Taxonomic Classification of Bacteria in Shotgun Metagenomic Samples

Evaluation of Species Level Classification in the Context of Pathogen Screening for Healthcare Applications

Master's thesis in Engineering Mathematics and Computational Science

IRIS GOLD RODAL

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Taxonomic Classification of Bacteria in Shotgun Metagenomic Samples

Evaluation of Species Level Classification in the Context of Pathogen Screening for Healthcare Applications

IRIS GOLD RODAL

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology
Gothenburg, Sweden 2023

Taxonomic Classification of Bacteria in Shotgun Metagenomic Samples Evaluation of Species Level Classification in the Context of Pathogen Screening for Healthcare Applications
IRIS GOLD RODAL

Supervisor: Oscar Aspelin, Bioinformatician at 1928 Diagnostics
Examiner: Erik Kristiansson, Professor in Biostatistics and Bioinformatics at Chalmers University of Technology

Master's Thesis 2023
Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Taxonomic Classification of Bacteria in Shotgun Metagenomic Samples
Evaluation of Species Level Classification in the Context of Pathogen Screening for Healthcare Applications
IRIS GOLD RODAL
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

The aim was to investigate taxonomic classification and removal of human host DNA in the context of a bioinformatic analysis pipeline for screening of pathogens. The examination was carried out using simulated short-read sequenced shotgun metagenomic samples. It was found that a majority of human origin DNA could be separated from bacteria using the K-mer based read classifier Kraken 2 with a custom built human only reference database. Effects on taxonomic classification performance were surveyed for variations in sample composition, parameter settings of the taxonomic classifier and reference database composition. Maintaining both high precision and recall for species level taxonomic classification of metagenomic samples was challenging for limited computational resources. A one-size-fits-all approach to taxonomic classification of any shotgun metagenomic sample would be near impossible with the tested K-mer based classifiers (Kraken 2 and Bracken) and instead specialized pipeline tracks optimized for different expected range of species, sequencing depth and abundance distributions could be a solution.

Keywords: metagenome taxonomic classification, shotgun metagenomics, WGS, Kraken 2, Bracken, taxonomic classifier, host removal.

Acknowledgements

This endeavor would not have been possible without collaboration with 1928 Diagnostics, who provided me with the opportunity to carry out my thesis project at the company and I am thankful for the resources they provided. First and foremost I would like to express my gratitude to my supervisor Oscar Aspelin for sharing expert bioinformatics knowledge and for his continued guidance and encouragement throughout the year. I am further thankful for the support from the development team at 1928 Diagnostics, who, in addition to helpful coding tips and tricks, gave me much appreciated insight into how problems similar to the one in this thesis are tackled for real world applications. I would also like to direct many thanks my examiner Erik Kristiansson for taking on this project and my opponent Jewahri Idris Yousuf, especially for thoughtful feedback and suggestions on the report.

Iris Gold Rodal, Gothenburg, October 2023

Contents

List of Figures

List of Tables

1

Introduction

Resistance to antibiotics among pathogenic bacteria is listed by the World Health Organization as one of the largest threats to global health [2]. When antibiotics stop working, previously easily treatable infections can become life-threatening to individuals and a dire problem for healthcare systems. Infections can afflict anyone however people who are immunocompromised, such as premature babies and organ transplant recipients, are particularly at risk and in need of quick diagnostic methods [3][4]. An important part of tracking and prevention of infectious disease outbreaks in hospitals is identification of both the bacterial species and the specific genetic components that can cause antibiotic resistance [5].

There are several different types of methods for identification of species, such as those based on direct laboratory tests or DNA sequencing. An emerging DNA sequencing based technology is shotgun metagenomic sequencing [6]. It is generally faster than single isolate sequencing, where a time consuming cell cultivation step is needed, and captures more information (such as resistance information) than the amplicon metagenomic sequencing method does. In shotgun metagenomics all the DNA in the sample is directly sequenced without prior cultivation and separating of species [7]. However, the drawback of the method is that the bioinformatic analysis can become complicated by having a mixture of sequences from different species and abundances, as well as containing contaminants such as DNA from human hosts.

1.1 Aim and Scope

The context of this work is the need for fast and comprehensive methods for screening for pathogens and associated antimicrobial resistance (AMR) in healthcare settings. A broader goal is to develop an automated bioinformatic pipeline for analysis of metagenomic samples, for example taken from patient's bodily fluids, skin swabs or hospital facilities. The main purpose of the pipeline is to take raw whole genome shotgun metagenomic sequencing data as input, do analyses and output a report with identified organisms and markers for AMR. The analysis steps in a minimal proposed prototype can be divided into pre-processing of the sample data, removal of human host DNA, taxonomic classification, genome assembly and identification of antimicrobial resistance markers, shown schematically in Figure 1.1. The scope of the project is however limited to implementing and evaluating two parts of the prototype shotgun metagenomics pipeline, the taxonomic classification and filtering of human DNA. A more in-depth description of the aims and scope pertaining to

the selected parts are given in sections 1.1.1 and 1.1.2 below.

Bioinformatic analyses can be inaccessible and expensive, especially of large whole genome sequenced metagenomic samples, as methods tend to be computationally intensive and require high memory usage[8]. To make the findings more applicable, the challenge is set for the pipeline to be adapted for running on a laptop (500 GB disc, 16 GB RAM and 8 CPUs). Furthermore, to achieve the functionalities, building blocks of open software programs and algorithms are used and there is thus no intention of constructing a fully original pipeline.

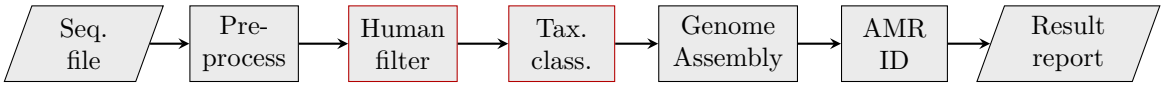


Figure 1.1: Main segments of the *in silico* pipeline for taxonomic classification and identification of antimicrobial resistance of shotgun metagenomic samples from the raw sequence files (FASTQ format). The samples are pre-processed with quality control of the reads in the sequence files, reads with human origin are thereafter removed in a filtering step before the sample is passed to taxonomic classification where the species in the sample are identified. The non-human reads in the sample are assembled, which is where the mixed sequence fragments are sorted and reassembled into the original coherent genome sequences for each organism. Using the genome assemblies, genes associated with antimicrobial resistance can be found. The focus of this thesis are the steps in red which are filtering of human host DNA and taxonomic classification.

1.1.1 Taxonomic Classification

Knowing which microorganisms, especially pathogens, are present in a sample and which are not, can be central to treatment of infections or other medical interventions [9]. The function of a taxonomic classification step is to identify and name the organisms in the metagenomic sample based on unique properties of their genome sequences. An advantage of shotgun metagenomics is the possibility of capturing genome sequences of pathogens from all domains of life at the same time, however, due to project limitations the investigation into the problem of taxonomic classification is constrained to bacteria. Furthermore, the lowest taxonomic rank the bacteria will be differentiated at is species and it is limited to previously characterized species. Ideal taxonomic classification, within the scope of the project, is thus defined as correctly classifying all non-novel bacteria in a shotgun metagenomic sample down to the species rank with zero false discoveries.

There are multiple factors involved in determining the difficulty level and success of taxonomic classification. These include, but are not limited to, sample content, taxonomic classifier and reference database. The aim is to describe the individual and combined effects of selected variables from each category and optimize for classification performance.

The main aspects considered surrounding sample composition is how the classification is effected by the number of species in the sample, similarity between species and sequencing depth of the organisms in the sample. The scope is limited to short-read samples sequenced using Illumina technologies, which is one of the commonly

used sequencing methods for whole genome shotgun metagenomics [10].

The examination of taxonomic classifiers is focused on the K-mer based DNA-to-DNA classification program Kraken 2 [11] and its companion program Bracken [12]. Kraken 2 has consistently been among the top performers in regards to computational time and classification recall, however lower on precision [8][13][1]. Bracken is a tool for Bayesian reestimation of abundance after classification with Kraken and has been shown to enhance the abundance estimates of the taxonomic classifications [12]. The objective is to investigate if parameter tuning of the Kraken 2 and the accompanying Bracken can improve classification performance.

Kraken 2 uses an indexed database of genetic reference sequences and for fast classification needs enough free memory to hold the database in memory or maps to disc as a slower alternatively [11]. As memory space is limited, there will for a fixed file size be a dichotomy between the amount of representation each species has and the number of species in the database. The aim is to survey if custom built reference databases can increase classification performance compared to a pre-built standard within the size constraints of being capable to run on a laptop.

1.1.2 Human DNA Filtering

Collateral human DNA is expected to be included when doing shotgun metagenomic sampling directly from a patient, for example of blood or skin and oral swabs [14]. Mixed human origin DNA is also likely to appear in varying amounts when sampling from an environment such as hospitals [15]. In the context of a bioinformatic pipeline intended for clinical application, it is therefor likely for incoming samples to contain substantial amounts of human DNA if not removed in the laboratory.

The content of the human DNA is not of interest for taxonomic classification of bacteria or identification of antibiotic resistance, however, storing and processing it would consume unnecessary resources as well as have implications for privacy issues as it can be linked to individuals. The bottleneck for computational resources is the genome assembly step. Genome assembly, where the mixed sequence fragments in the sample are sorted and reassembled into the original coherent genome sequences for each organism, is needed for the identification of antibiotic resistance genes. Depending on the program used for assembly, the time and resource usage can grow fast when the sequence complexity grows, such as when species and genome sizes are increased [16][17]. Filtering out human reads from the sample before assembly could reduce the complexity as the human genome is large compared to microorganisms. It is therefore relevant to investigate methods for filtering of human reads as a pipeline step when working with limited memory and computational resources.

Before sequencing the metagenomic sample, physical human DNA fragments can be removed or degraded using commercial human host depletion kits [14][18]. However, the processes of human DNA elimination can have biased effects on the recovery of microbial DNA for certain species due to their characteristics [14]. Filtering of human DNA reads in-silico, post sequencing, can be used as a compliment or preferred alternative to the depletion kits depending on the percentage human DNA in the sample and intended analysis [14]. There are established workflows for separating

human reads from the microbial reads in the sample FASTQ file based on direct alignment of reads to a human reference genome [19]. The k-mer based taxonomic read classifier Kraken 2 has been shown to be comparably fast at classifying microorganisms in metagenomic samples [11]. It is proposed here to take advantage of the speed of Kraken 2 for the extrapolated task of human read classification and removal. The aim pertaining to a pipeline step for filtering out human DNA is to investigate if Kraken 2 can be a faster alternative to alignment based methods while quality wise remaining competitive. Specifically the questions to answer are: Can human origin reads be filtered out from an Illumina short-read shotgun metagenomic sample using the K-mer based taxonomic classifier Kraken 2? Is filtering by Kraken 2 comparable to a method using the Burrows-Wheeler aligner BWA-mem 2?

2

Theory

The intent of the chapter is to provide a theoretical background on the methodology behind metagenomic sequencing data, reference databases and taxonomic classification programs.

2.1 Metagenomic Sequencing

Metagenomic sampling is used when characterizing microbiomes in both the natural and human environment [20][21]. Metagenomics is also emerging in the context of infectious disease detection for species identification and antimicrobial resistance prediction [9][6]. A metagenomic sample is a biological sample that contains genetic material from organisms, potentially multiple of different types, present in the sampled environment. When a metagenomic sample is collected, by for example a skin swab or tissue sample, first the genetic material, in this case DNA, is extracted from the cells [22]. To be able to interpret the information carried by the physical DNA strands they need to be converted into a readable form by a DNA sequencer. The extracted DNA strands could be sequenced by targeting certain regions using amplicon methodology, or non-selectively in so called shotgun sequencing where everything in the sample is sequenced [23].

The output from DNA sequenced with next generation sequencing technologies is a digital read file, typically in the format .FASTQ, where each *read* of a DNA fragment is a transcription into letters accompanied by Phred quality scores of each DNA nucleotide [24]. The reads can be of differing length depending on the type of sequencer. For example, Illumina short-read sequencers give read lengths of 50-300 bp and read lengths of 150 bp are recommended by Illumina for whole-genome sequencing [25][26]. Illumina sequencing can generate so called paired-end reads where a fragment sequenced in both directions gives rise to two paired reads and a sequenced sample can have two .FASTQ files, one for the reads of the forward strand and one for the reverse strand [24]. The average sequencing error rates for Illumina sequencers are less than 1 %, however, the errors are non-uniformly distributed across the reads and furthermore depend on the specific illumina platform used [27].

Information can be gained directly from the read, such as some types of taxonomic classification, or the reads can be assembled into the original genomes using the property of reads overlapping along the genome [23]. Following from that the organisms in the metagenomic sample are not separated before extraction and sequencing of the DNA, the reads from all organisms in the sample are mixed together in the

.FASTQ file which is in contrast to sequencing of single isolate cultures where all reads originate from one type of organism [22]. The relative abundance of cells from a certain species in the physical sample will approximately carry over to the relative abundance of reads from that species in the .FASTQ file, however, the read abundance is also influenced by the length of the species's genome. The average read coverage of a species genome is the average number of reads covering a nucleotide position in the genome. A metagenomic sample with a non-uniform distribution of species abundances will result in some genomes of species being covered by fewer reads than others, even below an average of 1X depending on the sequencing depth of the sample [28]. To increase the read coverage of the lowest abundance species the whole sample must be further sequenced which extends the sequencing time.

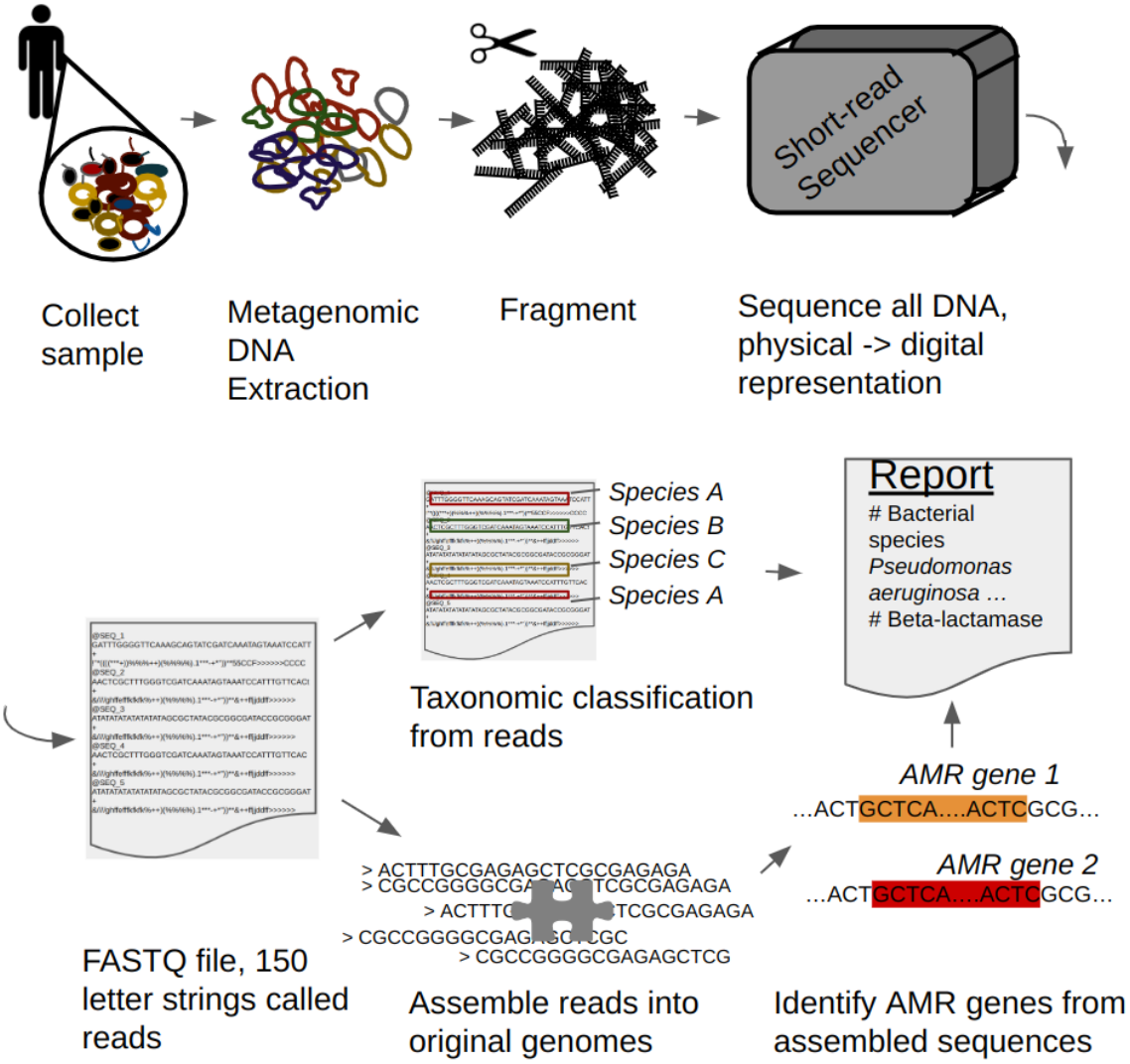
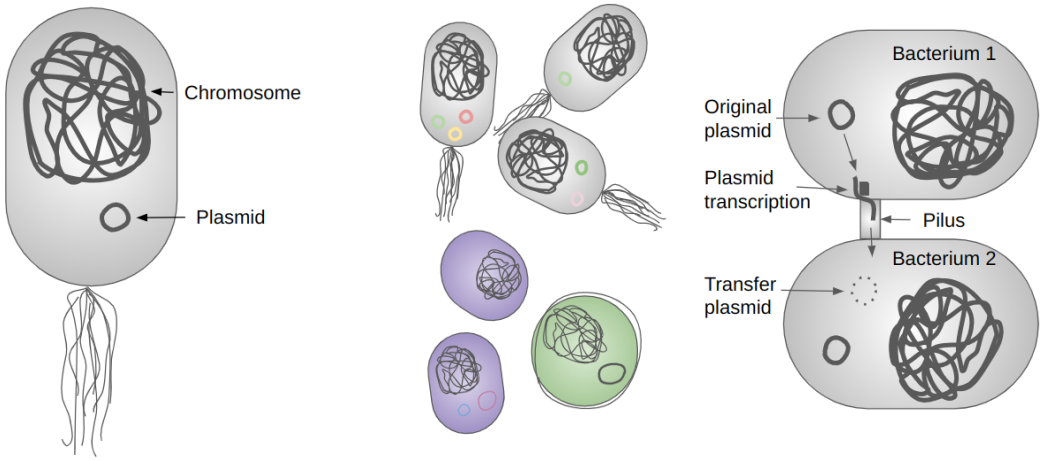


Figure 2.1: One of several possible metagenomic processes where DNA is extracted from a biological sample and converted to a digital format by a sequence. The sequenced DNA data can then be processed to extract information such as species of origin or genes.

2.2 Plasmids

A majority of a bacterium's genes are located on its chromosome, however, a subset of accessory genes may be found on a plasmid which is extrachromosomal circular DNA [29], see Figure 2.2 (a). The functionalities of plasmids are diverse as well as their size which ranges from hundreds of base pairs to over a million [30]. The prevalence and type of plasmids varies between bacterial species but can also differ within a species, as illustrated in Figure 2.2 (b). Similar as for the chromosome, plasmids are replicated and passed on to daughter cells in association with cell division, however, they also carry genes for independent transmission and replication [29]. Furthermore, plasmids are mobile genetic elements that can be horizontally transferred, by so called conjugation, between prokaryotic cells in the same generation, see Figure 2.2 (c). Conjugation of plasmids can occur naturally between both bacterial cells of the same species and of two different species [31]. Plasmids are therefore a common vector for spread of genes associated with antibiotic resistance properties [29]. The plasmid transfer is however not random as it may be induced by environmental factors and is furthermore unequally distributed across bacterial species [32]. For example, the F-type plasmids are mainly transferred among species within the *Enterobacteriaceae* family group which includes pathogenic strains of *Escherichia* and *Salmonella* [33].



(a) (b) (c)

Figure 2.2: Sketches of bacteria and with plasmids. (a) The plasmid is the smaller circular DNA element separate from the bacterial chromosome. (b) Bacteria may have zero, one or multiple plasmids of different types varying both intra and inter species. (c) Plasmid conjugation between two bacteria where the plasmid in Bacterium 1 is transcribed and then passed through a pilus temporary connecting 'tube' to Bacterium 2.

2.3 Genomic Reference Sequences and Taxonomy

The National Center for Biotechnological Information (NCBI) is maintained by the United States government and provides public access to biomedical and genomic in-

formation [34]. The NCBI online resources takes the form of reference databases, sequence analysis tools and repositories for research studies. NCBI genomic resources include genomes, mainly of human associated microorganisms, and are organized in databases with nucleotide sequences, genome assemblies and mapped annotations.

GenBank is an archival database of publicly available genetic sequences. It accepts submissions of whole genome sequences and annotated genome assemblies in addition to for example transcriptome assemblies and targeted locus studies. [35] Until the year 2022, 1,349,781 genome assemblies of cellular organisms were submitted of which 91 % were from bacteria [36]. RefSeq is another sequence collection distributed by NCBI and it can be described as a non-redundant and curated version of GenBank [37] [38]. In contrast to Genbank, the NCBI owns the records and can update, combine and improve the annotations and sequences [38]. The two databases have similar content but differ in format and how the sequences are annotated.

