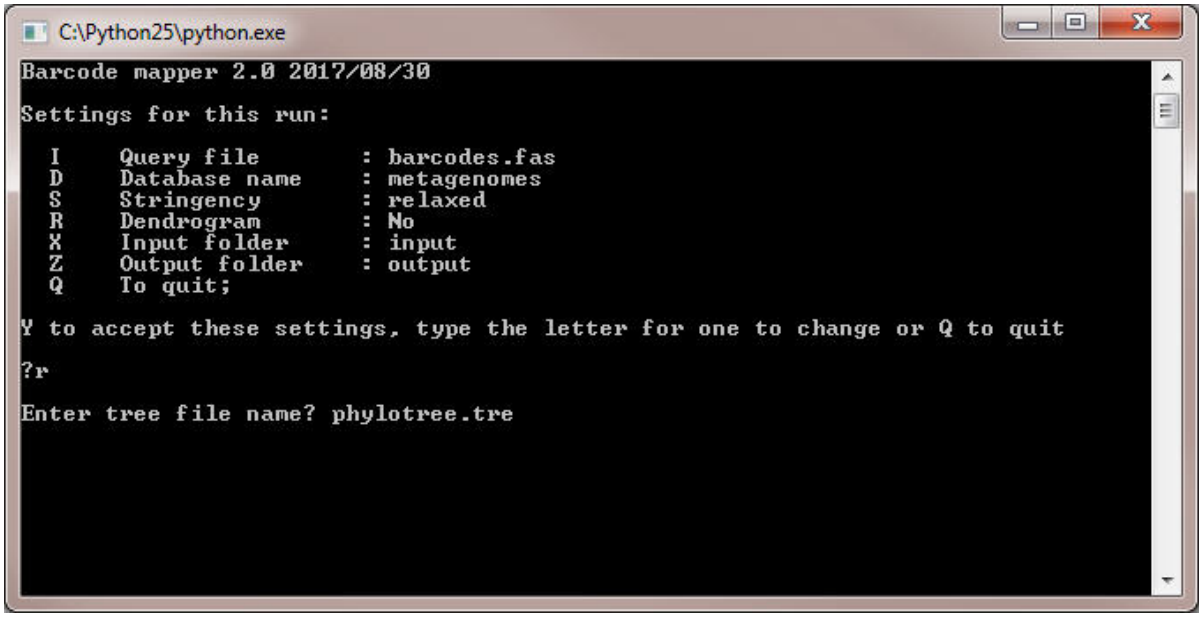



**Figure 4.7: Command line interface when the user wants to enter a phylogenetic tree file.**

If the program does not find the indicated file in the input folder, *no dendogram* setting is returned. However, if it does exist, the name will appear in the option set, as shown in Figure 4.8.

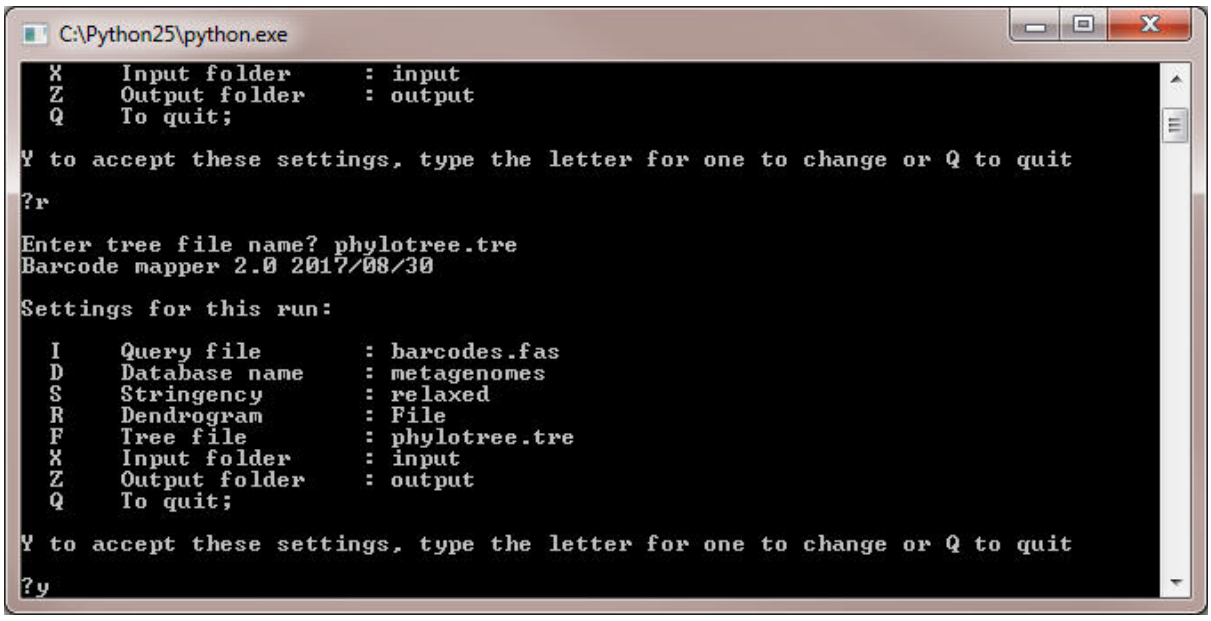



**Figure 4.8: Command line interface of Barcoding 2.0 with the phylogenetic tree file.**

When the user types **Y** and presses the Enter key, the program shows the progress made. The program generates the graphical file in SVG format and a text file. An example of identification of *Lactobacillus* species by generated barcode sequences in the phyllosphere 9673 metagenome publicly available from MG-RAST database is shown in Figure 4.9. The green columns indicate strains, which are likely to be present in the metagenome. The height of the columns depicts values of BarcodeScore2 (see equations 4-6, Chapter 3). The phylogenetic tree beneath the plot was generated by the SWPhylo program (http://swphylo.bi.up.ac.za/) by whole genome sequence comparison. Table 4.1 shows the text file generated.

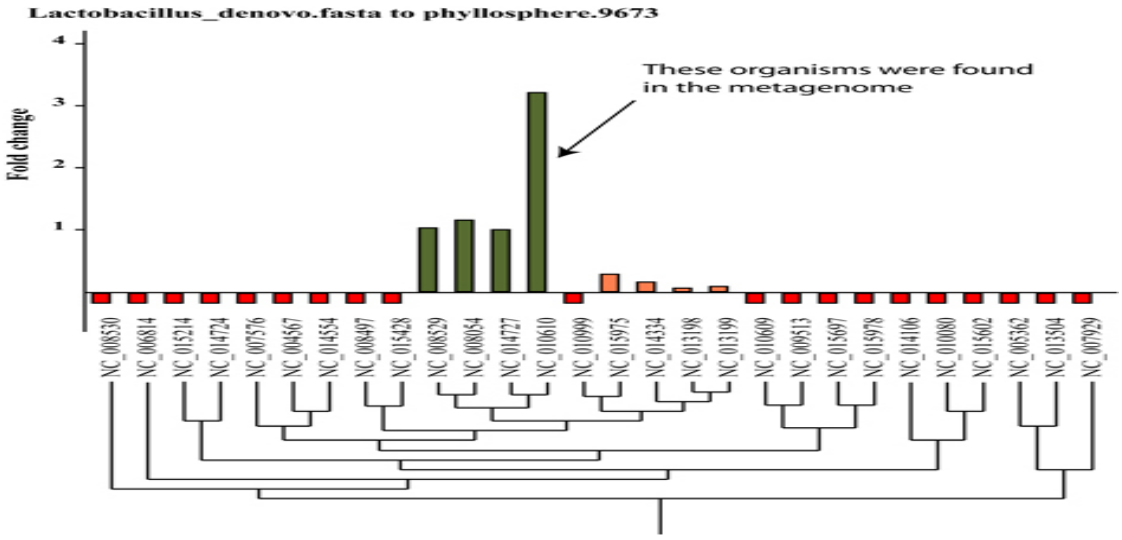



**Figure 4.9: Graphical file for *Lactobacillus.***

**Table 4.1: Shows the screenshot text file generated for Lactobacillus.**

| Strain | Count | CorrScore | Score Class | |
|---|---|---|---|---|
| NC_008530 | 11 | 0.000000 | 0.000000 | |
| NC_006814 | 10 | 0.000000 | 0.063199 | |
| NC_014106 | 7 | 0.000000 | 0.000000 | |
| NC_007576 | 16 | 0.000000 | 0.561481 | |
| NC_013504 | 14 | 0.000000 | 0.389877 | |
| NC_008497 | 15 | 0.000000 | 0.695068 | |
| NC_010610 | 54 | 3.223211 | 6.460976 | ** |
| NC_013199 | 24 | 0.091189 | 0.615796 | |
| NC_013198 | 23 | 0.062169 | 0.610398 | |
| NC_014554 | 10 | 0.000000 | 0.000000 | |
| NC_009513 | 9 | 0.000000 | 0.234012 | |
| NC_014724 | 8 | 0.000000 | 0.030455 | |
| NC_014727 | 36 | 1.006891 | 2.427307 | ** |
| NC_014334 | 24 | 0.168455 | 1.302963 | |
| NC_010080 | 11 | 0.000000 | 0.172688 | |
| NC_015978 | 12 | 0.000000 | 0.565240 | |
| NC_015975 | 26 | 0.284885 | 1.995275 | |
| NC_008529 | 37 | 1.031848 | 2.345540 | ** |
| NC_015697 | 10 | 0.000000 | 0.000000 | |
| NC_008054 | 38 | 1.159676 | 2.662743 | ** |
| NC_010999 | 20 | 0.000000 | 0.358525 | |
| NC_005362 | 14 | 0.000000 | 0.239091 | |
| NC_010609 | 8 | 0.000000 | 0.000000 | |
| NC_015428 | 20 | 0.000000 | 0.193746 | |
| NC_007929 | 17 | 0.000000 | 0.397608 | |
| NC_015214 | 9 | 0.000000 | 0.024616 | |
| NC_015602 | 9 | 0.000000 | 0.000000 | |
| NC_004567 | 10 | 0.000000 | 0.000000 | |

### 4.4 Help and downloads

All the necessary help and information needed are found on the webpage http://seqword.bi.up.ac.za/barcoder_help_download/index.html. In the webpage, the *readme.html* provides information about the Barcoder software tools (BarcodeGenerator and Barcoding 2.0). Lists of instructions are also provided on how to use the Barcorder software tools.

#### 4.4.1 Downloads

Downloads available on the webpage include: (i) the program Barcoding 2.0 (320 MB); (ii) example of input files of bacterial genomes to test the BarcodeGenerator; (iii) diagnostic barcodes created during the project; and (iv) examples of artificial metagenomes created for this project and discussed in Chapter 3. The program Barcoding 2.0 with the command-line user interface is available for download from the Barcoder webpage. The command-line program Barcoding 2.0 can be used for binning of reads of WGS metagenomes as explained earlier. Also available for download is an archived input file example with eight genbank (GBK) files of *Bacillus* genomes (*Bacilus amyloliquefaciens* NC_014551, *Bacillus clausii* NC_006582, *Bacillus coagulans* NC_015634, *Bacillus halodurans* NC_002570, *Bacillus licheniforms* NC_006322, *Bacillus pumilus* NC_009848, *Bacillus subtilis* NC_000964 and *Bacillus velezensis* NC_009725. This file was prepared as an example to test the BarcodeGenerator. To use the input file of the eight *Bacillus* genomes the user has to: (i) download the file *example.zip* to a computer; (ii) go to the Webpage http://bargene.bi.up.ac.za/; (iii) click the key *Browse* and select this file on the local computer; (iv) click the key *Upload files and barcode;* and (v) wait for a message to come to the user's e-mail with links showing: (i) the Core Gene plot (svg); (ii) barcode information; and (iii) barcode sequences in FASTA format. Examples of the expected output files are also available for viewing on the web-page http://seqword.bi.up.ac.za/barcoder_help_download/example/example.html.

The generated barcode sequences of different length for all organisms used in the case studies of this project (*Bacillus cereus*, *Escherichia* and *Shigella*, *Lactobacillus*, *Mycobacteria*, *Prochlorococcus*, *Salmonella*, *Shewanella*, *Streptococcus*) were made available for download at http://seqword.bi.up.ac.za/barcoder_help_download/barcodes/index.html, as shown in Table 4.2.

**Table 4.2: Different taxonomic groups for which barcode sequences were created and made available for download**

| Taxonomic group | Average length | Info & downloads |
|---|---|---|
| ***Bacillus cereus*** | 10kbp | Info download (92kb) |
| | 25kbp | Info download (211kb) |
| | 75kbp | Info download (596kb) |
| | 100kbp | Info download (781kb) |
| | 150kbp | Info download (781kb) |
| | 200kbp | Info download (1500kb) |
| | 250kbp | Info download (1863kb) |
| ***Escherichia* and *Shigella*** | 10kbp | Info download (92kb) |
| | 25kbp | Info download (293kb) |
| | 75kbp | Info download (805kb) |
| | 100kbp | Info download (1124kb) |
| | 150kbp | Info download (1636kb) |
| | 200kbp | Info download (2168kb) |
| | 250kbp | Info download (2702kb) |
| ***Lactobacillus*** | 10kbp | Info download (101kb) |
| | 25kbp | Info download (240kb) |
| | 75kbp | Info download (674kb) |
| | 100kbp | Info download (893kb) |
| | 150kbp | Info download (1325kb) |
| | 200kbp | Info download (1756kb) |
| | 250kbp | Info download (2180kb) |
| ***Mycobacteria*** | 10kbp | Info download (68kb) |
| | 25kbp | Info download (142kb) |
| | 75kbp | Info download (400kb) |
| | 100kbp | Info download (522kb) |
| | 150kbp | Info download (722kb) |
| | 200kbp | Info download (1019kb) |
| | 250kbp | Info download (1035kb) |
| ***Prochlorococcus*** | 10kbp | Info download (56kb) |
| | 25kbp | Info download (114kb) |
| | 75kbp | Info download (302kb) |
| | 100kbp | Info download (397kb) |
| | 150kbp | Info download (587kb) |
| | 200kbp | Info download (777kb) |
| | 250kbp | Info download (967kb) |
| ***Salmonella*** | 10kbp | Info download (54kb) |
| | 25kbp | Info download (178kb) |
| | 75kbp | Info download (539kb) |
| | 100kbp | Info download (705kb) |
| | 150kbp | Info download (1057kb) |
| | 200kbp | Info download (1407kb) |
| | 250kbp | Info download (1746kb) |
| ***Shewanella*** | 10kbp | Info download (97kb) |
| | 25kbp | Info download (206kb) |
| | 75kbp | Info download (541kb) |
| | 100kbp | Info download (711kb) |

| | | |
|---|---|---|
| | 150kbp | Info download (1039kb) |
| | 200kbp | Info download (1355kb) |
| | 250kbp | Info download (1668kb) |
| ***Streptococcus*** | 10kbp | Info download (180kb) |
| | 25kbp | Info download (482kb) |
| | 75kbp | Info download (1445kb) |
| | 100kbp | Info download (1918kb) |
| | 150kbp | Info download (2823kb) |
| | 200kbp | Info download (3636kb) |
| | 250kbp | Info download (4419kb) |

In the information section for each taxonomic group, the graphical output of the diagnostic barcode generated for each length is provided (Figure 4.10). Other information, such as the original genome, locus tag, annotation and location in the barcode of core/accessory genes, is provided (Figure 4.11). The information section is also linked to the NCBI. The NCBI offers an enormous collection of online resources for biological information and data, comprising the: (i) GenBank, (ii) nucleic acid sequence database; (iii) PubMed database of citations; and (iv) abstracts for published life science journals (NCBI Resource Coordinators, 2016). Over the years, the quantity and diversity of data that the NCBI sustains have expanded immensely and the data can commonly be divided into six groups: (i) Literature; (ii) Health; (iii) Genomes; (iv) Genes; (v) Proteins; and (vi) Chemicals (NCBI Resource Coordinators, 2016). The Entrez system (Schuler *et al.,* 2016) of the NCBI offers access to varied groups of 37 databases that together contain 2.1 billion records. Since the information section is linked to the NCBI, more detailed information about each genome is provided (Figure 4.12).

In the download section, the average length (10, 25, 75, 100, 150, 200, 250 kbp) of each barcode sequence in the different taxonomic group is available for download.

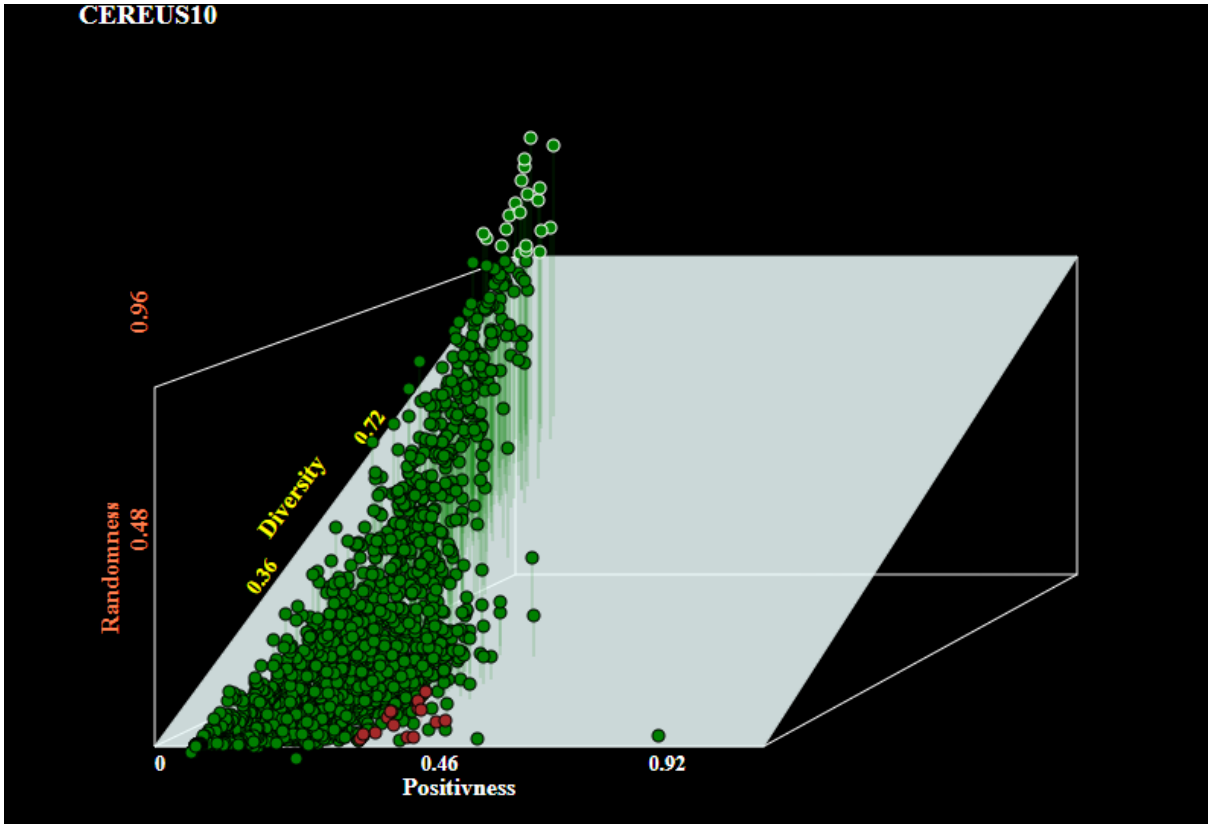



**Figure 4.10: Graphical output of clusters of orthologous genes selected for diagnostic barcodes generated for the group *Bacillus cereus* with an average length of10 kbp.**

*Genes with the same numbers are orthologs

| Genome | Locus tag | Annotation | Location in barcode | Core/Accessory gene |
|---|---|---|---|---|
| NC_004722 Bacillus cereus ATCC 14579, complete genome. | | | | |
| 1 | BC3924 | cytoplasmic protein (rev) [3907023..3907493] | [1-471] | Core |
| 2 | BC3159 | NAD-dependent epimerase/dehydratase / | [522-2378] | Core |
| 3 | BC0131 | rplC,50S ribosomal protein L3 (dir) [126823..127455] | [2429-3061] | Core |
| 4 | BC3130 | hypothetical protein (rev) [3100771..3101796] | [3112-4137] | Core |
| 5 | BC0117 | rplK,50S ribosomal protein L11 (dir) [110486..110911] | [4188-4613] | Core |
| 6 | BC3160 | TetR family transcriptional regulator | [4664-5248] | Core |
| 7 | BC1002 | anti-sigma B factor antagonist (dir) [983659..983997] | [5299-5637] | Core |
| 8 | BC2156 | hypothetical protein (dir) [2099430..2099813] | [5688-6071] | Core |
| 9 | BC0450 | protein tyrosine phosphatase (dir) [439787..440251] | [6122-6586] | Core |
| unique | BC2637 | hypothetical protein (rev) [2603031..2603225] | [6696-6941] | Accessory |
| unique | BC5124 | hypothetical protein (rev) [5025633..5027765] | [7246-9429] | Accessory |
| unique | BC3019 | hypothetical protein (rev) [2981452..2981655] | [11917-12171] | Accessory |
| unique | BC1875 | Phage protein (dir) [1827279..1827866] | [14913-15551] | Accessory |
| unique | BC2976 | hypothetical protein (dir) [2934856..2935068] | [18931-19194] | Accessory |
| unique | BC0943 | hypothetical protein (dir) [927449..927979] | [22837-23418] | Accessory |

**Figure 4.11: Information provided for each genome used to generate diagnostic barcode sequences available for download.**

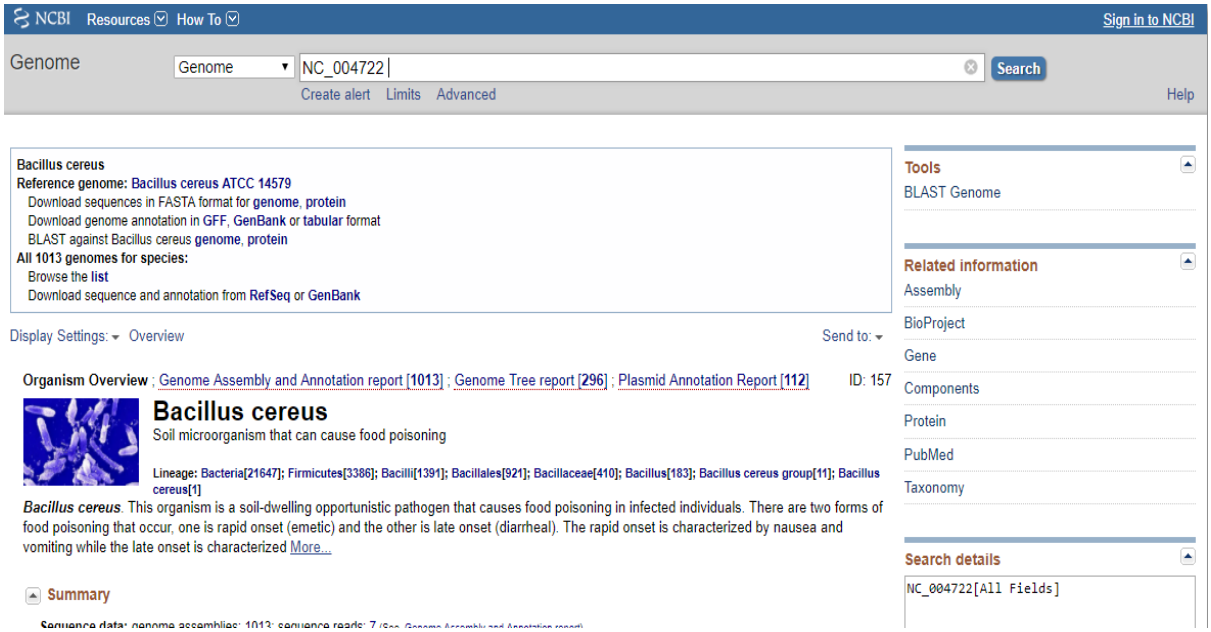



**Figure 4.12: Screenshot example of NCBI page linked to each genome used to generate barcode sequences.**

Also available for download are the sets of artificial metagenomic reads (200-500 bp) generated by the program MetaSim. Contents of artificial metagenomes provided on the Web-page http://seqword.bi.up.ac.za/barcoder_help_download/artificial_metagenomes/index.html are shown in Table 4.3.

