The evaluation of metagenomic analysis software, using *in-silico* and *in-vitro* mock community datasets, for the accurate study of bio-aerosol samples.

Submitted by Anthony Messer, to the University of Exeter as a thesis for the degree of Masters by Research in Biological Sciences, May 2019.

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

(Signature) …………………………………………………………………………….

Abstract

The bio-aerosol is an important medium for the potential dispersal of biological warfare agents within the battlefield space. In order to better protect the military personnel who work within this environment it is imperative that we increase our understanding of this matrix, especially the naturally occurring variation and its causes. Understanding the naturally occurring variation within the bio-aerosol will enable future and current biological detection platforms to be put through better test and evaluation processes, thus reducing the potential for false alarms and false negatives. Analysing bio-aerosol samples collected across a temporal gradient through a metagenomics approach will enable the natural variation to be better understood. However, metagenomic analysis tools have been shown to have contradictory reviews within the literature, it is therefore essential to identify the most suitable analysis approach.

Here I developed a metagenomic analysis pipeline which delivers high confidence taxonomic identification to species level, as well as accurate measures of diversity and homogeneity. The analysis pipeline that was developed takes the output from multiple tools thus reducing the number of false positives, delivering high confidence taxonomic identification. The analysis pipeline also gives a more accurate measure of diversity and homogeneity compared to any of the tools being used individually. This improved accuracy will deliver superior results when measuring the change in abundance of species identified within the bio-aerosol in sampling regimes carried out at Dstl. These improvements will lead to more accurate test bio-aerosols being developed for biological detection platform evaluation. Fundamentally this will improve the UK military's capability to detect biological warfare releases within the battlespace.

.

Table of Contents

List of Tables

List of Figures

Abbreviations

DSTL; Defence Science Technology Laboratories,

SARS; Severe Acute Respiratory Syndrome,

DNA; Deoxyribonucleic acid,

PCR; Polymerase chain reaction,

WGS; Whole Genome Shotgun,

VBNC; Viable but not Culturable,

PCR; Polymerase Chain Reaction,

ELISA; Enzyme linked Immunosorbent Assay,

PPE; Personal Protection Equipment,

NGS; Next-Generation Sequencing,

MetAMOS; Metagenomic assembly and analysis,

ABySS; Assembly By Short Sequencing,

SdBG; succinct de Bruijn graphs,

CPU; Central Processing Unit,

SNVs; single-nucleotide variants,

MetaPhlAn2; metagenomic phylogenetic analysis,

BLAST; Basic Local Alignment Search Tool,

NCBI; National Centre for Biotechnology Information,

NT; Nucleotide,

IMG/M; Integrated Microbial Genome/Metagenomes,

ER; Expert review,

LAST; Local Alignment Search Tool,

ML; Maximum Likelihood,

SUPRI; Sequence-based Ultra-Rapid Pathogen Identification,

RAM; Random Access Memory,

GB; Gigabyte,

CLoVR; Cloud Virtual Resource,

NR; Non-redundant,

MEGAN; Metagenomic Analyser,

LCA; Lowest Common Ancestor,

bp; Base Pair,

CLARK; Classifier based on Reduced K-mers, GOTTCHA; Genomic Origins Through Taxonomic Challenge,

BWA; Burrow-Wheeler Aligner,

SAM; Sequence Alignment Map,

LC; Linear Coverage,

SNP; Single Nucleotide Polymorphism,

specI; Species Identification;

PMGs; phylogenetic marker genes,

Phymm; Phylogenetic Markov Model,

IMM; Integrated Markov Model,

TACOA; Taxonomic Composition Analysis,

k-NN; k-nearest neighbour,

GFV; genomic feature vectors,

GC; guanine-cytosine,

BWT; Burrows-Wheeler Transform,

FM; Ferragina-Manzini

HTS; High Throughput Sequencing,

GUI; Graphical User Interface,

IDBA; Iterative De Bruijn Graph De Novo Assembler.

EDGE; Empowering the Development of Genomics Expertise,

CLIMB; CLoud Infrastructure for Microbial Bioinformatics

1 The evaluation of metagenomic analysis software, using *in-silico* and *in-vitro* mock community datasets, for the accurate study of bio-aerosol samples.

1.1 Overview

This thesis describes the process implemented to define a metagenomic analysis approach to evaluate the biological diversity of bio-aerosol samples. A range of sample types, including *in-silico* and *in-vitro* mock community samples, were generated to evaluate a range of metagenomic analysis algorithms. The tools were evaluated based on the true positives (a species identified by an algorithm that was present in the sample), false positives (a species identified by an algorithm that was not present in the sample) and false negatives (a species not identified by an algorithm that was present in the sample) that they identified. The time taken for the tools to run and their computational requirements were also of interest.

The output from this thesis will be directly exploited by a wider Defence Science and Technology Laboratory (Dstl) project looking to develop a greater understanding of the geographical and temporal variation of biological diversity within bio-aerosols. There is a long-term study to take bio-aerosol samples at a single location over the course of at least two years. These samples will be used to measure the variation of the biological content over temporal gradients. Further aerosol samples will be collected at multiple locations across the United Kingdom, Europe and potentially further afield. These samples will enable the measurement of variation between bio-aerosols across a variety of geographical regions. This greater understanding of the temporal and geographical variation of the bio-aerosol will be utilised when testing and evaluating new and existing bio-warfare detection platforms.

1.2 Introduction

1.2.1 Biological aerosols

Biological aerosols (bio-aerosols) are of interest to Dstl, and other military organisations, because a deliberate release of a biological warfare agent is likely to involve the aerosolisation of the agent [1]. As such, the bio-aerosol becomes an essential medium for monitoring any potential release. Bio-aerosols are also of importance to the wider population. There are numerous human health conditions that are impacted by the bio-aerosol, for example fungal spores are linked to asthma and other respiratory health conditions[2]. The bio-aerosol is also the transmission vector of airborne infections such as influenza[3], SARS[4] and tuberculosis[5].

The term bio-aerosol describes the biological content of the air within a specific location and point in time. The biological content includes single cell organisms, pollen, spores, plant and animal debris, metabolic products and extra cellular DNA[6]. The sources of the biological materials include the terrestrial and aquatic environments which can be released into the atmosphere through environmental actions[7], animals[8], and human behaviour[9]. The nature of the constituent parts that make up the bio-aerosol mean that they can travel large distances, via the wind, human movements or animal migrations[10]. Due to the complex factors involved in generating bio-aerosols and their distribution, both locally and globally, there is expected to be a large variation in the diversity of organisms found within the bio-aerosol depending on the time of year, geographical location and weather conditions. Looking for trends in how the diversity of the bio-aerosol changes over time and between locations is important not only for defence but also health professionals and agriculture.

The relatively recent advancements in sequencing techniques, including 16S ribosomal RNA (rRNA) and Whole Genome Shotgun (WGS) sequencing, have enabled the diversity of the bio-aerosols to be better appreciated[11]. Prior to the development of sequencing as a tool for microbial identification, analysis of the microbial load of environmental samples relied solely on direct culturing or polymerase chain reaction (PCR). Direct culturing as a technique has a few disadvantages, mainly around the viability of the organisms collected. Any organism that becomes non-viable during the collection phase will not be represented in the results. Due to the high airflow rates of some collectors, desiccation of the cells is a likely outcome impacting viability. However, any DNA collected would still be present in the sample and would therefore be identifiable as part of the bio-aerosol using a genomics approach. There is also the well-known phenomenon of viable but non culturable (VBNC) species which impact the results from an analysis approach based on culturing the sample[12]. By the year 2017 there were 85 bacterial species identified as being able to enter the VBNC state[13]. These species include *Escherichia coli*, *Listeria monocytogenes* and *Campylobacter sp*p. which have all been identified within bio-aerosol samples collected on this project[14]. Again, sequencing would be able to identify the VBNC species present within the sample as the DNA would be present providing a more accurate measure of the bio-aerosol[15]. The best advantage that culturing offers over a sequencing approach is in regards to true positives. If a species can be isolated through culturing, as long as good laboratory practice is followed, then it is safe to assume it was present in the sample. This cannot be said for the sequencing approach due to the high rates of DNA contamination within biological grade reagents[16]. Alongside the contamination issue there is the potential for high numbers of false positives identified by different analysis algorithms[17]. As an approach for identification PCR does not rely on culturing the cells so viability isn't an issue, however, the primers used to amplify the DNA have to be specific to the organism. Therefore PCR is a targeted approach to identify known organisms rather than a tool to gauge the unknown diversity of a sample.

The purpose of this work was to identify a taxonomic identification pipeline which reports the highest number of true positives with both the lowest number of false negatives and false positives. Identifying a bioinformatics approach to remove reagent contamination is also of interest for future work.

1.2.2 Biological warfare

Biological warfare is the intentional release of a biological agent that can cause death, incapacitation or territory denial. Biological warfare includes the release of wild type organisms or species which have been genetically modified to increase pathogenicity, transmissibility or antibiotic resistance[18]. Biological warfare was first truly investigated as an act of war during the First World War, with many nations having offensive programs[19]. However, there are reports of blankets contaminated with smallpox being given to Native Americans during the siege of Fort Pitt in 1763 and plague victims being catapulted into the sieged city of Caffa in 1346[20]. These reports suggest that biological warfare has been used as an offensive weapon for a much longer period of time.

As previously mentioned, the bio-aerosol would be a primary means for the dissemination of a biological warfare agent. The bio-aerosol route offers many advantages to a bio-terrorist or rogue state as it can be used to transit the biological agent from point of release to the intended victims. This enables the perpetrators to maintain a critical distance from the point of attack. Also, and potentially more importantly, the most pathogenic route of infection for the majority of biological warfare agents is through the respiratory tract[1]. This means an aerosolised agent transported through the bio-aerosol is likely to cause an infection in a large proportion of people exposed to the aerosol. Examples of biological warfare attacks using the bio-aerosol dispersal route include the Amerithrax letters in 2001; where letters laced with *Bacillus anthracis* spores were sent across the USA[21]. On opening the letter, the spores were aerosolised and inhaled by the victims, leading to 5 deaths and 17 infections. Another, less successful, attack took place in 1993 where the Japanese cult Aum Shinrikyo attempted to aerosolise a crude *B. anthracis* culture[22]. Despite successfully releasing the agent the cult had, fortuitously, acquired a vaccine strain of the organism and there were no reported human casualties, although there were reports of some animal deaths in the location.

1.2.3 Metagenomics

Metagenomics is the direct analysis of the genomes contained within a sample without the prior need for cultivating clonal cultures [23]. This direct approach of sequencing the DNA within a sample offers several benefits over the more conventional cultivation and sequencing methods. These include a reduction in the time taken for the process and the ability to identify non-culturable organisms within a sample. Metagenomic studies have been used to identify the organisms found within a variety of different environments, for example the Human Microbiome Project [24], the Yellowstone National Park project [25] and the Pacific Ocean Virome project [26]. Despite the advantages in the metagenomic approach the process can be biased. An example of this bias is during DNA extraction where sporulated, Gram-positive and Gram-negative bacteria will all require different methods for optimum DNA extraction. Nucleotide amplification steps in the sequencing process also risk adding bias to the sample. Additionally, the analysis of the data can add bias depending on not only the algorithms used for the analysis but how the software is used, due to a number of tools containing different parameter settings. This work will investigate if it is possible to use multiple tools and combine their output in an effort to minimise the bias that can be introduced through individual tools.

Metagenomic studies using current next-generation sequencing (NGS) platforms and chemistries have the potential to generate vast quantities of data. For example, a 2 x 125 base pair HiSeq run can generate up to 1 Tera base of data (http://www.illumina.com). These huge datasets have the potential to cause issues in data analysis and also place large burdens on data storage, computing resources and bioinformatics manpower. These burdens can result in the data analysis becoming a major bottleneck for projects. Therefore, speed and ease of use are important criteria to consider when assessing a metagenomics tool; however, the accuracy of the tool must remain the most important criteria upon which to judge the tools. The accuracy of a tool can be measured by its true positive rate (correctly identifying the presence of an organism), false positive rate (incorrectly identifying the presence of an organism not present in the sample) and false negative rate (unsuccessfully identifying an organism which is in the sample). These measures can then be used to calculate the sensitivity and precision of the tool (See section 2.2.4 for a definition).

1.2.4 Biological Detection

There are several methods for the identification of biological material within a bio-aerosol ranging from particle size and spectroscopy techniques[27], genetic detection[28], immunoassays[29], and culturing techniques[30]. Monitoring particle size and spectroscopic techniques offer near real time monitoring but do not offer specificity. At the other end of the spectrum are culturing techniques, which require long periods of time (up to 72 hours incubation) but offer high specificity for culturable organisms. In the middle of the spectrum are the genomic methods (such as sequencing and PCR) and the immunoassays (such as enzyme linked immunosorbent assay (ELISA)). These techniques offer a shorter time to result compared to culturing and are more specific than spectroscopic techniques.

In a military context, the aim of biological detection is to warn of the presence of bio-warfare agents. A detection event would lead to the donning of suitable personal protective equipment (PPE), decontamination of personnel and infrastructure, quarantine and medical surveillance and possible area avoidance. If the bio-detection system fails to alarm in the presence of a release the consequences have the potential to be grave; personnel would be exposed to the agent and would likely be infected. Due to an incubation period of several days and early symptoms presenting as undifferentiated febrile illnesses there is a high chance of the infection spreading between personnel in the local region[31]. There would also be a delay in administering suitable medical counter measures leading to a high chance of fatalities.

Conversely, if a detection platform falsely alarms then there could be dramatic consequences. The wearing of PPE, especially in hot environments, has been shown to increase the chance of potentially fatal heat injury[32]. There is also the financial cost associated with unnecessary decontamination and administering unnecessary medical treatment. Repeated false alarms also risk eroding confidence in the detection device, potentially leading to future alarms being ignored or the equipment not being used. It is therefore essential that every effort is made to ensure that the bio-detection platforms are highly accurate with minimal false positive or false negative alarms. This project will help build our knowledge of bio-aerosols and improve the ability to test the bio-detection platforms in a range of realistic, relevant and controlled bio-aerosols.

Bio-detection is also relevant to the wider community, especially in the field of agriculture where detection systems for plant pathogens are being developed[33]. This will enable more directed use of pesticides, reducing financial costs to the farmer and also be of benefit the environment. Bio-detection in the food animal industry has the potential to reduce the amount of antibiotics used each year. This approach may have a positive effect on the spread of antibiotic resistance genes through the environment, therefore impacting human health positively[34].

1.3 Literature Review

To ensure that the right metagenomic analysis tools were evaluated in this project a review of the literature was performed. A substantial number of tools were available for the analysis of metagenomic datasets, with the majority of the tools designed to perform one specific function in an analysis pipeline (e.g. assembly) rather than being multi-purpose tools. This review will only focus on those tools developed for the purposes of species identification within metagenomics, and therefore does not include tools designed for methods such as meta-transcriptomics. After an initial review of the literature over 70 tools, spanning the last 15 years, were identified as falling within the remit of this review (See Appendix 1 for a list of tools). The 70 tools were down-selected for further review based on several criteria:

- Relevance to the scientific community; only keeping tools which had more than one citation per month since their release.
- Whether the tool has been maintained since its release; only tools which have been actively maintained since their release will be considered for review.
- Whether the tool had been superseded by a newer release; if the tool had been superseded by a newer version the older version was not included.
- Whether the tool keeps the data private or releases to the public; online tools which don't have the option to keep the data private will not be included in this review.

This down-selection process reduced the initial list of over 70 tools to 24. The down-selection process also highlighted the speed at which this area is moving, at the time of writing this literature review one of the tools released in early 2015 was on its third version by the end of the year. Due to the speed at which metagenomic analysis tools are released, continual tech-watch will be required to stay up to date with new tools. It is important to note that tools were only considered for this review if they were released prior to the 31st December 2015.

In the following section of the review a brief description of each of the down-selected tools is provided, accompanied by a reference to the relevant literature regarding the tool. Additionally, the tools have been broadly collated, based upon their function, into two main categories: metagenomics assembly and metagenomics binning. A brief description of each of the categories is also provided, although some of the larger tools (e.g. MetAMOS [35]) can span both categories.

1.4 Metagenomics Binning Tools

Binning is an approach to metagenomic analysis where the sample reads are sorted into groups based on either their homology to a reference database or the composition of the sequences. In this section tools which cover both of these binning processes for assigning taxonomy to sequencing reads (or indeed contigs generated though metagenomics assembly) are discussed.

In homology binning, the reference databases can vary substantially in terms of their composition, from large databases of reference genomes to targeted databases of marker genes. Alignments of the reads against the databases are generated and are subsequently used to ‘bin’ the reads and assign taxonomy. It is important to note that some binning tools are able to use various alignment tools while others can only use one. However, a detailed review of alignment tools is beyond the scope of this current review.

In contrast to homology binning, composition binning looks for patterns within the sequence of the reads and uses these patterns to identify and taxonomically assign the reads. Typically the patterns looked for are in codon utilisation, single-nucleotide variants (SNVs) and k-mer content.

1.4.1 MetaPhlAn2 [36]

MetaPhlAn2 (metagenomic phylogenetic analysis) maps sample reads to a clade specific marker reference database and delivers strain level identification and relative quantitation for prokaryote, eukaryote and virus sample reads. A clade specific marker is a coding sequence that is strongly conserved within the clades genome and has no similarity to a known sequence outside of the clade. MetaPhlAn2 is an advanced version of MetaPhlAn [37] and includes an additional 1 million markers covering an additional 7,500 species to MetaPhlAn. MetaPhlAn2 supports parallelism and also incorporates the alignment tool BowTie2[38] which has increased the speed more than 10 times compared to MetaPhlAn. The MetaPhlAn2 clade specific reference database is greatly reduced in size compared to a full genome reference database and therefore can be mapped to sample reads very efficiently. Sequencing error is unlikely to result in an erroneous match to a marker sequence due to the uniqueness of the marker sequences and their small size. Therefore there is no requirement to perform any pre-processing such as assembly, gene annotation or error detection.

1.4.2 CONCOCT [39]

CONCOCT is an automatic algorithm that uses sequence composition and coverage to cluster contigs into genomes. The initial analysis of CONCOCT, as reported in the literature, used *in-silico* and mock community datasets The tool Ray [40] was used to generate the co-assemblies, with a k-mer length of 41, for this initial analysis. The mock datasets analysed in the paper consisted of ~750 million reads, required compute power of 2,048 cores and took 22.5 hours to assemble. Each of the assembled contigs were cut into 10 kilobase fragments to reduce compute requirements for alignment of reads to. BowTie2 was used to align the sample reads back against the contigs to determine coverage of the contigs per sample. Any contig which has more than 10 hits during the alignment step is labelled as a genuine contig. TAXAassign (https://github.com/umerijaz/TAXAassign) was used to determine the taxonomic classification of the contig; TAXAassign uses BLAST[41] to match against the NCBI NT database (http://www.ncbi.nlm.nih.gov/nucleotide).

1.4.3 IMG/M 4 [42]

IMG/M 4 (Integrated Microbial Genome/Metagenomes) has a number of analysis tools which are only available to registered users. The metagenomic tools available through IMG/M [43] are only accessible after the data has been uploaded to IMG/M ER version (Expert Review). The IMG/M database can be used to evaluate the metagenomes prior to public release on the IMG/M submission site. The paper is unclear who owns the data after it has been uploaded to this site; this raises concerns if sensitive samples are to be analysed. IMG/M accepts assembled or unassembled data and subjects them to a quality step which includes trimming, removal of replicates and masking of low complexity regions. Protein coding genes are detected using *ab initio* gene prediction tools: GeneMark [44], Metagene [45], Prodigal [46] and FragGeneScan[47]. Reads can be 100 – 800 bp in length and can be identified though comparison to the IMG [48, 49] protein database using BLASTX. IMG/M then assigns phylogenetic composition through comparisons of the samples protein coding genes to those in the IMG database and NCBIs RefSeq database.

1.4.4 MetaPhyler [50]

MetaPhyler uses a set of marker genes to form a taxonomic reference for homology-based classification. The marker gene reference is based on 31 protein-coding marker genes previously shown to be suitable for phylogenetic analysis. These marker genes are universal, only present once in each genome and are rarely subject to horizontal gene transfer pressure. MetaPhyler's database only enables identification of reads to genus level rather than species or even strain level. This level of taxonomic classification is not sufficient when identification to species level is required. For example, samples containing *Bacillus cereus* and *Bacillus* anthracis have the potential to lead to very different outcomes due to the health consequences of the organisms. The output of MetaPhyler displays the abundance of different genus within the population by calculating the coverage for each marker gene as a function of the total reads mapped.

1.4.5 PhyloSift [51]

PhyloSift has a standard database made up of 37 “elite” gene families (a subset of 40 elite genes previously reported) representing roughly 1% of the bacterial genome. The “elite” genes are defined as genes which are near universal and present in just single copies. Phylogenetic trees built using these genes individually are generally congruent with each other. In addition to these gene families there are four additional families; 16S and 18S rRNA genes; mitochondrial genes; eukaryote-specific genes; and viral genes. This equates to a set of 800 gene families, mostly viral genes. The PhyloSift work flow is broken into four distinct tasks; sequence identity search, alignment to reference multiple alignments, placement on phylogenetic reference tree and taxonomic summary of read placements. The sequence identity search task uses the LAST algorithm [52] to identify regions of the sample reads with homology to the reference database. PhyloSift uses hmmalign [53] for the alignment to reference multiple alignment tasks. Each of the alignments of the 37 elite gene markers are concatenated into a single row. For placement on a phylogenetic reference tree PhyloSift uses pplacer [54], which can be run in maximum likelihood (ML) or Bayesian mode. ML mode places the sample reads at the single most likely attachment point whereas the Bayesian mode shows all possible attachment points. The last task in the analysis workflow is the generation of a human friendly summary of the phylogenetic placements. Finally, PhyloSift produces Krona plots [55] for both the 37 elite gene families and the 37 elite gene families and additional families.

1.4.6 SURPI [56]

SURPI (Sequence-based Ultra-Rapid Pathogen Identification) utilises SNAP and RAPSearch [57]. In fast mode SURPI can detect viruses and bacteria from samples of 7-500 million reads in 11 minutes to 5 hours using a viral and bacterial database. The speed test detailed in the paper was performed using 64 core, 512 GB RAM, 3 x 4 TB hard drive. In comprehensive mode the sample reads are classified against the entire NCBI database. SURPI has a minimum hardware requirement of 60 GB RAM, 1 GB disk space for the program and 1 TB disk space for the reference data. SNAP is a hash-based nucleotide aligner developed for mapping a wide range of read lengths (50 –

10,000bp) to a reference genome. SNAP was specifically designed for human genome mapping and therefore a custom build was made for aligning to different reference databases. The SURPI pipeline accepts FASTQ files which are then trimmed on quality using cutadapt[58], low complexity sequences removed using DUST [59] and the reads are cropped to 75bp. The processed reads are then aligned to the human database followed by alignment to 29 indexed nucleotide sub-databases (bacterial, fungal, parasitic, other and viral). Matched reads are then taxonomically classified by looking up the GI numbers from NCBI. In comprehensive mode SURPI continues to *de novo* assembly using AbySS [8], AbySS is run several times to increase robustness of the de Bruijn graph based assembly and then RAPSearch translates the nucleotide output for viral or NCBI protein alignment.

1.4.7 CloVR [60]

CloVR (Cloud Virtual Resource) offers push-button automated sequence analysis utilizing either local or cloud computing resource. CloVR has four analysis protocols 1) BLAST search 2) comparative 16S rRNA analysis 3) Metagenomic comparative analysis 4) microbial genome assembly and annotation. The metagenome comparison pipeline takes FASTA input and processes through UCLUST [61] for clustering and replicate removal. Functional classification is performed using BLASTX against COG [62], taxonomic classification utilises BLASTN against RefSeq and Metastats[63] delivers comparative analysis and outputs summary reports and figures.

1.4.8 CARMA3 [64]

CARMA3 is a taxonomic classification method that can be used with assembled or unassembled sequences and has been developed to work with both BLAST and HMMER3 [53] homology searches. The first step for the BLAST component of the program is to use BLASTX for homologues to the metagenomic sequences in the NCBI NR database. After this initial search is complete all reads without taxonomic assignment or that have unclassified/other as the output are discarded from the process. The following step is to take all the hits that were generated in the first BLASTX search and generate a new database. This database is then searched against and the output is ranked with likely taxonomic classification enabling LCA (lowest common ancestor) classification. The main difference when using the HMMER3 alternative approach is that the metagenome sequence is translated into all six possible protein sequences before being searched against the pfam database [65]. The process then mirrors the BLAST component of the program but there is the added benefit of using the HMMER3 component which is the ability for functional classification of the reads. CARMA3 is not a fast program; it was reported in the paper that to run the complete CARMA3 pipeline on 10,000 reads using a dual core 8 GB RAM of computing resource took over 55 hours for the BLAST component and over 7 hours for the HMMER3 component [18].

1.4.9 MG-RAST [66]

MG-RAST is an online, open source environment where metagenomic sequence data is automatically analysed and annotated. MG-RAST accepts raw sequence data or assembled contigs and runs a prioritisation system where projects which allow their datasets to be made available to the public with metadata are processed with high priority and projects which keep their data private have a lower priority. The uploaded data is password protected and there is the option to give permission for the data to be accessible to colleagues or the MG-RAST user community. The first data manipulation step is to normalise the data by labelling the reads with unique IDs and removing duplicate reads. The normalised data is then screened for protein encoding genes using BLASTX against the SEED comprehensive non-redundant database [67]. At the same time the data is compared to accessory databases including rDNA databases such as GREENGENES [68], RDP-II [69], European 16S rRNA database [70], chloroplast database, mitochondrial database and ACLAME database [71] of mobile elements. The final data manipulation step is to calculate the taxonomic distribution and functional assignment of the sample. Taxonomic distribution is calculated using the phylogenetic information from the SEED database and the similarities to the 16S database. These results are shown in tabular format which can be mined to show specific taxonomic groupings. There is also a comparative metagenomic function which enables taxonomic variation between metagenome datasets to be compared against each other. The results are shown in a tabular heatmap, highlighting the differences.

1.4.10 CLARK [72]

The CLARK (CLAssifier based on Reduced K-mers) algorithm develops a reference index which contains the unique k-mers (k-mer = 31-mer) for each target/reference organism. The k-mers that appear in more than one target genome are removed. The target specific k-mers are put into a “dictionary” resulting in a discriminative database which unknown reads can be referenced against. Each unknown read is queried against the whole “dictionary” and is assigned to the target with the most hits using exact matching. Advantages of CLARK include the ability to remove k-mers from the unknown sample based on their abundance which enables the reduction of sequence error affecting the results. CLARK is also able to report confidence scores for each of the k-mer hits. In addition to the standard program, there are two variants of CLARK:

CLARK-l (a light version of CLARK) uses a k-mer length of 27 and skips four consecutive 27-mers which delivers a reduced compute requirement <4 GB RAM that is designed for laptop computing. CLARK-l, reportedly, still achieves high precision and high speed although lower precision when compared to CLARK.

CLARK-E (an express version of CLARK) is also available and uses a much smaller reference index which only uses non-overlapping k-mers and only queries a sample of the unknown reads and assigns the read to the first target k-mer that it hits.

1.4.11 GOTTCHA [73]

GOTTCHA (Genomic Origins Through Taxonomic CHAllenge) has a number of unique reference genome databases to differing levels of taxonomic levels. To generate the databases shared 24-mer sequences were removed from chromosomal and plasmid replicons. There are databases for bacteria at Class, Family, Genus, Order, Phylum, Species and Strain level and also viral databases at Genus, Species and Strain level. These databases are also available with all human 24-mers, which are derived from 3 human genomes, removed. GOTTCHA analysis takes reads trimmed on quality and then fragmented into non overlapping 30-mers. The short read aligner BWA[74] then aligns the fragmented sample reads to the chosen GOTTCHA reference database using the exact match option. The SAM alignment files are then profiled and filtered with the GOTTCHA profiler. GOTTCHA's primary classification parameter is the Linear Coverage (LC) defined as the percentage of the unique genome covered during the mapping stage. The LC must be greater than 0.5%. The ability of GOTTCHA to identify novel genomes was investigated; 2000 draft genomes were examined at a range of taxonomic levels and 92% of novel strains were correctly identified to the parent taxa. GOTTCHA reports a higher memory requirement and performs slower than other classifiers tested [22].

1.4.12 One Codex [75]

One Codex is a web-based platform using a k-mer based taxonomic classification algorithm. It contains two reference databases, the OneCodex database and an NCBI RefSeq database. The OneCodex database covers 40,000 bacterial, viral, fungal and protozoan genomes; the RefSeq database covers 8,000 microbial genomes. They are both developed using 31-mers, each read is broken down into overlapping 31-mers and then aligned to the 31-mers. The results are summarised as a "k-mer hit chain" showing the lowest taxonomic clade each read can be identified as. It is possible to review filtered and unfiltered output from the OneCodex database. The filtered output highlights high confidence results whereas the unfiltered results are all of the taxonomic identifications. One Codex is run as an open access program for research involving public data. Research institutes are limited to 50 runs per researcher and industrial users have to pay for access to the platform.

1.4.13 Kraken [76]

Kraken is described as an ultrafast tool delivering genus level sensitivity and precision comparable to the fastest BLAST program megaBLAST. BLAST was reported as one of the best tools for metagenomic alignment in 2009[77]. Kraken uses a database containing k-mer and lowest common ancestor (LCA) for all organisms for whose genome contains that k-mer. The Kraken database is built using a user defined set of genomes and the default setting for k is 31 which can be modified by the user. When classifying a sample each read is broken down into its k-mers and similar k-mers are grouped together for faster searching which is helped by using CPU cache rather than RAM.

Mini-Kraken utilises a 4 GB database, reducing the time and compute power required to search against it. To generate the reduced database size, 18 of every 19 k-mer records was removed. This reduction factor was chosen to reduce the database to below 4 GB so Mini-Kraken can be used on small personal computers. As reported in the literature, Kraken and Mini-Kraken offer comparable levels of precision (97.9% and 98.9% respectively), however, the sensitivity of Mini-Kraken is 66.6% compared to a sensitivity of 80.6% for Kraken[76].

1.4.14 ConStrains [78]

ConStrains (Conspecific Strains) has been developed to identify conspecific biological and archaeal strains using SNP patterns in a set of universal genes. ConStrain requires reference species to compare raw metagenomic reads to for SNP pattern identification. The ConStrain algorithm is broken into two operations; identification of species by which SNPs are detected followed by transforming individual SNPs into SNP profiles representing individual strains. MetaPhlAn is used for species composition profiling of the raw reads (ConStrains requires >10x coverage for each species) and a custom database is then built using the PhyloPhlAn marker set which the raw reads are mapped to using BowTie2. SAMtools [79] is subsequently used to generate a table of coverage by base position for SNP identification. ConStrains uses the identified SNPs to generate a ‘uniGcode’ which identifies SNPs spanning hundreds of genes. SNP-flow and SNP-type algorithms are used to create strain combination models and the relative abundance of each strain is estimated for each model. This step is repeated for each species with the required coverage. This approach was reported to deliver accurate results on *in-silico* datasets based on 36 strains of *E. coli* and also matched manual strain identification techniques on infant gut metagenome dataset[79].

1.4.15 Anvi'o [80]

Anvi’o is an analysis and visualisation application for ‘omics data and is available through command line or a graphical web-browser. Anvi’o is an assembly based metagenomic workflow which utilises human guided and automated sample binning, visualisation and reporting. Anvi’o generates community contigs from all or a subset of sample reads, and this contig database is used to store k-mer frequencies, functional annotation and GC content. The contig database is then mapped to each of the individual sample reads and properties such as mean coverage, for each contig in the read generated. The results from this alignment step are put through the CONCOCT [39] tool for binning of contigs into draft genomes. A limitation of Anvi’o is the high level of compute required for certain steps within the pipeline, cluster nodes of up to 512 GB memory were used for the analysis described in the paper.

1.4.16 TETRA [81]

TETRA takes DNA sequence reads and calculates the frequencies of each of the 256 possible tetranucleotides (a tetranucleotide is a 4-mer, a 4bp section of DNA). Using a maximal-order Markov model the expected frequencies are calculated based on the composition of bi- and tri- nucleotides. Divergence from the expected tetranucleotide frequency is converted into a z score; these z scores are compared in pairs using the Pearson’s correlation to calculate the relationship between the pairs. TETRA is available as a web-service where raw data is uploaded and the results are emailed when complete but is also available as a standalone program which has the benefit of the raw data being accessible to the user. TETRA is able to deliver full classification for reads in the region of 40 kb. However, it will struggle with reads in the region of 1 kb because the phylogenetic signal within tetranucleotide usage patterns is weak. Tetranucleotide usage patterns should also not be used to deduce phylogenetic relationships but rather as a fingerprinting technique and therefore it is unlikely to give high levels of resolution.

1.4.17 specI [82]

specI (species Identification tool) is based on pair wise average nucleotide identity of 40 protein coding phylogenetic marker genes (PMGs). The 40 PMGs were annotated for 3,496 prokaryotic genomes using SMASH routines [83] a tool for estimating and annotating metagenome phylogenetic and functional composition, and the sequences were obtained using the eggNOG database [84]. The pairwise distance for the PMGs was calculated using glsearch [85] to generate distance matrices. Clustering is performed using hcluster (https://code.google.com/p/scipy-cluster/) with the distance matrices using single, average and complete linkage. These clusterings were then transformed into discrete species level clusters. Using this technique, greater than 96% of assignments for 2,804 genomes from NCBI were correct.

1.4.18 PhymmBL expanded [86]

Phymm [77] (Phylogenetic Markov Models) was developed as a metagenomic phylogenetic classifier specifically for short reads. It has been shown to be accurate on reads as short as 100bp. Phymm was the first example of a phylogenetic classification tool that used interpolated Markov models (IMMs). IMMs are used to characterise oligonucleotides of different lengths into phylogenetic groups. In developing Phymm an IMM was generated for each of the genomes in the RefSeq database (1,146 at the time). It was also shown that utilising Phymm to classify the reads prior to analysis using BLAST improved the accuracy of using either Phymm or BLAST independently, this pipeline is referred to as PhymmBL. PhymmBL was expanded in 2011 and the output now includes confidence scores on the accuracy of the predictions. The PhymmBL database can now also be modified to include any amount of custom genomes (including eukaryotic and viral sequences), removing the reliance on RefSeq bacterial and archaeal genomes.