RefSeq contains genome assembly entries from all domains of life. For entries until and including the year 2022 there are 331,892 assemblies of cellular organisms of which 41 % is bacterial [36]. Furthermore, of the bacterial assemblies 33,364 are on complete genome level and when atypical genomes are excluded there are 32,645 assemblies remaining. More studied organisms are overrepresented in the number of assembly entries in RefSeq and the taxonomic resolution is often higher with entries categorized on lower taxonomic levels such as sub-species and/or strain level. For example a subset of seven pathogenic bacteria more closely examined in this thesis, *A. baumannii*, *E. faecium*, *E. fecalis*, *K. pneumoniae*, *P. aeruginosa*, *S. aureus* and *S. epidermidis*, make up 13.6 % of the complete level bacterial assemblies [36].

Every entry of a genome assembly in RefSeq is given a new unique RefSeq assembly accession alongside an individual accession for each FASTA sequence in the assembly. The assembly is tagged with its estimated level of completeness, where the alternatives in increasing order are contig, scaffold, chromosome and complete genome level. For a complete assembly of a bacterial organism the same number of sequences in the FASTA file are expected as the combined number of chromosomes and plasmids in the organism in question. The number and type of plasmids may not be internally consistent between exemplar of the same species as there can be a physiological variation in plasmid distribution.

A subset of the RefSeq assemblies are categorized by NCBI as *Representative genomes*, also referred to as *Reference genomes*, based on their sequence annotation quality, recognition as a community standard or medical importance [38][36]. As to not confuse terminology, any genome can be used as a reference, for example when aligning reads or building an indexed database, however, the RefSeq category refers to assemblies of particular importance and generally one per species. In exceptional cases there may be more than one reference genome per prokaryote species, for example *E. coli* has two. The number of bacterial reference genomes listed until 2022 are 16,765 spread across all assembly completeness levels. Another aspect of the RefSeq database entries are their status. A genome assembly may be assigned a suppressed status for reasons that include being replaced by another record to reduce redundancy and removal by curators due to lack of support [37]. Suppressed records are however still retrievable but have a marker.

The taxonomy system used by NCBI is a hierarchical organization of organism relations with the main ranks of super kingdom/domain, phylum, class, order, family, genus and species [39], sketched in Figure 2.3 (a). All NCBI records, such as genome assemblies, are mapped to nodes in the taxonomic tree. For cellular organisms the distance between two organisms across ranks in the taxonomic tree is not always consistent with the actual genetic similarity between organisms and the taxonomic tree is thus not a phylogentic tree [40]. To capture the variability there may be extra ranks added, such as species complexes, sub-species and strains. With new research discoveries the taxonomy is continuously updated by adding or merging taxonomic nodes [39]. Furthermore changes of nomenclature conventions may also trigger updates of the NCBI taxonomy. For example an inclusion of the phylum rank in the International Code of Nomenclature of Prokaryotes lead to renaming or new inclusion of 41 phyla in NCBI effecting millions of records in January of 2023 [41]. For evaluation of taxonomic classification it is thus important to correctly align the versions of taxonomy and current scientific names used by the classifier and the metadata of metagenomic sample.

2.4 Kraken 2 - Taxonomic Classifier

Kraken 2 is a faster and more memory efficient re-written version of the taxonomic classification program Kraken [11]. In Kraken 2 each read in a FASTQ sequence file of a metagenomic sample is classified independently (for short-read sequencing a read is 150 nucleotides long). Simplified it is based on the querying of read sub-sequences of length K against a specially built Kraken 2 database containing genomic reference sequences associated with nodes in a taxonomic tree, see Figure 2.3 (b). Some K-mers will uniquely map to nodes at the lowest taxonomic levels while others, by evolutionary relatedness or random chance, are shared by several lower taxonomic nodes and are only unique for a higher taxonomic node, here called a lowest common ancestor. For example, if the database was built with only one reference genome all the K-mers of that genome sequence would be uniquely mapped to its taxonomic node, if another genome sequence was added from a different node then fewer K-mers would be uniquely mapping to the first taxonomic node and for an infinite number of reference sequences no unique K-mers would be found.

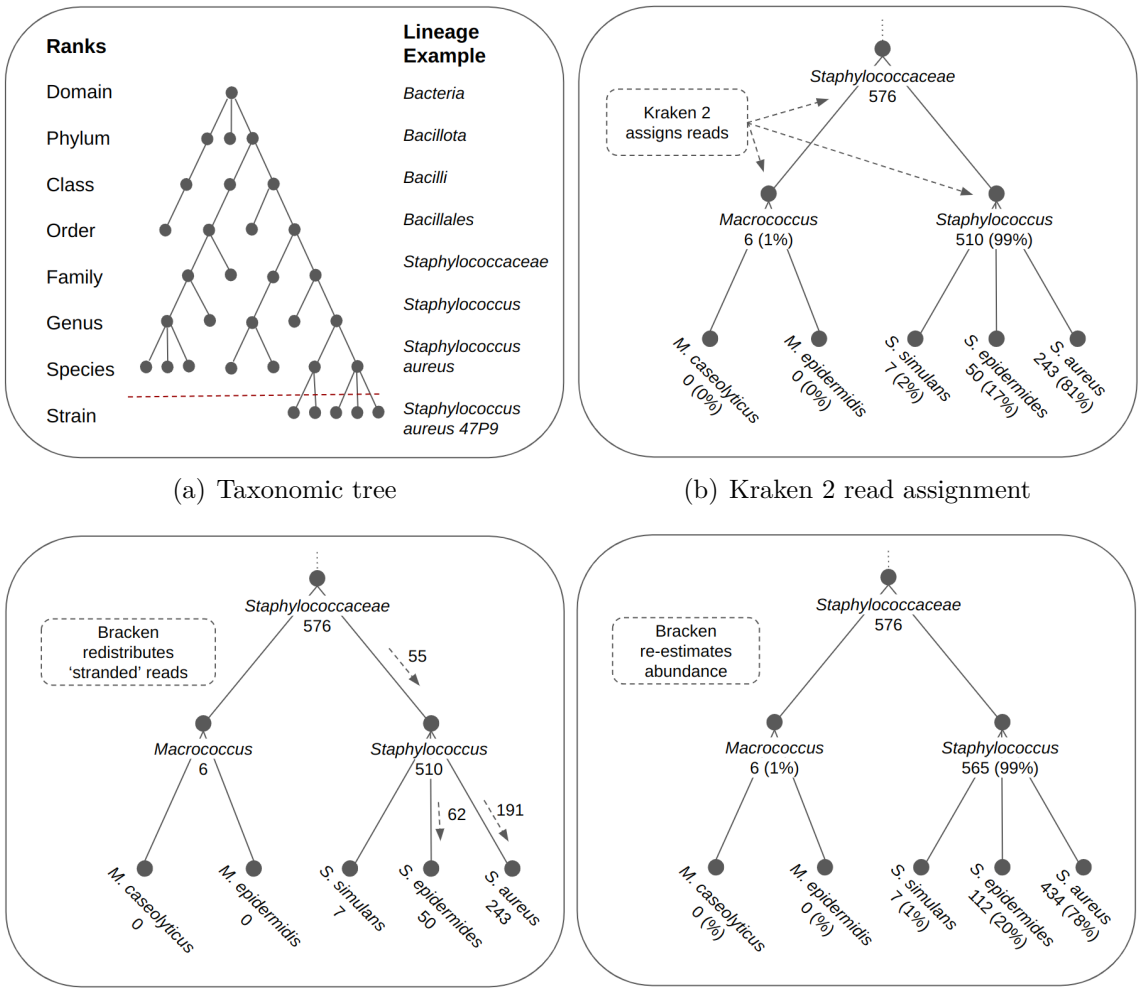
In reality the Kraken 2 database is a probabilistic compact hash table where only minimizers of the K-mers are stored, sub-strings $L<K$, and the actual querying of the K-mers in the reads against the K-mers in the database is done by minimizers on both ends using a spaced seed mask of length $S<L$ [11]. The default Kraken 2 settings of the K-mer length (`--kmer-len`), minimizer length (`--minimizer-len`) and spacer (`--minimizer-spacer`) is 35, 31 and 7 nucleotides respectively. Changes in the parameter values and their internal ratios can effect classification speed, memory usage and sensitivity [11]. The K-mer length, minimizer length and minimizer spaces are set in the database build step and are thereafter unchangeable when running classification of a sample against that particular database.

In contrast to K-mer lengths, the confidence (`--confidence`) threshold parameter can be set for each classification run and it represents the certainty of a read be-

longing to the assigned taxonomic node. The confidence score of a read label is the fraction of the number of K-mers of the read that match unambiguously to a lowest common ancestor taxonomic label in the reference database divided by the total number of K-mers on the read [42]. The total number of K-mers for a read is defined as including those that map to another label or are not present in the database but excludes K-mers with an ambiguous nucleotide. The score threshold has a non-probabilistic interpretation and can be specified in the interval [0, 1], where zero is the default. If the score of an assigned read label falls below the set confidence, the score of the next lowest common ancestor taxonomic node is calculated until the threshold is met. Increasing the confidence score setting will thus result in a trade off of recall in favor of precision [43]. For example, if the threshold was set at 0.1 and a given read scored 0.05 for species *Mycobacterium tuberculosis*, 0.15 for genus *Mycobacterium* and 0.18 for family *Mycobacteriaceae*, the read would be labeled *Mycobacterium* which is the lowest common ancestor passing the threshold. However, if the root of the taxonomic tree (domain/super kingdom) is reached without a high enough confidence score the read is considered unclassified.

In a database build step prior to the hash table, the genomic reference sequence accession numbers are mapped to the corresponding taxonomic ID number belonging to the organism they originated from. Any reference sequences and taxonomic system can be used to build custom Kraken 2 classification databases. However, Kraken 2 is adapted for default import of the current NCBI taxonomy, nomenclature and all genome sequence accessions mapped to taxonomic nodes. The NCBI RefSeq database status for each sequence accession is checked in association with the taxonomic mapping step and if the record has a suppressed status it is by default excluded from the Kraken 2 database build. For convenience there are regularly updated pre-built standard databases based on RefSeq complete genome level assemblies that include a range of mainly human associated microorganism [44]. When databases of different contents are compared, generally the more reference sequences with the higher variability used as a base for the database building gives better classification performance [43][13]. For example, in a study the best performing database was based on all the nuclotide sequences in RefSeq which resulted in a database size of >1TB [43].

When Kraken 2 is run, by default the entire Kraken 2 database is read into the random access memory (RAM) which can be problematic if the reference database is large. If instead the option `--memory-mapping` is used the database is instead read from disk however the runtime will increase as a result. For control of the database size, a limit of the final file size can be applied during the build step using the argument `--max-db-size` [11]. The same reference sequences will be represented as for the full size but with fewer than possible unique K-mers per sequence. In general, limiting the database size (for the same genome sequences) reduces the classification precision [43]. An alternative to limiting the database size in the build step is to start with fewer genomic reference sequences, possibly by representing less variation in for example each species or including genomes from fewer species in total.



(a) Taxonomic tree

(b) Kraken 2 read assignment

(c) Bracken reestimation 1

(d) Bracken reestimation 2

Figure 2.3: (a) Nodes in a taxonomic tree where the vertical level represent ranks together with an example of a ranked lineage for the bacterial species *Stphylococcus aureus* rooted at the rank of domain and branching out to the rank of strain. The taxonomic classification is limited to ranks of species and above. (b) The taxonomic classifier Kraken 2 assigns sequence reads to a node in the taxonomic tree. A read assigned to a lower rank will by default also be assigned to all the directly ascending nodes in the tree, however, the opposite is not true as reads can get 'stranded' at higher taxonomic ranks if the classification dose not reach the Kraken 2 confidence threshold for lower ranks. Note that the figure is a simplified sketch and would in reality contain all the ranks and nodes in the Kraken 2 reference databases used for the the taxonomic classification. (c) Bracken is applied on sample level after read assignment by a Kraken type classifier and it redistributes stranded reads from higher to lower taxonomic nodes. Note that the threshold for minimum number of prerequisite reads a node must have for Bracken to assign extra is here set at ten reads, and as such no stranded reads at the *Staphylococcaceae* node are redistributed to the *Macrococcus*. (d) A redistribution of reads by Bracken results in a reestimation of the sample abundance percentages, independent across each taxonomic rank, which are displayed in parenthesis.

2.5 Bracken - Abundance Reestimation

Bracken is a program for Bayesian reestimation of abundances of metagenomic samples after classification with Kraken, Kraken 2 or KrakenUniq [12][45]. Bracken has been shown to improve the overall classification performance in addition to abundance estimation when used in combination with Kraken type classifiers compared to use of only the classifier [8][43]. The tool is lightweight with runtimes less than a minute[13]. However to run Bracken, a reference database is required and it needs to be custom built for the specific Kraken, Kraken 2 or KrakenUniq database used for the classification.

Because the Kraken type classifiers perform read level classification based on the lowest common ancestor, a proportion of the reads can get stranded at nodes for higher taxonomic ranks resulting in few reads at lower levels [12]. The problem is accentuated for species in the database with high average nucleotide identity and can thus result in uncertainty in the prediction of a samples species level composition. Bracken solves the problem of low classification rates at lower taxonomic ranks by probabilistic redistribution of 'stranded' reads at nodes higher up in the taxonomic tree down towards the species rank (or any other requested taxonomic rank) [12], see Figure 2.3 (c) and (d). For a species node to receive additional reads it needs to be directly descending to the higher taxonomic node and have a minimum number of reads already assigned by the Kraken type classifier [45]. The threshold for minimum number of reads can be adjusted using the `--THRESHOLD` parameter.

3

Methods

Various reference databases, metagenomic sample datasets and software programs were used for the investigation of human DNA filtering and taxonomic classification. In this chapter, first the content of the databases and datasets are described together with the methods used to build them. Thereafter performance metrics for taxonomic classification are defined and lastly the experimental runs are described in terms of combinations of databases, datasets and classifiers used.

For orientation of where components place in the pipeline see the flowchart in Figure 3.1 where the main steps of a short-read analysis are displayed. A sample enters the pipeline in the form of a .FASTQ file, or two in the case of paired-end reads. The sample first is pre-processed by trimming low quality ends of reads and length filtering. Human origin reads can be removed from the sample by either Kraken 2 K-mer matching or genome alignment with BWA-mem2 using human reference databases. The remaining human free sample is past on to taxonomic classification and antibiotic resistance gene identification, however where only the former is further analyzed. The non-human reads are taxonomically classified with Kraken 2 using a bacterial reference database and the output is passed to Bracken for abundance re-estimation using a Kraken 2 complementary reference database. The results are collected and compared to possible sample metadata containing a theoretical sample species composition and known resistance genes. The pipeline output is a report per sample in addition to dataset summaries if samples are run in batch.

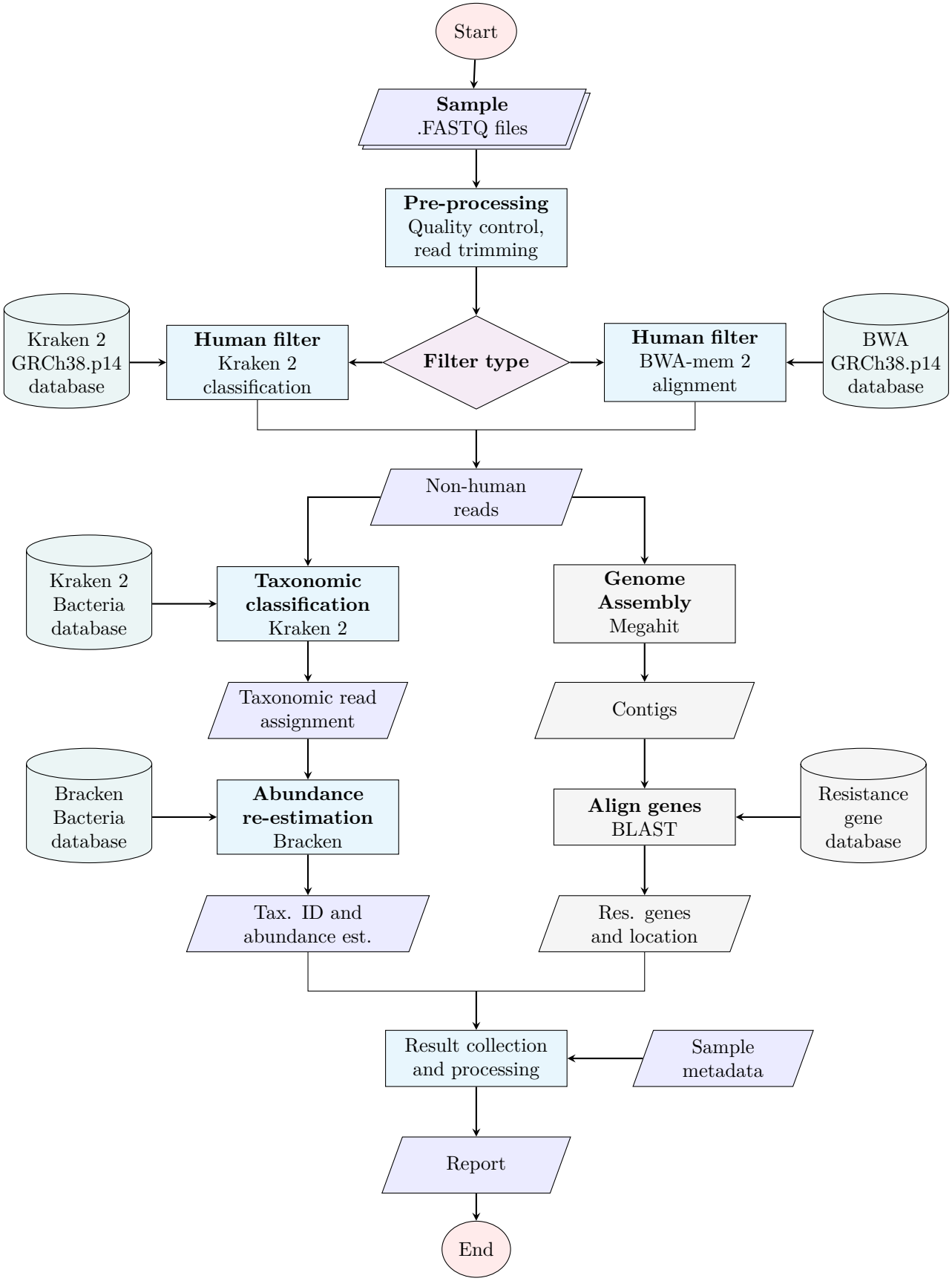


Figure 3.1: Pipeline scaffold for human host filtering, taxonomic classification and antibiotic resistance gene identification.

3.1 Databases

A collection of reference databases were built for assessing the impact of the database choice on human DNA filtering and the taxonomic classification of bacteria. An overview of the databases and their properties can be found in Table 3.1. Two databases, the Kraken 2 *db-k2-human* and the BWA-mem 2 *db-bwa-human*, were used for evaluating human origin DNA filtering. For general bacterial screening a prebuilt Kraken 2 and Bracken combined baseline database *db-standard* was compared against two pairs of bacteria only combined Kraken 2 and Bracken databases, where each set contains one with and one without plasmid sequences removed, *db-bac-comp* and *db-bac-comp-np* and then *db-bac-rep* and *db-bac-rep-np*. In addition the Kraken 2/Bracken *db-bac-subset* database containing only a subset of seven pathogenic bacteria was built for screening specifically for a selection of bacteria.

Database Name	Content	Plasmids?	Genomes	Size cap?	Size (GB)
db-k2-human	Human	-	1	no	4.6
db-bwa-human	Human	-	1	no	20.1
db-standard[44]	Archaea, bacteria, human, virus	yes	-	yes	8.0
db-bac-comp	Bacteria	yes	33k	yes	8.6
db-bac-comp-np	Bacteria	no	33k	yes	8.6
db-bac-rep	Bacteria	yes	17k	yes	8.0
db-bac-rep-np	Bacteria	no	17k	yes	8.0
db-bac-subset	Bacteria	yes	5k	no	0.4

Table 3.1: Databases for human DNA filtering, taxonomic classification of general bacteria and taxonomic classification of a subset of pathogenic bacteria. In the database names, Kraken 2 is abbreviated *k2* and BWA-mem 2 *bwa* and it indicates for which programs the databases were constructed while the other databases are combined for Kraken 2 and Bracken referencing.

3.1.1 Human Reference

Both K-mer and alignment based methods, used for detection of human reads in a sample, require an indexed database including human reference sequences. The same human genome assembly GRCh38.p14 [46] from the Genome Reference Consortium was used as the reference when building the K-mer based Kraken 2 database *db-k2-human* and the alignment based BWA-mem 2 database *db-bwa-human*.

The *db-k2-human* was built from the assembly with default database building parameter settings of a K-mer lenght of 35 nucleotides, minimizer length of 31 nucleotides and minimizer spaces of 7 nucleotides. Contrary to default settings, no masking of the human genome regions containing low complexity sequences, such as regions with repeated 'ACACACAC' or 'AAAAAAA', was performed before building the indexes with the argument `--no-mask`. Furthermore, no memory capping was applied to the build and the resulting database had the size of 4.6 GB.

For the *db-bwa-human* database each chromosome, and collected mitochondrial DNA, in the human genome assembly was indexed separately by BWA-mem 2 before being combined. The combined database size was 20.1 GB, although, as a result of the chromosome split no individual file exceeded 1 GB and could therefore be loaded entirely into memory when running on a laptop with 16 GB RAM.