**Table 4.3: Contents of the artificial metagenomes**

| ARTIFICIAL METAGENONES | | |
|---|---|---|
| ***Shewanella*** | | |
| *S. amazonensis* SB2B | NC_008700 | 5% |
| *S. frigidimarina* NCIMB400 | NC_008345 | 10% |
| *Shewanella sp.* MR4 | NC_008321 | 15% |
| ***Escherichia/Shigella*** | | |
| *E. coli* ATCC8739 | NC_010468 | 5% |
| *E. coli* BL21 | NC_012947 | 10% |
| *Shigella dysenteriae* Sd197 | NC_007606 | 15% |
| ***Lactobacillus*** | | |
| *L. sanfranciscensis* TMW1 | NC_015978 | 5% |
| *L. plantarum* NCFS1 | NC_004567 | 10% |
| *L. fermentum* IF03956 | NC_010610 | 10% |
| ***Mycobacterium*** | | |
| *M. avium* Env77 | NC_008595 | 5% |
| *M. abscessus* ATCC 1977 | NC_010397 | 10% |

### 4.4 SeqWord project

The Barcoder software tools are part of the SeqWord project (genome linguistic approaches for comparative genomics, phylogenomics and mobilomics). The SeqWord project addresses the development of an integrated research environment for data mining in DNA sequences by using genome linquistics. SeqWord projects are non-commercial academic software tools and web applications, which were developed with the support of the National Research Foundation of South Africa (NRF). The principal investigator of the SeqWord project is Prof. Oleg Reva. All tools in the SeqWord project were created by post-graduate students. Other tools available in the SeqWord project (http://seqword.bi.up.ac.za/) are Genome browser, Genomic Island Sniffer, Sniffer GI Browser, GI Databases, Interactive GI maps, SWPhylo, **GenomeBarcoder**, OligoDBViewer, MetaLingvo and LingvoCom.

### 4.6 Conclusion

This web interface provides users with online access to the program BarcodeGenerator, which creates diagnostic barcodes based on the genome sequences of species of interest submitted by users. The program also allows the addition of a genome-specific accessory to improve the sensitivity of the barcode sequences. Hence the BarcodeGenerator is an efficient approach for generating diagnostic barcode sequences. The BarcodeGenerator also has a local version, which is a command-line program. It was designed to select the most appropriate genes for genetic barcoding and generate barcode sequences, which can be used for the analysis and visualisation of metagenomic datasets by using the program Barcoding 2. Barcoding 2.0 is another program available from the same resources that enables efficient and practical use of metabarcodes for visualisation of distribution of organisms of interest in environmental and clinical metagenomic samples. The program Barcording 2.0 is a command-line program written in Python 2.7 and designed to align metagenomic reads generated by Roche 454 and/or Illumina against taxon-specific barcode sequences generated by the program BarcodeGenerator (locally or through the web interface). The Barcoder software tools (BarcodeGenerator and Barcoding 2.0) are available for download at http://bargene.bi.up.ac.az/. For framework data visualisation, *matplotlib* 1.5.1 Python module (https://matplotlib.org/1.5.1/index.html) was used. All the programs for this work are compatible with Python 2.5/2.7 and are made accessible at the website http://bargene.bi.up.ac.za/ through a PHP interface.

# CHAPTER 5 Evaluation of the program Barcoding 2.0 by binning real metagenomic reads


## Abstract

In this chapter the researcher gives an in-depth explanation of case studies of DNA reads of different metagenomics datasets from the MG-RAST database used together with barcode sequences generated for selected groups of microorganisms discussed in Chapter 2.


## 5.1 Introduction

Advancement in technology has made it possible for the genome sequencing project to move from the study of single genomes to the investigation of genomes in the community. Metagenomics allows culture-independent and sequence-based studies of microbial communities (Chan *et al.,* 2008). Metagenomics projects usually start by using shotgun WGS on environmental samples to conduct: (i) sequence reads; (ii) assembly of sequence reads; (iii) gene prediction; and (iv) functional annotation and metabolic pathway construction (Chan *et al.,* 2008).

An important step in metagenomics is called “binning”. The binning process sorts sequence fragments (either original reads generated by sequencers or assembled contigs) of various species obtained from WGS sequencing into phylogenetically related bins or groups (Mavromatis *et al.,* 2007). Normally, each sequence is either classified into a taxonomic group such as OTU, genus or family through association to some referential data, or clustered into groups of sequences that denote taxonomic groups centred on common characteristics such as the GC content (Sharpton *et al.,* 2014). Binning plays a key part in the analysis of metagenomes: (i) depending on the approach used, binning can give understanding into the presence of new genomes that are challenging to identify; (ii) it can be used to provide better insight into the unique numbers and kinds of taxa in a given community; and (iii) it can decrease the intricacy of data, as used in post-binning analysis such as assembly, which can be done autonomously on each set of the binned reads rather than on the whole population of data (Sharpton, 2014). Most of the present binning techniques involve assigning of sequence fragments by comparing sequence similarity or sequence composition with already sequenced

genomes that are still far from comprehensive (Chan *et al.,* 2008). Hence, most methods used for binning of metagenomic reads do not allow identification below the genus level and very often stop on the level of bacterial families (Thomas *et al.,* 2012).

In this chapter, the researcher discusses the results of different case studies where DNA reads of different metagenomic datasets from MG-RAST database were aligned with BLASTN using the novel Barcoding 2.0 program (Chapter 3) against taxon-specific barcode sequences generated by the online program BarcodeGenerator (Chapter 2).

## 5.2 Program implementation

Barcode sequences generated by BarcodeGenerator can be used for identification of species of interest in environmental metagenome samples sequenced by Roche 454 or Illumina technologies. Barcoding 2.0 is an application written in Python 2.5 (also compatible with Python 2.7) with a command-line user interface made available for downloading from the BarcodeGenerator website (http://bargene.bi.up.ac.za/). The program uses BLASTN to align reads against the generated barcode sequences and then calculates several parameters for scoring the results of the BLASTN alignment and individual barcodes.

## 5.3. Identification of barcoded sequences in real metagenomes

An attempt was made to test the barcode sequences created by the program BarcodeGenerator for various genomes of bacteria of industrial, medicinal and ecological importance on real metagenomic datasets available from NCBI and MG-RAST. The metagenomes used are divided into three groups: (i) symbiotic microbiomes (canine gut, human gut, mammalian blood, termite gut and cow gut); (ii) soil and rhizosphere microbiomes (desert soil, grassland, forest rhizosphere, phyllosphere, rain forest, soybean rhizosphere); and (iii) environmental microbiomes (anthropogenic estuariane, sludge, hydrothermal vent and mediterian bathypelagic).

### 5.3.1 Metagenome analyser

Since this is the first version of the Barcoding 2.0 program released, to validate the researcher's results and to determine how well the Barcoding 2.0 performed, the researcher first performed a BLASTN alignment of various metagenomic reads used in the case studies against a local copy of the NCBI *nt* database using the *blastn* for Linux implementation of the

alignment program installed on the computer server. The MEGAN 4.70.4 program was then used to estimate and interactively explore the taxonomical content of the dataset, using the NCBI taxonomy to summarise and order the results. MEGAN uses a simple algorithm that reads standard BLASTN output files and assigns each read to the LCA of the set of taxa that it hits in comparison. Hence, species-specific sequences are assigned to the taxa near the leaves of the NCBI tree, whereas widely conserved sequences are assigned to high order taxa closer to the root (Huson *et al.*, 2007). As discussed in Chapter 3, the 50-100 000 nucleotide long barcodes gave the best results. For all case studies the researcher used 100 000-nucleotide-long barcodes generated by the BarcodeGenerator. Results obtained from the MEGAN program are discussed in further detail below.

#### 5.3.1.1 Canine and cow intestinal microbiomes

All mammals are populated by groups of organisms vital to the typical form and function of the host. Regarding cellular composition, genetic diversity and metabolic capacity, the host mammal should be regarded as a multispecies hybrid organism made up of host and microbial cells functioning in vibrant and symbiotic symmetry (Turnbaugh *et al.*, 2007; Shreiner *et al.*, 2015; Barko *et al.*, 2018). The gastrointestinal microbiome is a varied conglomerate of bacteria, archaea, fungi, protozoa and viruses that occupy the gut of mammals. Research in humans and mammals has associated the microbiome in a series of physiologic processes that are important to host health, including energy homeostasis, metabolism, gut epithelial health, immunologic activity and neurobehavioral progress. The microbial genomes confer metabolic competences above those of the host organism alone, making the gut microbiome a dynamic contributor in host physiology (Barko *et al.*, 2018).

Figure 5.1 shows the results obtained for the MEGAN binning of reads of canine gut. From the figure one can see the different groups of bacteria that could possibly be identified in the canine metagenome. From the top of the phylogenetic tree, groups include the Bacteriodetes/Chlorobi group, Chylamydiae/Verrucomicrobia group, Fibrobacteres/Acidobacteria group and Proteobacteria group. The canine gut metagenomic dataset was selected for this case study as representative of rich symbiotic gut micro-flora enriched with many organisms, which were used in previous steps to generate diagnostic barcodes by the program BarcodeGenerator.

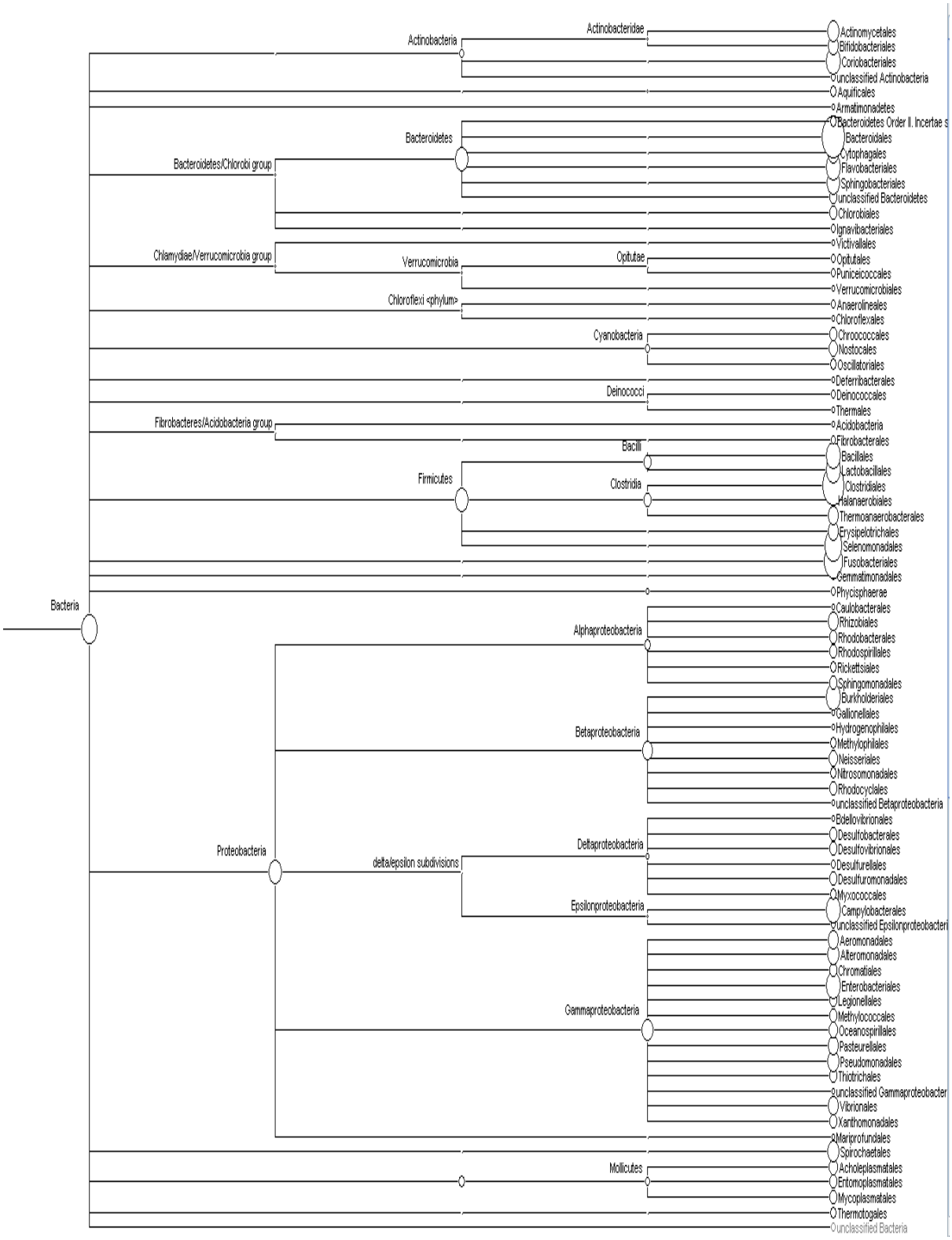



**Figure 5.1: MEGAN analysis of reads from the canine gut metagenome.**

### 5.3.1.2 Phyllosphere

The phyllosphere in the aerial surface of plants is an essential and pervasive habitat for bacteria (Vorholt, 2012). It is appraised on a universal scale that the phyllosphere spans more than $10^8$ $km^2$ and serves as home to approximately $10^{26}$ bacterial cells (Lindow and Brandl, 2003). Leaf-related bacteria epitomise a widespread and primeval symbiosis that can affect host growth and function in various ways, including the production of growth-promoting nutrients and hormones (Reed *et al.*, 2010) and protection of hosts against pathogen infection (Innerenbner *et al.*, 2011). The phyllosphere bacteria can influence plant biogeography and the ecosystem function through their influence on plant performance under different environmental conditions (Kembel *et al.*, 2012).

Figure 5.2 shows the results obtained for the MEGAN binning of reads of the phyllosphere. From the figure one can see the different groups of bacteria that could possibly be identified in the phyllosphere metagenome. From the top of the phylogenetic tree the groups include (i) Actinobacteria; (ii) Armatimonadete; (iii) Bacteroidetes; (iv) Verrucomicrobia; (v) Thermomicrobia; (vi) Cyanobacteria; (vii) Acidobacteria; (viii) Firmicutes; (ix) Planctomycetes; (x) Alphaproteobacteria; (xi) different subdivisions of proteobacteria; (xi) Mollicutes; and (xii) unclassified groups of bacteria. This metagenome was selected for the case studies as representative of species-rich environmental micro-flora to validate diagnostic barcodes prepared for the identification of *Mycobacteria* and *Shewanella*.

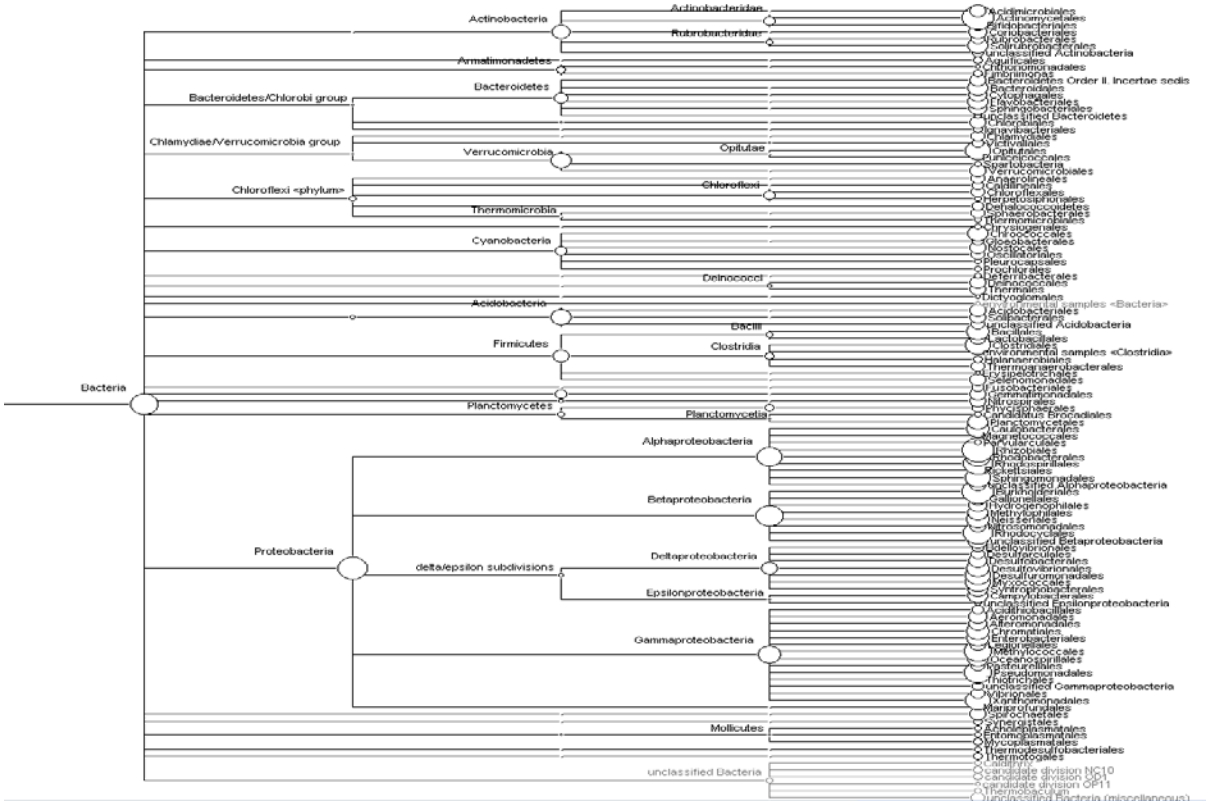



**Figure 5.2: MEGAN analysis of reads of the phyllosphere metagenome.**

### 5.3.1.3 Grassland

The significance of the soil microbiome in the cycling of important nutrients such as carbon and nitrogen is well understood. However, because of the microbiome's complexity, little is understood about how climate will affect the diversity, abundance and structure of the community (Shaver *et al.,* 2000). In grassland soils, experimental warming has been shown to increase bacterial biomass in winter and spring (Belay-Tedla *et al.,* 2009; Sheik *et al.,* 2011); nevertheless, warming affected bacterial biomass negatively and 16S rRNA gene abundance in summertime and early fall correspondingly, signifying that warming may have a seasonal effect on soil moisture (Castro *et al.,* 2010; Sheik *et al.,* 2011).

This metagenomic dataset was selected for the case studies as an example of species-rich micro-flora associated with a plant rhizosphere, first of all to test the identification of environmental Mycobacteria.

#### 5.3.1.4 Hydrothermal vent

The greatly varying chemical conditions present in different places above and below the sea floor at deep-sea hydrothermal vents and the often very steep gradients between different conditions generate a wide range of geochemical niches and potential energy sources for microorganisms (Fisher *et al.,* 2000). Primary production by chemolithoautotrophs sustains not only the heterotroph components in the microbial ecosystem, but also the animal communities through either symbioses or free-living bacteria that form the base of food webs (Fisher *et al.,* 2000). The pathways of inorganic carbon metabolism used for primary production by hydrothermal vent microbes are very diverse, which may reflect the diversity of physical and chemical microhabitats they occupy (Fisher *et al.,* 2000).

This metagenome was used as an example of an environment with a relatively limited number of specific bacterial species. Figure 5.3 shows the results obtained for the MEGAN binning of reads of the hydrothermal vent. From the figure one can see the different groups of bacteria that could possibly be identified in the hydrothermal vent metagenome, though not as rich and diverse as the canine gut and phyllosphere metagenome. From the top of the phylogenetic tree there are groups that include: (i) Actinobacteria; (ii) Bacteroidetes; (iii) Cyanobacteria; (iv) proteobacteria; and (v) mollicutes.