1.4.19 TACOA [87]

TACOA (TAxonomic COmposition Analysis) is a supervised taxonomic classification method utilising the k-nearest neighbour (k-NN) approach with a smoother kernel. The k-NN approach classifies unknown reads to the nearest known read, but there are limitations associated with this approach regarding high-dimensional input data. The smoother kernel reduces this issue by utilising the whole reference data during classification rather than a close neighbourhood. TACOA computes genomic feature vectors (GFV) for each genome in the reference database and for the reads to be classified. The GFVs are calculated as a ratio of observed oligonucleotide frequency within a genomic fragment compared to the expected frequency of the oligonucleotide given by the GC content. Reads with similar GFVs are classified together at different taxonomic ranks to genus level.

1.4.20 Centrifuge [88]

Centrifuge is a metagenomic classifier that aligns the sequencing reads to an index. The alignments are performed using a modified indexing scheme based on the Burrows-Wheeler transform[89] (BWT) and the Ferragina-Manzini[90] (FM) index. BWT and FM were developed to enable fast and low memory alignments. This approach enables a rapid identification of reads using a desktop computer. Centrifuge further reduced the size of the FM index by removing a large amount of repeated genomic information between highly similar bacterial strains and species. For example, the total sequence size for the 131 strains of *Salmonella enterica* was reduced from 661 to 74 Mega base pairs. This fast identification and small index size enables the tool to process over 500,000 reads per minute compared to MegaBLAST which processed 327 reads per minute. This tool falls outside of the initial selection criteria, however, due to the abundance data that it generates it was included into the study.

1.5 Metagenomic Assembly Tools

First generation sequencing techniques, such as Sanger sequencing, deliver longer sequencing read lengths compared to typical NGS platforms, such as the Illumina MiSeq and IonTorrent PGM. These newer platforms typically deliver a lower cost per base and produce a larger number of shorter reads than the original sequencing technologies. The longer read lengths of the first generation platforms and smaller reference databases in the past meant that aligning the reads to the database was a suitable approach for read identification. With orders of magnitude more genomes in the databases and substantially larger amounts of data generated with NGS, aligning the reads within a metagenomics sample to these large databases is computationally expensive and time consuming. The shorter read lengths generated by NGS platforms also means that aligning reads directly to databases can generate ambiguous taxonomic labelling. One potential mechanism to overcome these problems, prior to attempting to assign a read to a particular taxonomic group, is to assemble the reads into longer regions called contigs. These longer contigs can subsequently be used in the identification and taxonomic assignment process. The tools discussed in this part of the review all fall into this category and perform metagenomic assembly.

The assembly tools described below use De Bruijn graphs (dBG) in order to assemble contigs from the raw reads. A dBG works by dividing the sequencing reads into smaller, overlapping sections of a pre-defined length called k-mers. Figure 1 shows how four sequencing reads are broken down into k-mers of length 3 (3-mers); realistically the k-mer length is longer and generally over 20[91]. The dBG is then built with the k-mers representing nodes within the graph where the edges connect overlapping nodes. Sequencing errors, shown in red, may be ignored by the assembly tool as there will be fewer reads containing the error so any branches within the graph containing low frequency k-mers can be ignored. The overlapping nodes within the graph are then combined to generate the contigs[92].

Using dBG to assemble metagenomic datasets is made particularly difficult due to four main factors: sequencing errors, repeat regions, strain variants and varied depth of coverage[93]. Sequencing errors generate false k-mers which increase the size and complexity of the graph. This can lead to errors in the assembled contigs. Repeat regions generate large numbers of edges within the graph increasing the number of routes across it. Increasing the length of the k-mers will reduce the chance of error but does require a high level of compute resource. The fact that related strains are genetically similar means there will be a high number of similar nodes within the graph. Many of these nodes will be ambiguous leading to fragmented contigs. It is also challenging to differentiate between polymorphisms due to strain variants and sequencing error. Finally, the depth of sequencing coverage may impact the contigs assembled using a dBG approach. Low abundant species will have a lower depth of coverage compared to a highly abundant species. This may lead to some branches within the graph being erroneously deleted due to the low abundance of their nodes. The higher abundance nodes may also be misinterpreted as highly repeated regions leading to errors in the assembly.

For this study, the contigs generated by the assembly tools were aligned against the NR database using DIAMOND[94]. MEGAN5[95] was then used to assign taxonomic classification to the contigs that were generated by the de novo assembly tools discussed in this section. Using one standardised tool to assign taxonomic classification to all aligned contigs enabled the tools to be fairly compared against each other.

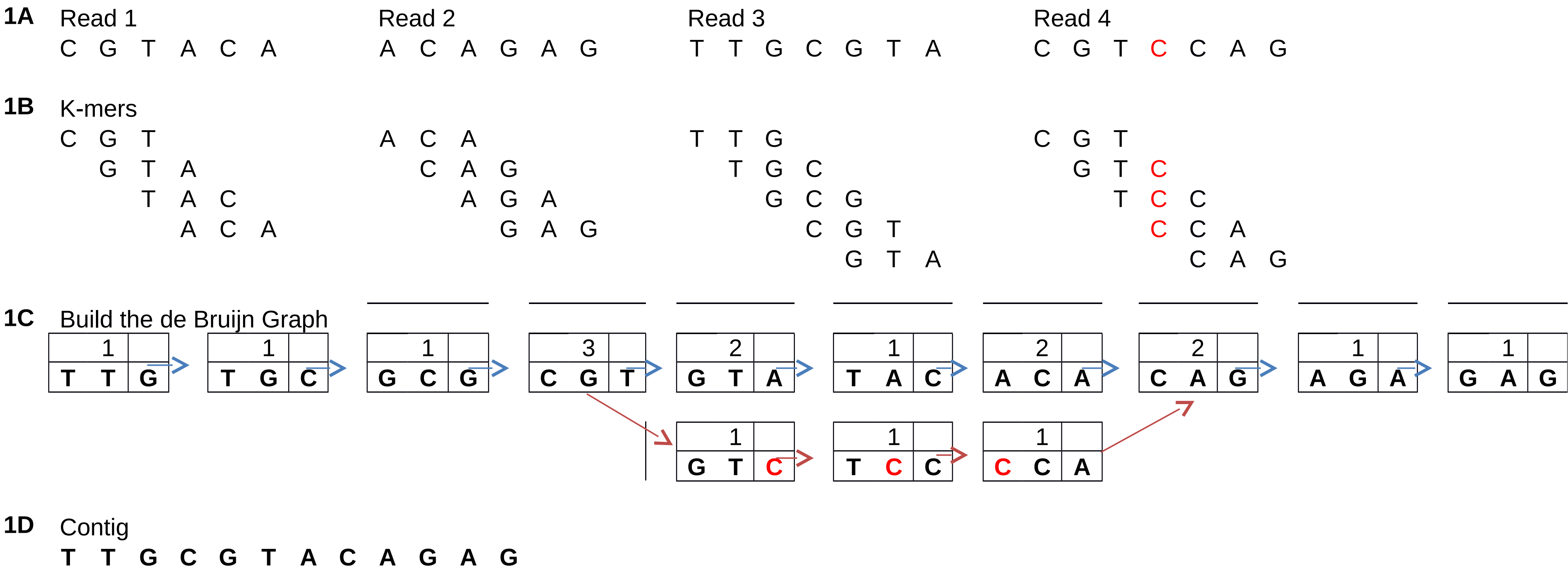


Figure 1: A simplistic representation of the generation of contigs from raw sequencing reads using de Bruijn Graphs, the process utilised by many de novo assembly tools. **1A** shows four short sequencing reads, read four has a sequencing error (denoted with red font); **1B** identifies all possible k-mers of 3 bases for the four reads; **1C** shows the built de Bruijn Graph, the nodes of the dBG represent the k-mers with their frequency denoted, each node shares an edge with a node with overlapping k-mers, due to the low frequency of the divergent branch it may be disregarded; **1D** shows the final contig.

1.5.1 MetAMOS [35]

MetAMOS (Metagenomic assembly and analysis) is a modular and customisable pipeline claiming to deliver a push button solution for metagenomic analysis. The entire pipeline is built around Bambus 2.0 [96], a metagenomic scaffolder but also supports a large number of tools, including eight assemblers, six contig annotation methods, three gene prediction tools, one abundance estimation method and an interactive tool for visualisation of taxonomic composition. The workflow can be broken down into three main sections of pre-processing, contig scaffolding and contig analysis. MetAMOS produces a report summarising the results from the sequencing run.

1.5.2 MetaVelvet-SL [97]

Typical single genome de novo assemblers are not capable of assembling multiple genomes from mixed sequence reads. MetaVelvet [98] is a metagenomic assembly tool that was developed from Velvet [99] (a single genome assembler) and utilises the de Bruijn graph approach. However, MetaVelvet shows low accuracy and sensitivity identifying chimeric nodes. Regarding metagenomic de novo assembly a chimeric node is a node within a dBG which is shared between different species. This could be due to orthologous sequences from closely related species, repeat sequences from a single species, conserved sequences or gene transfer.

To combat this issue MetaVelvet-SL has been developed which utilises supervised machine learning to classify every node as chimeric or non-chimeric in a de Bruijn graph. Metavelvet-SL consists of two modules: the supervised learning module which is required to develop the model for the classification of nodes, and secondly the assembly module. The taxonomic profile for the sample can either be inferred from taxonomic profiling methods (a pipeline is available connecting the MetaPhlAn [37] profiling tool and MetaVelvet-SL), estimated using prior knowledge of the target community or MetaVelvet-SL is supplied with a library of pre-trained classification models for different environments (e.g. soil, human blood and deep sea).

1.5.3 IDBA-UD [100]

IDBA-UD is a de novo assembler and is a further development to the IDBA and Meta-IDBA programs [101]. It has been specifically developed for assembling samples which have a very uneven depth of coverage across the genome. Three problems were identified for other de novo assembly tools using de Bruijn graphs: 1) sequencing errors generating incorrect k-mers (a k-mer is defined as a piece of DNA k nucleotides in length); 2) missing k-mers in low coverage areas; 3) increased branches in the de Bruijn graph due to small k-mers. IDBA-UD has been developed utilising a novel approach in the generation of the de Bruijn graph to overcome these problems, which are exacerbated with metagenomic datasets. To resolve problems 1 and 3, IDBA-UD uses variable thresholds for length k to enable the longest value for k to be used across the dataset. This reduces the number of branches generated but also enables the shortest k to reduce the chance of incorrect k-mers. To resolve problem 2 of missing k-mers, IBDA-UD uses paired end reads to form local assembly areas.

1.5.4 ABySS [102]

ABySS (Assembly By Short Sequencing) is a *de novo* assembler specifically developed for short reads and uses a unique version of de Bruijn graph theory. The algorithm is split into two parts where the first stage generates all possible k-mers for the sequence reads. The k-mers are processed with read errors removed and contigs are built. The second stage of the process then extends the length of the contigs using paired-end reads.

1.5.5 MEGAHIT [103]

MEGAHIT is a *de novo* assembly tool which uses succinct de Bruijn graphs (SdBG). SdBG offer advantages over de Bruijn graphs, but their construction is non-trivial. In order to overcome these challenges MEGAHIT exploits the parallelism of a graphics processing unit. This approach speeds the construction of the SdBG by up to 5 times compared to only using CPU. Sequencing error also has a large impact on MEGAHIT, so to reduce this k-mers that only appear once are disregarded prior to constructing the SdBG. This approach has the potential to lead to k-mers from low abundance organisms not being built into contigs. The results suggest that MEGAHIT generates longer and more accurate contigs compared to some of its contemporary assembly tools.

1.5.6 MEGAN [95]

MEGAN (MEtaGenomic ANalyzer) is a computer program designed to analyse metagenomic data with an unrestricted choice of alignment tools and databases. The MEGAN pipeline is now on version 5 with version 6 in beta test stage. MEGAN's pipeline starts with the comparison of raw sequence data or assembled contigs to one or more reference databases (e.g. NCBI NT) using a comparison tool (e.g. BLAST). The results from the comparison step are then collated and each hit has a taxon ID assigned to it based on NCBI taxonomy. A Lowest Common Ancestor (LCA) algorithm is then used to produce a summary of taxonomy. The LCA algorithm takes all of the matches for each read from the comparison to reference databases stage and calculates the lowest common ancestor for each of the hits. The output from the MEGAN pipeline can be displayed at any taxonomic level down to species and strain level. MEGAN is able to accurately assign reads as short as 35bp but reads this short can be aligned to a number of divergent genomes and therefore results in higher level taxonomic classification. It is suggested that reads of at least 200bp are used for accurate classification to species or strain level.

1.6 Work plan following the outcome of the literature review.

The tools detailed above highlight the different approaches available for metagenomic analysis. From the initial assessment of over 70 tools, 24 were down-selected for further review. Based on the literature review, future work was undertaken to gain a better understanding of how each of the down-selected tools works with data of interest. Attempts were made to download and install the tools described above. However, some were unavailable for download and some were unable to be installed correctly. The tools which were correctly installed were evaluated initially using simple *in-silico* datasets and the results from this initial evaluation enabled a further round of down-selection. The tools which performed best on the simple dataset were further evaluated using datasets of increasing complexity, culminating in the analysis of real-life data.

Evaluating the ability of each tool to successfully analyse metagenomic data will be based on the following metrics:

- The taxonomic level of identity; as previously stated it is essential that the tool is able to identify the genomes to species level.
- The accuracy and sensitivity of the tool will be defined in the following terms: true positive, false positives and false negatives.

The *in-silico* data were generated using MetaSim [104]. MetaSim is a tool designed to simulate genomic data using adaptable error models to represent real sequencing data. This approach enabled the comparison of tools using a fully characterised dataset. The initial datasets were designed to simulate Illumina MiSeq runs using a suitable error model. The datasets were made of eight and ten genomes, each with equal coverage. The Illumina MiSeq platform is a widely used sequencing platform so simulating the output based on this platform increases the utility of the *in-silico* datasets. To increase the complexity of the *in-silico* datasets the number of genomes was increased to 100 and the level of abundance manipulated for each genome. The 10 and 100 genome datasets mirrored real WGS data from bio-aerosol studies. The simulated datasets enabled evaluation of the tools without sample preparation bias affecting the results. Further analysis of the down-selected tools, using real WGS datasets, enabled the tools to be analysed on sequence data of known content. The culmination of the study was to evaluate bio-aerosol samples collected, over long and short term temporal studies, at Dstl.

2 Construction and analysis of two simple *in-silico* datasets using binning and assembly metagenomics tools.

2.1 Introduction

In order to understand the accuracy of different metagenomic analysis tools it is imperative to test them against known datasets. Each tool evaluated on this project has published data indicating the accuracy of its output; however some of these reports are contradictory. For example, Kraken is described as having high precision at the genus level (>90%) [76] but also as generating large numbers of false positives at the species level (> 4,000) [105]. It is therefore important to independently verify these reports with well defined datasets.

Sequencing a defined *in-vitro* mock community metagenomic sample could generate errors in the sequencing and therefore bias the analysis output. These errors could be caused by sequencing biases as well as unknown contaminants potentially being introduced during the sequencing process. These potential errors would reduce confidence that the sequencing output was a true representation of the sample and would therefore reduce the confidence of any results derived from their analysis. In order to overcome these issues, it was decided to build *in-silico* datasets, ensuring the genomic content of the sample waere well defined. The software package used for this process was MetaSim[104]. MetaSim was developed at Tubingen University to generate simulated metagenomic datasets for the benchmarking of metagenomic analysis software.

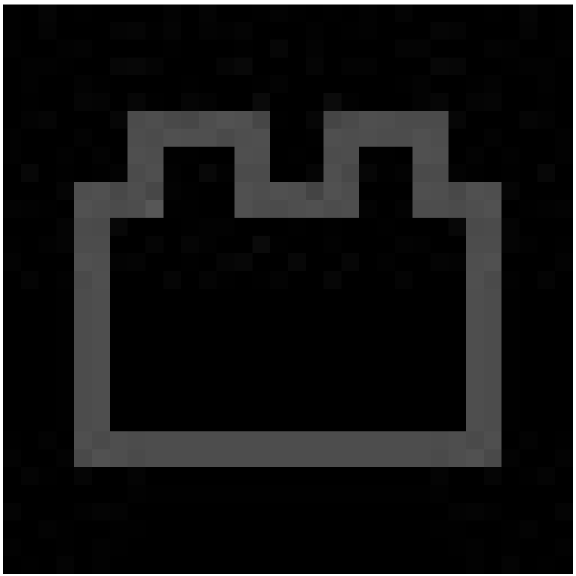
In order to build a dataset that represented a true aerosol metagenome it was decided to identify an aerosol sample from the open literature. Due to a lack of more suitable published datasets, Singapore Air Sample 2 published on Integrated Microbial Genomes and Microbes (https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=MetaDetail&page=metaDetail&taxon_oid=2003000007) was used. Ten organisms were selected to cover a range of GC content and include Gram-positive and Gram-negative organisms as these factors are considered to impact the results from sequencing[106]. The 10 organisms selected were highly abundant organisms from Singapore air sample 2. The organisms selected cover 7 genera to allow evaluation of the ability of the tested tools to differentiate between species from the same taxonomic lineage. This dataset was called AM_10G_10M.

The second dataset, termed Zymo_8G_10M, contained the eight bacterial organisms used by ZymoBiomics in their Community Standard (cat # D6300). Dstl have used the ZymoBiomics community standard as a positive control sample to analyse a range of sequencing platforms, DNA extraction methodologies and other aspects of the sequencing pipeline. It was therefore a logical step to build an *in-silico* dataset based on this well-utilised sample. The eight bacterial organisms cover a range of GC content from one third to two thirds and are also a mix of Gram-positive and Gram-negative species.

2.2 Methods

2.2.1 *In-silico* simulation of short read datasets for evaluation of metagenomic analysis tools.

MetaSim has a Graphical User Interface (GUI) which enables the user to select genomes of interest from its database in order to generate bespoke *in-silico* metagenomic samples. There is the option to alter the abundance of the selected genomes and also to simulate sequencing errors (indels and substitutions) to more accurately represent output from a sequencing platform. MetaSim was used to generate two *in-silico* datasets, one based on a real aerosol dataset and one based on a community standard often used at Dstl. Each of the datasets used the error profile for an Illumina MiSeq sequencing run with a read length of 300bp and a total of 10 million reads. Figure 2 shows the chance of error for each location of the read, highlighting the more error prone 5’ end of the read. For these simple *in-silico* datasets each of the datasets were designed with an even abundance.



Object 3

Figure 2: A representation of the rate of error assigned to each position of the 300bp error profile used with MetaSim to simulate the AM_10G_10M and Zymo_8G_10M *in-silico* datasets.

AM_10G_10M	**Gram +/-**	**GC content (%)**	**Reads**	**Genome size (Mb)**
Pseudomonas aeruginosa PAO1	-	66.6	1383193	6.26
Pseudomonas putida KT2440	-	61.5	1366898	6.18
Brevundimonas subvibrioides ATCC 15264	-	68.4	761685	3.45
Bacillus thuringiensis serovar konkukian str. 97-27	+	35.4	1174778	5.32
Bacillus cereus ATCC 14579	+	35.2	1200992	5.43
Bacillus anthracis str. Ames	+	35.4	1155395	5.23
Staphylococcus aureus subsp. aureus NCTC 8325	+	32.9	623083	2.82
Mycobacterium tuberculosis H37Rv	-	65.6	976654	4.41
Acholeplasma laidlawii PG-8A	-	31.9	330789	1.5
Escherichia coli str. K-12 substr. MG1655	-	50.8	1026533	4.64

Table 1: Strain level information for the 10 organisms used to construct the *in-silico* dataset AM_10G_10M using the metagenomic simulation package MetaSim.

Zymo_8G_10M	**Gram +/-**	**GC content (%)**	**Reads**	**Genome Size (Mb)**
Bacillus subtilis subsp. subtilis str. 168	+	43.5	1345867	4.22
Listeria monocytogenes EGD-e	+	38	940866	2.94
Staphylococcus aureus subsp. aureus NCTC 8325	+	32.8	901277	2.82
Enterococcus faecalis V583	+	37.3	1073728	3.37
Lactobacillus fermentum IFO 3956	+	51.5	672433	2.1
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	-	52.2	1581336	4.95
Escherichia coli str. K-12 substr. MG1655	-	50.8	1483062	4.64
Pseudomonas aeruginosa PAO1	-	66.6	2001431	6.26

Table 2: Strain level information for the 8 organisms used to construct the *in-silico* dataset Zymo_8G_10M using the metagenomic simulation package MetaSim.

2.2.2 Hosting, installing and running the down-selected binning tools

Both AM_10G_10M and Zymo_8G_10M were analysed using seven binning tools, listed below in Table 3. These tools were either hosted locally on a Dell XPS13 7 Gb RAM laptop, on EDGE (Empowering the Development of Genomics Expertise) hosted on the CLIMB (CLoud Infrastructure for Microbial Bioinformatics) network [107] or were on-line tools. EDGE is a software package which enables a number of pre-installed sequencing analysis tools to be run using a GUI. The metagenomic taxonomic analysis tools available through EDGE are GOTTCHA, MetaPhlAn, kraken-mini and BWA. CLIMB is a shared cloud compute resource developed for the sharing of bioinformatics pipelines and sequencing data. Due to the ease of running MetaPhlAn through EDGE it was decided to use it to analyse the initial datasets. This enabled a comparison of the output between a tool and the more up to date replacement.

Binning Tool	**Hosted**	**Command**
MetaPhlAn2	Locally	MetaPhlAn2.py sample.fastq --mpa_pkl /db_v20/mpa_v20_m200.pkl --bowtie2db

		/db_v20/mpa_v20_m200 --input_type fastq > sample.txt
MetaPhlAn	EDGE	Run using default parameters
OneCodex	On-line tool	Run using default parameters and results taken from the filtered and unfiltered output
Kraken-mini	Locally	Kraken –db /minikraken_20141208 sample.fastq –threads 4 > sample.kraken Kraken-report –db/minikraken_20141208/ sample.kraken
	EDGE	Run using default parameters
GOTTCHA	Locally	Bin/gottcha.pl –threads 4 –mode all –input sample.fastq –database /GOTTCHA_BACTERIA_c4937_k24_u30
	EDGE	Run using default parameters
Phylosift	Locally	Phylosift all –output=results-sample sample.fastq
Constrains	Locally	Constrains –c sample.conf –o sample-output –t 4
MG-RAST	On-line tool	Run using default parameters

Table 3: Information pertaining to the location the different binning metagenomic analysis tools were hosted and the commands required for their use.

2.2.3 Hosting, installing and running the down-selected assembly Tools

The assembly tools listed in Table 4 were used to analyse AM_10G_10M and Zymo_8G_10M. The assembly tools were hosted on the CLIMB cloud compute infrastructure. The contigs generated by each of the assembly tools were then aligned using DIAMOND [94] against the NCBI non redundant (nr) database. In order to generate true positive, false positive and false negative results taxonomic assignment was performed with MEGAN5 [95] using the default parameters. Due to the read lengths that had been simulated, IDBA-UD needed to have the source code modified; the constant kMaxShortSequence was increased from 100 to 300 to enable 300 bp reads to be processed.

Assembly Tool	**Hosted**	**Command**
Abyss	CLIMB	ABYSS –k ‘*k-mer_length’* –o *‘output_location’* ‘*read_location’*

IDBA-UD	CLIMB	idba_ud –r *‘read_location’* --num_threads 4 –o *‘output_location’*
CLARK	CLIMB	clark-l –k *‘k-mer_length’* –T *‘target-definition_location’* –D *‘database_location’* –O *‘read_location’* –R *‘output_location’* –g *‘gap_length’*
MegaHit	CLIMB	Megahit –r *‘read_location’* –m *‘mode’* –o *‘output_location’*
Diamond	Locally	./diamond blastx -d nr -q *‘read_location’* -o *‘output_location’*

Table 4: Information pertaining to the location the different assembly metagenomic analysis tools were hosted and the commands required for their use.

2.2.4 Results interpretation

The results from each analysis were reported as true positive, false positive and false negative taxonomic identifications.

The true positive, false positive and false negative results enabled the sensitivity and precision of the tools to be calculated[108]. Sensitivity is a measure of a tool’s ability to correctly identify that the organism is truly present in the sample; a tool with 100% sensitivity will return zero false negatives. Precision defines the proportion of true positives compared to false positives; a tool with 100% precision will have zero false positives.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \times 100$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \times 100$$

A tool with high sensitivity will have a low number of false negatives, giving confidence that organisms present in a sample will be identified. However, sensitivity does not take into account false positives. A tool with high precision will have a low number of false positives giving confidence that the organisms identified are present in the sample. As precision does not take false negatives into account there is potential for species present in the sample not to be identified by a tool with high precision. It is therefore important that tools are measured and shown to perform well against both metrics of precision and sensitivity.

2.3 Results and Discussion

2.3.1 Results from the analysis of simple *in-silico* datasets using binning tools.

Identification at the genus level was very accurate for the binning tools tested, all but one tool showed 100% sensitivity to both datasets. GOTTCHA, hosted locally, was the only tool unable to identify all genus present in the two datasets. MetaPhlAn2, MetaPhlAn, GOTTCHA and ConStrain reported zero false positives for either of the datasets. Conversely (the number of false positives shown in brackets refer to AM_10G_10M and Zymo_8G_10M respectively), One Codex (235 and 170), Kraken-mini (65 and 77), Kraken-mini (EDGE) (68 and 74), Phylosift (1278 and 1106) and MG-RAST (742) showed high levels of false positives see Figure 3A and 3B. MG-RAST reported high levels of false positives for AM_10G_10M, and took several weeks for the program to return the results. Due to the length of time to get the results it was decided that this tool is not worth progressing with so the Zymo_8G_10M dataset was not processed.

The ability of the binning tools to correctly identify the organisms to the species level in the two *in-silico* datasets varied across the seven tools tested. MetaPhlAn2, One Codex, Kraken and GOTTCHA (EDGE) were able to correctly identify all organisms at the species level for both datasets. As well as correctly identifying all species MetaPhlAn2 and GOTTCHA reported zero false positives. Kraken (128 and 154), Kraken (EDGE) (175 and 209), One Codex (810 and 528) and Phylosift (3264 and 2787) reported higher levels of false positives for AM_10G_10M and Zymo_8G_10M respectively (Figure 3A and 3B).

Of the seven binning tools tested only One Codex and Kraken were able to identify all organisms to strain level across the two datasets. The only other tools able to detect to strain level were MetaPhlAn2 (3 and 0), GOTTCHA (hosted locally) (4 and 5), Phylosift (6 and 6) and ConStrains (3 and 0). The figures in brackets refer to the number of strains correctly identified for AM_10G_10M and Zymo_8G_10M datasets respectively. The strain level classification generated the highest level of false positives; One Codex returned

1992 and 2514, Kraken returned 418 and 475, Kraken (EDGE) 322 and 406 and Phylosift had 3456 and 3088 false positives for AM_10G_10M and Zymo_8G_10M respectively (Figure 3A and 3B).

Of note were the differences in output from the same tool hosted on different compute platforms. GOTTCHA hosted locally out-performed GOTTCHA (EDGE) at strain level but not at species or genus level identification. Likewise, there were differences between the output of Kraken hosted locally and Kraken (EDGE) (Figure 4A and 4B). These differences could be explained with the tools using different versions of their databases or could be different release versions of the software. There is also the possibility that the tools hosted on-line have different options selected for the commands to run the tools. No information could be found pertaining to the database or commands used by the on-line tools.

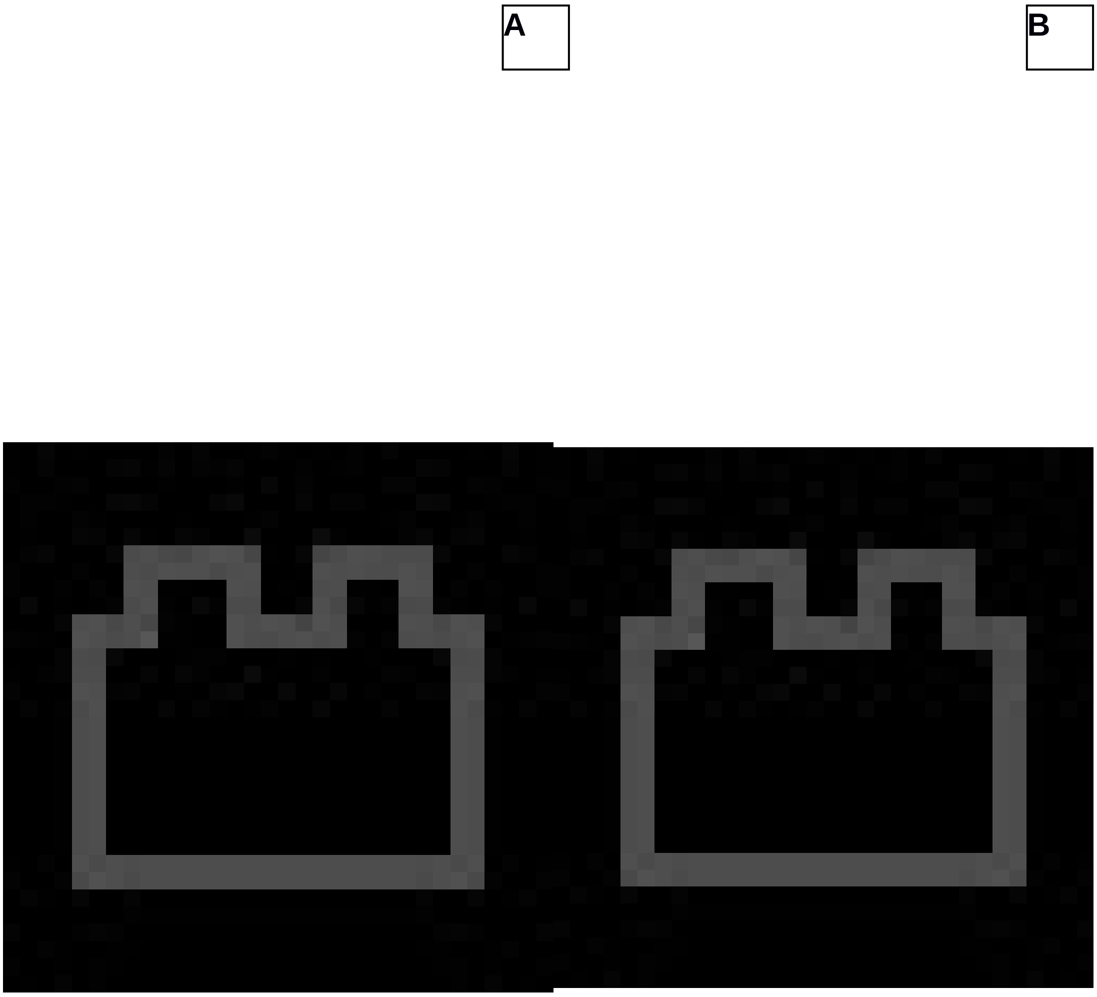


Figure 3: Total number of taxa identified in the *in-silico* datasets AM_10G_10M (A) and Zymo_8G_10M (B) using ten binning tools (MetaPhlAn2, MetaPhlAn (EDGE), One Codex (Filtered), One Codex, Kraken-mini, Kraken-mini (EDGE), GOTTCHA, GOTTCHA (EDGE), Phylosift and Constrains) at strain, species and genus taxonomic levels.

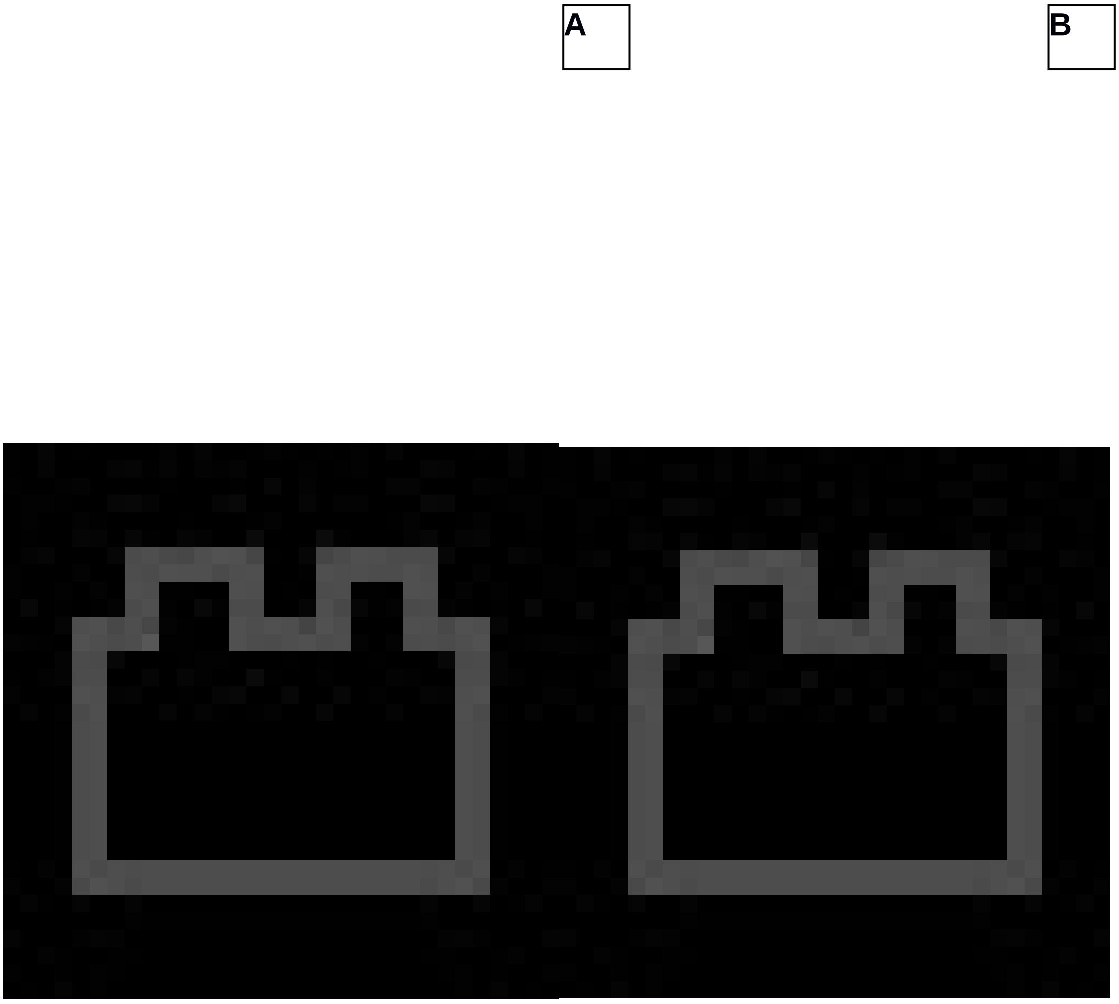


Figure 4: Percentage of true positive, false positive and false negative identifications for the *in-silico* datasets AM_10G_10M (A) and Zymo_8G_10M (B) using ten binning tools (MetaPhlAn2, MetaPhlAn (EDGE), One Codex (Filtered), One Codex, Kraken-mini, Kraken-mini (EDGE), GOTTCHA, GOTTCHA (EDGE), Phylosift and Constrains) at strain, species and genus taxonomic levels.