3.1.2 Taxonomic Classification - General Screening

The baseline database was fetched from the Index zone collection of prebuilt Kraken 2 and Bracken combined databases and here named *db-standard* (original name Standard-8, version 2022-12-09) [44]. The collection, which is regularly updated, contains databases with various content and sizes, and is maintained by Ben Langmead who is associated with the Kraken projects [44]. According to Index zone the Standard-8 database was built using all complete level assemblies for bacteria, archaea, plasmid and virus listed on RefSeq until 2022-12-9. The human genome (GRCh38) was also included in the Standard-8 and its Kraken 2 library was created with the `--no-mask` argument which means that low complexity regions of the genome were not masked. The NCBI UniVec_Core library containing vector, adapter, linker, and primer sequences were also included in the database. The Standard-8 is the 8 GB size capped version of the Standard 64 GB database and was constructed using the exact same genome assemblies, however with less of each included, which was accomplished with the `--max-db-size` argument. Lastly, the Kraken 2 complimentary Bracken database was added to the Standard-8 database. Apart from no masking of low complexity regions of the human genome and size capping, the Kraken 2 and Bracken databases were built with defaults settings.

The combined Kraken 2 and Bracken index database *db-bac-comp* was built using only sequences from the bacterial domain. Sequences downloaded were all RefSeq bacterial genome assemblies with assembly level of complete, recorded before 2022-12-31, and with an exclusion of genomes tagged as atypical. Assemblies with a RefSeq status of suppressed were automatically excluded by Kraken 2 during the build step, which amounted to 5.3 % of the total number of downloaded sequences. Default parameter settings were used in the Kraken 2 database build apart from applying the argument `--max-db-size`. The size of the hash-map was intended to be limited to 8 GB and the same size as the *db-standard*, however, the wrong gigabyte definition was applied which resulted in a size of 8.6 GB. A Bracken database was thereafter built as an attachment to the Kraken 2 database using default settings.

A combined Kraken 2 and Bracken bacterial database without plasmids, *db-bac-comp-np*, was constructed using the same bacterial genome assemblies and parameter settings as *db-bac-comp* with the exception that the plasmid sequences were excluded from the sequence library before the database build step. The plasmids were removed by searching and deleting sequences with the word 'plasmid' in the sequence header for each assembly FASTA file.

The databases *db-bac-rep* and *db-bac-rep-np* were built using the same methods as the databases *db-bac-comp* and *db-bac-comp-np* respectively except with a different set of genome sequences. The sequences used were bacterial assemblies in the Ref-

Seq category of *Representative/Reference genomes* listed up until 2022-12-31. The assembly level of completeness ranged from contig to complete. No records were suppressed out of the 923,148 individual sequences downloaded however 8 had unidentified sequence accessions and were therefor excluded from the Kraken 2 database build.

3.1.3 Taxonomic Classification - Pathogen Subset Screening

A reference database, *db-bac-subset*, for Kraken 2 with supplementary Bracken database was built for a subset of 7 bacterial pathogens. The sequences used consisted of all RefSeq complete level assemblies of the bacteria *A. baumannii*, *E. faecium*, *E. fecalis*, *K. pneumoniae*, *P. aeruginosa*, *S. aureus* and *S. epidermidis* listed until 2022-12-31. Of the downloaded sequences, 10.2 % were automatically excluded from use in the databases as a result of having a NCBI suppressed status. The *db-bac-subset* database was built using default settings including the default of no size capping. Without size limitations the resulting database was 0.4 GB.

3.2 Simulated Metagenomic Datasets

Read simulation was used to create a set of metagenomic samples with known composition, for which taxonomic classification could be evaluated. Indeed the exact sample composition, in terms of species and their proportions, was in it self required in order to obtain high resolution classification performance metrics. The datasets of raw read shotgun metagenomic samples used in this project were all generated by simulation from already taxonomically profiled reference genomes. One dataset was adapted from an 2021 article on the evaluation of taxonomic classifiers (Parks et al. [1]) and four were simulated anew in the project using the illumina short-read simulation program InSilicoSeq (v.1.5.4) [47]. The samples in the InSilicoSeq datasets (*ds-ln-low*, *db-human*, *ds-un-low* and *ds-param-opt*) were simulated using distinct genome sequences from those used for building the reference databases, however, there was overlap in source sequences between databases and the article adapted dataset (*ds-ln-high*). An overview of the five datasets, that in total contain 250 samples, can be found in Table 3.2.

Dataset name	Number of samples	Sequence in databases?	Bacterial species per sample
ds-ln-high [1]	140	yes	50-600
ds-ln-low	40	no	10-80
ds-human	20	no	20-40
ds-un-low	44	no	10
ds-param-opt	6	no	10-80

Table 3.2: Overview of simulated datasets used.

3.2.1 Many Species Dataset - *ds-ln-high*

The *ds-ln-high* consist of 140 simulated Illumina short-read (150 bp) samples of mock prokaryotic communities adapted from Parks et al. [1]. The reads were simulated using in part the same RefSeq genome assemblies as used for the reference database build (*db-standard*). However, as the average nucleotide identity (ANI) to the reference genomes was modified in the simulation the authors estimated that using in part the same original genome assemblies for databases and sample construction would not significantly biaas the classification results.

The dataset was divided into 14 groups of 10 samples with replicate compositions, meaning that the replicates were drawn from the same distributions for factors such as species diversity, number of species and average nucleotide identity to closest reference genome. The samples included mainly bacteria but also archaea with number of species per sample draw from a normal distribution with either $\mu = 100, \sigma = 25$ or $\mu = 500, \sigma = 100$, log-normal species abundance distributions and varying diversity in strains per species. The number of nucleotides simulated per sample ($2.1{*}10^9$ bp/sample) was independent of the number of species in the sample. Following that the samples had a fixed number of reads and that the within sample species abundances were log-normally distributed, some species in samples with a high number of species had relative abundances as low as $1.9 * 10^{-6}\%$. For context, a bacterial species, that on average has a genome size of 5 million bp [48], was on the extreme low side represented in a sample by only 40 reads of 150 bp which means a «1X coverage. For a display of all sample group variations see Table A.1 in Appendix A and Parks et al. original article for a full explanation on the dataset simulation methodology.

Before the 140 samples were used the metadata was modified with new taxon names. Since the article was published in 2021 the scientific names of individual species have changed as well as major updates to phylum level names have occurred. The metadata was therefore updated using the same NCBI taxonomy as was used for building the classifier reference databases.

3.2.2 Fewer Species Dataset - *ds-ln-low*

The *ds-ln-low* dataset was created to cover a lower species complexity range in combination with a higher mean read coverage per specie genome than the *ds-ln-high*. As can be seen in Table 3.3, the dataset contains 40 metagenomic samples of log-normal distributed bacterial communities, divided into 8 groups of number of speceis and read coverage profiles. Each group has 5 replicate samples with a fixed number of species, 10, 20, 40 or 80, with a species average read coverage drawn from a log-normal distribution with the normal μ either being 10X or 20X average genome coverage. The randomness in the sampling process that effected the replicates came from which species were drawn and their specific average read coverage. Sequences used as simulation references were bacterial complete genome level assemblies in the NCBI sequence archive Genbank that were not used for building the classification reference databases. At the time of download (2023-03-17) there were 1076 bacterial species represented among the complete level assemblies, however, only 390

intersected with the species in the classification databases. The pool of genome assembly specimens drawn from to make up a sample were thus belonging to those 390 species found both among the GenBank complete level assemblies (not used for building classification databases) and the genome assemblies in the classification databases.

As an example, a sample with 10 species and a coverage μ of 20X (group *s10-c20* in Table 3.3) was generated by first sampling without replacement 10 species out of the 390 available. The corresponding genome assemblies were then identified for each of the 10 species and if there were more than one assembly per specie a random specimen was selected. Thereafter, 10 values were sampled from a log-normal distribution with $\mu = 20$ and assigned one each to the species as an average read coverage of the genome. The accessions of each sequence in the 10 assemblies (for example chromosome 1, plasmid 1A, plasmid 1B, chromosome 2, plasmid 2A, ..., chromosome 10, plasmid 10A) were paired with their assigned average genome coverage (for example 5.3X, 5.3X, 5.3X, 19.2X, 19.2X, ..., 22X, 22X) and given as input to the InSilicoSeq read simulator. The result was that the plasmids and chromosome of a bacteria had the same average coverage. Furthermore, sampling for average read coverage per species, instead of number of reads per species, took into consideration that species have differing genome sizes. InSilicoSeq then simulated the sequencing of a physical shotgun metagenomic sample of the 10 species with the listed community fractions. It generated double stranded 150 basepairs long reads covering the genome assemblies the assigned average times (5.3X, 19,2, ..., 22X) while applying sequencing error patterns according to a model trained on the Illumina NovaSeq sequencer. All the generated reads were then finally collected into two FASTQ files, one for each strand. Note that the total number of reads generated per metagenomic sample was a result of the input parameters (more species and higher coverage resulted in more reads) and thus also the size of the FASTQ files.

Sample group	**Replicates**	**No. species per sample**	**Coverage distribution** μ	**Mean no. reads per sample**
s10-c10	5	10	10x	2.2M
s10-c20	5	10	20x	4.0M
s20-c10	5	20	10x	4.6M
s20-c20	5	20	20x	9.0M
s40-c10	5	40	10x	7.0M
s40-c20	5	40	20x	15.4M
s80-c10	5	80	10x	20.1M
s80-c20	5	80	20x	31.5M

Table 3.3: The *iss-ln-rand* dataset. Log-normal distribution of community fraction taking into account that species have different genome sizes. The species pool drawn from are represented in the reference databases while the specific genome assembly specimens for those species are strictly different from the genome assemblies used for building the reference databases. Simulated paired end Illumina reads using InSilicoSeq with NovaSeq model. The samples have separate files for forward and reverse strands. Note that the listed read count per sample is for one strand.

3.2.3 Human Spike Dataset - *db-human*

The *db-human* dataset was generated by using 20 samples from the *ds-ln-low* dataset, specifically the sample groups *s20-c10*, *s20-c20*, *s40-c10* and *s40-c20*, as a bacterial base and then spiking them with human reads. The human reads were added such that the final composition of each sample was 10 % human and 90 % bacteria. The two chromosome level assemblies used for generating human reads were the KOREF_S1v2.1 [49] assembly of a Korean man and the Ash1_v2.2 [50] assembly of an Ashkenazi Jewish man. No value was placed on the specific ethnicity of the assemblies, but rather that they were from distinct individuals and of differing ethnicity from the GRCh38.p14 assembly which was used in the classification and alignment databases *db-k2-human* and *db-bwa-human*. From each assembly, 10 M NovaSeq illumina paired end reads were generated using InSilicoSeq. The required number of human reads was thereafter added to the bacterial samples by random draw from the in total 20 M simulated read pairs (sampling with replacement between each metagenomic sample).

3.2.4 Pathogen Subset Dataset - *ds-un-low*

A dataset was needed for testing if the content of classification databases could be adjusted for better detection of a subset of pathogenic bacteria. The content and structure of the *ds-un-low* dataset was thus designed in relation to the classification database *db-bac-subset*. The classification database was built on genomes from seven bacterial species, *A. baumannii*, *E. faecium*, *E. fecalis*, *K. pneumoniae*, *P. aeruginosa*, *S. aureus* and *S. epidermidis*, which belong to the five genera *Acinetobacter*, *Enterococcus*, *Klebsiella*, *Pseudomonas*, and *Staphylococcus*. The approach to the dataset is to have samples with different amounts of species and genera overlapping with the content of the databases.

To limit the size of the dataset, only two of the seven species, *A. baumannii* and *S. aureus*, were chosen as bases for sample construction. In total, the dataset has 44 samples divided into nine groups with five replicates each (one group has four samples due to a error in simulation). Each sample consists of a different set of 10 species with a fixed uniform average read coverage per genome of 20X.

As described in Table 3.4, in sample group *a-baumannii-g0* all samples contain *A. baumannii*, which is in the database *db-bac-subset*, plus nine random bacterial species which are not in the database nor belong to a genus that is present in the databases. The samples in group *a-baumannii-g1* all contain *A. baumanni*, one other random *Acinetobacter* species and then eight random bacterial species not in the database or have a genus in the database. Last in the *A. baumannii* series is group *a-baumannii-g4* that contains samples with *A. baumannii*, four other random *Acinetobacter* species plus five other random species not in the database or have a genus in the database. A similar series arrangement was thereafter done for the samples in groups *s-aureus-g0*, *s-aureus-g1* and *s-aureus-g4* with *S. aureus* as a base. As *S. epidermidis* also belongs to the genus *Staphylococcus* it was excluded from the genus draws for the *S. aureus* groups. In addition a negative control series was made. In group *control-g0* there is no overlap of species nor genera between the

samples and databases. Groups *control-g1* and *control-g4* consist of samples with one or four random species from the genera in the databases, but not the specific species, plus nine or six random species of a genera not in the database.

After the species composition of each sample was determined, corresponding reference genome assemblies were selected. With a list of accessions of the sequences in the assemblies and a fixed average read coverage of 20X, the metagenomic samples were then simulated using InSilicoSeq in same way as for *ds-ln-low*. As can be seen in Table 3.4, even if the number of species and coverage is uniform, the group mean of the number of reads per sample varies between 2.3M an 2.8M because of the different genome lengths of the species in each sample.

Sample group	**Replicates**	**Species in DB**	**Genus in DB**	**Species *not* in DB**	**Genus *not* in DB**	**Mean no. reads**
control-g0	5	0	0	10	10	2.6M
control-g1	5	0	1	10	9	2.8M
control-g4	4	0	4	10	6	2.5M
a-baumannii-g0	5	1	1	9	9	2.7M
a-baumannii-g1	5	1	2	9	8	2.8M
a-baumannii-g4	5	1	5	9	5	2.5M
s-aureus-g0	5	1	1	9	9	2.3M
s-aureus-g1	5	1	2	9	8	2.4M
s-aureus-g4	5	1	5	9	5	2.4M

Table 3.4: The *iss-path-subset* dataset contains 44 samples with InSilicoSeq simulated paired end Illumina reads. One sample in group *control-g4* was assigned the wrong content when simulating and therefor excluded. Uniform average genome coverage of 20x for all species in the samples. Five replicates per subgroup, ten species in each sample with reads drawn for a uniform coverage distribution of 20X genome. Note that the listed read count per sample is for one strand.

3.2.5 Parameter Optimization Dataset - *ds-param-opt*

Extra metagenomic samples were created for parameter tuning of the taxonomic classifier. The *ds-param-opt* dataset is a mixed collection of six InSilicoSeq simulated samples similar to the samples of *ds-ln-low* and *ds-un-low*. The samples have 10-80 species randomly drawn from the same pool as for the *ds-ln-low* dataset with either log-normal or fixed uniform abundance distributions. For the composition of each sample see Table 3.5.

Sample Name	No. species	Average genome coverage	No. reads
sample-ln-s10-c20	10	from log-normal dist. $\mu = 20$x	2.9M
sample-ln-s15-c20	15	from log-normal dist. $\mu = 20$x	1.7M
sample-un-s20-c15	20	fixed uniform of 15x	1.7M
sample-ln-s25-c10	25	from log-normal dist. $\mu = 10$x	4.3M
sample-ln-s40-c15	40	from log-normal dist. $\mu = 15$x	11.4M
sample-un-s80-c20	80	fixed uniform of 20x	20.4M

Table 3.5: Composition of the *db-param-opt* dataset.

3.3 Classification Metrics

When a metagenomic sample in the digital FASTQ format is passed through the taxonomic classification step the simplified result is a list of predicted species (predictions are also made at all other taxonomic ranks) and the number of reads classified per species by Kraken 2 and then the re-estimated number by Bracken. As the samples used in this project were simulated, the true species composition was known which is in contrast to real world samples taken from physical microbiomes. In theory each simulated read fragment could be traced to its origin organism, however, in this project the species metadata was on sample level with only the total number of reads per species known. As such the classification performance was also measured on sample level and not if each individual read fragment was correctly classified.

A True Positive (TP) observation is here defined as a species predicted by the Kraken 2 and Bracken combination that was actually in the sample, irregardless of if the predicted and actual abundances (>0) were the same. The question of precision in abundance prediction was not considered in the project scope. A False Positive (FP) observation is defined as occurring when a species is predicted that was not actually in the sample and reversely a False Negative (FN) is when a species actually in the sample was not predicted by the taxonomic classifier. The True Negative (TP) observations would theoretically be all the, potentially thousands, species in the database not actually present in the sample and that were also not predicted to be in the sample. It is common to apply a post classification filter to discount predicted species below a certain threshold read count or abundance percentage before calculating the TP, FP and FN [8]. However, the raw output from Bracken was here considered the predicted species if not otherwise specified.

Classification metrics can be derived from the sum of the TP, FP and FN observations for a sample. The metrics applied here for measurement of taxonomic classification performance were True Positive Rate (TPR), Positive Predictive Value (PPV), False Discovery Rate (FDR) and F-score (F_β). The true positive rate, Equation 3.1, also known as recall, is defined as the the number of true species found by the classifier divided by the actual number of species in the sample.

$$TPR = \frac{TP}{TP + FN} \tag{3.1}$$

Positive Predictive Value, Equation 3.2, also know as precision, is defined as the proportion of species that the classifier correctly predicts being in the sample over the total number of species that are predicted.

$$PPV = \frac{TP}{TP + FP} \tag{3.2}$$

False Discovery Rate, Equation 3.3, is the number of species that the classifier falsely predicts being in the sample divided by the total number of predicted species. The false discovery rate is the inverse of the positive predictive value.

$$FDR = \frac{FP}{TP + FP} = 1 - PPV \tag{3.3}$$

The F_β score, Equation 3.4, is a weighted mean of the true positive rate and positive predictive value, where β is the factor of importance of the true positive rate in relation to the positive predictive value.

$$F_\beta = \frac{(1 + \beta^2) * TP}{(1 + \beta^2) * TP + \beta^2 * FN + FP} \tag{3.4}$$

The F_β score for $\beta = 1.5$ was mainly used here, where the true positive rate was weighted 1.5 times the positive predictive value. Although, the more commonly used metric for comparing taxonomic classifiers is the F_1 score and it was used in addition to $F_{1.5}$ [8].

3.4 Experimental Runs

In this project, a sample *run* refers to two FASTQ files with simulated paired-end reads of a shotgun metagenomic sample being processed through the pipeline together with a metadata file on the samples theoretical content. Because the focus of the investigation was on human DNA filtering and taxonomic classification, the analysis steps surrounding genome assembly and identification of antibiotic resistance genes were switched off to save time. However, minimal pre-processing in the form of read quality trimming and filtering was applied for each run using the program Fastp (v.0.23.2) [51].

The samples in the *db-human* dataset was run with both the Kraken 2 (v.2.1.2) and BWA-mem 2 (v.2.2.1) human read filtering methods. The reads in the samples were first classified with Kraken 2 (confidence parameter = 0.2) and the *db-k2-human* reference database as either human or non-human. The non-human reads were then passed on to taxonomic classification with the Kraken 2 (confidence parameter = 0.075) and Bracken (v.2.8) (read threshold = 750 reads) using the *db-standard* database. Thereafter all the reads, divided into the categories *db-k2-human* classified, *db-standard* classified and *db-standard* unclassified, were aligned to the human genome using BWA-mem 2 and the *db-bwa-human* database. The total number of reads aligned per sample and Kraken 2 classification categories could then be established.

The three larger datasets, *ds-ln-high*, *ds-ln-low* and *ds-un-low*, were run through the pipeline for taxonomic classification with Kraken 2 and Bracken using the default parameter settings of Kraken 2 (confidence = 0.0) and Bracken (read threshold = 10 reads)[1]. The runs were repeated for each of the five bacterial screening databases *db-standard*, *db-bac-comp*, *db-bac-comp-np*, *db-bac-rep*, *db-bac-rep-np*. Additionally, even though the datasets theoretically do not include human genomes, they were all run with prior Kraken 2 human filtration step (confidence parameter = 0.2).

The six samples in the *ds-param-opt* dataset were run repeatedly for Kraken 2 confidences of 0.0-0.2 and Bracken read thresholds of 10-1250 reads for the *db-standard* reference database. The datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* were thereafter re-run for all the screening databases with the combination of the Kraken 2 confidence (0.075) and Bracken read threshold (750 reads) which gave the highest $F_{1.5}$ score for *iss-aram-opt*. Furthermore the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* were run again for the five screening databases but with a more conservative parameter value increase of Kraken 2 confidence = 0.05 and Bracken minimum number reads = 50, compared to the optimized. Lastly the pathogen *ds-un-low* dataset was taxonomically classified with the bacterial subset databas *db-bac-subset* using Kraken 2 with confidence = 0.3 and Bracken with a minimum number read threshold of 10,000 reads.

[1]The default minimum number of reads classified by Kraken 2 at a taxon node for Bracken to redistribute them is set to 0 in the source code but stated as 10 in the documentation (v.2.8) [45]. The threshold of 10 reads is in this project used as the default.