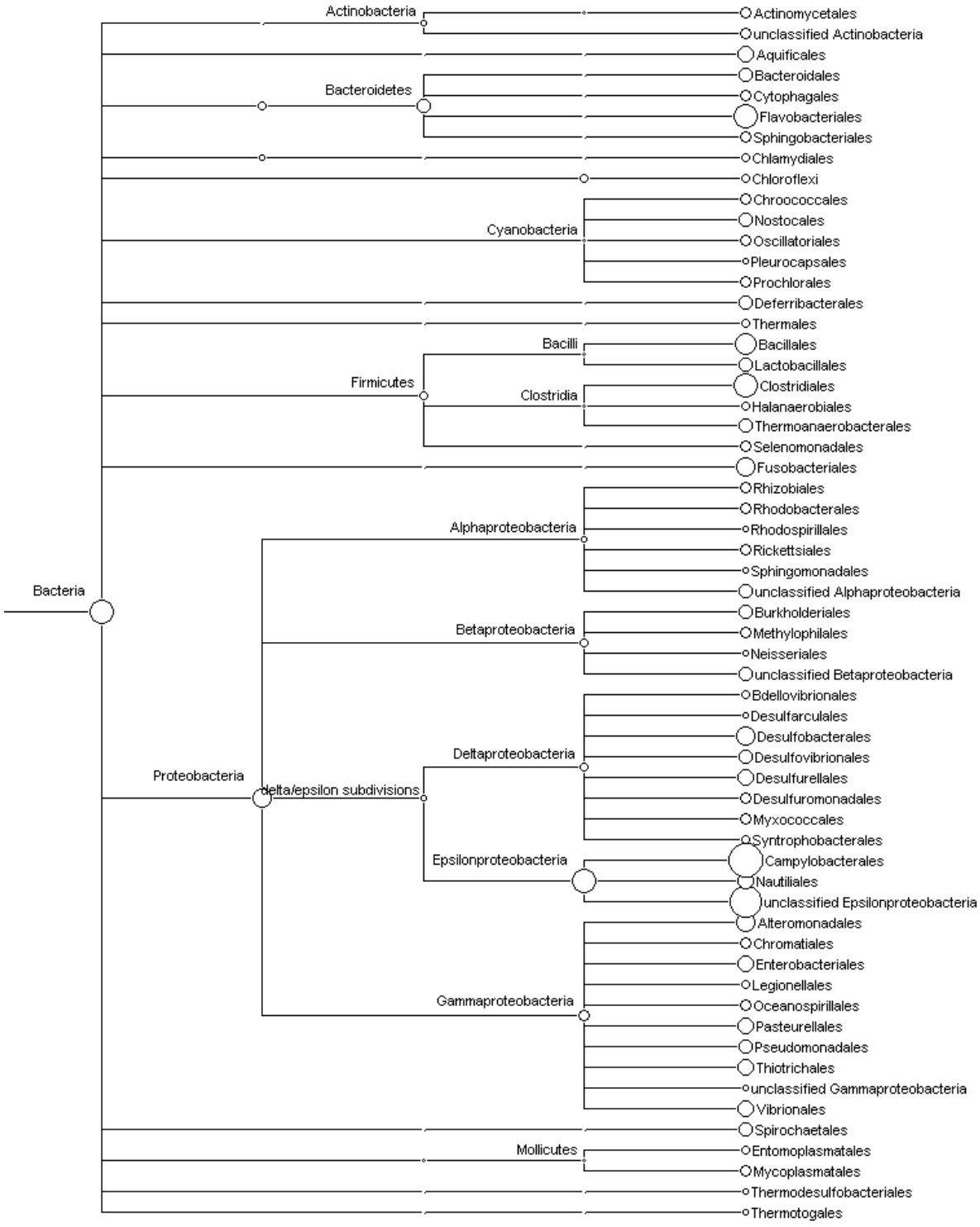



**Figure 5.3: MEGAN analysis of reads of the hydrothermal vent metagenome.**

### 5.3.1.5 Mammalian blood

Mammalian blood is believed to be sterile, except in cases of bacteremia and contamination, when microorganisms are isolated in mammalian blood. Hence, MEGAN analysis for mammalian blood in Figure 5.4 showed very few microorganisms. In this study the researcher used mammalian blood as a form of negative control.

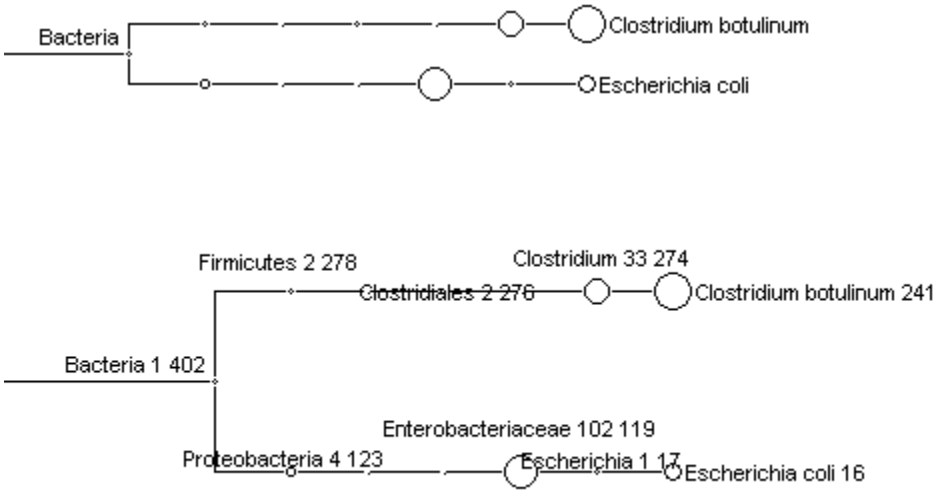



**Figure 5.4: MEGAN analysis showing the bacteria seen in the mammalian blood metagenome.**

### 5.3.1.6 Other metagenomes used in this study

In addition to metagenomes described above, subsets of several other metagenomic datasets available from the MG-RAST database were used to evaluate the diagnostic barcodes generated for this study (Chapter 3), which were made available from the project web-site at http://seqword.bi.up.ac.za/barcoder_help_download/barcodes/index.html. A description of all the metagenomic datasets used is given in Table 5.1.

**Table 5.1: Samples of metagenomic datasets from MG-RAST database used in this study.**

| Name of metagenome | Number of reads | Average length of reads in bp | MG-RAST sample number or reference |
|---|---|---|---|
| **SYMBIOTIC MICROBIOMES** | | | |
| Canine gut | 583,523 | 400 | Swanson *et al*., 2010 |
| Human gut | 500,000 | 1,365 | mgs79383 |
| Mammalian Blood | 92,948 | 1,915 | mgs81295 |
| Termite gut | 99,776 | 856 | Singh *et al*., 2015 |
| Cow gut | 264,849 | 100 | mgs17404 |
| **SOIL AND RHIZOSPHERE MICROBIOMES** | | | |
| Desert soil | 85,549 | 65 | mgs64929 |
| Grassland | 976,268 | 374 | Delmont *et al*., 2012 |
| Forest rhizosphere | 561,526 | 148 | mgs50708 |

| | | | |
|---|---|---|---|
| Phyllosphere | 1,026,982 | 386 | mgs9673 |
| Rain Forest | 782,404 | 418 | mgs6030 |
| Soybean rhizosphere | 151,054 | 523 | soyjp1 |
| **ENVIRONMENTAL MICROBIOMES** | | | |
| Anthropogenic Estuariane | 526,919 | 358 | Kisand *et al*., 2012 |
| Sludge | 96,563 | 1,056 | Guo *et al*., 2017. |
| Hydrothermal Vent | 293,065 | 1,008 | mgs18062 |
| Mediterian Bathypelagic | 9,047 | 797 | mgs2358 |

### 5.3.2 BARCODING 2.0

In the next step, an attempt was made to test the barcode sequences created by the program BarcodeGenerator for selected genomes of various bacteria of industrial, medicinal and ecological importance on real metagenomic datasets obtained from the MG-RAST database. The results of identification of bacterial taxa by aligning reads against diagnostic barcode sequences by means of the program Barcoding 2.0 are shown in Table 5.2. The identification is depicted by (++), meaning that at least one organism was reliably identified by *BarcodingScore*1 and *BarcodingScore*2, as indicated on the output graphs by green bars; (+) indicates that at least one organism was identified with the BarcodingScore1 above 1, which is depicted by an orange bar; (+/-) means at least one organism was identified with *BarcodingScore*1 below 1 and the (-) sign means that no organisms were identified in the sample. The results obtained for each taxonomic group are discussed below.

**Table 5.2: Results obtained with Barcoding 2.0 program for different metagenomes.**

| **METAGENOMES** | **BARCODES** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **SYMBIOTIC MICROBIOMES** | *Bacillus cereus* | *Escherichia-Shigella* | *Lactobacillus* | *Mycobacteria* | *Prochlorococcus* | *Salmonella* | *Shewanella* | *Streptococcus* |
| Canine gut | + | + | + | — | + | + | — | + |
| Human gut | +/— | +/— | ++ | +/— | +/— | — | +/— | ++ |
| Mammalian Blood | — | — | — | — | — | — | — | ++ |
| Termite gut | — | +/— | ++ | — | +/— | — | +/— | +/— |
| Cow gut | — | +/— | ++ | +/— | + | — | +/— | ++ |
| **SOIL AND RHIZOSPHERE** | | | | | | | | |

| **MICROBIOMES** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Desert soil | — | — | + | + | +/— | — | +/— | — |
| Grassland | +/— | +/— | ++ | + | ++ | ++ | + | +/— |
| Forest rhizosphere | +/— | +/— | ++ | + | + | +/— | +/— | +/— |
| Phyllosphere | +/- | +/— | ++ | + | ++ | +/— | ++ | +/— |
| Rain Forest | — | +/— | ++ | + | ++ | +/— | ++ | +/— |
| Soybean rhizosphere | — | +/— | ++ | +/— | + | +/— | +/— | — |
| **ENVIRONMENTAL MICROBIOMES** | | | | | | | | |
| Anthropogenic Estuariane | — | +/— | ++ | +/— | +/— | +/— | +/— | ++ |
| Sludge | — | +/— | + | + | +/— | + | — | + |
| Hydrothermal Vent | +/— | ++ | ++ | — | + | — | +/— | ++ |
| Mediterian Bathypelagic | — | — | — | — | — | — | — | — |

**++ - at least one organism is reliably identified (green bar);**
**+ - at least one organism is identified with the score above 1 (orange bar);**
**+/— - at least one organism is identified with the score below 1 (orange bar);**
**— - no organism was identified (all short red bars);**

### 5.3.2.1 Bacillus cereus

*Bacillus cereus* is closely related to *Bacillus anthracis* and the insect pathogen *Bacillus thuringiensis* (Ivanova *et al;* 2003). *Bacillus anthracis* are dangerous zoonotic pathogens while *Bacillus thuringienis* are used in pesticides. *Bacillus anthracis* and *Bacilllus thuringienis* do contain plasmid borne-specific toxins and this is usually used to differentiate them from *Bacillus cereus* (Ivanova *et al.,* 2003).

In symbiotic microbiomes *Bacillus cereus* was only identified in canine gut (+) and in human gut (+/-). For soil and rhizosphere microbiomes *Bacillus cereus* was identified in grassland (+/-), the forest rhizosphere (+/-) and phyllosphere (+/-). In the environmental microbiomes, *Bacillus cereus* was only identified in a hydrothermal vent (+/-). This indicates that these bacteria are widely distributed in nature but are not abundant in selected habitats.

#### 5.3.2.2 Escherichia coli/Shigella

*Escherichia coli* are a known commensal of the gastrointestinal tract of warm-blooded animals and are used as the everyday laboratory mainstay. However, pathogenic *E. coli* has also been reported, which causes human diseases ranging from disorders of the gastrointestinal tract to ones affecting extra-intestinal sites such as the urinary tract, bloodstream and the central nervous system (Kaper *et al.,* 2004; Croxen and Finlay 2010; Croxen *et al.,* 2013). Though various aetiological agents have been reported as the cause of diarrhoea, pathogenic *E. coli* stands out among others as a major cause. Entero-invasive *E.coli/Shigella* spp are described as facultative intracellular pathogens and the aetiological agents of bacillary dysentery, also known as shigellosis. *Bacillus dysenteriae,* also called *Shigella,* was first identified in 1897 by Kiyoshi Shiga during an epidemic in Japan, where it infected more than 91 000 people, causing a mortality rate of more than 20% (Trofa *et al.,* 1999).

*Escherichia coli/Shigella* was identified in all the symbiotic microbiomes with the exception of mammalian blood, which is very much anticipated, except in cases of bacteraemia or contamination. In soil and rhizosphere microbiomes, *Escherichia coli/Shigella* was identified in all the microbiomes with the exception of desert soil. However, *Escherichia coli/Shigella* is not usually isolated from desert soils. *Escherichia coli/Shigella* was identified in all environmental microbiomes. However, most of the identifications of *Escherichia coli/Shigella* in the metagenomes had scores below 1, depicted by a short orange bar (+/-), meaning they were not abundant in any of these habitats.

#### 5.3.2.3 Lactobacillus

Lactic acid bacteria were mostly seen in various natural environments and were represented by precise *lactobacilli* compositions such as *L. acidophilus, L. delbrueckii* spp. *bulgaricus. Lactobacillus helveticus* are the classic representatives of the micro-flora of fermented milk products such as yoghurt and kefir, while the *L. casei* group, comprising *L. casei, L. paracasei* and *L. rhamnosus,* can be found in various types of cheese (Bouton *et al.,* 2002; Markiewicz *et al.,* 2010). *Latobacillus delbrueckii* has also been illustrated as a strain producing biosurfactants and crude oil biodegrading compounds (Thavasi *et al.,* 2006). *Lactobacillus* is known to help prevent infections of the urogenital and intestinal tracts as

well. The dominance of *Lactobacillus* in the vagina is linked with a reduced risk of bacterial vaginosis and urinary tract infections (Reid and Burton, 2002).

*Lactobacillus* species were reliably identified (++) in each of the metagenomes used in this work, except for mammalian blood and mediterranean bathypelagic.

#### 5.3.2.4 Mycobacteria

Most mycobacterial species are ubiquitous and can be found in water, soil, food and vegetation. *M. bovis* infection has been developed by consuming unpasteurised milk. Bacilli Calmette-Guérin, which is a strain of *M. bovis,* is widely used for immunisation against tuberculosis. It is also administered as an immunotherapeutic agent for the treatment of superficial bladder carcinoma or melanoma. *Mycobacterium fortuitum* has been described as a normal flora of the skin (Eisenstadt, 1995). Pathogenic isolates of *Mycobacterium* include (i) *M. tuberculosis* ─ the causative agent of human tuberculosis; (ii) *M. bovis* ─ the causative agent of bovine tuberculosis; (iii) *M. leprae* ─ the causative agent of leprosy; (iv) *M. ulcerans,* which causes Buruli ulcers and is the third most common form of mycobacterial disease in humans; and (v) *M. marium* ─ the causative agent of fish tank granuloma in humans and granulomatous lesions similar to those of *M. tuberculosis* in zebra fish (Demangel *et al.,* 2009; Rahman *et al.,* 2014). The non-pathogenic groups are *M. gilvum*, *M. vanbaalenii* and *M. smegmatis* (Raham *et al.,* 2014). Opportunistic pulmonary infections are mostly caused by members of the MAC that includes *M. avium* and *M. avium-M. intracellulare*, while Crohn's disease in humans is suspected to be caused by the third member of the MAC group, *Mycobacterium avium* subsp. *paratuberculosis* (Cook, 2010; Chiodini *et al.,* 2012).

In the symbiotic microbiomes, *Mycobacteria* were identified in human gut (+/-) and cow gut (+/-), which was to be expected, for in the soil and rhizosphere microbiomes *Mycobacteria* were identified in all microbiomes; in environmental microbiomes *Mycobacteria* were identified in the anthropogenic estuarine (+/-) and sludge (+) environments.

### 5.3.2.5 Prochlorococcus

*Prochlorococcus* is a unicellular marine cyanobacterium, which is found throughout the euphotic zone of open ocean between $45^0$N and $40^0$S, where it carries out a significant portion of global photosynthesis (Partensky *et al.,* 1999; Flombauum *et al.,* 2013; Biller *et al.,* 2014). The genome of *Prochlorococcus* is the smallest of any known free-living photosynthetic cells, ranging from 1.6 to 2.7 Mbp (Kettler *et al.,* 2007). Though the core set of genes present is shared by all strains, notable diversity in the gene content was reported among isolates. The *Prochlorococcus* group has an open pan-genome, such that each newly sequenced genome typically contains various novel genes never identified before (Kettler *et al.,* 2007).

In the symbiotic microbiomes, *Prochlorococcus* was identified with a score above 1 with an orange bar (+) for canine gut/cow gut and human gut and (+/-) in termite gut, which was rather unexpected for these microbiota. Signature sequences of *Prochlorocossus* were also found in the soil and rhizosphere microbiomes, grassland, phyllosphere and rain forest had at least one organism reliably identified with a green bar (++); the forest rhizosphere returned a (+) identification, while for desert soil the result was (+/-). For environmental microbiomes, a hydrothermal vent showed a (+) identification, while the anthropogenic estuarine environment and sludge showed a (+/-) identification.

### 5.3.2.6 Salmonella

Whole genome sequencing of pathogens, immunological trials and characterisation of bacteria-host interactions at the cellular, humoral and mucosal level helped to generate a comprehensive view of the evolution and emergence of pathogens (Feasey *et al.,* 2012). *Salmonella typhimurium* or *Salmonella enterica* var Enteritidis (*S. enteritidis*), which are non-typhoidal *Salmonella,* have been reported to be the major cause of disease across Africa (Feasey *et al.,* 2012). Researchers have also reported disease outbreaks associated with the following serotypes: (i) *Salmonella enterica* var lsangi (*S. isangi*) in South Africa (Wadula *et al.,* 2006); (ii) *Salmonella enterica* var concord (*S. concord*) in Ethopia (Beyene *et al.,* 2011); and (iii) *Salmonella enterica* var Stanleyville (*S. stanleyville*) and *Salmonella enterica* var Dublin (*S. dublin*) in Mali (Tennant *et al.,* 2010). Non-typhoidal *Salmonella* have been established as a major HIV-related pathogen in sub-Saharan African adults (Gilks *et al.,* 1990). While non-typhoidal *Salmonella* have a broad range of hosts among humans and

animals, the typhoidal serotypes *S. typhi* and *S. paratyphi A* are totally host-constrained to people, causing invasive disease in immune-competent individuals (Feasey *et al.*, 2012).

In this work, *Salmonella* was in the canine gut (+) micro-flora representing symbiotic microbiomes. In the soil and rhizosphere microbiomes, *Salmonella* strains were identified in grassland (++). Weak signals of the presence of Salmonella were also seen in the forest rhizosphere (+/-), phyllosphere (+/-), rain forest (+/-) and soybean rhizosphere (+/-) metagenomes. *Salmonella* was not identified in desert soil. In the environmental microbiomes, *Salmonella* was identified in the sludge metagenome (+) and probably in the anthropogenic estuarine (+/-) environment.

#### 5.3.2.7 Shewanella

*Shewanella* genus microorganisms are facultative anaerobes, Gram-negative gamma-Proteobacteria found in various environments, but mostly in marine sediments and frequently in association with fish (Ivanova *et al.*, 2004; Dikow 2011; Wright *et al.*, 2016). *Shewanella* species signify a vital family of dissimilatory metal-reducing bacteria, which enables the transmission of metabolically produced electrons from a cell interior to external electron acceptors such as solid metal oxides during anaerobic respiration (Fredrickson *et al.*, 2008; Wang *et al.*, 2011).

Weak signals of the presence of *Shewanella* were unexpectedly recorded in this study in symbiotic human gut, termite gut and cow gut with a (+/-). In the soil and rhizosphere metagenome, the rain forest and phyllosphere had a (++) identification, grassland (+), forest rhizosphere (+/-), desert soil (+/-) and soybean rhizosphere (+/-). Again unexpectedly, in the environmental microbiomes *Shewanella* was identified in the anthropogenic estuarine and hydrothermal vent with a weak signal (+/-).

#### 5.3.2.8 Streptococcus

The genus *Streptococcus* currently comprises more than 100 recognised species and the number of these species is expected to rise with the increasing availability of next-generation sequencing technologies (Spellerberg and Brandt, 2015). *Streptococcal* bacteria are linked to the development of dental caries. Oral cavity microbes are usually referred to as viridans

streptococci because of the greenish pigmentation produced by these bacteria when grown on blood agar. This reaction is often termed alpha-haemolysis and is suggestive of the presence of hydrogen peroxide production (Nobbs *et al.,* 2009).

*Streptococcus pneumonia* is a Gram-positive coccus and a member of the lactic acid bacteria, which has been described as one of the foremost sources of morbidity and mortality worldwide. The WHO reported that approximately 1 million children die of pneumococcal disease every year in third-world countries (Hoskin *et al.,* 2001; WHO/UNICEF, 2005; WHO, 2007). Pneumococcal infections have been reported to be the foremost cause of death from vaccine-preventable illnesses in children younger than five years (CDC, 2006). Invasive diseases caused by pneumococci include meningitis and pneumonia associated with bacteraemia and emphysema. The risk factors for developing invasive IPD include age, with the highest risk of incidence among young children less than two years old and elderly people older than 65 years; ethnicity and geographic location, with the ability to attend care centres being the main factor; as well as associated chronic sickness (Fletcher *et al.,* 2006; WHO, 2007; Isaacman *et al.,* 2010).

*Streptococcus* was well identified in symbiotic microbiomes in this study, with most having a (+) and (++) identification. However, the mammalian blood metagenome, which is believed to be sterile, also showed a (++) sign, which could be possible in cases of bacteraemia and contamination. In the soil and rhizosphere microbiomes, most identification was of the weak (+/-), signifying that at least one organism had been identified with a score below 1, with an orange bar. *Streptococcus* was not identified in desert soil and the soybean rhizosphere. In environmental microbiomes, *Streptococcus* was well identified in anthropogenic estuariane (++), hydrothermal vent (++) and sludge (+) environments.

In Table 5.3, the researcher looked at the total number of reads in some of the metagenomes (canine gut, grassland, phyllosphere and hydrothermal vent) used, how many of those reads were binned to species of interest using BLASTN alignment against the *nt* NCBI database and to which extent these results corresponded with the barcoding results of this study. The total number of reads in the canine gut metagenome used in this study was 99 125, of which 83 were aligned by BLASTN to *Bacillus cereus,* 138 to *Escherichia coli and Shigella,* 544 to *Lactobacillus,* 18 to *Mycobacteria,* 0 to *Prochlorococcus,* 10 to *Salmonella,* 0 to *Shewanella* and 256 to *Streptococcus* (see Figure 5.1). Alignment of these reads against the diagnostic

barcodes by the program Barcoding 2.0 confirmed their presence in *Bacillus cereus*, *E. coli / Shigella*, *Lactobacillus*, *Salmonella* and *Steptococcus*, and also the absence of *Shewanella* in this metagenome. However, *Mycobacteria* were not identified in canine gut micro-flora by the program, despite the presence of mycobacterial reads identified by the program MEGAN (Figure 5.1). The reason for this is that all mycobacterial reads binned by MEGAN were assigned to the species *M. terrae*, which was not present among diagnostic barcodes generated for the group Mycobacteria. The Barcoding program identified some signals of *Prochlorococcus.* Indeed, the BLASTN search identified 133 reads of Cyanobacteria in canine gut mycroflora, of which 37 were binned to *Chroococcales*; however, MEGAN did not resolve the taxonomy of these reads to the species level. One has to conclude that the presence of *Prochlorococcus* species in the canine gut was not confirmed.