The sensitivity of the tested binning tools was consistently high (mean sensitivity of 94.8%) at the genus level with all but one of the tools (GOTTCHA hosted locally – 53% sensitivity) delivering 100% sensitivity. At species level, MetaPhlAn2, One Codex (filtered), One Codex, Kraken-mini, Kraken-mini (EDGE), GOTTCHA (EDGE) and Phylosift were all 100% sensitive. The remaining tools were still able to identify a majority of species; ConStrains – 90 and 100%, MetaPhlAn – 80 and 100% and GOTTCHA – 40 and 62.5% sensitive for the two datasets. The binning tools were less sensitive at identification to the strain level. MetaPhlAn, GOTTCHA (EDGE) and One Codex (filtered) had 0% sensitivity at strain level. MetaPhlAn2 and ConStrains were 30% and 0% sensitive at strain level identification for the two datasets which was bettered by GOTTCHA – 40% and 62.5% and PhyloSift – 60% and 75% sensitive to the two datasets. However, OneCodex, Kraken and Kraken (EDGE) were 100% sensitive at the strain level for both *in-silico* datasets (Figure 5A and 5B).

The binning tools could be split into two cohorts when looking at the precision of the tools. OneCodex, Kraken, Kraken (EDGE) and Phylosift had much lower precision in comparison to MetaPhlAn2, MetaPhlAn, GOTTCHA, GOTTCHA (EDGE) and ConStrains. Apart from One Codex (filtered) which only returned results at the species level, the higher performing tools showed 100% precision at genus level. In comparison Kraken, Kraken (EDGE) and Phylosift showed precision rates of 9.7 and 9.4%, 9.3 and 9.7%, and 0.5 and 0.7% respectively. At species level the high performing tools had a range of 72% – 100% precision compared to 0.3% – 7.2% for the lower cohort. The range of precision for the lower cohort at strain level identification was 0.3% – 3.0%. In comparison MetaPhlAn2, GOTTCHA and ConStrains showed 100% precision at strain level identification. GOTTCHA (EDGE), MetaPhlAn and One Codex (filtered) did not give any results for identification at the strain level (Figure 6A and 6B).

Binning tools did not offer accurate strain identification. Constrain, MetaPhlAn2 and GOTTCHA (hosted locally) showed 100% precision but their sensitivity was consistently below 50%. This means that for every strain identified in the sample there was at least one more unidentified strain. Conversely, One Codex, Kraken and Kraken (EDGE) were 100% sensitive, with Phylosift showing 67% sensitivity at strain identification but the precision was below 1%, meaning only 1 in 100 strains identified were actually present in the sample. This low precision was also observed at species and genus identification. MetaPhlAn2 and GOTTCHA (EDGE) were able to identify at species and genus level to 100% precision and sensitivity.

The results generated using the simple *in-silico* datasets AM_10G_10M and Zymo_8G_10M generated results that resolved contradictory reports in the literature for some tools and confirmed reports for others. As mentioned in the introduction, Kraken-mini is an interesting case with conflicting results reported in the literature. It is described as having high precision at the genus level (>90%) [76] but also as generating large numbers of false positives at the species level (> 4,000) [105]. The results from this work show (Figure 3A) that the tool generates a large number of false positives (209 at species level identification) confirming the results from the later paper rather than the introductory Kraken paper [88]. Similarly, when MetaPhlAn2 was first described in the literature it was shown that its levels of precision and sensitivity were equally high [36]. The tool described 10 false positives and 12 false negatives at species level across 24 *in-silico* datasets. These results corroborate the results from this work shown in Figure 3.

GOTTCHA was described in the literature as returning higher levels of false negatives (7) compared to false positives (0) across six synthetic datasets[105]. Figure 4 confirms the results described in the literature; zero false positives were returned for the two simple *in-silico* datasets generated and analysed on this project.

OneCodex (filtered) was reported in the literature to have higher accuracy compared to Kraken, CLARK and OneCodex[75]. These findings were confirmed here using the simple *in-silico* datasets shown in Figures 3, 5 and 6.

The only tool that gave substantially differing results between this work and the literature was Constrains. Constrains was reported as being able to identify and differentiate between strains of *E. coli* [78]. So it was a surprising result that

Constrains was only able to identify 17% of the strains within the two *in-silico* samples (Figure 4).

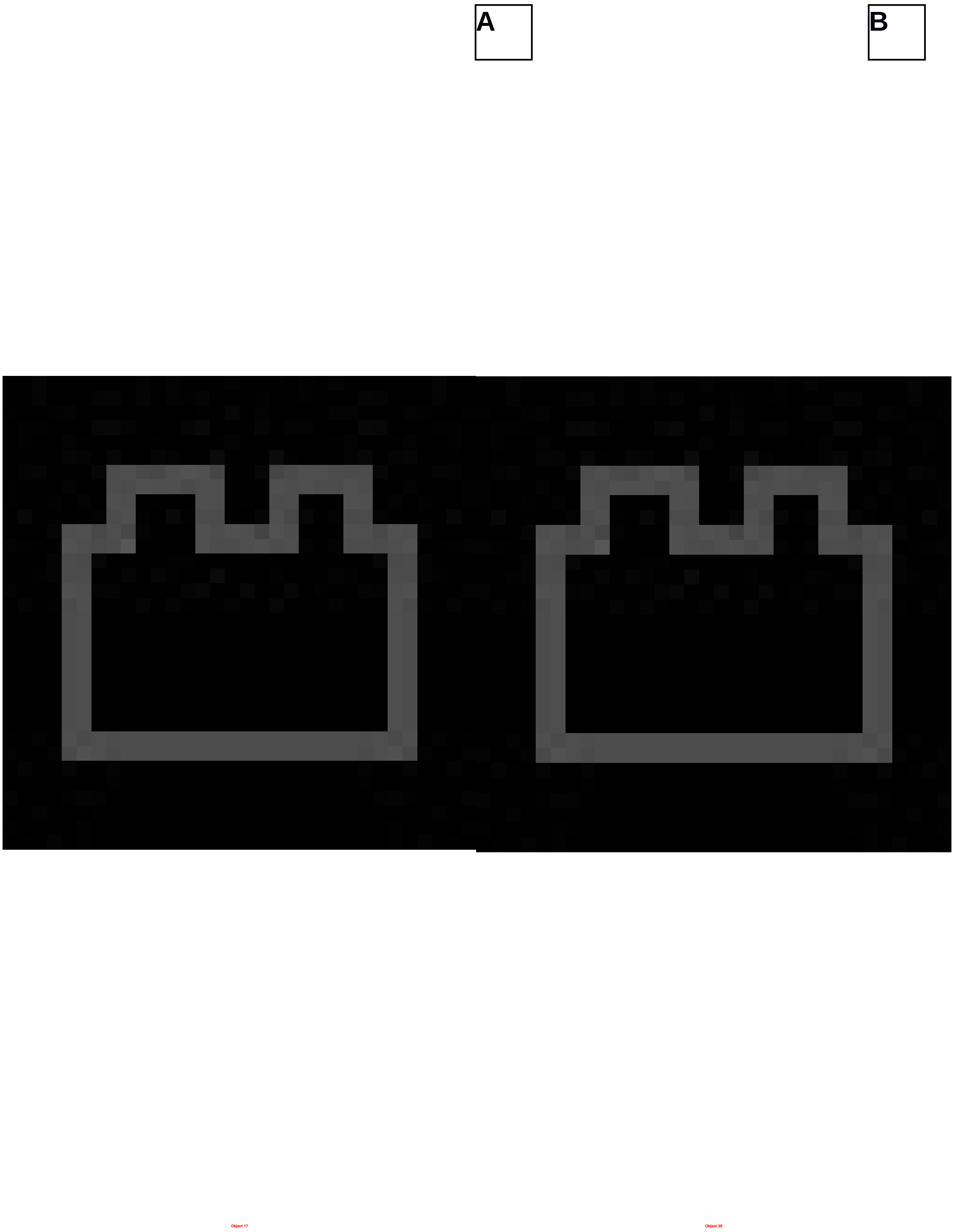


Figure 5: Sensitivity for 10 binning tools (MetaPhlAn2, MetaPhlAn (EDGE), One Codex (Filtered), One Codex, Kraken-mini, Kraken-mini (EDGE), GOTTCHA, GOTTCHA (EDGE), Phylosift and Constrains) post analysis of the *in-silico* datasets AM_10G_10M (A) and Zymo_8G_10M (B) at the strain, species and genus taxonomic levels

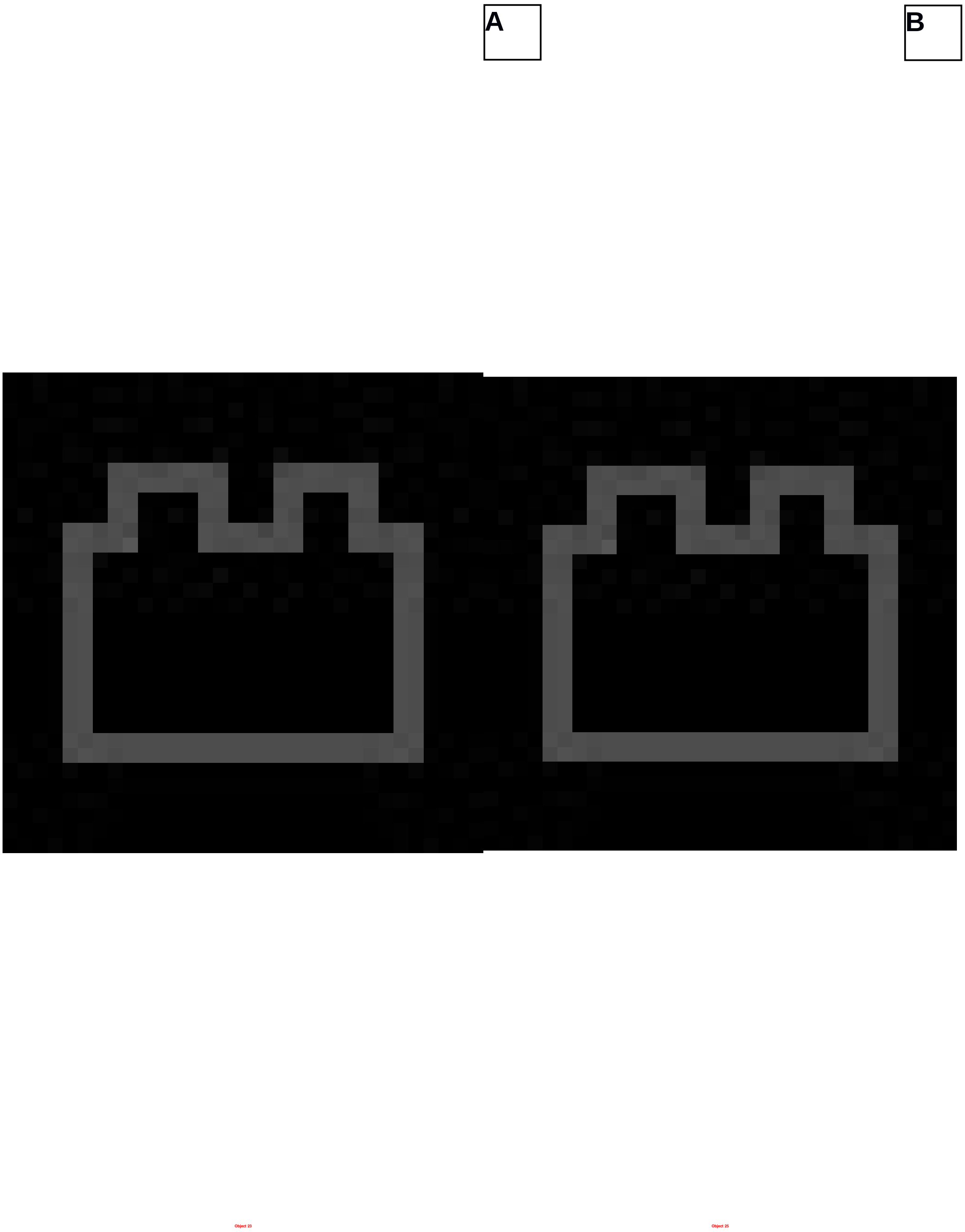
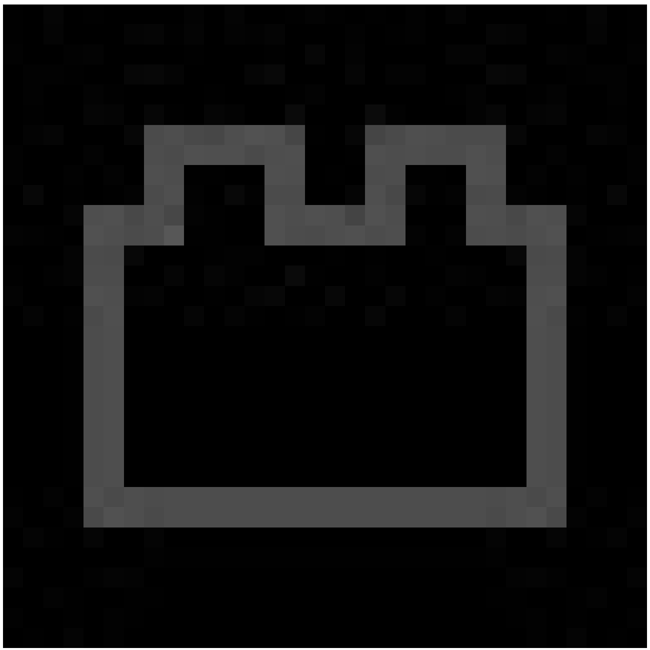


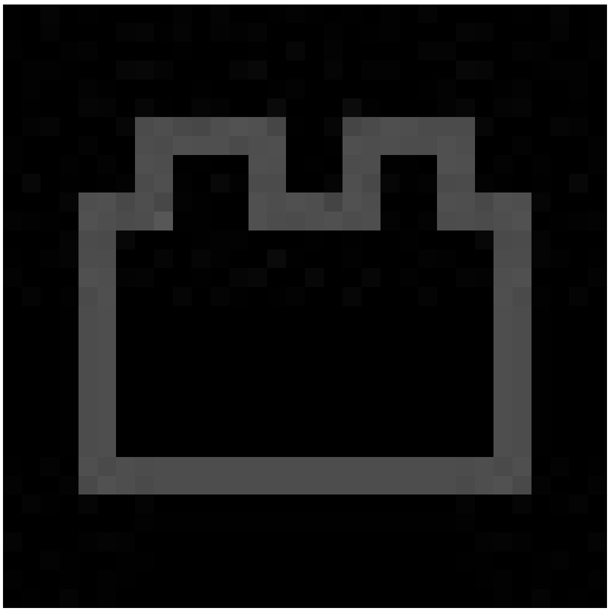
Figure 6: Precision for 10 binning tools (MetaPhlAn2, MetaPhlAn (EDGE), One Codex (Filtered), One Codex, Kraken-mini, Kraken-mini (EDGE), GOTTCHA, GOTTCHA (EDGE), Phylosift and Constrains) post analysis of the *in-silico* datasets AM_10G_10M (A) and Zymo_8G_10M (B) at the strain, species and genus taxonomic levels.

2.3.2 Results from the analysis of simple *in-silico* datasets using assembly tools.

The taxonomic classifications by MEGAN on the contigs generated with the assembly tools were more consistent compared to those generated with the binning tools. Similar levels of true positives, false positives and false negatives were seen across all tools tested (Figures 7 and 8). The four assembly tools tested were able to correctly identify all the organisms present in both datasets at species and genus level. The number of false positives ranged from 12 – 20 genera at the genus level across both datasets and 18 – 45 species at the species level. There were 2 false negative identifications for ABYSS, Megahit and IDBA-UD at strain level for AM_10G_10M and 1 false negative for Zymo_8G_10M, CLARK does not generate strain level results (Figure 7).

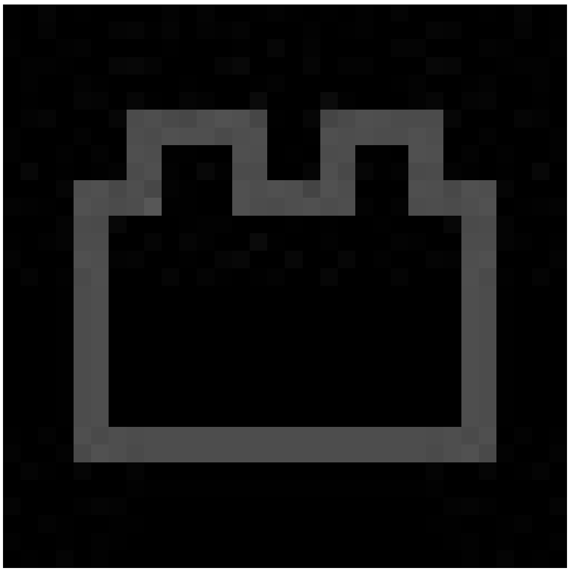


Object 27



Object 30

Figure 7: Number of taxa identified from the analysis of the *in-silico* datasets AM_10G_10M (A) and Zymo_8G_10M (B) using four assembly tools (IDBA-UD, Megahit, ABYSS and CLARK) at strain, species and genus taxonomic levels.



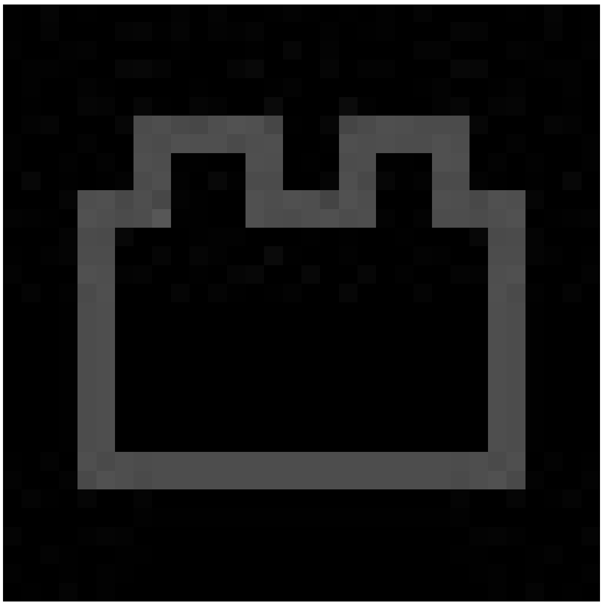
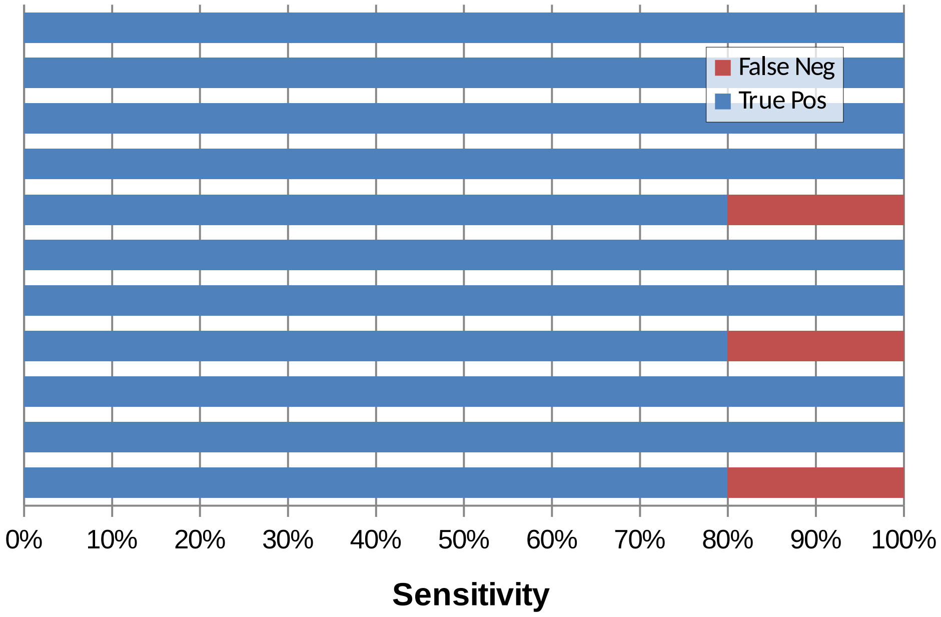
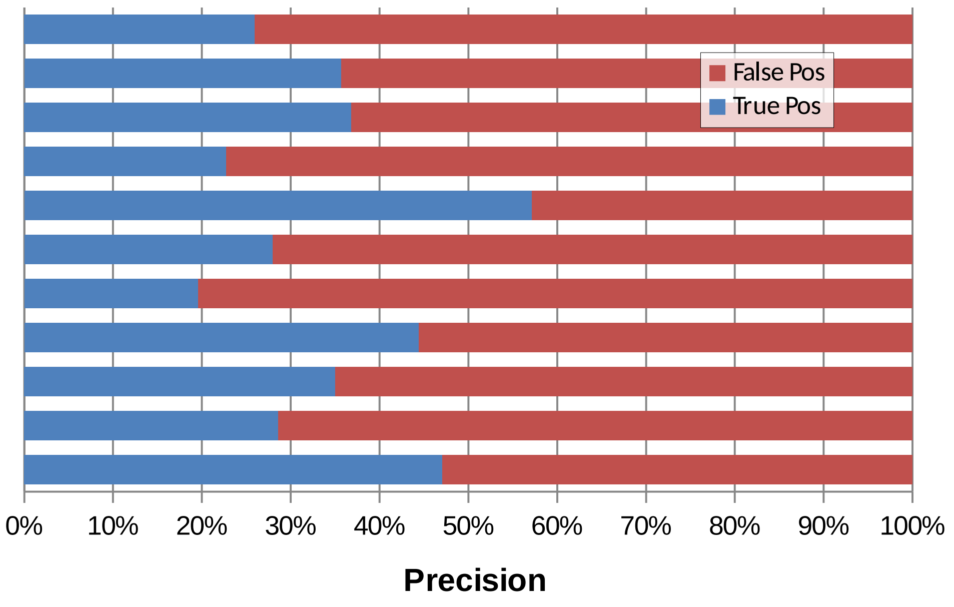


Figure 8: Percentage of true positive, false positive and false negative identifications for the *in-silico* datasets AM_10G_10M (A) and Zymo_8G_10M (B) using four assembly tools (IDBA-UD, Megahit, ABYSS and CLARK) at genus, species and strain taxonomic levels.

Taxonomic identification from MEGAN using the contigs from the assembly tools delivered low numbers of false positives ensuring high levels of sensitivity. The contigs generated by IDBA-UD, Megahit, ABYSS and CLARK all gave

100% sensitivity at genus and species level. The three tools that report results at strain level had a mean sensitivity of 80% for AM_10G_10M and 88% for Zymo_8G_10M (Figure 9).


False Neg
True Pos
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Sensitivity



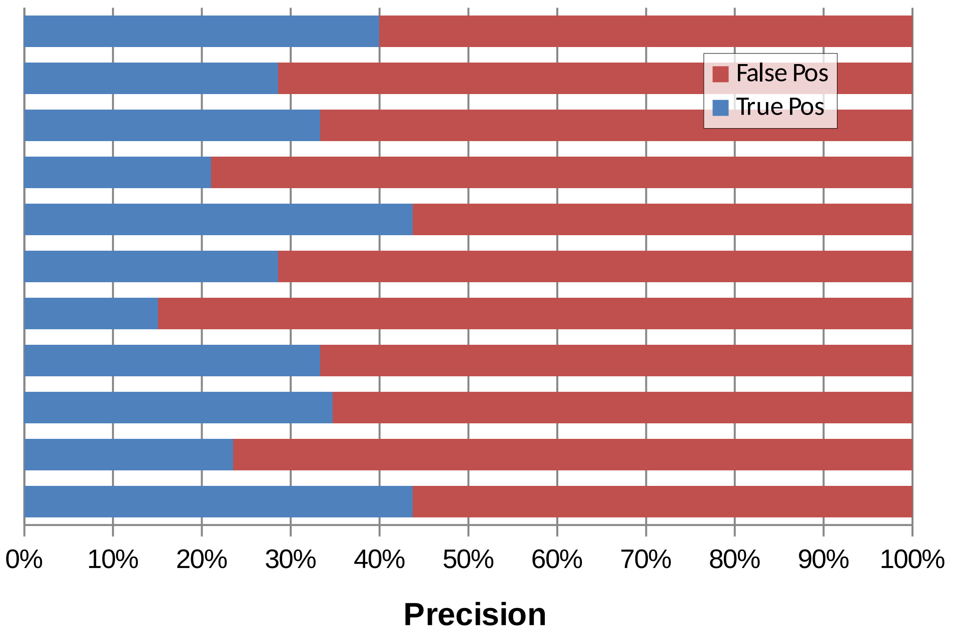
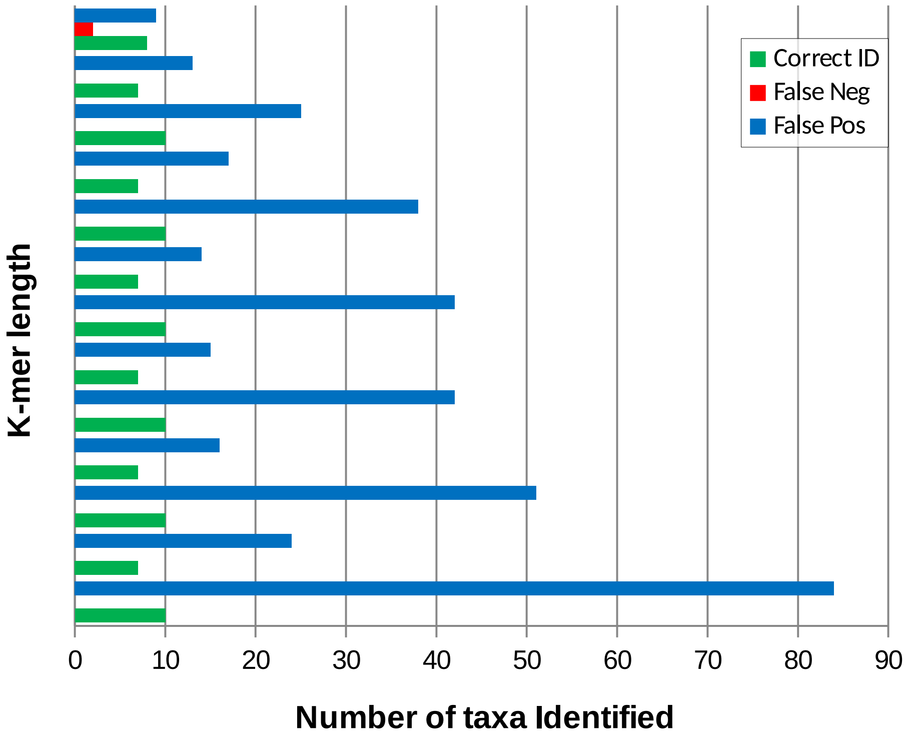


Figure 10: Precision of four assembly tools (IDBA-UD, Megahit, ABYSS and CLARK) for analysis of AM_10G_10M (A) and Zymo_8G_10M (B) at different taxonomic levels.

2.3.3 Results from the optimisation of assembly tools using simple *in-silico* datasets.

The assembly tools offer the ability to optimise some of the parameters in an effort to improve the accuracy of their output.

The output of IDBA-UD was optimised through altering the k-mer length parameter (lengths of 20, 40, 60, 80, 100 and maximum 'contig' were tested). The k-mer length is the length the reads are broken down into to build the dBG for assembly (See Figure 1). The lower the k-mer the less computationally expensive the assembly becomes as fewer edges are stored in the dBG. However, reducing the k-mer length increases the number of ambiguous nodes within the dBG which will reduce the length and quality of the contigs generated. Increasing the k-mer length from 20 to the maximum contig length reduced the false positives at the species level from 84 to 25 for AM_10G_10M dataset and from 59 to 26 for Zymo_8G_10M. However, at the longest k-mer length false negatives were reported, with two for the AM_10G_10M dataset and one in the Zymo_8G_10M dataset (Figure 11).


K-mer length
0
10
20
30
40
50
60
70
80
90
Number of taxa Identified
Correct ID
False Neg
False Pos

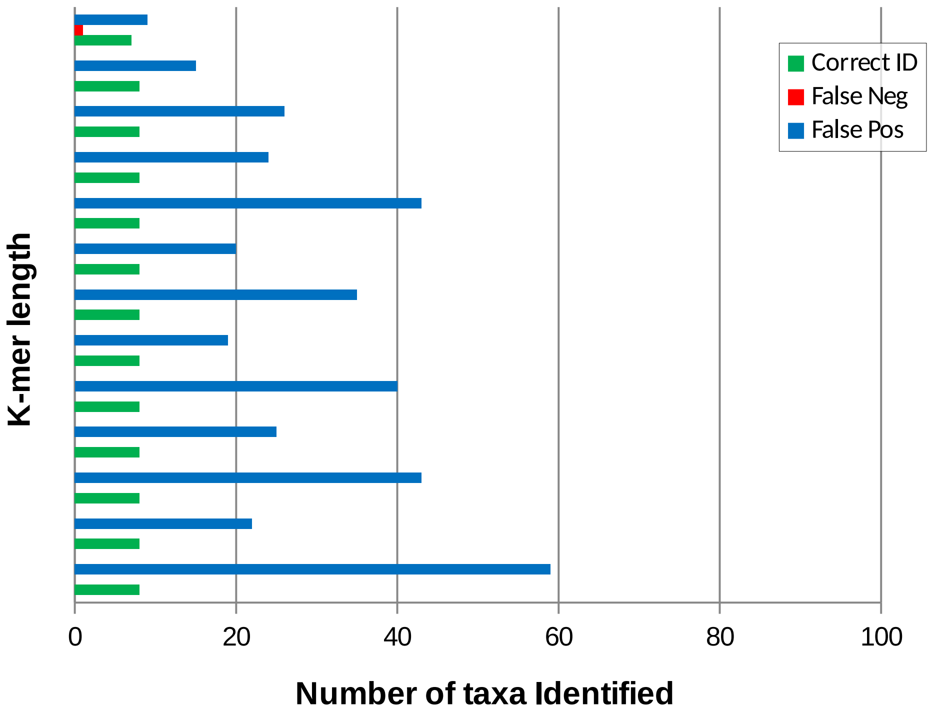
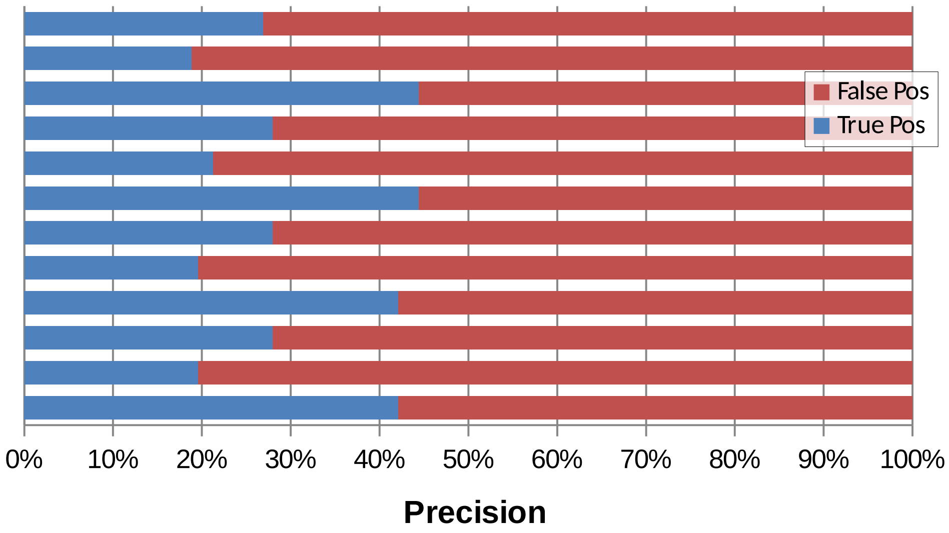

K-mer length
0
20
40
60
80
100
Number of taxa Identified
Correct ID
False Neg
False Pos

Figure 11: Impact k-mer length has on the accuracy of output when using the assembler IDBA-UD on the AM_10G_10M (A) and Zymo_8G_10M (B) datasets.

Megahit has predefined modes for different sample types: standard, meta, meta-large and meta-sensitive. The meta mode was develop to analyse standard metagenomic samples and uses a range of k-mer sizes from 21-99 increasing the k-mers by 20. The mode meta-large was designed for large/complex samples such as soil samples; its range of k-mer lengths is 27-87 and increases in steps of 20. The meta-sensitive mode is designed to give a sensitive but slower analysis of the sample; the k-mer lengths range from 21-99 but this time the step is set at 10, this approach generates twice the data as twice the k-mer lengths are explored. A comparison of these different modes using the two *in-silico* datasets showed they had a very subtle impact on the observed results. At strain level the precision for the standard mode was 42% and 32% and was improved to 44% and 33% with the meta-large mode for the two datasets. The meta-large mode also offered the optimum precision at species level of 21% and 15% - an improvement from 19% and 15% for the sensitive mode. The optimum precision at genus level of 28% and 30% came from the standard and meta mode, where-as the precision for the meta-sensitive mode reduced to 27% and 30% (Figure 12). There was no difference between the sensitivity of any of the pre-set modes, with all showing 100% sensitivity.