4

Results and Discussion

4.1 Human DNA Filtering

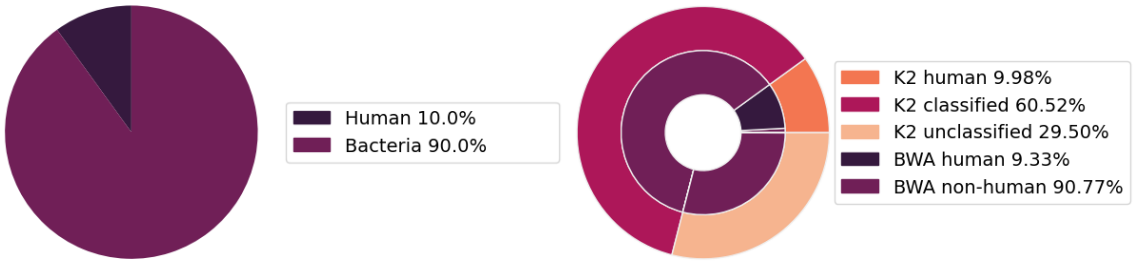
Assembly of the genomes in the metagenomic sample, that is reconstructing the genomes of the organisms from the reads fragments in the sample, is a step needed before the search for antibiotic resistance genes. However, the process of assembling genomes from a metagenomic sample is expensive in terms of computational time and memory and it is therefore resourceful to not assemble reads that have a non-bacterial origin, such as human, as they do not carry the resistance genes. Ideally exactly all human reads would be removed and none of the reads with other origins, however, in reality that is hard to achieve. The threshold could either be set to miss some human reads or incorrectly filter out some bacterial reads. Arguably it would be better to have the assembly take slightly longer time by removing most but not all human reads compared to potentially removing bacterial reads carrying important information.

The first approach to filtering out human reads was to do it in association with taxonomic classification using Kraken 2 and the *db-standard* database as it contains a human reference in addition to microorganisms. The filtration was unsatisfactory as it was observed that a substantial proportion of the human reads were classified as microorganisms, mostly fungi, or were unclassified. Adjusting the Kraken 2 confidence parameter for better human filtration would however at the same time impact the classification of bacteria. As the *db-standard* database was size restricted when building, it was hypothesized that running against an unrestricted K-mer indexation of the human genome, that is the *db-k2-human* database, could catch a larger proportion of the human reads. Although, building a Kraken 2 database with only one species is contrary to the principle of Kraken 2 classification based on a database with distinct K-mers brought out by the internal uniqueness of K-mers among the genomes used to build the database. To counter balance for only having one species, and thus low variance, in the database, the Kraken 2 confidence parameter was increased from default of 0.0 to 0.2 which would require more human K-mer matches per read in order for the read to be classified. In addition, applying a human read filtering step separate from the taxonomic classification enables independent parameter settings.

An alternative method to K-mer based classification of human reads is direct alignment of reads to the human genome and it was tested for comparison purposes. The BWA-mem 2 sequence alignment program was chosen because it is a widely

used short-read aligner and is among the top performers. bwa-mem was shown to have a lower runtime per read than one of the closest competitors Bowtie 2, while having similar alignment performances [52]. Thereafter, BWA-mem 2 was released and shown to be faster than bwa-mem [53].

The theoretical composition of each of the 20 samples in the *ds-human* dataset is 10 % human origin reads and 90 % bacteria. The result of running the samples in the dataset for the Kraken 2 database *db-k2-human* (confidence = 0.2) and then *db-standard* (confidence = 0.075) on the non-human filtered reads was an average 9.98 % of the reads classified as human by *db-k2-human*, 60.52 % classified by *db-standard* and 29.50 % were unclassified by *db-standard*, see Figure 4.1. When all the reads in the samples then were aligned to the human genome with BWA-mem 2, on average 9.33 % of the reads aligned. Moreover if the aligned reads were subdivided into the Kraken 2 classification categories, as seen in Table 4.1, then 92.84 % of the Kraken 2 *db-k2-human* classified reads aligned to human genome, 0.03 % of the *db-standard* classified aligned and 0.16 % of the *db-standard* unclassified. As the the human reads were not traced on read level one can not draw the immediate conclusion that because 9.98 % of the reads in the samples were classified as human by Kraken 2 that they truly were human reads. That on average 92.84 % of the Kraken 2 human classified reads in turn also did align to the human genome using BWA-mem 2 strengthens the indication that the majority of reads filtered out are truly human. Another indication is that <1% of the reads classified by Kraken 2 to not be human (either unclassified or classified as a microorganism) aligned to the human genome using BWA-mem 2.



(a) Theoretical composition

(b) Kraken 2 outer, bwa-mem 2 inner

Figure 4.1: Average percent of reads filtered out as human in the *iss-human* dataset where (a) each sample independent of size has a theoretical composition of 10 % simulated human reads and 90 % bacterial reads. The outer doughnut in (b) represents the average proportions of reads that are filtered out by *k2-human* and classified or unclassified by *k2-b-standard*. The inner doughnut represents the reads that do or do not align to the human genome using BWA-mem2 in relation to the outer doughnut. It can be seen that the human aligned reads are concentrated by the Kraken 2 human classified reads.

Sample group	% K2 human filter of total	% BWA align to human of total	% BWA align of K2 human filter	% BWA align of K2 classified	% BWA align of K2 unclassified
s20-c10-h10	9.98 [9.97,9.99]	9.68 [9.53,9.78]	96.34 [95.04,97.44]	0.03 [0.02,0.03]	0.18 [0.09,0.38]
s20-c20-h10	9.97 [9.97,9.98]	9.37 [8.91,9.56]	93.21 [88.73,95.28]	0.03 [0.02,0.05]	0.18 [0.13,0.25]
s40-c10-h10	9.97 [9.97,9.98]	9.35 [9.13,9.56]	93.17 [91.10,95.34]	0.03 [0.02,0.03]	0.14 [0.10,0.20]
s40-c20-h10	9.98 [9.97,9.98]	8.91 [8.74,9.13]	88.66 [87.00,90.95]	0.03 [0.02,0.04]	0.14 [0.12,0.16]
Full dataset	9.98 [9.97,9.99]	9.33 [8.74,9.78]	92.84 [87.00,97.44]	0.03 [0.02,0.05]	0.16 [0.09,0.38]

Table 4.1: Divided by sample group (5 replicates i each group) of the *ds-human* dataset the average [min, max] percentages of reads classified by Kraken 2 (K2) and bwa-mem 2 (BWA). The theoretical composition of the samples are 10 % human and 90 % bacteria.

The percentages of reads filtered with Kraken 2 and BWA-mem 2 are consistent across sample with differing number of reads and number of other bacterial species in the sample, which can be seen across the sample groups in Table 4.1. The result is expected because the classification and alignment is independent for each read and more or less other bacterial reads should not effect the human reads.

The parameter settings for Kraken 2 and BWA-mem 2 were not optimized for human filtration and it is likely that both approaches could be equally as good at filtering if a rigorous parameter optimization was performed. Kraken 2 however strongly outperformed BWA-mem 2 for speed as it on average, for the 20 samples run, was 94 times faster.

Can the results be applied to real world samples from the human microbiome? The main limitation is that the human reads in the test dataset were simulated and were of a relatively high sequencing quality which would be expected to vary in real samples. Furthermore, when simulating the human reads, the coverage was uniformly distributed across the genome, which is unlikely for a real world human genome. Also noteworthy is that the Kraken 2 confidence was set for classification with the *db-k2-human* database for separating human reads from bacterial reads, hence it is questionable if the high performance withholds for human filtering from more similar types of organisms such as other Eukaryotes.

In a broader perspective, if a similar pipeline were to be used for example in a veterinary setting then the genomes of animals in question would be used to build the filtration database. This is relevant since the problem of antibiotic resistance is present in agriculture too. It could possibly be easier to construct a Kraken 2 database with the animal species in question compared to development of a physical host DNA extraction kit to apply pre-sequencing.

4.2 Classifier, Parameters and Thresholds

Kraken 2 is know for its speed and high recall for microbial classification but also for its low precision [1][43]. These features were observed in this project, where false discovery rates of higher than 90 % were common for default Kraken 2 and Bracken parameter settings which meant that a majority of the species identified by the classifier were false positives. What an acceptable false discovery rate is would depend on the specific intended application of the classification results, however, a rate of > 90 % is likely unacceptable for most clinical screening purposes. Implementing measures to control the false discovery rate would generally also effect the true positive rates. The goal was therefor to find an approach that reduced the false discovery rate as much as possible without severely reducing the true positive rate. The main classification metric used for harmonizing the two was the $F_{1.5}$ score. The $F_{1.5}$ score, where true positive rate is weighted higher than the positive predictive value, was chosen over the more common F_1 score because it was considered more important to not miss potentially dangerous pathogens than to have false alarms. The intention was for the taxonomic classification step to be part of a fast screening pipeline and if bacterial species of interest were found then further tests with higher precision could possibly be done for increased certainty.

There are several approaches to the problem of lowering the false discovery rate. For example, adjustments can be applied to parameters in the Kraken 2 read classification, which reads to be considered for abundance re-estimation by Bracken and by setting post classification filtering thresholds of minimum estimated community abundance or minimum number of estimated reads assigned to a taxon for it to be considered found. Taking it further, the pre-processing steps before taxonomic classification, such as the length and quality trimming, could also play a role but were not considered within the project scope. Tuning of Kraken 2 and Bracken parameters were first considered.

4.2.1 Kraken 2 and Bracken Parameters

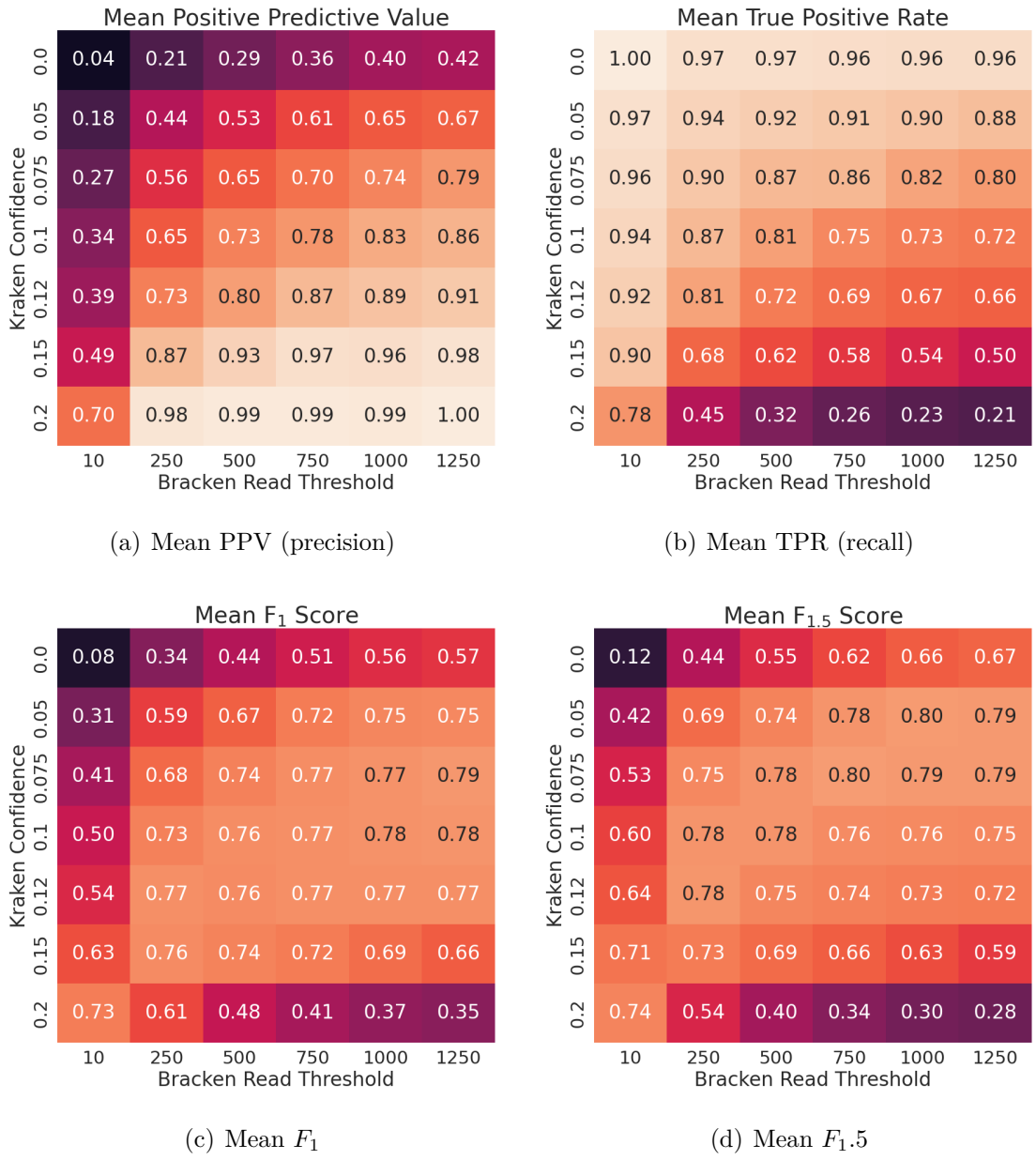
The parameters considered were the read classification confidence for Kraken 2 and the threshold for minimum number classified reads required for abundance reestimation by Bracken. Other Kraken 2 parameters, such as K-mer length or minimizer length, are also relevant for classification performance as was shown in the Kraken 2 method article [11]. These parameters are however tied to the Kraken 2/Bracken database construction which in turn can be cumbersome to rebuild for a variety of settings with limited computational resources. The confidence and read threshold were on the other hand adjustable for each sample run without re-building the reference databases.

Ideally a dataset with a large variety in sample complexity, distribution and species diversity, would have been used for finding the optimal parameter settings for a general metagenomic sample or a sub-group of samples. However, due to restrictions in time and sample simulation abilities the number of samples in the dataset used for parameter optimization *ds-param-opt* was held low and with fewer species. It is

important to note that the six sample dataset *ds-param-opt* used for the optimization was not a representation of all possible shotgun metagenomics samples, and contains samples with lower species diversity than for example the *ds-ln-high* and simulated from a pool of only 390 species. The motivation behind preforming an optimization was not to find the universal optimal metagenomic parameter settings but rather to demonstrate effects of varied parameter settings for the same dataset in relation to the default.

The samples in the *ds-param-opt* dataset were taxonomically classified using the *db-standard* database for a range of Kraken 2 confidence settings and thresholds for Bracken minimum number of reads required for read redistribution. In Figure 4.2 the average positive predictive value, true positive rate, F_1 and $F_{1.5}$ scores for the samples in the dataset are shown for Kraken 2 confidences [0.0, 0.2] and Bracken read thresholds [10, 1250]. The results are only in relation to bacteria and possible false positives from other domains were not include. For default settings, the dataset average true positive rate was 100 % which means that all species in the samples were identified. The positive predictive value was however only 4 % which means that 96% of the species predicted by the classifier were false positives. The resulting average $F_{1.5}$ score was 12 %. Increasing either the confidence or the read threshold increased the positive predictive value and decreased the true positive rate. Applying both a high confidence (0.2) and read threshold (1250 reads) instead gave an average positive predictive value of 100 % however with the cost of decreasing the true positive rate to 21 %, resulting in a $F_{1.5}$ score of 28 %. The $F_{1.5}$ optimized parameter settings were a Kraken 2 confidence of 0.075 and Bracken read threshold of 750 reads which gave an average $F_{1.5}$ score of 80 %. It was thus clear that the default parameter settings were not optimal for the *ds-param-opt* dataset.

A reason for why false positive observations occur, aside from database misannotation, is because a portion of the reads from a species truly in the sample have a higher K-mer match to other closely related species. If a species truly in the sample is in high abundance then the proportion of reads originating from it that are misclassified can appear significant even if the classification error rate is low on read level. Furthermore, the misclassifications have been shown to be systematic where finding a certain species in high abundance is correlated to also falsely finding specific species in low abundance [54]. The systematic error was more prevalent in Kraken 2 compared to marker gene type taxonomic classifiers such as MetaPhlAn [54]. That there are systematic errors are however not surprising since the species have varying levels of phylogenetic relatedness. The problem of systematic misclassifications could then worsen if the abundance of the false positives are artificially inflated by Bracken's redistribution of reads at higher taxonomic levels to lower levels. A threshold for Bracken abundance reestimation in the form of a minimum fraction of reads classified by Kraken 2 as a specific taxon could be more scalable than using an absolute number of reads. If a choice had to be made between adjusting Kraken 2 or Bracken, arguably the Kraken 2 confidence would be increased form default because its effects are on read level and are independent of the number of species or the number of reads in the sample.



(a) Mean PPV (precision)

(b) Mean TPR (recall)

(c) Mean F_1

(d) Mean $F_1.5$

Figure 4.2: Average species level classification metrics for samples in the dataset *ds-param-opt* re-run for Kraken 2 confidence parameter settings between 0.0-0.2 on the x-axis and a Bracken read threshold between 10 and 1250 on the y-axis. The reference database used was the *db-standard*. The highest $F_{1.5}$ score was observed for a Kraken 2 confidence of 0.075 in combination with a Bracken minimum number read threshold of 750 reads.

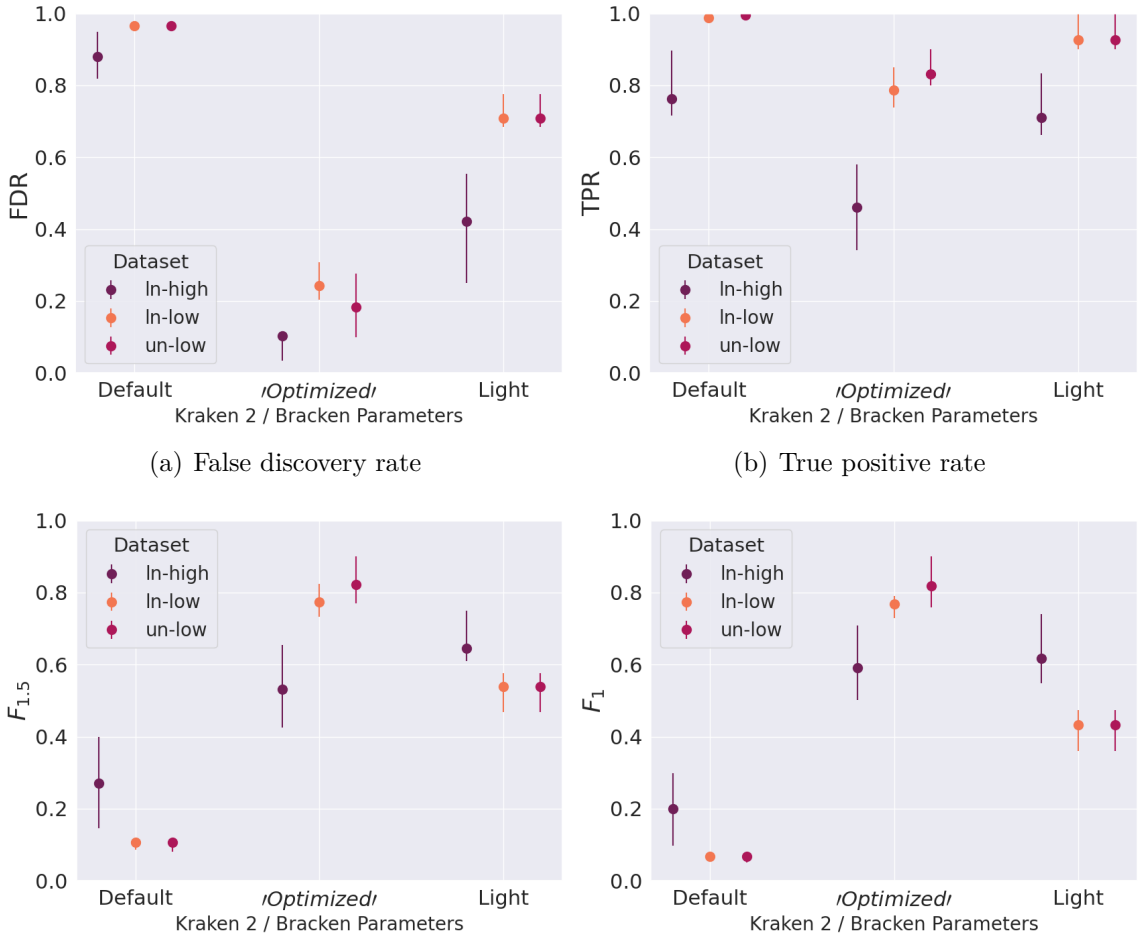
The samples in the larger *ds-ln-high*, *ds-ln-low* and *ds-un-low* datasets were taxonomically classified with the Kraken 2 and Bracken combination using different databases for the default parameter settings and then re-run for 'optimized' parameter settings (confidence = 0.075, read threshold = 750). In Table 4.2 the resulting dataset averages of species $F_{1.5}$, true positive rates and false discovery rates are listed for five databases while Figure 4.3 graphically shows trends only for the baseline *db-*

standard database.[1] For default Kraken 2/Bracken parameters and the *db-standard* database, the average $F_{1.5}$ scores were 27.1 %, 10.6 % and 10.6 % for the *ds-ln-high*, *ds-ln-low* and *ds-un-low* datasets respectively. When instead optimized parameters were used for classification the average scores increased to 53.1 %, 77.4 % and 82.2 % for the three datasets. That the relative improvement was larger for the *ds-ln-low* and *ds-un-low* datasets, compared to *ds-ln-high*, was not surprising because they were more similar to the dataset used for $F_{1.5}$ optimization in terms of species diversity, sequencing depth and program used for simulation. Furthermore, the datasets average false discovery rates decreased from 88.0 %, 96.5 % and 96.5 % for the default parameters to 10.3 %, 24.2 % and 18.3 % for the optimized. In contrast, the largest decrease was observed for the *ds-ln-high*, however, the true positive rate also decreased the most from 76.4 % to 46.0 %. A probable reason for both lower true positive rates and false discovery rates for the *ds-ln-high* is that the samples have a higher proportion of species with low genome sequence coverage (less than the Bracken 750 reads threshold) compared to the *ds-ln-low* and *ds-un-low*.