For the soil and rhizosphere metagenome, the grassland metagenome used yielded a total of 134 368 reads; 24 of them were aligned to *Bacillus cereus,* seven to *Escherichia coli/Shigella,* 23 to *Lactobacillus,* 4 796 to *Mycobacteria,* five to *Prochlorococcus,* 13 to *Salmonella,* 19 to *Shewanella* and seven to *Streptococcus*. In total 4 907 reads were binned to microorganisms of interest used in the case studies. The phyllosphere metagenome yielded a total of 1 933 702 reads, of which nine aligned to *Bacillus cereus*, 13 to *Escherichia coli/Shigella,* 40 to *Lactobacillus,* 4 024 to *Mycobacteria,* 12 to *Prochlorococcus*, 29 to *Salmonella*, 60 to *Shewanella* and 34 to *Streptococcus*. In total 4 221 reads were assigned to the microorganisms of interest. Mycobacteria were the most abundant among bacterial taxa barcoded in this study; however, they were not scored high by the program Barcoding. The reason for this may be that the genes selected for barcodes were too conserved among mycobacterial species and aligning of reads among multiple barcodes would diminish the final score. The current version of the program does not allow any direct control of the level of sequence similarity between diagnostic barcode sequences at the time of their generation by the program BarcodeGenerator. This function will be added to the next version of the program to improve the sensitivity of diagnostic barcodes.

Barcoding of other taxonomic groups confirmed the binning results obtained by MEGAN; however, the barcode scoring did not correlate with the numbers of reads binned to these taxa by the program MEGAN. This may be explained by the fact that MEGAN usually does not

allow resolving of taxonomy below the genus level and the barcode sequences generated for these case studies were species- and strain-specific.

In the environmental microbiomes, the hydrothermal vent metagenome comprised 11 326 reads, 28 of which were aligned to *Bacillus cereus,* 0 to *Escherichia coli/Shigella,* 14 to *Lactobacillus,* 0 to *Mycobacteria,* 19 to *Prochlorococcus,* 0 to *Salmonella,* 15 to *Shewanella* and 11 to *Streptococcus*. In total 87 reads were assigned by MEGAN to the microorganisms of interest. Very unexpected were strong signals of *E. coli* and *Streptococcus* returns by the program Barcoding for this rather exotic environment, where none of these microorganisms could be expected. Binning of the reads by BLASTN against the *nt* database confirmed the presence of *Streptococcus,* but not *E. coli*. This discrepancy can be explained by contamination of the sample with a small amount of *E. coli* DNA from humans working with these samples. Barcodes are sensitive and can identify the presence of this DNA, while MEGAN reports taxonomic units only if the number of assigned reads is above a certain cutoff value.

A stronger scoring of Shewanella barcodes could be expected for this marine environment. MEGAN has identified DNA reads generated from *S. violacea,* which was barcoded for this study. Low scoring may be explained by the same problem that was reported for Mycobacteria – a high level of sequence similarity between the generated diagnostic barcodes that reduces the specificity of the barcodes.

**Table 5.3: Shows the total number of reads for some of the metagenomes used and the results obtained.**

| **METAGENOMES** | **BARCODES** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Bacillus cereus* | *Escherichia-Shigella* | *Lactobacillus* | *Mycobacteria* | *Prochlorococcus* | *Salmonella* | *Shewanella* | *Streptococcus* |
| Canine gut | + 83 | + 138 | + 544 | — M.terrae - 18 | + 0 | + S.enetrica - 10 | — 0 | + 256 |
| Grassland | +/— 24 | +/— 7 | ++ 23 | + 4796 | ++ 5 | ++ 13 | + S.baltica - 19 | +/— 7 |
| Phyllosphere | +/— 9 | +/— 13 | ++ 40 | + 4024 | ++ 12 | +/— 29 | ++ 60 | +/— 34 |
| Hydrothermal Vent | +/— 28 | ++ 0 | ++ 14 | — 0 | + 19 | — 0 | +/— S.violacea - | ++ S.anginosus |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 11326 | | | | | | | 15 | - 11 |

In the following sections, the consistency of identification of different barcoded taxonomic groups by Barcoding 2.0 was determined.

### 5.4. Consistency of identification of taxonomic groups in real metagenomes

The following study was conducted with the aim to validate identification of species and predict the species content of metagenomic samples.

#### 5.4.1 Analysis of *Lactobacillus* in different metagenomes

##### 5.4.1.1 Gut micro-flora

Twenty-eight strains from the *Lactobacillus* group representing different species and subspecies, including commercial and biotechnological potential strains, were used to generate genetic barcodes for *Lactobacillus* (Chapter 2). When DNA reads from the canine gut metagenomic datasets obtained from the MG-RAST database were aligned with BLASTN using the Barcoding 2.0 program, strong signals were obtained for strains of the following *Lactobacillus* species (Figure 5.5A): (i) *Lactobacillus salivarius* [strain UCC118 (NC-007929)] (ii) *Lactobacillus delbruecki* subsp. bulgaricus [strain ATCC BAA-365 (NC-008529)] (iii) *Lactobacillus fermentum* [strain IFO 3956 (NC-010610)] (iv) *Lactobacillus casei* [strain. Zhang (NC-014334)] (v) *Lactobacillus brevis* [strain ATCC 367 (NC-008497)] (vi) *Lactobacillus reuteri* [strain JCM 1112 (NC-010609)] (vii) *Lactobacillus sanfranciscensis* [strain TMW (NC-015978)] and (viii) *Lactobacillus kefiranofaciens* [strain ZW3 (NC-015602)]. This is in agreement with other studies, which show the abundance of *L. acidophilus, L. rhamnosus, L. salivarius, L. fermentum* and *L. reuteri* in canine gut (Pasupathy *et al.*, 2001; Beasley *et al.*, 2006; McCoy and Gilliland, 2007). *Lactobacillus fermentum* has been researched to help prevent and treat urogenital infections and to be effective in inhibiting the growth of harmful bacteria in the canine body (Beasley *et al.*, 2006). *Lactobacillus reuterii* has been studied in dogs and cats; it is known to inhibit the growth of harmful bacteria as well as support the production of natural antibiotic-like substances (McCoy and Gilliland, 2007). *Lactobacillus salivarus* produces a large quantity of lactic acid, which helps to stop the growth of *Helicobacter pylori,* hence reducing the inflammation and risk of dogs with peptic ulcers and irritable bowel syndrome (Beasley *et al.*, 2006).

However, the metagenome analyser MEGAN identified 25 different species of *Lactobacillus* (Figure 5.6) in the canine gut, including those not identified by Barcoding 2. The researcher believes that a subsequent version of Barcoding 2.0 released would identify many other species/strains that was not picked up by this version.

Compared to canine gut, *Lactobacillus amylovorus* [strain GRL 1112 (NC-014724)] yielded a very strong signal with a green bar (Figure 5.5B). *Lactobacillus amylovorus* is a widely abundant specie of *Lactobacillus* found in the intestines of piglets and cows. It is known to have various probiotic properties, such as antimicrobial activity against enteric pathogens (Kant *et al.*, 2011). *Lactobacillus amylovorus* has also been isolated in the bovine uterus, possessing immunomodulatory properties of endometrial cells (Gärtner *et al.*, 2015). Some species of *Lactobacillus* were identified in both canine and cow gut: (i) *Lactobacillus salivarius* [strain UCC118 (NC-007929)]; (ii) *Lactobacillus gasseri* [strain ATCC 33323 (NC-008530)]; (iii) *Lactobacillus fermentum* [strain IFO 3956 (NC-010610)]; (iv) *Lactobacillus casei* [strain Zhang (NC-014334)]; and (v) *Lactobacillus reuteri* [strain JCM 1112 (NC-010609)].

In human gut metagenome *Lactobacillus casei* [strain Zhang (NC-014334) and *Lactobacillus sanfranciscensis* [strain TMW (NC-015978)] showed very strong signals with green bars (Figure 5.5C). A study by Zhang et al. in 2017 proved that *L. casei zhang* or vitamin k12 could significantly alleviate the intestinal tumour burden in mice (Zhang *et al.*, 2017). *Lactobacillus sanfranciscensis* is the main bacterium and probably the most requently used species in the production of traditionally fermented sourdoughs. *Lactobacillus sanfranciscensis* contributes to dough rheology and flavour properties owing to solid acidification by an optimised carbohydrate metabolism and the liberation of precursors of volatile compounds by a proteolytic system and the catabolism of specific amino acid (Vogel *et al.*, 2011). Other *Lactobacillus* species that showed strong signals in orange in the human gut were: (i) *Lactobacillus crispatus* [strain ST1 (NC-014106)]; (ii) *Lactobacillus salivarius* [strain UCC118 (NC-007929)]; (iii) *Lactobacillus fermentum* [strain IFO 3956 (NC-010610)]; (iv) *Lactobacillus buchneri* [strain NRRL B-30929 (NC-015428)]; (v) *Lactobacillus rhamnosus* [strain Lc 705 (NC-013199)]; (vi) *Lactobacillus casei* [strain BL23 (NC-010999)]; (vii) *Lactobacillus sakei* subsp. sakei 23k chromosome (NC-007576); (viii) *Lactobacillus reuteri* [strain JCM 1112 (NC-010609)]; and (ix) *Lactobacillus johnsonii* [strain F19785 (NC-013504)], while *Lactobacillus plantarum* [strain WCFS1 (NC-004567)]

yielded a strong signal with a score above 1 with an orange bar. In termite gut *Lactobacillus casei* str. Zhang (NC-014334) produced strong signals with green bars and scores above 1 (Figure 5.5D).

In this study, *Lactobacillus casei* str. Zhang (NC-014334) was identified in the four metagenomes (canine, cow, human and termite) used. All species of *Lactobacillus* identified in the gut micro-flora represented strains and species that were of commercial and biotechnological importance. The performance of Barcoding 2.0 was average in the identification of different *Lactobacillus* species/strains in the four metagenomes used for gut micro-flora in this study

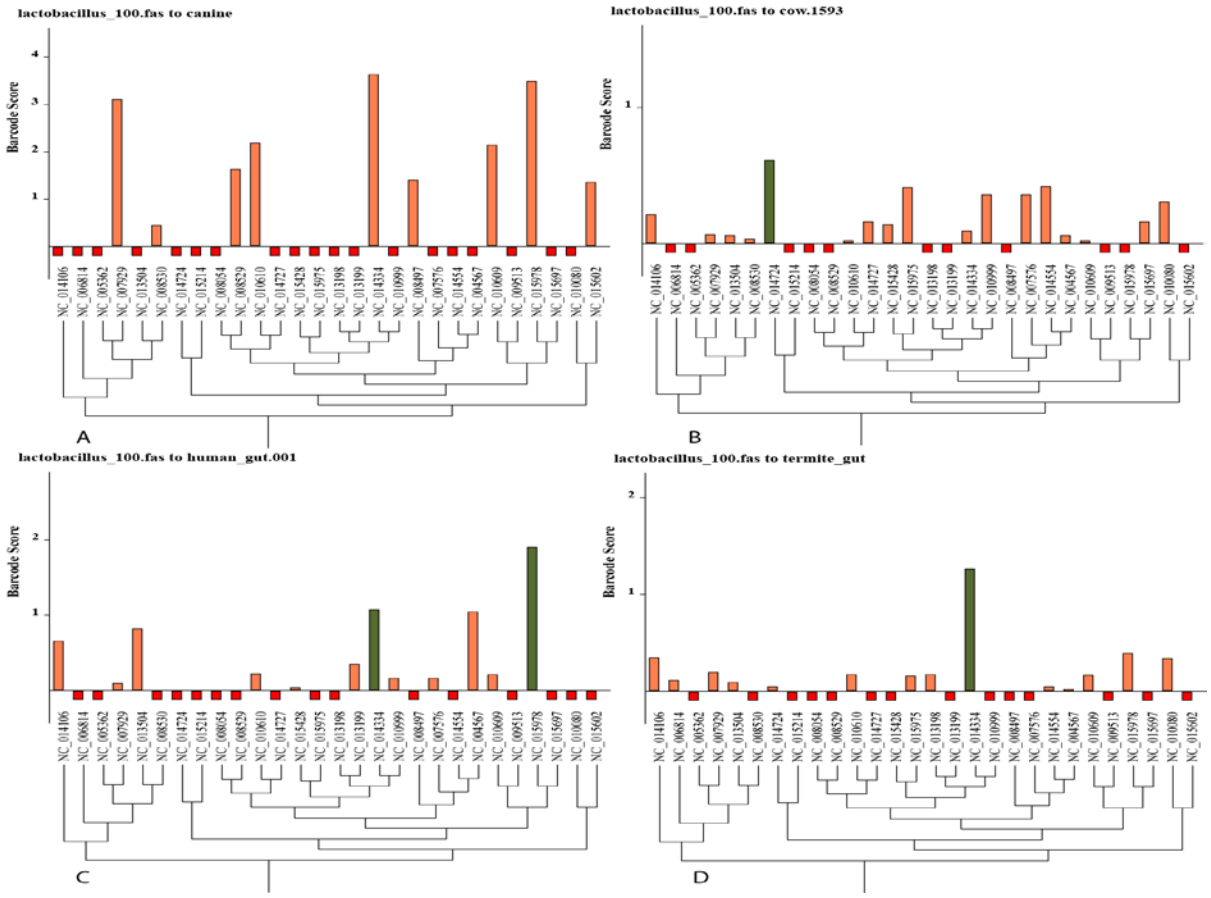



**Figure 5.5: Lactobacillus specie profile in: (A) canine gut (B) cow gut (C) human gut and (d) termite gut.**

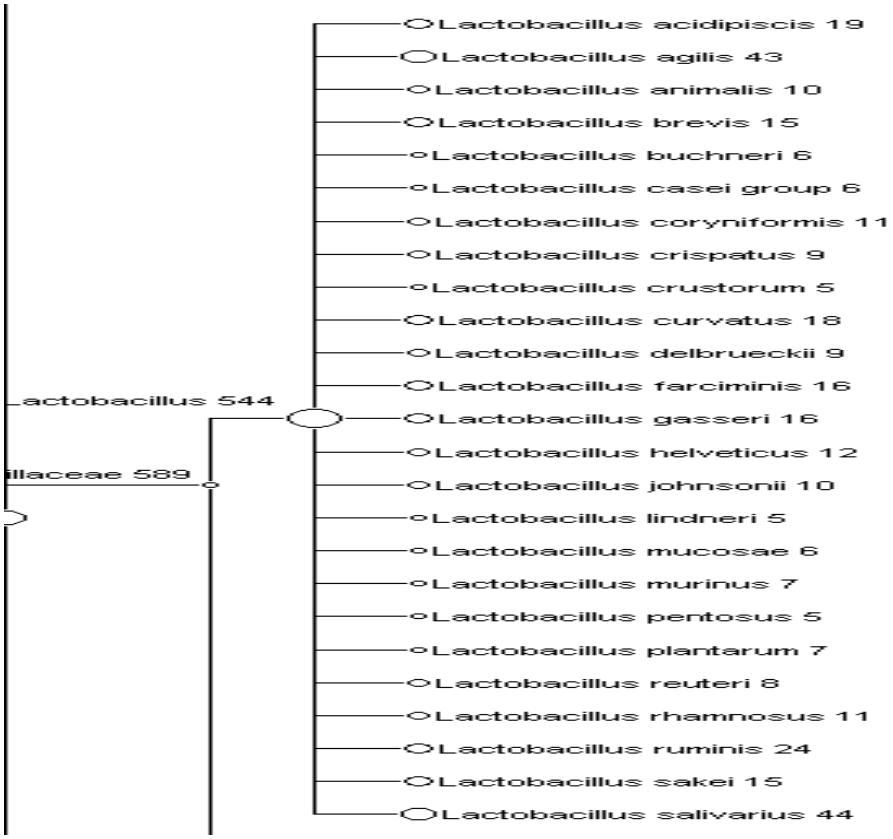



**Figure 5.6: MEGAN analysis showing the different species of *Lactobacillus* in the canine gut.**

### 5.4.1.2 Plant-associated micro-flora

Barcoding 2.0 performed averagely in identifying the species and strains of *Lactobacillus* in plant-associated micro-flora. In plant-associated micro-flora, the *Lactobacillus* specie profile was quite similar for all metagenomes used in this study. However, the *Lactobacillus* specie profile for desert soil was slightly different, as *Lactobacillus* species are not commonly isolated in desert soils. Studies from across the world show that desert soil typically contains a number of ubiquitous phyla, which include actinobacteria, bacteriodetes and proteobacteria (Channal *et al.*, 2006; Makhalanyane *et al.*, 2015). This explains why *Lactobacillus* specie profiles in desert soil were mostly orange bars with scores above 1 (Figure 5.7A)

*Lactobacillus* species with stronger signals with scores above 1 and green bars were mostly identified in the grassland forest, forest rhizosphere, phyllosphere, rain forest and soybean rhizosphere. Strains of *Lactobacillus* species with strong signals in the grassland metagenome include: (i) *Lactobacillus amylovorus* GRL 1112 (NC-014724); (ii) *Lactobacillus buchneri* NRRL B-30929 (NC-015428); (iii) *Lactobacillus casei* BL23 (NC-010999); and (iv) *Lactobacillus sakei* subsp. sakei 23k chromosome (NC-007576).

*Lactobacillus amylovorus* is known to show features of a common homofermentative *Lactobacillus* species, such as the production of enormous quantities of lactic acid and small amounts of acetic acid, but no gas from glucose. *Lactobacillus amylovorus* is one of the main S-layer-carrying *Lactobacillus* species in pigs. It shows strong adherence to pig intestinal epithelial cells and is of interest because of its potential probiotic properties (Kant *et al.,* 2011). Since most pigs do readily graze on grassland, this probably explains why *Lactobacillus amylovorus* GRL 1112 (NC-014724) was clearly identified in the grassland metagenome used in this study.

*Lactobacillus buchneri* is a specie relevant for commercial silage, bioethanol and vegetable fermentations (Briner and Barrangou, 2014). Under anaerobic conditions *Lactobacillus buchneri* is known to use lactic acid to ferment cucumber (Franco *et al.,* 2012). *Lactobacillus casei* is a specie of *Lactobacillus* that is used in several foods, agricultural and industrial fermentations. This leaves room for genetic manipulation of *L. casei,* which can be undertaken to understand their physiological and biochemical properties and allows for the progress of industrial strains (Welker *et al.,* 2014). *Lactobacillus sakei* is a pscychrotrophic lactic acid bacterium found naturally on fresh meat and fish. *Lactobacillus sakei* strain 23k identified in the grassland metagenome in this study was originally isolated from French sausage with specific reference to survival aspects and competition with other meat-borne bacteria. *Lactobacillus sakei* is generally mostly used in the manufacture of fermented meats and has biotechnological potential in biopreservation and food safety (Chaillou *et al.,* 2005) (Figure 5.9B).

For other metagenomes (forest rhizosphere, phyllosphere, rain forest and soybean-rhizosphere, Figure 5.7C - Figure 5.7F) used for plant-associated micro-flora, the specie composition for *Lactobacillus* specie was similar to that of the grassland metagenome. All species/strains of *Lactobacillus* identified in the plant-associated micro-flora metagenome in this study are of biotechnological and commercial importance.

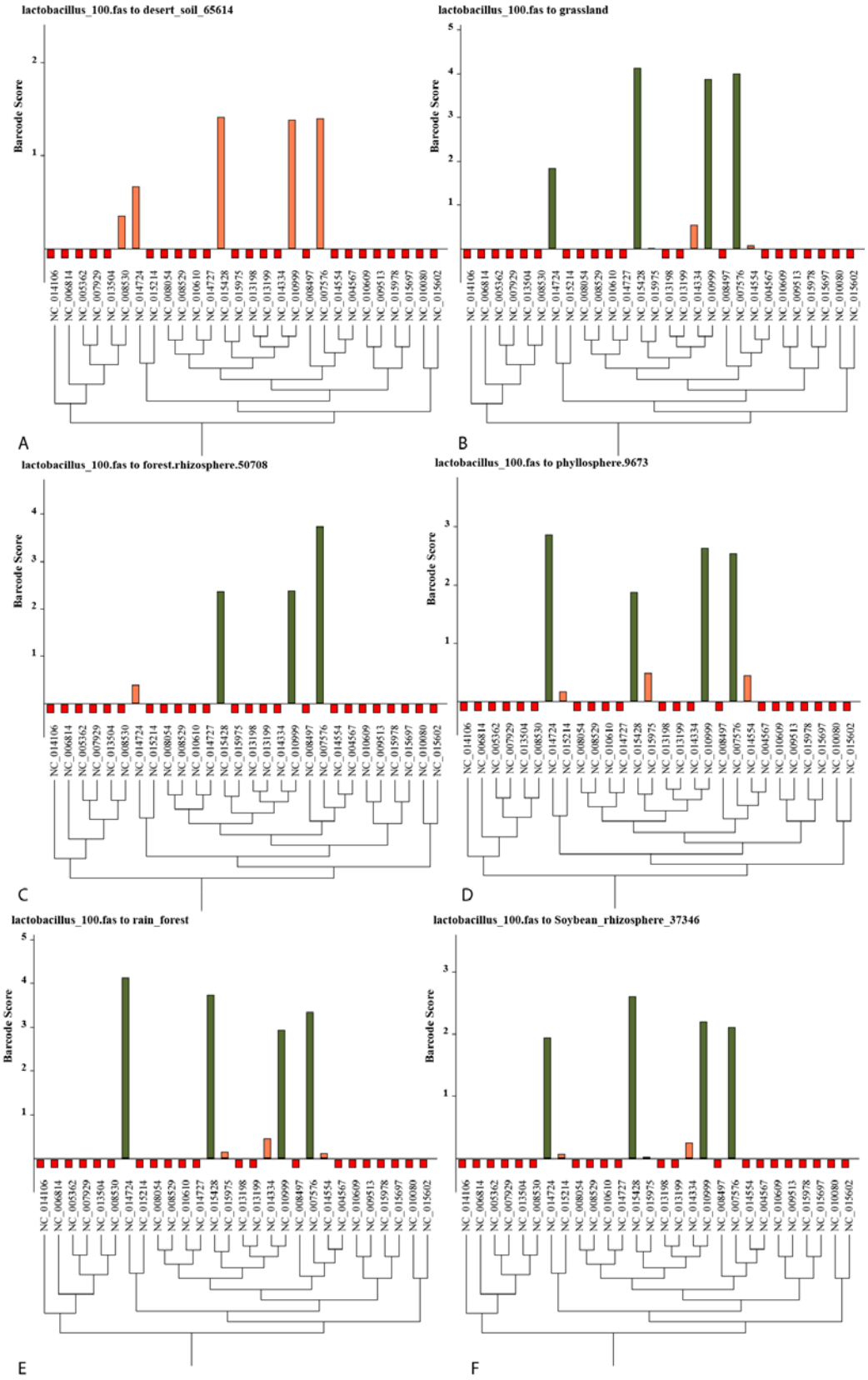



**Figure 5.7: *Lactobacillus* specie profile in: (A) desert soil (B) grass land (c) forest rhizosphere (d) phyllosphere (e) rain forest and (f) soybean -rhizosphere.**

#### 5.4.1.3 Environmental micro-flora

The composition of estuary water is known to be very complex and to vary depending on the degree of anthropogenic inference (Edet *et al.*, 2018). Strong signals were obtained for *Lactobacillus salivarius* [strain UCC118 (NC-007929)] and *Lactobacillus sanfranciscensis* [strain TMW (NC-015978)] in the anthropogenic-estuarine metagenome in this study (Figure 5.8A). *Lactobacillus salivarius* helps to inhibit the growth of *H. pylori*, hence reducing the associated inflammation and risk for dogs with peptic ulcers and irritable bowel syndrome (Beasely *et al.*, 2006), while *Lactobacillus sanfranciscensis* has been described as probably the best adapted specie and is also regarded as the autochthonous key organism of sourdough microbiota (Vogel *et al.*, 2011). *Lactobacillus* specie composition was similar for the anthropogenic-estuarine metagenome and the hydrothermal vent; a strong signal was also obtained in the hydrothermal vent for *Lactobacillus sanfranciscensis* (Figure 5.8B).