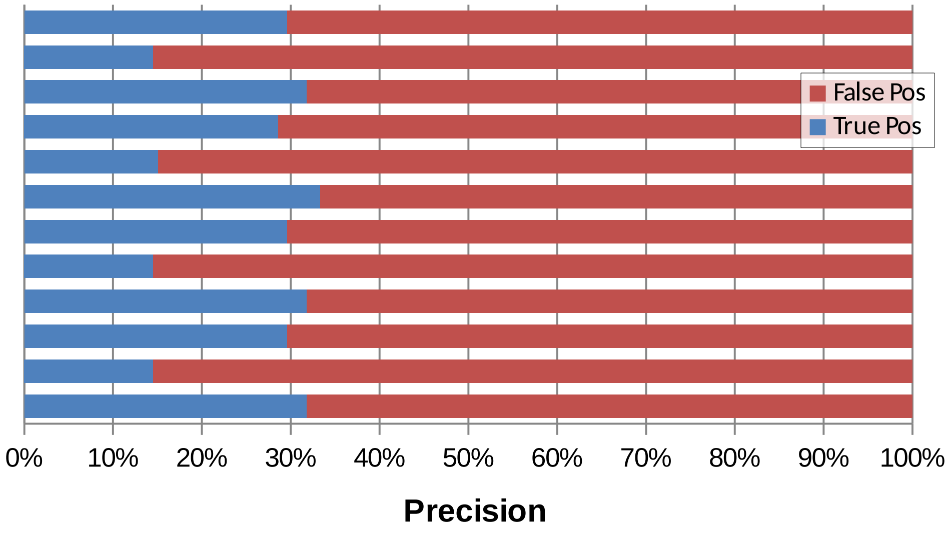
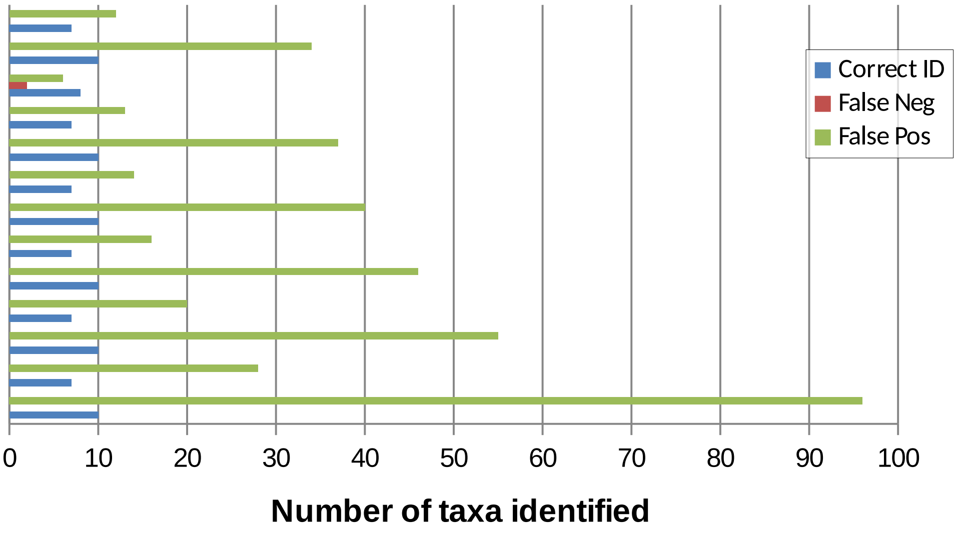


Figure 12: Precision of Megahit using preset modes for different taxonomic levels for datasets AM_10G_10M (A) and Zymo_8G_10M (B).

Changes to the k-mer length parameter impacted the output from ABYSS resulting in improved accuracy for taxonomic identification. The minimum k-mer length of 20 generated 96 and 80 false positives at species level and 28 and 25 false positives at genus level for the *in-silico* datasets AM_10G_10M and Zymo_8G_10M respectively. Selecting the longest k-mer length reduced the false positives at species level to 34 and 30 and at genus level to 12 and 16. This was a reduction in false positives of 64% at species level and 47% at genus level. As with IDBA_UD the optimum k-mer length for precision had a negative impact on sensitivity. There were 3 false positives reported at strain level across both datasets for k64 compared to no false negatives for the other k-mer lengths (Figures 13 and 14).



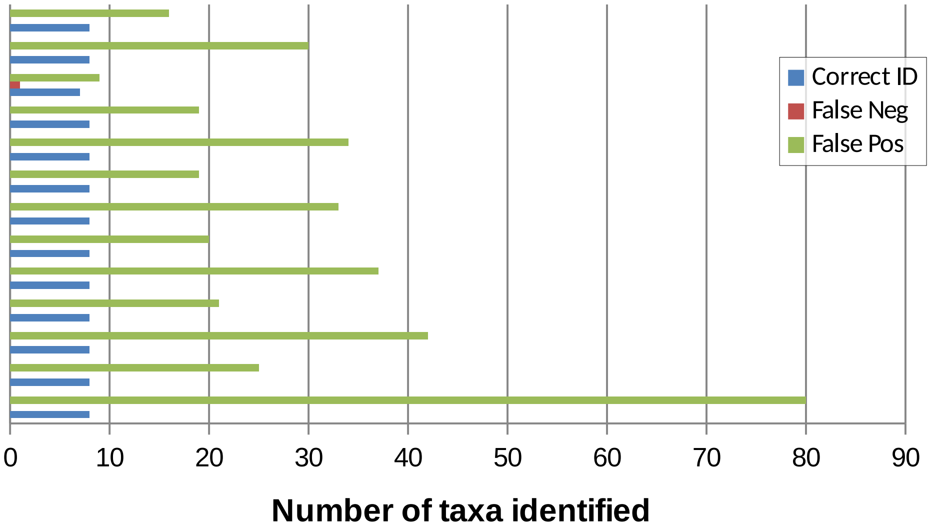
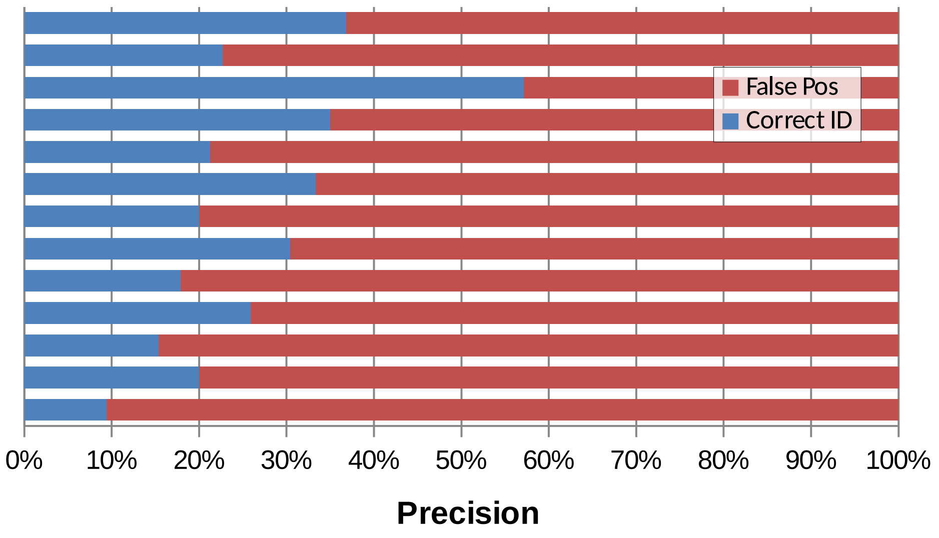


Figure 13: Impact the k-mer length has on the accuracy for the assembly tool ABYSS on the *in-silico* datasets AM_10G_10M (A) and Zymo_8G_10M (B) at the strain, species and genus taxonomic level.



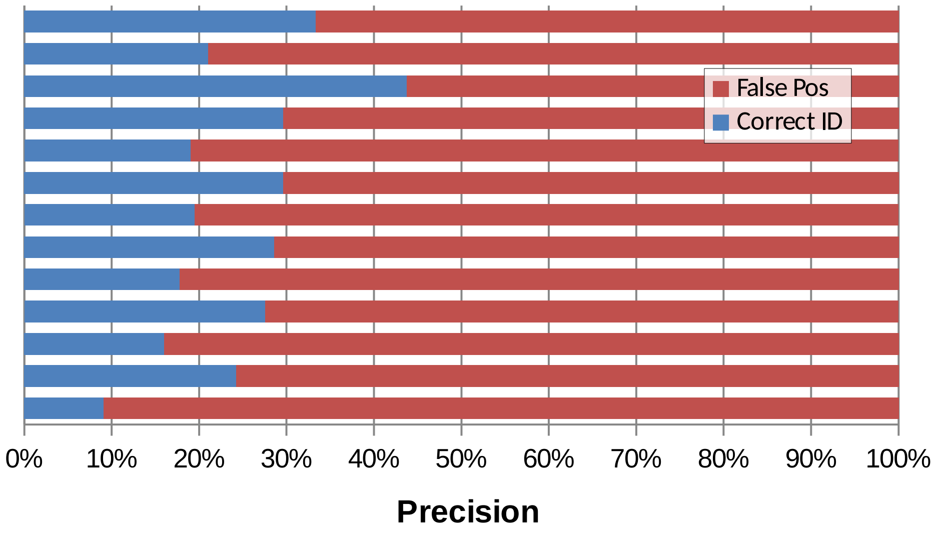


Figure 14: Impact the k-mer length has on precision for the taxonomic identification using contigs generated by the assembly tool ABYSS on the *in-silico* datasets AM_10G_10M (A) and Zymo_8G_10M (B) at the strain, species and genus taxonomic level.

The two parameters that can be optimised for CLARK are k-mer length and gap length. Changing the k-mer length had no impact on the output (results not shown). The gap length refers to the number of non-overlapping k-mers to pass. Increasing the gap length will reduce the RAM usage but is also reported to decrease sensitivity. Optimising the gap length reduced the number of false positives without impacting the number of false negatives; however, it was not a linear relationship between gap length and false positives. The general trend was that by increasing the gap length the number of false positives reduced. A gap length of 1 returned the lowest precision at species and genus level for both datasets. The precision for AM_10G_10M dataset at species level rose from 20% with a gap length of 1 to 45% with a gap length of 6 or 10, the precision at genus level rose from 20% to 37% with a gap length of 9. For the dataset Zymo_8G_10M the gap length increased from 16% to 35% with a gap length of 9, the precision at genus level rose from 17% to 67% with a gap length of 9 or 10 (Figure 15).

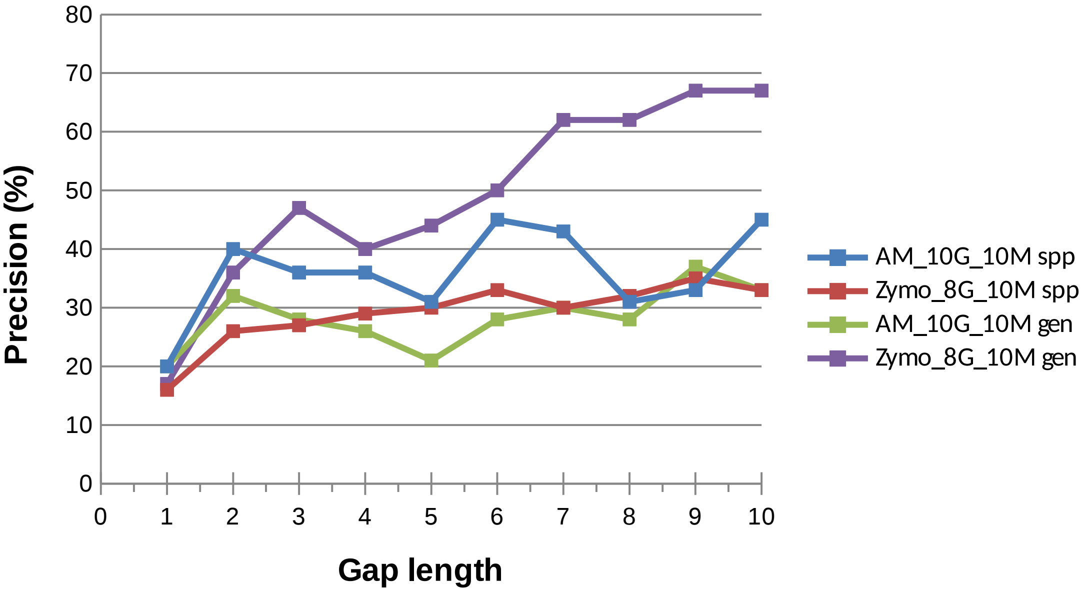
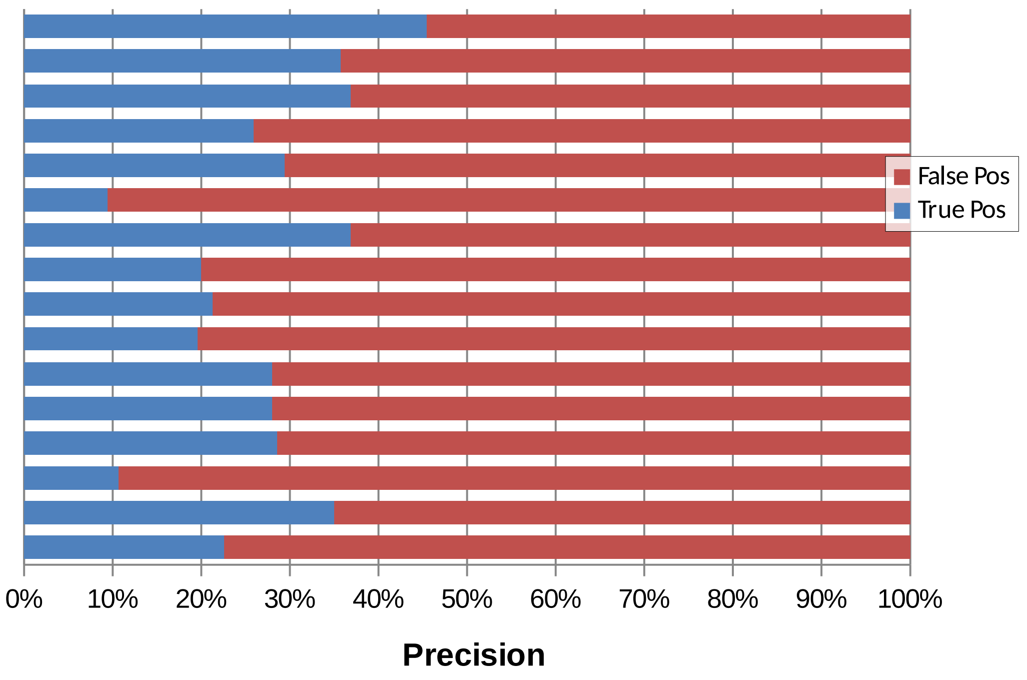


Figure 15: Precision of CLARK using different gap lengths at different taxonomic levels for *in-silico* datasets AM_10G_10M and Zymo_8G_10M.

The results from the assembly tool optimisation show that all tools tested had 100% sensitivity at species and genus level. However, altering the parameters can cause an increase in false negatives at strain level as seen with IDBA_UD and ABYSS. The pre-set options for Megahit had the least effect on results, with the precision only improving for species level identification with the AM_10G_10M dataset from 20 to 21%. CLARK showed the next lowest improvement to precision with increases just over 20% for species level and 42% and 68% at the genus level for the two datasets. ABYSS and IDBA-UD showed similar levels of improvements to their precision scores with an average increase of 144% and 132% at the species level and 62% and 41% for the two datasets respectively. Despite these large improvements to the precision of ABYSS and IDBA-UD it was CLARK that was the most precise tool both pre and post optimisation for both datasets at species and genus level (Figure 16).



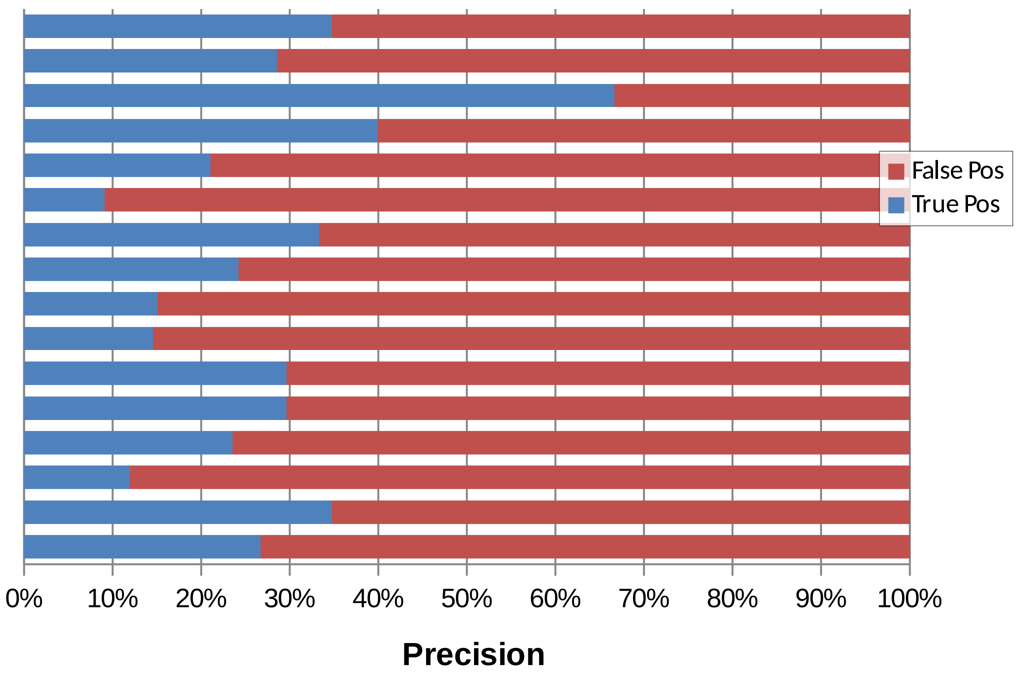


Figure 16: Results of assembly tool optimisation for four assembly tools (IDBA-UD, Megahit, ABYSS and CLARK) on precision for *in-silico* datasets

AM_10G_10M (A) and Zymo_8G_10M (B) at species and genus level identification.

2.4 Conclusion

The analysis of small *in-silico* datasets using Kraken, One Codex and Phylosift generated large numbers of high false positives. Despite their high sensitivity they were disregarded for future analysis due to their low levels of precision. Kraken-mini (EDGE) however was retained to determine how more complex datasets impact its sensitivity and accuracy. Kraken-mini (EDGE) was selected as it had one of the higher precision levels of the poorer performing tools. It also had a high ease of use; it was run when running GOTTCHA (EDGE). One Codex was discounted for further analysis. Despite its filtered results generating high sensitivity and precision, it was deemed that the high cost for each analysis made the tool unsuitable for further use.

Due to the inability of the analysis tools to accurately identify strains, future work was planned to focus on species and genus level identification. If the aim of this project was to identify the bio-warfare agents from a bio-aerosol sample then accurate identification to strain level would be a major goal. However, as this project is aiming to develop a pipeline to accurately measure the variation of bio-aerosols, strain identification is not an essential aspect. ConStrains was originally evaluated for its ability to add strain level identification to MetaPhlAn2 analysis and so it was not used for future analysis. Further work continued to identify new binning tools and evaluate their ability to identify species and genera present in *in-silico* datasets. MetaPhlAn2 and GOTTCHA (EDGE) were down-selected for further evaluation due to their 100% sensitivity and precision at species and genus identification.

Results from this initial analysis of simple *in-silico* datasets showed the sensitivity of the assembly tools at species and genus level were equal to the best performing binning tools. However, despite efforts to optimise their performance, the precision of the assembly tools at all taxonomic levels was lower than the binning tools. The assembly tools also require a higher compute resource and take considerably longer to run compared to the binning tools. Taking this into account, only the best assembly tool, which this initial analysis showed to be CLARK, was considered for further analysis. If there was a requirement for functional analysis of the metagenome then the assembly tools would offer advantages over the binning tools[109, 110], however, the purpose of this study is taxonomic identification of organisms present within the aerosol microbiome.

To further evaluate the down-selected analysis tools, more complex datasets were subsequently generated. Increasing the number of organisms in the sample will determine whether sample diversity impacts numbers of false positives. Varying the abundance of the organisms present within the sample will inform as to whether organism abundance will impact detectability. Due to the length of time that some bioaerosol samples will be left on the collector (potentially up to 8 hours) with 300 litres of air/minute passing over them, the micro-organisms could become degraded through dehydration and desiccation pressures[15]. The degraded cells could lyse causing the DNA to be exposed to stresses leading to fragmentation impacting the length of the output from the sequencers. Reducing the length of the reads in the *in-silico* datasets will show whether the length of sequence output has an impact on precision and sensitivity of the tools. This approach has the added benefit that the Illumina NextSeq generates shorter read lengths than the Illumina MiSeq so will help to identify whether sequencing platform will impact the output from the tools.

3 The construction and analysis of complex *in-silico* datasets using down-selected metagenomic analysis tools.

3.1 Introduction

In order to further examine the sensitivity and precision of the metagenomics analysis tools down-selected in the previous section, more complex datasets were designed. These datasets included increasing the diversity of the species present, altering the abundance of the species present and reducing the read lengths. The modifications were designed to generate a suite of datasets that more closely resembled a real bio-aerosol sample.

After analysis of the Singapore air samples 1 and 2 it is expected that there will be numerous species present in bio-aerosol samples from the same genus. The datasets with increased diversity were designed to test the tools' abilities to differentiate organisms that are closely related.

The datasets previously used to perform the initial tool down-selection (Chapter 2) were designed with even abundance for each of the represented species. This even representation of organisms is an unlikely reflection of reality. It is important to gain an understanding of a tool's ability to identify organisms that are under-represented compared to the more dominant species within the sample. It is also of interest to see if more dominant species mask the lower abundant species.

The selected sequencing platform for this project was the Illumina MiSeq using NexteraXT library preparation reagents to generate 300bp reads. In order to replicate this read length the datasets for the initial down-selection of analysis tools reflected the 300bp reads. In reality the sequencer generates reads with a variety of read lengths; the majority of reads may be 300bp but there are also a number of shorter reads. The potential for cell lysis on the collection filters could lead to the degradation of DNA prior to sequencing, impacting the read lengths generated. Generating datasets with shorter read lengths will enable the impact read length has on the output from the down-selected analysis tools to be investigated.

3.2 Methods

3.2.1 *In-silico* simulation of complex datasets for evaluation of metagenomic analysis tools.

All of the *in-silico* datasets described here were generated using MetaSim as described in Chapter 2.

3.2.1.1 Generating a dataset with increased diversity

This dataset was again based on the Singapore air sample (used for AM_10G_10M); the top 100 abundant species were included and covered 73 genera (Appendix 2). The dataset was designed with an even coverage of all species. The previously described Illumina error profile was used, with a read length of 300bp and a total of 10 million reads. This dataset was designed to be comparable to AM_10G_10M and Zymo_8G_10M and was referred to as:

- AM_100G_10M

3.2.1.2 Generating datasets of variable abundance

MetaSim was used to modify the three previously described datasets (AM_100G_10M, AM_10G_10M and Zymo_8G_10M) with mixed abundance of the organisms present (abundances shown in Appendix 2). The previously described Illumina error profile, 300bp read length and 10 million read options were maintained. The resulting abundances for AM_100G_10M-V ranged from 0.3%-1.9% with 10 species at each of the abundance. Dataset AM_10G_10M-V's abundance ranged from 1%-19% and Zymo_8G_10M-V ranged from 1%-24% abundance. These new datasets were titled:

- AM_100G_10M-V
- AM_10G_10M-V
- Zymo_8G_10M-V

3.2.1.3 *Generating datasets with shorter read length*

MetaSim was used to generate datasets with read lengths of 150bp. These datasets were based on the previously described datasets AM_100G_10M, AM_10G_10M and Zymo_8G_10M. In order to build these datasets the Illumina error profile that had been previously used had to be redeveloped. The error profile works by defining the chance of error at each position of the synthesised read. In order to ensure that the error for each synthesised read was the same as the 300bp error model the error profile couldn't just be divided in two; if the 3' end of the error profile was used then the error would be too low, likewise if the 5' end was used then the error would be too high. Instead, every odd base pair was removed from the 300bp error model ensuring that the error for the whole read remained the same (Figure 17).

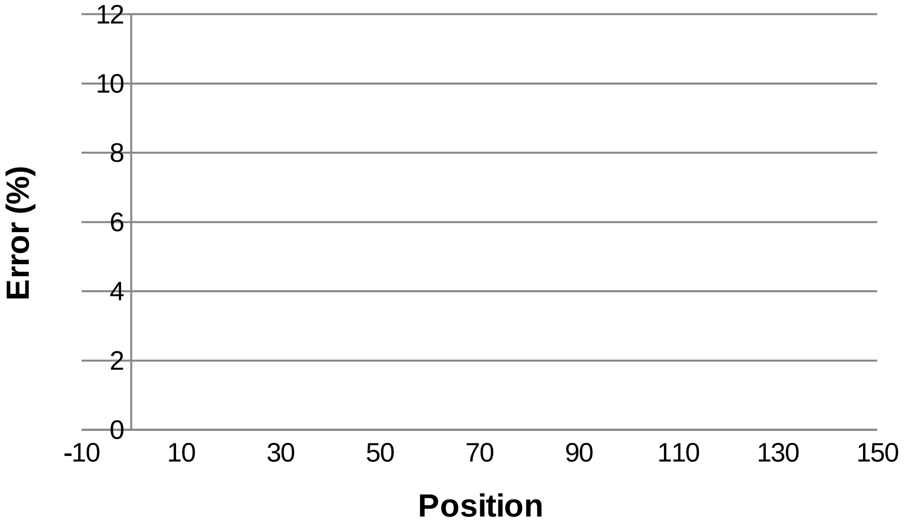


Figure 17: A representation of the rate of error assigned to each position of the 150bp error profile used with MetaSim to simulate the AM_100G_10M-150, AM-10G_10M-150 and Zymo_8G_10M-150 *in-silico* datasets.

Each of the species in the datasets had an equal abundance and there were 10 million reads generated for each of the datasets. These three new datasets were labelled:

- AM_100G_10M-150
- AM_10G_10M-150

- Zymo_8G_10M-150

3.2.2 Analysis tools

The analysis tools used to evaluate the more complex datasets were three binning tools (MetaPhlAn2, GOTTCHA-EDGE and Kraken-mini) and one assembly tool (CLARK). The commands described in Tables 3 and 4 were followed for all analysis performed in this section.

3.2.3 Results interpretation

The results from this analysis will be interpreted as previously described for the simple *in-silico* datasets (Section 2.2.4).

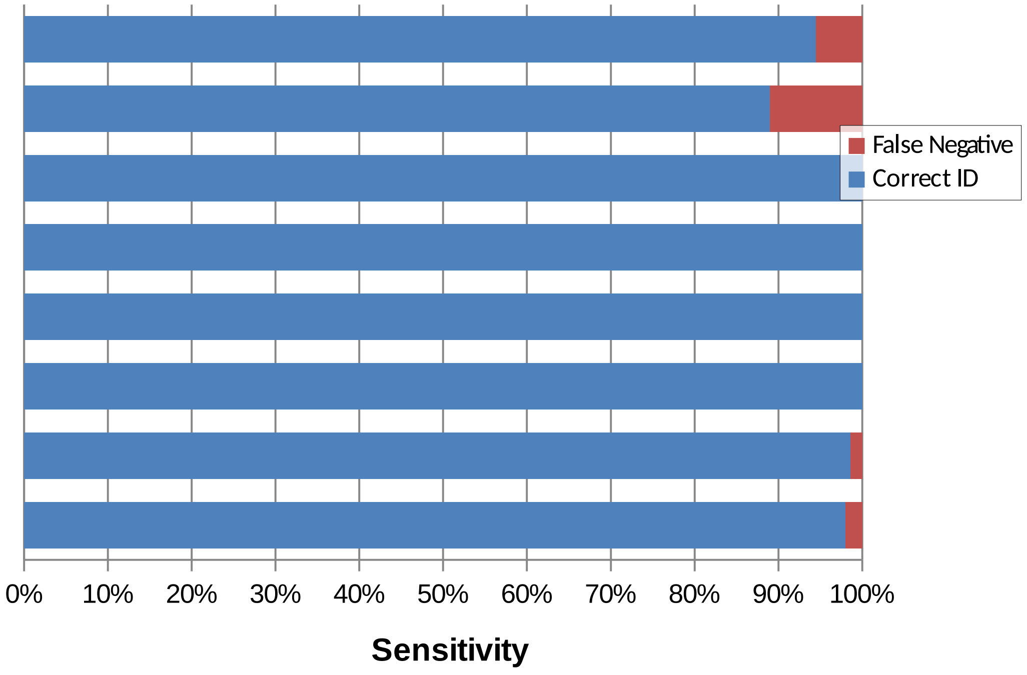
3.3 Results and Discussion

3.3.1 Results from dataset with increased diversity

Increasing the number of species present in the *in-silico* datasets had no impact on the sensitivity for GOTTCHA or Kraken-mini, with 100% sensitivity across all three datasets at both species and genus level for both tools (Figures 5 and 18). There was a slight drop in sensitivity for MetaPhlAn2 from 100% for both AM_10G_10M and Zymo_10G_10M (Figure 5) at species and genus level to 98% for species level identification and 99% for identification at the genus level. The precision of MetaPhlAn2 and GOTTCHA dropped slightly at species level from 100% for both of the smaller datasets to 97% and 99% precision respectively. In contrast to this trend of poorer results with a more complex dataset, Kraken's precision increased from 5.4% for AM_10G_10M and 3.7% for Zymo_8G_10M to 36% for AM_100G_10M at species level, with a similar increase for genus level identification. The high numbers of false positives reported by Kraken are generally closely related to species within the sample. As the *in-silico* dataset was designed to have multiple species from the same genus it is possible that a number of species are misidentified but the false positives are fortuitously present in the sample.

The results for CLARK show a reduction in sensitivity from 100% at genus and species level for the two smaller datasets (Figure 9) to 95% and 89% for the AM_100G_10M dataset (Figure 18A). The precision of CLARK for the larger dataset did not change greatly from the optimised analysis of the smaller datasets. The precision of identification at species level for AM_10G_10M was 45% and Zymo_8G_10M was 35% compared to 45% for the larger dataset. Whereas for genus level identification the precision was 37% and 67% for the smaller datasets compared to 58% for AM_100G_10M (Figures 16 and 18B).

Due to CLARK's low level of sensitivity and precision it did not offer sufficiently high enough accuracy, it was therefore disregarded for future analysis.


False Negative
Correct ID
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Sensitivity

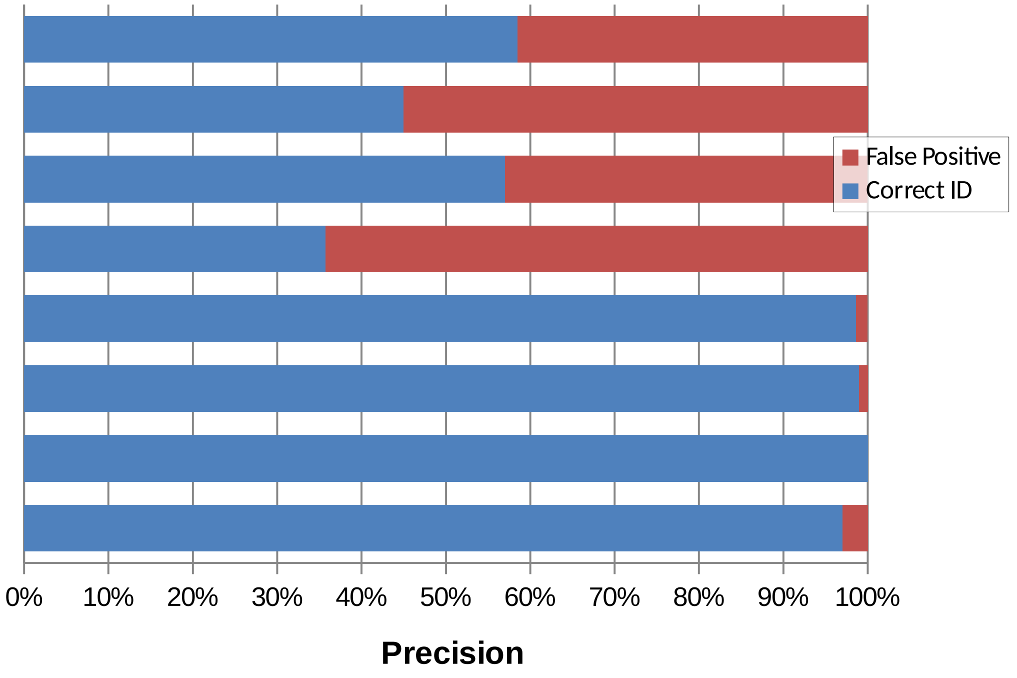

False Positive
Correct ID
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Precision

Figure 18: The sensitivity (A) and precision (B) of four metagenomic analysis tools (MetaPhlAn2, GOTTCHA, Kraken and CLARK) for the *in-silico* dataset AM_100G_10M.

3.3.2 Results for datasets with shorter read lengths

Altering the read length for the *in-silico* datasets had the most impact on the accuracy of the tools tested. When comparing dataset AM_100G_10M-150 (Figure 19) with AM_100G_10M (Figure 18) the sensitivity was slightly lower for the three tools tested at species and genus level identification. For example MetaphlAn2 reduced in sensitivity from 98% to 96% for species identification and GOTTCHA's sensitivity reduced from 100% to 97% for genus identification. The precision for MetaPhlAn2 and GOTTCHA showed very little difference in identification at the species or genera level of identification. However, Kraken's precision dropped from 36% to 29% for species identification and from 57% to 43% for genus identification. This equates to a reduction in precision of 19% and 25% respectively.