As the exact same sample datasets were used in each experimental runs, the averages of classification performance metrics can be compared between different databases and parameter settings, however it is important to note that the spread across samples within a dataset can be large, as can be observed in Table 4.2 where approximately 10-20 percentage points differ between the 25 % and 75 % percentiles. Also noteworthy is that the average classification performance of samples in a dataset would not be relevant when for example screening a patient and a confidence interval on performance for a certain type of sample would be more applicable.

For a future unknown metagenomic sample, with possibly a low sequencing depth, using a Kraken 2 confidence 0.075 and Bracken read threshold of 750 reads which was optimal for the *ds-param-opt* dataset could be too aggressive. As an example take the the sharp decrease in true positive rate for the *ds-ln-high*. For the *ds-param-opt*, the mean $F_{1.5}$ increased steeply for increased Bracken read thresholds between 10 and 250 reads (not shown in the results) and it was seen that a small change in parameter values could make a difference. It was therefore hypothesized that a smaller parameter adjustment could increase the $F_{1.5}$ score while being 'safe' for more diverse types of sample compositions. The datasets were re-run again for parameter settings of Kraken 2 confidence 0.05 and Bracken read threshold of 50 reads. The results were that the *ds-ln-high* performed better while the *ds-ln-low* and *ds-un-low* performed worse compared to classification with 'optimal' parameter settings. Classification with a conservative parameter adjustment however performed better for $F_{1.5}$ than with default settings for all the datasets.

[1] Additional F_1 scores are listed in Appendix B together with the genus level classification performance for each combination of parameter setting, dataset and database.



(a) False discovery rate

(b) True positive rate

(c) $F_{1.5}$ score

(d) F_1 score

Figure 4.3: Mean [25 percentile, 75 percentile] classification metrics for Kraken 2 confidence and Bracken minimum number of reads threshold parameter settings of default (conf. = 0.00, min reads = 50), optimized (conf. = 0.075, min reads = 750) and light (conf. = 0.05, min reads = 10) across datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* run with the *db-standard* reference database. (a) false discovery rates (FDR), (b) true positive rates (TPR), (c) $F_{1.5}$ scores and (e) F_1 scores.

The conclusion is that the default Kraken 2 confidence and Bracken read threshold may be far from optimal. What is optimal parameter settings can vary depending on sample type and that for future implementation it could be good to optimize specifically for the range of sequence coverage, species number and diversity that is relevant for the use case. For example if the pipeline would be used for identifying bacteria in skin swabs, samples from a skin microbiome could be simulated and used for a parameter optimization. A conservative threshold increase from default to confidence of 0.05 and read threshold of 50 reads could be a compromise for unknown samples.

	Dataset	Database	$F_{1.5}$ (%)	TPR (%)	FDR (%)
Default (conf. 0, min reads 10	ds-ln-high	db-standard	27.1 [14.6,40.0]	76.3 [71.5,89.6]	88.0 [81.8,94.8]
		db-bac-comp	25.3 [13.1,37.3]	75.1 [69.7,89.1]	89.0 [83.6,95.5]
		db-bac-comp-np	24.2 [12.6,35.6]	75.1 [69.7,89.1]	89.7 [84.6,95.7]
		db-bac-rep	24.5 [14.2,37.6]	91.5 [88.5,97.8]	90.0 [83.7,95.1]
		db-bac-rep-np	24.0 [13.9,36.8]	91.5 [88.5,97.9]	90.3 [84.1,95.2]
	ds-ln-low	db-standard	10.6 [8.7,11.4]	98.7 [98.4,100.0]	96.5 [96.2,97.2]
		db-bac-comp	10.5 [8.6,11.4]	99.5 [100.0,100.0]	96.5 [96.2,97.2]
		db-bac-comp-np	9.5 [7.8,10.2]	99.5 [100.0,100.0]	96.9 [96.6,97.5]
		db-bac-rep	5.0 [4.0,5.7]	93.4 [90.0,97.8]	98.4 [98.2,98.7]
		db-bac-rep-np	4.8 [3.9,5.6]	93.4 [90.0,97.8]	98.4 [98.2,98.8]
	ds-un-low	db-standard	10.6 [8.0,11.8]	99.5 [100.0,100.0]	96.5 [96.1,97.4]
		db-bac-comp	10.3 [8.2,11.7]	100.0 [100.0,100.0]	96.6 [96.1,97.3]
		db-bac-comp-np	9.3 [7.1,10.9]	100.0 [100.0,100.0]	96.9 [96.4,97.7]
		db-bac-rep	5.1 [3.9,5.9]	95.2 [90.0,100.0]	98.4 [98.1,98.8]
		db-bac-rep-np	4.9 [3.8,5.6]	95.2 [90.0,100.0]	98.4 [98.2,98.8]
Optimal (conf.0.075, min reads 750)	ds-ln-high	db-standard	53.1 [42.5,65.4]	46.0 [34.1,58.0]	10.3 [3.4,5.4]
		db-bac-comp	55.8 [46.3,68.3]	49.4 [37.9,62.2]	13.2 [4.7,19.8]
		db-bac-comp-np	56.1 [46.8,68.7]	49.7 [38.4,62.4]	13.5 [4.7,19.4]
		db-bac-rep	45.5 [32.6,59.2]	38.0 [25.6,51.8]	7.7 [3.4,9.3]
		db-bac-rep-np	45.4 [32.6,59.2]	38.0 [25.2,51.2]	8.0 [3.6,9.5]
	ds-ln-low	db-standard	77.4 [73.2,82.4]	78.7 [73.8,85.0]	24.2 [20.4,30.8]
		db-bac-comp	76.8 [73.8,80.8]	81.9 [80.0,85.0]	31.5 [27.2,37.5]
		db-bac-comp-np	75.7 [72.6,79.3]	82.4 [80.0,86.2]	35.2 [28.4,41.2]
		db-bac-rep	72.4 [67.5,76.8]	70.9 [67.2,75.0]	22.1 [13.5,26.7]
		db-bac-rep-np	71.9 [67.9,76.6]	70.4 [66.2,75.6]	22.5 [15.7,26.8]
	ds-un-low	db-standard	82.2 [77.0,90.0]	83.2 [80.0,90.0]	18.3 [10.0,27.6]
		db-bac-comp	82.6 [77.6,90.0]	87.7 [80.0,90.0]	25.6[16.1,35.7]
		db-bac-comp-np	81.1 [73.8,88.0]	87.7 [80.0,90.0]	29.5 [23.1,38.5]
		db-bac-rep	78.2 [71.7,85.2]	76.4 [70.0,80.0]	15.7 [9.8,20.6]
		db-bac-rep-np	78.0 [71.7,85.2]	76.1 [70.0,80.0]	15.7 [9.8,20.0]
Light (conf.0.05, min reads 50)	ds-ln-high	db-standard	64.5 [60.9,74.9]	71.0 [66.2,83.4]	42.2 [25.0,55.4]
		db-bac-comp	61.7 [57.0,73.8]	70.8 [64.0,83.5]	47.9 [29.8,62.0]
		db-bac-comp-np	61.1 [55.3,73.0]	70.9 [64.1,83.8]	49.0 [30.5,63.0]
		db-bac-rep	70.9 [64.2,78.0]	81.5 [77.2,89.6]	40.4 [25.1,54.6]
		db-bac-rep-np	70.6 [63.8,77.9]	81.4 [77.3,89.4]	41.0 [26.0,55.3]
	ds-ln-low	db-standard	53.9 [46.9,57.6]	92.7 [90.0,100.0]	70.9 [68.5,77.5]
		db-bac-comp	49.7 [46.0,53.9]	93.5 [90.0,96.2]	75.5 [72.3,78.3]
		db-bac-comp-np	47.6 [44.4,52.5]	92.9 [90.0,96.2]	77.0 [73.7,79.8]
		db-bac-rep	45.0 [41.7,50.3]	87.4 [83.8,91.6]	77.9 [74.8,80.9]
		db-bac-rep-np	44.5 [40.9,47.7]	87.2 [82.5,91.6]	78.2 [76.2,82.2]
	ds-un-low	db-standard	53.9 [46.9,57.6]	92.7 [90.0,100.0]	70.9 [68.5,77.5]
		db-bac-comp	49.4 [42.7,54.9]	93.2 [90.0,100.0]	75.1 [72.6,80.1]
		db-bac-comp-np	47.2 [40.8,51.3]	94.1 [90.0,100.0]	77.1 [74.7,80.9]
		db-bac-rep	44.2 [37.9,49.3]	90.2 [90.0,100.0]	79.0 [76.1,83.5]
		db-bac-rep-np	43.6 [37.3,48.5]	90.5 [90.0,100.0]	79.5 [76.8,83.9]

Table 4.2: Species level taxonomic classification metrics $F_{1.5}$ score, true positive rate (TPR) and false discovery rate (FDR) (dataset sample mean [25 percentile, 75 percentile]) for different combinations of datasets and general bacterial screening databases, repeated for three Kraken 2 and Bracken parameter settings (left most column).

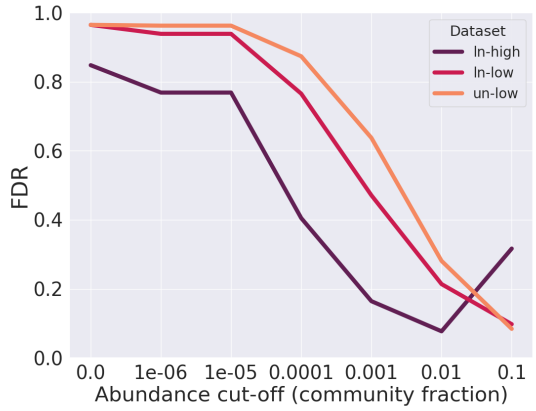
4.2.2 Post Classification Filtering

Following that false positive species predictions by Kraken 2 are over represented at low abundances it is common to apply a post classification filter for removing species with low estimated abundance or low read counts from the list of predicted species [1]. When Kraken 2 and Bracken are applied for taxonomic classification of samples from single isolate bacterial cultures, high false discovery rates at low abundances could be tolerable because it would still be clear what the main cultured bacteria was if a species abundance was estimated to for example 95 % and then 10 other species to a combined 5 %. For metagenomic samples from a microbiome the true species abundances are however expected to have a non-uniform distribution often with a heavy low abundance tail [55][56]. Distinguishing true positive low abundance species from false positives in metagenomic samples is thus an added challenge compared to samples from single isolate cultures.

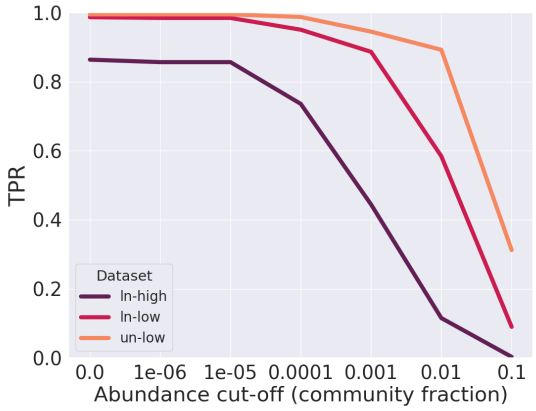
Post classification filtering was tested by disregarding species in the Bracken result output below abundance thresholds, that is community fractions, for a range up to 10 % of the sample. In Figure 4.4 average classification metrics for the samples in the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low*, when run with the *db-standard* database for default parameter settings, are plotted for increasing cut-off values in Bracken estimated abundance. The average false discovery rate start at 88 % for the *ds-ln-high* and 97 % for both the *ds-ln-low* and *ds-un-low* datasets and then decrease with an increasing abundance threshold to being below 30 % for an abundance threshold of 1 % (that is species estimated to be less than 1 % of the sample are disregarded) as can be seen in Figure 4.4 a). The average true positive rates unfortunately also decrease towards zero with an increasing abundance cut-off threshold as displayed in Figure 4.4 b). The average true positive rate for the datasets can differ for the same abundance threshold, for example at a 1 % threshold the difference is approximately 80 percentage points between the *ds-ln-high* and the *ds-un-low*. An optimal threshold can be found when weighting the false discovery rate and true positive rate in a $F_{1.5}$ score. Without post classification filtering and default values the average $F_{1.5}$ scores for the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* are 27 %, 11 % and 11 % and peak at 65 %, 71% and 82 % for the species abundance thresholds of 0.01 %, 0.1 % and 1 % respectively. Noteworthy is that the $F_{1.5}$ score optima occur at different abundance thresholds for the three datasets.

Some variation between the datasets was expected as the samples were of different content and species abundance distributions. Filtering out species based on estimated community fraction was expected to be more stable across datasets than filtration based on an absolute number of reads in Kraken 2 or Bracken result output because of independence of the number of reads in the sample, however, when comparing the filtering methods appear similarly to give varying results depending on sample type.[2]

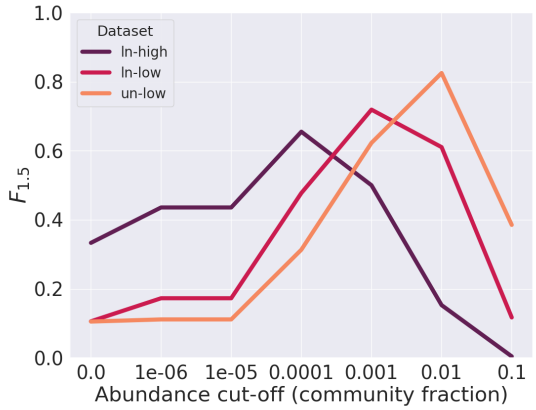
[2]The false discovery rates, true positive rates and $F_{1.5}$ scores for the *ds-ln-high*, *ds-ln-low* and *ds-un-low* when instead applying filtering thresholds in absolute read counts from both raw Kraken 2 output and Bracken reestimated read counts can be found in Figure B.1 Appendix B



(a) False discovery rate



(b) True positive rate



(c) $F_{1.5}$ score

Figure 4.4: Average (a) false discovery rates (FDR), (b) true positive rates (TPR) and (c) $F_{1.5}$ scores for the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* over read classification abundance filtering for each species. Note that the x-axis are logarithmic.

For samples with a species distribution including low abundances, such as log-normal or exponential, applying a cut-off based on estimated community fraction or number of reads classified of a species would inherently also filter out low abundance true positives. It would instead be a detection limit of sorts because false positives or true positives below a threshold would not be reported as found, however it would not translate to a detection limit for the true abundance since the probability of a species with a true abundance of for example 1 % to fall above or below a limit of estimated abundance of 1 % would be dependent of the abundance distribution of the other species in the sample. The use of post classification filtering would theoretically be more applicable for samples with a uniform species distribution. As can be seen in Figure 4.4 c), the dataset *ds-un-low* with uniform samples of 10 species at 10 % community fraction each had the largest relative increase in $F_{1.5}$ score of approximately 70 percentage points between no abundance cut-off to optimum at a 1 % threshold. Adjustment of the Kraken 2 confidence threshold although still appears as a superior alternative to post classification filtering because the confidence is on

read level and is therefor independent of factors such as distribution of abundance fractions, number of species in the sample or the absolute number of reads per species which all might be unknown for a real world metagenomic sample.

4.3 Database Content

It has previously been established that generally the larger the Kraken 2/Bracken databases the better classification performance [43]. Apart from the file size, a database can be 'large' in the aspect of the number of taxonomic nodes represented, such as number of genera, species and strains, the number of genome assemblies per taxonomic node or the ratio of K-mers mapped per node or assembly. Furthermore, the quality of the genome assemblies used for the database build could effect the classification, both the aspects of level of completeness and annotation quality. As the database is required to be fully loaded into memory when running Kraken 2 (in the most time efficient default mode), a limit on memory resource allocation will in turn limit the maximum file size the database can have. The question then became how to balance the database aspects of depth and spread for optimal classification of metagenomic samples.

Taxonomic classification of the samples from the *ds-ln-high*, *ds-ln-low* and *ds-un-low*, using the *db-standard* database with default parameter settings of Kraken 2 and Bracken, resulted in $F_{1.5}$ scores of on average 27.1 %, 10.6 % and 10.6 % for the three datasets. As can be seen in Table 4.2, the true positive rates were relatively high of 76.3 %, 98.7 % and 99.9 % while the false discovery rates were on average 88.0 %, 96.5 % and 96.5 %. The focus was on finding a database composition that would decrease the false discovery rate as it was the main driver behind the low $F_{1.5}$ scores.

4.3.1 Bacteria Only Databases

The K-mer minimizers stored in the baseline *db-standard* database were 89.6 % from bacterial genome assemblies while the rest were from human, archaeal and viral reference. Since the project scope was to only taxonomically classify bacteria, and because the filtering of human reads was moved to a separate step with its own Kraken 2 database, it was hypothesized that the non-bacterial content could be taking up unnecessary space. A database (*db-bac-comp*) was then built with only the bacterial genomes of the *db-standard* to allow for more K-mers per genome and to increase the chance of a K-mer matching the bacteria in the sample.

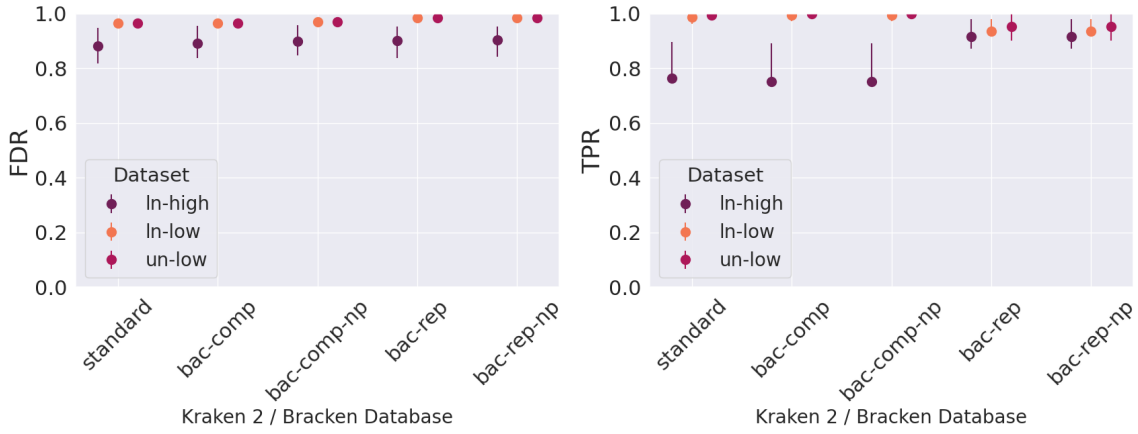
The results were, for default parameter settings of Kraken 2 and Bracken, that the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* on average had lower classification performance for the bacteria only *db-bac-comp* database compared to the *db-standard*. The differences in $F_{1.5}$ were however negligible, 1.8 %, 0.1 % and 0.2 % percentage points for the three datasets, see Table 4.2 and Figure 4.5 for a graphical comparison. Here a small increase in number of minimizers/K-mers per bacterial genome assembly did not completely compensate for the loss of diversity in the taxonomic nodes represented. However, re-running the datasets with non-default classifier pa-

rameter setting of Kraken confidence of 0.075 and Bracken threshold of 750 reads instead resulted in two out of three having a slightly higher $F_{1.5}$ score for the *db-bac-comp* compared to the baseline. It was therefor concluded that the databases had approximately equal performance for bacterial classification although with the *db-standard* having the extra capability of also classifying non-bacteria.

Another bacteria only Kraken 2/Bracken database was tested for the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low*. The database was built from all the RefSeq bacterial genome assemblies categorized as *Reference/Representative sequences* that were found across different levels of assembly completeness. Compared to *db-bac-comp*, the *db-bac-rep* represents more species but was built using half the number of genome assemblies, see Table 3.1 for database overview. Furthermore, each species in the *db-bac-comp* is, however, only represented by approximately one genome while genomes from more studied species are over-represented in the *db-bac-comp* and *db-standard* databases. An additional hypothesized benefit of using only the representative genome sequences was that they were less contaminated with other inserts from other species (estimated to be 74 % cleaner than non-representative genomes on 2023-04-15, a later date from database building[57]).