The *Lactobacillus* specie composition for the sludge metagenome was slightly different from that of the anthropogenic-estuarine and the hydrothermal vent (Figure 5.8C). Strong signals were obtained for the following strains of *Lactobacillus* species in the sludge: *L. gasseri* [strain ATCC 33323 (NC-008530)], *L. amylovorous* [strain GRL 1112 (NC-014724)] and *L. acidophilus* [strain 30s (NC-015214)] (Figure 5.8C). *Lactobacillus acidophilus* can be isolated everywhere, both in humans and dogs. It is known to have the ability to cling to the intestinal wall without harming it. *Lactobacillus acidophilus* is a stable for any probiotic supplement (Pasupathy *et al.*, 2001), while *L. gasseri* is another specie of *Lactobacillus* widely used as a probiotic for fermented products (Tada *et al.*, 2017).

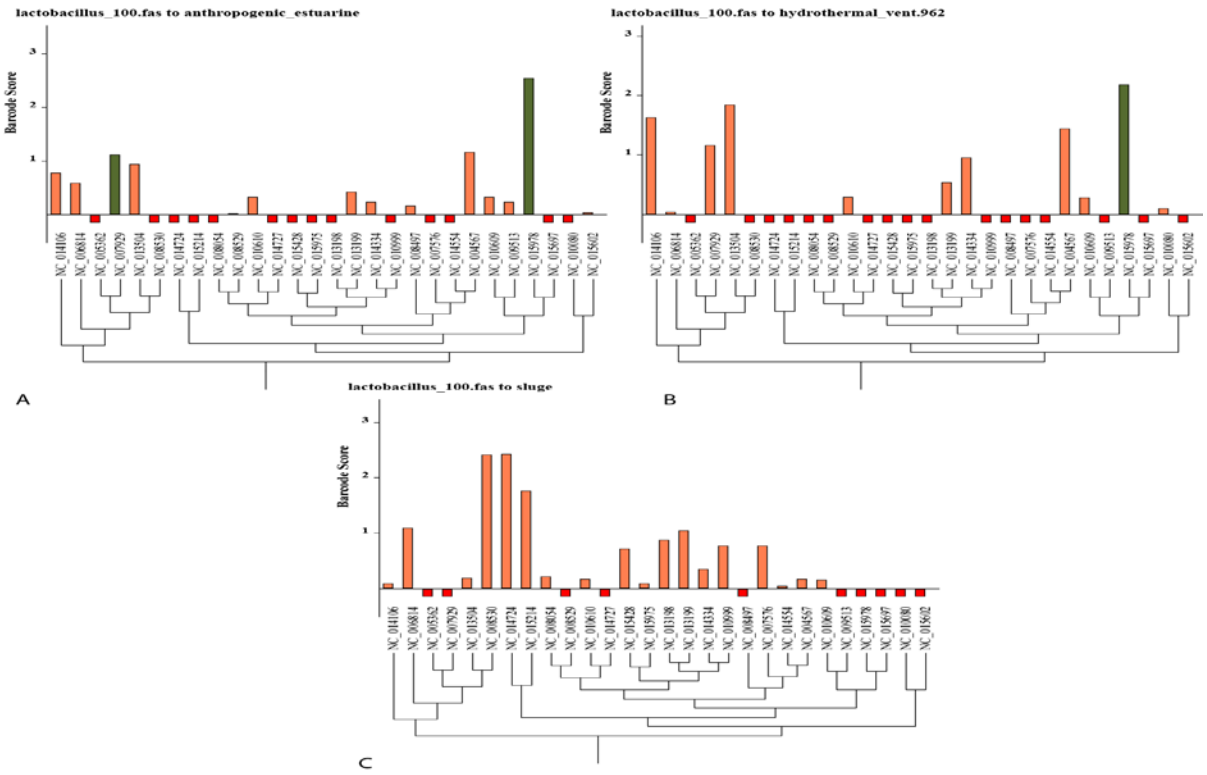



**Figure 5.8: *Lactobacillus* specie profile in: (A) anthropogenic estuarine (B) hydrothermal vent and (C) sludge.**

### 5.4.2 Analysis of *Mycobacteria* in the phyllosphere and grassland

Sixteen strains from the *Mycobacteria* group representing different species and subspecies, including commercial and biotechnological potential strains, were used to generate genetic barcodes for *Mycobacteria* (Chapter 2). When DNA reads from the phyllosphere and grassland metagenomic datasets obtained from the MG-RAST database were aligned with BLASTN using the Barcoding 2.0 program, the *Mycobacteria* species profile obtained in the phyllosphere and grassland were quite similar to the same species of *Mycobacteria*. This is, however, not surprising, as these two metagenomes are somewhat similar. The phyllosphere represents the interface between the above-ground parts of the air. A conservative estimate shows that roughly 1 billion square kilometres of the worldwide leaf surfaces host more than $10^{26}$ bacteria, which are the most abundant colonizers of the habitat (Delmotte *et al.*, 2009). Grasslands are among the largest ecosystems in the world; their area is estimated at 52.5 million square kilometres or 40.5 percent of the terrestrial area, excluding Greenland and Antarctia.

Most species of *Mycobacteria* are ubiquitous and can be found in water, soil, food and vegetation (Eisenstadt, 1995). Signals were obtained for the following strains of *Mycobacteria* species in grassland and the phyllosphere metagenome: (i) *Mycobacterium vanbaalenii* [strain PYR-1 (NC-008726)]; (ii) *Mycobacterium sp.* [strain JDM601 (NC-015576)]; (iii) *Mycobacterium abscessus* (NC-010397); (iv) *Mycobacterium marinum* M (NC-010612); (v) *Mycobacterium intracellulare* MOTT-02 (NC-016947); (vi) *Mycobacterium* sp. MOTT36Y (NC-017904); (vii) *Mycobacterium avium* 104 (NC-008595); (viii) *Mycobacterium avium* subsp. paratuberculosis k-10 (NC-002944); (ix) *Mycobacterium* sp. MCS (NC-008146); (x) *Mycobacterium* sp. KMS (NC-008705); and (xi) *Mycobacterium sp.* JLS (NC-009077) (Figure 5.9).

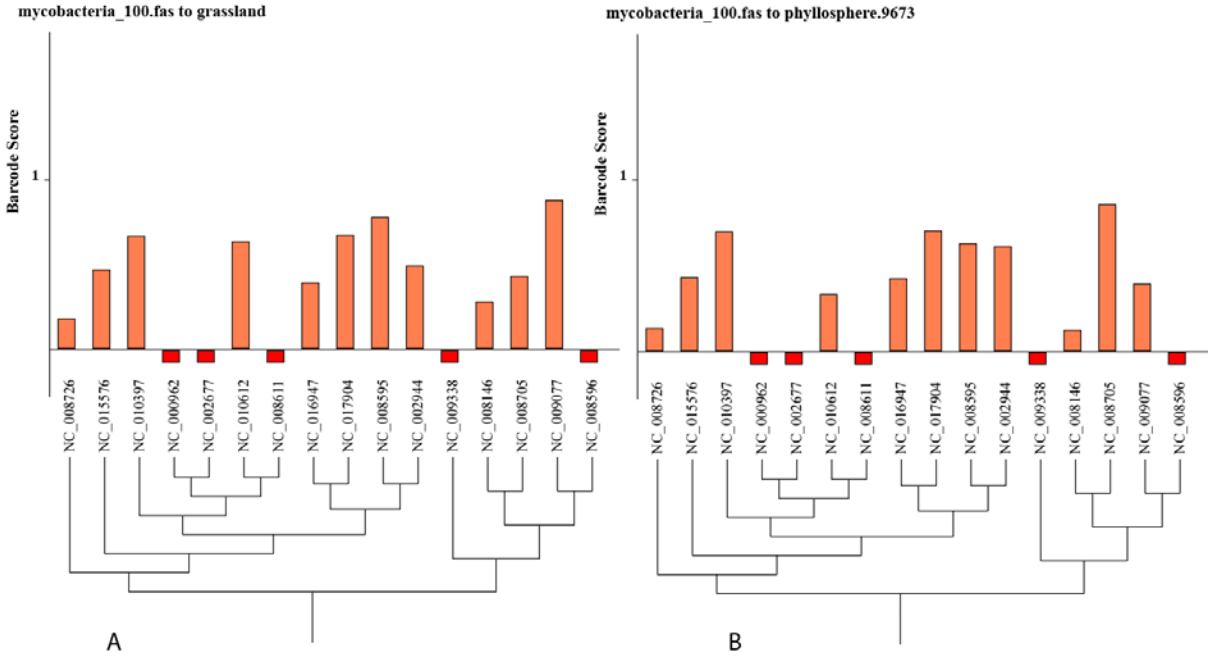
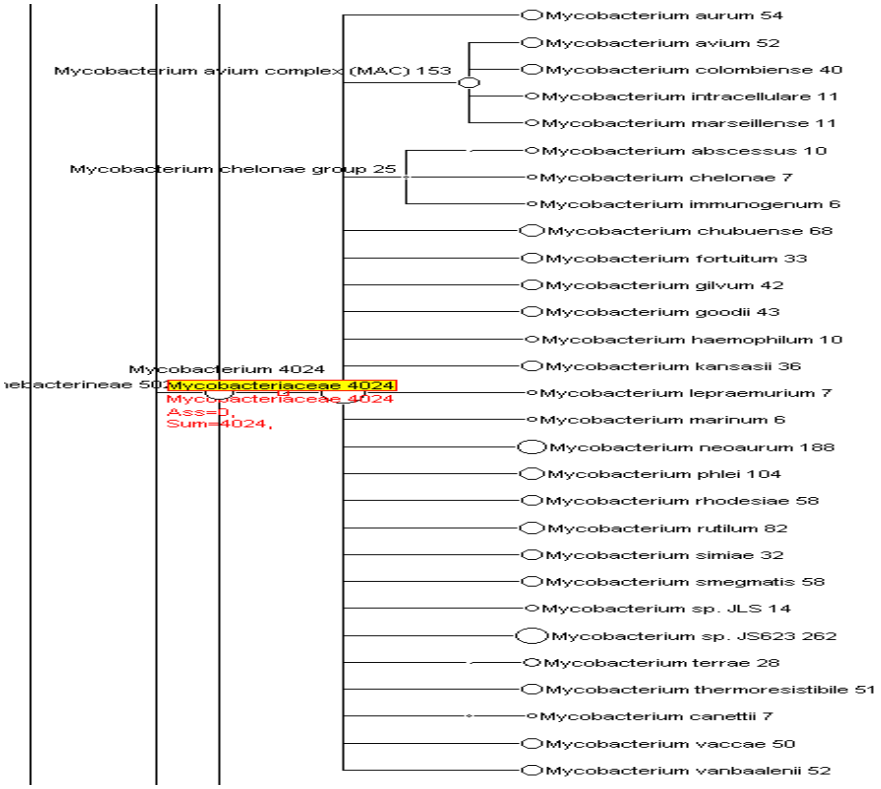



**Figure 5.9: *Mycobacteria* specie profile in the (A) grassland and (B) phyllosphere metagenome.**

However, the metagenome analyser MEGAN identified 29 different species of the *Mycobacteria* phyllosphere and 30 in the grassland (Figure 5.10A/5.10B), including those not identified by Barcoding 2. The researcher believes that a subsequent version of Barcoding 2.0 released would identify many other species/strains that were not picked up by this version.




**A**

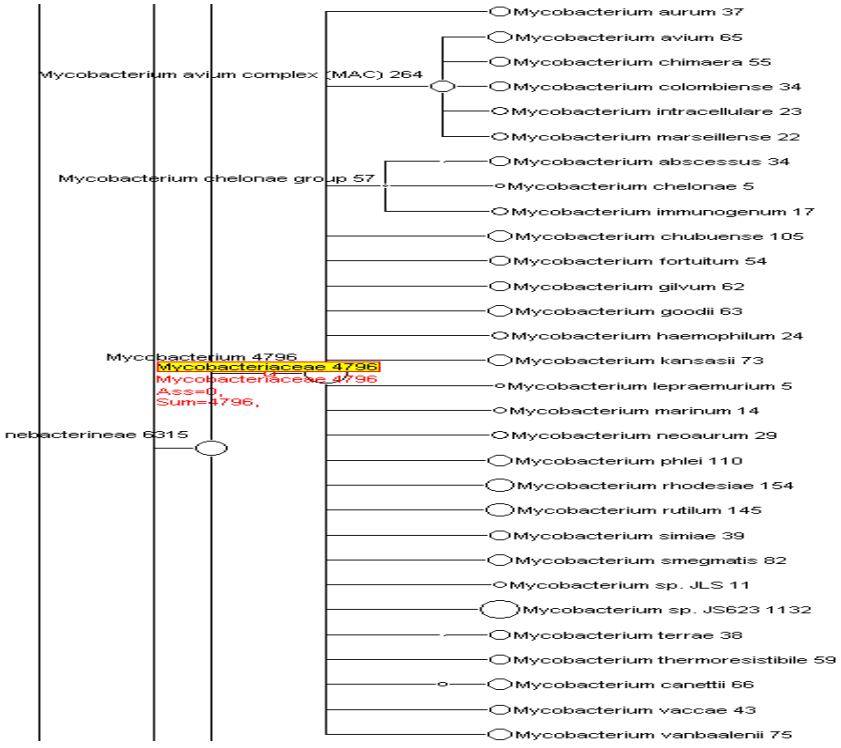



**Figure 5.10: MEGAN analysis showing the different species of Mycobacteria in A) phyllosphere and B) grassland metagenome.**

### 5.4.3. Identification of *Streptococcus* in various metagenomes

#### 5.4.3.1 Analysis of *Streptococcus* in symbiotic microbiomes

The gut metagenome is rich with diverse groups of microorganisms. Hence, it was interesting to investigate if the species of *Streptococcus* seen in human gut, cow gut and canine gut were similar, using the Barcoding 2.0 program. In this study very strong signals were obtained of *Streptococcus equi* subsp. zooepidemicus (NC-012470) and *Streptococcus oralis Uo5* (NC-015291) in human gut. *Streptococcus equi* subsp. zooepidemicus (NC-012470) infection is uncommon in humans. However, a case-control study by Bordes-Benítez *et al*. in 2006 proved that the consumption of inadequately pasteurised cheese of a specific brand was associated with *Streptococcus equi* subsp. zooepidemicus disease (Bordes-Benítez *et al*., 2006). *Streptococcus oralis* is a commensal specie of the human oral cavity and belongs to the Mitis group of *Steptococci* ((Reichmann *et al.,* 2011). *Streptococcus oralis Uo5* is a known high-level penicillin- and multiple-antibiotic-resistant isolate from Hungary. It is competent for genetic transformation under laboratory conditions. Hence, the comparative and functional genomics of *Streptococcus oralis Uo5* will be of importance in understanding the evolution of pathogenesis among the Mitis *Streptococci* and their potential to engage in interspecies gene transfer (Reichmann *et al.,* 2011). Eleven species of *Streptococcus* were picked up in canine gut by the MEGAN (Figure 5.12)

In the cow metagenome, *Streptococcus pyogenes* [strain M1 GAS (NC-002737)], *Streptococcus pyogenes* [strain MGAS8232 (NC-003485)] and *Streptococcus macedonicus* [strain ACA-DC (NC-016749)] yielded strong signals with green bars and scores above 1 (Figure 5.11B). *Streptococcus pyogenes,* also referred to as GAS, for harbouring Lancefeild group A antigen, is a clinically important human pathogen commonly associated with skin or throat infections, but can also cause life-threatening situations including sepsis, streptococcal toxic shock syndrome and necrotising fasciitis (Ibrahim *et al.,* 2011). Physical contact between man and cow probably explains why it was identified in the cow metagenome used in this study. *Streptococcus macedonicus* belongs to the *Streptococcus bovis*/*Streptococcus equinus* complex (SBSEC) and is mostly isolated from fermented foods, mainly of dairy origin. Members of the SBSEC have been implicated in human endocarditis and colon cancer (Papadimitriou *et al.,* 2014). The *Streptococcus* specie composition in human gut and the canine metagenome was quite similar. However, in the canine metagenome, the majority of *Streptococcus* species identified had strong signals with orange bars and scores above 1.

In mammalian blood *Streptococcus equi* subsp. zooepidemicus (NC-012470) was very strongly dominant with green colours and scores of 1. However, the researcher believes that this was a form of contamination (Figure 5.11C).