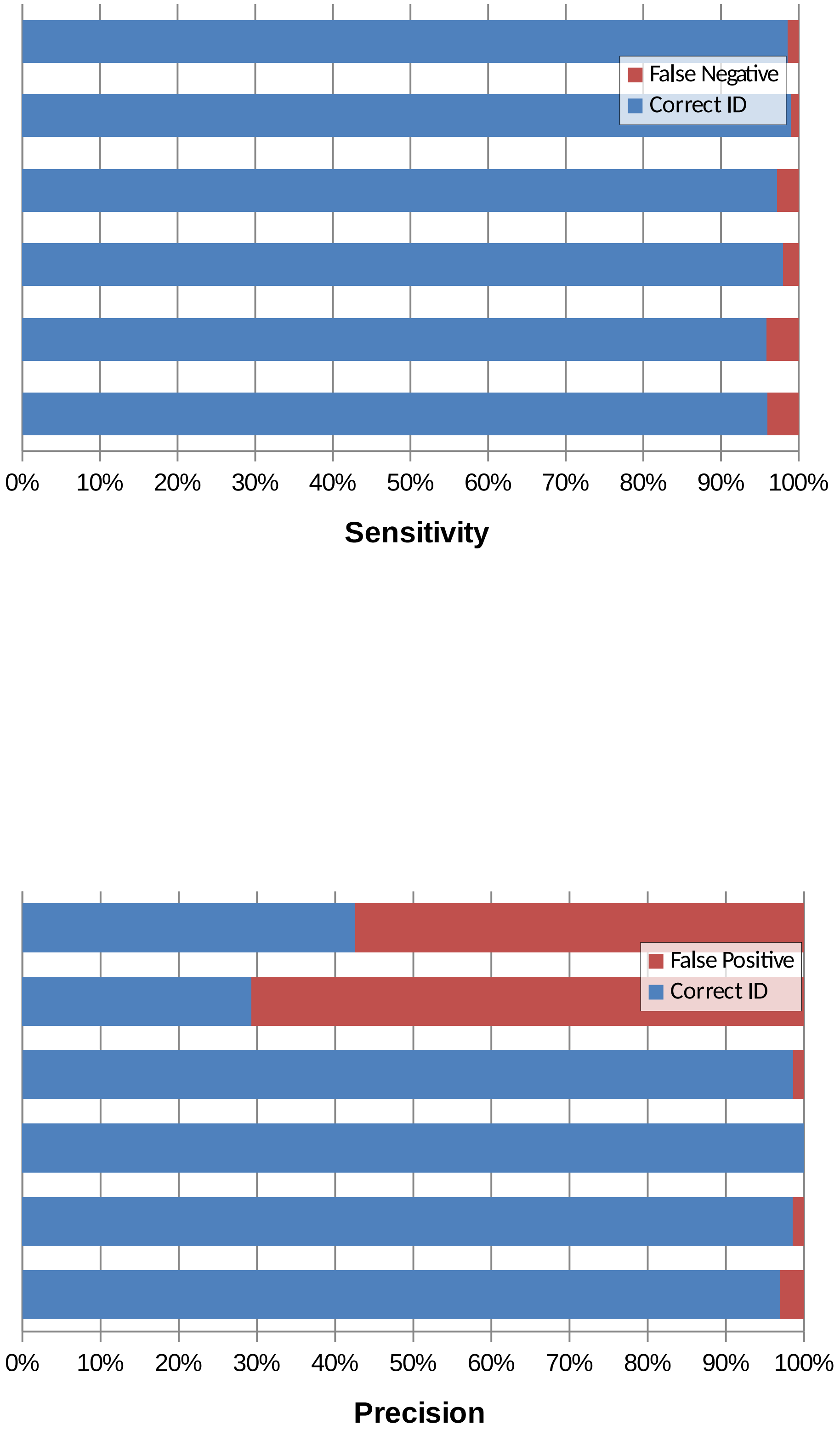


Figure 19: The sensitivity (A) and precision (B) of three metagenomic analysis tools (MetaPhlAn2, GOTTCHA and Kraken) for the *in-silico* dataset AM_100G_10M-150.

All tools tested returned 100% sensitivity when analysing the AM_10G_10M and AM_10G_10M-150 datasets at species and genus level. However, Kraken's precision showed the greatest difference with the species identification reduced by 26% to 4% and there was a reduction of 44% to 5.2% at the genus level (Figure 20).

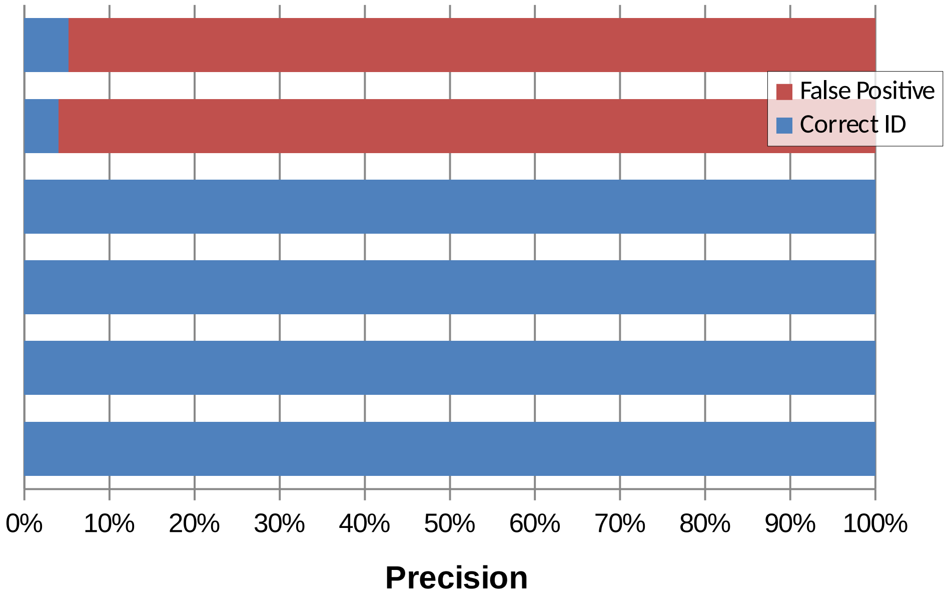


Figure 20: The precision of three metagenomic analysis tools (MetaPhlAn2, GOTTCHA and Kraken) for the *in-silico* dataset AM_10G_10M-150.

The results for the comparison between the Zymo_8G_10M (Figures 5 and 6) and Zymo_8G_10M-150 (Figures 21) datasets were very similar to those of the AM_10G_10M datasets. All tools tested returned 100% sensitivity at both specie and genus level for the Zymo_8G_10M and Zymo_8G_10M-150 datasets. There was also no difference in precision between the two datasets for MetaPhlAn2 and GOTTCHA. The precision was lower for Zymo_8G_10M-150 compared to Zymo_8G_10M, with a reduction of 21.6% to 2.9% precision at species level and a reduction of 40.8% to 5.8% precision for identification at the genus level.

Reducing the read length from 300 bases to 150 bases had a similar result to the two previous parameters that were modified. There was no difference for MetaPhlAn2 and GOTTCHA to the output for the 2 smaller datasets at species and genus level identification. Kraken showed no difference in sensitivity at species or genus level, with all results at 100%. Kraken did show subtle reductions in precision when the read length was reduced and these reductions were more pronounced for genus level identification. The precision for AM_10G_10M-150 at genus level was 5.2% compared to 9.3% for AM_10G_10M and it reduced from 9.8% Zymo_8G_10M to 5.8% for Zymo_8G_10M-150. At species level Kraken's precision reduced for all three datasets. AM_100G_10M-150 was 29% compared to 36% for the 300bp alternate dataset, AM_10G_10M-150 was 25% lower with a precision of 4%, the output for Zymo_8G_10M-150 was 2.9% compared to 3.7% for Zymo_8G_10M.

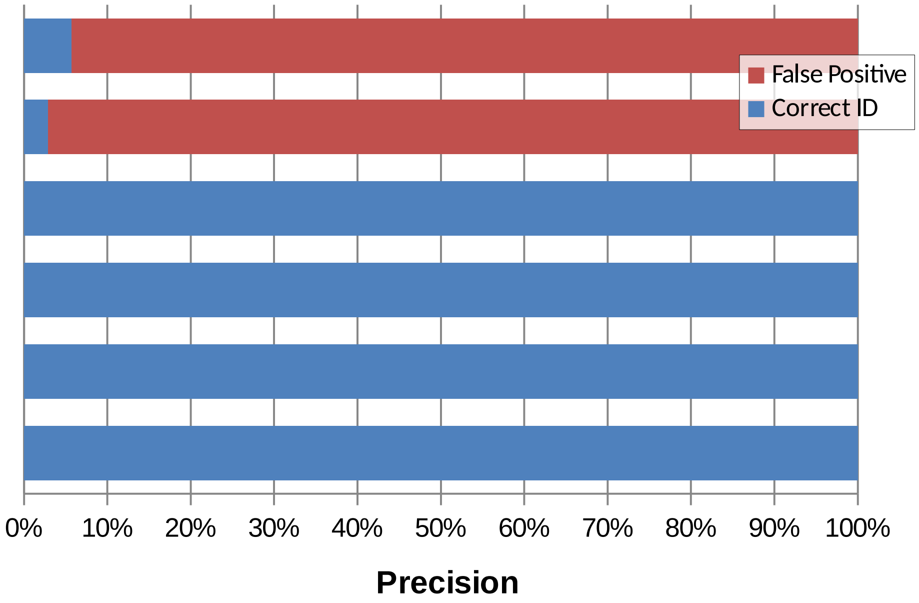


Figure 21: The precision of three metagenomic analysis tools (MetaPhlAn2, GOTTCHA and Kraken) for the *in-silico* dataset Zymo_8G_10M-150.

3.3.3 Results from datasets with variable abundance

Altering the abundance of the organisms present in the *in-silico* datasets had little impact on the output from the analysis tools. For the smaller datasets (AM_10G_10M, AM_10G_10M-V, Zymo_8G_10M and Zymo_8G_10M-V) there was no difference to either the sensitivity or precision for MetaPhlAn2 and GOTTCHA at species and genus level (Figures 5, 6, 22 and 23). Likewise, there was no difference to the sensitivity of Kraken at species or genus level. There was only a very slight difference to Kraken's precision at species level (from 5.4% for AM_10G_10M to 5.6% for AM_10G_10M-V at species level and from 3.7% for Zymo_8G_10M to 4.2% for Zymo_8G_10M-V at species level).

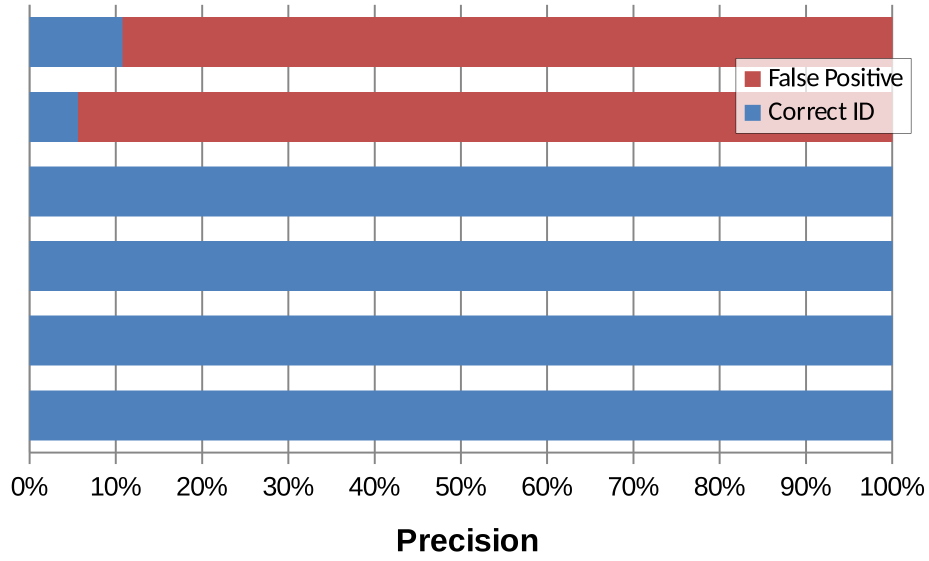


Figure 22: The precision of three metagenomic analysis tools (MetaPhlAn2, GOTTCHA and Kraken) for the *in-silico* dataset AM_10G_10M-V.

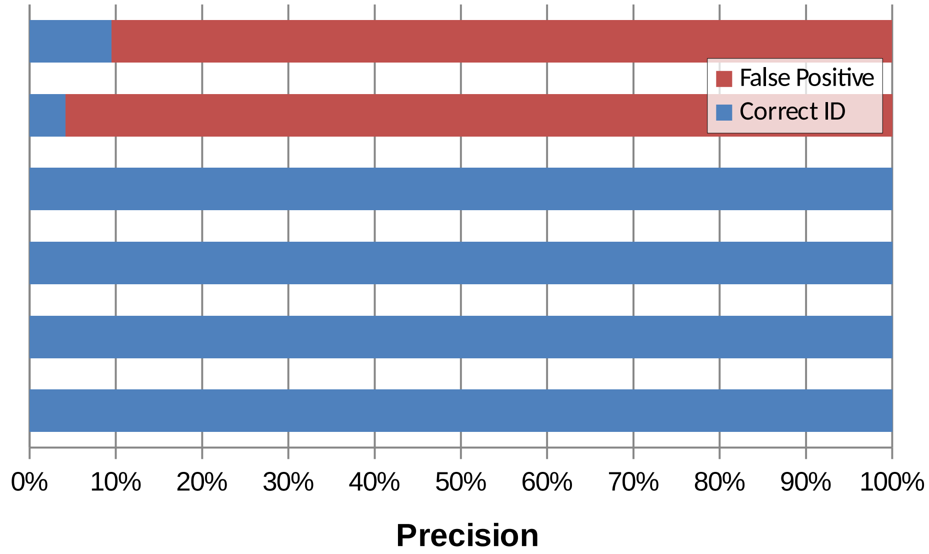


Figure 23: The precision of three metagenomic analysis tools (MetaPhlAn2, GOTTCHA and Kraken) for the *in-silico* dataset Zymo_8G_10M-V.

The results followed a similar trend when comparing the effect varying the species abundance had on the larger dataset (AM_100G_10M and AM_100G_10M-V). GOTTCHA produced the same results for sensitivity and precision at species and genus level for AM_100G_10M-V and AM_100G_10M (Figures 18, and 24). MetaPhlAn2 results showed very little difference between the two datasets, the greatest difference being a 1.3% reduction to the sensitivity at genus level. Regarding Kraken's results, the sensitivity did not alter when the species abundance was changed at species or genus level identification. In addition, the precision increased slightly when the abundance was changed. The mean increase in precision for the three datasets at species level identification was 1.6% and for genus level identification the mean increase in precision was 4.5%.

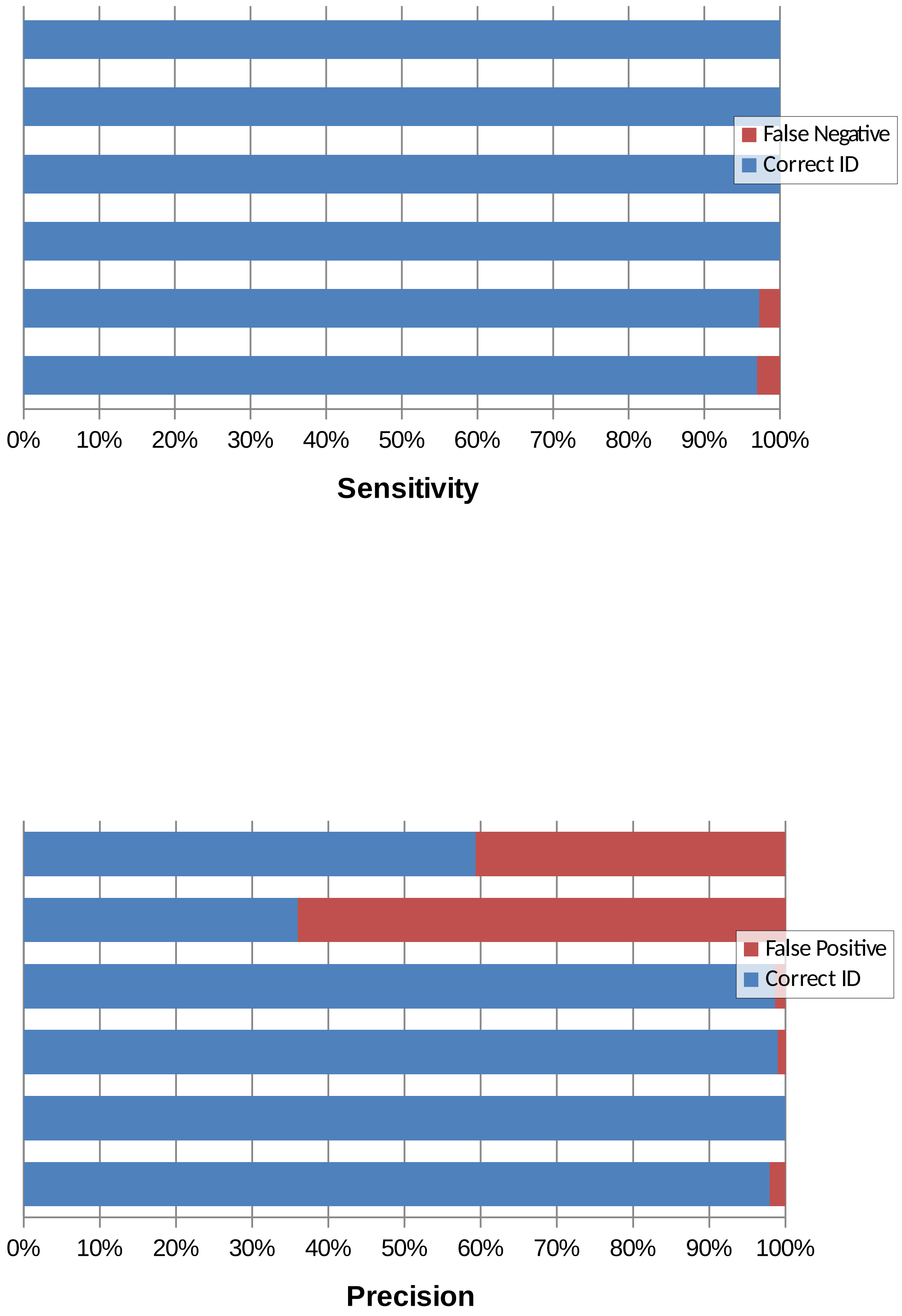


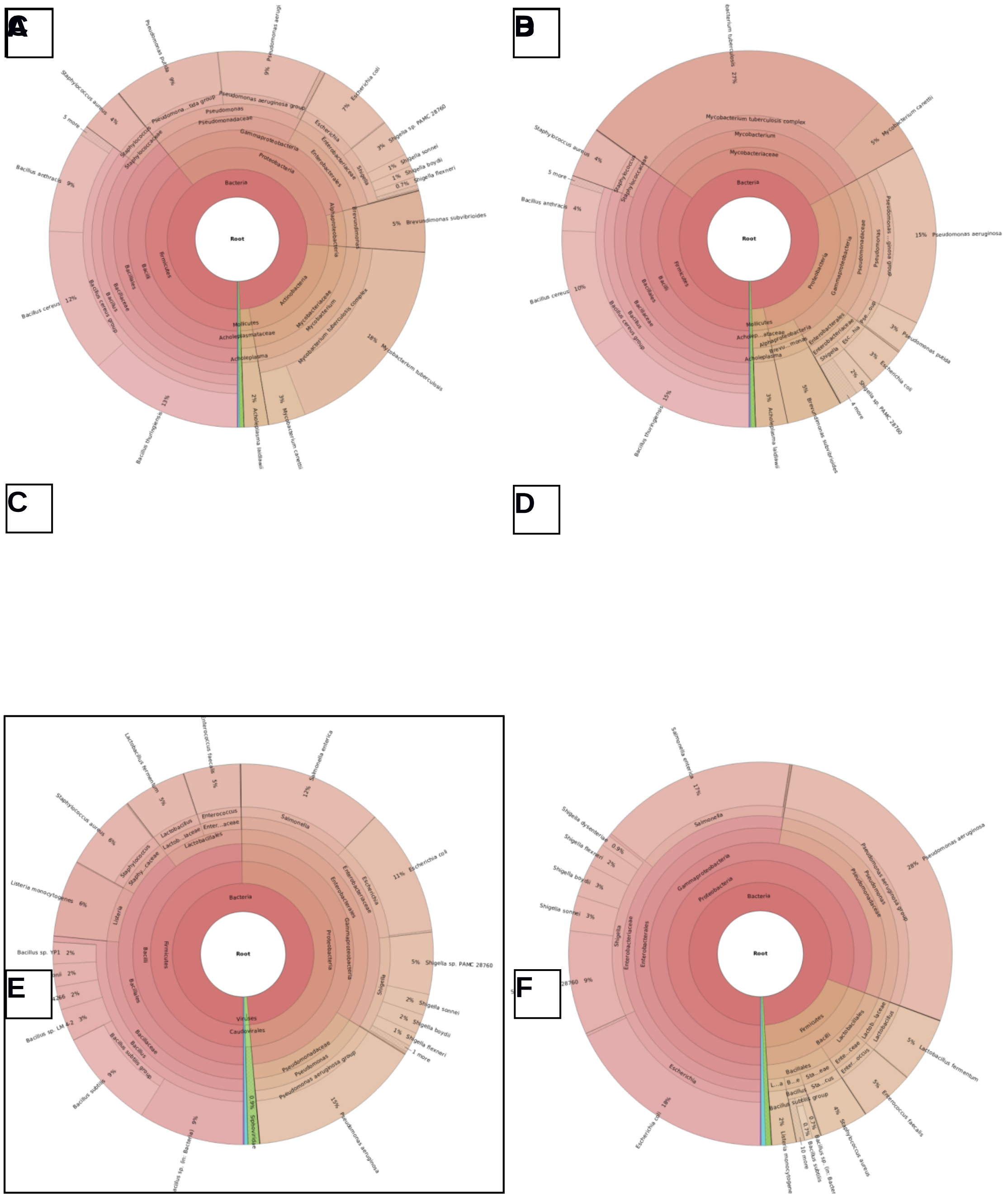
Figure 24: The sensitivity (A) and precision (B) of three metagenomic analysis tools (MetaPhlAn2, GOTTCHA and Kraken) for the *in-silico* dataset AM_100G_10M-V.

These results show that altering the abundance of the species present in these *in-silico* datasets has very little impact on the output of the tools. Having a more divergent abundance of species could have a more dramatic effect on the tools output. For example contaminating DNA from laboratory reagents are a major issue, especially for low biomass environments[111]. Suggesting samples with high levels of DNA are less likely to be impacted by contaminating DNA as the low levels are masked by highly abundant species. However, due to a lack of available time, this avenue was not explored further in this piece of work.

3.3.4 Accuracy of metagenomic analysis tools to estimate the relative abundance of taxon within *in-silico* samples.

True positive, false positive and false negative identifications are an important metric for the evaluation of metagenomic analysis tools. However, when monitoring how organisms change over time simple identification is not sufficient. Of the three down-selected tools two, MetaphlAn2 and GOTTCHA, output relative abundance results. Because Kraken doesn’t output the relative abundance of the species identified, alternative metagenomic analysis tools were required. Centrifuge was identified as a suitable tool (Appendix 3) which outputs relative abundance data and performs similarly to Kraken. See Appendix 3 for the output of Centrifuge’s analysis of the *in-silico* datasets previously described.

The relative abundance data produced by Centrifuge can be used to generate Krona plots, as shown in Figure 25. Interactive versions of these Krona plots are available at the following link: . The Krona plots give an easily interpretable visual representation of the abundance of different species within the sample. The plots shown in figure 25 illustrate Centrifuges ability to differentiate the abundance of the species within the sample. Figures 25 A, C and E are from samples with equal abundance and the majority of species identified are shown to have an even level of abundance. Figures 25 B, D and F are from datasets with varied abundance, Figures 25 B and D most clearly show the abundance for species identified is varied.


A
B
C
D
E
F
Root
Bacteria
Proteobacteria
Gammaproteobacteria
Firmicutes
Bacilli
Bacillales
Bacillaceae
Bacillus
Bacillus cereus group
Pseudomonas aeruginosa
Pseudomonas aeruginosa group
Pseudomonadaceae
Pseudomonas
Mycobacterium tuberculosis complex
Mycobacterium tuberculosis
Mycobacteriaceae
Mycobacterium
Actinobacteria
Mollicutes
Acholeplasma
Acholeplasmataceae
Acholeplasma laidlawii
Brevundimonas subvibrioides
Enterobacteriaceae
Enterobacterales
Escherichia
Escherichia coli
Shigella
Shigella sp. PAMC 28760
Shigella sonnei
Shigella boydii
Shigella flexneri
Shigella dysenteriae
Staphylococcus aureus
Staphylococcaceae
Staphylococcus
Bacillus anthracis
Bacillus thuringiensis
Bacillus cereus
Bacillus subtilis
Bacillus subtilis group
Mycobacterium canettii
Pseudomonas putida
Salmonella
Salmonella enterica
Listeria monocytogenes
Listeria
Lactobacillus fermentum
Lactobacillus
Lactobacillales
Enterococcus faecalis
Enterococcus
Viruses
Caudovirales
Siphoviridae
5 more
4 more
1 more
10 more

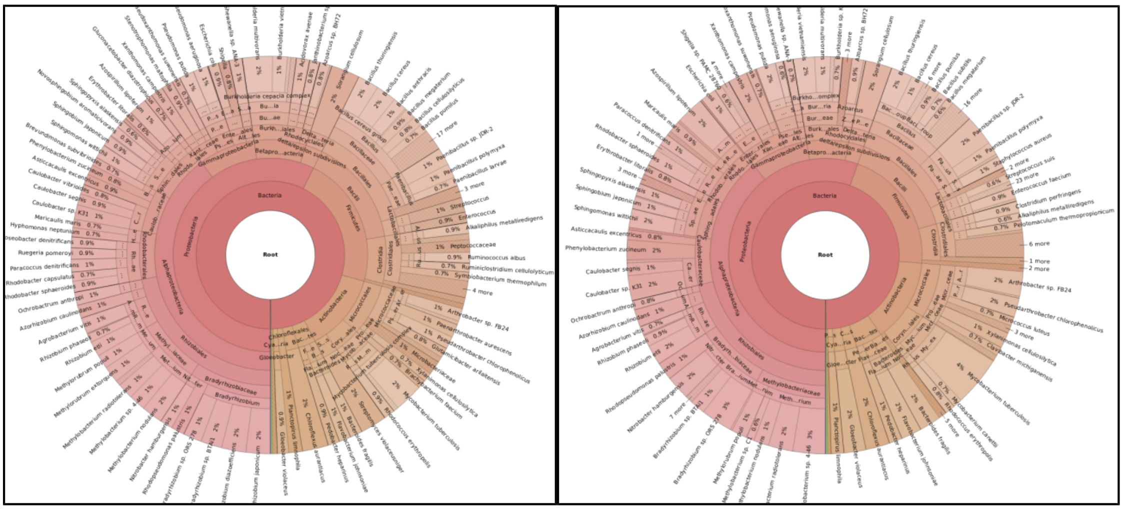


Figure 25: Krona plots showing relative abundance determined by Centrifuge at species level for datasets AM_10G_10M (A) and AM_10G_10M-V (B), Zymo_8G_10M (C) and Zymo_8G_10M-V (D), AM_100G_10M (E) and AM_100G_10M-V (F).

The abundance output for the three down-selected tools was able to predict the relative abundance for each of the species present with varying degrees of accuracy.

Figure 26A shows the relative abundance as calculated by MetaPhlAn2, GOTTCHA and Centrifuge for the even abundance dataset AM_10G_10M. GOTTCHA shows the most accurate prediction of abundance with a mean of 10% ± 0.7 (95% confidence interval) for the 10 species present in the dataset. MetaPhlAn2 is the next most accurate tool with an average abundance for the 10 species of 9.6% ± 2.3 and Centrifuge showed the lowest accuracy with a mean abundance of 6.3% ±2.7. The lower accuracy for Centrifuge can be explained by the large number of false positives. These false positives will all have an abundance associated with them and therefore reduce the accuracy of the relative abundancy output.

Figure 25B shows the tools predictions for the varied abundance dataset AM_10G_10M-V. The tools are all able to demonstrate the increase in abundance. Again, GOTTCHA is the more accurate tool with a mean difference to the actual abundance of 0.6%. MetaPhlAn2 shows a mean difference to the mean of 3.15% and Centrifuge, again the least accurate tool, with a mean difference to the known abundance of 4.88%.

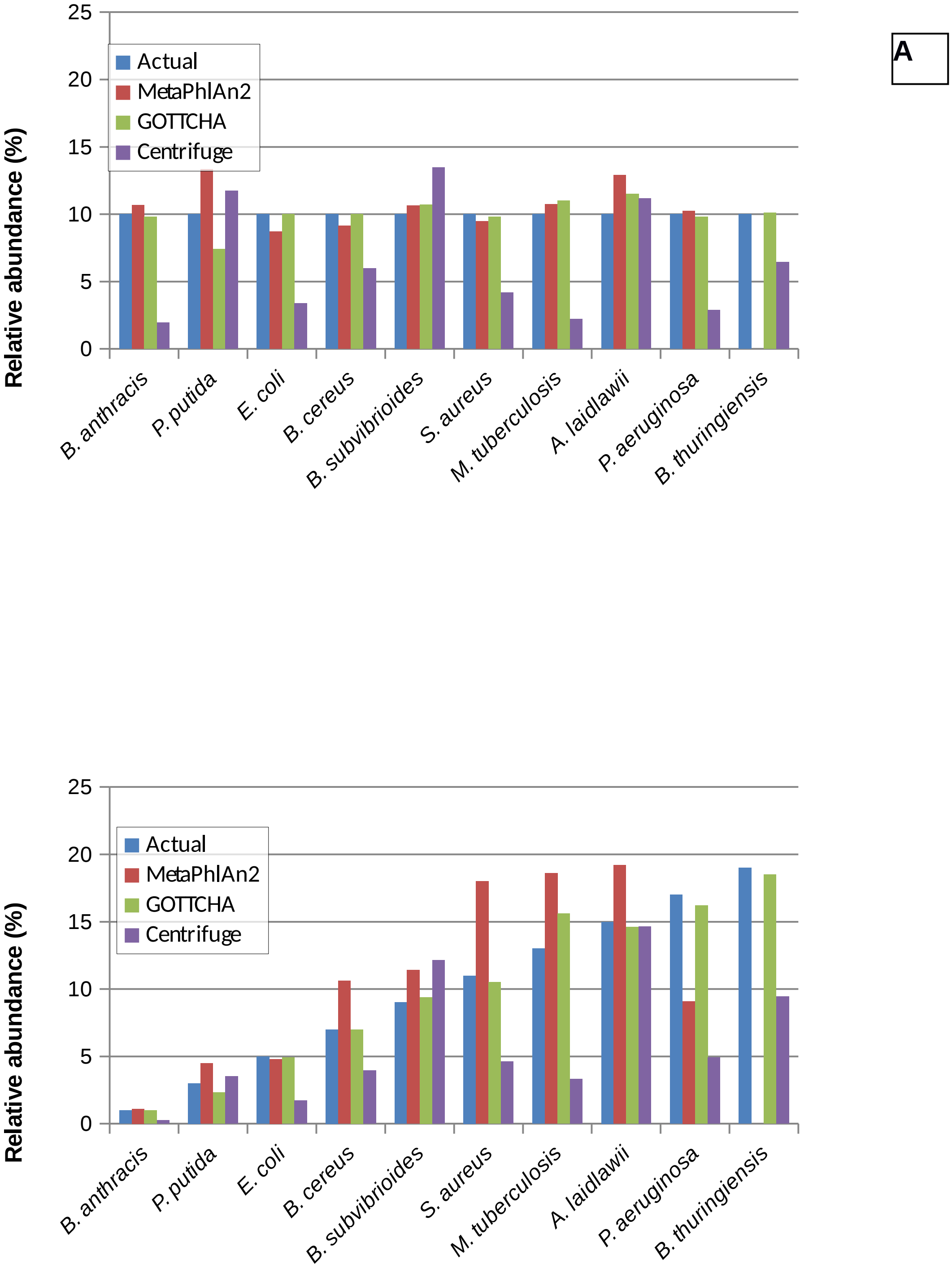


Figure 26: Relative abundance as calculated by three tools (MetaPhlAn2, GOTTCHA and Centrifuge) for the datasets AM_10G_10M (A) and AM_10G_10M-V (B) compared to the actual abundance and the mean abundance for the three tools.

The analysis of the constant abundant dataset Zymo_8G_10M showed very similar results to the AM_10G_10M dataset. GOTTCHA was again the more accurate tool with a mean relative abundance for the 8 species of 12.5% ±1.4. MetaPhlAn2 predicted a mean relative abundance for the dataset of 11.2% ±1.9 with Centrifuge, again, under-predicting the abundance with a mean relative abundance of 6.6% ±3.5 (Figure 27A).

Figure 27B shows that the three tools are able to describe the difference in abundance with the Zymo_8G_10M-V dataset with varying degrees of accuracy. GOTTCHA showed a mean difference to the known abundancies of 1.1%. MetaPhlAn2 was slightly less accurate with a mean difference to the known abundance of 1.73% and Centrifuge, once more, was the least accurate with a mean difference of 6.17%.