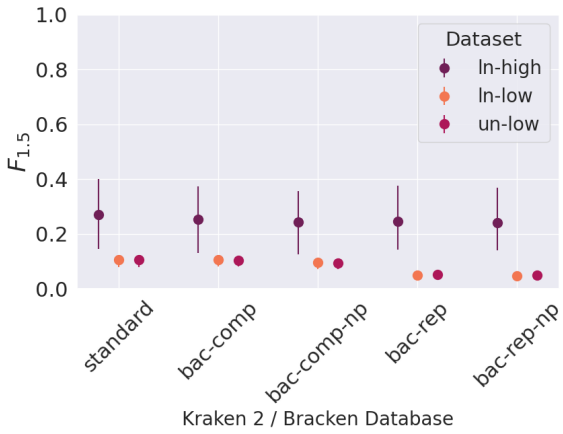
For default Kraken 2/Bracken parameter settings the *db-bac-rep* performed worse than the *db-bac-comp* and *db-standard* with average $F_{1.5}$ scores up to 5 percentage points lower. The average $F_{1.5}$ scores were 24.0 %, 5.0 % and 5.1 % for the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* run against the *db-bac-rep* database with default Kraken 2/Bracken parameter settings. However, surprisingly for the *ds-ln-high*, with high species diversity and low genome coverage, the *db-bac-rep* gave a higher true positive rate of 91.5 % compared to 76.3 % for the *db-standard*, although, this did not apply to the other datasets. With optimized Kraken 2/Bracken parameter settings an average $F_{1.5}$ score of 78.2 % for the *ds-un-low* was achieved however this indicates that the effect of parameter adjustment appear to be more powerful than the change of database content. The *db-bac-rep* had the highest average false discovery rates for default Kraken 2/Bracken parameter settings, although it only differed ≤ 2 percentage points compared to the *db-bac-comp* and *db-standard* databases across all datasets. In contrast, the false discovery rates for the optimized classifier parameter settings were lower for the *db-bac-rep* with up to 10 percentage points difference, however, the true positive rates were also lower which resulted in *db-bac-rep* still having the lowest $F_{1.5}$ scores. Moreover the lowest average false discovery rate of the tested dataset, database and parameter settings combinations was 7.7 % with the *db-bac-rep* for the *ds-ln-high* and optimized classifier parameters.

The conclusion was that the *db-bac-rep* performed worse for taxonomic classification of bacteria compared to the baseline *db-standard* although the relative performance between the databases differed for varying parameter settings and datasets. Because the factors of fewer genomes, balanced species representation and varied assembly level completeness were confounded the reasons for poorer performance could not be determined. For deeper investigation into the effects of various databases the sets of false positive observations could be compared between databases, and further if the correlations between observations of high abundance true positive species and low abundance false positives differed between datasets.



(a) False discovery rate

(b) True positive rate



(c) $F_{1.5}$

Figure 4.5: Mean [25 percentile, 75 percentile] (a) false discovery rates (FDR), (b) true positive rates (TPR) and (c) $F_{1.5}$ scores for the databases *db-standard*, *db-bac-comp*, *db-bac-comp-np*, *db-bac-rep* and db-bac-rep-np across datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* when classified with default Kraken 2 and Bracken.

4.3.2 Plasmid Depleted Databases

The combined Kraken 2 and Bracken databases were built with plasmid sequences removed form the assemblies to test if plasmids increased the false discovery rate. The reasoning was that possible misannotation of plasmids due to inter-species plasmid transfers and intra-species variability in the number and type of plasmids could cause confusion in the classification. The idea behind investigating the effects of removing plasmid came from misclassifications of single isolate samples suspected to be caused by inter-species plasmid transfers. To demonstrate with an example, consider the result of the Bracken abundance estimation of a sample where it was 96.3 % *Salmonella enterica*, 3.2 % *Escherichia coli* and other mixed bacteria together summing to <1 %. Ideally there should only be one species in the sample because it is taken from a single culture colony. A possible explanation for finding more than one species is that the sample was contaminated before sequencing and

that the contaminants were sequenced giving reads that thereafter were classified correctly. Another explanation could be that there was no lab contamination but instead some *S. enterica* reads were misclassified by Kraken 2, possibly due to high average nucleotide identity and a low Kraken 2 confidence setting. In an attempt to distinguish between the explanation options, the reads assigned directly by Kraken 2 to *E. coli* were aligned against the combined Genbank and RefSeq databases using BLAST. What was found was near perfect alignment to a plasmid annotated as belonging to *E. coli* and not the chromosome. Since plasmids are known to be transferred between *S. enterica* and *E. coli* [58] it is possible that the specific *S. enterica* sequenced had acquired a plasmid from *E. coli* which had not been added to the NCBI database yet. The phenomenon of the reads assigned to the second most abundant species aligning to plasmids of another specie was found in some single isolate samples. Note that the samples were not selected at random, but instead specifically chosen because of the results showing a second bacteria at an unexpected high percentage (higher than 1 %). The evidence for a part of the false positives originating from inter-species transferred plasmids was on an anecdotal level but nevertheless interesting enough to prompt investigation for metagenomic samples.

The Kraken 2/Bracken databases with plasmid sequences removed *db-bac-comp-np* and *db-bac-rep-np* were built from the same bacterial assemblies as for *db-bac-comp* and *db-bac-rep* respectively. When run with the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* for different Kraken 2 and Bracken parameter settings, the reference databases without plasmids had a consistently sightly worse classification performance when compared with the corresponding database with plasmids. As can be observed in Table 4.2, the average F_1 scores for *db-bac-comp-np* and *db-bac-rep-np* across the datasets are approximately one percentage point lower than for *db-bac-comp* and *db-bac-rep* respectively. Furthermore, the increase in false discovery rate was stronger than the decrease in true positive rate.

A possible explanation for why the resulting false discovery rate was higher when using a database without plasmids compared to the hypothesized lower is that the real world single culture samples were outliers (only a handful of single culture samples were examined) or that the genome assemblies used for simulating metagenomic samples did not include plasmids annotated differently from those in the assemblies used for the databases. Furthermore, when removing plasmids from all reference genomes it reduces the number of possible unique nucleotide K-mers per genome and species, in the sense of lowering the pool of genetic material per species. The gain from reducing false positives caused by inter-species plasmid transfer could then have been overridden by the loss of the uniqueness that non-transferring and correctly annotated plasmid variants brought which in turn caused increased misclassification among similar species. The effects of possible misannotated plasmids could additionally be obscured by there being many species in the metagenomic samples compared to single culture samples. If the sample contained both the species of which a plasmids is listed in the reference database as and other species where it is present then there would not be a false positive since the classification results are reported on sample level and not read level.

The conclusion for general metagenomic samples, without pre-indication of mainly

containing species known for jumping plasmids, is that a Kraken 2/Bracken reference database with plasmids give better classification performance than without. For future studies plasmids could however still be interesting as they are more unstable genetic elements than chromosomes and perhaps instead of removing plasmids weighting their K-mer matches differently from the chromosome.

4.3.3 Database for Pathogen Subset

In the case of a taxonomic classification pipeline being used for discovery of pathogens, misclassifying just one as false negative could potentially have disastrous consequences if the sample is from a critically ill patient. When optimizing the taxonomic classification procedure for a F-score, the close to 100 % true positive rate observed for the *db-standard* with default Kraken and Bracken parameter settings inevitably decreased in favor of a lower false discovery rate. A compromise solution could be to include two taxonomic classification steps, one which is F-score optimized for producing a general taxonomic profile of the sample and one in which a subset of pathogens of particular interest are screened for at a higher acceptance threshold for true positive rate.

Aside from adjusting the Kraken and Bracken parameters alone, a hypothesized solution was to increase the number of minimizers in the classification database specifically belonging to the nodes of the pathogens in consideration. It was concluded in Section 4.3.1 that increasing bacterial minimizer counts with a bacteria only database (*db-bac-comp*) was not greatly superior compared to the more diverse one with genomes from several kingdoms (*db-standard*). However, this was for databases with minimizers down-sampled to 8 GB and increasing minimizers without down-sampling could possibly have a larger effect.

The idea was tested using the dataset *ds-un-low* and database *db-bac-subset* which was built with only genome assemblies from seven pathogens but without down-sampling the minimizers. A similar parameter optimization search as for the *db-standard* was done also for the *db-bac-subset* after which the samples in the dataset *ds-un-low* were run for a Kraken confidence parameter of 0.3 and Bracken minimum threshold of 10,000 reads. In Table B.4 the classification results for each sample group are displayed together with the results for the *db-standard* (run with $F_{1.5}$ optimized parameters and with the accommodation of only considering the same subset of species as in *db-bac-subset* when reporting the positive and negative observations). Among the group of 7 pathogens in the subset database, *S. aureus* and *A. baumannii* were selected for detection testing and were successfully found in all samples where they were expected when run with *db-bac-subset*. When run with the *db-standard A. baumannii* was found in all samples but *S. aureus* was only correctly classified in 8 out of 15 samples where it should have been found. For the negative control samples (14 total), with no species in the pathogen sub-set but 0, 1 or 4 species sharing the same genus as species in the subset, there were 1 or 2 false positive classifications per sample when run with the *db-bac-subset* while only 1 false positive for a single sample when run with *db-standard* (for $F_{1.5}$ optimized parameters). Although the reference database built with a subset of bacteria did not have zero false discoveries, it had a lower false discovery rate than when running the *db-standard* even

for default parameters while still finding the correct pathogen in every sample. If there is not access to sufficient RAM to hold the full standard database without down-sampling, a database built with fewer taxonomic nodes without minimizer restriction could be used in compliment to the $F_{1.5}$ optimized taxonomic classification step.

4.4 Metagenomic Sample Composition

A sample of genetic material being *shotgun metagenomic* speaks to its collection and sequencing method rather than its composition, the contents could for example be similar to a single culture isolate where one main species is expected or contain hundreds of species with a nonuniform abundance distribution and have some species represented by only few reads. The intent was to subdivide metagenomic samples bases on attributes such as abundance distribution, species diversity and genome coverage and then evaluate classification performance by category. However, the results and discussion are only partially covering the topic due to inadequate experimental planning.

4.4.1 The Fewer Species Dataset *ds-ln-low*

For the 40 sample dataset *ds-ln-low* with log-normally distributed species abundances the classification metrics divided by the eight sample group were calculated for the dataset run with the *db-standard* database and 'optimized' parameters (Kraken 2 confidence = 0.075, Bracken read threshold = 750 reads). As can be seen in Table 4.3, the average classification performance does not appear to vary systematically across the groups of five replicates of species counts between 10-80 and log-normal coverage distributions with μ either 10X or 20X. The within group variations were larger than the between group differences. For example, the average $F_{1.5}$ scores were between 72.3% and 82.0 % for the groups and for *s20-c20* the difference between the minimum and maximum sample scores were $92.9 - 63.2 = 29.3$ percentage points. The coverage and number of species appear to have had less of an impact on classification metrics compared to the actual composition of species types. However, for the *ds-ln-low* the abundances of the species in the samples were sampled form log-normal distributions and for each replicate in the sub-groups both the exact abundances and species in the sample would differ. Which factor, number of species, species coverage or the specific species, that caused the difference in classification performance was thus convoluted. For future dataset simulations it could be more informative to simulate the same set of species but with different ordering of abundances and average coverages among themselves for each replicate.

Group	F_1 (%)	$F_{1.5}$ (%)	TPR (%)	FDR (%)
s10-c10	73.8 [66.7,88.9]	72.3 [63.9,85.2]	70.0 [60.0,80.0]	21.4 [0.0,30.0]
s10-c20	82.4 [77.8,88.9]	82.0 [74.6,85.2]	82.0 [70.0,90.0]	14.8 [0.0,30.8]
s20-c10	78.0 [66.7,87.2]	79.4 [66.0,89.5]	82.0 [65.0,95.0]	25.1 [10.5,32.0]
s20-c20	78.1 [59.6,94.7]	79.4 [63.2,92.9]	82.0 [70.0,90.0]	24.3 [0.0,48.2]
s40-c10	77.4 [69.1,88.0]	77.2 [69.5,85.8]	77.0 [70.0,82.5]	22.0 [5.7,31.7]
s40-c20	71.5 [62.1,75.3]	73.1 [64.0,78.4]	76.0 [67.5,85.0]	32.2 [27.9,42.6]
s80-c10	78.3 [76.2,82.2]	78.8 [76.2,82.8]	79.8 [76.2,83.8]	23.1 [19.3,25.6]
s80-c20	74.6 [70.5,78.4]	76.9 [72.6,81.2]	80.7 [76.2,87.5]	30.5 [27.1,34.4]

Table 4.3: Species level taxonomic classification metrics F_1, $F_{1.5}$, true positive rate (TPR) and false discovery rate (FDR) for the dataset *ds-ln-low* run for the *db-standard* database with Kraken 2 confidence = 0.075 and Bracken read threshold = 750. The values listed are means of the five replicates in each dataset group together with the minimum and maximum [min, max] of the sample group.

4.4.2 The Pathogen Subset Dataset *ds-un-low*

The average genome coverage for the species in the samples of the *ds-un-low* was fixed to 20X (which makes the species abundances uniform) as well as the number of species per sample, 10 species, and the difference between the samples was thus the particular set of species. It can be noted that the species composition in the *ds-un-low* was not fully random for each sub-group as the dataset was created to test the pathogen subset databases. How representative the genome assemblies used for read simulation of a species actually were for the indented species could also possibly influence the classifiers ability to recognize reads as coming from the species. Since the genomes used for simulating samples were all species identified complete level assemblies the risk of completely misannotated sequences (that the assembly was mislabeled as a different species) was however considered lower than for general GenBank assemblies.

The results of running the *ds-un-low* for the *db-standard* database with 'optimized' Kraken 2/Bracken parameter settings (as any other sample, not with regard to the specific subset of pathogen in the samples), was that there still was large variation between replicate samples from the same groups, Table 4.4. The differences of $F_{1.5}$ scores and true positive rates between samples in a dataset group (replicates drawn from the same pools of species) were up to 50 percentage points. For example for sub-group *a-baumannii-g0* the average true positive rate was 78.0 % but where the minimum was 50 % and the maximum was 100 %, meaning that for one sample half of the species were not identified and for another sample all the species were identified. The conclusion was that which species happened to be in the sample would have large effects on the taxonomic classification of the sample. Which species that were more difficult to classify than others was not investigated further in this project.

Sample group	F_1 (%)	$F_{1.5}$ (%)	TPR (%)	FDR (%)
control-g0	87.1 [70.0,94.7]	87.4 [70.0,94.2]	88.0 [70.0,100.0]	13.3 [0.0,30.0]
control-g1	82.7 [75.0,87.0]	85.3 [80.1,91.6]	90.0 [80.0,100.0]	22.6 [11.1,35.7]
control-g4	79.4 [69.6,85.7]	81.4 [73.2,87.3]	85.0 [80.0,90.0]	24.6 [11.1,38.5]
a-baumannii-g0	81.3 [58.8,100.0]	79.9 [55.1,100.0]	78.0 [50.0,100.0]	14.1 [0.0,28.6]
a-baumannii-g1	78.4 [63.6,88.9]	78.1 [65.9,85.2]	78.0 [70.0,90.0]	19.8 [0.0,41.7]
a-baumannii-g4	78.8 [69.6,85.7]	80.6 [73.2,87.3]	84.0 [80.0,90.0]	25.2 [11.1,38.5]
s-aureus-g0	84.7 [73.7,94.7]	85.0 [72.2,92.9]	86.0 [70.0,90.0]	15.6 [0.0,35.7]
s-aureus-g1	79.2 [55.6,90.0]	78.7 [53.3,90.0]	78.0 [50.0,90.0]	19.2 [10.0,37.5]
s-aureus-g4	83.9 [62.5,100.0]	82.9 [57.0,100.0]	82.0 [50.0,100.0]	11.5 [0.0,30.8]

Table 4.4: Species level taxonomic classification metrics F_1, $F_{1.5}$, true positive rate (TPR) and false discovery rate (FDR) for the dataset *ds-un-low* run for the *db-standard* database with Kraken 2 confidence = 0.075 and Bracken read threshold = 750. The values listed are means of the five replicates in each dataset group together with the minimum and maximum [min, max] of the sample group. Note that the group names are in reference to the pathogen subset databases and all bacteria in the *db-standard* were in this case considered for prediction.

In a study benchmarking taxonomic classifiers for ten common bloodstream pathogens and contaminates it was found that the species had different probabilities of being correctly identified across different taxonomic classification programs including the Kraken 2 and Bracken combination (Govender 2022) [13]. It was also investigated which species were correlated as false positives for high abundances of specific true positive species. As an example *E. coli* had a high probability of being misclassified although if it was listed in the result as a true positive then *Escherichia fergusonii* and *Shigella flexneri* were most likely to appear as false positives. The article furthermore suggested to use species dependent adjustment factors to multiply with a fixed post classifier minimum abundance threshold for improved classification performance. Correlation analyses and tables of probabilities of a species being a true or false positive, given the presence and estimated abundances of another species in a sample, would be possible for a subset of species of interest however appears to be less feasible for all species in a database such as *db-standard*.

4.4.3 The Many Speceis Dataset *ds-ln-high*

A conclusion drawn mainly from the dataset *ds-un-low* was that the specific species composition of the samples effect the classification success. Nonetheless, for the *ds-ln-high* other factors appear to be separating the groups than species. The replicates in the groups of 10 samples have more similar within group classification metrics than for the datasets *ds-ln-low* and *ds-un-low*, see Table B.5 and Table B.6 in Appendix A. For example, groups of samples simulated from genomes with different average nucleotide identities (ANI) were distinct from each other. A possible explanation for lower species composition dominance is that difficult to classify species have less impact on the sample metrics because of the hundreds of species per sample compared to only 10 species. There are however many differences between the simulation techniques of the *ds-un-low* and the *ds-ln-high* leading to a need for

cautious comparison. See the original Parks et al. article [1] for more in-depth between group comparisons and note that only classification of bacteria was recorded in the project compared to all species, bacteria and archaea, in the article.

One of the differences in the simulation method for the *ds-ln-high* was the use of a fixed number of reads per sample independent of the number of species. Not adjusting the number of reads for sample complexity could be more similar to real world samples where the species count is of course unknown before sequencing. A fixed sample size can however generate samples with some species having few reads, which in the case of some samples in the *ds-ln-high* was less than 100 hundred reads. Considering the read length of 150 bp and a bacterial genome being millions of basepairs long the fraction of the genome was covered is low. Assuming that the database does not contain all K-mers to completely cover each genome the probability of identifying a species would go down with the fraction of its genome covered by reads. Furthermore, even if a low coverage species is correctly identified it would be difficult to distinguish from other low abundance species that are false positives, especially since Bracken can inflate false discovery rates by faulty read redistribution. If a simple true or false for a species presence in a sample is sufficient, without the addition of a percentage abundance estimate, then the use of Bracken read redistribution after Kraken 2 classification would increase uncertainty without adding information.

4.4.4 Reference Genome Considerations

A potential contributing reason for the high false discovery rates, which was not considered in the project, was contamination or otherwise misannotation of assembly sequences in the NCBI genome repositories GenBank and RegSeq that were used for building the classifier reference databases and simulation of samples. In this context contamination referrers to when sequences from genomes of other species are accidentally integrated into a genome assembly labeled as only being of one organism [59]. As an example, a *S. aureus* reference could have short sequences from the human genome and *E. coli* inserted throughout its assembly but be treated as a pure assembly in the metadata resulting in an increased probability of the contaminates appearing as false positive observations. Another type of misassembly could be chimeric where genomes from multiple organisms of the same type are assembled into one individual [59].

Publications from recent years have reported ongoing issues with NCBI database contamination and screening algorithms [60][61][62]. For example, in 2020 a study 114,035 and 2,161,746 contaminated sequences were found in RefSeq and GenBank respectively [61], although from across all domains of life and levels of assembly completeness, and in 2021 it was estimated that 5.7 % of genomes in GenBank were chimeras [62]. NCBI are themselves aware of the problem and identified contaminants in approximately 10 % of the prokaryote genomes in 2022 [57]. The NCBI Foreign Contamination Screen tool suite has a newly developed genome contamination screener FCS-GX (pre-print published 2023-06-06) that can be applied at the scale of the public databases for improvement compared to their legacy screening pipelines [57].

When selecting genome assemblies for building classification reference database and simulating metagenomic samples, the category of *Atypical Assemblies* that according to NCBI encompasses chimeric and contaminated assemblies was excluded [59]. Following that GenBank and RegSeq rapidly increases in size and are continuously curated, the exact status of the levels of genome assembly contamination and how successful the filtering of atypical assemblies was at the time of access is unknown. Including additional pre-screening for contamination of genome assemblies from public databases before use for classification reference databases or sample simulation could be a consideration for future projects.

5

Conclusion

In conclusion it was challenging to taxonomically classify shotgun metagenomic samples with both high precision and recall for limited computational resources. Filtering of human DNA using Kraken 2 was on the other hand successful for simulated human reads and found to be comparable in removal to filtering using the BWA-mem 2 aligner while being more time efficient. The best overall classification performance was observed from the baseline standard database although similar performances were observed among the other tested size capped reference databases, different bacterial composition and with and without plasmids. Changing the parameter settings of the classifier (Kraken 2 read confidence and Bracken read threshold for read redistribution) could drastically change which species were classified and it was found that the default settings may be far from optimal depending on the intended application. The influence of factors relating to metagenomic sample composition such as genome coverage, abundance distribution and number of species were not fully understood, however, it was observed that the composition of species in the sample could have drastic effects on how they were classified. For the tested classification related variables, changes in content of a size capped reference database had a lower effect on classification performance compared to adjusting classifier parameter settings while the largest impact appeared to be from the specific composition of species in the sample. A one-size-fits-all approach to taxonomic classification of any shotgun metagenomic sample would be near impossible with the tested K-mer based classifier and a solution could be to have specialized pipeline tracks optimized for samples with different expected range of species, sequencing depth and abundance distributions.