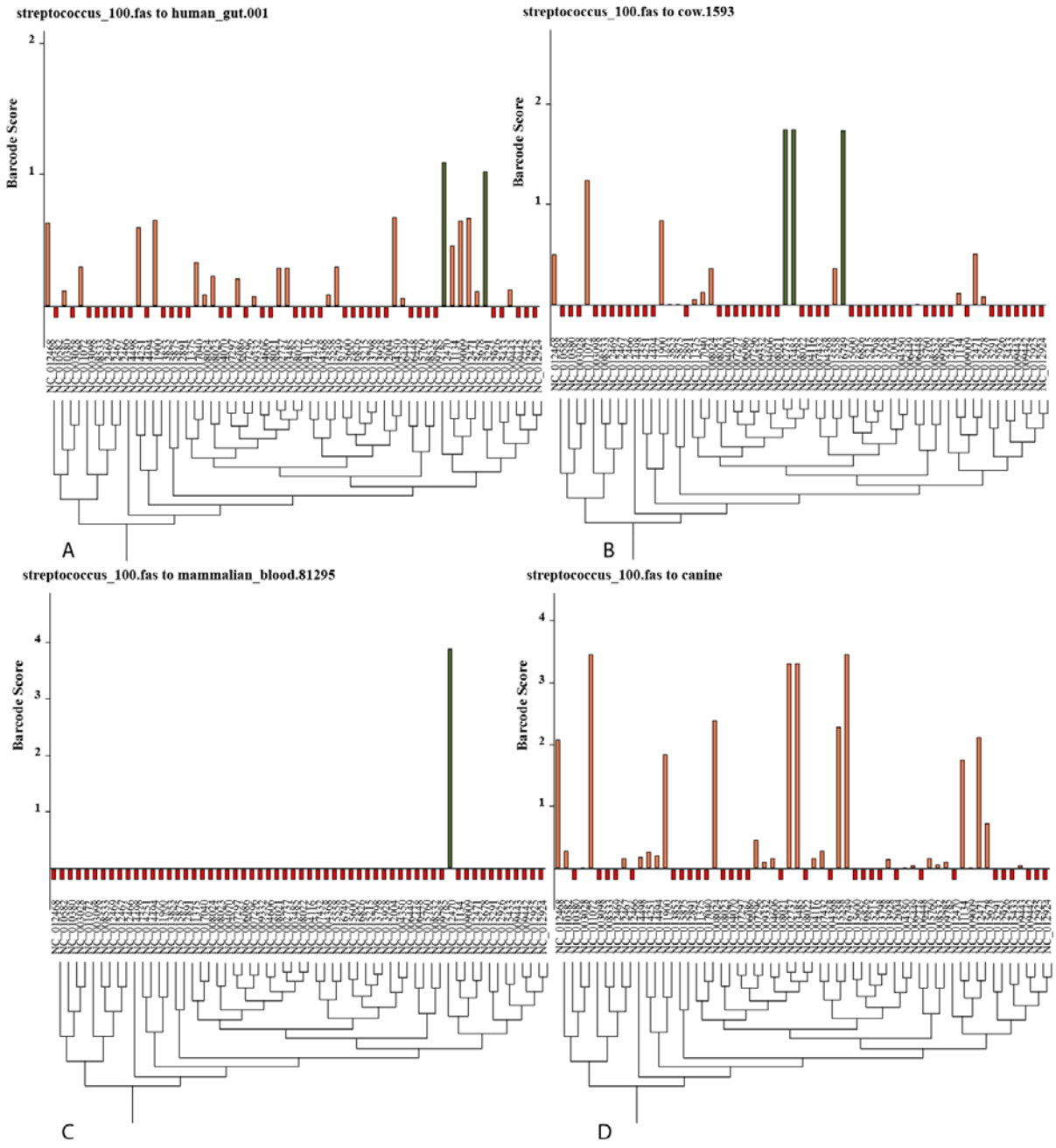



**Figure 5.11: *Streptococcus* specie profile in (A) human gut (B) cow gut (C) mammalian blood and (D) canine gut.**

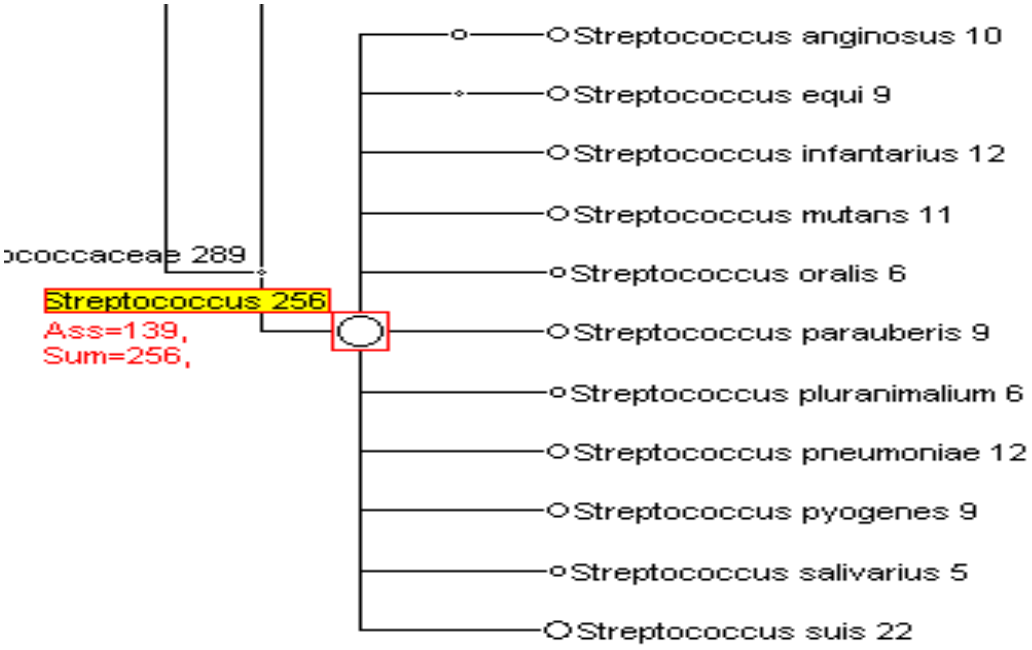



**Figure 5.12: MEGAN analysis showing species of *Streptococcus* in the canine gut.**

### 5.4.3.2 Analysis of *Streptococcus* in environmental metagenomes

The composition of estuary water is known to be very complex and to vary depending on the degree of anthropogenic inference (Edet *et al.*, 2018). In the anthropogenic estuarine environment *Streptococcus pyogenes* [strain MGAS10750 (NC-008024)] yielded a very strong signal with a green bar and score above 1. *Streptococcus pyogenes* is a clinically important human pathogen commonly associated with skin or throat infections, but can also cause life-threatening situations including sepsis, streptococcal toxic shock syndrome and necrotising fasciitis (Ibrahim *et al.*, 2011). In the hydrothermal vent metagenome a strong signal was obtained for *Streptococcus equi* subsp. zooepidemicus (NC-012470). However, the researcher believes that the identification of *Streptococcus equi* subsp. zooepidemicus (NC-012470) in the hydrothermal vent was a form of contamination. Moreover, *Streptococcus* is not a known hydrothermal microorganism. The *Streptococococcus* specie composition for the anthropogenic estuarine and sludge environments was similar. In the sludge, signals were obtained for: (i) *Streptococcus pneumonia* [strain G54 (NC-0011072)]; (ii) *Streptococcus pyogenes* [strain MGAS10750 (NC-008024)]; (iii) NC-002737; (iv) *Streptococcus pyogenes* [strain MGAS8232 (NC-003485)]; (v) *Streptococcus parauberis* KCTC 11537 (NC-015558); (vi) *Streptococcus macedonicus* [strain ACA-DC (NC-016749)]; and (vii) *Streptococcus gordonii* Str. Challis substr. CH1 (NC-009785) with orange bars and scores above 1 (Figure 5.13).

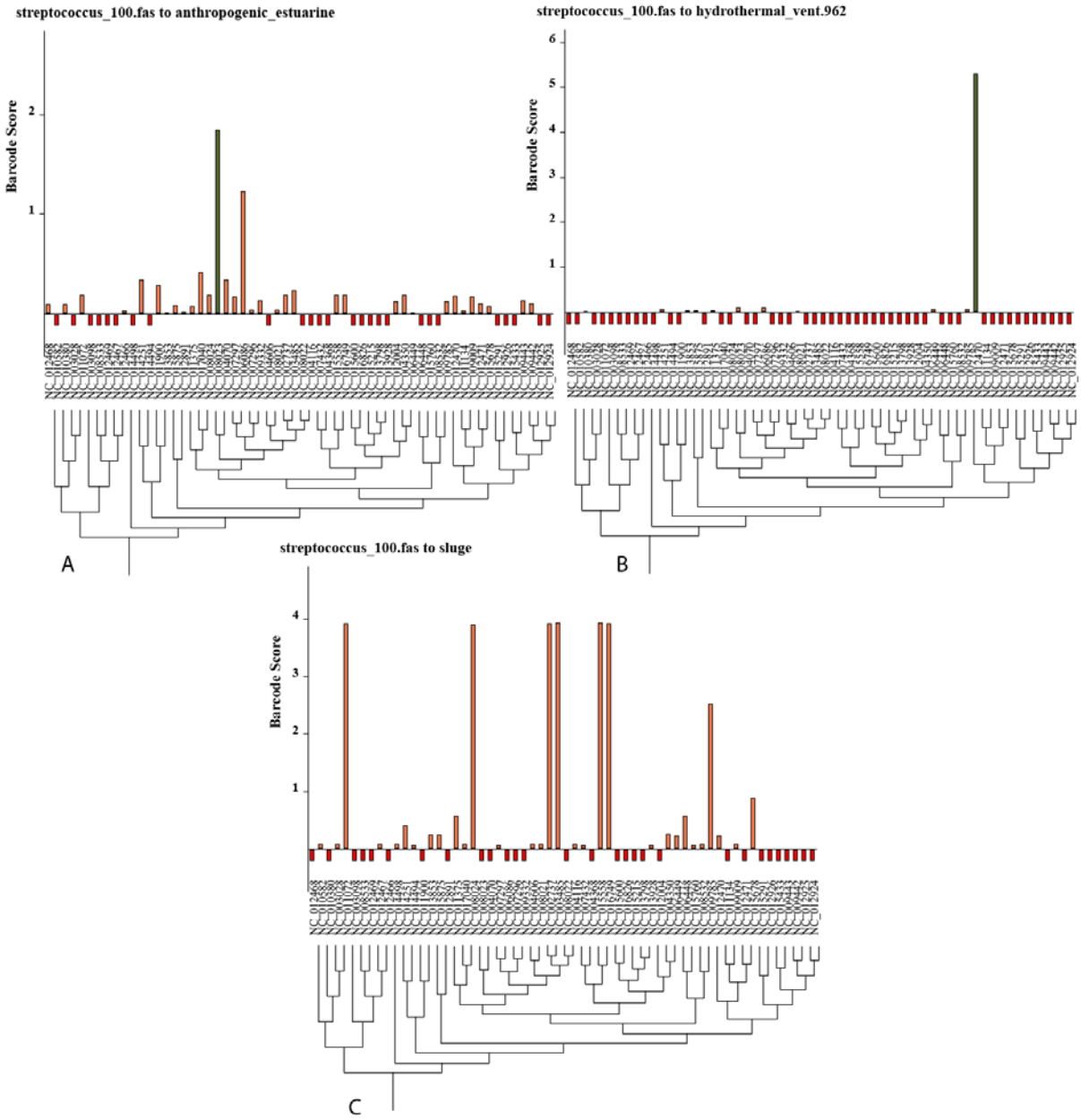



**Figure 5.13: *Streptococcus* specie profile in: (A) anthropogenic estuarine (B) hydrothermal vent and (C) Sludge.**

### 5.4.4 Analysis of *Escherichia coli/Shigella* in the hydrothermal vent metagenome.

Thirty-seven strains from the *Escherichia coli/Shigella* group representing different species and subspecies, including commercial and biotechnological potential strains, were used to generate genetic barcodes for *Escherichia coli/Shigella* (Chapter 2). When DNA reads from the hydrothermal vent metagenomic datasets obtained from the MG-RAST database were aligned with BLASTN using the Barcoding 2.0 program, strong signals were obtained for strains of the following *Escherichia coli/Shigella* species shown in Figure 5.14: (i)

*Escherichia coli* [strain BW2952 (NC-012759)]; (ii) *Escherichia coli* O127:H6 [strain E2348/69 (NC-011601)]; (iii) *Escherichia coli* [strain 536 (NC-008253)]; (iv) *Escherichia coli* [strain IA139 (NC-011750)]; and (v) *Shigella boydii* [strain Sb227 (NC-007613)] as the most dominant species in the hydrothermal vent, with green bars and scores above 1. Species of *Escherichia coli* and *Shigella* are mostly isolated in clinical samples from patients with diarrhoea and show high resistance to antibiotics (Nguyen *et al.*, 2005). Thermophiles are the most common microorganisms seen in hydrothermal vents. The presence of *Escherichia coli/Shigella* in the hydrothermal vent could be due to pollution from the human gastrointestinal tract.

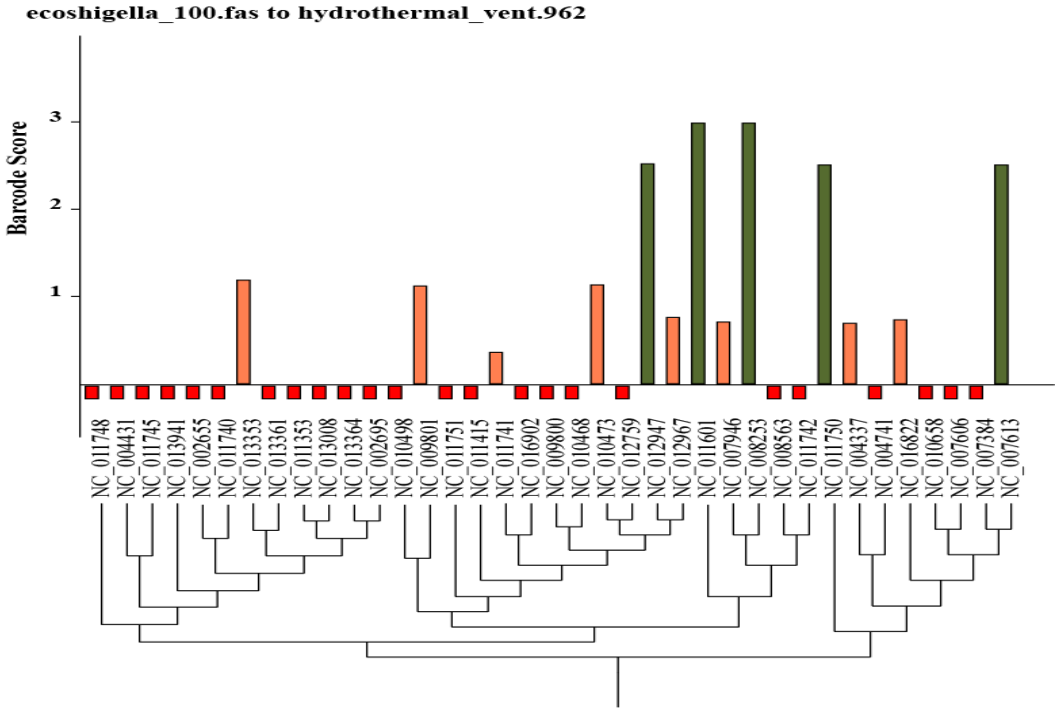



**Figure 5.14: *Escherichia coli/Shigella* specie profile in the hydrothermal vent metagenome.**

### 5.4.5 Analysis of *Shewanella* in the phyllosphere and rain forest

*Shewanella* is a mostly aquatic microorganism. It may, however, be isolated or carried by insects, worms or amphibians to areas such as the phyllosphere and rain forest. *Shewanella* was clearly identified in the phyllosphere and rain forest in this study using the Barcoding 2.0 program. *Shewanella baltica* [strain OS195 (NC-00997)] and *Shewanella violacea* [strain DSS12 (NC-014012)] were identified in the phyllosphere and rain forest metagenome with green bars. However, *Shewanella violacea* [strain DSS12 (NC-014012)] was seen in a very

small proportion of the rain forest (Figure 5.15). *Shewanella* specie composition was different in the hydrothermal vent compared to the phyllosphere and rain forest. MEGAN analysis picked up three species of *Shewanella* in the phyllosphere, namely *Shewanella amazonensis, Shewanella baltica* and *Shewanella lohica* (Figure 5.16)

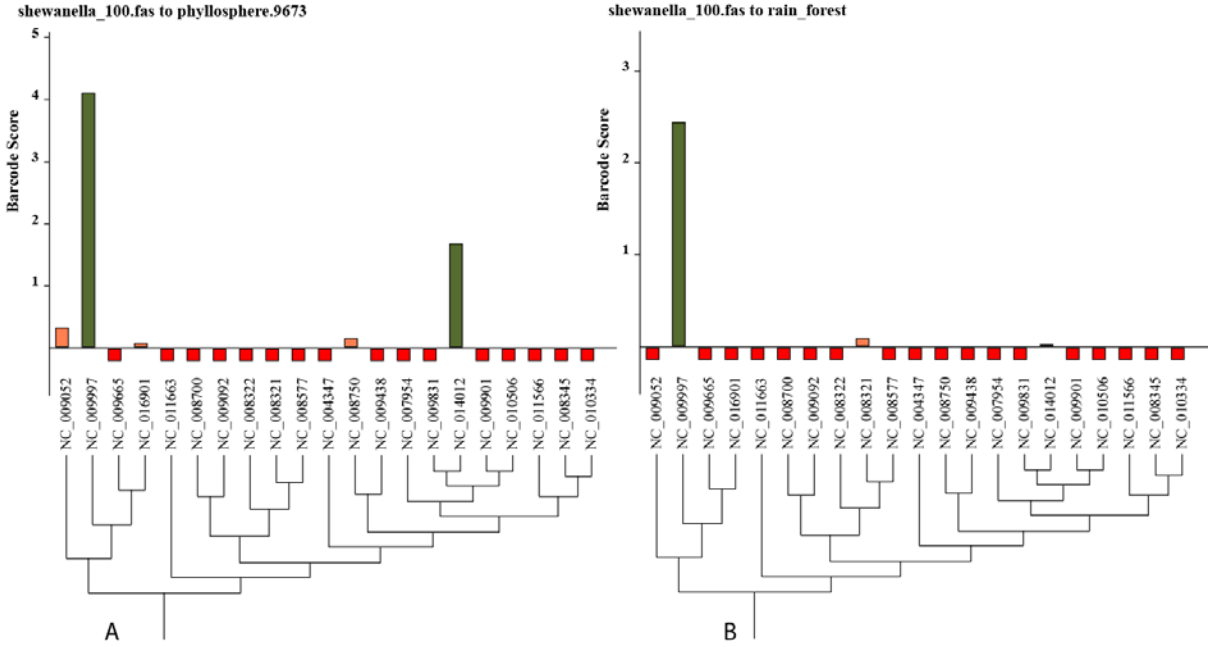



**Figure 5.15: *Shewanella* species profile in the phyllosphere and rain forest metagenome.**

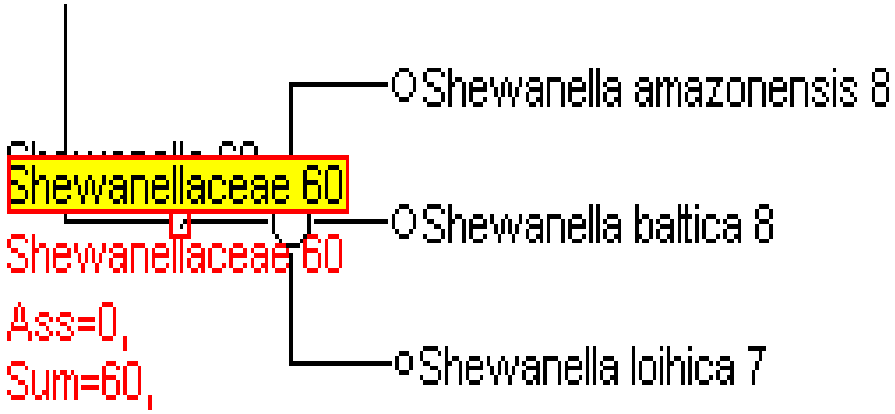



**Figure 5.16: MEGAN analysis showing species of *Shewanella* in the phyllosphere.**

### 5.4.6 Analysis of *Prochlorococcus* in the gut and environmental metagenomes

*Prochlorococcus* are well known marine cyanobacteria. Barcode sequences were created for 12 strains of *Prochlorococcus marinus* (Chapter 2). When DNA reads from cow gut metagenomic datasets obtained from the MG-RAST database were aligned with BLASTN using the Barcoding 2.0 program, strong signals were obtained for *Prochlorococcus marinus*

str.MIT 9301 (NC-009091) in the cow gut. *Prochlorococcus marinus* is the dominant photosynthetic organism in the ocean. It is usually seen in two major ecological forms: HL-adapted genotypes in the upper part of the water column and LL-adapted genotypes at the bottom of the illuminated layer (Dufresne *et al.*, 2003). The grazing of cows in grassland areas close to streams and oceans probably explains why *Prochlorococcus marinus* str.MIT 9301 (NC-009091) was identified in the cow metagenome. *Prochlorococcus* specie composition in canine gut was very different compared to cow gut, as more *Prochlorococcus marinus* strains were identified: (i) *Prochlorococcus marinus* str. NATL2A (NC-007335); (ii) *Prochlorococcus marinus* str. MIT 9313 (NC-005071); (iii) *Prochlorococcus marinus* subsp. marinus str. CCMP1375 (NC-005042); (iv) *Prochlorococcus marinus* str. MIT 9312 (NC-007577); and (v) *Prochlorococcus marinus* str. MIT 9215 (NC-009840). Wild canine animals do feed on everything (vegetable matter, rotten fruit and semi-digested contents of their prey's stomach) and also drink water from streams and oceans, hence this explains why different strains of *Prochlorococcus marinus* were identified in canine gut.

The specie composition of *Prochlorococcus* in the grassland, phyllosphere and rhizosphere metagenome was similar; *Prochlorococcus marinus* str.MIT 9301 (NC-009091) yielded strong signals in the three metagenomes. These metagenomes are similar and close to the streams and oceans where *Prochlorococcus* are mostly found (Figure 5.17 and Figure 5.18).

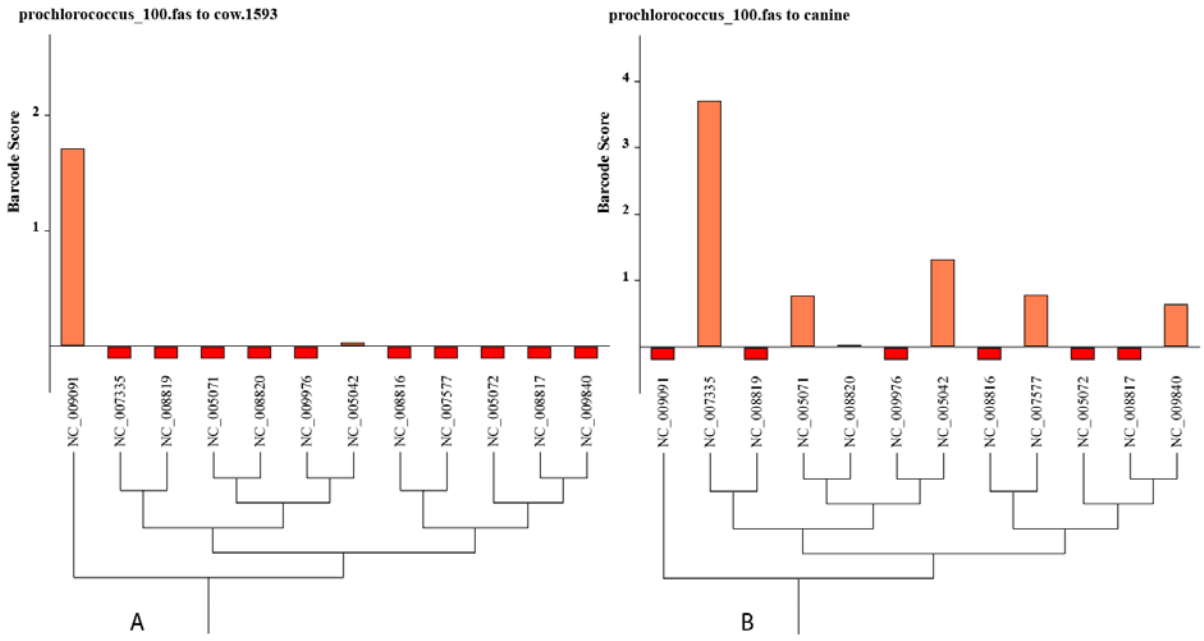



**Figure 5.17: *Prochlorococcus* specie profile in the cow and canine gut metagenome.**

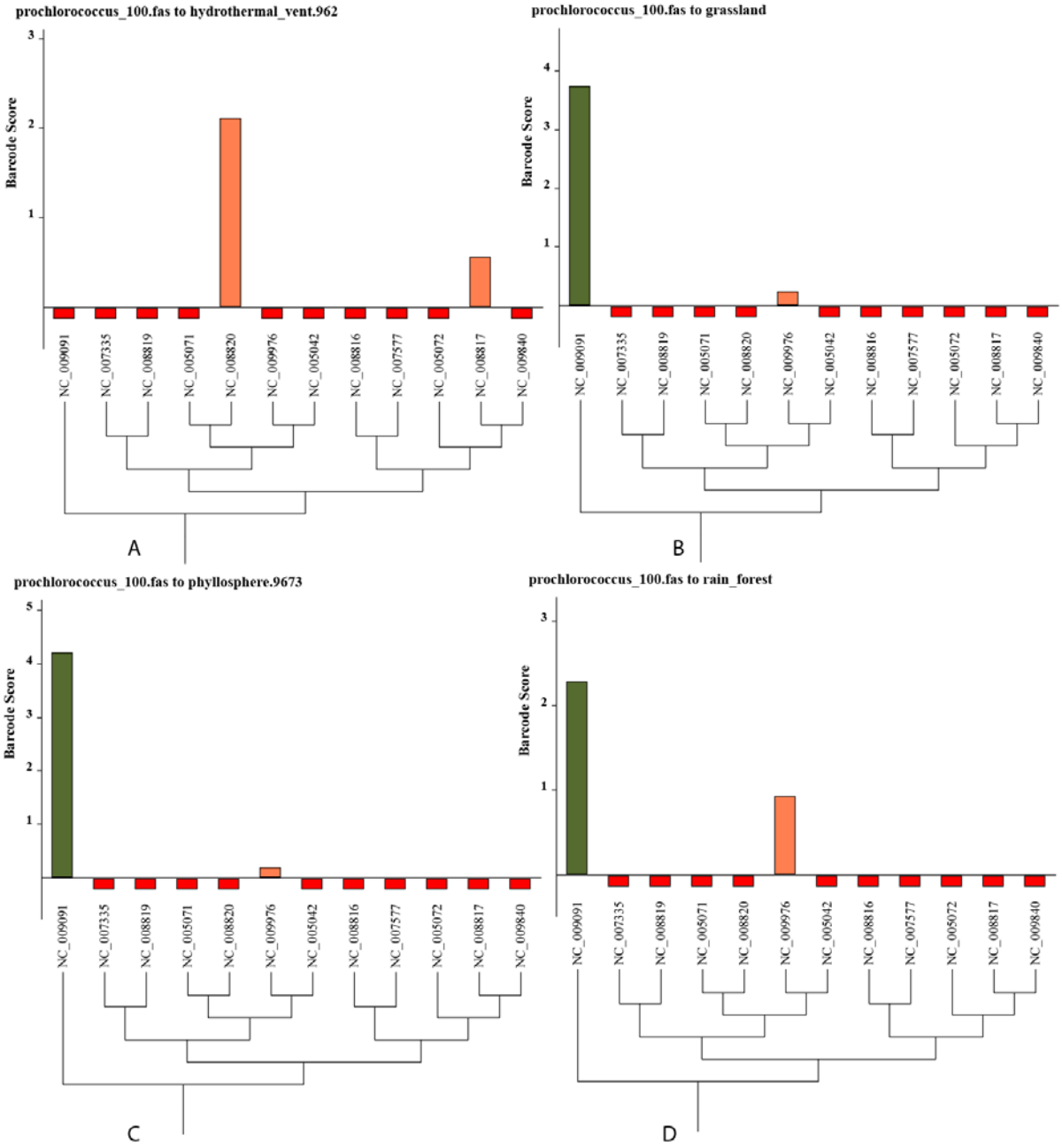



**Figure 5.18: *Prochlorococcus* specie profile in: (A) hydrothermal vent (B) grassland (C) phyllosphere and (D) rain forest.**

## 5.5 Conclusion

In this chapter, the novel command-line program Barcoding 2.0 (Chapter 3) was used for binning of metagenomic reads of different metagenomes obtained from MG-RAST.