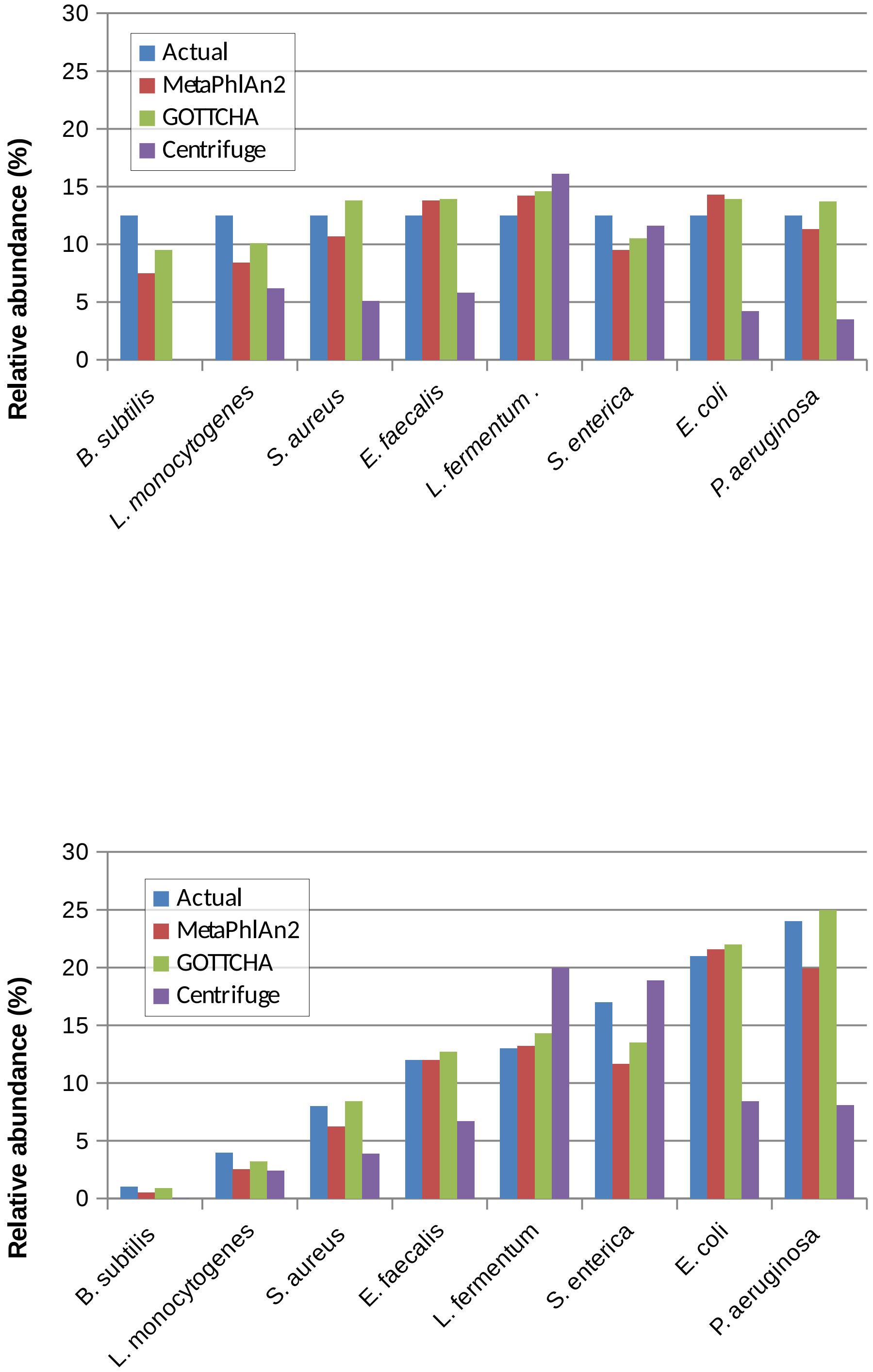
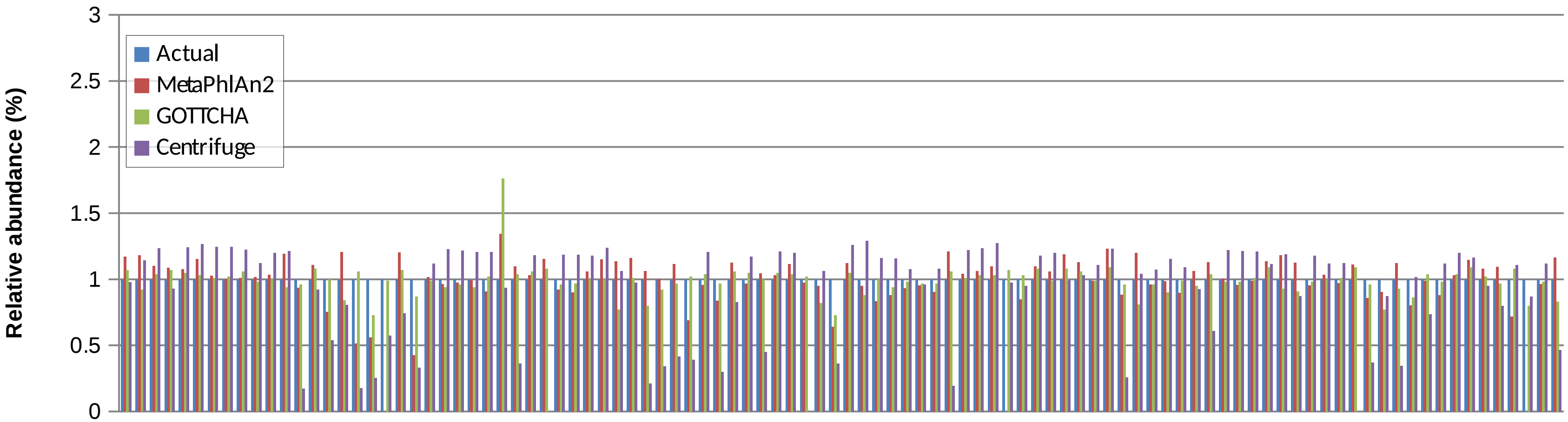


Figure 27: Relative abundance as calculated by three tools (MetaPhlAn2, GOTTCHA and Centrifuge) for the datasets Zymo_8G_10M (A) and Zymo_8G_10M-V (B) compared to the actual abundance and the mean abundance for the three tools.


Relative abundance (%)
3
2.5
2
1.5
1
0.5
0
Actual
MetaPhlAn2
GOTTCHA
Centrifuge

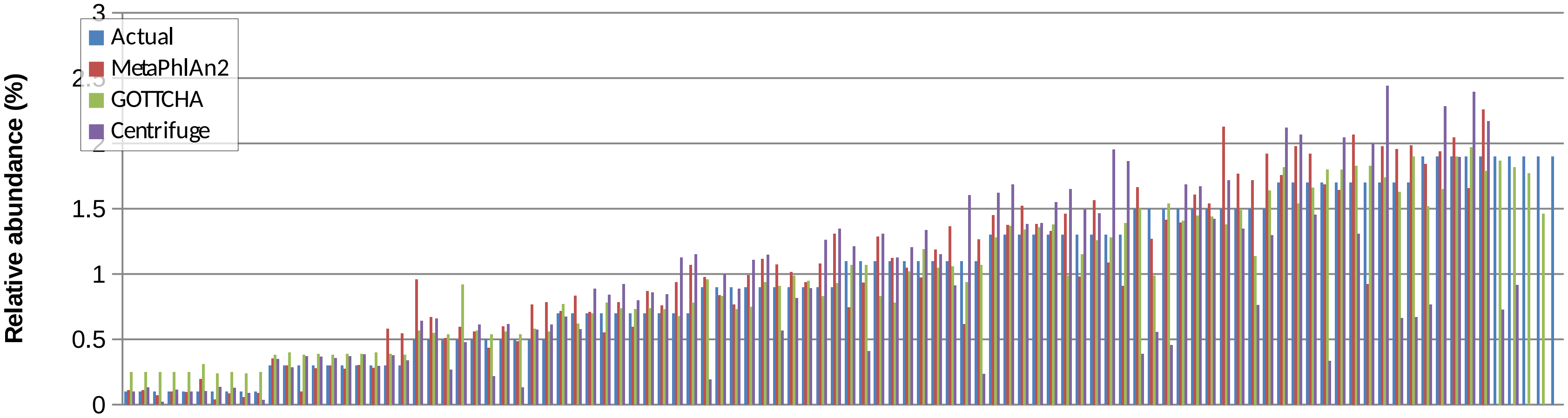


Figure 28: Relative abundance, as calculated by three tools (MetaPhlAn2, GOTTCHA and Centrifuge) for the datasets AM_100G_10M (A) and AM_100G_10M-V (B) compared to the actual abundance and the mean abundance for the three tools. Species details have been removed for clarity but they have been ordered in increasing abundance from left to right (See Appendix 2-A).

The large datasets AM_100G_10M, with a constant abundance, followed the same trends as the smaller datasets (Figure 28A). GOTTCHA was the most accurate tool for describing the abundance of the species with a mean abundance for all 100 species of 0.99% ±0.02. MetaPhlAn2 had a mean abundance of 0.98% ± 0.05 and Centrifuge was the least accurate tool for describing abundance with a mean abundance of 0.92% ±0.07. Of note was over represented abundance of *B. multivorans* by both GOTTCHA (1.76%) and MetaPhlAn2 (1.34%).

Figure 28B displays the general trend for the tools to accurately identify the species of greater abundance within a sample. The mean difference from the known abundance for all 100 species using GOTTCHA was 0.12%, for MetaPhlAn2 it was 0.17% and Centrifuge it was 0.31%. This trend is further demonstrated in Figure 29 where the mean relative abundances, for the 10 species at each of the known abundances, are shown. The mean relative abundance for the three tools closely resembled the actual abundance. The range that the mean relative abundance varied from the actual mean was from 0.01% for the 0.9% cohort of species to 0.2% for the 1.9% cohort.

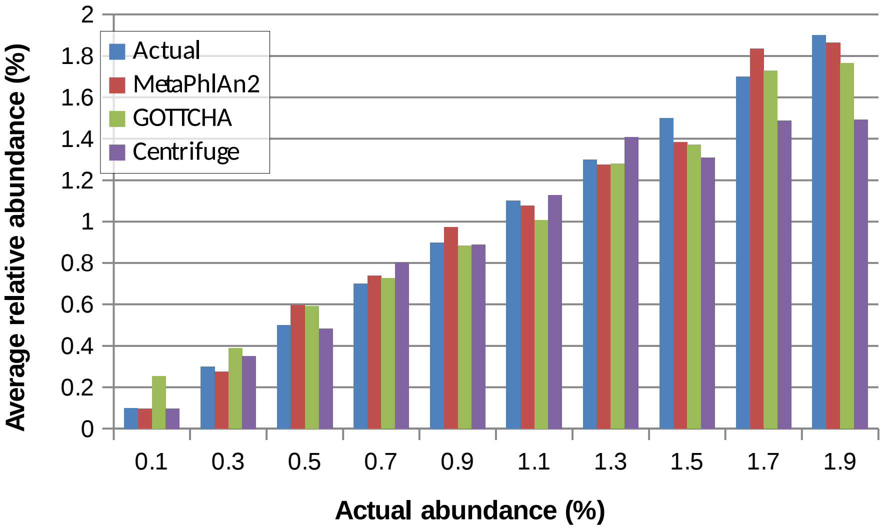


Figure 29: Mean abundance as calculated by three analysis tools (MetaPhlAn2, GOTTCHA and Centrifuge) for the 10 species at each level of abundance in the *in-silico* dataset AM_100G_10M-V (Error bars equal 1 standard deviation).

3.4 Conclusions

This section of work has investigated how the complexity of *in-silico* generated datasets impacts on the accuracy of metagenomic analysis tools for taxonomic identification and estimated relative abundance. Increasing the diversity of the sample to 100 species had minimal impact on the output for MetaPhlAn2 and GOTTCHA when compared to the less diverse datasets. The sensitivity did not drop below 98% and precision did not drop below 97%, which is deemed a high level of accuracy. The sensitivity of Kraken was unaltered, however, its precision increased by roughly 8-fold for species identification and roughly 6-fold for genus level identification. These changes to the accuracy of the tools do not raise any concerns for their use to evaluate the content of bio-aerosol samples. Regarding Kraken, it has been shown here that the accuracy of the tool increases with sample diversity; likely due to the design of the datasets and the false positives generated being closely related to species within the sample.

Altering the abundance of the species within the samples had no impact on the results for any of the tools tested at any taxonomic level. This would imply that 10 million reads is a sufficient number of reads to identify species, even when they only represent 0.03% of the sample (seven species with relative abundance of 0.1% were identified by all tools tested). Future sequencing runs will aim to deliver 10 million reads per sample, which equates to a full MiSeq cartridge for each sample. This is an area that should be investigated in the future so as to ascertain whether it is possible to run multiple samples on a MiSeq cartridge. If the full diversity of a sample can be described through running multiple samples on one cartridge, this would reduce the cost of sequencing dramatically.

For the low diversity datasets, MetaPhlAn2 and GOTTCHA were unaffected by read length. Kraken's sensitivity was also unaffected; however, Kraken's precision was reduced by up to 44%. For the more diverse dataset there were reductions in sensitivity and precision for all tools at both the species and genus taxonomic levels. The most notable reduction was for Kraken's precision which reduced by 25% for genus level identification. Because reducing the read length had a negative impact on all tools this is clearly an important parameter for accurate taxonomic classification from sequence data and should be maximised where possible. This research has not investigated long read sequencing platforms, but these results clearly raise the question as to whether there is a pay off between the accuracy of short read Illumina sequencing and the longer but less accurate reads generated by long read sequencing [112] such as the MinION produced by Oxford Nanopore Technologies[113].

The final aspect of this section looked at the accuracy of the tools to describe the relative abundance of the species within an *in-silico* dataset. GOTTCHA was shown to be the more accurate tool with mean differences from the known abundance of 0.6%, 1.1% and 0.12% for the datasets AM_10G_10M-V, Zymo_8G_10M-V and AM_100G_10M-V respectively. This compares to Centrifuge which showed the lowest level of accuracy and had a mean difference from the known abundance of 4.88%, 6.17% and 0.31%. The lower accuracy for Centrifuge is due to the large number of false positives it reports. These false positives all have an abundance assigned to them so will reduce the accuracy of the abundance for the true positive results.

This section of work has concluded that the three evaluated metagenomic analysis tools are able to accurately identify the organisms, at species and genus level, for *in-silico* datasets at a range of complexities. It has also concluded that the relative abundances assigned to the identified species are an accurate representation of the true abundance.

The next section of work focused on the analysis of real sequence data. An *in-vitro* mock metagenomic community was generated for another project at Dstl and the data was obtained for use on this project. The criterion of 10 million reads and 300bp read lengths was adhered to and the species were used with equal quantities.

4 Analysis of two previously generated *in-vitro* mock community datasets using down-selected metagenomic analysis tools.

4.1 Introduction

To verify the results obtained from analysing *in-silico* datasets it was important to test the down-selected metagenomic analysis tools against 'real' sequencing data. Previous work at Dstl generated two mock metagenomics communities which were sequenced on the Illumina MiSeq platform. These mock communities were developed using DNA extracted from 51 species, with DNA extracted from either *Bacillus anthracis* or *Yersinia pestis* included, all at even abundancies. By using quantified DNA extracted from each species, any bias introduced to a metagenomics pipeline through the DNA extraction process was removed. Non-biased DNA extraction of metagenomic samples is an important area of research that this work has not attempted to investigate[114, 115]. Post sequencing, the reads were quality checked and trimmed to ensure only data of a suitable quality was analysed by the tools.

4.2 Methods

4.2.1 Generating *in-vitro* mock community datasets

Sequence data from a previous project undertaken at Dstl was obtained. In brief, the data was generated by sequencing two mock community mixes using the Illumina MiSeq sequencing platform with Nextera XT library preparation reagents. The *in-vitro* mock community DNA mixes was made using DNA extracts from 52 species (Appendix 4). The DNA was extracted following optimised methods for each organism. The DNA from each organism was then pooled prior to sequencing, see Appendix 4 for the quantities and relative abundances of each species added. The two *in*-vitro mock community datasets varied by just one organism, one contained *B. anthracis* and the other contained *Y. pestis*. These two *in-vitro* mock community metagenomic mixes were processed for and sequenced on an Illumina MiSeq platform. Library preparations were performed using the Nextera XT method generating 2x300bp reads. The two datasets will simply be referred to as:

- Ba [*Bacillus anthracis*]
- Yp [*Yersinia pestis*]

4.2.2 Initial analysis of sequencing data

The data generated from the sequencing run was initially passed through FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to ensure that the run had generated data of a suitable quality based on the 11 matrices measured. Post FastQC, the reads were trimmed based on their quality score using seqtk (https://github.com/lh3/seqtk). The seqtk trimfq command removes bases from the reads with a quality score lower than that defined in the command line. A base with a quality score lower than 20 was removed from the reads using the following command:

- seqtk trimfq –q 0.05 *sample.fq.gz* > *sample-trim.fq*

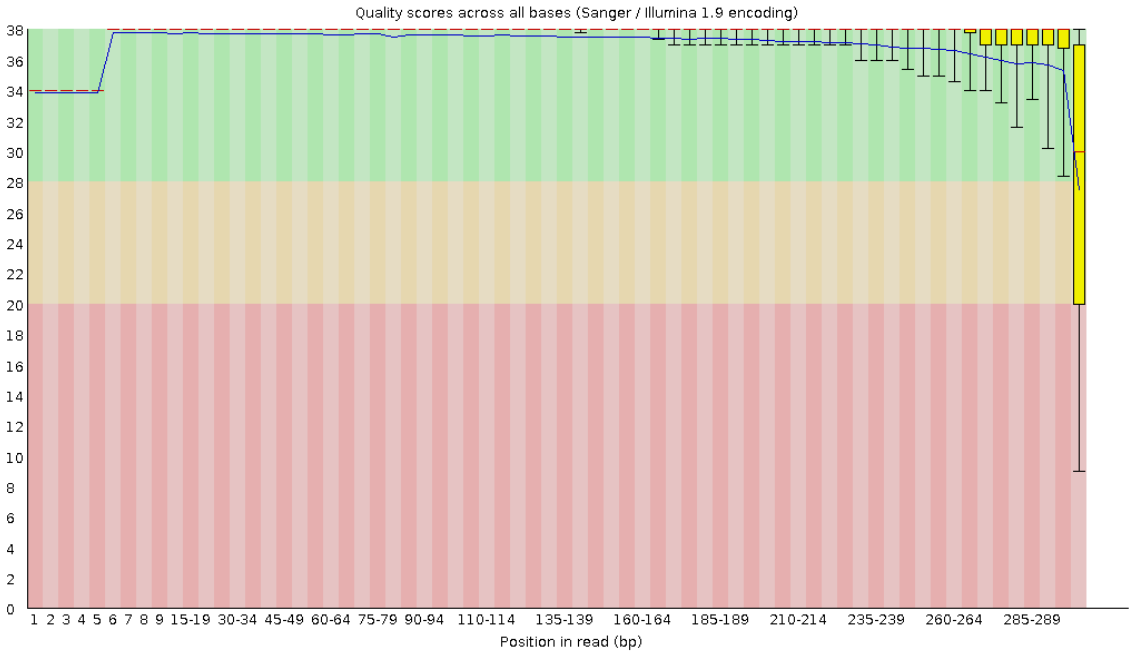


Figure 30: FastQC output file highlighting how the quality score reduces across the length of the read, matching the error profile used to generate the *in-silico* datasets (Figures 2 and 17).

4.2.3 Analysis tools

The analysis tools used to evaluate the more complex datasets were three binning tools (MetaPhlAn2, GOTTCHA-EDGE and Centrifuge). The commands described in Tables 3 were followed for the MetaPhlAn2 and GOTTCHA analysis performed in this section. The command to run Centrifuge was

- centrifuge –p 4 –min-hitlen 250 –x centrifuge-1.0.3-beta/p_compressed+h+v -1 *sample-R1-trim.fq.gz* -2 *sample-R2-trim.fq.gz* –report-file *sample-report* –S *sample.out*

4.2.4 Results interpretation

The results from this analysis will be interpreted as previously described in the simple *in-silico* dataset section (2.2.4).

Additionally, the terms “High Confidence” identification and “Combined” identification will be used to define the output of results detailed in this section. The additional approaches were the outcome of combining the three tools used to analyse the data. For an identification to be defined as High Confidence it had to be identified by all three tools or by both MetaPhlAn2 and GOTTCHA. Organisms identified by Centrifuge and MetaPhlAn2, or Centrifuge and GOTTCHA, were pooled with the High Confidence identifications to produce the Combined identification output.

Taxonomic identification approach	**Combination of output from tools**
High Confidence	MetaPhlAn2 + GOTTHCA
	MetaPhlAn2 + GOTTCHA + Centrifuge
Combined	Centrifuge + MetaPhlAn2
	Centrifuge + GOTTCHA
	High Confidence

Table 5: The incorporation of individual tools into the newly described High Confidence and Combined taxonomic identification approaches.

4.3 Results and discussion

The analysis of the *in-vitro* mock community datasets returned less accurate results compared to all of the *in-silico* datasets previously analysed. There were more false negatives for the *in-vitro* mock community datasets (Figure 28) compared to the *in-silico* datasets, which will impact on the sensitivity of the tools. The lowest sensitivity for species identification across all *in-silico* datasets for all three tools was 96% for MetaPhlAn2's analysis of AM_100G_10M-150. This compares to the highest sensitivity for species identification across the *in-vitro* mock community datasets of 89.3% achieved by Centrifuge's analyses of both Ba and Yp. For genus level identification the lowest sensitivity for the *in-silico* datasets was 95.9%, compared to the highest for *in-vitro* mock community datasets being 92.3%. The precision was more comparable between the *in-silico* and *in-vitro* mock community datasets, although the output from the *in-vitro* mock community datasets was slightly lower.

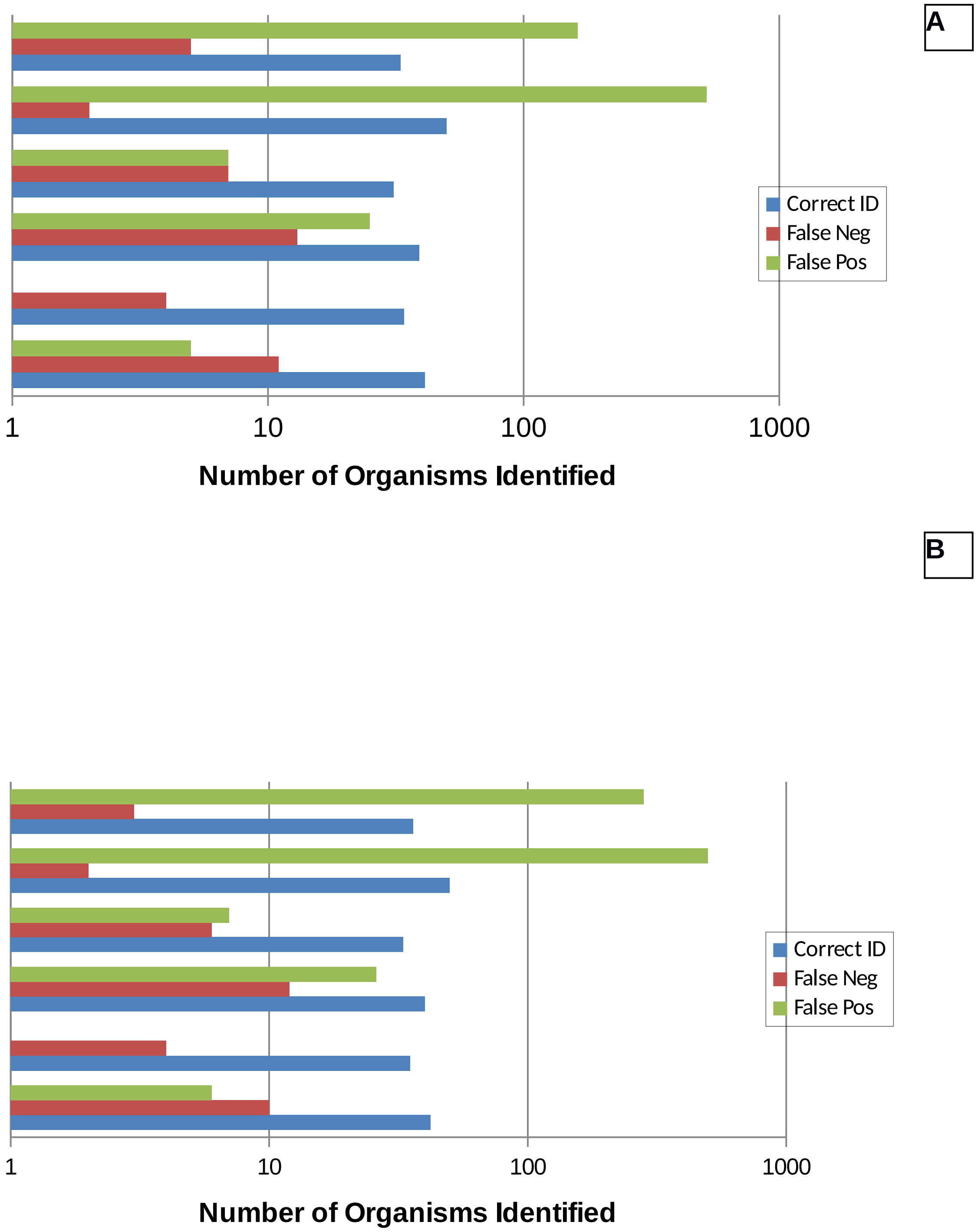


Figure 31: Results from the analysis of the *in-vitro* mock community datasets Ba (A) and Yp (B) using three binning tools (MetaPhlAn2, GOTTCHA and Centrifuge) at the species and genus taxonomic levels.

Due to the lower accuracy for the analysis of the *in-vitro* mock community datasets this work investigated taking the output from the tools with high precision (MetPhlan2 and GOTTCHA) and combining it with the output of the sensitive tool (Centrifuge). If a species or genus were identified by MetaPhlAn2 and GOTTCHA and Centrifuge, or by MetaPhlAn2 and GOTTCHA, then they were deemed a High Confidence identification, whereas, if they were identified by Centrifuge and MetaPhlAn2, or Centrifuge and GOTTCHA, they were pooled with the High Confidence results to produce the Combined identification output (Figures 32-35).

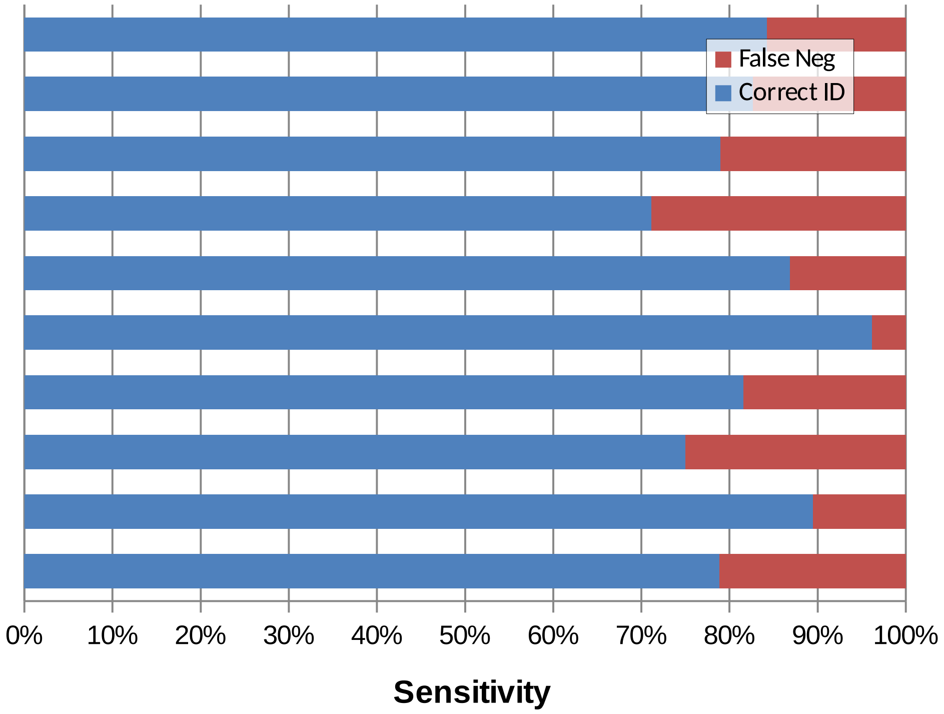


Figure 32: Sensitivity of five identification approaches, including three binning tools (MetaPhlAn2, GOTTCHA and Centrifuge) and two combined approaches (High confidence and combined) for analysis of the *in-vitro* mock community dataset Ba at species and genus level.

The high confidence output for the Ba dataset enabled 71.2% (Figure 32) of the species present in the sample to be identified, and of those species identified 92.5% (Figure 33) were true positives. These results compare favourably to MetaPhlAn2 which showed 78.8% sensitivity but 89.1% precision at species level. GOTTCHA showed 75% sensitivity and 60.9% precision to the Ba dataset at species level. At the genus level the high confidence output showed 78.9% sensitivity of the diversity identified with 96.8% of those genera identified being true positives. This precision is close to that obtained when analysing the *in-silico* datasets.

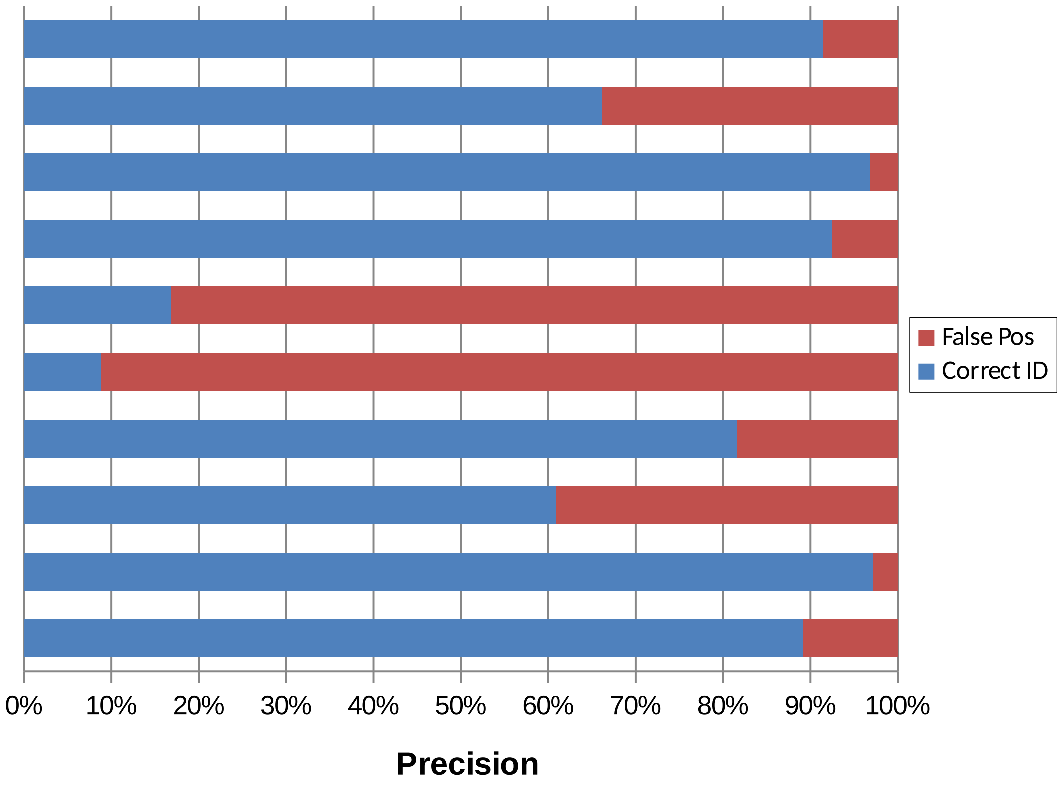


Figure 33: Precision of five identification approaches, including three binning tools (MetaPhlAn2, GOTTCHA and Centrifuge) and two combined approaches (High confidence and combined) for analysis of the *in-vitro* mock community dataset Ba at species and genus level.

The tool with the highest sensitivity for the Yp dataset was Centrifuge which was able to identify 96.2% of the species within the sample (Figure 34). However, only 9.1% of the results were true positives. Using the High Confidence approach, as described for the Ba dataset, the precision of the output got as high as 92.7% for species identification (Figure 35). This approach was more precise than MetaPhlAn2 and GOTTCHA with 87.5% and 60.6% respectively. Although the high confidence approach does not describe the full diversity of the sample, it does increase the confidence that the species identified are in fact true positives.

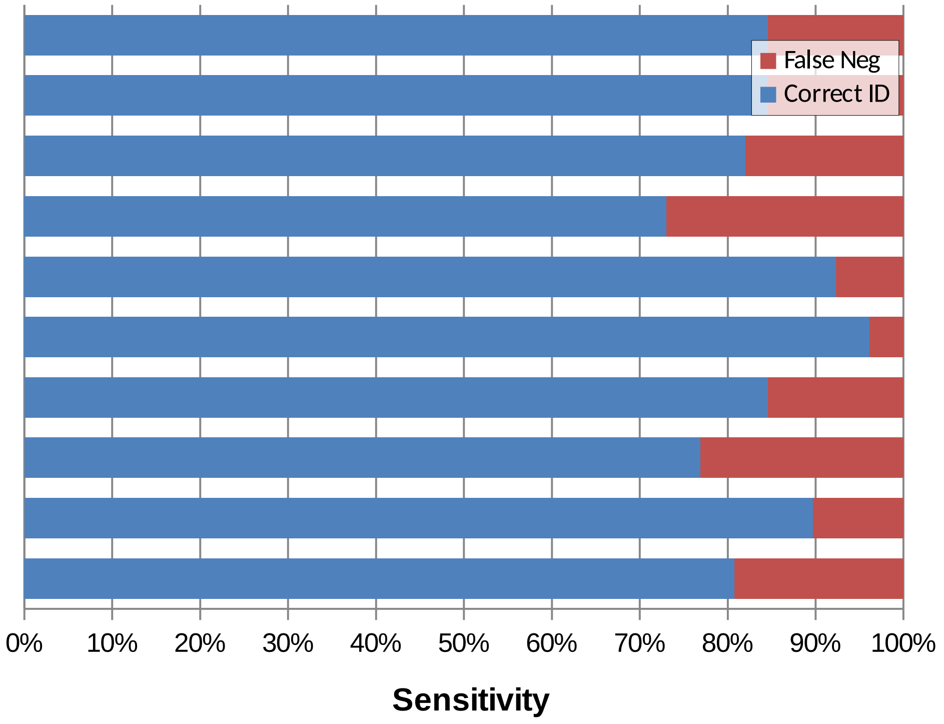


Figure 34: Sensitivity of five identification approaches, including three binning tools (MetaPhlAn2, GOTTCHA and Centrifuge) and two combined approaches (High confidence and combined) for analysis of the *in-vitro* mock community dataset Yp at species and genus level.

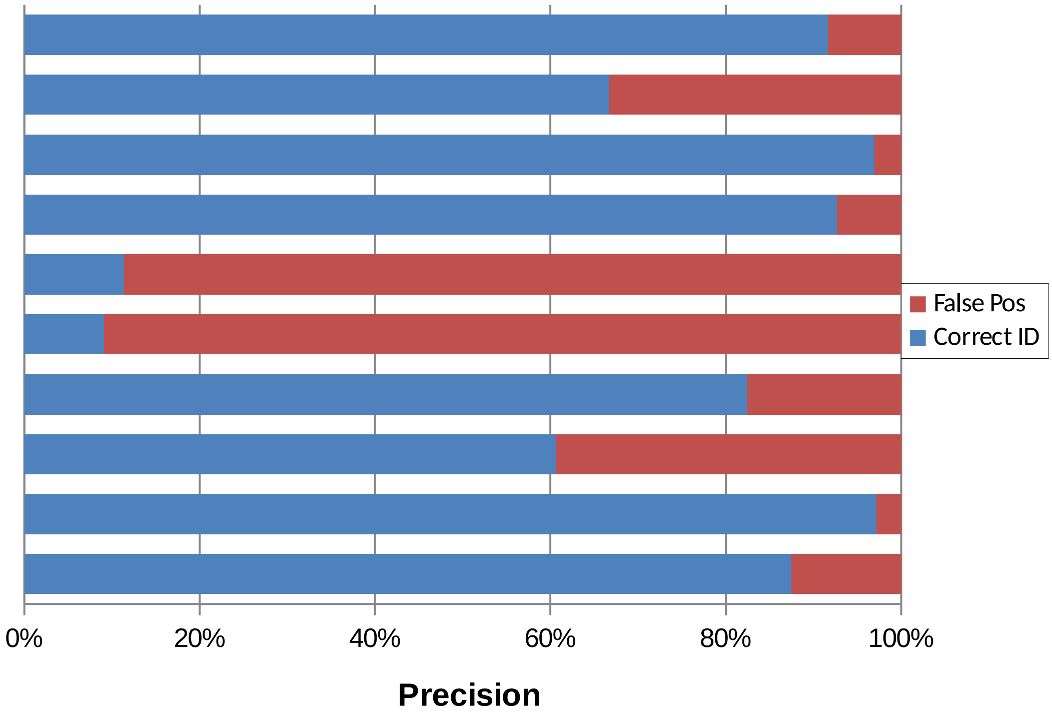


Figure 35: Precision of five identification approaches, including three binning tools (MetaPhlAn2, GOTTCHA and Centrifuge) and two combined approaches (High confidence and combined) for analysis of the *in-vitro* mock community dataset Yp at species and genus level.

4.4 Conclusions

This section of work has shown that the *in-silico* datasets generated for the initial assessment of the bioinformatics analysis tools (Section 2 and 3) did not fully replicate the intricacies of real sequencing data. The *in-silico* datasets were suitable for the initial down-selection of the poorest performing tools. This was shown as the low precision and sensitivity of the weakest tools tested were highlighted with the simple *in-silico* datasets. However, the high accuracy displayed by some of the tools with the *in-silico* datasets were not replicated when analyzing the *in-vitro* mock community datasets. However, using the *in-vitro* mock community dataset would be a suitable dataset to evaluate future tools. The only advantage that *in-silico* datasets provide is the ease at which the parameters can be modified, i.e. read length or abundance.

To fully utilize the potential benefits offered through *in-silico* dataset generation further work is required to enhance the error model that was used to generate the *in-silico* datasets used for this project. Alternatively, other *in-silico* metagenomic dataset simulator software scould be investigated such as CAMSIM[116] or InSilicoSeq[117] [118].

The work from this section has shown the output from an individual tool does not offer the required accuracy when analyzing real sequencing data. Combining the output from three tools increases the precision and does impact the sensitivity. The overall aim of this project is to be able to accurately monitor the variation of the bio-aerosol. Having confidence that the species identified in a sample are true positives is of greater importance than being confident that you have identified all species within the sample. To this end a bioinformatics approach that delivers high precision will be deemed superior to an approach that offers high sensitivity. If the aim of the project was to deliver a bioinformatics approach that was monitoring the bio-aerosol for threat agents then tools that offered high sensitivity would likely be favored over tools with high precision. Improving the confidence in the identification using low precision tools could be accomplishedby aligning the reads back to a reference database of known threat agents. However, this approach would add compute resource and time penalties.

5 Analysis of bio-aerosol samples collected at Dstl.

5.1 Introduction

The final aim of this work was to use the metagenomic analysis pipeline to describe the microbial diversity of bio-aerosol samples. A bio-aerosol collection methodology was developed by Dstl and collaborators to collect bio-aerosols onto dry filters using a SASS 3100 instrument. The SASS 3100 instrument passes air through a dry filter at a rate of 300 litres per minute. The instrument was located at a specific location within the grounds of Dstl and was used to collect samples to answer two initial questions:

1. How does the bacterial diversity of the bio-aerosol change across a long temporal gradient?
2. How does the bacterial diversity of the bio-aerosol change across a short term temporal gradient?

In order to answer these questions with the highest level of granularity, work was undertaken by colleagues to ascertain the required duration of sampling. Collecting samples for too long will not enable the bio-aerosol's variation of diversity to be fully resolved. Collecting samples for too short a period of time will not deliver sufficient DNA to enable successful sequencing. Samples were collected for 0.5, 2 and 4 hours, the DNA was extracted from the filters using the method described in section 5.2.2.1 and quantified using the Qubit high sensitivity double stranded DNA kit. The result of this work identified that a 4 hour sample collection regime delivers sufficient DNA for successful sequencing. However, the quantities were too low for the Nextera XT library preparation kit so the ThruPLEX DNA-seq kit was identified as an alternative. The ThruPLEX DNA-kit can work with an input quantity of between 50 pg and 50 ng of DNA compared to the Nextera XT kit which is designed for an input of 1 ng.

To answer the first question the SASS 3100 collector was used to collect two consecutive 4 hour samples on a monthly basis. Moving forward this will change to a bi-monthly sampling strategy, with this regime planned to continue until March 2020 at the earliest. The second approach was to collect six consecutive 4 hour samples covering a 24 hour period. The analysis of the variation across a short term temporal gradient is planned to be repeated at least annually for the next two years.

An additional challenge to the work described previously in this thesis was how to extract DNA from the filter. An unpublished method was developed by Dstl and one of its collaborators, with thorough research under taken to ensure that the DNA extraction method was effective on Gram positive and Gram negative cells and also vegetative and sporolated cells. This delivered an approach minimising the level of bias that a poorly developed extraction method would introduce to the sequencing of a metagenomic sample.

The extracted DNA was sequenced using the Illumina NextSeq platform, in contrast to the MiSeq platform which had previously been used. The NextSeq generates shorter reads (150bp) compared to the MiSeq (300bp) which has a small impact on the accuracy of the metagenomic analysis tools; data in section 3.3.2 shows a slight drop in sensitivity demonstrated by *in-silico* datasets with shorter read lengths but almost no impact on precision. However, the NextSeq produces a greater number of reads compared to the MiSeq. This enabled multiple samples to be sequenced on each sequencing run, thus reducing the sequencing costs. This financial saving will enable a greater number of samples to be collected and sequenced for the remainder of the project, moving from one to two sampling days each month increasing the statistical validity of the findings of the study.

The sequencing reads were analysed using the pipeline as described in section 4.2.4 and Table 5. The output from MetaPhlAn2, GOTTCHA and Centrifuge was combined to deliver High Confidence and Combined taxonomic identification. These identifications were used to measure how the abundance of species varied over time, how the diversity of the samples changed over time and how the homogeneity of the samples changed. The measure of diversity was based on the Shannon index[119], a widely used measure of diversity in many ecological studies. When the Shannon Index score increases it is a measure that the level of diversity has increased. It is then possible to calculate the homogeneity of the sample using the Shannon index and the maximum possible Shannon score for the sample size. This measure runs from zero, where the sample shows minimum homogeneity, to one, where the sample shows maximum homogeneity. These scoring methods will enable the diversity and homogeneity of the different samples to be compared more accurately.

5.2 Methods

5.2.1.1 Aerosol Collection

The SASS 3100 aerosol collector was used to collect samples onto dry SASS 3100 standard filter cartridges. The collector was secured to a tripod and fixed 1.5 meters above the ground. The collector was inverted at 45º to ensure that no biological material landed on the filter and to avoid the filter getting wet.

The collection location was defined and marked at the start of the collection regime. It was in a remote part of the site in order to avoid it being disturbed by the mass transit of employees and vehicles on site. Towards the end of the trial a gazebo was erected during the collection periods to reduce the chance of inclement weather inhibiting sample collection.

In order to answer the two questions regarding long and short term temporal gradients two collection regimes were deployed:

- Two consecutive 4 hour samples were collected on the 21st of each month at 08:00 – 12:00 and 12:00 – 16:00 from May to August 2018 (this monthly sampling routine will continue for a minimum of two years). Unfortunately, rainfall and high wind inhibited the 21st May 08:00 – 12:00 sample and the 21st July 08:00 – 12:00 sample being collected.
- Six 4 hour samples were collected over a 24 hour period at 00:00 – 04:00, 04:00 – 08:00, 08:00 – 12:00, 12:00 – 16:00, 16:00 – 20:00 and 20:00 – 00:00 (this sampling regime will be repeated at least annually).

Post sample collection, the filters were stored at -80ºC until they could be processed for sequencing.

5.2.1.2 Metadata collection

Alongside the SASS 3100 collectors a suite of meteorological detectors were also deployed. These monitored the air and soil temperature, humidity, wind direction and wind speed, particle number and particle size. Notes were also made of any activities that were taking place around the collection site, such as mowing the grass, low flying aircraft or insects being found on the filter which could have an effect on the DNA collected.

5.2.2 Sequencing of bio-aerosol samples

5.2.2.1 DNA extraction

The DNA extraction of all samples was performed by colleagues at Dstl. The DNA extraction process was developed by Dstl and one of its collaborators and the method is awaiting publication. Further details will be provided in the forthcoming publication, but briefly:

1. The filters were incubated in lysis buffer followed by centrifugation and the supernatant set aside
2. The pellet from step 1 was put through a MetaPolyzyme pre-treatment before centrifugation
3. The pellet from step 2 then went through a lysis, DNA extraction and inhibitor removal process including bead beating
4. The supernatants from step 1 and 3 were combined and magnetic beads used to isolate the DNA.

5.2.2.2 Library preparation

As described in section 5.1 the DNA quantities isolated from the bio-aerosol collection filters waere too low for the Nextera XT kit (used to generate the sequencing libraries in Chapter 4). So the ThruPLEX DNA-seq kit was used following the manufacturer’s guidelines. All library preparation procedures were performed by colleagues at Dstl.

5.2.2.3 Sequencing Platform

As described previously in section 5.1 the NextSeq platform from Illumina was used with a multiplexed approach. Colleagues performed the sequencing using the NextSeq 500/550 High Output Kit v2.5 (300 cycles) with a dual indexed workflow following the manufacturer's guidelines.

5.2.3 Initial analysis of sequencing data

As described previously in section 4.2.2, the sequence data was analysed using FASTQC to ensure the quality of the run was suitable for further analysis. The reads were then trimmed using the seqtk trimfq command.

5.2.4 Sequence analysis tools

The metagenomic analysis tools used to evaluate the samples were the same as those described in section 4.2.3. MetaPhlAn2, GOTTCHA and Centrifuge were used, running the previously described methods. The only modification to the Centrifuge command was changing the minimum hit length parameter (min-hitlen) from 250 to 100. This was to ensure that the shorter reads generated using the NextSeq platform were evaluated.

5.2.5 Interpretation of sequence analysis

As these samples are true unknowns it is impossible to measure the precision or sensitivity of the tools as previously described. The High Confidence and Combined output (described in section 4.2.4) were used to describe the bacterial content of the samples to species level.

The diversity and homogeneity of the samples were calculated using the Shannon index ($H^{'}$) and relative diversity ($J^{'}$). The Shannon index gives a measure of diversity for the sample; as the Shannon index score increases so does the sample diversity. The equation to calculate the Shannon Index is shown below (www.real-statistics.com/descriptive-statistics/diversity-indices/shannons-diversity-index/describes the Shannon index), where *k* is the number of species and p_i is the proportion of observations in the i th of k categories.

Shannon Index

$$H^{'}=-\sum_{i=1}^{k} p_i \log p_i$$

The relative diversity ($J^{'}$) is a measure of evenness and is calculated by working out the proportion the Shannon index is of its maximum. In order to calculate $J^{'}$ the maximum Shannon index ($H^{'}_{max}$) must first be calculated. $H^{'}_{max}$ describes a position where all species are evenly observed and so p_i is 1.

Maximum Shannon Index

$$H^{'}_{max}=-\sum_{i=1}^{k} p_1 \log p_1=-k\, p_1 \log p_1=-\log p_1=-\log\left(\frac{1}{k}\right)=\log k$$

It is then a simple case of performing the following equation to calculate the relative diversity:

Relative diversity

$$J^{'}=\frac{H^{'}}{\log k}$$

5.3 Results and Discussion

5.3.1 Analysis of samples from the monthly collection regime

The results from the High Confidence approach are shown in Appendix 5-A. Due to inclement weather conditions two samples were not able to be collected (21/05/2018 0800-1200 and 21/07/2018 0800-1200). 27 different species were identified across the four months of sampling and had a range in relative abundance from 0.53 – 12.4%. Of the 27 species identified, 41% of the species were identified in just one sample and 7% of the species were identified in all 6 samples. The two species that were present in all samples were *Stenontrophomonas maltophilia* (which ranged in abundance from 0.53% - 5.83%) and *Clavibacter michiganensis* (with a range in abundance from 1.59% - 6.15%). The species which had the highest abundance was *Lactobacillus amylovorus* with 12.4% and the species with the lowest abundance across the 4 months was *Terriglobus roseus* with an abundance of 0.53%.

Table 6 shows the different number of species identified by the different taxonomic identification methods for each of the samples. The High Confidence method continuously reports fewer species compared to the other approaches, whereas the Combined approach and GOTTCHA report a similar number of species for each sample. As seen with the *in-silico* datasets Centrifuge reported a far higher number of species for each of the long term temporal gradient samples. The results from the *in-silico* datasets would imply that the vast majority of these Centrifuge results are false positives. The sample with the highest number of species, as identified by all tools, was collected on 21st June between 12:00 and 16:00. There was also a good trend between the numbers of species identified for each sample across all tools.

	21/05/2018		21/06/2018		21/07/2018		21/08/2018	
	0800-1200	**1200-1600**	**0800-1200**	**1200-1600**	**0800-1200**	**1200-1600**	**0800-1200**	**1200-1600**
High Conf	-	7	9	19	-	10	13	6
Combined	-	32	28	61	-	37	52	36
MetaPhlAn 2	-	17	14	41	-	27	37	24
GOTTCHA	-	30	29	60	-	34	48	36
Centrifuge	-	1120	1064	1385	-	1210	1274	1176

Table 6: Number of species identified by the different taxonomic identification strategies for the samples collected on the long term temporal study.

The number of species identified is not a true measure of diversity and so the Shannon index was used to measure the difference in diversity across the six samples collected across the four month period (Figure 36). Due to the large number of False Positives Centrifuge generates it will have a biased measure of diversity. This is represented in Centrifuge's high Shannon Index score. Surprisingly the Shannon Index for GOTTCHA was similarly high to Centrifuge which would suggest that it also reports a higher than authentic level of diversity within the sample. The Shannon Index score for the High Confidence approach was similar to MetaPhlAn2's score. This suggests the diversity of the samples were at the lower end of the spectrum of the tools tested. The level of diversity ranged from 0.687 – 1.105 for the High Confidence approach, 1.343 – 1.655 for the combined approach, 1.169-1.799 for Centrifuge, 0.877-1.158 for MetaPhlAn2 and 1.459-1.766 for GOTTCHA. The highest measures of diversity were for June 21st 1200-1600 and August 21st 0800-1200 however the general trend was for the diversity to increase through year. The High Confidence approach had a Shannon index score of 0.803 in May compared to 1.024 in August resulting in an R^2 score of 0.918 with a p-value of 0.042. Likewise Centrifuge had a Shannon Index score of 1.169 in May compared with 1.799 in August giving an R^2 score of 0.927 with a p-value of 0.37. These variations in diversity need to be measured against the metadata that was also collected at the time of sampling to see if any correlations can be made between weather conditions and diversity.

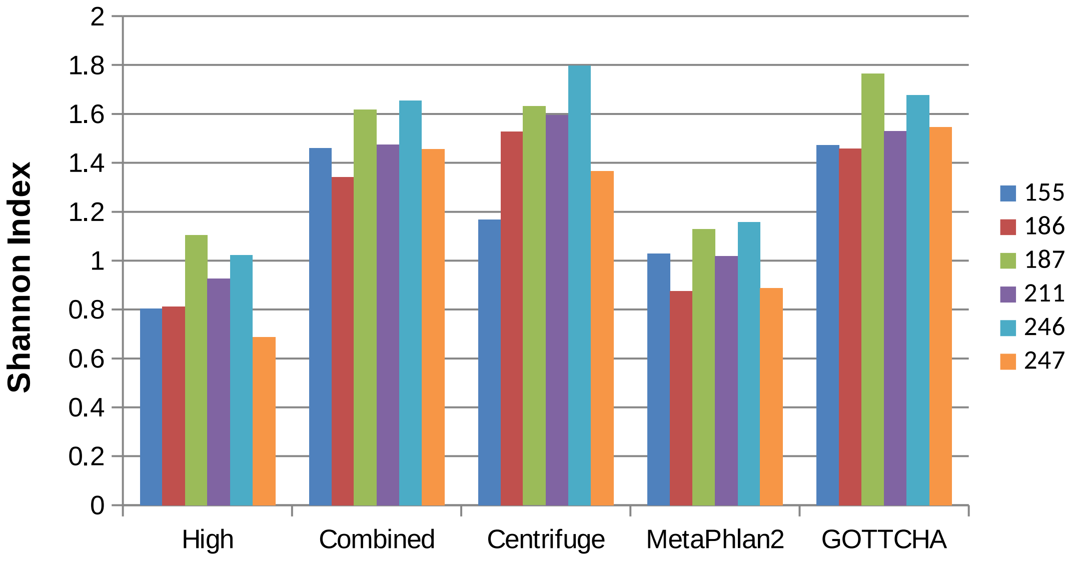


Figure 36: Shannon index for the taxonomic identification strategies run on the samples collected on the long term temporal study. The legend relates to sample numbers (155 = May 21st 1200-1600, 186 = June 21st 0800-1200, 187 = June 21st 1200-1600, 211 = July 21st 1200-1600, 246 = August 21st 0800-1200 and 247 = August 21st 1200-1600).

The relative diversity of the samples is a measure of homogeneity and runs on a scale between 0 and 1. A score of 1 represents a sample with a completely homogenous group of species present. Figure 37 shows the range of relative diversity scores for all of the taxonomic approaches used. Again, due to the high level of false positives that Centrifuge is known to report its relative diversity is likely to be skewed towards a more uneven distribution and as the results show Centrifuge did have the lowest relative diversity scores. The lowest score was 0.494 for the sample taken 1200-1600 on the 21st August. In fact the five lowest relative abundance scores were from Centrifuge's output. The taxonomic identification approach with the next lowest relative abundance was MetaPhlAn2, with 5 of the 6 samples falling within Centrifuges range of Relative abundances (0.644 and 0.765). This result suggests that MetaPhlAn2 also over estimates the relative diversity of a sample. At the other end of the scale GOTTCHA's relative abundance score for all six of the samples was 0.993 or greater. This implies that all of the species detected across all six samples have near identical levels of diversity which is very unlikely. With the High Confidence output generating relative abundance scores between those of

Centrifuge, MetaPhlAn2 and GOTTCHA it would suggest that the output is potentially more accurate.

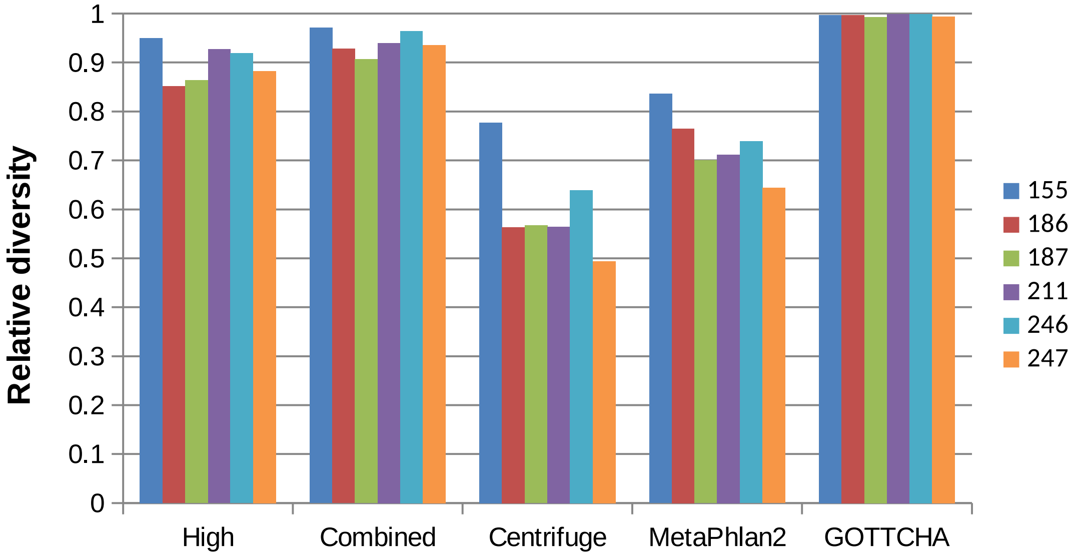


Figure 37: Relative diversity for the different taxonomic identification strategies run on the samples collected for the long term temporal study. The legend relates to sample numbers (155 = May 21st 1200-1600, 186 = June 21st 0800-1200, 187 = June 21st 1200-1600, 211 = July 21st 1200-1600, 246 = August 21st 0800-1200 and 247 = August 21st 1200-1600).

5.3.2 Analysis of samples from the 24 hour collection regime

The results for the analysis of the samples collected over a 24 hour period using the High Confidence taxonomic identification approach are shown in Appendix 5-B. Over the 24 hours 34 different species were identified with a range of abundancies from 0.14% - 19.14%. Of the 34 species identified 26% were identified in only one sample, whereas 15% of the species were identified in all six samples. The species that were present in all samples were *Lactobacillus amylovorus,* with a relative abundance ranging from 3.28 – 19.14%, *Lactobacillus reuteri* which ranged in abundance from 1.68 – 3.39%, S. maltophilia with a range of abundances from 0.53 – 3.18%, *Brachybacterium* *faecium* ranging from 0.39 – 2.94% and *C. michiganensis* with a range of relative abundance from 0.64 – 2.11%. The species with the highest abundance across the 24 hour period was *L. amylovorus* at 19.14% for a single sample. The species with the lowest abundance was *Bacteroides salanitronis* with an abundance of 0.14%.

Table 7 shows the number of species identified for each of the samples collected across the 24 hour period. The High Confidence approach returned the lowest number of species across all samples, with Centrifuge showing the largest number of species. As with the long term temporal study GOTTCHA and the combined approach returned a similar Shannon Index score for all samples. Surprisingly, the sample which showed the greatest number of species identified was the 2000 – 0000 sample. A review of the Metadata associated with this collection is required to understand the reasons behind this increase in diversity. The samples with the lowest variety of species were collected between 0400 and 1200. There was also agreement between the different numbers of species identified across the samples for the different taxonomic identification approaches.

	0000-0400	**0400-0800**	**0800-1200**	**1200-1600**	**1600-2000**	**2000-0000**
High Conf.	15	13	9	19	17	23
Combined	57	38	28	61	72	96
MetaPhlAn2	34	27	14	41	42	60
GOTTCHA	54	38	29	60	69	87
Centrifuge	1242	1082	1064	1385	1476	1599

Table 7: Number of species identified by the different taxonomic identification strategies for the samples collected on the short term temporal study.

As seen for the long term temporal study, the Shannon Index for the High Confidence taxonomic approach mirrored that of MetaPhlAn2 analysis of the short term temporal study (Figure 38). The High Confidence approach generated a range of Shannon Index scores of 0.812 – 1.107 and MetaPhlAn2 had a range from 0.877 – 1.275. The Combined taxonomic identification approach and Centrifuge also had similar levels of diversity with Shannon index scores ranging from 1.343 – 1.619 and 1.215 – 1.665 respectively. On this occasion GOTTCHA showed the highest level of diversity with the highest

Shannon Index scores ranging from 1.459 – 4.866. The general trend for the diversity across the 24 hours for all analysis approaches showed a drop in abundance for the 0800 – 1200 collection. This was followed by a more subtle drop in the Shannon Index for the collection from 1600 – 2000. These timings correlate with the majority of staff entering and leaving the site but further investigation of the metadata would be needed to confirm this hypothesis. Any trends will need to be confirmed with more repeats of this short term temporal study.

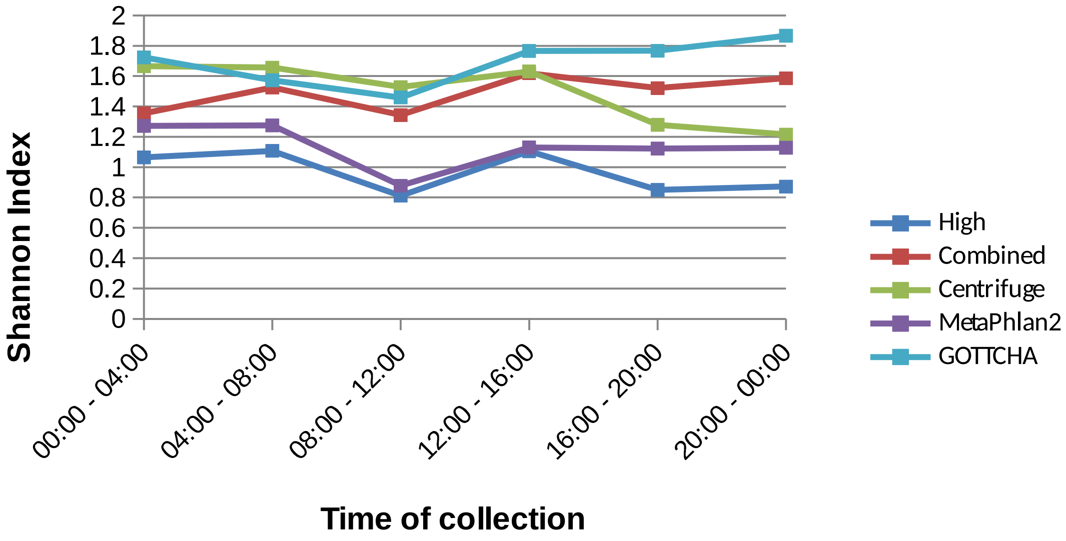


Figure 38: Shannon index for the taxonomic identification strategies run on the samples collected on the short term temporal study.

The Relative abundance results for the short term temporal study are similar to the long term study results. GOTTCHA generates high scores, between 0.961 and 0.997, which would suggest there is no variation in the abundance of the species present in the samples. The relative index scores for Centrifuge are the lowest of all approaches tested (ranging from 0.412 – 0.601). This low level of homogeneity is likely skewed by the large numbers of false positives that Centrifuge has been shown to report. The High Confidence approach shows a Relative abundance score range of 0.641 – 0.964. These results are between the unlikely scores from GOTTCHA and the negatively biased Centrifuge suggesting they are a more accurate representation of the homogeneity of the species within the samples. The trend for the Relative Diversity for all analysis approaches was for the score to decrease over the 24 hours. For example

GOTTCHA dropped from 0.995 – 0.962, Centrifuge reduced from 0.592 – 0.412 and the High Confidence approach went from 0.905 – 0.635 over the 24 hour period (Figure 39). Time will need to be spent trying to correlate these results with changes in the meteorological conditions. The short term temporal study also needs to be repeated to see if this is a general trend or if it is a one off phenomenon.

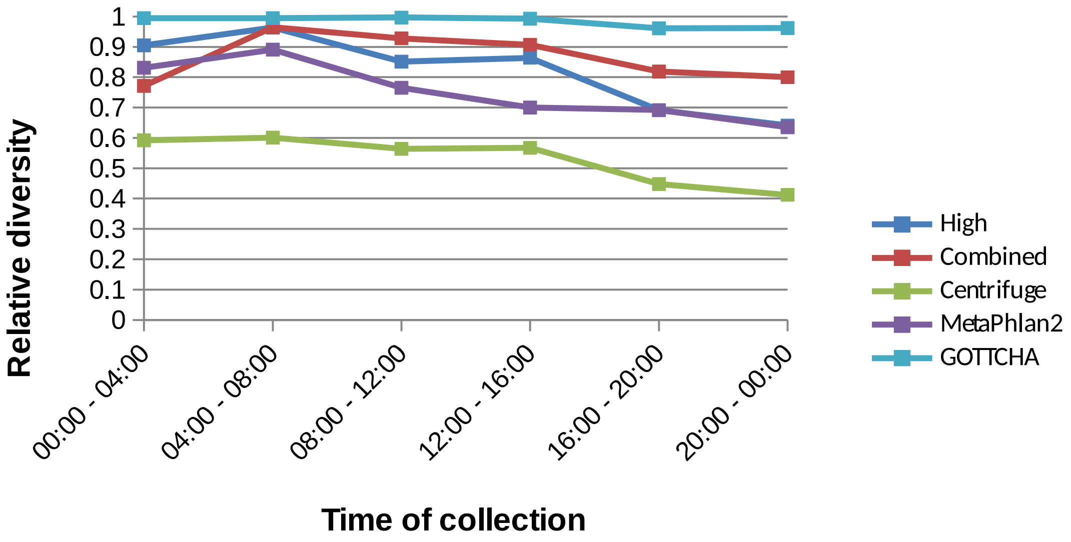


Figure 39: Relative diversity for the different taxonomic identification strategies run on the samples collected for the short term temporal study.

5.4 Conclusions

The analysis of the long term temporal study samples showed that the High Confidence taxonomic identification results accurately describe the diversity of the samples. This finding was confirmed with the analysis of the short term temporal study samples. This summation was based on the fact that Centrifuge has been shown in this study to generate a large number of false positives which inflate the Shannon index score, inaccurately describing the level of diversity within the sample. The Shannon Index score for GOTTCHA was similarly high to Centrifuge suggesting that GOTTCHA also over estimates the level of diversity within a sample. MetaPhlAn2 and the High Confidence approach both returned Shannon Index scores lower than Centrifuge or GOTTCHA thus implying they more accurately describe the level of abundance within a sample.

The high number of false positives also negatively impact Centrifuge's ability to accurately predict the homogeneity of a sample. The results for the long and short term temporal studies show that Centrifuge gave the highest level of unevenness to all of the samples, confirming this predicted bias. MetaPhlAn2 gave measures of relative diversity similar to Centrifuge for the long and short term temporal studies. These results bring into question MetaPhlAn2's ability to accurately describe the homogeneity of the samples. The results for GOTTCHA's analysis of the long and short term temporal studies show a very high level of relative diversity (0.961 – 0.997) suggesting that all the species identified are evenly distributed within the sample. This is a very unlikely outcome, and therefore calls into question GOTTCHA's ability to accurately define diversity. The High Confidence results fall between these extremes shown from Centrifuge and GOTTCHA, suggesting it is able to more accurately describe the homogeneity of the samples.

The results from the previous sections show that the High Confidence taxonomic identification approach offers the most accurate method for identifying the species present in a sample. The results shown in this section now demonstrate that the High Confidence approach is also the best method for

accurately describing the diversity and homogeneity of the sample compared to running the tools individually.
From the work undertaken within this study, based on three different measurements using the output from the High Confidence taxonomic identification approach, there was a greater diversity of species within the 24 hour study compared to the long term temporal study. More species were identified across the six samples taken across 24 hours (n=34) compared to the six samples collected over 4 months (n=27). The mean Shannon index for the long term temporal study was 0.893 compared to 0.968 for the short term study. Also, the homogeneity was greater in the 4 month study with a mean relative diversity score of 0.899 compared to 0.819 for the 24 hour study. These findings need to be confirmed with further repeats of the short term temporal study, which are planned to take place until at least March 2020. Ultimately, these findings show that there is a requirement to monitor the bio-aerosol across the full 24 hour period in order to fully understand the variation rather than just sampling the same time through the year.

It is now imperative that the meteorological data, collected alongside the aerosol collections, are analysed. Accurately identifying meteorological conditions which cause differences in species abundance in the aerosol microbiome are now achievable due to this work.

6 Final Conclusions and future work.

This work has confirmed that there is variance within the accuracy in the available metagenomic analysis tools. The tools which performed well in this study were able to demonstrate 100% accuracy at the species level for simple *in-silico* datasets. The poorer performing tools showed much lower levels of accuracy, with precision as low as 1.3%, for the same simple *in-silico* datasets.

In-silico datasets are a suitable mechanism for performing initial down-selection of metagenomic analysis tools. Using the simple *in-silico* datasets 10 binning tools and 4 assembly tools were able to be down-selected to just 3 binning tools and 1 assembly tool. The results from analysing the simple *in-silico* datasets discussed in this thesis clarify the results published in the literature, therefore providing confidence in the down-selection of the tools.

Increasing the complexity of the *in-silico* dataset, by increasing the diversity of species present, only had a minor effect on accuracies. The only major difference between the smaller datasets, AM_10G_10M and Zymo_8G_10M, and the larger dataset, AM_100G_10M, was a near 10-fold increase in the precision of Kraken. This increase in precision for Kraken is likely due to the miss identified taxa actually being present in the sample. These are important findings and give confidence that the accuracy of species identification will not differ if the bio-aerosol complexity changes through the year.

Reducing the read length of the *in-silico* datasets from 300bp to 150bp had an impact on the accuracy of the metagenomic tools. This reduction in read length was more substantial than any of the other variables tested for the large dataset. There was a drop in sensitivity for all three tools, but only a minimal reduction in precision. However, for the smaller *in-silico* datasets the read length had no impact on the accuracy of the tools. The reduction in accuracy caused by reduced read lengths is because the shorter read lengths are less likely to contain the unique sections of the genomes required for successful taxonomic identification to the species level.

Altering the abundance of organisms represented in the *in-silico* datasets made little difference to the accuracy of the analysis tools. The sensitivity and the precision of the tools remained similar with the varied abundance datasets compared to the datasets with even abundance. These results suggest that species with low abundance will still be identified from bio-aerosol samples; to confirm this assumption future work using *in-vitro* mock community samples of varied abundance is required. Identification of low abundant species is positive as bio-aerosols are thought to have a low biological content. However, this increases the need to fully understand and control the impact reagent contaminants will have on the results.

Analysing the tools ability to accurately predict the relative abundances of the species present was successful. All three tools were able to predict the abundance of the species in the even and varied abundance *in-silico* datasets. GOTTCHA was shown to be the tool that most accurately estimated the abundance of the species in the *in-silico* datasets, with Centrifuge showing the lowest level of accuracy. The ability to accurately estimate the abundance of species is an essential element of monitoring changes in the diversity of the bio-aerosol so it was imperative that this aspect of the analysis was a success.

Generating and analysing real sequencing data gave the best measure of a tool's accuracy. The accuracy of all tools tested was impacted negatively when using *in-vitro* mock community datasets compared to *in-silico* datasets. Due to the cost benefits of using an *in-silico* approach and the ease in analysing different variables, future time could be invested in investigating different metagenomic simulation tools to devise more realistic datasets for use as initial down-selection analysis. This approach could then be followed with the analysis of a well characterised *in-vitro* mock community sample for the finer selection of any new tools of interest. Other areas of interest would be to investigate the tool's abilities to accurately describe the abundance of the organisms present within an *in-vitro* mock community sample.

Combining the output from the three different metagenomic analysis tools increased the accuracy compared to using the tools independently. This approach reduced the number of false positives and therefore increases confidence that the species identified in unknown samples are in fact true positives. This is critical for Dstl as it will allow an accurate analysis of bio-aerosols in military relevant environments, enabling future bio-detection platforms to be evaluated in a well-defined, mock bio-aerosol environment.

The culmination of this work produced an analysis pipeline which delivered accurate taxonomic identification and relative abundance prediction of species. This pipeline was used, and will continue to be used, for the analysis of real bio-aerosol samples collected as part of a long term and a short term temporal gradient study designed to measure the variation in species abundance in bio-aerosol samples. The results from the real data analysis show that the High Confidence taxonomic identification approach delivers the most accurate prediction of diversity and homogeneity of all the analysis tools tested.

The application of the High Confidence taxonomic identification pipeline has led to changes to the sampling regime at Dstl. Due to higher levels of variation measured outside of 0800 – 1600 there will be an increase in the number of short term, 24 hour temporal studies through to March 2020. This will enable a measure of the full diversity and homogeneity of the bio-aerosol.

The output of this work is a metagenomic taxonomic identification pipeline that is already employed on a bio-aerosol analysis project at Dstl. The aim of the project is to measure the temporal and geographical variation of the bio-aerosol. In order to understand any variation it is imperative that a well characterised bioinformatics approach is used. This work has delivered that well characterised analysis pipeline utilising tools selected on their performance. The implementation of the pipeline developed through this research will have a direct impact on the evaluation of future and current biological warfare detection platforms, leading to the improved safety for military personnel.

7 References

1. Vijayaraghavan, D.T.R., *Biological warfare agents.* Pharmacy & BioAllied Sciences, 2010. **2**(3): p. 179-188.
2. Hughson, *Health Effects of Indoor Fungal Bioaerosol Exposure.* Applied Occupational and Environmental Hygiene, 2010. **18**(7): p. 535-544.
3. Blachere, F., *Bioaerosol sampling for the detection of aerosolized influenza virus.* Influenza and other Respiratory Viruses, 2007. **1**(3).
4. Li, Y., *Role of air distribution in SARS transmission during the largest nosocomial outbreak in Hong Kong.* Indoor Air, 2004. **15**(2).
5. Gard, E., *Bioaerosol Mass Spectrometry for Rapid Detection of Individual Airborne Mycobacterium tuberculosis H37Ra Particles.* Applied and Environmental Microbiology, 2005.
6. Cox, C., *Bioaerosol handbook.* 1995.
7. Dowd, S., *Bioaerosol Transport Modeling and Risk Assessment in Relation to Biosolid Placement.* Journal of Environmental Quality Abstract, 2000. **29**(1): p. 343-348.
8. Pillai, S.D. and S.C. Ricke, *Review / SynthèseBioaerosols from municipal and animal wastes: background and contemporary issues.* Canadian Journal of Microbiology, 2002. **48**(8): p. 681-696.
9. Prussin, A.J. and L.C. Marr, *Sources of airborne microorganisms in the built environment.* Microbiome, 2015. **3**(1): p. 78.
10. Lighthart, B., *Distribution of Microbial Bioaerosols.* Springer, 1994.
11. O'Brien, K.M., *High throughput genomic sequencing of bioaerosols in broiler chicken production facilities.* Microbial Biotechnology, 2016. **9**(6).
12. Oliver, J., *The Viable but Nonculturable State in Bacteria.* The Journal of Microbiology, 2005. **43**(S): p. 93-100.
13. Xihong Zhao, et al., *Current Perspectives on Viable but Non-culturable State in Foodborne Pathogens.* frontiers in Microbiology, 2017. **8**(580).
14. Li, L., et al., *The importance of the viable but non-culturable state in human bacterial pathogens.* frontiers in Microbiology, 2014. **5**(258).
15. Chi, M.-C. and C.-S. Li, *Analysis of Bioaerosols from Chicken Houses by Culture and Non-Culture Method.* Aerosol Science and Technology, 2006. **40**(12): p. 1071-1079.

16. Chiodini, R., *Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples.* Gut Pathogens, 2016. **8**(24).

17. Lindgreen, S., *An evaluation of the accuracy and speed of metagenome analysis tools.* nature scientific reports, 2016. **6**(19233).

18. Petro, j., *Biotechnology: Impact on Biological Warfare and Biodefense.* Biosecurity and Bioterrorism, 2004. **1**(3).

19. Roffey, R., *Biological warfare in a historical perspective.* Clinical Microbiology and Infection, 2002. **8**(8): p. 450-454.

20. Roffey, R., *Biological weapons and bioterrorism preparedness: importance of public-health awareness and international cooperation.* Clinical Microbiology and Infection, 2002. **8**(8).

21. Rasko, D.A., et al., *Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation.* Proceedings of the National Academy of Sciences, 2011. **108**(12): p. 5027-5032.

22. Keim, P., et al., *Molecular Investigation of the Aum Shinrikyo Anthrax Release in Kameido, Japan.* Journal of Clinical Microbiology, 2001. **39**(12): p. 4566-4567.

23. Oulas, A., et al., *Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies.* Bioinformatics and Biology Insights, 2015. **9**: p. 75-88.

24. Turnbaugh, P.J., et al., *The Human Microbiome Project.* Nature, 2007. **449**(7164): p. 804-810.

25. Inskeep, W.P., et al., *The YNP metagenome project: environmental parameters responsible for microbial distribution in the Yellowstone geothermal ecosystem.* Frontiers in Microbiology, 2013. **4**.

26. Hurwitz, B.L. and M.B. Sullivan, *The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology.* Plos One, 2013. **8**(2).

27. Holzinger, M., *Nanomaterials for biosensing applications: a review.* Frontiers in Chemistry, 2014.

28. Mukoda, T., L. Todd, and M. Sobsey, *PCR and gene probes for detecting bioaerosols.* Journal of aerosol science, 1994. **25**(8): p. 1523-1532.

29. Speight, S.E., et al., *Enzyme-linked immunosorbent assay for the detection of airborne microorganisms used in biotechnology.* Journal of aerosol science, 1997. **28**(9): p. 483-492.

30. G, R., P. P, and S. C, *Bacterial aerosol emission from wastewater treatment plants:Culture methods and bio-molecular tools.* Aerobiologia, 2000. **16**(1): p. 39-46.

31. Theodore, J., *Medical consequences of biological warfare: The Ten Commandments of Management.* Military Medicine, 2001. **166**(12): p. 2-11.

32. Paull, J., *Heat Strain and Heat Stress for Workers Wearing Protective Suits at a Hazardous Waste Site.* American Industrial Hygiene Association Journal 1987. **48**(5): p. 458-463.

33. Sindhuja Sankaran, A.M., Reza Ehsani, Cristina Davis,, *A review of advanced techniques for detecting plant diseases.* Computers and Electronics in Agriculture, 2010. **72**(1): p. 1-13.

34. Durso, L., *Distribution and Quantification of Antibiotic Resistant Genes and Bacteria across Agricultural and Non-Agricultural Metagenomes.* plos one, 2012.

35. Treangen, T.J., et al., *MetAMOS: a modular and open source metagenomic assembly and analysis pipeline.* Genome Biology, 2013. **14**(1): p. 20.

36. Truong, D.T., et al., *MetaPhlAn2 for enhanced metagenomic taxonomic profiling.* Nature Methods, 2015. **12**(10): p. 902-903.

37. Segata, N., et al., *Metagenomic microbial community profiling using unique clade-specific marker genes.* Nature Methods, 2012. **9**(8): p. 811-+.

38. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature Methods, 2012. **9**(4): p. 357-U54.

39. Alneberg, J., et al., *Binning metagenomic contigs by coverage and composition.* Nature Methods, 2014. **11**(11): p. 1144-1146.

40. Boisvert, S., et al., *Ray Meta: scalable de novo metagenome assembly and profiling.* Genome Biology, 2012. **13**(12).

41. Stephen F. Altschul, et al., *Basic Local Alignment Search Tool.* J Mol Biol., 1990. **215**(3): p. 403-410.

42. Markowitz, V.M., et al., *IMG/M 4 version of the integrated metagenome comparative analysis system.* Nucleic Acids Research, 2014. **42**(D1): p. D568-D573.

43. Markowitz, V.M., et al., *IMG/M: the integrated metagenome data management and comparative analysis system.* Nucleic Acids Research, 2012. **40**(D1): p. D123-D129.

44. Zhu, W., A. Lomsadze, and M. Borodovsky, *Ab initio gene identification in metagenomic sequences.* Nucleic Acids Research, 2010. **38**(12).

45. Noguchi, H., J. Park, and T. Takagi, *MetaGene: prokaryotic gene finding from environmental genome shotgun sequences.* Nucleic Acids Research, 2006. **34**(19): p. 5623-5630.

46. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* Bmc Bioinformatics, 2010. **11**.

47. Rho, M., H. Tang, and Y. Ye, *FragGeneScan: predicting genes in short and error-prone reads.* Nucleic Acids Research, 2010. **38**(20).

48. Markowitz, V.M., et al., *IMG: the integrated microbial genomes database and comparative analysis system.* Nucleic Acids Research, 2012. **40**(D1): p. D115-D122.

49. Markowitz, V.M., et al., *IMG 4 version of the integrated microbial genomes comparative analysis system.* Nucleic Acids Research, 2014. **42**(D1): p. D560-D567.

50. Liu, B., et al., *Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences.* Bmc Genomics, 2011. **12**.

51. Darling, A.E., et al., *PhyloSift: phylogenetic analysis of genomes and metagenomes.* Peerj, 2014. **2**.

52. Kielbasa, S.M., et al., *Adaptive seeds tame genomic sequence comparison.* Genome Research, 2011. **21**(3): p. 487-493.

53. Eddy, S.R., *Accelerated Profile HMM Searches.* Plos Computational Biology, 2011. **7**(10).

54. Matsen, F.A., R.B. Kodner, and E.V. Armbrust, *pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree.* Bmc Bioinformatics, 2010. **11**.

55. Ondov, B.D., N.H. Bergman, and A.M. Phillippy, *Interactive metagenomic visualization in a Web browser.* Bmc Bioinformatics, 2011. **12**.

56. Naccache, S.N., et al., *A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples.* Genome Research, 2014. **24**(7): p. 1180-1192.

57. Zhao, Y., H. Tang, and Y. Ye, *RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data.* Bioinformatics, 2012. **28**(1): p. 125-126.

58. Marcel, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17**(1): p. 10-12.

59. Schmieder, R. and R. Edwards, *Quality control and preprocessing of metagenomic datasets.* Bioinformatics, 2011. **27**(6): p. 863-864.

60. Angiuoli, S.V., et al., *CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing.* Bmc Bioinformatics, 2011. **12**.

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

62. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* Bmc Bioinformatics, 2003. **4**.

63. Joseph N Paulson, M. Pop, and H.C. Bravo,

Metastats: an improved statistical method for analysis of metagenomic data. Genome Biology, 2011. **12**(1): p. 17.

64. Gerlach, W. and J. Stoye, *Taxonomic classification of metagenomic shotgun sequences with CARMA3.* Nucleic Acids Research, 2011. **39**(14).

65. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Research, 2010. **38**: p. D211-D222.

66. Meyer, F., et al., *The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.* Bmc Bioinformatics, 2008. **9**.

67. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Research, 2005. **33**(17): p. 5691-5702.

68. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.* Applied and Environmental Microbiology, 2006. **72**(7): p. 5069-5072.

69. Cole, J.R., et al., *The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data.* Nucleic Acids Research, 2007. **35**: p. D169-D172.

70. Wuyts, J., et al., *The European database on small subunit ribosomal RNA.* Nucleic Acids Research, 2002. **30**(1): p. 183-185.

71. Leplae, R., et al., *ACLAME: A CLAssification of mobile genetic elements.* Nucleic Acids Research, 2004. **32**: p. D45-D49.

72. Ounit, R., et al., *CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers.* Bmc Genomics, 2015. **16**.

73. Freitas, T.A.K., et al., *Accurate read-based metagenome characterization using a hierarchical suite of unique signatures.* Nucleic Acids Research, 2015. **43**(10).

74. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.

75. Samuel S. Minot, N.K.a.N.B.G., *One Codex: A Sensitive 1 and Accurate Data Platform for Genomic Microbial Identification.* bioRxiv, 2015: p. 23.

76. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome Biology, 2014. **15**(3).

77. Brady, A. and S.L. Salzberg, *Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models.* Nature Methods, 2009. **6**(9): p. 673-U68.

78. Luo, C., et al., *ConStrains identifies microbial strains in metagenomic datasets.* Nature Biotechnology, 2015. **33**(10): p. 1045-+.

79. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

80. Eren, A.M., et al., *Anvi'o: an advanced analysis and visualization platformfor 'omics data.* PeerJ, 2015. **3**.

81. Teeling, H., et al., *TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences.* Bmc Bioinformatics, 2004. **5**.

82. Mende, D.R., et al., *Accurate and universal delineation of prokaryotic species.* Nature Methods, 2013. **10**(9): p. 881-+.

83. Arumugam, M., et al., *SmashCommunity: a metagenomic annotation and analysis tool.* Bioinformatics, 2010. **26**(23): p. 2977-2978.

84. Powell, S., et al., *eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges.* Nucleic Acids Research, 2012. **40**(D1): p. D284-D289.

85. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proceedings of the National Academy of Sciences of the United States of America, 1988. **85**(8): p. 2444-2448.

86. Brady, A. and S. Salzberg, *PhymmBL expanded: confidence scores, custom databases, parallelization and more.* Nature Methods, 2011. **8**(5): p. 367-367.

87. Diaz, N.N., et al., *TACOA - Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach.* Bmc Bioinformatics, 2009. **10**.

88. Kim, D., *Centrifuge: rapid and sensitive classification of metagenomic sequences.* Genome research, 2016. **26**: p. 1-9.

89. M, B. and W. D, *A block-sorting lossless data compression algorithm.* Technical Report 124 Digital Equipment Corporation, 1994.

90. Ferragina, P. and G. Manzini, *Opportunistic data structures with applications.* Proceedings of the 41st IEEE symposium on foundations of computer science, 2000.

91. Walt, A.J.v.d., et al., *Assembling metagenomes, one community at a time.* BMC Genomics, 2017. **18**(521).

92. Ayling, M., R.M. Leggett, and M.D. Clark, *New approaches for metagenome assembly with short reads.* 2019.

93. Hill, C.M., et al., *Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes.* 2017.

94. Buchfink, B., C. Xie, and D. Huson, *Fast and sensitive protein alignment using DIAMOND.* nature Methods, 2015. **12**: p. 59-60.

95. Huson, D.H., et al., *MEGAN analysis of metagenomic data.* Genome Research, 2007. **17**(3): p. 377-386.

96. Koren, S., T.J. Treangen, and M. Pop, *Bambus 2: scaffolding metagenomes.* Bioinformatics, 2011. **27**(21): p. 2964-2971.

97. Afiahayati, *MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning.* DNA Research, 2015. **22**(1): p. 69-77.

98. Namiki, T., et al., *MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads.* Nucleic Acids Research, 2012. **40**(20).

99. Zerbino, D.R. and E. Birney, *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.* Genome Research, 2008. **18**(5): p. 821-829.

100. Peng, Y., et al., *IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.* Bioinformatics, 2012. **28**(11): p. 1420-1428.

101. Peng, Y., et al., *Meta-IDBA: a de Novo assembler for metagenomic data.* Bioinformatics, 2011. **27**(13): p. I94-I101.

102. Simpson, J.T., et al., *ABySS: A parallel assembler for short read sequence data.* Genome Research, 2009. **19**(6): p. 1117-1123.

103. Li, D., *MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p. 1674-4676.

104. Richter, D.C., et al., *MetaSim-A Sequencing Simulator for Genomics and Metagenomics.* Plos One, 2008. **3**(10).

105. Tracey Allen K. Freitas, P.-E.L., Matthew B. Scholz and Patrick S. G. Chain, *Accurate read-based metagenome characterization using a hierarchical suite of unique signatures.* Nucleic Acids Research, 2015. **43**(10): p. 14.

106. Pinard, R., et al., *Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing.* BMC Genomics, 2006. **7**(1): p. 216.

107. Connor, T., *CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community.* Microb Genom, 2016. **2**(9).

108. Peabody, *Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities.* Bmc Bioinformatics, 2015. **16**(363).

109. Ehrlich, D., *Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes.* Nature Biotechnology, 2014. **32**: p. 822-828.

110. Harris, S. and M. Hunt, *ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads.* Microbial Genomics, 2017. **3**(10).

111. Susannah J Salter, et al., *Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.* BMC Biology, 2014. **12**(87).

112. Levy, S. and R. Myers, *Advancements in Next-Generation Sequencing.* Annual Review of Genomics and Human Genetics, 2016. **17**: p. 95-115.

113. Sanderson, N.D., et al., *Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices.* BMC Genomics, 2018. **19**(1): p. 714.

114. Yadav, D., *Comparative assessment of methods for metagenomic DNA isolation from soils of different crop growing fields.* 3 Biotech, 2016. **6**(2): p. 220.

115. Das, B., *An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples.* nature scientific reports, 2016. **6**(26775).

116. Fritz, A. and A. McHardy, *CAMSIM: Simulating metagenomes and microbial communities.* Microbiome, 2019. **7**(17).

117. Bongcam-Rudloff, H.G.E., *Simulating Illumina metagenomic data with InSilicoSeq.* Bioinformatics, 2019. **35**(3): p. 521-522.

118. Escalona, M., S. Rocha, and D. Posada, *A comparison of tools for the simulation of genomic next-generation sequencing data.* Nature Reviews Genetics, 2016. **17**: p. 459.

119. SHANNON, C.E., *A Mathematical Theory of Communication.* The Bell System Technical Journal, 1948. **27**: p. 379-423, 623-656.

Appendix 1 – Initial list of metagenomic analysis tools identified for potential evaluation

ABySS
Anvi'o
BLAST
CARMA
CARMA2
CARMA3
CLARK
CLARK-S
CLC Genomics Workbench
CloVR
CloVR-Meta
CONCOCT
ConStrains
DIAMOND
DiScRIBinATE
ESOM
GaSiC
GATTACA
Genometa
GOTTCHA
GroopM
GSMer
IDBA-UD
IMG
IMG-4
IMG-M
IMG-M4
Kraken
Kraken-mini
LMAT
megaBLAST
MEGAHIT
MEGAN
MetaAMOS
MetaBin
MetaCV
MetaFlow
Meta-IDBA
MetaPhlAn
MetaPhlAn2
MetaPhyler
metaSPAdes
MetaVelvet
MetaVelvet-SL
metBEETL
MG-RAST
MOCAT
NBC
One-Codex
PathoScope
PathoScope2
PhyloPhlAn
Phylopythia
PhyloSift
PhymmBL
PhymmBL_expanded
Ray Meta
RITA
Sequedex
ShotMAP
SIGMA
SOAP
SOAP2
SPAdes
SPANNER
specl
SURPI
TACOA
TaxyPro
TETRA
Velvet

Appendix 2 – List of organisms selected to build *in-silico* datasets, including the abundances for the varied abundance *in-silico* datasets AM_100G_10M (Appendix 2-A); AM_10G_10M (Appendix 2-B) and Zymo_8G_10M (Appendix2-C).

Appendix 2-A: List of species for *in-silico* dataset AM_100G_10M and the abundance of species for dataset AM_100G_10M-V.

Genus	Species	AM_100G_10M-V Abundance (%)
Acholeplasma	*Acholeplasma laidlawii*	0.1
Acidovorax	*Acidovorax avenae*	0.3
Agrobacterium	*Agrobacterium vitis*	0.5
Alkaliphilus	*Alkaliphilus metalliredigens*	0.7
Arthrobacter	*Arthrobacter arilaitensis*	0.1
	Arthrobacter aurescens	0.5
	Arthrobacter chlorophenolicus	1.5
	Arthrobacter sp. FB24	1.9
Asticcacaulis	*Asticcacaulis excentricus*	0.9
Azoarcus	*Azoarcus sp. BH72*	1.1
Azorhizobium	*Azorhizobium caulinodans*	1.3
Azospirillum	*Azospirillum sp. B510*	1.5
Bacillus	*Bacillus anthracis*	0.1
	Bacillus cellulosilyticus	0.3
	Bacillus cereus	0.5
	Bacillus megaterium	0.7
	Bacillus pumilus	1.3
	Bacillus subtilis	1.5
	Bacillus thuringiensis	1.7
Bacteroides	*Bacteroides fragilis*	1.7
Borrelia	*Borrelia burgdorferi*	1.9
Brachybacterium	*Brachybacterium faecium*	0.1
Bradyrhizobium	*Bradyrhizobium japonicum*	0.3
	Bradyrhizobium sp. BTAi1	0.7
	Bradyrhizobium sp. ORS 278	1.7
Brevundimonas	*Brevundimonas subvibrioides*	0.3
Burkholderia	*Burkholderia multivorans*	0.5
	Burkholderia vietnamiensis	1.5
Candidatus Desulforudis	*Candidatus Desulforudis audaxviator*	0.5
Candidatus Protochlamydia	*Candidatus Protochlamydia amoebophila*	0.7
Caulobacter	*Caulobacter crescentus*	0.3
	Caulobacter segnis	1.1
	Caulobacter sp. K31	1.7

Chloroflexus	*Chloroflexus aurantiacus*	0.9
Clavibacter	*Clavibacter michiganensis*	1.1
Clostridium	*Clostridium cellulolyticum*	0.1
	Clostridium difficile	0.9
	Clostridium perfringens	1.9
Enterococcus	*Enterococcus faecalis*	0.5
	Enterococcus faecium	1.9
Erythrobacter	*Erythrobacter litoralis*	1.3
Escherichia	*Escherichia coli*	1.5
Flavobacterium	*Flavobacterium johnsoniae*	1.7
Gloebacter	*Gloeobacter violaceus*	1.9
Gluconacetobacter	*Gluconacetobacter diazotrophicus*	0.1
Hyphomonas	*Hyphomonas neptunium*	0.3
Janthinobacterium	*Janthinobacterium sp. Marseille*	0.5
Lactobacillus	*Lactobacillus crispatus*	0.7
Leifsonia	*Leifsonia xyli*	0.9
Listeria	*Listeria monocytogenes*	1.1
Maricaulis	*Maricaulis maris*	1.3
Methylobacterium	*Methylobacterium extorquens*	0.1
	Methylobacterium nodulans	0.7
	Methylobacterium populi	1.1
	Methylobacterium radiotolerans	1.5
	Methylobacterium sp. 4-46	1.9
Micrococcus	*Micrococcus luteus*	1.5
Mycobacterium	*Mycobacterium tuberculosis*	1.7
Nitrobacter	*Nitrobacter hamburgensis*	1.9
Novosphingobium	*Novosphingobium aromaticivorans*	0.3
Ochrobactrum	*Ochrobactrum anthropi*	0.7
Paenibacillus	*Paenibacillus larvae*	0.3
	Paenibacillus polymyxa	0.9
	Paenibacillus sp. JDR-2	1.7
Paracoccus	*Paracoccus denitrificans*	0.9
Pedobacter	*Pedobacter heparinus*	1.1
Pelotomaculum	*Pelotomaculum thermopropionicum*	1.3
Phenylobacterium	*Phenylobacterium zucineum*	1.9
Planctomyces	*Planctomyces limnophilus*	0.9
Pseudomonas	*Pseudomonas aeruginosa*	0.5
	Pseudomonas putida	1.3
Pseudoxanthomonas	*Pseudoxanthomonas suwonensis*	1.1
Rhizobium	*Rhizobium etli*	1.3
Rhodobacter	*Rhodobacter capsulatus*	0.7
	Rhodobacter sphaeroides	1.5
Rhodococcus	*Rhodococcus erythropolis*	0.9
Rhodopseudomonas	*Rhodopseudomonas palustris*	1.3
Roseobacter	*Roseobacter denitrificans*	0.1
Ruegeria	*Ruegeria pomeroyi*	0.3
Ruminococcus	*Ruminococcus albus*	0.5
Shewanella	*Shewanella sp. ANA-3*	0.7
Sorangium	*Sorangium cellulosum*	0.9
Sphingobium	*Sphingobium japonicum*	1.1

Sphingomonas	*Sphingomonas wittichii*	1.3
Sphingopyxis	*Sphingopyxis alaskensis*	1.5
Spirosoma	*Spirosoma linguale*	1.7
Staphylococcus	*Staphylococcus aureus*	1.9
Stenotrophomonas	*Stenotrophomonas maltophilia*	0.1
Streptococcus	*Streptococcus agalactiae*	0.1
	Streptococcus pneumoniae	1.1
	Streptococcus suis	1.7
Streptomyces	*Streptomyces violaceusniger*	0.3
Symbiobacterium	*Symbiobacterium thermophilum*	0.5
Syntrophothermus	*Syntrophothermus lipocalidus*	0.7
Thermaerobacter	*Thermaerobacter marianensis*	0.9
Thermoanaerobacter	*Thermoanaerobacter italicus*	1.1
Veillonella	*Veillonella parvula*	1.3
Xanthomonas	*Xanthomonas campestris*	1.5
Xylanimonas	*Xylanimonas cellulosilytica*	1.7
Zymomonas	*Zymomonas mobilis*	1.9

Appendix 2-B: Abundance of species for dataset AM_10G_10M-V.

AM_10G_10M-V species	**Abundance %**
Pseudomonas aeruginosa	24.2
Bacillus thuringiensis	22.9
Acholeplasma laidlawii	15
Bacillus cereus	8.6
Staphylococcus aureus	7.1
Brevundimonas subvibrioides	7
Escherichia coli	5.3
Mycobacterium tuberculosis	4.4
Pseudomonas putida	4.2
Bacillus anthracis	1.2

Appendix 2-C: Abundance of species for dataset Zymo_8G_10M-V.

Zymo_8G_10M-V species	**Abundance %**
Pseudomonas aeruginosa	34.31
Escherichia coli	22.26
Salmonella enterica	19.22
Enterococcus faecalis	9.2
Lactobacillus fermentum	6.21
Staphylococcus aureus	5.15
Listeria monocytogenes	2.69
Bacillus subtilis	0.96

Appendix 3 – Output from analysis of *in-silico* datasets using the metagenomic analysis tool Centrifuge

In-silico Dataset	True Positive	False Positive	False Negative	Sensitivity (%)	Precision (%)
AM_10G_10M	10	252	0	100	3.8
AM_10G_10M-V	10	232	0	100	4.1
AM_10G_10M-150	10	374	0	100	2.6
Zymo_8G_10M	8	307	0	100	2.5
Zymo_8G_10M-V	8	305	0	100	2.6
Zymo_8G_10M-150	8	466	0	100	1.7
AM_100G_10M	97	635	3	97	13
AM_100G_10M-V	97	606	3	97	14
AM_100G_10M-150	94	916	6	94	9.3

Appendix 4 – List of organisms selected to build the *in-vitro* mock community datasets Ba and Yp.

Genus	Species	Quantity (µg)	Genome size (Mb)	Relative abundance (%)
Actinobacillus	*Actinobacillus pleuropneumoniae*	10	2.3	0.04
Aeromonas	*Aeromonas hydrophila*	10	4.9	0.02
Alcaligenes	*Alcaligenes faecalis*	10	3.9	0.02
Arcanobacterium	*Arcanobacterium pyogenes*	10	2.3	0.04
Bacillus	*Bacillus cereus*	10	5.8	0.02
	Bacillus halodurans	5	4.2	0.01
	Bacillus subtilis	5	4.1	0.01
	Bacillus thuringiensis	10	6.1	0.02
Bordetella	*Bordetella bronchiseptica*	5	5.2	0.01
	Bordetella parapertussis	5	4.8	0.01
	Bordetella pertussis	5	4.1	0.01
Citrobacter	*Citrobacter freundii*	10	5.3	0.02
Clostridium	*Clostridium difficile*	5	4.2	0.01
	Clostridium perfringens	5	3.5	0.01
Cupriavidus	*Cupriavidus metallidurans*	5	7.0	0.01
Deinococcus	*Deinococcus radiodurans*	10	3.2	0.03
Derxia	*Derxia gummosa*	10	5.2	0.02
Enterobacter	*Enterobacter aerogenes*	10	5.3	0.02
	Enterobacter cloacae	10	4.9	0.02
Enterococcus	*Enterococcus*	5	3.0	0.02

	faecalis			
	Enterococcus faecium	10	2.9	0.03
Escherichia	*Escherichia coli*	5	5.1	0.01
Geobacillus	*Geobacillus stearothermophilus*	10	2.9	0.03
Klebsiella	*Klebsiella oxytoca*	10	6.0	0.02
	Klebsiella pneumonia	5	5.6	0.01
Legionella	*Legionella pneumophila*	5	3.4	0.01
Listeria	*Listeria innocua*	10	2.9	0.03
	Listeria monocytogenes	10	3.0	0.03
Mannheimia	*Mannheimia haemolytica*	10	2.6	0.04
Morganella	*Morganella morganii*	10	4.0	0.02
Pantoea	*Pantoea agglomerans*	10	4.9	0.02
Pectobacterium	*Pectobacterium atrosepticum*	10	5.0	0.02
Plesiomonas	*Plesiomonas shigelloides*	10	3.8	0.02
Porphyromonas	*Porphyromonas gingivalis*	10	2.3	0.04
Propionibacterium	*Propionibacterium acnes*	5	2.5	0.02
Proteus	*Proteus mirabilis*	10	4.0	0.02
	Proteus vulgaris	10	4.0	0.02
Providencia	*Providencia stuartii*	10	4.4	0.02
Pseudomonas	*Pseudomonas aeruginosa*	5	6.6	0.01
	Pseudomonas fluorescens	10	6.3	0.01
	Pseudomonas putida	5	6.0	0.01
Rahnella	*Rahnella aquatilis*	10	5.4	0.02
Rhizobium	*Rhizobium*	10	5.6	0.02

	radiobacter
Salmonella	*Salmonella enterica*
Serratia	*Serratia marcescens*
Shewanella	*Shewanella oneidensis*
Shigella	*Shigella flexneri*
Staphylococcus	*Staphylococcus aureus*
Streptococcus	*Streptococcus pneumoniae*
Vibrio	*Vibrio fischeri*
	Vibrio parahaemolyticus
Bacillus	*Bacillus anthracis*
Yersinia	*Yersinia pestis*

Appendix 5 – High Confidence Taxonomic identification of bio-aerosol samples collected over a long term temporal study (Appendix 5-A) and over a short term temporal study (Appendix 5-B).

Appendix 5-A: High Confidence taxonomic identification of bio-aerosol samples collected over a long term temporal study.

Sample ID	**155**	**186**	**187**	**211**	**246**	**247**
Date	**21/05/18**	**21/06/18**		**21/07/18**	**21/08/18**	
Collection time	**1200-1600**	**0800-1200**	**1200-1600**	**1200-1600**	**0800-1200**	**1200-1600**
Lactobacillus amylovorus	-	12.4123867	8.343856667	-	1.85819	-
Stenotrophomonas maltophilia	4.69644	1.82406333	0.533871667	4.90813	5.825053333	2.96190333
Clavibacter michiganensis	1.5921456	2.1093543	1.832943	1.82923666	2.53020666	6.15399

	7	3		7	7	
Pseudomonas poae	-	-	2.143286667	4.946746667	-	7.51603
Pseudomonas syringae	-	2.419126667	2.23864	2.66321	-	2.06946333
Lactobacillus reuteri	1.91240333	3.03222333	1.682636667	1.273882333	1.415793333	-
Brachybacterium faecium	3.55100333	1.53823333	0.609757333	2.062173333	1.11978	-
Megasphaera elsdenii	-	2.825376667	2.131496667	-	-	-
Propionibacterium acnes	2.551536667	-	0.767475667	-	1.500546667	-

Eubacterium rectale	-	2.040125	0.957675	-	-	-
Escherichia coli	-	-	-	1.264448	-	1.71671167
Corynebacterium efficiens	-	1.43220667	0.723621	-	0.710475333	-
Lactobacillus salivarius	1.85378	-	-	-	0.999376333	-
Terriglobus roseus	1.258786	-	0.528376667	-	1.05332	-
Lactobacillus johnsonii	-	-	-	1.43175	1.31283	-
Pantoea vagans	-	-	0.795341667	1.265456	-	-
Staphylococcus	-	-	-	-	1.27850333	-

saprophyticus					3	
Erwinia billingiae	-	-	1.188046667	-	-	-
Sanguibacter keddieii	-	-	-	-	-	1.170034
Lactobacillus crispatus	-	-	-	-	1.16491	-
Lactobacillus helveticus	-	-	-	1.043987	-	-
Erwinia tasmaniensis	-	-	0.947076667	-	-	-
Arthrobacter arilaitensis	-	-	-	-	0.871963333	-
Jonesia denitrificans	-	-	0.780666333	-	-	-

Corynebacterium glutamicum	-	-	0.678162	-	-	-
Buchnera aphidicola	-	-	0.639882	-	-	-
Acholeplasma laidlawii	-	-	0.554955	-	-	-

Appendix 5-B: High Confidence taxonomic identification of bio-aerosol samples collected over a short term temporal study.

Collection time	0000-0400	0400-0800	0800-1200	1200-1600	1600-2000	2000-0000
Lactobacillus amylovorus	7.19791333	3.27983	12.4123867	8.343857	16.9548533	19.1439333
Lactobacillus reuteri	2.43749333	3.39356	3.03222333	1.682637	2.94471667	3.05811
Stenotrophomonas maltophilia	2.55310333	3.17570667	1.82406333	0.533872	0.61193367	0.53766
Brachybacterium faecium	1.36035	2.937	1.53823333	0.609757	0.390704	0.42012367
Clavibacter michiganensis	0.63513333	0.82501267	2.10935433	1.832943	0.80031133	0.64158433
Pseudomonas syringae	1.2543833	-	2.41912667	2.23864	0.65299467	0.4360466

	3					7
Corynebacterium efficiens	0.20550543	-	1.43220667	0.723621	1.26763867	1.244324
Megasphaera elsdenii	-	-	2.82537667	2.131497	2.77635	2.7496
Lactobacillus johnsonii	2.07319	4.25058333	-	-	0.8436	0.83179667
Eubacterium rectale	-	-	2.040125	0.957675	1.386925	1.244505
Lactobacillus salivarius	1.34865333	3.88342	-	-	-	0.31305267
Lactobacillus crispatus	0.84138	3.113485	-	-	-	0.516985
Erwinia billingiae	-	-	-	1.188047	2.43878667	0.29819367

Jonesia denitrificans	-	-	-	0.780666	0.82722767	0.81013267
Acholeplasma laidlawii	-	-	-	0.554955	0.571679	0.42232843
Staphylococcus saprophyticus	0.48177	2.83818	-	-	-	-
Corynebacterium urealyticum	0.34619367	2.073143	-	-	-	-
Propionibacterium acnes	-	1.27939333	-	0.767476	-	-
Lactobacillus helveticus	0.431452	1.45699167	-	-	-	-
Escherichia coli	0.30113833	1.17852	-	-	-	-
Corynebacterium glutamicum	-	-	-	0.678162	-	0.67655433

Desulfovibrio desulfuricans	-	-	-	-	0.571823	0.50621767
Acidaminococcus fermentans	-	-	-	-	0.403092	0.357667
Bifidobacterium thermophilum	-	-	-	-	0.38063133	0.35457433
Lactobacillus delbrueckii	-	-	-	-	0.34675507	0.30004383
Pseudomonas poae	-	-	-	2.143287	-	-
Erwinia tasmaniensis	-	-	-	0.947077	-	-
Pantoea vagans	-	-	-	0.795342	-	-
Acidaminococcus intestini	-	-	-	-	-	0.65218
Buchnera aphidicola	-	-	-	0.639882	-	-

Terriglobus roseus	-	-	-	0.528377	-	-
Arthrobacter arilaitensis	-	-	-	-	-	0.47326217
Brachyspira pilosicoli	-	-	-	-	-	0.30845467
Bacteroides salanitronis	0.141275	-	-	-	-	-