Bibliography

[1] Parks DH, Rigato F, Vera-Wolf P, Krause L, Hugenholtz P, Tyson GW, et al. Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome. Front. Microbiol. 2021;12:731. DOI: 10.3389/FMICB.2021.643682/

[2] World Health Organization. Antibiotic resistance. [Internet]. Geneva: World Health Organization; 2020. [cited 2023 Oct 29] Available from: https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance

[3] Rong Z, Yan Z, Zheng Hui X, Cai Yun L, Fan Z, Wei Qing, et al. Diagnosis and Surveillance of Neonatal Infections by Metagenomic Next-Generation Sequencing. Front. Microbiol. 2022;13. DOI: 10.3389/FMICB.2022.855988

[4] Parize P, Pilmis B, Lanternier F, Lortholary O, Lecuit M, Muth E, et al. Untargeted next-generation sequencing-based first-line diagnosis of infection in immunocompromised adults: a multicentre, blinded, prospective study. Clinical Microbiology and Infection. 2017;23(8). DOI: 10.1016/J.CMI.2017.02.006

[5] Waddington C, Carey ME, Boinett CJ, Higginson E, Veeraraghavan B, Baker S. Exploiting genomics to mitigate the public health impact of antimicrobial resistance. Genome Medicine. 2022;14:1-14. DOI: 10.1186/S13073-022-01020-2

[6] Govender KN, Street TL, Sanderson, Nicholas D. and Eyrea, David W. Metagenomic Sequencing as a Pathogen-Agnostic Clinical Diagnostic Tool for Infectious Diseases: a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. Journal of Clinical Microbiology. 2021;59(9). DOI: 10.1128/JCM.02916-20

[7] Balloux F, Brønstad Brynildsrud O, van Dorp L, Shaw LP, Chen H, Harris KA, et al. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. Trends in Microbiology. 2018;26(12):1035-48. DOI: 10.1016/J.TIM.2018.08.004

[8] Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell Press. 2019;178:779-94. DOI: 10.1016/J.CELL.2019.07.010

[9] Moragues-Solanas L, Scotti R, O'Grady J. Rapid metagenomics for diagnosis of bloodstream and respiratory tract nosocomial infections: cur-

rent status and future prospects. Expert Review of Molecular Diagnostics. 2021;21(4):37-80. DOI: 10.1080/14737159.2021.1906652

[10] Sevim V, Lee J, Egan R, Clum A, Hundley H, Lee J, et al. Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. Scientific Data. 2019;6(1):1-9. DOI: 10.1038/s41597-019-0287-z

[11] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology. 2019;20(1):1-13. DOI: 10.1186/S13059-019-1891-0

[12] Lu J, Breitwieser FP, Thielen Peter, Salzberg SL. Bracken: Estimating species abundance in metagenomics data. PeerJ Computer Science. 2017;3:e104. DOI: 10.7717/PEERJ-CS.104

[13] Govender KN, Eyre DW. Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications. Microbial Genomics. 2022;8(10). DOI: 10.1099/MGEN.0.000886

[14] Shi Y, Wang G,Lau HCH, Yu J. Metagenomic Sequencing for Microbial DNA in Human Samples: Emerging Technological Advances. International Journal of Molecular Sciences. 2022;23(4). DOI: 10.3390/IJMS23042181

[15] Moya A, Medina DA, Francisco Vazquez-Castellanos J, Leuven K, Lai PS and Sui H, et al. Impact of DNA Extraction Method on Variation in Human and Built Environment Microbial Community and Functional Profiles Assessed by Shotgun Metagenomics Sequencing. Front. Microbiol. 2020;11. DOI: 10.3389/fmicb.2020.00953

[16] Pell J, Hintze A, Canino-Koning R, Howe A, Tiedje JM, Brown C, et al. Scaling metagenome sequence assembly with probabilistic de Bruijn graph. Proc Natl Acad Sci USA. 2012;109(33):13272-7. DOI: 10.1073/PNAS.1121464109/-/DCSUPPLEMENTAL

[17] Sohn JI, Nam JW. Scaling metagenome sequence assembly with probabilistic de Bruijn graph. Briefings in Bioinformatics. 2018;19(1):23-40. DOI: 10.1093/BIB/BBW096

[18] Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome. 2018;6(1). DOI: 10.1186/S40168-018-0426-3

[19] Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, et al. Metagenome analysis using the Kraken software suite. Nat Protoc. 2022;12(17):2815–39. DOI: 10.1038/s41596-022-00738-y

[20] Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Research. 2017;27(4):626-38. DOI: 10.1101/GR.216242.116

[21] Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nature Biotechnology.

2017;35(9):833-44. DOI: 10.1038/nbt.3935

[22] Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front. Plant Sci. 2014;5. DOI: 10.3389/FPLS.2014.00209

[23] Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics. 2019;20(4):1125-36. DOI: 10.1093/bib/bbx120

[24] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2010;38(6):1767-71. DOI: 10.1093/NAR/GKP1137

[25] Illumina. Sequencing Read Length | How to calculate NGS read length. [Internet]. Illumina. [cited 2023 Aug 20]. Availible from: https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html

[26] Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014;30(5):614-20. DOI: 10.1093/BIOINFORMATICS/BTT593

[27] Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics and Bioinformatics. 2021;3(1). DOI: 10.1093/NARGAB/LQAB019

[28] Smith RH, Glendinning L, Walker AW, Watson M. Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome. Animal Microbiome. 2022;4(57). DOI: 10.1186/S42523-022-00207-7

[29] MacLean RC, San Millan A. Microbial Evolution: Towards Resolving the Plasmid Paradox. Current Biology. 2015;25(17):PR764-7. DOI: 10.1016/J.CUB.2015.07.006

[30] Top EM, Mizrahi I, Oliveira C, Johnson TJ, Shintani M, Sanchez ZK, et al. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front. Microbiol. 2015;6. DOI: 10.3389/fmicb.2015.00242

[31] Bañuelos-Vazquez LA, Torres Tejerizo G, Brom S. Regulation of conjugative transfer of plasmids and integrative conjugative elements. Plasmid. 2017;91:82-9. DOI: 10.1016/J.PLASMID.2017.04.002

[32] Stecher B, Denzler R, Maier L, Bernet F, Sanders MJ, Pickard DJ, et al. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. PNAS. 2012;109(4):1269-74. DOI: 10.1073/pnas.1113246109

[33] Koraimann G. Spread and Persistence of Virulence and Antibiotic Resistance Genes: A Ride on the F Plasmid Conjugation Module. EcoSal Plus. 2018;8(1). DOI: 10.1128/ecosalplus.esp-0003-2018

[34] National Center for Biotechnology Information (NCBI) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. [cited 2023 Jun 12]. Available from: https://www.ncbi.nlm.nih.gov/

[35] National Center for Biotechnology Information (NCBI). GenBank Overview [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. [updated 2022 Dec 8; cited 2023 Jun 2]. Available from: https://www.ncbi.nlm.nih.gov/genbank/

[36] National Center for Biotechnology Information (NCBI). Genome.[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. [cited 2022 Jun 12]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/

[37] National Center for Biotechnology Information (NCBI). About RefSeq [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. [updated 2021 Mar 19; cited 2023 Jun 2]. Available from: https://www.ncbi.nlm.nih.gov/refseq/about/

[38] National Center for Biotechnology Information (NCBI). Prokaryotic RefSeq Genomes [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. [updated 2022 Oct 6; cited 2023 Jun 2]. Available from: https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

[39] Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database: The Journal of Biological Databases and Curation. 2020;2020:baaa062. DOI: 10.1093/DATABASE/BAAA062

[40] Conrad Schoch. Frequently Asked Questions - Taxonomy Help - NCBI Bookshelf. [Internet]. Bethesda (MD): National Center for Biotechnology Information; 2011. [updated 2020 Feb 21; cited 2023 May 23]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK54428/

[41] National Center for Biotechnology Information (NCBI). Prokaryotic phylum name changes coming soon! - NCBI Insights [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2022. [cited 2023 Jun 6]. Available from: https://ncbiinsights.ncbi.nlm.nih.gov/2022/11/14/prokaryotic-phylum-name-changes/

[42] Lu J. Kraken2 Manual [computer program documentation]. Kraken2, version 2.1.2 Manual, Confidence Scoring. [Updated 2020 Dec 1, cited 2023 Jun 6]. Available from: https://github.com/DerrickWood/kraken2/wiki/Manual#confidence-scoring

[43] Wright RJ,Comeau AM, Langille MGI. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microbial Genomics. 2023;9(3). DOI: 10.1099/MGEN.0.000949

[44] Langmead B. Index zone by BenLangmead. Kraken 2 / Bracken Refseq indexes [Internet]. [cited 2023 Apr 4] Available from: https://benlangmead.github.io/aws-indexes/k2

[45] Lu J, Breitwieser F. Bracken, version 2.8 [computer program documentation]. [Updated Oct 11 2022, cited 2023 Jun 6]. Available from: https://github.com/jenniferlu717/Bracken

[46] National Center for Biotechnology Information (NCBI). Genome, Homo sapiens genome assembly GRCh38.p14. Accession No. GCF_000001405.40 [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2022. [cited 2023 Jun 6]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome /GCF_000001405.40/

[47] Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics. 2019;35(3):521-2. DOI: 10.1093/BIOINFORMATICS/BTY630

[48] Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15:141-61. DOI: 10.1007/s10142-015-0433-4

[49] Kim HS, Jeon S, Kim Y, Kim C, Bhak J, Bhak J. KOREF_S1: phased, parental trio-binned Korean reference genome using long reads and Hi-C sequencing methods. GigaScience. 2022;11:giac022. DOI: 10.1093/GIGASCIENCE/GIAC022

[50] Shumate A, Zimin AV, Sherman RM, Puiu D, Wagner JM, Olson ND, et al. Assembly and annotation of an Ashkenazi human reference genome. Genome biology. 2020;21(129). DOI: 10.1186/S13059-020-02047-7

[51] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. DOI: 10.1093/BIOINFORMATICS/BTY560

[52] Musich R, Cadle-Davidson L, Osier MV. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. Front. Plant Sci. 2021;12. DOI: 10.3389/FPLS.2021.657240

[53] Vasimuddin Md., Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 20-24 May 2019; Rio de Janeiro, Brazil. IEEE; 2019. DOI: 10.1109/IPDPS.2019.00041

[54] Johnson J, Sun S, Fodor AA. Systematic classification error profoundly impacts inference in high-depth Whole Genome Shotgun Sequencing datasets. bioRxiv. 2022. DOI: 10.1101/2022.04.04.487034

[55] Sloan WT, Woodcock S, Lunn M, Head IM, Curtis TP. Modeling taxa-abundance distributions in microbial communities using environmental sequence data. Microb Ecol. 2007;53:443-55. DOI: 10.1007/S00248-006-9141-X

[56] Curtis TP, Sloan WT, Scannell JW. From the Cover: Estimating prokaryotic diversity and its limits. 2002;99(16):10494-9. DOI: 10.1073/P-NAS.142680199

[57] Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. bioRxiv. 2023. DOI: 10.1101/2023.06.02.543519

[58] McMillan EA, Jackson CR, Frye JG. Transferable Plasmids of Salmonella enterica Associated With Antibiotic Resistance Genes. Front. Microbiol. 2020;11. DOI: 10.3389/FMICB.2020.562181

[59] National Center for Biotechnology Information (NCBI). Documentation, Genome Notes [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. [cited 2023 Aug 18]. Available from: https://www.ncbi.nlm.nih.gov/datasets/docs/v2 /policies-annotation/genome-processing/genome_notes/

[60] Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, Baurain D, Cornet Luc. Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. Front. Microbiol. 2021;12. DOI: 10.3389/FMICB.2021.755101

[61] Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biology. 2020;21(1). DOI: 10.1186/S13059-020-02023-1

[62] Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology. 2021;22(178). DOI: 10.1186/S13059-021-02393-0

A

Article Dataset

Group name	ANI similarity	Species diversity	Strain diversity	ANI to closest ref. genome (%)	AF to closest ref. genome (%)[a]	No. species	Strain per species	Species abun. (%)
ani100_stTrue	Identical[b]	Medium	Single	100	100	106±15.8	1	26.8 to 3.3×10^{-4}
ani100_stFalse	Identical	Medium	Multiple	100	100	92±22.5	2.6±0.16	62.9 to 9.0×10^{-5}
ani100_stTrue	Identical	High	Single	100	100	490±96.0	1	26.9. to 1.9×10^{-6}
ani100_stFalse	Identical	High	Multiple	100	100	505±74.8	2.5±0.07	13.2. to 6.1×10^{-6}
ani99_cLOW_stTrue	High[c]	Medium	Single	99.4±0.22	94.5±3.02	99±21.3	1	38.0 to 2.4×10^{-4}
ani99_cLOW_stFalse	High	Medium	Multiple	99.3±0.22	94.4±3.06	106±29.7	4.7±0.33	39.4 to 3.2×10^{-4}
ani99_cHIGH_stTrue	High	High	Single	99.3±0.22	94.5±2.93	499±86.1	1	60.3 to 1.6×10^{-5}
ani99_cHIGH_stFalse	High	High	Multiple	99.4±0.22	94.4±3.00	450±116	4.0±0.32	18.4 to 1.3×10^{-5}
ani97_cLOW_stTrue	Moderate[d]	Medium	Single	98.3±0.54	90.9±4.41	104±24.3	1	62.3 to 2.3×10^{-4}
ani97_cLOW_stFalse	Moderate	Medium	Multiple	98.4±0.52	91.2±3.92	106±19.6	4.7±0.16	29.6 to 3.2×10^{-4}
ani97_cHIGH_stTrue	Moderate	High	Single	98.3±0.54	90.8±4.23	509±58.6	1	23.2 to 2.8×10^{-5}
ani97_cHIGH_stFalse	Moderate	High	Multiple	98.3±0.52	91.1±4.19	532±70.9	3.8±0.26	10.0 to 9.2×10^{-6}
ani95_cLOW_stTrue	Low[e]	Medium	Single	96.4±0.50	87.9±4.56	93±32.9	1	80.5 to 2.8×10^{-4}
ani95_cLOW_stFalse	Low	Medium	Multiple	96.3±0.52	88.0±4.33	109±26.6	3.2±0.23	36.6 to 1.4×10^{-4}

[a] AF, alignment fraction, i.e., percentage of orthologous regions shared between two genomes,
[b] 100% ANI similarity, [c] [99%, 99.75%] ANI similarity,
[d] [97%, 99%) ANI similarity, [e] [95%, 97%) ANI similarity.

Table A.1: Dataset from the article *Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome* by Parks et al. [1] adapted into the *ds-ln-high* dataset. Table recreated from *Table 2* in the article.

B

Additional Results

	Dataset	Database	F_1 (%)	TPR (%)	FDR (%)
Default (conf. 0, min reads 10)	ds-ln-high	db-standard	20.0 [9.7,29.9]	76.3 [71.5,89.6]	88.0 [81.8,94.8]
		db-bac-comp	18.5 [8.5,27.7]	75.1 [69.7,89.1]	89.0 [83.6,95.5]
		db-bac-comp-np	17.6 [8.2,26.0]	75.1 [69.7,89.1]	89.7 [84.6,95.7]
		db-bac-rep	17.4 [9.3,27.5]	91.5 [88.5,97.8]	90.0 [83.7,95.1]
		db-bac-rep-np	17.0 [9.1,26.9]	91.5 [88.5,97.9]	90.3 [84.1,95.2]
	ds-ln-low	db-standard	6.8 [5.5,7.3]	98.7 [98.4,100.0]	96.5 [96.2,97.2]
		db-bac-comp	6.7 [5.5,7.3]	99.5 [100.0,100.0]	96.5 [96.2,97.2]
		db-bac-comp-np	6.1 [5.0,6.5]	99.5 [100.0,100.0]	96.9 [96.6,97.5]
		db-bac-rep	3.1 [2.5,3.6]	93.4 [90.0,97.8]	98.4 [98.2,98.7]
		db-bac-rep-np	3.0 [2.4,3.5]	93.4 [90.0,97.8]	98.4 [98.2,98.8]
	ds-un-low	db-standard	6.8 [5.1,7.6]	99.5 [100.0,100.0]	96.5 [96.1,97.4]
		db-bac-comp	6.6 [5.2,7.5]	100.0 [100.0,100.0]	96.6 [96.1,97.3]
		db-bac-comp-np	6.0 [4.5,7.0]	100.0 [100.0,100.0]	96.9 [96.4,97.7]
		db-bac-rep	3.2 [2.4,3.7]	95.2 [90.0,100.0]	98.4 [98.1,98.8]
		db-bac-rep-np	3.1 [2.4,3.5]	95.2 [90.0,100.0]	98.4 [98.2,98.8]
Optimal (conf.0.075, min reads 750)	ds-ln-high	db-standard	59.1 [50.1,70.9]	46.0 [34.1,58.0]	10.3 [3.4,5.4]
		db-bac-comp	61.1 [53.3,72.5]	49.4 [37.9,62.2]	13.2 [4.7,19.8]
		db-bac-comp-np	61.3 [53.4,73.1]	49.7 [38.4,62.4]	13.5 [4.7,19.4]
		db-bac-rep	52.1 [39.5,65.3]	38.0 [25.6,51.8]	7.7 [3.4,9.3]
		db-bac-rep-np	52.1 [39.6,65.2]	38.0 [25.2,51.2]	8.0 [3.6,9.5]
	ds-ln-low	db-standard	76.8 [72.8,79.0]	78.7 [73.8,85.0]	24.2 [20.4,30.8]
		db-bac-comp	74.1 [70.0,77.4]	81.9 [80.0,85.0]	31.5 [27.2,37.5]
		db-bac-comp-np	72.2 [67.2,76.2]	82.4 [80.0,86.2]	35.2 [28.4,41.2]
		db-bac-rep	73.7 [69.8,78.1]	70.9 [67.2,75.0]	22.1 [13.5,26.7]
		db-bac-rep-np	73.2 [70.2,77.9]	70.4 [66.2,75.6]	22.5 [15.7,26.8]
	ds-un-low	db-standard	81.8 [75.9,90.0]	83.2 [80.0,90.0]	18.3 [10.0,27.6]
		db-bac-comp	80.0 [75.0,87.7]	87.7 [80.0,90.0]	25.6[16.1,35.7]
		db-bac-comp-np	77.6 [69.6,86.0]	87.7 [80.0,90.0]	29.5 [23.1,38.5]
		db-bac-rep	79.6 [72.2,88.9]	76.4 [70.0,80.0]	15.7 [9.8,20.6]
		db-bac-rep-np	79.3 [72.2,88.9]	76.1 [70.0,80.0]	15.7 [9.8,20.0]
Light (conf.0.05, min reads 50)	ds-ln-high	db-standard	61.8 [54.8,74.0]	71.0 [66.2,83.4]	42.2 [25.0,55.4]
		db-bac-comp	58.0 [49.0,71.4]	70.8 [64.0,83.5]	47.9 [29.8,62.0]
		db-bac-comp-np	57.3 [47.7,70.9]	70.9 [64.1,83.8]	49.0 [30.5,63.0]
		db-bac-rep	66.7 [58.7,76.6]	81.5 [77.2,89.6]	40.4 [25.1,54.6]
		db-bac-rep-np	66.3 [58.4,76.4]	81.4 [77.3,89.4]	41.0 [26.0,55.3]
	ds-ln-low	db-standard	43.3 [36.0,47.4]	92.7 [90.0,100.0]	70.9 [68.5,77.5]
		db-bac-comp	38.6 [35.1,42.9]	93.5 [90.0,96.2]	75.5 [72.3,78.3]
		db-bac-comp-np	36.7 [33.3,41.1]	92.9 [90.0,96.2]	77.0 [73.7,79.8]
		db-bac-rep	34.8 [31.6,39.5]	87.4 [83.8,91.6]	77.9 [74.8,80.9]
		db-bac-rep-np	34.4 [30.2,37.3]	87.2 [82.5,91.6]	78.2 [76.2,82.2]
	ds-un-low	db-standard	43.3 [36.0,47.4]	92.7 [90.0,100.0]	70.9 [68.5,77.5]
		db-bac-comp	38.6 [32.6,42.8]	93.2 [90.0,100.0]	75.1 [72.6,80.1]
		db-bac-comp-np	36.4 [31.1,39.6]	94.1 [90.0,100.0]	77.1 [74.7,80.9]
		db-bac-rep	33.8 [27.3,38.3]	90.2 [90.0,100.0]	79.0 [76.1,83.5]
		db-bac-rep-np	33.1 [27.3,37.2]	90.5 [90.0,100.0]	79.5 [76.8,83.9]

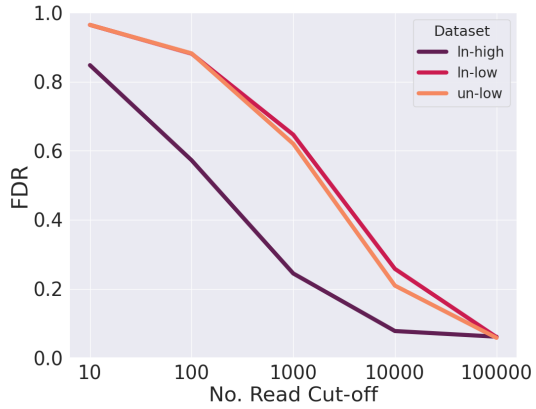
Table B.1: Species level taxonomic classification metrics F_1, true positive rate (TPR) and false discovery rate (FDR) (dataset mean [25 percentile, 75 percentile]) for different combinations of datasets and general bacterial screening databases, repeated for three Kraken 2 and Bracken parameter settings (left most column).