The MEGAN program was first used to estimate and interactively explore the taxonomical content of the dataset used, by using the NCBI taxonomy to summarise and order the results.

The results from the MEGAN analysis gave researchers an idea of how well the novel program Barcoding 2.0 performed.

An attempt was then made to test the barcode sequences created by the program BarcodeGenerator (Chapter 2) for selected genomes of various bacteria used as case studies on real metagenomic datasets obtained from MG-RAST using the Barcoding 2.0 program. Compared to the MEGAN program, Barcoding 2.0 performed averagely in the identification of microorganisms. The microorganisms identified by Barcoding 2.0 were of biotechnological and commercial importance. However, the researcher believes that newer versions of Barcoding 2.0 released in the near future will perform much better.

## References


Barko PC, McMichael MA, Swanson KS and Williams DA (2018). The Gastrointestinal microbiome: A review (2018). *The Journal of Veterinary Internal Medicine,* 32, pp. 9-25

Beasley SS, Mannienen TJ and Saris PE (2006). Lactic acid bacteria isolated from canine faeces. *Journal of Applied Microbiology,* 101, pp. 131-138

Belay-Tedla A, Zhou X, Su B, Wan S and Luo Y (2009). Labile recalcitrant, and microbial carbon and nitrogen pools of a tall grass prairie soil in the US Great Plains subjected to experimental warming and clipping. *Soil Biology and Biochemistry Journal,* 41, pp. 110-116

Beyene G, Nair S, Asrat D, Mengistu Y, Engers H and Wain J (2011). Multidrug resistant Salmonella concord is a major cause of salmonellosis in children in Ethopia. *The Journal of Infection in Developing Countries,* 5, pp. 23-33

Biller SJ, Berube PM, Berta-Thompson JW, Kelly L, Roggendack SE, Awad L, Roache-Johnson KH, Ding H, Giovannoni SJ, Rocap G, Moore LR and Chisholm SW (2014). Genomes of diverse isolates of marine cyanobacterium *Prochlorococcus. Scientific Data,* 1:140034

Bordes-Benítez A, Sànchez-Onoro M, Suàrez-Bordón P, Garcià-Rojas AJ, Saéz-Nieto JA, Gonzàlez-Garcia A, Álamo-Antúnez I, Sànchez-Maroto A and Bolaños-Rivero M (2006). Outbreak of *Streptococcus equi* subsp. Zooepidemicus infections on the island of Gran Canaria associated with the consumption of inadequately pasteurized cheese. *European Journal of Clinical Microbiology and Infectious Diseases,* 25, pp. 242-246

Bouton Y, Guyot P, Beuvier E, Tailliez P and Grappin R (2002). Use of PCR-based methods and PFGE for typing and monitoring homofermentative *Lactobacilli* during Comte' cheese ripening. *International Journal of Food Microbiology,* 76, pp. 27-28

Briner AE and Barrangou R (2014). *Lactobacillus bucheneri* gentotyping on the basis of clustered regularly interspaced Short Palindromic Repeat (CRISPR) Locus Diversity. *Applied and Environmental Microbiology,* 80, pp. 994-1001

Castro HF, Classen AT, Austin EE, Norby RJ and Schadt CW (2010). Soil microbial community responses to multiple experimental climate change drivers. *Applied and Environmental Microbiology,* 76, pp. 999-1007

Centers for Disease Control and Prevention (2006). Vaccine preventable deaths and the global immunization vision and strategy. *Morbidity and Mortality Weekly Report,* 55, pp. 511-515

Chaillou S, Champomier-Vergés MC, Cornet M, Coq AM, Dudez AM, Martin V, Beaufils S, Darbon-Rongere E, Bossy R, Loux V and Zagorec M (2005). The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23k. *Nature Biotechnology,* 23, pp. 1527-1533

Chan CKK, Hsu AL, Halgamuge SK and Tang SL (2008). Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics,* 9:215

Chanal A, Chapon V, Benzeraran K, Barakat M, Christen R, Achouak W, Barras F and Heulin T (2006). The desert of Tataouine: an extreme environment that hosts a wide diversity of microorganisms and radiotolerant bacteria. *Environmental Microbiology,* 8, pp. 514-525

Chiodini RJ, Chamberlin WM, Sarosiek J and McCallum RW (2012). Crohn's disease and the mycobacterioses: A quarter century later. Causation or simple association? *Critical Reviews in Microbiology,* 38, pp. 52-93

Cook JL (2010). Nontuberculous mycobacteria: opportunistic environmental pathogens for predisposed hosts. *British Medical Bulletin,* 96, pp. 45-59

Croxen MA and Finlay BB (2010). Molecular mechanisms of *Escherichia coli* pathogenicity. *Nature Reviews Microbiology,* 8, pp. 26-28

Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M and Finlay BB (2013). Recent advances in understanding enteric pathogenic *Escherichia coli. Clinical Microbiology Reveiws,* 26, pp. 822-880

Delmont TO, Prestat E, Keegan KP, Faubladier M, Robe P, Clark IM, Pelletier E, Hirsch PR, Meyer F, Gilbert JA, Le Paslier D, Simonet P and Vogel TM, (2012). Structure, fluctuation and magnitude of a natural grassland soil metagenome. *The ISME Journal*; 6, pp. 1677–1687

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, von Mering C and Vorholt JA (2009). Community proteogenomics reveals insights into phyllosphere bacteria. *Proceedings of the National Academy of Sciences of the United States of America,* 106, pp. 16428-16433

Damangel C, Stinear TP and Cole ST (2009). Buruli ulcer: reductive evolution enhances pathogenicity of *Mycobacterium ulcerans. Nature Reviews Microbiology,* 7, pp. 50-60

Dikow RB (2011). Genome-level homology and phylogeny of *Shewanella* (Gammaproteobacteria: Iteromonadales: Shewanellaceae). *BMC Genomics,* 12:237

Dufresne A, Salanoubat M, Paertensky F, Artiguenave F, Axmann IM, Duprat S, Galperin MY, Koonin EV, Le Gall F, Makarova KS, Ostrowski M, Oztas, Robert C, et al., (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* ss120, a nearly minimal oxyphototrophic genome. *Proceedings of the National Academy of Sciences of the United States of America*, 100, pp. 10020-10025

Edet UO, Antai SP, Brooks AA, and Asitok AD (2018). Microbiological examination and physiological analysis of Estuary water used as a point of source drinking water. *International Journal of Pathogen Research,* 1, pp. 1-18

Eisenstadt and Hall DGS (1995). Microbiology and classification of *Mycobacteria. Clinics in Dermatology,* 13, pp. 197-206

Feasey NA, Dougan G, Kingsley RA, Heyderman RS and Gordon MA (2012). Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa. *Lancet,* 379, pp. 2489-2499

Fisher CR, MacDonald IR, Sassen R, Young CM, Macko SA, Hourdez S, Carney RS, Joye S and McMullin (2000). Methane Ice Worms: *Hesiocaeca methanicola* colonizing fossil fuel reserves. *Naturwissenschaften,* 87, pp. 184-187

Fletcher MA, Laufer DS, Mclntosh EDG, Cimino C and Malinoski FJ (2006). Controlling invasive pneumococcal disease: is vaccination of at-risk groups sufficient. *The International Journal of Clinical Practice,* 60, pp. 450-456

Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jia ON, Karl DM, Li WKW, Lomas MW, Veneziano D, Vera CS, Vrugt JA and Martiny AC (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus. Proceedings of the National Academy of Sciences of the United States of America,* 110, pp. 9824-9829

Franco W, Pérez-Diaz IM, Johanningsmeier SD and McFeeters RF (2012). Characteristics of spoilage-associated secondary cucumber fermentation. *Applied and Environmental Microbiology,* 78, pp. 1273-1284

Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, Gardner TS, Nealson KH, Osterman AL, Pinchuk G, Reed JL, Rodionov DA, Rodrigues JLM, Saffarini DA, et al., (2008). Towards environmental systems biology of *Shewanella. Nature Reviews Microbiology,* 6, pp. 592-603

Gärtner MA, Bondzio A, Braun N, Jung M, Einspanier R and Gabler C (2015). Detection and characterisation of *Lactobacillus spp.* In the bovine uterus and their influence on bovine endometrial cells *in vitro. PLoS ONE,* 10:e0119793

Gilks CF, Brindle RJ, Newnham RS, Watkins WM, Waiyaki PG, Were JBO, Otieno LS, Simani PM, Bhatt SM, Lule GN, Okelo GBA, Brindle RJ, Newnham RS, Gilks CF and Warrell DA (1990). Life-threatening bacteremia in HIV-1 seropositive adults admitted to hospital in Nairobi, Kenya. *The Lancet,* 336, pp. 545-549

Guo J, Ni B-J, Han X, Chen X, Bond P, Peng Y, Yuan Z (2017). Data on metagenomic profiles of activated sludge from a full-scale wastewater treatment plant. *Enzyme and Microbial Technology*. 102, pp. 16-25.

Hoskins J, Alborn JR WE, Arnold J, Blaszczak LC, Burgett S, Dehoff BS, Estrem ST, Fritz L, Fu DJ, Fuller W, Geringer C, Gilmour R, Glass JS, et al., (2001). Genome of the bacterium *Streptococcus pneumoniae* strain R6. *Journal of Bacteriology,* 183, pp. 5709-5717

Huson DH, Auch AF, Qi J and Schuster SC (2007). MEGAN analysis of metagenomic data. *Genome Research,* 17, pp. 377-386

Ibrahim J, Eisen JA, Jospin G, Coil DA, Khazen G and Tokajian S (2011). Genome analysis of *Streptococcus pyogenes* associated with pharyngitis and skin infections. *PLoS ONE,* 11:e0168177

Innerebner G, Knief C and Vorholt JA (2011). Protection of *Arabidopsis thaliana* against leaf pathogenic *Pseudomonas syringe* by *Sphingomonas* strains in a controlled model system. *Applied and Environmental Microbiology,* 77, pp. 3202-3210

Isaacman DJ, Mclntosh D and Reinert RR (2010). Burden of lnvasive pneumococcal disease and serotype distribution among *Streptococcus pneumoniae* isolates in young children in Europe: impact of the 7-valent pneumococcal conjugate vaccine and considerations for future conjugate vaccines. *International Journal of Infectious Diseases,* 14, pp. e197-e209

Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V, Bhattacharyya A, Reznik G, Mikharlova N, Lapidus A, Chu L, Mazur M, Goltsman, et al., (2003). Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis. Letters to nature,* 423, pp. 87-91

Ivanova EP, Gorshkova NM, Bowman JP, Lysenko AM, Zhukova NV, Sergeev AF, Mikhailov VV and Nicolau DV (2004). *Shewanella pacific asp*. Nov., a polyunsaturated fatty acid-producing bacterium isolated from sea water. *International Journal of Systematic and Evolutionary Microbiolgy,* 54, pp. 1083-1087

Kant R, Paulin L, Alatalo E, de Vos WM and Palva A (2011). Genome sequence of *Lactobacillus amylovorus* GRL 1112. *Journal of Bacteriology,* 3, pp. 789-790

Kaper JB, Nataro JP and Mobley HL (2004). Pathogenic *Escherichia coli. Nature reviews Microbiology,* 2, pp. 123-140

Kembel SW, O'Connor TK, Arnold HK, Hubbell SP, Wright SJ and Green JL (2012). Relationship between phyllosphere bacterial communities and plant functional traits in neotropical forest. *Proceedings of the National Academy of Sciences of the United States of America,* 111, pp. 13715-13720

Kettler GC, Martiny AC, Haung K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P and Chisholm SW (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus. PLoS Genetics,* 3:e231

Kisand E, Valente A, Lahm A, Tanet G, Lettieri T (2012). Phylogenetic and Functional Metagenomic Profiling for Assessing Microbial Biodiversity in Environmental Monitoring. *PLoS ONE*; 7: e43630.

Lindow SE and Brandl MT (2003). Microbiology of the phyllosphere. *Applied and Environmental Microbiology,* 69, pp. 1875-1883

Makhalanyane TP, Valverde A, Gunnigle E, Frossard A, Ramond JB and Cowan DA (2015). Microbial ecology of host desert edaphic systems. *FEMS Microbiology Reviews,* 39, pp. 203-221

Markiewicz LH, Biedrzycka E, Wasilewska E and Bielecka M (2010). Rapid molecular identification and characteristics of *Lactobacillus* strains. *Folia Microbiologica,* 55, pp. 481-488

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P and Kyrpides N (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods,* 4, pp. 495-500

McCoy S and Gilliland SE (2007). Isolation and characterization of *Lactobacillus* species having potential for use as probiotic cultures for dogs. *Journal of Food Science,* 72, pp. M94-M97

Nguyen TV, Le PV, Le CH and Weintraub A (2005). Antibiotic resistance in Diarrheagenic *Escherichia coli* and *Shigella* strains isolated from children in Hanoi, Vietnam. *Antimicrobial Agents and Chemotheraphy,* 49, pp. 816-819

Nobbs AH, Lamont RJ and Jenkinson HF (2009). *Streptococcus* adherence and colonization. *Microbiology and Molecular Biology Reviews,* 73, pp. 93-104

Papadimitriou K, Anastasiou R, Mavrogonato E, Blom J, Papandreous NC, Hamodraka SJ, Ferreira S, Renault P, Supply P, Pot B and Tsakalidou E (2014). Comparative genomics of diary isolate *Streptococcus macedonicus* ACA-DC 198 against related members of *Streptococcus bovis/Streptococcus equinus* complex. *BMC Genomics* 15:272

Partensky F, Hess WR and Vaulot D (1999). *Prochlorococcus,* a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews,* 63, pp. 106-127

Pasupathy K, Sahoo A and Pathak NN (2001). Effect of *lactobacillus* supplementation on growth and nutrient utilization in mongrel pups. *Archiv für Tierenährung,* 55, pp. 243-253

Rahman SA, Singh Y, Kohli S, Ahmad J, Ehtesham NZ, Tyagi AK Hasnain SE and Lee SY (2014). Comparative analyses of non-pathogenic, opportunistic, and totally pathogenic mycobacteria reveal genomic and biochemical variabilities and highlight the survival attributes of *Mycobacterium tuberculosis. mBio,* 5:e02020

Reed SC, Townsend AR, Cleveland CC, and Nemergut DR (2010). Microbial community shifts influence patterns in tropical forest nitrogen fixation. *Oecologia* 164, pp. 521-531

Reid G and Burton J (2002). Use of *Lactobacillus* to prevent infection by pathogenic bacteria. *Microbes and Infection,* 4, pp. 319-324

Riechmann P, Nuhn M, Denapaite D, Brücker R, Henrich B, Maurer P, Rieger M, Klages S, Reinhard R and Hakenbeck R (2011). Genome of *Streptococcus oralis* strain Uo5. *Journal of Bacteriology,* 193, pp. 2888-2889

Sharpton TJ (2014). An introduction to the analysis of shotgun metagenomic data (2014). *Frontiers in Plant Science,* 5:209

Shaver GR, Canadell J, Chapin III FS, Gurevitch J and Henry G (2000). Global warming and terrestrial ecosystems, a conceptual framework for analysis. *BioScience,* 50:871

Sheik CS, Beasely WH, Elshahed MS, Zhou X, Luo Y and Krumholz LR (2011). Effect of warming and drought on grassland microbial communities. *The ISME Journal,* 5, pp. 1692-1700

Shreiner AB, Kao JY and Young VB (2015). The gut microbiome in health and disease. *Current Opinion in Gasteroenterology*, 31, pp. 69-75

Singh A, Singh DP, Tiwari R, Kumar K, Singh RV, Singh S, Prasanna R, Saxena AK, Nain L (2015). Taxonomic and functional annotation of gut bacterial communities of *Eisenia foetida* and *Perionyx excavates*. *Microbiological Research*, 175, pp. 48-56

Spellerberg B and Brandt C (2015). *Streptocococcus. In* Jorgensen J, Pfaller M, Carroll K, Funke G, Landry M, Richter S, Warnock D (ed), *Manual of clinical Microbiology,* Elventh Edition. ASM Press, Washington, DC. Doi:10.1128/97815558173811. Chapter 22, pp. 383-402

Swanson KS, Dowd SE, Suchodolski JS, Middelbos IS, Vester BM, Barry KA, Nelson KE, Torralba M, Henrissat B, Coutinho PM, Cann IK, White BA and Fahey Jr GC (2011). Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. *The ISME Journal*, 5, pp. 639-649

Tada I, Tanizawa Y, Endo A, Tohno M and Arita M (2017). Revealing the genomic differences between two subgroups in *Lactobacillus gasseri. Bioscience of Microbiota, Food and Health*, 36, pp. 155-159

Tennant SM, Diallo S, Levy H, Livio S, Sow SO, Tapia M, Fields PI, Mikoleit M, Tamboura B, Kotloff KL, Nataro JP, Galen JE and Levine MM (2010). Identification by PCR of non-typhoidal *Salmonella enterica* serovars associated with invasive infections among febrile patients in Mali. *PLoS Necglected Tropical Diseases*, 4:e621

Thavasi R (2006). Biosurfactants from marine hydrocarbonoclastic bacteria and their application in marine oil pollution abatement. Ph.D Thesis, Annamali University, India p.162

Thomas T, Gilbert J and Meyer F (2012). Metagenomics-a guide from sampling to data analysis. *Microbial Informatics and Experimentation,* 2:3

Trofa AF, Ueno-Oslen H, Oiwa R and Yoshikawa M (1999). Dr. Kiyoshi Shiga: discoverer of the dysentery *bacillus. Clinical Infectious Diseases,* 29, pp. 1303-1306

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R and Gordon JI (2007). The human microbiome project. *Nature*, 449, pp. 804-810

Vogel RF, Pavlovic M, Ehrmann MA, Wiezer A, Liesegang H, Offschanka S, Voget S, Angelov A, Böcker G and Liebl W (2011). Genomic analysis reveals *Lactobacillus sanfranciscensis* as stable element in traditional sourdoughs. *Microbial Cell Factories,* 10, Suppl 1: S6

Vorholt JA (2012). Microbial life in the phyllosphere. *Nature Reveiws Microbiology,* 10, pp. 828-840

Wadula J, von Gottberg A, Kilner D, de Jong G, Cohen C, Khoosal M, Keddy K and Crewe-Brown H (2006). Nosocomial outbreak of extended spectrum β-lactamase-producing *Salmonella isangi* in pediatric wards. *The Pediatric Infectious Disease Journal,* 25, pp. 843-844

Wang G, Qian F, Saltikov CW, Jiao Y and Li Y (2011). Microbial reduction of Graphene oxide by *Shewanella. Nano Research,* 4, pp. 563-570

Welker DL, Hughes JE, Steele JL and Broadben JR (2014). High efficiency electrotransformation of *Lactobacillus casei. FEMS Microbiology letters,* 362, pp. 1-6

World Health Organization (2007). Pneumococcal conjugate vaccine for childhood immunization: WHO position paper. *Weekly Epidemiological Record,* 82, pp. 93-104

World Health Organization/United Nations Children's Fund (WHO/UNICEF) (2005). Global immunization vision and strategy. Washington DC: World Health Organization

Wright MH, Farooqui SM, White AR, and Greene AC (2016). Production of manganese oxide nanoparticles by *Shewanella* species. *Applied and Environmental Microbiology,* 85, pp. 5402-5409

Zhang W, Guo H, Cao C, Li L, Kwok L-Y, Zhang H and Sun Z (2017). Adaptation of *Lactobacillus casei zhang* to gentamycin involves an alkaline shock protein. *Frontiers in Microbiology,* 8:231

# CHAPTER 6 General conclusion

## 6.1 Summary

Bacteria and archae (prokaryotes) make up a substantial percentage of the living biomass on earth and help to sustain the geochemical element cycles: an enormously complicated, planetary scale metabolic network (Strous *et al.*, 2012). Prokaryotes form intricate ecological communities comprising an assembly of different species and only a small quota of these species has been cultivated in the laboratory, has been studied experimentally and has a known genome sequence (Strous *et al.*, 2012).

Hence, with the express development of next-generation sequencing methods, metagenomics, also known as environmental metagenomics, has emerged as a thrilling research area that allows the analysis of the microbial environment which humans live (Leung *et al.*, 2011). The DNA fragments of a metagenomics project are usually from several genomes and most of the genome sequences are unidentified. A vital phase in metagenomic analysis is grouping DNA fragments from similar species together, which is referred to as "binning" (Mavromatis *et al.*, 2007), to determine the microbe distribution of the sample and classify species within the sample. Subject to different research requirements, the binning process could be done on different taxonomic levels ranging from kingdom to species (Leung *et al.*, 2011).

Moreover, the choice of binning methods for any metagenomic dataset is steered by the length of reads that make up the datasets. The performance of binning methods centred on the comparison of compositional characteristics of sequences is reliable only for longer sequences, having sufficient lengths to stem a robust compositional signature (Dutta *et al.*, 2014). Sequence alignment-centred binning methods perform better for an extensive range of read lengths. Of all the NGS technologies, Roche 454, Illumina and Ion Torrent systems are most frequently used for metagenomic samples (Mardis 2010; Metzker 2010). The Illumina sequencers offer excellent sequencing throughput and result in read lengths of about 100-200 bp; Roche-454 sequencing technology produces relatively longer reads around 400-600 bp, with suggestively lower throughput (Dutta *et al.*, 2014). Recently, Roche 454 became obsolete and gave way to new technologies: PacBio, MinION and Oxford Nanopore, which produce long reads up to 20 Gbp. However, public databases still contain many metagenomic datasets generated by older technologies.

The aim of this study was to create multi-locus genetic barcodes for identification and

tracking down of biotechnological and/or pathogenic strains in the environment and development of software tools and databases for design and utilisation of genetic barcodes for application in biotechnology and medicine. To achieve this aim, an attempt was made to improve the metagenome binning resolution by creating genome-specific barcodes based on larger selections of core and accessory gene sequences. This protocol was implemented in novel software tools available for use and downloaded from http://bargene.bi.up.ac.za/. BarcodeGenerator is a novel online software tool, which for any given set of genomes would compare all pairs of genes in the genomes and select the most appropriate COG for diagnostic barcodes. The appropriateness of COG for barcoding is estimated by an analysis of alignments of respective DNA and protein sequences, as described in detail in Chapter 2. It was assumed that barcodes have to comprise important core genes under pressure of positive evolutionary selections, which should be distinguishable on the level of species and subspecies, but sufficiently conserved to allow unambiguous identification. To improve the sensitivity of the barcode sequences created, accessory genes that are genome-specific may be added to the barcode sequences. Hence, the program allows the addition of accessory genes to constitute the barcodes in user-defined proportions. The program was implemented in the form of a web application that allows uploading of genome sequences of organisms of interest and then returns a link to the user's e-mail address to the generated barcode sequences in FASTA format, information on the genes selected for barcodes and a graphical file in SVG format. BarcodeGenerator was used to create barcode sequences for different microorganisms used as case studies in this work; all of them were made available from the project website (http://seqword.bi.up.ac.za/barcoder_help_download/index.html). The strains used in the case studies represent different species and subspecies, including pathogenic and biotechnological strains.

To check the consistency of the selection of marker genes, the researcher also investigated the evaluation of ontology terms of genes selected for barcodes in different groups of microorganisms by the BarcodeGenerator. Among the core genes selected for barcodes, the most abundant groups was the gene-encoding ribosomal proteins, enzymes of purine and pyrimidine biosynthetic pathways, ABC transporters, tRNA synthetases and amidotransferases, various oxidoreductases, acyl carrier proteins and several other functional categories. Constituents of the central metabolic pathways were expected to be among the conserved genes involved in bacterial speciation and suitable for barcoding. For example, ribosomal proteins comprised up to 15% of the sequences selected for barcodes by the

program BarcodeGenerator. This finding was in agreement with many publications reporting ribosomal proteins as the most suitable taxonomic and phylogenetic markers used in rMLST (Jolley *et al.*, 2012; Glaeser and Kämpfer, 2015).

Among accessory genes selected for barcodes, the most frequently selected ones were IS1 and IS2 transposases and Orf2/OrfB genes, *Ynhf*-type membrane proteins, phage-related transcriptional regulators and capsular polysaccharide biosynthesis proteins and other mobile elements abundant in bacterial populations. The case study with closely related organisms of *Escherichia* and *Shigella* demonstrated that including accessory genes in barcodes worsens the specificity of the methods, as the mobile elements were shared by all these microorganisms in a random fashion. Accessory genes may improve the sensitivity of the methods when more diverse organisms are to be distinguished, as mobile genetic elements are mostly clade- and species-specific.

Barcode sequences mined by BarcodeGenerator can be used for the identification of species of interest in Roche 454 or Illumina metagenomics datasets. Barcoding 2.0 is an application written in Python 2.5/Python 2.7 with a command-line user interface made available for downloading from the BarcodeGenerator website (http://bargene.bi.up.ac.za/). The program uses BLASTN to align reads against the generated barcode sequences and then calculates several parameters for scoring the results of the BLASTN alignment and individual barcodes. After scoring all the aligned reads, the program calculates scores for every barcode to identify organisms present in metagenome samples. Taxonomic units are identified by comparison of calculated barcode scores to standard cut-off values set by default. This approach may not be applicable for the analysis of metagenomes generated by PacBio and Oxford Nanopore technologies owing to the high rate of sequencing errors and computational inefficacy of BLAST alignment of long reads.

With the aim to determine whether the length of barcode sequences and the number of reads in a metagenomic dataset influence the sensitivity and specificity of the method, artificial metagenomes of different sizes with a pre-defined composition of reads generated from several reference microorganisms were aligned against barcodes of various lengths. In the first experiment, metagenomic datasets of varying sizes from 10 000 to 500 000 reads were aligned against barcodes of the same length (50 kbp). The researcher found that the sample size (the number of reads in a metagenome) has basically no effect on the sensitivity or

specificity of the algorithm in the given range of sizes (see Chapter 3 for more details). In this range of values, the percentage of TPs increased with the sample size proportional to the number of false positives. The ratio of TPs over false prediction was higher in smaller metagenomes. When varying lengths of barcode sequences (10 to 250 kbp) were aligned against an artificial metagenomic dataset of 500 000 reads, the sensitivity and specificity also remained unchanged. However, the ratio of TPs over FNs was optimal when the barcode sequences were in the range from 100 to 200 kbp. All generated barcodes and artificial metagenomic datasets were made available for download from the project website (http://seqword.bi.up.ac.za/barcoder_help_download/index.html).

Receiver operating characteristic curves of the algorithm performance were calculated for different microorganisms used in the artificial metagenomics datasets. Distinguishing between species of the same genus or family by the program was close to perfect, but the program performed worse in distinguishing between strains of *Escherichia coli* and *Shigella*. Closely related organisms could be identified better when barcodes were based solely on core genes.

The web interface provides users with an online access to the program BarcodeGenerator, which creates diagnostic barcodes based on the genome sequences of species of interest submitted by users. To generate a set of barcode sequences, the user uploads the corresponding genome sequences in GenBank format in a single archived ZIP, TAR or GZ file (maximum file size is 500 MB). For barcode generation, the Barcode mode set by default is used as the mode of operation. The user then selects the proportion of accessory genes in the generated barcode sequences and the approximate length of the barcode sequences. The project name is entered alongside an e-mail address, which is used to receive a link to the file with the generated barcode sequences. Having uploaded the input file, the program starts generating barcode sequences according to the program run parameters set by the user. A local version of the program BarcodeGenerator with a command line interface was also made available for advanced users at http://seqword.bi.up.ac.za/barcoder_help_download/barcodegenerator.html.

To perform metagenomic read binning against the generated diagnostic barcode, another command-line program, Barcoding 2.0, written in Python 2.5/Python 2.7, was designed and made available for downloading from http://bargene.bi.up.ac.za/. The command-line

program Barcoding 2.0 can be used for binning reads of WGS metagenomes. The program Barcording 2.0 is a command-line program written in Python 2.7 that aligns metagenomic reads of Roche 454 and/or Illumina against taxon-specific barcode sequences generated by the online program BarcodeGenerator. The program performs BLASTN alignment of reads against barcode sequences and then scores every barcode in a set and every taxonomic unit represented by a corresponding diagnostic barcode. The program workflow and the scores calculated by the program were explained in detail in Chapter 3.

Also available for download from the website are: (i) examples of all generated barcode sequences for all organisms used as case studies; (ii) graphical output of the diagnostic barcode generated for each length (10, 25, 75, 100, 150, 200, 250 kbp); (iii) artificial metagenomes created with the Metasim program with the supporting information regarding the contents of the artificial metagenomes; and (iv) hyperlinks to NCBI resources to provide more detailed information about each barcoded organism and genes selected for generated barcodes. All the programs for this work were written on Python 2.5/Python 2.7 and made accessible at the website http://bargene.bi.up.ac.za/ through a PHP.

An attempt was then made to evaluate the barcode sequences created by the program BarcodeGenerator (Chapter 2) for selected genomes in real metagenomic datasets obtained from MG-RAST database using the Barcoding 2.0 program. However, since this is the first version of the Barcoding 2.0 program released, to validate the results and to determine how well Barcoding 2.0 performed, the researcher first performed a BLASTN alignment of various metagenomic reads used in the case studies against a local copy of the NCBI *nt* database using the BLASTN for Linux implementation of the alignment program installed on the computer server. The MEGAN 4.70.4 program was then used to estimate and interactively explore the taxonomical content of the dataset, using the NCBI taxonomy to summarise and order the results. MEGAN uses a simple algorithm that reads standard BLASTN output files and assigns each read to the LCA of the set of taxa that it hits in comparison. Hence, species-specific sequences are assigned to the taxa near the leaves of the NCBI tree, whereas widely conserved sequences are assigned to high order taxa closer to the root (Huson *et al.,* 2007). Compared to MEGAN, Barcoding 2.0 also performed averagely in identification of microorganisms. All species/strains identified by Barcoding 2.0 represented strains and species that were of commercial and biotechnological importance.

The results from this work showed that the novel online tool BarcodeGenerator

(http://bargene.bi.up.ac.za/) is an efficient approach to generating barcode sequences from a set of complete genomes provided by users. The Barcoding 2.0 program made available from the same resource enabled efficient and practical use of metabarcodes for visualisation of distribution of organisms of interest in environmental and clinical samples.

## 6.2 Conclusion

In this work, the researcher created an interactive web application and software tools for identification of the most suitable marker sequences for DNA-based multi-local barcoding. The basic idea was to allow the selection and use of different marker genes for identification of organisms of interest on different taxonomic levels in environmental samples. The program BarcodeGenerator, available online at http://bargene.bi.up.ac.za, creates genome-specific barcodes based on the core and accessory genes for genome sequences provided by users. Another command-line application, Barcoder 2.0, available for download from the same website, performs binning of metagenomics reads against generated barcodes and visualises the results. It should be noted that these software tools were developed exclusively for metabarcoding, i.e. for identification of strains and species of interest in environmental samples by binning of metagenomics reads, but not for phylogenetic inferences. However, the program Barcoder 2.0 allows the alignment of identified organisms along phylogenetic trees generated by other programs and provided in PHYLIP/Newick format together with other input files.

This type of research is unique, useful and necessary because:

(i) Research on bacterial DNA barcoding is yet very limited and still in its infancy.

(ii) There are no standard interactive computational services for the identification of the most suitable marker sequences for DNA-based multi-local barcoding.

(iii) Most methods used for binning metagenomic reads do not allow identification below the genus level and very often stop on the level of bacterial families.

(iv) There are many bacterial and fungal cultures that have shown significant enzymatic, antibacterial and hormonal activities, which may be of importance for the medical, biotechnological and agricultural industries.

**References**

Dutta A, Tandon D, Mohammed MH, Bose T and Mande SS (2014). Binpairs: Utilization of IIIumina paired-end information for improving efficiency of taxonomic binning of metagenomic sequences. *PLoS One*, 9:e114814

Glaser PS and Kämfer P (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonmy. *Systematic and Applied Microbiology*, 38, pp. 23

Huson DH, Auch AF, Qi J and Schuster SC (2007). MEGAN analysis of metagenomic data. *Genome Research,* 17, pp.377-386

Joelly KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarattttthna H, Harrison OB, Sheppard SK, Cody AJ and Maiden MCJ (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology,* 158, pp.1005-1015

Leung HCM, Yiu SM, Yang B, Peng Y, Wang Y, Liu Z, Chen J, Qin J, Li R and Chin FYL (2011). A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics,* 27, pp. 1489-1495

Mardis ER (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9, pp.387-402

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P and Kyrpides N (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4, pp.495-500

Metzker ML (2010). Sequencing technologies-the next generation. *Nature Review Genetics,* 11, pp.31-46

Strous M, Kraft B, Bisdorf R and Tegetmeyer HE (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. *Microbial Physiology and Metabolism*, 3, pp. 1-11

**Research output**

This research has been presented in local and international conferences and published in peer reviewed scientific journal

**Publications:**

Rotimi AM, Pierneef R and Reva ON (2018). Selection of marker genes for genetic barcoding of microorganisms and binning of metagenomic reads by Barcoder software tools. *BMC Bioinformatics*, 19:309

**Presentations:**

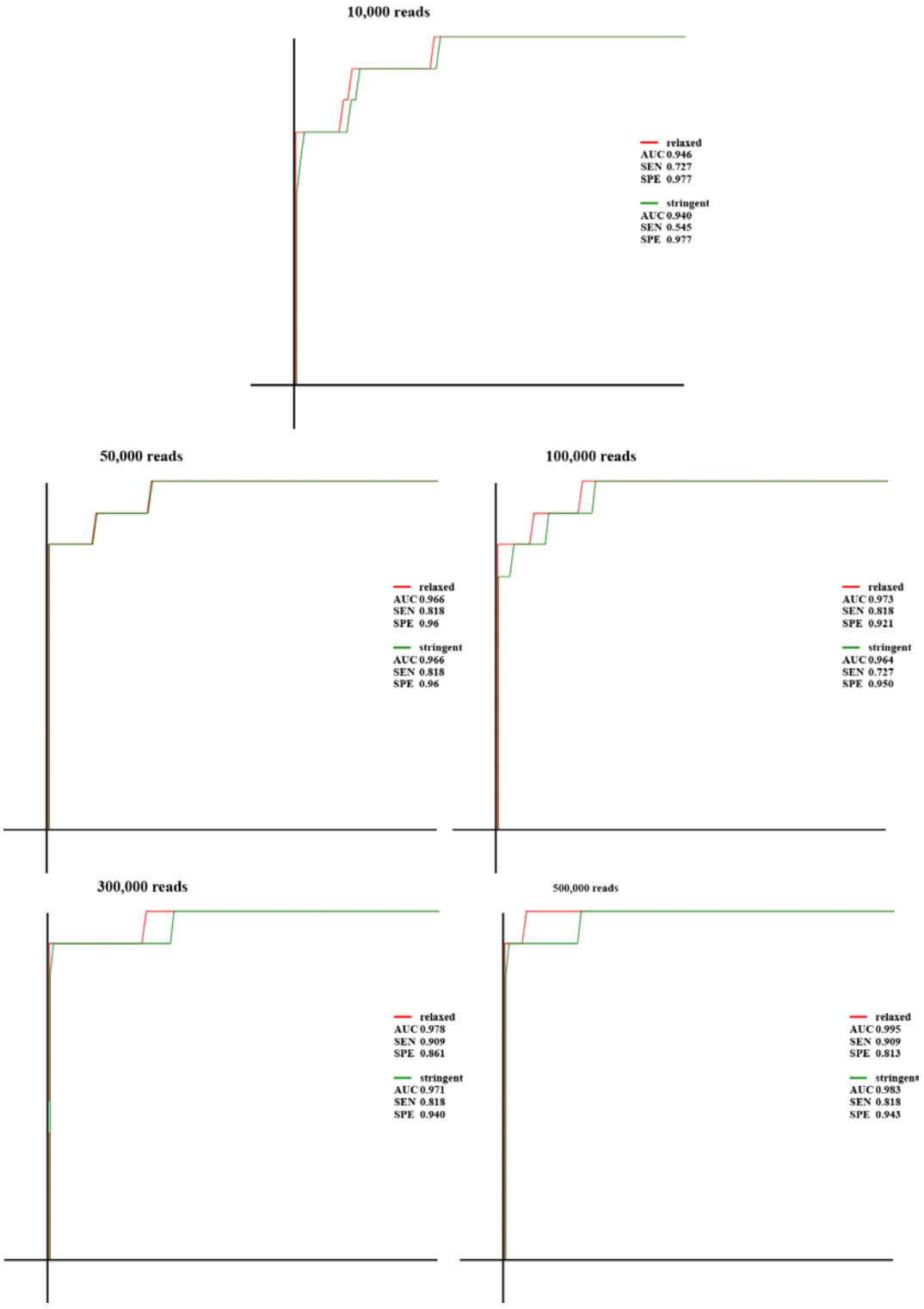
- Rotimi Adeola and Oleg Reva (2016) Genetic barcoding and metabarcoding in Biotechnology. Presented at SASBi (South African Society for Bioinformatics) (Oral Presentation)

- Rotimi Adeola and Oleg Reva (2016) Genetic barcoding and metabarcoding in Biotechnology. Presented at the GRI Symposium University of Pretoria (Oral Presentation)

- Rotimi Adeola and Oleg Reva (2017) Genetic barcoding and metabarcoding in Biotechnology. Presented at the 69$^{th}$ Annual Meeting of the DGHM Microbiology and infection 2017, Wuerzburg, Germany, 5 to 8 March 2017 (Oral presentation).

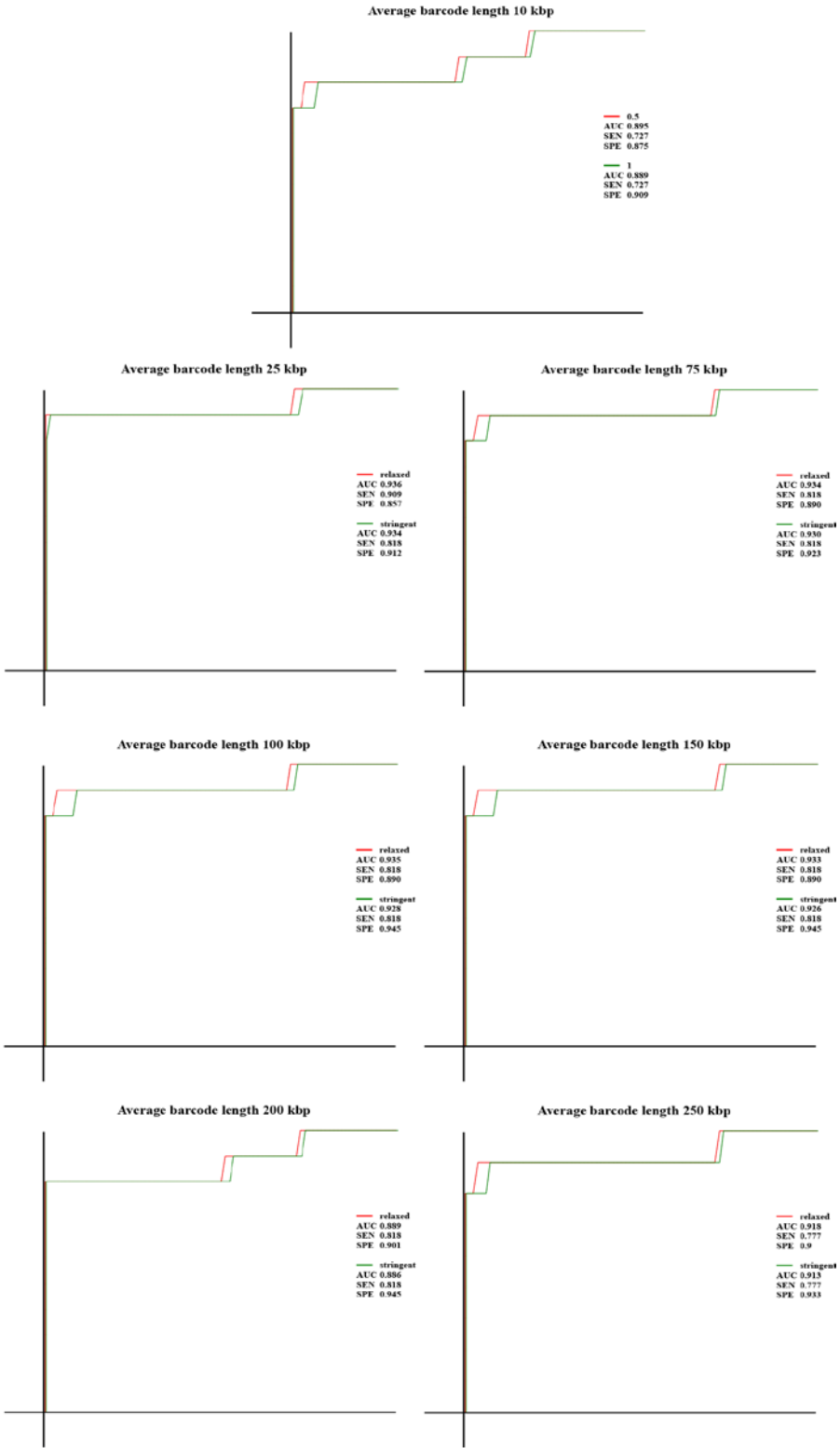
- Rotimi Adeola and Oleg Reva (2017) Genetic barcoding and Metabarcoding in Biotechnology. Presented at the Biochemistry Symposium, University of Pretoria. (Oral Presentation).

## Appendix 1




**ROC diagrams calculated for artificial metagenomic datasets of different sizes for 50,000 bp long barcode sequences. The following parameters were calculated: AUC – area under the curve; SEN – sensitivity; and SPE – specificity**.

## Appendix 2




**ROC diagrams calculated for artificial metagenomic datasets of 500,000 randomly generated reads with barcode sequences of different lengths. The following parameters were calculated: AUC – area under the curve; SEN – sensitivity; and SPE – specificity**.