	Dataset	Database	$\mathbf{F}_1$ (%)	$\mathbf{F}_{1.5}$ (%)
Default (conf. 0, min reads 10)	**ds-ln-high**	**db-standard**	42.0 [29.9,55.5]	51.9 [40.1,65.6]
		db-bac-comp	39.4 [27.6,51.8]	49.3 [37.7,62.6]
		db-bac-comp-np	37.2 [25.5,48.5]	47.1 [35.3,59.7]
		db-bac-rep	34.2 [21.2,49.2]	44.5 [30.4,60.7]
		db-bac-rep-np	33.2 [20.5,47.3]	43.5 [29.6,59.2]
	ds-ln-low	**db-standard**	20.7 [15.5,23.6]	29.3 [22.9,33.4]
		db-bac-comp	20.2 [15.3,24.7]	28.9 [22.6,34.8]
		db-bac-comp-np	17.5 [13.1,19.7]	25.4 [19.7,28.6]
		db-bac-rep	8.3 [6.4,8.7]	12.7 [10.0,13.4]
		db-bac-rep-np	8.1 [6.2,8.7]	12.4 [9.6,13.3]
	ds-un-low	**db-standard**	18.2 [12.0,22.3]	26.2 [18.2,31.8]
		db-bac-comp	17.9 [12.1,21.4]	25.7 [18.2,30.7]
		db-bac-comp-np	15.5 [11.2,19.4]	22.7 [17.1,28.1]
		db-bac-rep	6.7 [4.7,8.5]	10.5 [7.4,13.2]
		db-bac-rep-np	6.5 [4.5,7.8]	10.0 [7.1,12.1]
Optimal (conf.0.075, min reads 750)	**ds-ln-high**	**db-standard**	77.9 [71.4,87.1]	72.8 [64.4,83.3]
		db-bac-comp	78.9 [73.2.87.7]	74.0 [66.4,84.4]
		db-bac-comp-np	79.0 [72.9,88.2]	74.2 [66.1,84.8]
		db-bac-rep	75.0 [66.4,84.2]	69.0 [58.9,79.6]
		db-bac-rep-np	75.0 [66.4,83.7]	69.1 [58.9,79.5]
	ds-ln-low	**db-standard**	95.5 [92.8,98.6]	96.2 [93.9,99.1]
		db-bac-comp	95.5 [93.2,98.1]	96.5 [94.2,98.5]
		db-bac-comp-np	94.2 [92.0,97.6]	95.7 [94.2,98.5]
		db-bac-rep	94.1 [91.5,98.2]	94.3 [91.7,98.0]
		db-bac-rep-np	94.3 [91.7,98.2]	94.4 [91.8,98.0]
	ds-un-low	**db-standard**	97.1 [94.7,100.0]	97.5 [96.6,100.0]
		db-bac-comp	95.6 [92.3,100.0]	97.0[94.9,100.0]
		db-bac-comp-np	94.1 [90.0,100.0]	96.0 [93.6,100.0]
		db-bac-rep	96.9 [94.7,100.0]	96.7 [92.9,100.0]
		db-bac-rep-np	97.2 [94.7,100.0]	96.8 [92.9,100.0]
Light (conf.0.05, min reads 50)	**ds-ln-high**	**db-standard**	85.0 [81.3,90.6]	84.8 [80.1,91.2]
		db-bac-comp	83.7 [79.4,89.7]	84.2 [80.9,90.3]
		db-bac-comp-np	82.8 [79.0,89.2]	83.6 [79.7,89.8]
		db-bac-rep	88.5 [86.0,91.8]	89.9 [87.9,92.5]
		db-bac-rep-np	88.2 [85.5,91.8]	89.7 [87.5,92.6]
	ds-ln-low	**db-standard**	79.8 [75.9,84.5]	86.2 [83.4,89.9]
		db-bac-comp	66.7 [59.8,72.8]	76.0 70.8,81.3]
		db-bac-comp-np	71.9 [67.4,77.6]	80.2 [77.1,84.9]
		db-bac-rep	70.1 [65.0,77.1]	78.5 [74.2,84.6]
		db-bac-rep-np	69.5 [63.2,76.2]	78.0 [72.9,83.9]
	ds-un-low	**db-standard**	77.8 [69.2,84.6]	84.5 [78.5,89.9]
		db-bac-comp	73.5 [66.0,80.5]	81.4 [75.9,87.0]
		db-bac-comp-np	66.7 [59.8,72.8]	76.0 [70.8,81.3]
		db-bac-rep	67.8 [60.0,76.9]	76.8 [70.9,84.4]
		db-bac-rep-np	66.4 [58.1,75.0]	75.7 [69.2,83.0]

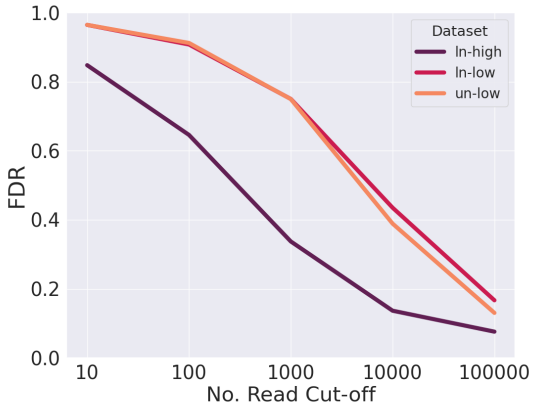
Table B.2: Genus level taxonomic classification metrics $F_{1.5}$ and F_1 (dataset mean [25 percentile, 75 percentile]) for different combinations of datasets and general bacterial screening databases, repeated for three Kraken 2 and Bracken parameter settings (left most column).

	Dataset	Database	TPR (%)	FDR (%)
Default (conf. 0, min reads 10)	ds-ln-high	db-standard	92.4 [89.7,96.4]	71.3 [60.7,81.9]
		db-bac-comp	92.2 [90.1,96.6]	73.6 [64.4,83.9]
		db-bac-comp-np	92.2 [90.2,96.6]	75.7 [67.1,85.2]
		db-bac-rep	98.3 [97.5,99.4]	78.4 [67.3,88.1]
		db-bac-rep-np	98.3 [97.5,99.4]	79.1 [68.9,88.5]
	ds-ln-low	db-standard	99.6 [100.0,100.0]	88.2 [86.6,91.6]
		db-bac-comp	100.0 [100.0,100.0]	88.5 [85.9,91.7]
		db-bac-comp-np	100.0 [100.0,100.0]	90.3 [89.1,93.0]
		db-bac-rep	99.6 [100.0,100.0]	95.7 [95.4,96.7]
		db-bac-rep-np	99.6 [100.0,100.0]	95.8 [95.5,96.8]
	ds-un-low	db-standard.	99.8 [100.0,100.0]	89.8 [87.5,93.6]
		db-bac-comp	100.0 [100.0,100.0]	90.0 [88.0,93.6]
		db-bac-comp-np	100.0 [100.0,100.0]	91.5 [89.3,94.0]
		db-bac-rep	99.6 [100.0,100.0]	96.5 [95.5,97.6]
		db-bac-rep-np	99.6 [100.0,100.0]	96.6 [95.9,97.7]
Optimal (conf.0.075, min reads 750)	ds-ln-high	db-standard	66.2 [55.7,78.0]	1.4 [0.0,2.2]
		db-bac-comp	67.7 [57.9,79.5]	1.7 [0.0,2.5]
		db-bac-comp-np	67.9 [57.6,80.0]	1.9 [0.0,2.7]
		db-bac-rep	61.5 [49.9,73.0]	0.9 [0.0,1.8]
		db-bac-rep-np	61.6 [49.9,73.0]	1.0 [0.0,1.8]
	ds-ln-low	db-standard	97.2 [94.6,100.0]	5.9 [0.0,10.2]
		db-bac-comp	98.3 [97.0,100.0]	6.9 [0.0,9.8]
		db-bac-comp-np	98.4 [97.1,100.0]	9.4 [3.6,13.8]
		db-bac-rep	94.7 [91.6,100.0]	6.0 [0.0,7.5]
		db-bac-rep-np	94.7 [91.6,100.0]	5.6 [0.0,7.9]
	ds-un-low	db-standard	98.3 [100.0,100.0]	3.9 [0.0,9.3]
		db-bac-comp	99.4 [100.0,100.0]	7.4 [0.0,11.9]
		db-bac-comp-np	99.4 [100.0,100.0]	10.0 [0.0,18.2]
		db-bac-rep	96.4 [90.0,100.0]	2.3 [0.0,0.0]
		db-bac-rep-np	96.4 [90.0,100.0]	1.7 [0.0,0.0]
Light (conf.0.05, min reads 50)	ds-ln-high	db-standard	84.9 [79.4,93.2]	13.4 [6.2,18.4]
		db-bac-comp	85.4 [80.4,93.7]	16.4 [8.4,24.2]
		db-bac-comp-np	85.5 [80.6,93.7]	18.1 [8.4,25.3]
		db-bac-rep	92.7 [89.2,97.1]	14.3 [6.0,21.9]
		db-bac-rep-np	92.7 [89.2,97.1]	14.8 [6.2,22.6]
	ds-ln-low	db-standard	99.3 [100.0,100.0]	32.6 [26.8,38.8]
		db-bac-comp	100.0 [100.0,100.0]	48.7 [42.8,57.3]
		db-bac-comp-np	99.7 [100.0,100.0]	42.8 [36.4,49.2]
		db-bac-rep	98.8 [98.4,100.0]	44.7 [37.2,50.9]
		db-bac-rep-np	98.8 [98.4,100.0]	45.3 [38.5,52.9]
	ds-un-low	db-standard	99.8 [100.0,100.0]	34.4 [25.0,47.1]
		db-bac-comp	100.0 [100.0,100.0]	40.4 [32.7,50.7]
		db-bac-comp-np	100.0 [100.0,100.0]	48.7 [42.8,57.3]
		db-bac-rep	99.4 [100.0,100.0]	47.4 [37.5,57.1]
		db-bac-rep-np	99.4 [100.0,100.0]	49.1 [40.0,59.1]

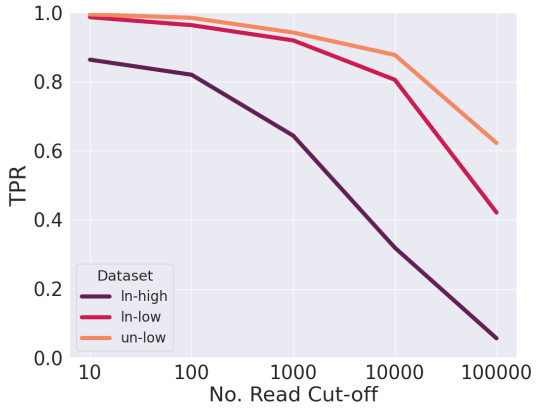
Table B.3: Genus level taxonomic classification metrics true positive rate (TPR) and false discovery rate (FDR) (dataset mean [25 percentile, 75 percentile]) for different combinations of datasets and general bacterial screening databases, repeated for three Kraken 2 and Bracken parameter settings (left most column).



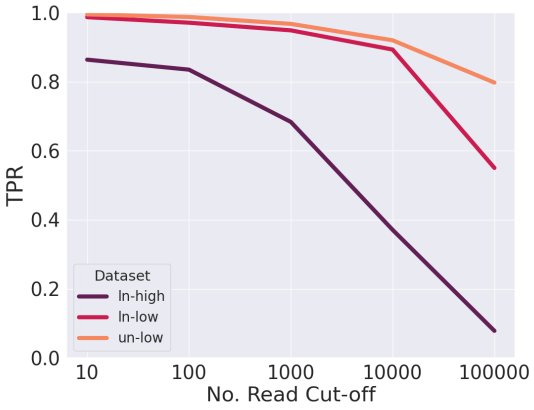
(a) FDR for Kraken 2 assigned reads



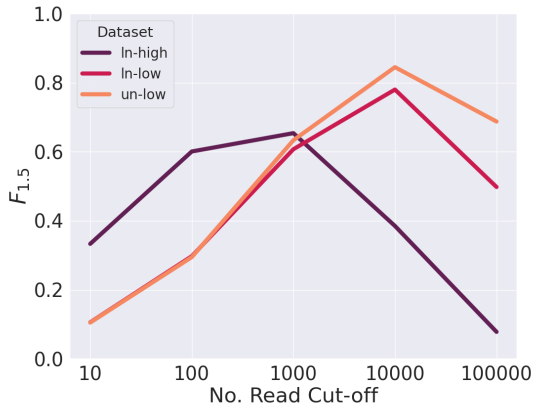
(b) FDR for Bracken estimated reads



(c) TPR for Kraken 2 assigned reads



(d) TPR for Bracken estimated reads



(e) $F_{1.5}$ for Kraken 2 assigned reads

(f) $F_{1.5}$ for Bracken estimated reads

Figure B.1: Average (a) false discovery rates (FDR) Kraken 2 read cut-off, (b) false discovery rates (FDR) Bracken read cut-off, (c) true positive rates (TPR) Kraken 2 read cut-off, (d) true positive rates (TPR) Bracken read cut-off, (e) $F_{1.5}$ scores Kraken 2 read cut-off and (f) $F_{1.5}$ scores Kraken 2 read cut-off, for the datasets *ds-ln-high*, *ds-ln-low* and *ds-un-low* over read classification abundance filtering for each species. The classification reference database used for (a)-(f) is the baseline *db-standard*. Note that the x-axis are logarithmic.

DB	Sample Group	P	N	TP	FP	FN
db-standard	**control-g0**	0	10	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
	control-g1	0	10	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
	control-g4	0	10	0.0 [0.0,0.0]	0.25 [0.0,1.0]	0.0 [0.0,0.0]
	a-baumannii-g0	1	9	1.0 [1.0,1.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
	a-baumannii-g1	1	9	1.0 [1.0,1.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
	a-baumannii-g4	1	9	1.0 [1.0,1.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
	s-aureus-g0	1	9	0.6 [0.0,1.0]	0.0 [0.0,0.0]	0.4 [0.0,1.0]
	s-aureus-g1	1	9	0.4 [0.0,1.0]	0.0 [0.0,0.0]	0.6 [0.0,1.0]
	s-aureus-g4	1	9	0.6 [0.0,1.0]	0.0 [0.0,0.0]	0.4 [0.0,1.0]
db-bac-subset	**control-g0**	0	10	0.0 [0.0,0.0]	1.0 [1.0,1.0]	0.0 [0.0,0.0]
	control-g1	0	10	0.0 [0.0,0.0]	1.25 [1.0,2.0]	0.0 [0.0,0.0]
	control-g4	0	10	0.0 [0.0,0.0]	1.5 [1.0,2.0]	0.0 [0.0,0.0]
	a-baumannii-g0	1	9	1.0 [1.0,1.0]	0.2 [0.0,1.0]	0.0 [0.0,0.0]
	a-baumannii-g1	1	9	1.0 [1.0,1.0]	0.6 [0.0,1.0]	0.0 [0.0,0.0]
	a-baumannii-g4	1	9	1.0 [1.0,1.0]	0.2 [0.0,1.0]	0.0 [0.0,0.0]
	s-aureus-g0	1	9	1.0 [1.0,1.0]	0.2 [0.0,1.0]	0.0 [0.0,0.0]
	s-aureus-g1	1	9	1.0 [1.0,1.0]	0.6 [0.0,1.0]	0.0 [0.0,0.0]
	s-aureus-g4	1	9	1.0 [1.0,1.0]	0.2 [0.0,1.0]	0.0 [0.0,0.0]

Table B.4: Species level taxonomic classification of the *ds-un-low* dataset where only the subset of 7 species were considered. The average [min, max] number of true positive (TP), false positive (FP) and false negative (FN) for the five replicates in each dataset group are listed together with the theoretically expected number of positives (P) and negatives (N). The dataset was run for the databases *db-standard* with Kraken 2 confidence = 0.075 and Bracken read threshold = 750 and the *db-bac-subset* with Kraken 2 confidence = 0.3 and Bracken read threshold = 10,000.

Sample Group	$\mathbf{F}_1$ (%)	$\mathbf{F}_{1.5}$ (%)	TPR (%)	FDR (%)
ani100__cLOW__stTrue	68.3 [63.0,74.0]	62.4 [57.0,68.1]	54.9 [49.4,60.5]	8.9 [7.3,11.6]
ani100__cLOW__stFalse	33.1 [26.5,40.9]	27.7 [21.5,34.9]	22.0 [16.4,28.2]	30.6 [22.2,34.4]
ani100__cHIGH__stTrue	53.5 [51.8,54.2]	45.8 [44.1,46.4]	37.2 [35.6,37.7]	4.5 [3.5,4.9]
ani100__cHIGH__stFalse	25.9 [24.5,28.7]	20.6 [19.3,23.0]	15.5 [14.4,17.5]	19.3 [16.3,21.6]
ani99__cLOW__stTrue	80.8 [78.6,83.4]	76.3 [75.0,79.2]	70.2 [68.2,73.0]	4.3 [2.4,5.3]
ani99__cLOW__stFalse	68.2 [66.9,71.8]	62.8 [60.5,67.5]	55.8 [52.4,61.5]	11.7 [7.5,14.7]
ani99__cHIGH__stTrue	66.8 [63.6,71.0]	59.6 [56.1,64.4]	50.9 [47.1,56.1]	2.2 [1.3,3.0]
ani99__cHIGH__stFalse	46.0 [42.7,48.7]	38.5 [35.3,41.1]	30.5 [27.6,32.8]	4.7 [4.0,5.3]
ani97__cLOW__stTrue	82.9 [80.4,84.7]	78.8 [75.6,81.5]	73.0 [68.8,76.4]	3.9 [2.2,4.7]
ani97__cLOW__stFalse	67.1 [66.3,68.3]	62.0 [61.1,64.8]	55.3 [53.7,59.1]	13.6 [11.0,17.3]
ani97__cHIGH__stTrue	61.3 [57.5,63.8]	53.6 [49.5,56.2]	44.6 [40.6,47.2]	1.6 [1.0,1.7]
ani97__cHIGH__stFalse	50.4 [48.5,51.4]	42.6 [40.8,43.6]	34.2 [32.5,35.1]	4.3 [3.2,5.1]
ani95__cLOW__stTrue	66.4 [61.6,69.1]	61.2 [55.4,65.0]	54.4 [48.0,59.9]	13.7 [9.9,17.5]
ani95__cLOW__stFalse	57.0 [52.6,61.5]	51.6 [47.3,57.2]	45.1 [40.5,51.1]	20.3 [14.4,24.9]

Table B.5: Species level taxonomic classification metrics F_1, $F_{1.5}$, true positive rate (TPR) and false discovery rate (FDR) for the bacteria in the dataset *ds-ln-high* run for the *db-standard* database with Kraken 2 confidence = 0.05 and Bracken read threshold = 50. The values listed are means of the 10 replicates in each dataset group together with the minimum and maximum [min, max] of the sample group.

Sample Group	$\mathbf{F}_1$ (%)	$\mathbf{F}_{1.5}$ (%)	TPR (%)	FDR (%)
ani100__cLOW__stTrue	58.4	65.6	82.7	54.4
	[52.9,63.6]	[60.8,70.6]	[81.9,84.8]	[50.2,60.8]
ani100__cLOW__stFalse	63.2	63.8	64.8	38.0
	[59.1,67.3]	[61.4,67.0]	[62.1,67.8]	[34.4,42.7]
ani100__cHIGH__stTrue	74.9	75.3	76.0	26.1
	[73.7,76.4]	[74.1,76.6]	[75.3,76.9]	[23.8,27.5]
ani100__cHIGH__stFalse	66.3	63.1	58.5	23.5
	[64.5,69.1]	[61.3,65.8]	[56.5,61.1]	[21.1,24.8]
ani99__cLOW__stTrue	61.7	70.7	93.7	53.2
	[54.9,65.9]	[65.8,74.0]	[92.3,94.1]	[48.8,61.6]
ani99__cLOW__stFalse	66.9	74.2	90.1	46.4
	[63.7,71.4]	[72.2,77.5]	[87.6,92.5]	[42.9,50.9]
ani99__cHIGH__stTrue	79.3	81.5	85.5	25.8
	[78.6,81.3]	[81.1,81.6]	[83.8,87.1]	[22.0,27.7]
ani99__cHIGH__stFalse	79.5	78.8	77.7	18.4
	[77.1,81.8]	[76.8,81.4]	[75.7,80.3]	[15.6,20.8]
ani97__cLOW__stTrue	55.5	65.8	93.8	60.3
	[51.1,59.7]	[62.3,69.2]	[92.3,95.7]	[55.9,64.9]
ani97__cLOW__stFalse	64.4	72.3	90.5	49.5
	[60.8,67.7]	[69.6,75.5]	[89.9,92.8]	[46.3,54.3]
ani97__cHIGH__stTrue	77.3	79.3	82.9	27.5
	[76.9,78.1]	[78.7,79.7]	[81.3,84.5]	[25.9,29.2]
ani97__cHIGH__stFalse	77.3	78.2	79.7	24.8
	[76.6,78.5]	[77.3,79.3]	[78.1,81.2]	[23.3,25.7]
ani95__cLOW__stTrue	53.1	61.7	83.9	60.8
	[50.7,55.2]	[60.2,63.5]	[82.9,85.8]	[58.7,64.1]
ani95__cLOW__stFalse	55.7	62.9	80.7	56.4
	[54.3,59.9]	[59.5,66.9]	[78.0,85.0]	[52.5,59.2]

Table B.6: Species level taxonomic classification metrics F_1, $F_{1.5}$, true positive rate (TPR) and false discovery rate (FDR) for the bacteria in the dataset *ds-ln-high* run for the *db-bac-rep* database with Kraken 2 confidence = 0.05 and Bracken read threshold = 50. The values listed are means of the 10 replicates in each dataset group together with the minimum and maximum [min, max] of the sample group.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY