Norwegian University
of Life Sciences

Master's Thesis 2022 30 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Classification and quantification of the contents of fishmeal using hash-based classification algorithms on metagenomic sequence data

Halvor Ekeland
Chemistry and Biotechnology

Acknowledgements

This master thesis is the culmination of my six years of studies at the Norwegian University of Life Sciences and I would like to give a huge thanks to my supervisor Hilde Vinje for all the constructive feedback and help! You have encouraged me to think, discuss and have an opinion and have always lifted my spirit when my hope was waning!
In addition, I would like to thank my co-supervisors Lars Snipen at NMBU and Jamie Ortiz at Cornell, who's understanding of the subject, the data and the code have been of great help when I got lost.

Lastly, I would like to thank my mom and dad. I remember like it was yesterday the evening we discussed if this was the study program for me, and it feels surreal to be at the end of the journey that started that evening. Thank you! None of this would have been possible without your endless love and support.

Abstract

With the abundance of sequence data and the development of effective computer algorithms, new practical applications using sequence data are possible, among the authentication of the contents of food and feed. This study aims to explore if metagenomic sequence data can be used to classify and quantify the contents of fishmeal samples. In this study the hash-based classification software Kraken2+Bracken was used to classify and quantify the contents of fishmeal samples. The study utilized simulated sequences as a best-case scenario and real DNA sequence data and assess the ability of Kraken2+Bracken to classify and quantify the contents with the measurements purity, completeness, Bray-Curtis dissimilarity and Principal component analysis. The results indicate that Kraken2+Bracken is able to classify the DNA sequence in the samples, but that the accuracy of the classifications and quantifications are dependent on the threshold and settings used, and for the samples based on real meatgenomic data, sample preparation likely affect the accuracy.

Abstrakt

Med den store mengden sekvensdata og utviklingen av effektive data algoritmer, nye praktiske anvendelser av sekvensdata er mulige, blant dem autentisering av innholdet i mat og fòr. Denne studien har som mål å utforske om metagenomisk sekvensdata kan brukes til å klassifisere og kvantifisere innholdet i fiskemel blandinger. I denne studien ble den hash-baserte klassifiserings programvaren Kraken2+Bracken som ble brukt til å klassifisere og kvantifisere innholdet i fiskemel blandinger. Studien benyttet simulert data som et ideal-tilfelle og ekte DNA sekvensdata og vurderte Kraken3+Bracken sin kvantifiserings og kvantifiseringsevne ved å bruke målene purity, completeness, Bray-Curtis ulikhet og principal component analysis. Resultatene antyder at Kraken2+Bracken er i stand til å klassifisere DNA sekvensene i blandingene, men at nøyaktigheten til klassifiseringen og kvantifiseringen er avhengig av terskelverdier og innstilingene brukt, og for blandingene basert på ekte metagenomisk data er det sannsynlig at hvordan prøvene tilberedes påvirker nøyaktigheten.

1 Contents

2 Introduction

Around 10 billion to 15 billion dollars yearly is lost to food fraud according to the American grocery manufacturers association (Johnson 2014). Food fraud does not only come at a financial cost, but also at a health risk and an environmental cost. An example is the use of mustard oil adulterated with argemone oil, which can cause serious health problems (Challagundla Kishore Babu 2007) and inflated estimates of fish stocks as a result of mislabeling, leading to overfishing (Blanco-Fernandez, Garcia-Vazquez et al. 2021). These examples of food fraud illustrate the importance of trusted and reliable authentication methods for agricultural and aquaculture products. New methods to ascertain the contents of food items have been developed that can be done in a laboratory and is therefore not based on trust or documentation accompanying the product, but rather a reproducible result acquired by testing the product itself.

Global demand for meat and seafood has increased over the last decades and is expected to rise in coming decades as the number of consumers of such products grows larger (Salter 2017) (Delgado, Wada et al. 2003). As a part of the industrial agriculture and aquaculture needed to meet the increased demand for meat and seafood, fishmeal is used as an additive to animal and fish feed. Fishmeal is rich in proteins, with a good amnio acid composition and rich in energy. This makes the fishmeal an ideal component of animal feed as it means less feed is required and that higher growth rates for animals and fish can be achieved (Miles and Chapman 2006). This is beneficial as less feed is required for the same nutritional needs to be met compared to feed from other sources, this intern reduces the harmful waste produced by the fish (Mente, Pierce et al. 2006). Fishmeal is today made mostly from fish that are not useable for human consumption and some is made from the byproducts from the production of fish products for human consumption.

Development and innovation in the field of DNA sequencing has been moving at breakneck speeds the last two decades. Not long ago, sequencing was a labor intensive and argues process, where reading gels was common practice. Now the computer is king and the most important tool for analysis of genetic information (Heather and Chain 2016). Modern high-throughput sequencing machines can generate billions of base-pairs requiring little time, and with relative ease (Levy and Myers 2016). The massive generation of high-quality sequence data is now making it possible to reliably, in a cost-effective way, utilize sequence data for new data intensive applications, also within food authentications. Metagenomic analysis used for food authentication would provide governments with a new way to assess and facilitate sustainable food production, give businesses the secure knowledge and ability to avoid fraud and allow consumers to make more informed decisions when choosing what food items to purchase (Haynes, Jimenez et al. 2019).

As genome databases have improved and with increased ease of DNA-data proliferation, new and fast tools have been developed to utilize the massive increase in sequence data. Kraken2 is an example of a DNA sequence classifier that uses hash algorithms, similar to those of search engines, to match DNA-sequences quickly and accurately. This is an improvement from alignment-based approaches which are extremely accurate, but practically impossible to utilize with the amount of data now available. Therefore, Kraken2 promises to be a fast, effective, and sufficiently accurate way to utilize sequence data for food (Haynes, Jimenez et al. 2019).

2.1.1 Aim of study

This study is a subproject of an ongoing project by Orivo As and Patogen to explore the possibility for developing a laboratory-based certifications scheme for feed products used in aquaculture. As stated in the application to the Research Council of Norway the goal of the project is to *"make it possible to identify and quantify which species they [the feed products] consists of with high accuracy"*.

The aim for this study is to explore if DNA sequence data from metagenomic shotgun sequencing can be used to classify and quantify the contents of fishmeal reliably and accurately. This will be done by using real metagenomic sequence data and simulated DNA sequences. Simulated sequences will be generated using genome assemblies available in the NCBI genbank as a template and the Art program. Real data stems from various laboratories and utilizes different sequencing technologies. This will indicate how the differences affects the ability to correctly classify and quantify the contents of different real samples compared to samples based on perfect simulated DNA sequences. The classifications will be done by Kraken2+Bracken, which is a hash-based classification algorithm that classifies DNA sequences to different species. To assess the ability of Kraken2+ Bracken to classify DNA sequences, purity and completeness will be calculated for each sample. The study will utilize Bray-Curtis dissimilarity to measure the ability to correctly predict the composition of the samples. To study Kraken2+Bracken's ability to classify reads, PCA will be used on centered log ration (CLR) transformed classification data to see how different confidence requirements in the classification prosses affects the classifications.

3 Theory

The following section will present the theory and techniques that makes the metagenomic classification and quantification of the contents of fishmeal possible. The first section presents how fishmeal is made and how DNA is degraded. Section 3.2 presents how DNA is sequenced and sequence data obtained. In section 3.3 the theory behind the tools utilized in this study is laid out. The following two sections, 3.4 and 3.5, present the ways the classification and quantification will be measured and theory relevant to factors that might affect the results.

3.1 Fishmeal

The production of fishmeal is to a large extent standardized across the world. The process stars with mincing of raw materials to reduce the size of particles before the heating procedure. To extract moisture and oils from the minced raw materials and to inactivate microbes and viruses present in the material that can ruin it, heat is applied. The temperature is usually around 75 degrees for about 20 – 25 minutes depending on what the raw material are made of. The heated material is then sent to a strainer to remove additional moisture in the form of press liquor, the heated and demoisturized materials are then sent to a dryer where it is again heated. The drying is done either by a drying screw heated by steam or by hot air. The drying reduces the moisture from around 60% to around 12%. When the materials leave the dryer after approximately 30 min, the temperature of the materials is around 80 degrees Celsius. During cooling, additional moisture is removed and the temperature of the materials is reduced from 80 degrees Celsius to room temperature. After cooling, the materials are grinded so that the particles size becomes homogeneous, before packing the now finished fishmeal is packaged. (Marvin Ingi Einarsson 2019)

3.1.1 DNA Degradation during processing

When food is processed the DNA will deteriorate. Degradation for DNA is primarily caused by changes in temperature and pH, where pH has the biggest effect according to (Torsten Bauer 2003). Low pH had a big effect on the rate of DNA degradation and high temperatures some effect (Torsten Bauer 2003). This degradation was measured using HPLC and quantified using the ratio between the area of circular covalently bonded DNA and total area of plasmid structures. Bauer et al. finds that larger fragments over 1000 bp are no longer detectable using a primer-based approach but notes that in soy; fragments of around 700bp are detectable after a rigorous process involving alkaline lysis, removal of non-alkaline soluble components and acidic precipitation of proteins and spray drying, indication that DNA is still precent in fragments that could be sequenced.

3.2 DNA and DNA sequence data

DNA is the fundamental unit that carries biological information from generation to generation. For many years the organization of heritable traits were unknow, but in the latter half of the twentieth century the discovery of the double helix gave a model for future understanding of the human genome. Today DNAs composition and ways of replicating are well established, and modern sequencing technologies gives the possibility of studying the sequence of DNAs four constituent bases in incredible scale and detail.

3.2.1 DNA Sequencing

The first sequencing system was developed by Fredrick sanger, and although a brake though, it was a time and labor-intensive process. Today DNA sequencing is done with automated, miniaturized and

massively paralleled systems that produce huge amounts of data. Two of the biggest players in the DNA sequencing industry is Illumina sequencing and Thermo fisher, each with their own sequencing approach. This study uses simulated DNA sequences and real DNA sequences from two different sequencing systems, Illumina and ion-torrent sequencing from Thermo fisher.

Illumina pair-end sequencing is based on a sequencing approached called sequencing by synthesizing and is arguably the most common sequencing system for short reads. Before the sample can be sequenced, the DNA in a sample must be prepared and a library of DNA fragments created. the standard library preparations require fragments of length 200 – 500, though up to 1000 base pairs is possible (Quail, Swerdlow et al. 2009) It uses a cyclical program on a flowcell containing millions of primers permanently attached to the flowcell. These attach to and fix the fragments of DNA that are to be sequenced. The DNA sequence is read using a laser to emit a differently colored light signal for each of the four fluorescently tagged bases precent in the solution surrounding the flowcell. This light signal is detected, and a base is called depending on the color of the light emitted when the tagged bases are hit by the laser. The primers used in the sequencing are oligonucleotides and come in two varieties, one for the forward strand and the other for the revers strand. The process results in read pairs which makes it possible two associate two sequences, with length equal to the number of cycles chosen (Hu, Chitnis et al. 2021). The illumine sequence data simulated and the data from real sources used int this study consist of reads with a length of around 150 nucleotides for each read in the read pair.

Ion torrent is a system developed by Thermo Fisher and uses the changes in pH to make the base calls. In an ion torrent system, the DNA fragments that are to be sequenced get attached to primers on a bead, the bead will be deposited into a well on a microchip that can detect changes in pH. The fragments are DNA from the samples to be sequenced. For ion torrent, the length of the DNA fragments should ideally be between 100 – 600 depending on the library preparation protocol (Quail, Smith et al. 2012) (Forth and Höper 2019). The Sequenced of the fragment that make up the library is sequenced one cycle at a time. In every cycle, the chip with millions of wells is “washed” with a solution containing one of the four different nucleotides. When one or more nucleotides are added to the DNA-fragment the pH will change and will change the voltage of a current that can be measured by the chip (Merriman, D Team et al. 2012). The Ion torrent sequencing system usually produces reads from length 200 – 600 bases.

Data from next generation sequencing technologies are usually formatted into either the fasta format or the FASTq format, and FASTq has become widely used because of the addition of information on the quality for the read. The fastq file has 4 lines for each read it contains, one containing the read id and additional data about the read, the second line contains the sequence, the third is usually a placeholder line containing only a plus sign, and the fourth line contains the per nucleotide quality information. This quality information is a score for every base in a given read using the Phred system (Cock, Fields et al. 2009). There are several variations for the fastq file format, developed by different entities with different compatibilities in mind. In this study the fastq format used by Illumina is used.

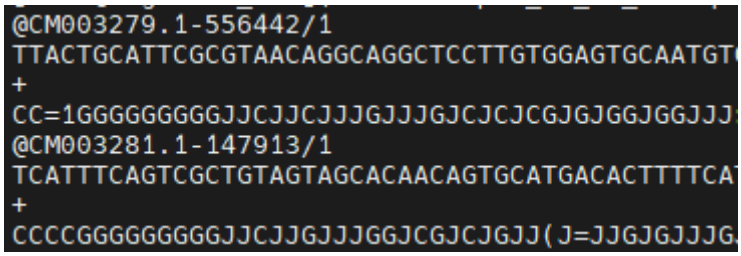

Figure 1: The 8 lines of 2 reads in a fastq file. Two reads are partly shown (the full sequence is note visible)

3.2.2 DNA sequence quality

The fastq format variant is characterized by Phred-scores with an encoding of +33, meaning that the scores integer value go from 0 to 42, starting at the 33 symbol of the ASCII scoring system (illumina 2021). Phred scores are called for each based and gives information on the probability of that base being called correctly.

Equation 1 is used to calculate the probability that a base was called erroneously, where P represents the probability of an erroneously called base, Q represents the integer value of the corresponding Phred quality score to the base (Cock, Fields et al. 2009).

Equation 1: Calculating P, probability that a base was called erroneously, where Q is the Phred score from 0 to 42

$$P = 10^{\frac{-Q}{10}}$$

A score of " ! " corresponds to Q = 0, giving a P of 1, complete certainty of base-call error. A score of "K" corresponds to Q = 42, giving a P = 0.00006.

3.2.3 Read simulation with Art

To study the ability to classify and quantify the contents of samples based on their metagenomic sequence, simulated reads can be used as a "best case scenario". The simulated reads are made from genomes in the NCBI GenBank (Coordinators 2016). The simulated reads, and the sample based on them will not have any errors or biases introduced by the steps necessary to sequence a real sample, and therefor allow the study of how well a metagenomic profiler dose under ideal conditions. Since the reads simulated are based on the same genomes that are in the database, the sequences are the same and unambiguous classifications should be easier to achieve.

Art is a software made to simulate the read sequences from different sequencing technologies (Huang, Li et al. 2012). It utilizes a reference genome and allows the user to choose which sequencing technology to simulate. In the script the user can specify which sequencing platform is use, the program then uses the error profile for that platform when constructing the fastq file for the simulated reads (Tille 2014). This allows the simulated reads to have an error profile consistent with what Is expected for the given platform. As seen in Figure 2, the genome is made into reads of the chosen length (step 1) and then the quality profile is used to add substitution errors (step 2, errors marked in orange), where the chance that a base will be changed is based on the Phred score of that base in the quality profile, shown as colored bars under the reads. The reads are then added to a fastq file (step 3). The quality profile is created from real sequencing data and is a probability distribution of error for each of the bases in the read. (Huang, Li et al. 2012).

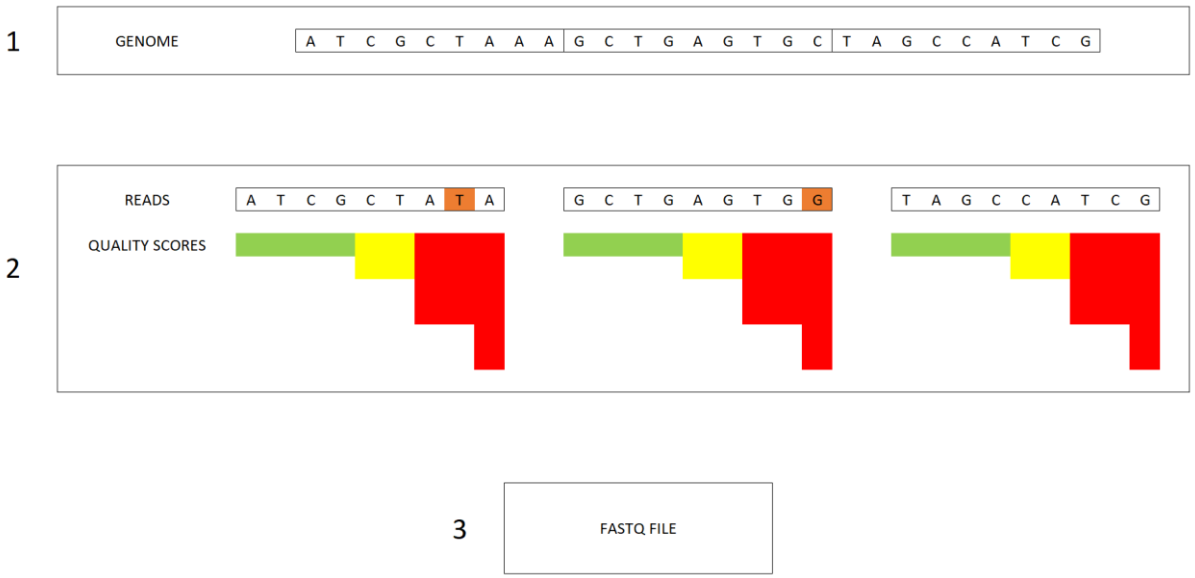

Figure 2: The process of read simulation using Art.

3.3 Taxonomic classification of DNA sequence

Classification, also known as discrimination, is when data points are assigned to a category based on one or more traits. This can be classifying cars based on the type of energy source they use or classifying tumors to different types of cancer based on the expression of different genes in their tissue. Classifications can be made by many different methods; the proximity to observations belonging to known categories (Keller, Gray et al. 1985), or how they relate to a line or plane in a space defined by the traits (Noble 2006), or by a decision tree (Dreiseitl and Ohno-Machado 2002). Traits can be anything that is considered usable to distinguish observations belonging to different defined categories.

3.3.1 k-mers

There are many ways to determine if a sequence belongs to a species. Some tools search data bases by aligning sequences, like Blast (Johnson, Zaretskaya et al. 2008). Other tools use comparison of specific parts of the DNA, a DNA barcode, with a database of barcodes for known species (Weitschek, Fiscon et al. 2014). The tool used in this study utilize k-mers as the mode of comparing sequences. A k-mer is a sequence of length k derived from the reads or genome. As the amount of sequence data in repositories have grown and tools have been developed that can utilize large amounts of sequence data in bioinformatics analysis, the old ways of querying and comparing sequences have become untenable as they become too computational and/or time demanding. k-mers represent a streamline and efficient way of studying sequence similarity among other things (Marchet, Boucher et al. 2021). k-mers are created by defining a length k and making k-length sequences out of a DNA sequence. if k = 3 then a sequence of n = 10 bases would contain N number of k-mers, where N can be calculated by the following formula:

$$N = n - k + 1$$

Equation 2: Number of k-mers N, resulting for sequence of length n with k as length of k-mer

The k-mers can be said to be made by sliding a k-length window along a DNA sequence and making a subsequence for every base and the k-1 subsequent bases as Illustrated in Figure 3.

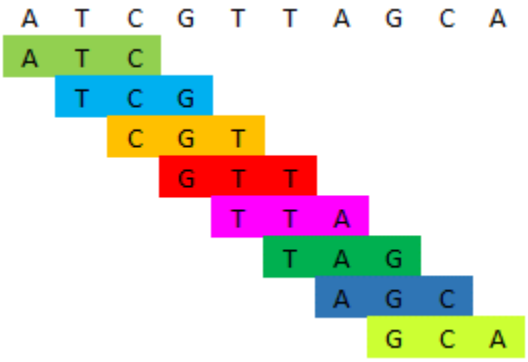

Figure 3: The N = 8 k-mers (3-mers) resulting from a sequence of n=10 bases when k = 3. Colored to distinguish unique k-mers

3.3.2 Kraken2

The taxonomic classification of DNA sequences is the process of classifying the reads that compose the metagenomic DNA sequence of a sample, to different taxonomic ranks. The categories that a sequence can be classified to are the different taxa within a rank, i.e. different species or different genera, for example species or genus. The taxonomic ranks used in this study is the NCBI taxonomic ranks. The traits are the the k-mers of the reads. The k-mers are not classified, but their assignment to the taxonomic tree determines the classification of the read that the k-mers originate form. Therefore, in the text, reads get classified, k-mers get assigned and species get predicted as present, or not present, based on the classification of the reads.

This study aims to classify and quantify the contents of fishmeal using metagenomic sequence data. To classify the reads, Kraken2 was chosen as the tool used for classifying the contents of the samples. Kraken2 is a software for taxonomic classification of biological sequences, most commonly DNA sequences (Wood, Lu et al. 2019). The goal is to take a sequence, compare it to a database and assign it a taxonomic ID. The k-mers are compared to k-mers from the genomes in the database used for the classification. This is done by using kraken2s hash function to create hash-values, each k-mer gets a unique hash-value. The hash-based approach allows for fast comparison of k-mers from samples and genomes. The database can be custom built, containing only genomes that the user finds relevant to look for or a standard that can be easily downloaded. Since the contents of the mixtures of reads that are to be classified in this study are known, a custom database consisting of 19 genomes from the NCBI genome assembly database will be used.

The database requires that we have reference genomes with known taxonomy. Kraken2 databases uses NCBI taxonomical ID. This is used to assign the k-mers to a rank in taxonomic tree. The assignment of the k-mers to the different ranks is used as the basis for the classification.

The Kraken database consists of k-mers from the genome provided. These k-mers are created and counted by the jellyfish software. The database stores information on k-mer count and what lowest common ancestor I associated with the k-mer. For k-mers found in more than one species, the lowest common ancestor will be at a rank higher than species. This association of species with k-mers is

facilitated though the use of taxonomic ids, which is the reason the database requires genomes to have a taxonomical id when being added to the Kraken2 database (Wood, Lu et al. 2019).

Each k-mer is associated with the lowest common ancestor (LCA) taxon. The k-mers of a read are associated with different ranks of the taxonomic tree. The weight of each rank is the number of k-mers that fall into that level. The classification is based on the lowest level in a branch which has k-mers associated with it (Wood, Lu et al. 2019).

Some of the reads are classified as higher ranks by kraken2. Normally the standard database used for Kraken2 is biased towards genomes that stem from an area of study where a lot of similar genomes have been assembled and added to the database. Since these genomes are more alike each other, a larger number of reads will be identical resulting in an increased chance that a given k-mer is assigned/associated with a higher rank (Wood, Lu et al. 2019). For this study, the use of a custom database circumvents this issue. The selection of genomes is done to fit the samples and goal of the analysis, and not just based on the genomes added to the standard Kraken2 database.

3.3.3 Bracken

Estimating abundance of a given species requires the reads to be assigned to taxa at species level. Kraken2 will assign the reads at higher levels if there is ambiguity as to which species, genus ect. that read belongs to. Because of these higher-level classifications, abundance estimation based on the kraken2 results alone becomes inaccurate as many reads will be classified to a rank shared by many species. Therefore, Bracken is used to re-classify the reads of higher rank to a specific species. Bracken stands for “Bayesian Reestimation of Abundance after Classification with KrakEN” and as the name implies uses Bayesian probabilities to re-classify reads of a higher rank than species (Lu, Breitwieser et al. 2017).

3.4 Classification assessment

To assess the classifications made by Kraken2+Bracken the Open-community Profiling Assessment tooL (OPAL) framework will be used as a starting point. A first step in the analysis of the data is to study the ability of Kraken2+Bracken to classify the reads. This is done by measuring the purity and completeness as defined by Meyer et al (Meyer, Bremges et al. 2019). This is done by counting what species in a sample have reads assigned to them and which do not. A True Positive (TP) is defined as a species that is present in the sample with reads from a sample assigned to it, equally; A False positive (FP) is defined as a species not present in the sample with reads from the sample assigned to it. A True False negative is when a species present in a sample that has no reads assigned to it.

Table 1 shows explains how True Positives (TP), false positives (FP), false negatives (FN) and True Negatives (TN) are counted. The columns pertain to presence in the sample, the rows pertain to the discovery of reads belonging to that species by Kraken2+Bracken

	In the sample	Not in the sample
Found by KB	True Positive	False Positive
Not found by KB	False Negative	True Negative

In this study the phrase “predicted to be present” refers to a species that has reads classified to it above a defined threshold. If the threshold is 10 reads, then 10 reads or more must be classified as a

givens species for that species to be predicted as present. If that species is also present meant to be present in the sample, it is counted as a true positive.

3.4.1 Purity

Meyer et al. defines purity as the fraction correctly predicted taxa divided by the total number of taxa predicted to be present made for a given rank, in this study rank is species. Purity is therefore a measurement of what fraction of the species predicted to be in the sample that are in the sample. A score of 0 indicates that there are no true positives, in other words: none of the species in the sample are predicted to be present. A score of 1 indicated that there are no false positives, i.e., all species predicted to be present in the sample, are in the sample. Equation 3 states how purity is calculated.

Equation 3: Equation for calculating the purity of the sample, where purity is the share of species predicted to be present that are present in the sample

$$Purity = \frac{TP}{TP + FP}$$

3.4.2 Completeness

Completeness is defined as the number of correctly predicted taxa for a given sample divided by the total number of taxa in that sample. Completeness therefore indicates the ability of KB to find every taxon that is in the sample, in this study the rank is always species. A score of 0 indicates that none of the species know to be in the sample are predicted to be present. A score of 1 indicates that every species present in the sample is predicted to be present by Kraken2+Bracken, meaning also that there are no false negatives, i.e., species in the sample, that are not predicted to be present by Kraken2+Bracken. Completeness is calculated using the following equation, Equation 4:

$$Completenss = \frac{TP}{TP + FN}$$

Equation 4: Equation for the completeness of a sample, where completeness is the fraction of species present in the sample with reads classified to that species

3.4.3 Bray-Curtis dissimilarity

Meyer et al. Uses Bray-Curtis dissimilarity as a measure of the profiles ability to correctly approximate the abundance of the species in the sample, in this case the ability of Kraken2+Bracken to classify the correct number of reads when compared with the true number of reads for that species. The true number of reads or fraction of contents is referred to in the text as “the gold standard”. Bray-Curtis dissimilarity gives a number between 0 and 1, where 0 indicates perfect reconstruction of abundances, and 1 indicates complete dissimilarity, meaning that none of the reads for a given species can be found in the gold standard.

Equation 5 is used for calculating the Bray-Curtis dissimilarity, where i is the taxon, ie. Species, in the samples being compared, x_i is the number of reads in the gold standard and x_i^* is the number of predicted reads for species i. The nominator calculated the absolute value of the difference in number of reads between the sample and the gold standard for taxon i, while the denominator calculates the total difference in the number of reads for all the taxon in the samples being compared.

Equation 5: Equation calculating the Bray-Curtis Dissimilarity (BCD) for taxon i at rank r. Bray-Curtis dissimilarity becomes a value between 0 and 1, 0 being perfect similarity and 1 being complete dissimilarity.

$$BCD = \frac{\sum_i |(x_r)_i - (x_r^*)_i|}{\sum (x_r)_i - (x_r^*)_i}$$

3.4.4 Confidence in the classification by Kraken2

To study how the strictness of Kraken2 effects the classifications we increase the strictness when running Kraken2 by giving it different levels of "confidence". Confidence is an option when running Kraken2 where an input of a number between 0 and 1 is given, where 0 is the least strict and 1 is the strictest. This number denotes a fraction of k-mers that must be assigned to a species, for that read to be classified as the species. If the threshold is 0.5, then 50% of k-mers that were assigned a species, must be assigned to the same species for read to be classified. If the sequence does not pass this threshold the read remains unclassified (Krakcen2 manual) (Wood, Lu et al. 2019). Therefor as the threshold increase the number of unclassified reads is likely to increase. This will affect the Bracken classification as well, since bracken uses the distribution of k-mers for a given read in the taxonomic hierarchy as inputs to the bayes-based algorithm that re-classifies the read as a given species (Lu, Breitwieser et al. 2017). Reads that are unclassified by kraken2 do not get classified by Bracken.

3.4.5 Principal component analysis

Principal component analysis is a powerful and versatile statistical tool that can be used to analyze how variables relate to each other and to cluster similar samples (Abdi and Williams 2010). In this study it will be used for the latter purpose. As the number of variables in for each observation increase it becomes less and less possible to discern meaningful insights from data, but at the same time a large number of variables can capture important variation among the samples (Francis and Wills 1999). Principal component analysis can be used to reduce dimensionality as variables are used for the calculation of principal components that capture the variation in the data. These components are then used to gain insight instead of the original observed variables (Abdi and Williams 2010)

The principal components may capture the variation that exist in the data. A "good" principal component analysis will result in a few variables that explain a lot of the variation. The number of principal components is equal to the number of variables used in the analysis. The components are created by a linear combination of the variables in the analysis and represent the relationship between the variables (Francis and Wills 1999).

The property of PCA analysis that is most relevant for this study is its ability to cluster similar observations when plotting using the two first principal components, which are the components that capture most of the variation. This can be used to visualize the effect that a different level of confidence in Kraken2 can have on the ability to accurately quantify the contents of the samples and compare them to them with the gold standard. Samples that are in close proximity to the gold standard indicate a more accurate reproduction of the relative contents of the samples.

3.5 Compositional and evolutionary relationships in DNA sequence data

3.5.1 Centered Log Ratio

Atchison suggest that a transformation can be utilized to mitigate the problems with compositional data. These problems stem from the fact that the data exist in a positive simplex where the different traits studied, in this case the number of reads or relative abundance have an upper limit. The simplex is a sample space with positive values, as no abundance can be negative (Aitchison 1983).

The species is either absent with a relative abundance of 0, or present with a relative abundance greater than 0. The abundances are therefore not independent from each other as a change in one will affect the others, since there is a total, whether that is 1 when using relative abundances or the total number of reads in the sample when using read counts.

The sequences resulting from the sequencing machine are compositional data as a result of there being a limit to the number of DNA fragments a given sequencing system can sequence. This means that there is a relationship between the number of sequences from different sources. It is a null-sum scenario where to get one more read of one species, there must be one less from one of the other species (Gloor, Macklaim et al. 2017). This means that the abundances are not independent of each other, and there for the assumption of independent observations is not met. As a result of this many statistical analyses cannot be carried out with satisfying reliability. A transformation of data is therefore necessary and Gloor et al. suggest that a centered log ratio transformation can be utilized so that the data better meet the requirements of a common approaches like PCA. The Centered log ration is calculated using the following equation:

Equation 6: Equation for transforming the number of reads x, for the species j, in the sample s.

$$s_{clr} = [\log\left(\frac{x_j}{G(s)}\right), \log\left(\frac{x_{j+1}}{G(s)}\right), \dots, \log\left(\frac{x_n}{G(s)}\right)]$$

Where s_{clr} is a vector of the CLR-transformed read numbers x, for species j in sample s. G(s) is the geometric mean of the sample calculated by the equation below.

Equation 7: Equation for calculating the geometric mean

$$G(s) = \log\left(\sqrt[n]{x_j * x_{j+1} * \dots * x_n}\right)$$

Where x is the number of reads classified to species j, for sample s with n species.

3.5.2 Estimation of evolutionary relationships using mash distances

When classifying using k-mers or DNA sequences as the basis for classifications, species with similar genomes have a larger chance of being misclassified. Mutations, insertions, deletion, and transposons all change the DNA sequence of a species, and this variation increases with time. Species with a recent common ancestor is therefore more likely to share a larger part of their DNA sequence than distant related species. A close evolutionary relationship will therefore increase the probability that a given k-mer is found in more than one of the species. To study this, mash distance was used to see how the species chosen for the database relate to each other. Mash was chosen as it uses a similar k-mer based approach as kraken2. (Ondov, Treangen et al. 2016)

Mash distances are based upon comparisons of lexicographically sorted hash vales made using k-mers from different genomes of interest as the inputs to the hash function. A given k-mer produces a hash of a defined length that is unique to that sequence. A bottom sketch of s hashes is created, it is this sketch that will be used when computing the distances. Only hashes that have a smaller value than that of the maximum hash value of the sketch are included in the bottom sketch. This either removes the hash with the maximum value to keep the bottom sketch size at s or it adds the hash until the sketch reaches size s mash (Ondov, Treangen et al.).

The benefits of this approach it the reduced computational needs and time usage. The k-mers based hash comparison strategy means that there is no longer a need to compare the sequenced, which for applications with large genomes would be unpractical.

The lexicographically sorting of hashes when comparing means that a small percentage of the total sequence can quickly and accurately be used to estimate evolutionary distance. The principal here is that similar genomes will have more shared k-mers resulting in more shared hashes. The number of identical hashes is used for the calculation of mash distances. This calculation is based on the Jaccard index, which is the ratio of shared hashes to the total number of hashes compared. This is formulated by Ondov et al. as |A ∩ B| = shared k-mers, with |A ∪ B|= number for total k-mers in the genomes. As formulated in Equation 8

Equation 8: Jaccard index defined as the relationship between shared hashes (numerator) and total hashes compared (denominator)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Since each unique k-mer produces a unique hash-value, |A ∩ B| can be said to approximately equal the number of shared hash-values, and |A ∪ B| can be said to approximately equal to the total number of hash-values from the genomes. The number of k-mers in a genome is dependent on the size of the genome, and this will affect the results when comparing a much smaller genome to a larger one. Among the genomes selected for this study is the genome of Atlantic salmon. This genome is significantly larger than the others in the study because of a duplication event 80 million years ago coinciding with increased transposon activity (Lien, Koop et al. 2016). This difference in size is accounted for when using mash as it uses the average size of the genomes, n.

the Jaccard index can be framed and calculated using average genome size, which is useful when there are large differences in the genomes size. In Equation 9, w is shared k-mers, and n is the average size of the genomes (Ondov, Treangen et al. 2016).

Equation 9: Jaccard index j, framed in terms of average genome size, where w is number of shared k-mers, n is average number of k-mers pr genome compared.

$$j = \frac{w}{2n - w}$$

The accuracy for the estimation of distance is subject to the chosen length of k-mers, the longer the k-mers the greater the ability to separate closely related genomes. For this study the length k chosen is 32 base pairs, meaning 32-mers are utilized to create the hashes. The probability of observing two identical k-mers given no relationship among the genomes is four to the power of k, where four is equal to alphabet size. When applied to DNA sequences the alphabet size is equal to four as there are for primary nucleotides, A,T,C and G. k is the same k as in k-mers, in this case k = 32. The following equation is used to calculate the mash distance:

Equation 10: Equation for calculating mash distance for k length k-mers, with a Jaccard index of j

$$D = -\frac{1}{k}\ln\left(\frac{2j}{1+j}\right)$$

Where J Is the Jaccard index framed in terms of average genome size, this can be reordered to the following equation which is then incorporated to the Mash-distance equation above

Equation 11: the relationship between w, number of shared k-mers, n average genome size and the Jaccard index j

$$\frac{w}{n} = \frac{2j}{1+j}$$

For both formulas, w is the number of shared k-mers, n is the average genome size and j is the Jaccard index factoring in genome size. (Ondov, Treangen et al. 2016). From

The mash distance uses a poission distribution to estimate the rate of random site mutation in the genome. It does not model the many processes of evolution and measures distance based on similarity/dissimilarity, which in the context of this study is useful information when looking at which species get misclassified as which. (Ondov, Treangen et al. 2016).

3.5.3 Pearson Correlation

To study how sequence quality affects the classification and quantification accuracy, Pearsons correlation coefficient was used. This basic statistical measurement will indicate if there is any relationship between the quality of the reads and the ability to quantify, measured using Bray-Curtis dissimilarity.

Person correlations coefficient (PCC) is a number between -1 and 1, where -1 is a linear relationship that is inversely proportional and where 1 is a linear relations ship that is proportional. The PCC is calculated by dividing the covariance of to variables by the product of their individual standard deviations. Equation for calculating PCC is given by Equation 12.

Equation 12: Equation for calculation the Pearson correlation coefficient between two data series, A and B. cov(A,B) being the covariance between A and B, σ_A is the standard deviation of data series A and σ_B is the standard deviation of data series B

$$PCC = corr(A, B) = \frac{cov(A, B)}{\sigma_B \sigma_B}$$

4 Methods

This section explains how the different theory and techniques were used to classify and quantify the contents for the samples that were analyzed. The section has four sections; “4.2 Software and data” which states what software was used including which version, in addition to explaining the data that was used in the study. “4.3 Taxonomic classification of DNA sequences” presents how the software presented in sections 3 was used and with what settings. “4.4 Assessing metagenomic classification of DNA sequences” explains who the measurements in section 3 were used to assess the performance of Kraken2+Bracken on the different data presented in section 4.1. The last section, ”4.5 Exploring the effects of evolutionary relationships and data quality” explains how the software minhash was used to calculate evolutionary distance and how the Pearson correlation coefficient between sample quality and quantifications measured with Bray-Curtis dissimilarity was calculated.

The aim of this study is to explore whether Kraken2+Bracken can be used to accurately classify and quantify the contents of real and simulated samples of fishmeal. The order of the sections is meant to mimic a guide to reproducing this study, staring with data collection and creation, and ending with assessing the resulting classifications and quantification in addition to exploring how some relevant factors might affect the results.

4.1 Sample compositions

Table 2 shows the composition of the samples that were classified using Kraken2+Bracken. Each column is a sample, and each row denotes a species that is in the database. The numbers refer to a percentage of the contents of a sample.

Sample 1, 2, 3, 4, 5, 6 and 7 have a simulated variant and real sequence variant from ion torrent. Sample 1, 8 and 15 have a simulated variant and a real sequence variant from paired-end illumina. Sample 9, 10, 11, 12, 13 and 14 only have a simulated variant, these samples are the samples that contain some species with reads in amounts of 5%.

Sample 1 and 15 are identical and contain only *Salmo salar*, Atlantic salmon. These to samples along with sample 08 are the only samples that contain only one species. Sample 8 containing only sequences from *Gadus morhua*, known as Atlantic Cod.

Sample 1 and sample 15 have a simulated variant, two illumina variants and one ion torrent variant.

Table 2: Compositions of samples in the study. Columns denote a sample, rows denote the species. Numbers given as percentages

species	sample_01	sample_02	sample_03	sample_04	sample_05	sample_06	sample_07	sample_08	sample_09	sample_10	sample_11	sample_12	sample_13	sample_14	sample_15
Gadus_morhua	0	0	0	40	5	10	80	100	0	0	0	0	0	0	0
Melanogrammus_aeglefinus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Salmo_salar	100	80	50	30	90	80	10	0	80	50	10	5	80	5	100
Thunnus_albacares	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trachurus_trachurus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ctenopharyngodon_idella	0	0	0	0	0	0	0	0	5	0	0	5	0	0	0
Hypophthalmichthys_molitrix	0	0	0	0	0	0	0	0	5	0	0	0	0	5	0
Hypophthalmichthys_nobilis	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0
Ictalurus_punctatus	0	0	0	0	0	0	0	0	5	0	5	0	0	5	0
Sardina_pilchardus	0	0	0	0	0	0	0	0	0	40	0	5	10	60	0
Oncorhynchus_mykiss	0	20	50	30	0	0	0	0	0	0	0	0	0	0	0
Gadus_chalcogrammus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Oreochromis_niloticus	0	0	0	0	0	0	0	0	0	5	0	5	5	5	0
Mallotus_villosus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Clupea_harengus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Scomber_colias	0	0	0	0	0	0	0	0	0	0	80	80	0	20	0
Cyprinus_carpio	0	0	0	0	0	0	0	0	0	5	5	0	5	0	0
Hippoglossus_hippoglossus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gallus_gallus	0	0	0	0	5	10	10	0	0	0	0	0	0	0	0
Pollachius_virens	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Boreogadus_saida	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4.2 Software and data

4.2.1 Software

Analysis in this study were performed with R version 4.0.4 using jupyterhub (Team 2021) as a part of the Orion computing cluster at NMBU, unless otherwise explicitly stated. Kraken2 and Bracken (Wood, Lu et al. 2019) (Lu, Breitwieser et al. 2017), Art and reduction in the number of reads for simulated data were done using Orion with mobaXterm as editor. The R-packages listed in Table 3 were utilized in R version 4.0.4.

Table 3: R packages and version used.

Package	Version
Tidyverse	1.3.0
microseq	2.1.5
microclass	1.2
pals	1.7
vegan	2.5.7
readxl	1.3.1
writexl	1.4.0

Kraken2, Bracken, and Art were used through the containers found at singularity (Kurtzer, Sochat et al. 2017).

All shell-scripts run on Orion requires a header with settings for various components of the cluster. These vary somewhat and are to some extent tailored to the task the shell-script preforms. Since these setting do not affect the results of the use of Kraken2+Bracken, Art or mash, these will not be discussed in detail, but all settings are stated in the shell scripts found in the appendix and in the GitHub repository.

All scrips utilized to conduct this study are available at the following GitHub repository: [halvor-ekeland/Master: Master i bioinformatikk/anvendt statistikk (github.com)]

4.2.2 Genomes used in the study

Building a custom database used for classification of reads and simulating reads using Art requires genome assemblies to be used as a source for k-mers when classifying and as template for simulating reads. 21 genomes were selected to comprise the database used for classification with Kraken2+Bracken. These 21 genomes were downloaded from the NCBI GenBank, where assemblies for many different species can be found. Table 4 contains the names of the species whose genomes were used to create the database. The genomes in Table 4 were chosen based on a list by Orivo AS of species that were of interest to them.

The genomes were downloaded on the 6. January 2022

Table 4: Table of genomes added to the database used for classifying reads with Kraken2+Bracken.

Latin name	Common name	Taxonomic ID	NCBI accession number
Gadus morhua	Atlantic cod	8049	GCA_905250895.1
Melanogrammus aeglefinus	Haddock	8056	GCA_900291075.1
Salmo salar	Atlantic salmon	8030	GCA_000233375.4
Thunnus albacares	yellow tuna	8236	GCA_914725855.1
Trachurus trachurus	Atlantic horse mackerel	36212	GCA_905171665.2
Ctenopharyngodon idella	grass carp	7959	GCA_019924925.1
Hypophthalmichthys molitrix	silver carp	13095	GCA_004764525.1
Hypophthalmichthys nobilis	Bighead carp	7965	GCA_004193235.1
Ictalurus punctatus	channel catfish	7998	GCA_004006655.3
Sardina pilchardus	European pilchard	27697	GCA_900499035.1
Oncorhynchus mykiss	Rainbow trout	8022	GCA_013265735.3

Gadus chalcogrammus	Alaska pollock	1042646	GCA_900302575.1
Oreochromis niloticus	Nile tilapia	8128	GCA_922820385.1
Mallotus villosus	capelin	30960	GCA_903064625.1
Clupea harengus	Atlantic herring	7950	GCA_000966335.1
Scomber colias	Atlantic chub mackerel	338315	GCA_021039115.1
Cyprinus carpio	European carp	7962	GCA_018340385.1
Hippoglossus hippoglossus	Atlantic halibut	8267	GCA_009819705.1
Gallus gallus	chicken	9031	GCA_016700215.2
Pollachius virens	saithe	8060	GCA_900312635.1
Boreogadus saida	polar cod arctic cod	44932	GCA_900302515.1

4.2.3 Calculating sample quality

To study how the quality of the sequences affect the ability of kraken2 to quantify the quality of the samples, the average Phred score for all reads in a fastq file was used. First the average Phred-scores for the reads in a samples fastq file was calculated. Then the average of those scores were calculated, resulting in one quality score for each sample's fastq file. The quality score of a fastq file can therefore be described as a mean of means. This is done to have one number associated with each variation of a sample.

To convert the quality scores of a read from symbols to numbers that can then be averaged, the function "`strtoi`" in R is used. The input to this function is the quality symbols and a base, the base being the encoding 33, specifying which version of ASCii to use when converting the symbols to numbers. This creates a vector of numeric quality scores for each base in a read, the mean of these scores being the quality score for that read. To find the quality for a sample, all the read quality scores are averaged using the `mean` function. The script for this can be found in the GitHub repository for this study, link to which can be found on page 60, in section 10.1.

A bespoke function was made to select a set of random reads from the fastq file and the selected reads were used as the basis for the quality calculation of that samples fastq file. In the study, 10 000 randomly chosen reads were used for each sample.

When sequences stem from paired-end sequencing systems, the sequences are represented by R1 and R2 fastq files. There is usually a sample difference in Phred between these files, R1 having a slightly higher average Phred score than R2. For the sake of simplicity, the quality of the R1 file will be used for comparison.

4.2.4 Data simulation

This study used simulated data as a gold standard to assess the ability of kraken2 + Bracken to predict the presence of species and accurately approximate the composition of the samples. The composition of the fastq files created for each mixture is based on the assumption that there is a one-to-one relationship between the relative abundance of amount in grams of a given fish and the number of read pairs in the fastq file for each mixture. Meaning that if 50% of the weight of a sample is salmon, then 50% of the read pairs in the fastq file is from salmon. This assumption is made to simplify the setup of the study and may not reflect the reality of the metagenomic sequence of real-world samples. This approach was used by Heiminen et al. (Haiminen, Edlund et al. 2019)in their study of food authentication using metagenomic shotgun sequencing.

The reads simulated in this study mimic Illumina HiSeq 2500 (Illumina 2022), which is one of the most used sequencing systems in the world. In this study the simulated reads have a length of about 150

base pairs each. The mean length for the Art simulated reads were set to 150. Fragment size was set to 200 and the standard deviation was set to 10 base pairs. If sequencing real samples are to yield reads of a similar length, the library preparation must result in fragments of around 350 – 400 base pairs. This is based on the approach chosen by Head et al (Head, Komori et al. 2014).

Paired end sequencing was used resulting in a R1 fastq file and an R2 fastq file. Fold coverage was set to 1 as this was the lowest possible, meaning that the number of bases in the reads equal the number of bases in the genome the reads were simulated from.

To minimize the computational requirements needed to prepare, make, and run the different components of the Kraken2 + Bracken analysis, the fastq files resulting from the simulation of reads using Art was cropped to contain only 1 million reads. Cropping the number of reads was done so that the creation in the fastq files for the different simulated samples would take less time. All reads after the first 1 million in the fastq file were removed using the `sed` command in linux using the mobaXterm editor. This assumes that the order of the simulated reads in the fastq files resulting from Art are in random order.

The simulated samples were created by randomly sampling the reads resulting from Art. The fastq files for each simulated sample was made using the tidyverse package and the microseq packages for R. The function "`readFastq`" from microseq was used to load the simulated reads made from the 21 genomes into R, resulting in a data frame with 3 columns. One with the Read ID, one with the DNA sequence and a column with the Phred scores associated with the bases in the sequence. This was done for both R1 and R2 reads. One read from the fastq files is one row in the data frames, a read pair shares the same row index. Using the function "`sample`" with bounds defined by the number of reads in the genomes and the number of reads to be added to the sample, a vector of integers between 0 and the number of simulated reads was created. The integers in this vector is used by the functions "`slice`" to extract rows from the data frames containing the simulated reads.
The reads that are chosen, then get exported by using the function "`writeFastq`", which writes a fastq file with the reads from the table created from the simulated reads using the "`sample`" and "`slice`" functions. The samples made with reads simulated with Art will in the text be referred to as "the simulated samples"

4.2.5 Real DNA sequence data

For the real data the reads are accompanied by metadata stating which fastq files contains reads from what samples and what species those samples contain. The real DNA sequence data were provided by Orivo AS and came in two variants, illumnina pair-end sequences and ion torrent sequences

The illumina sequence data are a part of a study examining the effects of different DNA extraction methods. These methods referred to as CHAN and Mericon in the metadata have different focuses. CHAN is a method for extracting DNA from samples using magnetic beads with the goal to achieve high quality and purity DNA fragments for sequencing. Mericon focuses on extracting DNA from highly processed samples, such as fishmeal and uses modified cetyltrimethylammonium bromide to extract DNA. The chemistry of these extraction is not directly relevant to this study, but the different extraction methods give an opportunity to see if this could affect the classification and quantification of the sample contents. Sample 1 and Sample 8 use the CHAN method and sample 15 uses the Mericon method. This data was sequences using illumina hi-seq 2500 and is paired-end, resulting in two fastq files for each sample, the R1 file and the R2 file. Samples composed of this sequence data will in the text be referred to as "the illumina samples"

The samples sequenced using ion torrent originate from Patogen. They were made by extracting DNA and then creating samples with the compositions listed in Table 2. Ion torrent samples only have one fastq file, unlike the simulated samples and the illumina samples due to the differences in sequencing procedure described in section 3.2.1

4.3 Taxonomic classification of DNA sequences

4.3.1 Creating a database for Kraken2 and Bracken

Metagenomic classifications using kraken2 + Bracken is done by comparing k-mers made from genomes added to a database and k-mers made from the reads contained in the fastq-files that are to be classified. The database can be standard or as in the case, a custom-made database. The genomes chosen for this database were from a list provided by Orivo AS in November of 2021 and later more genomes were added so that more real samples could be run. The 21 genomes in the final database can be seen in Table 4.

The database was made using the "`kraken-build function`" and the "`Bracken-build`" function in Linux. Genomes were added using the "`—add-to-library`" function and the path for the directory where the fasta for the genomes were stored. When using the "`bracken-build`" function several options can be customized. In this study k-mer size was set to 35, and the average length of the reads where set to 100. This is less than the length chosen when using Art to simulate reads and was done since the real samples are likely to have shorter reads.

The database made for this study needs around 80 Gb of disc space. Scripts for creating the database can be found in the GitHub repository for this study, se section 10.1 at page 60.

4.3.2 Classification of reads using Kraken2+Bracken

Kraken2+Brakcen is the hash-based algorithm that was used in this study to classify the reads in the samples, both real and simulated, to the 21 selected species that the custom database was composed of (Wood, Lu et al. 2019). Kraken2+Bracken containers for Singularity was used.

The database for Kraken2+Bracken was made with the "`kraken2-build`" function and the "`bracken-build`" function in Linux using the 21 genomes found at the NCBI genbank (Coordinators 2016). Taxonomic information was downloaded using "`kraken2-build`" and specifying the option "`--download-taxonomy`" and the path of the database. Genomes were added to the database using the same function with the option "`--add-to-library`".

Kraken2+Bracken was used on the Orion computing cluster and utilized a shell script and a list of samples to be classified. Three different lists of samples were used, one for simulated reads, one for illumina pair-end reads and one for the ion-torrent reads. They were run in three different "jobs" to separate simulated from real and because pair-end reads utilize two fastq files, one with R1-reads and another with R2 reads. Sequence data from ion torrent comes in one fastq file and therefore requires a separate script.

Kraken2 was run using 10 threads on Orion for all samples. For the simulated data as well as the real sequences from illumina, the "`paired`" options were utilized and "`r1`" and "`r2`" specified as R1

reads and R2 reads respectively. For the real illumina data the "`--gizp`" setting was put to decompress as these fastq files were compressed with the suffix "gz". For real data from ion torrent, only one fastq file variable was specified.

Bracken was run with the same settings on Orion for all samples. The option "`target rank`" allows the user to specify what taxonomic rank reads are to be classified to, in this study this was set to "`s`", specifying that all reads are to be classified at species level. Bracken allows for a threshold to be imposed on the classifications made by kraken2. For additional reads classified to a higher rank by Kraken2 to be reclassified to a given species by Bracken, that species must have reads at or in excess of the threshold. The default threshold is 10, but in this study this threshold was set to 1, meaning that any read classified to a species by Kraken2 would not be discarded by Bracken.

4.4 Assessing metagenomic classification of DNA sequences

4.4.1 Purity

Purity was calculated using a bespoke function created in R based on Equation 3. The functions find true positives, the number of species that are predicted to be in the sample by Kraken2 + bracken, that are also present in the gold standard. Additionally, it calculates the number of false positives, meaning the number of species that are predicted to be present in the sample by Kraken2+Brakcen, but are not present in the gold standard.

The function for purity has a threshold that can be set by the user specifying the number of reads that must be classified to a species for that species to be counted as present. Threshold at 0 means that if a species has one read or more classified to that species, it is predicted as present. If the threshold is 1000 reads, then a species needs at least 1000 classified to it to be predicted as present. To observe how the threshold affects the purity of a sample, the threshold was set at different levels. The levels being a fraction of the total number of reads classified by Kraken2+Bracken for that sample. The fractions were: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1} and 0. If a sample has 1 million reads classified by Kraken2+Bracken, then a fraction of 10^{-2} corresponds to a threshold of 10 000 reads.

The resulting purity for the samples at the different thresholds are visualized using a line plot. The x-axis denotes the threshold as the fraction of the number of classified reads for that sample, the y-axis denotes purity from 0 to 1. Each sample has a line colored based on sample.

4.4.2 Completeness

To calculate completeness, R was utilized with a bespoke function that calculated the number of species that where in both the gold standard and the species predicted by kraken2+Bracken it be present, the number of true positives and the number of species not predicted to be in the sample, but that are present in the gold standard, the false negatives. With the number of true positives and true negatives counted, the completeness of a given sample was calculated using Equation 4.

As with the function for purity, a threshold can be specified by the user. The threshold is the number of reads that must be classified as a species for that species to count as present in the sample. This was set to the same increasing thresholds between 0 and 0.1 as for purity, for all samples.

The input for the function was a long-form table of the results of the classifications made by Kraken2+Bracken and the mixes table, consisting of read numbers stemming from the gold standard relative content of each species multiplied with the total number of reads classified by Kraken2+Bracken for that sample

The calculated completeness for the samples at the different thresholds are visualized using a line plot. The x-axis denotes the threshold as the fraction of the number of classified reads for that sample, the y-axis denotes completeness from 0 to 1. Each sample has a line colored based on sample.

4.4.3 Bray-Curtis Dissimilarity

In this stud Bray-Curtis dissimilarity was calculated using the gold standard as the benchmark and this was compared to simulated illumina read pairs, real illumine read pairs and real ion torrent reads using the function "`vegdist`" from the "vegan" r-package, with the method set to "`bray`". Specifying method as "`bray`", calculates Bray-Curtis dissimilarity using Equation 5.

The approach was based on comparing the gold standard for a sample with the simulated and real data equivalent for that sample. The number of reads in the gold standard was set to equal the number of reads classified by Kraken2+Bracken for that sample.

The results of the Bray-Curtis dissimilarities were plotted using a column plot where each sample is represented by a column. The y-axis goes from 0 to 1 and represent the dissimilarity. The heigh of each column is determined by the dissimilarity between the sample and the gold standard.

4.4.4 PCA for samples with different confidence

To study how confidence affects the ability for Kraken2+Bracken to classify the reads correctly PCA was used to cluster the different versions for each sample. Classification with Kraken2+Bracken was done at different levels of confidence ranging from 0 to 1, by steps of 0.1. This was done with the 15 simulated samples. Each sample was run though Kraken2+Bracken 11 times, each time at different confidences resulting in 11 variations of results for the 15 simulated samples that were analyzed.

The results were then CLR-transformed before PCA was used. The variations of the results for each sample were used as the samples (rows in the matrix), the number of reads for the different species were used as the traits (columns in the matrix). Standard setting was used for PCA.

The results of the PCA are plotted as a score plot with the first principal component, referred to as PC1, on the x-axis and the second principal component, PC2, on the y-axis. If the two first components capture the vast majority of variation similar samples will cluster in the score plot. To study the amount of variation explained by each principal component, a scree plot I utilized. The scree plot is a column plot where each principal component is represented by a column. The fraction of total variation is denoted on the y-axis, and the height of the columns correspond to that component's faction of total variation.

4.5 Exploring the effects of evolutionary relationships and data quality

4.5.1 Calculating mash distance using minHash

The mash distance is an estimation of the evolutionary distance measured as the mutation rate between two sequences. A smaller mutation rate means a closer evolutionary relationship than a larger one. This is indicated by a larger amount of shared k-mers and a smaller D, where D is the mash distance (Ondov, Treangen et al. 2016).

When calculating mash distance using minHash for linux, the function mash is are used with two options, "`sketch`" and "`dist`". Sketch creates a k-mer sketch for each genome, "`dist`" takes these sketches and compares them to calculate the mash distance D. The sketch size, the number of

k-mers in the sketch was set to 10000. The maximum k-mer length for minHash is 32, so k was set to k = 32.

4.5.2 Pearson correlation

To study how the quality of the data affects the ability of Kraken2+Bracken to classify reads, Pearson correlation coefficient was used. This resulted in a correlation coefficient, that represents the linear relationship between quality and Bray-Curtis Dissimilarity

In R the function "`cor`" from the preloaded "stats" package, was used with the option "method" set to "pearson". Inputs to the function were the quality of the samples calculated using the average Phred score of the reads for the sample, referred to as quality, and the Bray-Curtis Dissimilarity for each sample calculated using the gold standard.

5 Results

The results are structured in the following sections: 5.1 present the data about the samples in the study, quality of sequences belonging to those samples and the mash distance between genomes. In section 5.2 the classifications made by Kraken2+Bracken are stated and in section 5.3 the classification and quantification assessments are presented. The last section, 5.4, presents the correlations between the quality for the reads and the number of reads with the results from Kraken2+Bracken.

The tables in this section will use the category “sample”. A sample refers to what the composition the mix has. There is a sample_01 from simulated, illumine and ion torrent, all of these have the same composition, but a different source. Sample 1 – 7 have an ion torrent equivalent, samples 1, 8 and 15 have an illumina equivalent and samples 9 – 14 have only a simulated variant. While the category “source” indicates the source of the data, being either simulated, illumina or Ion torrent.

Since the sequencing systems, the number of reads and the quality of the reads differ, the results are presented separately. Results stemming from different sources will not be represented in the same graph or table unless explicitly stated in the title and/or caption.

5.1 Sample Overview

5.1.1 Simulated samples

Table 5 shows the quality and number of reads contained in the R1 fastq files for the simulated samples, the composition of which can be found in Table 2. The scores are calculated by using the method described in section 4.2.3.

Table 5: Table show average Phred score and number of reads in fastq file for samples simulated using Art.

Sample	Quality	nr. reads
Sample 01	134,15	1000000
Sample 02	134,12	1000000
Sample 03	134,13	1000000
Sample 04	134,15	1000000
Sample 05	134,13	1000000
Sample 06	134,15	1000000
Sample 07	134,15	1000000
Sample 08	134,14	1000000
Sample 09	134,13	1000000
Sample 10	134,13	1000000
Sample 11	134,14	1000000
Sample 12	134,14	1000000
Sample 13	134,14	1000000
Sample 14	134,14	1000000
Sample 15	134,15	1000000

5.1.2 Real samples sequenced using illumina

Table 6 shows the quality and number of reads contained of the R1 fastq files for the real samples sequences using paired-end illumina, the composition of which can be found in Table 2. The scores are calculated by using the method described in section 4.2.3.

Table 6: Table show average Phred score and number of reads in fastq files for real samples sequenced using paired-end sequencing by illumina

Sample	Quality	nr. reads
Sample 01	131,77	80722
Sample 08	130,06	163094
Sample 15	132,57	86472

5.1.3 Real samples sequences using ion torrent

Table 7 shows the quality and number of reads contained fastq files for the real samples sequences from ion torrent, the composition of which can be found in Table 2. The scores are calculated by using the method described in section 4.2.3.

Table 7: Table show average Phred score and number of reads in fastq files for real samples sequenced using ion torrent

Sample	Quality	nr. reads
Sample 01	106,98	3594432
Sample 02	107,03	4316287
Sample 03	107,39	5190303
Sample 04	106,83	4186946
Sample 05	107,25	4388427
Sample 06	107,23	4644081
Sample 07	107,36	5637665

5.1.4 Mash distance between the 21 genomes

Figure 4 is the distance matrix between the 21 genomes used to construct the database and simulate samples. The mash distance goes from 0 to 1. 0 indicates that the genomes are close to or identical and 1 meaning complete difference. The diagonal has only values of 0 as this is the mash distance for the same genome. In Figure 4 blue indicates values closer to 0 and yellow indicates values closer to 1. The distance matrix is symmetrical along the diagonal. *Gallus gallus* is the most evolutionary distant from the other 20 species of fish. *Gadus chalcogrammus* and *Gadus morhua* appears to be the most similar based on the deep blue.

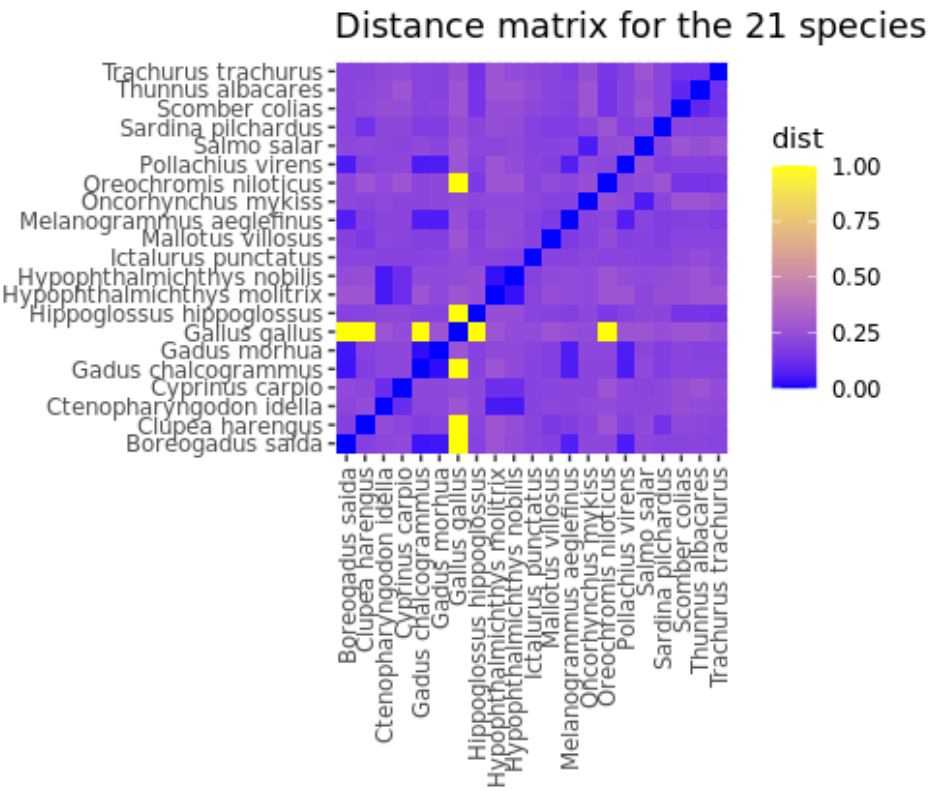

Figure 4: Distance matrix for the 21 genomes used for read simulation and database creation. color in corresponding to the distance between genomes, yellow indication larger evolutionary distance, blue indicating smaller evolutionary distance

5.2 Classification of reads using Kraken2+Bracken

5.2.1 Kraken2+Bracken classification results

The fraction of the total number of reads classified by Kraken2+Bracken to a given species is stated in the three following tables.

5.2.1.1 Kraken2+Bracken results for simulated samples

Table 8: Classification results for simulated samples stated fraction of total reads classified for a given sample. Columns refers to sample, rows to species.

sample	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6	sample 7	sample 8	sample 9	sample 10	sample 11	sample 12	sample 13	sample 14	sample 15
Gadus morhua	0,00018	0,00020	0,00022	0,37201	0,00462	0,00921	0,74613	0,93858	0,00019	0,00015	0,00005	0,00006	0,00017	0,00011	0,00018
Melanogrammus aeglefinus	0,00031	0,00034	0,00037	0,00494	0,00080	0,00138	0,00960	0,01198	0,00030	0,00026	0,00016	0,00013	0,00026	0,00022	0,00031
Salmo salar	0,96703	0,77873	0,49451	0,29939	0,87097	0,74777	0,10057	0,00155	0,78782	0,48211	0,09633	0,04817	0,77267	0,04925	0,96703
Thunnus albacares	0,00027	0,00033	0,00038	0,00062	0,00031	0,00033	0,00080	0,00097	0,00027	0,00024	0,00116	0,00115	0,00025	0,00046	0,00027

Trachurus trachurus	0,00027	0,0031	0,0040	0,0059	0,0026	0,0027	0,0077	0,0094	0,0028	0,0027	0,0036	0,0033	0,0024	0,0032	0,0027
Ctenopharyngodon idella	0,00032	0,0030	0,0027	0,0049	0,0033	0,0036	0,0070	0,0084	0,12892	0,0031	0,0084	0,05024	0,0032	0,03700	0,0032
Hypophthalmichthys molitrix	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
Hypophthalmichthys nobilis	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
Ictalurus punctatus	0,00035	0,0034	0,0036	0,0040	0,0038	0,0039	0,0046	0,0048	0,05108	0,0024	0,04973	0,0038	0,0029	0,05026	0,0035
Sardina pilchardus	0,00021	0,0024	0,0030	0,0043	0,0021	0,0024	0,0050	0,0060	0,0020	0,39962	0,0015	0,04983	0,10030	0,60369	0,0021
Oncorhynchus mykiss	0,02930	0,21725	0,50084	0,30294	0,02640	0,02356	0,00399	0,00130	0,02399	0,01476	0,00314	0,00170	0,02338	0,00175	0,00293
Gadus chalcogrammus	0,00011	0,0012	0,0015	0,00917	0,00123	0,00236	0,01817	0,02279	0,00011	0,00010	0,00011	0,00009	0,00011	0,00111	0,00011
Oreochromis niloticus	0,00019	0,0021	0,0026	0,0044	0,0021	0,0022	0,0060	0,0074	0,0026	0,05064	0,0034	0,05037	0,05073	0,05102	0,0019
Mallotus villosus	0,00013	0,0016	0,0021	0,0027	0,0014	0,0014	0,0037	0,0043	0,0015	0,0012	0,0006	0,0006	0,0013	0,0010	0,0013
Clupea harengus	0,00027	0,0029	0,0035	0,0055	0,0027	0,0032	0,0072	0,0089	0,0027	0,0044	0,0013	0,0013	0,0027	0,0052	0,0027
Scomber colias	0,00026	0,0031	0,0038	0,0053	0,0030	0,0030	0,0074	0,0088	0,0027	0,0026	0,79727	0,79640	0,0023	0,02215	0,0026
Cyprinus carpio	0,00029	0,0031	0,0036	0,0060	0,0031	0,0032	0,0076	0,0091	0,05366	0,05005	0,04974	0,00555	0,05016	0,02655	0,0029
Hippoglossus hippoglossus	0,00021	0,0023	0,0025	0,0038	0,0021	0,0022	0,0048	0,0058	0,0021	0,0017	0,0028	0,0024	0,0019	0,0018	0,0021
Gallus gallus	0,00005	0,00006	0,00006	0,00010	0,05048	0,10091	0,10236	0,00015	0,00007	0,00005	0,00005	0,00005	0,00006	0,00006	0,00005
Pollachius virens	0,00013	0,0013	0,0017	0,0254	0,0040	0,0076	0,0510	0,0630	0,0013	0,0011	0,0006	0,0005	0,0012	0,0009	0,0013
Boreogadus saida	0,00010	0,0011	0,0016	0,0360	0,0052	0,0098	0,0717	0,0907	0,0011	0,0008	0,0005	0,0005	0,0010	0,0006	0,0010

5.2.1.2 Kraken2+Bracken results for samples sequenced using illumina

Table 9: Classification results for samples sequenced using illumina stated fraction of total reads classified for a given sample columns revers to sample, rows to species.

sample	sample 1	sample 8	sample 15
Gadus morhua	0.00041	0.00489	0.00030
Melanogrammus aeglefinus	0.00079	0.00619	0.00057
Salmo salar	0.68730	0.00494	0.65343
Thunnus albacares	0.00051	0.00261	0.00061
Trachurus trachurus	0.00042	0.00286	0.00050
Ctenopharyngodon idella	0.00041	0.00189	0.00039
Hypophthalmichthys molitrix	0.00000	0.00000	0.00000
Hypophthalmichthys nobilis	0.00000	0.00000	0.00000
Ictalurus punctatus	0.00056	0.00173	0.00049
Sardina pilchardus	0.00775	0.03303	0.00995
Oncorhynchus mykiss	0.10732	0.00362	0.07134
Gadus chalcogrammus	0.07596	0.48895	0.16561
Oreochromis niloticus	0.00033	0.00237	0.00019
Mallotus villosus	0.00036	0.00189	0.00025
Clupea harengus	0.00268	0.16900	0.00351
Scomber colias	0.00084	0.00301	0.00090
Cyprinus carpio	0.00055	0.00266	0.00049
Hippoglossus hippoglossus	0.00033	0.00594	0.00038
Gallus gallus	0.00001	0.00006	0.00002
Pollachius virens	0.00052	0.03719	0.01299
Boreogadus saida	0.11280	0.22710	0.07797

5.2.1.3 Kraken2+Bracken results for samples sequenced using ion torrent

Table 10: Classification results for samples sequenced using ion torrent stated fraction of total reads classified for a given sample columns revers to sample, rows to species.

sample	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6	sample 7
Gadus morhua	0.00014	0.00013	0.00013	0.24229	0.03229	0.05982	0.60621
Melanogrammus aeglefinus	0.00053	0.00051	0.00049	0.00908	0.00156	0.00249	0.02213
Salmo salar	0.95212	0.76616	0.48083	0.32884	0.88202	0.81288	0.13432
Thunnus albacares	0.00016	0.00014	0.00017	0.00017	0.00013	0.00013	0.00018
Trachurus trachurus	0.00020	0.00020	0.00023	0.00020	0.00016	0.00018	0.00024
Ctenopharyngodon idella	0.00016	0.00013	0.00012	0.00017	0.00015	0.00015	0.00029
Hypophthalmichthys molitrix	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Hypophthalmichthys nobilis	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Ictalurus punctatus	0.00030	0.00027	0.00021	0.00018	0.00027	0.00025	0.00013
Sardina pilchardus	0.00015	0.00013	0.00012	0.00015	0.00013	0.00014	0.00023
Oncorhynchus mykiss	0.04516	0.23134	0.51677	0.36283	0.04069	0.03791	0.00660
Gadus chalcogrammus	0.00004	0.00003	0.00003	0.04196	0.00545	0.01030	0.10416

Oreochromis niloticus	0.00013	0.00012	0.00012	0.00012	0.00011	0.00011	0.00013
Mallotus villosus	0.00019	0.00020	0.00017	0.00018	0.00018	0.00019	0.00022
Clupea harengus	0.00010	0.00008	0.00008	0.00010	0.00008	0.00010	0.00014
Scomber colias	0.00011	0.00010	0.00010	0.00022	0.00011	0.00012	0.00031
Cyprinus carpio	0.00027	0.00022	0.00023	0.00024	0.00023	0.00024	0.00021
Hippoglossus hippoglossus	0.00013	0.00013	0.00011	0.00042	0.00012	0.00011	0.00010
Gallus gallus	0.00002	0.00002	0.00002	0.00011	0.03458	0.07178	0.09297
Pollachius virens	0.00004	0.00004	0.00004	0.00331	0.00045	0.00082	0.00805
Boreogadus saida	0.00004	0.00004	0.00004	0.00942	0.00127	0.00228	0.02338

5.2.2 Classification rate

The classification rate refers to the fraction of reads that are classified as one for the 21 species in the database by Kraken2+Bracken. With the settings used in this study, stated in section 4.3.2 Classification of reads using Kraken2+Bracken, only the reads not classified by Kraken2 are discarded, no reads are discarded by Bracken.

5.2.2.1 Classification rate for simulated samples

Table 11Table 12 states the number of reads in the samples fastq files, the number of reads classified and the classification rate for the simulated samples. The 15 simulated samples have classification rates ranging from 96.89% to 99.69% of the reads. Sample 12 has the highest classification rate, while sample 15 has the lowest classification rate.

Table 11: Table showing the number of reads in each sample, the number of classified reads for that sample and the classification rate for samples simulated using Art.

Sample	nr. reads	n reads classified	classification rate
Sample 01	1000000	987520	98,75 %
Sample 02	1000000	984846	98,48 %
Sample 03	1000000	980958	98,10 %
Sample 04	1000000	976241	97,62 %
Sample 05	1000000	987086	98,71 %
Sample 06	1000000	986563	98,66 %
Sample 07	1000000	973679	97,37 %
Sample 08	1000000	968882	96,89 %
Sample 09	1000000	970246	97,02 %
Sample 10	1000000	991120	99,11 %
Sample 11	1000000	995832	99,58 %
Sample 12	1000000	996884	99,69 %
Sample 13	1000000	989059	98,91 %
Sample 14	1000000	983921	98,39 %
Sample 15	1000000	987520	98,75 %

5.2.2.2 Classification rate for samples sequenced using illumina

Table 12 states the number of reads in the samples fastq files, the number of reads classified and the classification rate for the samples based on real sequenced using illumina. The 3 illumina samples have classification rates around 96%, and are very similar for all three samples. Sample 1 has the

highest with 98.91% of the reads classified, while sample 15 has the lowest classification rate, with 96.81%.

Table 12: Table showing the number of reads in the sample, the number of classified reads for that sample and the classification rate for samples sequenced using paired-end illumina

Sample	nr. reads	n reads classified	classification rate
Sample 01	80722	78224	96,91 %
Sample 08	163094	157987	96,87 %
Sample 15	86472	83710	96,81 %

5.2.2.3 Classification rate for samples sequenced using ion torrent

Table 13 states the number of reads in the samples fastq files, the number of reads classified and the classification rate for the samples based on real sequenced using ion torrent. The 7 ion torrent samples have classification rates ranging from 92.35% and 93.69%. Sample 1 having the lowest classification rate and sample 3 having the highest.

Table 13: Table showing the number of reads in the sample, the number of classified reads for that sample and the classification rate for samples sequenced using ion torrent

Sample	nr. reads	n reads classified	classification rate
Sample 01	3594432	3319475	92,35 %
Sample 02	4316287	4009584	92,89 %
Sample 03	5190303	4862858	93,69 %
Sample 04	4186946	3892348	92,96 %
Sample 05	4388427	4086773	93,13 %
Sample 06	4644081	4316599	92,95 %
Sample 07	5637665	5258177	93,27 %

5.3 Assessing the classifications made by kraken2+Bracken

This section presents the measurements of purity, completeness and Bray-Curtis dissimilarity. These are used to assess the ability of Kraken2+Bracken to classify and quantify the contents that the samples comprise of. How completeness and purity is calculated is stated in section 0 and 3.4.2, with Equation 4 and Equation 3, and how these measures are implemented is explained in section 4.4.1 and 4.4.2

The Bray-Curtis dissimilarity is used to assess the ability to quantify the contents of the samples in comparison to a gold standard. How the Bray-Curtis dissimilarity is calculated is stated in Equation 5 found in section 3.4.3.

5.3.1 Purity

Purity is a measure of the fraction of species predicted to be present by Kraken2+Bracken, that are present in the sample. Purity is calculated using Equation 3 and utilized in the way laid out in section 4.4.1. Purity is a number from 0 to 1, 0 indicating that no of the species predicted to be present are in the sample, and 1 indicates that all species predicted to be present in the sample are present in the sample.

5.3.1.1 *Purity for simulated samples with varying thresholds*

Figure 5 show the purity for the 15 simulated samples as the threshold is changed. All samples start with a purtiy ranging from 0 to 2.6 for a threshold of 0 and all samples accheave a threshold of 1 with the threshold set to a fraction of 0.1 of total number of reads classified.

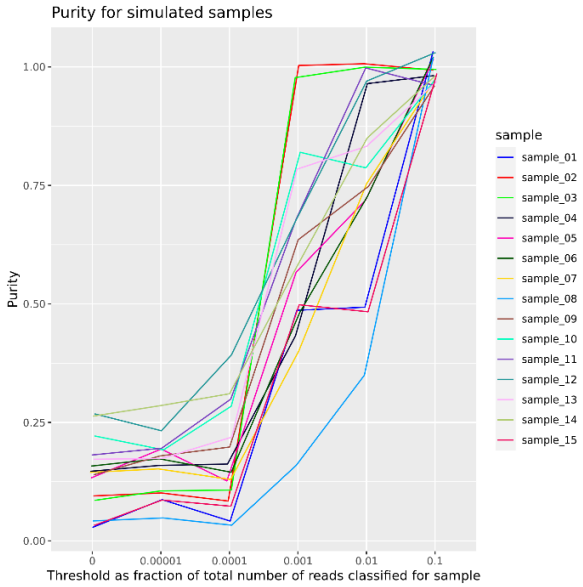

Figure 5: Purity for the 15 simulated samples at different threshold. The threshold was set as different factions of the total number of reads, increasing from 0 to 0.1. X-axis denotes the threshold as a fraction of the total number of reads classified for each sample, the y-axis denotes the purity and the legend states what samples correspond with what color.

5.3.1.2 *Purity for samples sequenced using illumina*

Figure 6 shows the purity for the 3 samples sequenced using paired-end illumina sequencing. All samples start with a purity of 0.052 and the maximum purity reached by any sample is 0.5. the lowest is 0 indicating that there are no true positives.

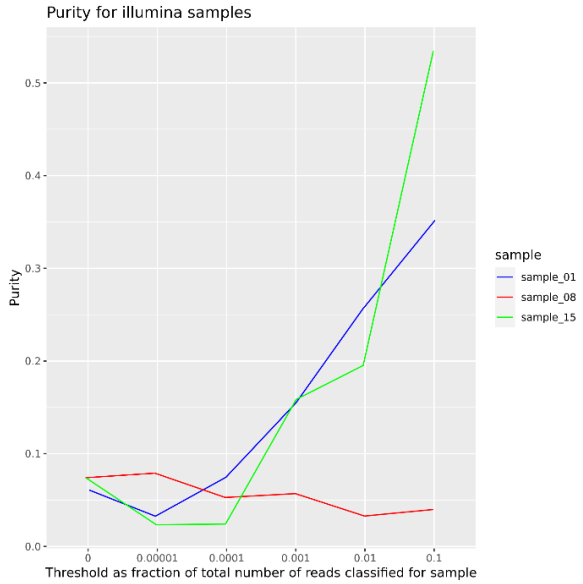

Figure 6: Purity for the 3 samples sequenced using paired-end illumina sequencing at different threshold. The threshold was set as different factions of the total number of reads, increasing from 0 to 0.1. X-axis denotes the threshold as a fraction of the total number of reads classified for each sample, the y-axis denotes the purity and the legend states what samples correspond with what color.

5.3.1.3 *Purity for samples sequenced using ion torrent*

Figure 7 show the purity for the 7 samples sequenced using ion torrent. At threshold 0 the purity ranges from 0.05 to 0.16, and at a threshold of 0.1 all samples except sample 7 reach a purity of 1. Sample 7 reaches a purity of 0.75.

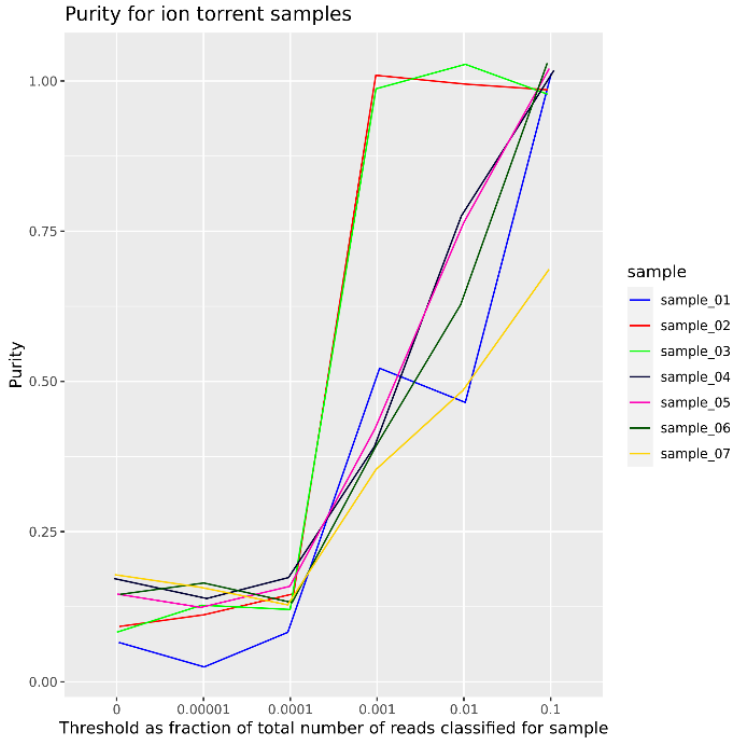

Figure 7: Purity for the 7 samples sequenced using ion torrent sequencing at different threshold. The threshold was set as different factions of the total number of reads, increasing from 0 to 0.1. X-axis denotes the threshold as a fraction of the total number of reads classified for each sample, the y-axis denotes the purity, and the legend states what samples correspond with what color.

5.3.2 Completeness

Completeness is a measure of the fraction of the species present in the sample are predicted to be present by Kraken2+Bracken. Completeness is calculated using Equation 4 found in section 3.4.2. Completeness is a number from 0 to 1, where 0 indicates that none of the species present in the sample are predicted to be present in the sample and 1 indicates that all of the species present in the sample are predicted to be present.

For the plots visualizing completeness, a jitter was introduced to the lines so that it would be easier to see the results for all samples. The small variations in the lines, stem from the jitter, not the measured completeness for a given threshold.

5.3.2.1 *Completeness for simulated samples*

Figure 8 show the completeness for the 15 simulated samples at different thresholds, ranging from 0 reads to a fraction of 0.1 of total number of classified reads for a given sample. The completeness of the samples ranges from 0.6 to 1 when the threshold is 0, and 0.2 and 1 when threshold is 0.1. Sample 12 has the biggest drop in completeness as the threshold increases, while sample 1 to 5 see no change in completeness as the threshold increases.

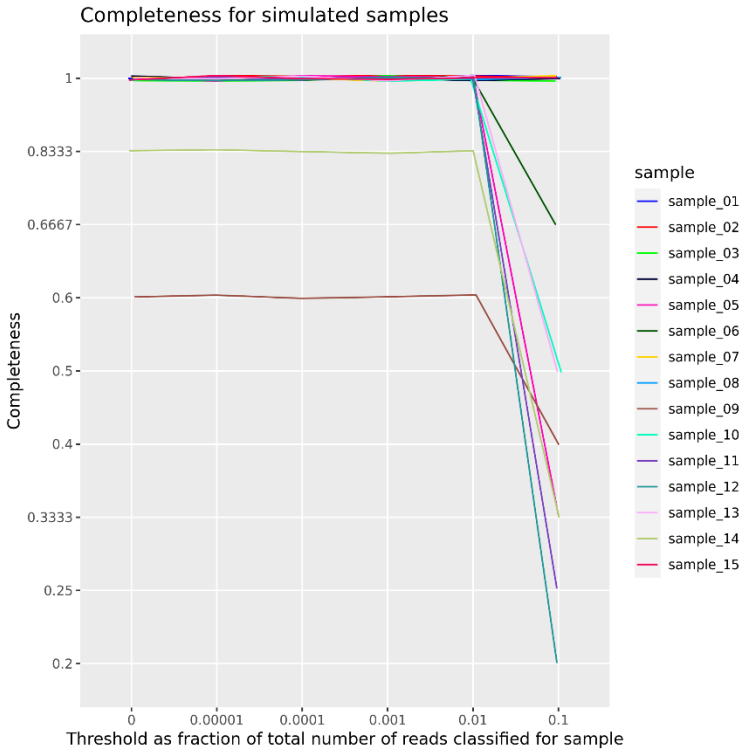

Figure 8: Completeness for the 15 samples made from reads simulated with Art. The x-axis denotes the threshold as a fraction of total number classified for a given sample. the y-axis denotes the completeness of the sample, and the legend states what color refers to what sample.

5.3.2.2 Completeness for samples sequenced using illumina

Figure 9 shows the completeness for samples sequenced using illumina sequencing. All samples start with a completeness of 1 with a threshold of 0 reads. As the threshold increases, all samples decrease in completeness, sample 8 at a threshold of 0.01 and sample 1 and 15 at a threshold of 0.1

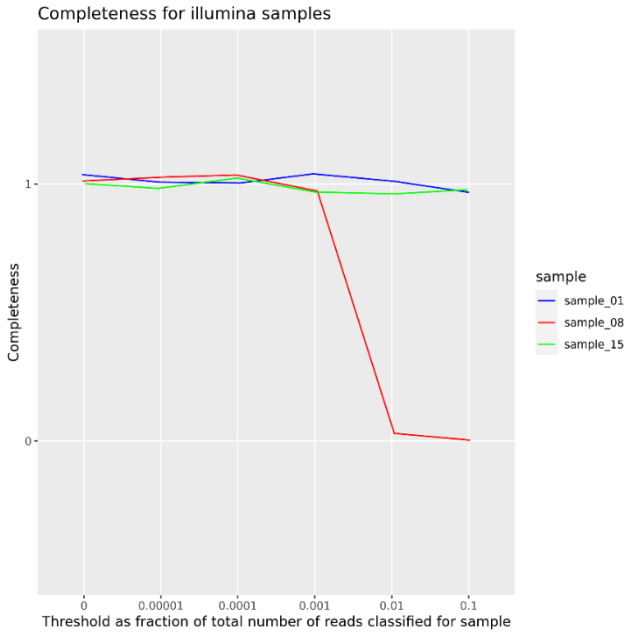

Figure 9: Completeness for the 3 samples sequenced using illumina paired-end sequencing. The x-axis denotes the threshold as a fraction of total number classified for a given sample. the y-axis denotes the completeness of the sample, and the legend states what color refers to what sample.

5.3.2.3 *Completeness for samples sequenced using ion torrent*

Figure 10 show the completeness for samples sequenced using ion torrent. All samples have a completeness of 1 at a threshold of 0, and as the threshold increases the completeness is stable for all samples except for sample 5, 6 and 7. At threshold of 0.1 sample 5 and have a completeness of 0.33 and sample 7 has a completeness of 0.66

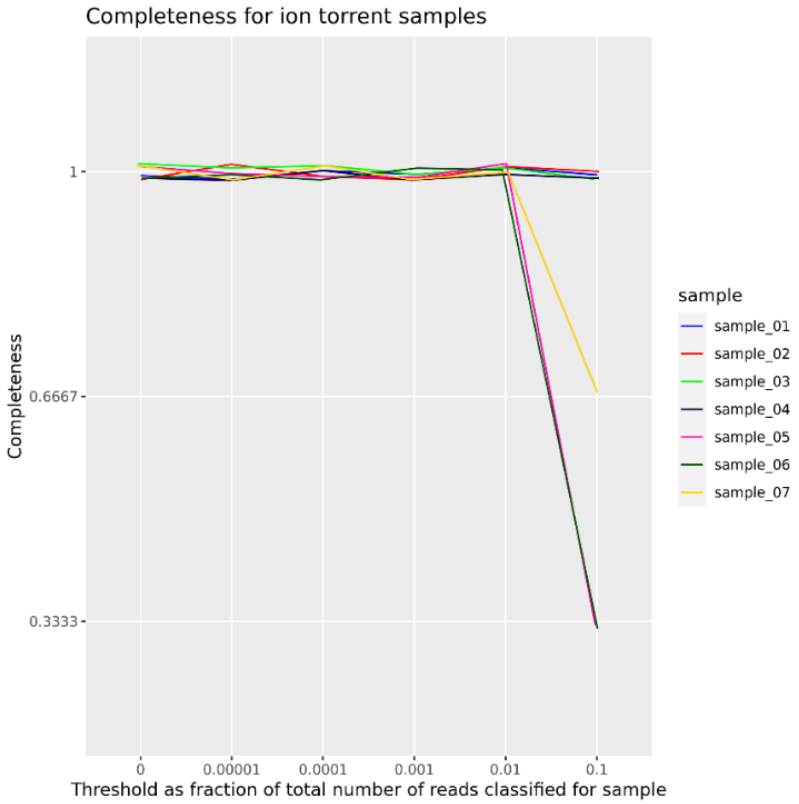

Figure 10: Completeness for the 7 samples sequenced using ion torrent sequencing. The x-axis denotes the threshold as a fraction of total number classified for a given sample. the y-axis denotes the completeness of the sample, and the legend states what color refers to what sample.

5.3.3 Bray-Curtis dissimilarity

Bray-Curtis dissimilarity is a measure of how similar or dissimilar two samples are. This is done using a number from 0 to 1, 0 being completely different and 1 being completely identical. The equation for calculating Bray-Curtis dissimilarity is stated in Equation 5. For all the samples, both simulated and real, the gold standard is used as the basis for comparison. Sample 1 is compared with the gold standard for sample 1, and so on.

5.3.3.1 *Bray-Curtis dissimilarity for the simulated samples*

The Bray-Curtis dissimilarity for the 15 simulated samples when compared with the gold standards are visualized in Figure 11. Each sample is represented by one colored column in the plot. Sample 9 has the highest dissimilarity when compared to the gold standard, calculated to be 0.11, while sample 3 has the lowest.

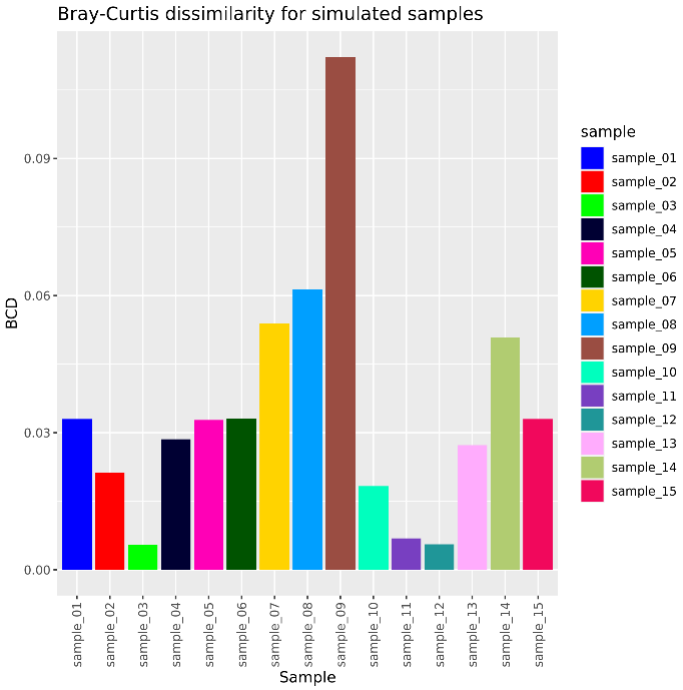

Figure 11: Bray-Curtis dissimilarity for the 15 simulated samples. Each column represents a sample, the high of the column corresponds to the Bray-Curtis dissimilarity for that sample when compared to the gold standard, the y-axis denotes the Bray-Curtis dissimilarity. The legend states what color corresponds with what sample.

5.3.3.2 Bray-Curtis dissimilarity for the samples sequenced using illumina

Figure 12 show the Bray-Curtis dissimilarity for the 3 samples sequenced using illumina pair-end sequencing. Each column represents one sample. The lowest dissimilarity is 0.312, measured for sample 1, while sample 8 has the highest dissimilarity compared with the gold standard, with a dissimilarity of 0.99.

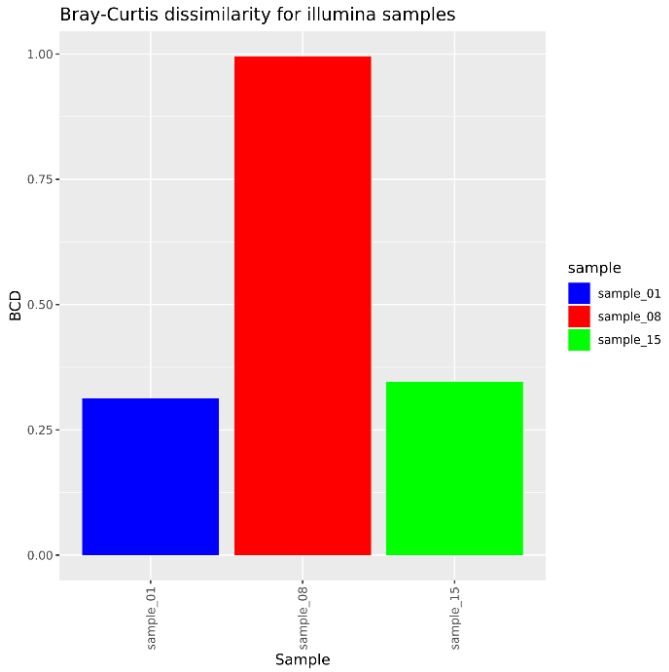

Figure 12: Bray-Curtis dissimilarity for the 3 samples sequenced using illumina paired-end sequencing. Each column represents a sample, the high of the column corresponds to the Bray-Curtis dissimilarity for that sample when compared to the gold standard, the y-axis denotes the Bray-Curtis dissimilarity. The legend states what color corresponds with what sample.

5.3.3.3 *Bray-Curtis dissimilarity for the samples sequenced using ion torrent*

The Bray-Curtis dissimilarity for the 7 samples sequenced using ion torrent are shown in Figure 13. Each sample is represented a column, the height of which corresponds to the Dissimilarity of that sample when compared to the gold standard. The dissimilarities for the 7 samples vary between 0.02 and 0.2, sample 3 having the lowest and sample 7 having the highest.

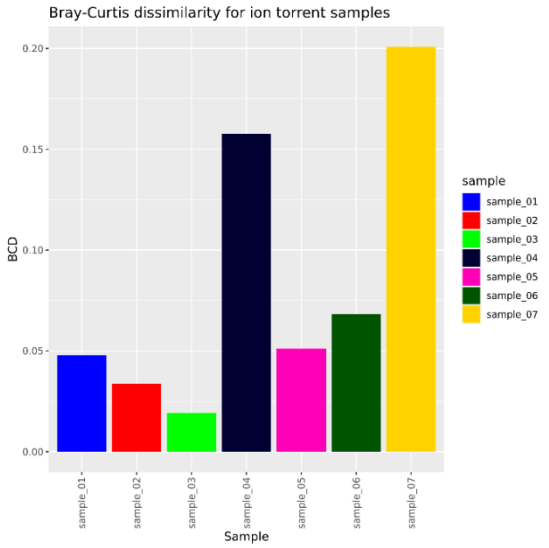

Figure 13: Bray-Curtis dissimilarity for the 3 samples sequenced using ion torrent sequencing. Each column represents a sample, the high of the column corresponds to the Bray-Curtis dissimilarity for that sample when compared to the gold standard, the y-axis denotes the Bray-Curtis dissimilarity. The legend states what color corresponds with what sample.

Table 14: Bray-Curtis dissimilarity for all samples from all three sources in the study. column name indicates what source for the samples with the dissimilarity stated in the cells. The table shows the Bray-Curtis dissimilarity for all samples analyzed in the study. The number in the cell represents the dissimilarity when compared with the gold standard, source for the sample is given in the column hears of column 2,3, and 4 counted form the left.

Table 14: Bray-Curtis dissimilarity for all samples from all three sources in the study. column name indicates what source for the samples with the dissimilarity stated in the cells.

Sample	Simulated samples	illumina samples	Ion torrent samples
sample 01	0,033	0,313	0,048
sample 02	0,021		0,034
sample 03	0,005		0,019
sample 04	0,029		0,158
sample 05	0,033		0,051
sample 06	0,033		0,068
sample 07	0,054		0,201
sample 08	0,061	0,995	
sample 09	0,112		
sample 10	0,018		
sample 11	0,007		
sample 12	0,006		
sample 13	0,027		
sample 14	0,051		
sample 15	0,033	0,347	

5.3.4 Effect of changing settings when using Kraken2+Bracken

In the section above, the purity measurement utilized different thresholds to see how classification accuracy would be affected, but more stringent requirements for counting a species as present can also be imposed directly when running Kraken2+Bracken.

The confidence option for Kraken2 was utilized to test how a stricter confidence level would affect the classifications. This was done as stated in section 4.3.2 on the 15 samples of simulated reads. results from a selection of the PCA analyses for the 15 samples are presented here. These are sample 1, 2, 9, 12, 14, 15 and were chosen because the clearly illustrate some aspects relevant to the data, the measurements or the software utilized in this study.

5.3.4.1 Classification rate

The classification rate is the number of the reads in a samples fastq files that are classified to a species. In Figure 14 the effect that changing the confidence option for Kraken2 has on the number of reads that get classified by Kraken2. This is visualized as a line plot where each sample is represented by a colored line.

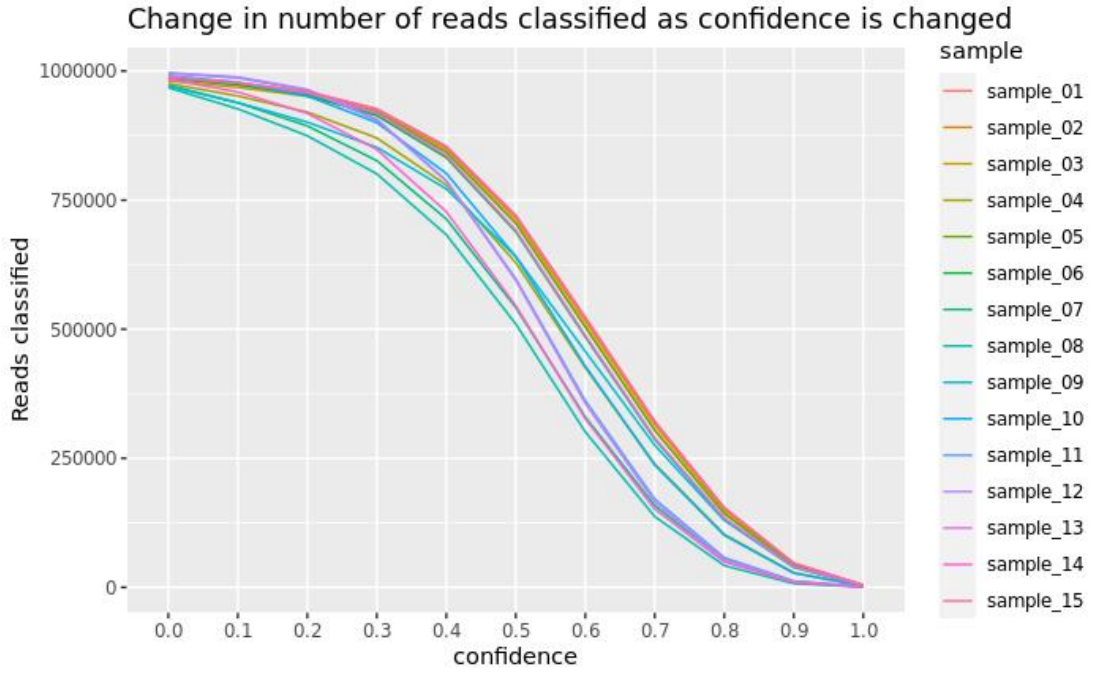

Figure 14: Effects of changing the confidence option on the number of reads classified by Kraken2 for the 15 samples simulated with Art. The x-axis denotes the input used for the confidence option for Kraken2, the y-axis denotes the number of reads classified and the legend gives an overview of what color is assisted with what sample.

5.3.4.2 Purity

The purity for the 15 simulated samples with different configuration of the confidence option for Kraken2 is shown in Figure 15. Each line represents a sample. The effect of changing the confidence can be seen as there is an increase in purity for the samples as confidence improves. The purity at confidence set to 0 ranges from 0.058 to 0.29 and at confidence set to 1, all samples have a purity of 1.

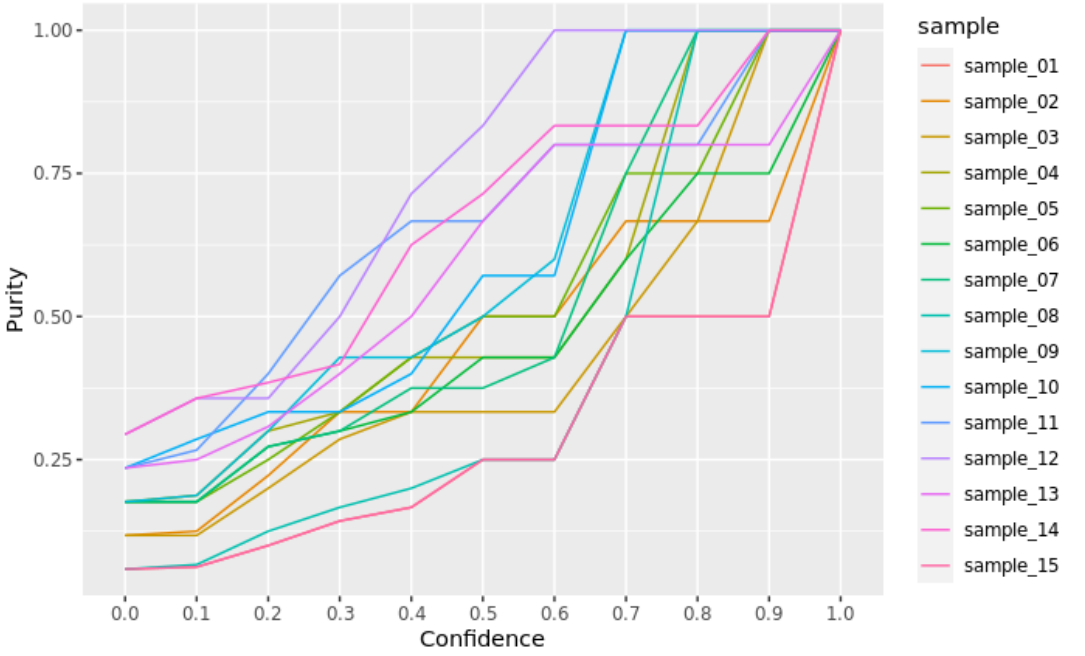

Figure 15: The effects of changing the confidence option for Kraken2 on purity for the 15 samples simulated with Art. The x-axis denotes the input used for the confidence option for Kraken2, the y-axis denotes the purity of the simulated samples at a given confidence, and the legend gives an overview of what color is assisted with what sample.

5.3.4.3 Completeness

Figure 16 visualizes the completeness of the 15 simulated samples when the confidence setting for Kraken2 is changed from 0 to 1, with increments of 0.1. all samples se no change in their confidence as the confidence is changed. Only two samples do note achieve a completeness of 1, that being 9 and 14, at 0.6 and 0.833 respectively

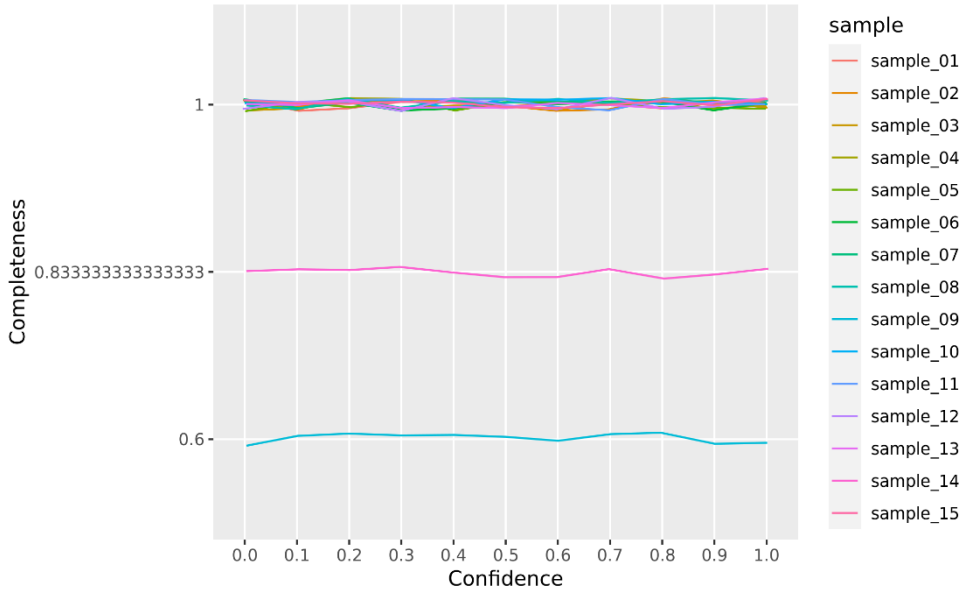

Figure 16: Completeness for the 15 simulated samples at different confidence settings. X-axis denotes the confidence setting; y-axis denotes the completeness of the samples. The legend states what sample a given colored line represents.

5.3.4.4 PCA

The PCA for the samples visualized in the section was performed on the number of reads classified to a given species by Kraken2+Bracken after centered log ratio transformations had been performed to

account for the compositional nature of the data. The samples selected to be presented here is 1, 2, 9, 12, 14 and 15. Sample 1 and 15 have identical composition. Sample 9 and 14 appear to represent cases where the classifications appear to have been somewhat unsuccessful compared to the other samples. Sample 2 and 12 appears to be the samples where the classifications were the most successful.

The following plots (Figure 17, Figure 18, Figure 19, Figure 20, Figure 21 and Figure 22) show the Score plots resulting from the PCA of the after centered log ratio transformed classification results from Kraken2+Bracken for the different confidences, from 0 to 1. The points have been labeled with the confidence and the gold standard is added to visualize the true values compared to the values stemming from classification by Kraken2+Bracken

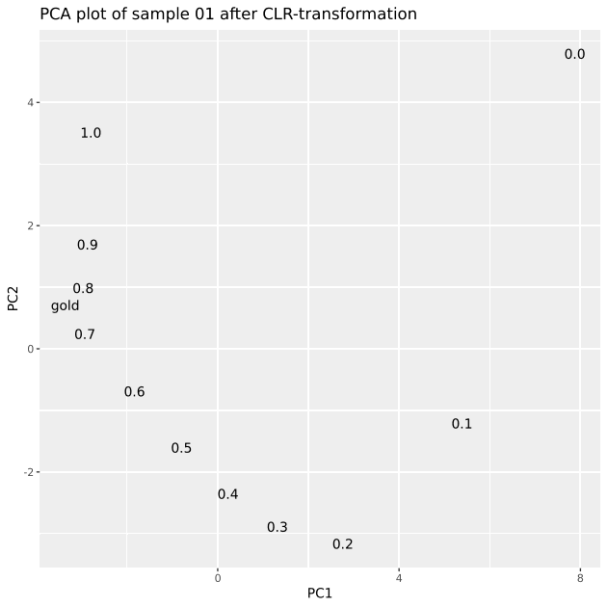

Figure 17: Score-plot for PCA of simulated sample 1. Principal component 1 (PC1) denoted on the x-axis, Principal component 2 (PC2) denoted on the y-axis. Each observation as a point labeled with the confidence setting used on Kraken2.

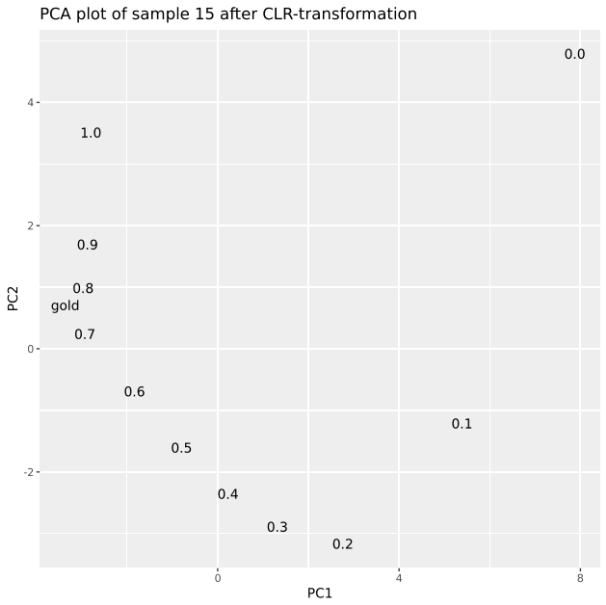

Figure 18: Score -plot for PCA of simulated sample 15. Principal component 1 (PC1) denoted on the x-axis, Principal component 2 (PC2) denoted on the y-axis. Each observation as a point labeled with the confidence setting used on Kraken2.

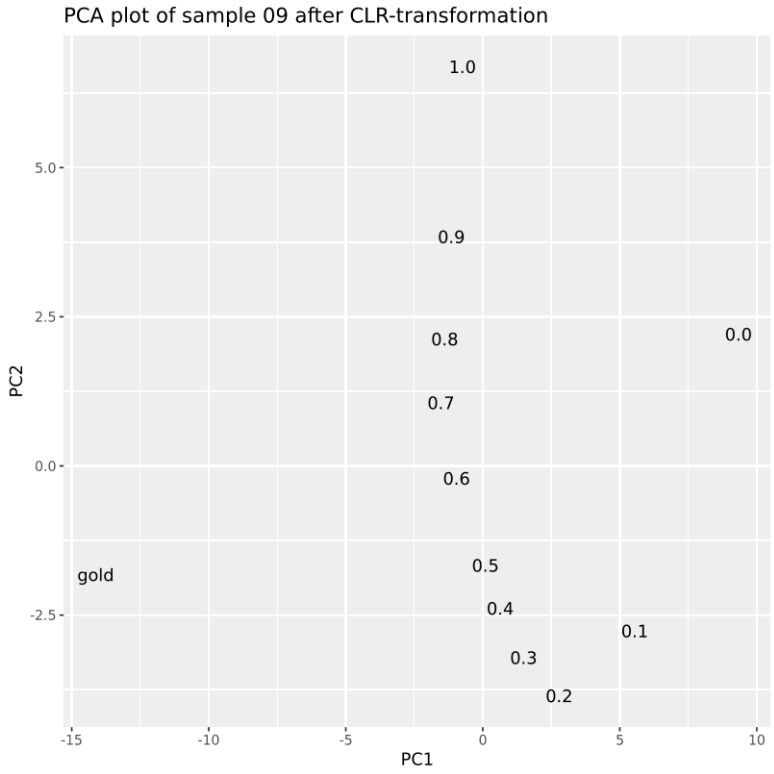

Figure 19: Score -plot for PCA of simulated sample 9. Principal component 1 (PC1) denoted on the x-axis, Principal component 2 (PC2) denoted on the y-axis. Each observation as a point labeled with the confidence setting used on Kraken2.

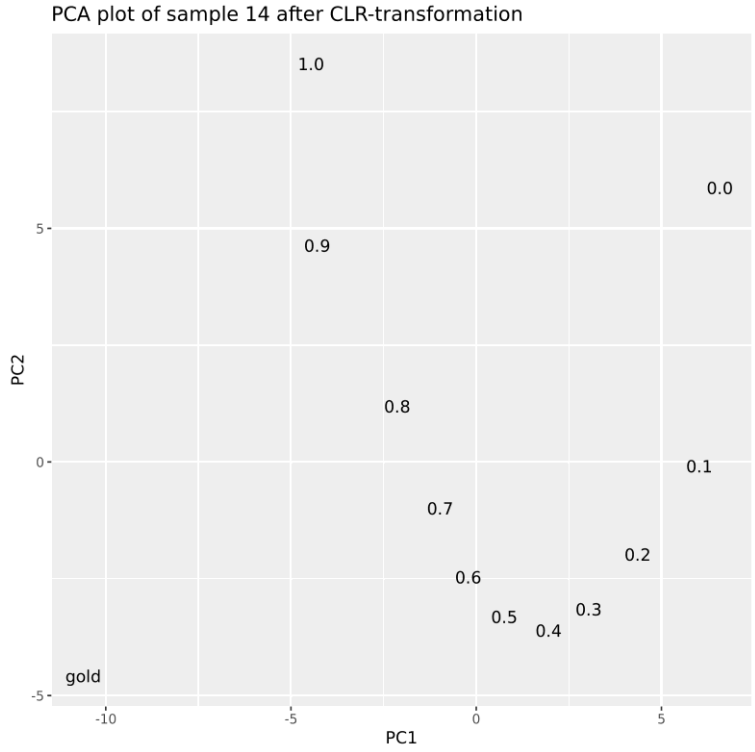

Figure 20: Score -plot for PCA of simulated sample 14. Principal component 1 (PC1) denoted on the x-axis, Principal component 2 (PC2) denoted on the y-axis. Each observation as a point labeled with the confidence setting used on Kraken2.

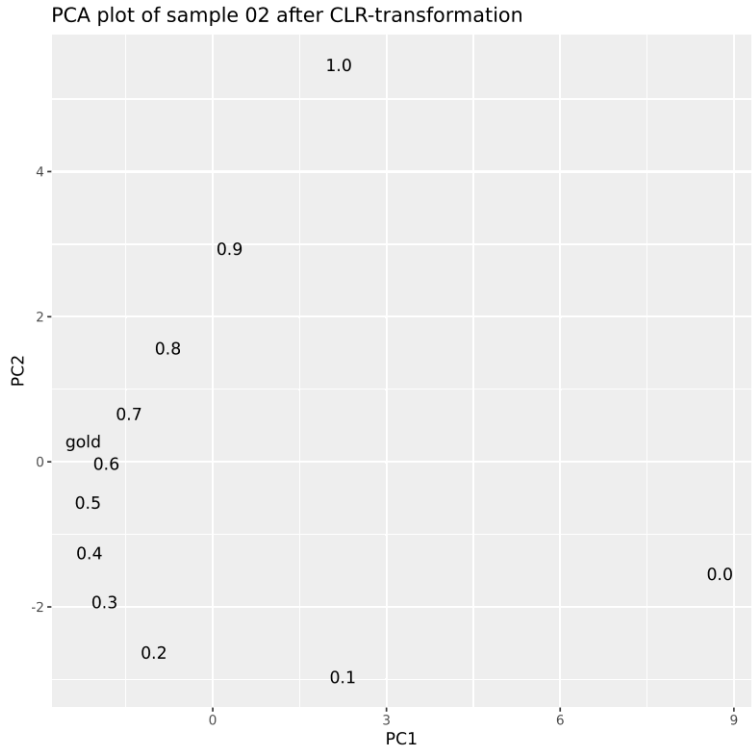

Figure 21: Score -plot for PCA of simulated sample 2. Principal component 1 (PC1) denoted on the x-axis, Principal component 2 (PC2) denoted on the y-axis. Each observation as a point labeled with the confidence setting used on Kraken2.

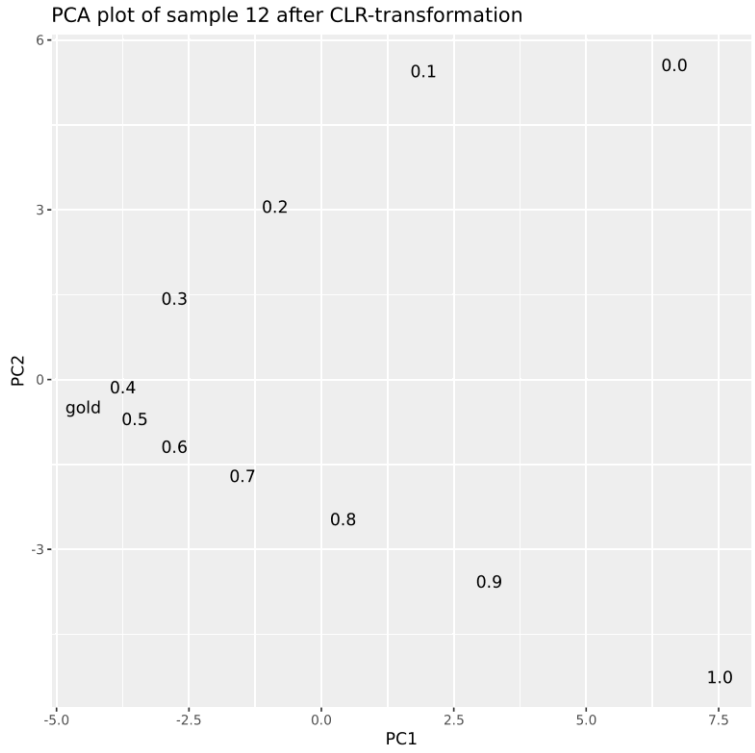

Figure 22: Score -plot for PCA of simulated sample 12. Principal component 1 (PC1) denoted on the x-axis, Principal component 2 (PC2) denoted on the y-axis. Each observation as a point labeled with the confidence setting used on Kraken2.

5.4 Other results

5.4.1 Correlation between quality and quantification

To see if there is a relationship between the quality of the reads and the ability of Kraken2+Bracken to quantify the contents, the Pearson correlation coefficient was calculated. The quality of the samples were calculated as explained in section 4.2.3 and the quantification was measured using Bray-Curtis dissimilarity as presented in section 4.4.3. Using Equation 12, the Pearson coefficient was calculated to be 0.0123.

5.4.2 Effects of random sampling when calculating Phred quality

Figure 23 shows the effects of sampling a random set of reads when calculating the Phred score for a sample. The lines represent the two approaches, random and non-random sampling of reads. From the figure the effect of read quantity when calculating quality for a sample can be discerned. The plot is based reads from sample 1 of the reads sequenced using illumina sequencing.

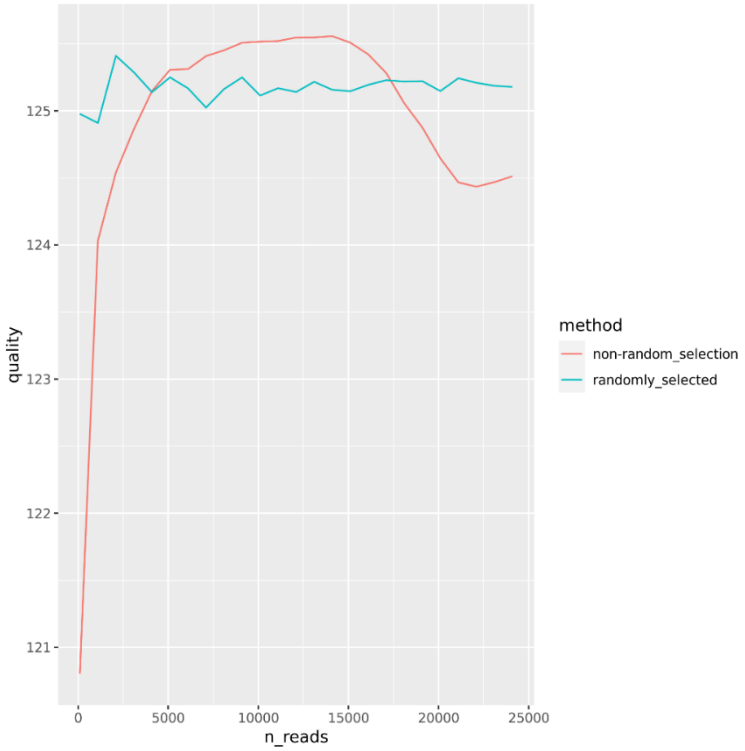

Figure 23: The effect of random sampling when calculating the sample quality based on the average Phred scores of the reads in the sample. x-axis denotes number of reads, y-axis denotes the quality for the sample and the lines represent the different approaches to sampling, random and non-random.

6 Discussion

The aim of the study was to explore if Kraken2+Bracken could accurately and reliably classify and quantify the contents of fishmeal samples based on their metagenomic DNA sequence. This was done by taxonomically classifying reads to species in a custom database built with relevant genomes from the NCBI Genbank. The assessment of the classification would utilize the measurements presented in by Meyer et al. in their article titled "Assessing taxonomic metagenome profilers with OPAL". These measurements were Purity and completeness to assess the ability to classify and Bray-Curtis dissimilarity to assess the ability to quantify the contents.

To study how changing the options on Kraken2 affects the ability to classify and quantify, the centered log ratio transformed results and PCA was utilized to analyze and visualize what level of confidence would result in the most accurate reproduction of the gold standard.

In this section the results of the analysis preformed on the data will be discussed in context of the aim of the study and then a discussion about the broader aspects of what could have been done differently and how what effects that might have on the results. First, in section 6.1 the results will be discussed in context of the aim of study. In section 6.2 the measurements for assessing classifications and quantifications will be discussed in greater detail and simulated samples will be compared with the real samples. Then section 6.3 will discuss the effects of changing the confidence settings on Kraken2 and how the dependencies in the data can be addressed using centered log ratio transformation. The effect of samples preparation is discussed in section 6.4 and how the composition of the database affects the classifications and quantifications is discussed in 6.5. Lastly, the effect of sample quality is discussed in 6.6

6.1 Contents classification and quantification by Kraken2+Bracken

The aim of the study was to classify and quantify the contents of fishmeal reliably and accurately using Kraken2+Bracken on metagenomic sequences. Simulated and real sequence data was used, and the results vary based on whether the samples are simulated or real, and within the real samples likely due to differences in the DNA extraction methods. Kraken2+Bracen is somewhat successful, but the results indicate that attention must be paid to the database compositions and thresholds used when analyzing the data.

Purity and completeness were used to measure the ability of Kraken2+Bracken to classify the contents of the samples. Seeing all samples, both simulated and real, as one it becomes clear that if all species with reads classified to them count as present, then the ability of Kraken2+Bracken to correctly predict what species are present in the sample is poor. As seen in Figure 5, Figure 6 and Figure 7 the purity will improve by imposing a threshold of a set number of reads that must be classified to a species for that species to be counted as present. This improvement in purity is caused by a reduction in the number of false positives, i.e., species incorrectly predicted to be present. Completeness sees the opposite pattern; as the threshold reaches a certain level, species that are in the samples will be predicted as not present, also referred to as a false negative. When studying the contents of fishmeal samples for the purpose of food authentication, what thresholds are used becomes important. A threshold that is too low will result in a higher likelihood of false positives, indicating that the sample contains species that should not be present. A threshold that is too high will result in a reduction in the ability to detect adulterations and falsely indicate that the species the contains, are absent. For the approach used in this study to be a tenable option for food

authentication, thresholds need to be determined so that the results of the analysis obtain adequate reliability.

Bray-Curtis dissimilarity was used to measure the accuracy of the quantification of the contents of the samples. The dissimilarity is calculated by comparing the number of reads classified to a given species by Kraken2-Bracken and the number of reads of that species in a gold standard, the true mix. This assumes a one-to-one relationship between the fraction of reads classified to a species and the faction of the contents in a sample of that species. The study utilized samples from three different origins; simulated samples based, samples sequenced with illumina and samples sequenced using ion torrent. When comparing samples with different origins it becomes clear that there are differences in the ability of Kraken2+Bracken to classify and quantify reads of different origins. Table 14 shows the Bray-Curtis dissimilarity for the samples, for all three origins. Kraken2+Bracken is more successful with the simulated samples, than with the real samples. The ion torrent samples are better than the illumina samples, which have large Bray-Curtis dissimilarities indicating a failure to correctly quantify the contents of the samples. This difference between the simulated and the real samples is expected as the simulated samples represent the best possible scenario, as the sequences being classified stem from the DNA sequences of the 21 genomes used to make the database. For the real samples there is an added element of complexity since the DNA extraction methods and sequencing systems use appear to have an effect on the classifications and quantifications. When using Kraken2+Bracken as a tool for classification and quantification of the contents for fishmeal it is important to be aware that the results are affected by the sample preparation. The simulated samples and the ion torrent samples have been prepared in a way that is in line with the assumption of a one-to-one between the fraction of reads from a species, and the fraction of contents in the sample for that species. The illumina samples are fishmeal mixes that were prepared in a way more closely resembling how samples would be prepared in a real-world setting, and these samples have the highest dissimilarity of all samples used in the study. The ability of Kraken2+Bracken to quantify is therefore dependent on the one-to-one assumption to be true, or there would need to be a way to correct for differences in genome composition that could account for this.

The contents of the samples in this study are classified and quantified using a database created from 21 genomes downloaded from the NCBI GenBank, all of which are listed in Table 4. When comparing the results from the classifications in Table 8, Table 9 and Table 10 with the sample compositions in Table 2 it is clear that some species are easier to classify more accurately than others. This stems from the fact that some of the 21 genomes in the database are more similar, measured by the mash distance. This similarity represents a higher number of shared k-mers. When using a k-mer based approaches to classifying the reads, such as Kraken2+Bracken, this similarity will lead to more misclassifications of one species to a closely related species. Reads from species with a greater evolutionary distance to the other species in the database, like *Gallus gallus*, will be easier to classify correctly. This can be observed especially well in the classification results for the simulated reads in Table 8, where the fraction of reads that are classified is very close to the gold standard, presented in Table 2. This indicates that when using Kraken2+Bracken for the purpose of food authentication, the composition of the database becomes important to the reliability of the results. Sample 9 and 14 contain *Hypophthalmichthys molitrix* and *Hypophthalmichthys nobilis* which has no reads classified to them by Kraken2+Bracken, stemming from a failure when constructing the database. Sample 9 and 14 both have lower completeness for all thresholds as seen in Table 8. This is because the two species will be counted as false negatives for sample 9 and *Hypophthalmichthys molitrix* will be counted as a false negative for sample 14. The reads that stem from these species will then either be discarded or misclassified as other species in the database. This results in a higher Bray-Curtis dissimilarity when compared to the other simulated samples, especially for sample 9, as seen in

Table 11. The takeaway from the findings is that composition of the database will have a large influence on the classification and quantification of the contents. When used for authenticating the contents of fishmeal, the database must be constructed based on knowledge of what species are meant to be present in the samples, what species might be added in cases of fraud and what species shouldn't be present from a sustainability perspective.

Orivo AS is the provider of a certification of, among other products, animal feed, and the logo of their certification utilize the word pure. Claiming that something is "pure" can be interpreted in many ways, for example, in this study samples with only one species were described as "pure". From the perspective of this study "pure" could be interpreted as; without false positives or misclassified reads, but these to standards are very different. The approach used in the study can achieve a result with no false positives using different thresholds or options for the classification algorithm as seen in the results presented in section 5.3.4. If on the other hand "pure" is interpreted to mean without misclassification, then the k-mer based approach is unlikely to ever be satisfactory as the genomes in a database fit for the purpose of food authenticating is highly likely to contain genomes similar enough to each other so that one read is misclassified. Whether or not the approach used in this study can give results that are of high enough accuracy as stated as the goal in the application for the research project therefor depends on what "high accuracy" entails. For some simulated samples, representing an unrealistic best-case scenario, the results approach appears to be able to classify and quantify. For the real samples the results are more mixed, with some samples being classified and quantified with some success, while another real sample is an example of a failure to classify and quantify.

6.2 Results and measurements

The ability of Kraken2+Bracken to classify the contents of the samples was assessed by using the measurements of purity and completeness. In sections 5.3.1 and 5.3.2 the results for assessing the classifications made by Kraken2+Bracken are presented. In Figure 5, Figure 6 and Figure 7 the purity for the simulated and real samples is visualized and in Figure 8, Figure 9 and Figure 10 the completeness of the simulated and real samples are visualized.

In the study, purity was used to measure what fraction of the species detected by Kraken2+Bracken were present in a given sample. A low purity indicates low accuracy and many false positives. At a threshold of 0, all samples have low purities. The simulated samples range from having a purity of 0.05 to 0.26, illumina all have 0.05 and the ion torrent samples range from 0.05 to 0.15 when the threshold is 0. The differences here stem form the number of species present in the gold standard. In Figure 5, showing simulated reads, it is apparent that sample 1, 8 and 15 start off with the lowest purity. In Figure 7, showing purity for ion torrent reads, sample 1 has the lowest purity. All versions of sample 1, 8 and 15 start with a purity of 0.0526, which at further inspection is the number of species in these samples, 1, divided by the number of species the database is able to classify reads to, 19 since two species never get any reads classified to them. This error in the database is discussed in further detail in section 6.5. Based on this observation it is apparent that a threshold of 0 reads is too low, since what determines the purity of the samples is the number of genomes in the database and in the gold standard for the sample. Is also reveals the weakness of purity as a measurement of the ability to classify. The k-mer based approach is likely to results in at least 1 read being classified to all species present in the database, this is compounded by the fact that Bracken also forces more ambiguous reads classified to a higher rank down to species rank, increasing the risk that reads are misclassified. Because of the nature of the k-mare approach, the purity is largely determined by the threshold imposed, and not so much the classification ability for kraken2.

Purity dose not take into account the false negatives and is therefor dependent on being used in conjunction with a measurement that does. In this study completeness was chosen to study the ability of Kraken2+Bracken to detect all species in a sample. Like purity, completeness is also not ideal when paired with a k-mer based classification method, since all species are likely to have one read classified to it. As seen in Figure 8, Figure 9 and Figure 10, showing the completeness of the simulated samples, illumina samples and ion torrent samples respectively, all samples except sample 9 and 14 form the simulated samples start with a completeness of 1. The completeness remains stable until the threshold reaches 0.1, meaning that any species with less than 10% of the reads classified to it will be counted as a false negative. This show, as with purity, that the thresholds to a large extent decide the completeness of the sample. The measurement becomes to blunt to assess the ability of Kraken2+Bracken to detect all species present in the sample, since

Both purity and completeness do not consider the number of reads classified to a species beyond the thresholds imposed. In the study, Bray-Curtis dissimilarity was used to measure the ability of Kraken2+Bracken to quantify the contents of the samples. This is done by compering how well the classifications match up with the gold standard for a given sample, assuming that there is a one-to-one relationship between the fraction of reads classified to a species, and the fraction of that species in the sample. Assuming this assumption to be true, the measurement is not optimal for measuring the quantifications. This stems from the fact that the read-counts are compositional data due to the way the sequencers work. The compositional aspect means that even if there was a one-to-one relationship between read composition and sample contents, there is a limit to how many DNA fragments are sequenced by the sequencing system, and for the quantification to be accurate the composition of the reads therefor must be assumed to be representative of the contents of the samples being sequenced. If one species was harder to sequence, that species would end up being underestimated, and the quantification would be inaccurate.

The ways of measuring the ability for Kraken2+Bracken to classify and quantify the contents of the samples was chosen based on Meyer et al.'s article about assessing metagenomic classifiers. Purity, completeness, and Bray-Curtis dissimilarity are uncomplicated measurements that are easy to implement for different metagenomic classification algorithms and therefor useful when studying how accurately different classifiers are. It is important to be aware of the shortcomings presented above when viewing the results of the study.

6.3 Effects of changing the confidence on Kraken2

In section 5.3.4 the results of the analysis exploring the effect of changing the confidence option are presented. This part of the study was done using only the 15 samples simulated using Art to study how changing the options would affect the classifications and quantifications. From Figure 14 it is clear that changing the confidence setting results in a clear drop in the number of reads classified by Kraken2+Bracken. As the strictness of the confidence option is increased the more ambiguous reads will not be classified and this represents a loss in the ability of Kraken2+Bracken to detect species with closer evolutionary relationships. The evolutionary relationship between the species in this study is characterized by a mash distance, which is a measure of shared k-mers. When a read contains k-mers that are shared by several species, the classification of the read will be less definitive, and might not pass a strict confidence. This might limit the usability for detecting adulterations in food and feed, especially those from species with close evolutionary relationships to other species.

As stated in the section above, Bray-Curtis dissimilarity does not adequately address the compositional nature of metagenomic sequence data. This was discussed by Gloor et al. and a centered log ration transformation was suggested as a way to account for the dependencies in the

data. To assess the quantification ability to of Kraken2+Bracken at the different confidence settings, PCA was used on the centered log ration transformed data to see what confidence would result in results closes to those observed for the gold standard. Figure 17 and Figure 18 shoe the results for sample 1 and 15, identical pure samples and show that using PCA to analyze the quantifications work, since the two identical samples results identical plots. The proximity of the points for the results at different confidences to the point for the gold standard represents how accurate the quantifications were. This is clearly illustrated by Figure 19 and Figure 20, showing the results of the PCA analysis on the centered log ratio transformed data of sample 9 and 14. Both these samples contain the two species the database was unable to classify reads to, leading to an incorrect quantification of the samples. This can be seen in the score plots as the increased distance between the points for the results and the gold standard. Figure 21 and Figure 22 show the results for sample 2 and 12, which represents the samples with what appears to be the most accurate quantification for the 15 samples. All the results indicate that the centered log ratio transformation manages to sufficiently address the compositional nature of the data so that PCA can be used and that other useful statistical tools can be utilized as well in future research.

From the 15 score plots, 11 of which can be found in section 10.2.1, it appears that a confidence setting of 0.6 or 0.7 provides the most accurate quantifications. This entails those quantifications based only on reads that have 60% or 70% of k-mers assigned to one species will be the most accurate. In future research it could be interesting to see if changing the confidence to 0.6 or 0.7 and preforming a centered log ratio transformation would result in improved Bray-Curtis dissimilarities for the samples, both simulated and real.

6.4 Effect of sample preparation

No conclusive picture about the different extraction methods can be drawn from the results for the illumina samples. illumina Sample 8 and illumina sample 15 both utilize the CHAN method but have greatly varying classification and quantification performances measured with Purity, Completeness and Bray-Curits dissimilarity. Illumina Sample 1 was prepared using the maricon method and has similar performance as sample 15, making it unclear if the DNA extraction technique has a great effect on the ability of Kraken2+Bracken to classify and quantify the contents of the samples. The method used DNA extraction is likely to have an effect, the real samples behave very different from both the simulated samples and each other. Studying how different protocols for preparing samples could affect the ability of Kraken2+Bracken to classify and quantify the samples was not a part of this study and a more structured approach to this is likely to yield more useful insights.

For the samples sequenced with ion torrent, the preparation of the samples resembles how the simulated samples were made. Just like reads where added to fastq files according to the sample compositions stated in Table 2 when making the simulated samples, DNA was first extracted and then added to mixes according to predefined compositional fractions, as those found in Table 2, then the mixes were sequenced. This was done to remove the differences in the amount of DNA for different species and the effect of extraction method. These samples can therefore be seen as the best case for real DNA sequences, but not a good representation for how real samples would be prepared when testing fishmeal. The differing genome sizes for species in fishmeal might also affect how the metagenomic sequence data is composed. This would affect the results of Kraken2+Bracken since the underlying assumptions for the quantification of the contents in the samples is that the fraction of reads from a species is analog to faction of a sample made up by that species. For the composition of the simulated samples and the ion torrent samples this assumption holds up, but not necessarily for the illumina samples, because of the more conventional way the illumina samples

were prepared. There are ways to normalize the data to account for different genome sizes and this, but such methods were not utilized in this study. Future research should attempt to take this into account and explore ways to account for genome size and how it could affect the classifications and quantifications.

6.5 Database composition

In this study a database for kraken2+Bracken was created. This custom database is used for assigning k-mers to taxonomic ranks, that then form the basis for the classifications of reads in the sample by Kraken2 and subsequently Bracken to make all classifications at the rank of species. This can be considered as a supervised classification algorithm and as such the classification are dependent on the composition of the database. It is therefore reasonable to assume that if the composition of the database was changed, different results would be obtained when analyzing the same samples.

An effect of the genomes in the database can be observed for the species *Hypophthalmichthys molitrix* and *Hypophthalmichthys nobilis* as no reads for any sample is classified to these species, while all other species have some reads classified to them. From the completeness and the Bray-Curtis dissimilarity seen in Figure 8 Figure 11 respectively, it is clear that this affects the measurements. Sample 9 and sample 14 of the simulated samples do not have a completeness of 1 for at any threshold. Sample 9 has the highest dissimilarity of all simulated samples, which is expected since the two species account for 10% of the contents in this sample. Sample 14 contains 5% *Hypophthalmichthys molitrix* and has one of the highest dissimilarities among the simulated samples, effected by the problems with the database.

When using the function "`kraken2-inspect`" in Linux to see what species have k-mers and minimizers in the database, it becomes apparent that *Hypophthalmichthys molitrix* and *Hypophthalmichthys nobilis* are not in the database. Using the "`grep`" command with the taxonomical id for these samples on the taxonomy downloaded from NCBI during the creation of the database, they appeared as one would expect. To be sure that the species are related to the taxonomy, the fasta files for the genomes of these species were manually altered so that the taxonomical id for the species where in the headers, as stated in the manual. A new database was created after the tax ids were added, but when inspected the genomes where not present. Why this is, is not clear, but it attests to the fact that using algorithms like Kraken2 and Bracken is not "plug and play", requiring tinkering and troubleshooting.

6.6 The effect of sample quality

To study how the quality of the samples affect the quantification of the contents, the Pearson correlation coefficient was calculated. In as presented in section 5.4.1, the coefficient was calculated to be 0.0123 which indicates no relationship between the quality of the samples and the ability of Kraken2+Bracken to quantify the contents. This is expected as the measurement for quality is not a good measurement for sample quality, but rather indicates the quality of the sequencing of the reads. As detailed in section 3.2.2, the Phred-scores states the probability of a sequencing error. Using the this as a stand in for the quality of the sequences that were sequenced is not a sound approach. The reason for choosing this approach in the study is that is easily implemented, especially when the preparation of the samples differ to a great extent, as is the case for the real samples in this study.

In section 5.4.2, the effect of random sampling when calculating the sample quality is visualized in Figure 23. The lines in Figure 23 show that random sampling improved how well the average quality of the sample represents the average quality of the reads. The line for random read sampling is more stable between 125 and 125.5. This indicates that the average quality of a read is not uniform throughout.

7 Further studies

This section presents elements of this study that should be studied in greater detail to gain more useful insights into the use of Kraken2+Bracken to accurately classify and quantify the contents of fishmeal samples.

7.1 presents different measurements that could be used to assess the classifications. Section 7.2 discusses why sample preparation should be studied further and 7.3 presents different classification algorithms for metagenomic data that have useful abilities that can address some of the elements discussed in the study.

7.1 Additional measurements to assess the classifications

In the discussion several species are discussed as evolutionary distant or close to others as this is likely to affect the ability of Kraken2+Bracken to classify the reads of those species correctly. A way to quantify the differences in accuracy for the species, a measure of pr. species inaccuracy could have been calculated. This could be something as simple as the average deviation from the gold standard for each species or a more advanced measurement. This would have made it easier to see what species where harder for Kraken2+Bracken to correctly classify and what species were easier.

It would also be interesting to study the effect Bracken has on the results. Bracken is used to re-classify reads classified to species level from higher taxonomic ranks so that all the reads are classified to species. It could be interesting to study if a greater rate of re-classifications would be correlated to a more inaccurate classification and quantification of the samples. It might also be possible to extract the result of the Bayesian calculation that the re-classifications are based on and study what species have mor uncertain re-classifications.

Different measures than the three chosen could also have been utilized. In the article by Meyer et al. where purity, completeness and Bray-Curtis dissimilarity were presented, other ways to assess the classifications were also presented. One of them was the Weighted UniFrac distance, which ranges from 0 to double the height of the taxonomic tree it utilizes. A low abundance indicates that the predictions are taxonomically similar to the gold standard for the sample being assessed. This takes into account more information about the relationship between the species than the measurements used in this study.

7.2 The effect of sample processing and preparation

In the introduction of the study, the ways DNA are degraded where presented in the context of having a possible effect on the metagenomic sequencing. Though the literature reviewed indicated that the DNA fragments should be of sufficient length to be sequenced, a more comprehensive approach to the study of DNA degradations effect on contents classification and quantification should be conducted. Different species might be easier or harder to detect after processing, parts of the DNA with a different structure might be more susceptible to degradation or other factors that could affect what DNA fragments are present for sequencing. These are unknown factors that should be explored in future research.

As discussed, it appears from the results that sample preparation could have an effect on the ability of Kraken2+Bracken to classify and quantify the contents of the samples. As a further step in exploring the approach to contents authentication laid out in this study, a systematic study of different protocols for preparing samples and its affect on the classification would be useful. The approach used for the ion torrent samples is not comparable to what would be done in practice.

Samples made from simulated reads also don't reflect reality as these samples are an unrealistic best-case scenario. The three samples that could be considered as representative for how samples would be handled in a real-world setting were not enough to gain any insight into how sample preparation might affect the classifications, just that it is likely to have an effect. For the illumina samples, more samples than could be utilized in this study were provided but lacked available genomes in the NCBI GenBank for all species in those samples. To gain insight into the effect of sample preparation on the ability of Kraken2+Bracken, additional genome assemblies should be created for relevant species so that the effect of sample preparation can be study using the existing sequence data, and ideally additional data.

7.3 Different ways to classified metagenomic sequence data

In this study Kraken2+Bracken was used to classify the contents of the samples analyzed, but there are other classification tools available, such as metaphaln which use marker genes to classify the DNA sequences. It would be interesting to see if different classifiers with different approaches to classifying sequences than the k-mer based approach used by Kraken2+Bracken would yield different results. A classified based on direct sequence alignment, more in line with BLAST might yield a more accurate classification of reads than the k-mer based used in this study. These suggestions show that there are several different ways to classify and quantify the contents, and a systematic comparison would be useful to see if one has strength that might way up for another's weakness, so that when they are used in combination, they provide reliable results.

8 Conclusion

Food authentication using metagenomic sequence is a promising way to increase the safety and reliability of food and feed products. This study finds that the hash based taxonomic classifier Kraken2+Bracken can classify and quantify the sequences in a metagenomic sample, but that a good understanding of the data that is being analyzed and the settings used for analysis is necessary. This study used simulated sequence data and real sequence data to study how the data quality would affect the accuracy of the results. The simulated data represents the best-case scenario for this approach, with samples-based sequences simulated form the same genomes used for the classification algorithm. The results for the simulated reads indicate that Kraken2+Bracken is able to classify the sequences and quantify the contents, but that accuracy is highly dependent on the setup of the software and the thresholds imposed when the results are analyzed. This is true for the real samples as well, but with the added level of complexity that data from real samples brings. For the real data the classifications and quantifications were less accurate, and the results indicate that sample preparation has an effect on the ability of Kraken2+Bracken. The real data also indicates that the assumption of a one-to-one relationship between the fraction of reads from a species in a sample and the contents of that species is unrealistic when analyzing real metagenomic sequence data. Using Kraken2+Bracken for food authentication has the potential to be an easy and reproduceable way to authenticate the contents of feed and food, but more structured research is needed to address the many factors that could affect the classification and quantification of the contents of the samples studied.

9 References

Abdi, H. and L. J. Williams (2010). "Principal component analysis." Wiley interdisciplinary reviews: computational statistics **2**(4): 433-459.

Aitchison, J. (1983). "Principal component analysis of compositional data." Biometrika **70**(1): 57-65.

Blanco-Fernandez, C., et al. (2021). "Seventeen years analysing mislabelling from DNA barcodes: Towards hake sustainability." Food Control **123**: 107723.

Challagundla Kishore Babu, S. K. K., Dr. Mukul Das (2007). Adulteration of Mustard Cooking Oil with Argemone Oil: Do Indian Food Regulatory Policies and Antioxidant Therapy Both Need Revisitation? Antioxidants & Redox Signaling. **9**.

Cock, P. J. A., et al. (2009). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." Nucleic Acids Research **38**(6): 1767-1771.

Coordinators, N. R. (2016). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Research **44**(D1): D7-D19.

Delgado, C. L., et al. (2003). Outlook for fish to 2020: meeting global demand, Intl food policy res inst.

Dreiseitl, S. and L. Ohno-Machado (2002). "Logistic regression and artificial neural network classification models: a methodology review." Journal of Biomedical Informatics **35**(5): 352-359.

Forth, L. F. and D. Höper (2019). "Highly efficient library preparation for Ion Torrent sequencing using Y-adapters." BioTechniques **67**(5): 229-237.

Francis, P. J. and B. J. Wills (1999). "Introduction to principal components analysis." arXiv preprint astro-ph/9905079.

Gloor, G. B., et al. (2017). "Microbiome Datasets Are Compositional: And This Is Not Optional." Frontiers in Microbiology **8**.

Haiminen, N., et al. (2019). Food authentication from shotgun sequencing reads with an application on high protein powders. npj Science of Food.

Haynes, E., et al. (2019). "The future of NGS (Next Generation Sequencing) analysis in testing food authenticity." Food Control **101**: 134-143.

Head, S. R., et al. (2014). "Library construction for next-generation sequencing: overviews and challenges." BioTechniques **56**(2): 61-passim.

Heather, J. M. and B. Chain (2016). "The sequence of sequencers: The history of sequencing DNA." Genomics **107**(1): 1-8.

Hu, T., et al. (2021). "Next-generation sequencing technologies: An overview." Human Immunology **82**(11): 801-811.

Huang, W., et al. (2012). "ART: a next-generation sequencing read simulator." Bioinformatics (Oxford, England) **28**(4): 593-594.

illumina (2021, 26.10.2021). "FASTQ files explained." from https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html.

Illumina (2022). "Power and efficiency for large-scale genomics."

Johnson, M., et al. (2008). "NCBI BLAST: a better web interface." Nucleic Acids Research **36**(suppl_2): W5-W9.

Johnson, R. (2014). Food Fraud and "Economically Motivated Washingtion DC, Congressional Research Service.

Keller, J. M., et al. (1985). "A fuzzy k-nearest neighbor algorithm." IEEE transactions on systems, man, and cybernetics(4): 580-585.

Kurtzer, G. M., et al. (2017). "Singularity: Scientific containers for mobility of compute." PLOS ONE **12**(5): e0177459.

Levy, S. E. and R. M. Myers (2016). "Advancements in Next-Generation Sequencing." Annual Review of Genomics and Human Genetics **17**(1): 95-115.

Lien, S., et al. (2016). "The Atlantic salmon genome provides insights into rediploidization." Nature **533**(7602): 200-205.

Lu, J., et al. (2017). "Bracken: estimating species abundance in metagenomics data." PeerJ Computer Science **3**: e104.

Marchet, C., et al. (2021). "Data structures based on k-mers for querying large collections of sequencing data sets." Genome Research **31**(1): 1-12.

Marvin Ingi Einarsson, A. J., Anne Mette Bæk, Anne Mette Bæk, Søren Anker Pedersen, Tor Andreas Samuelsen, Jóhannes Pálsson, Odd Eliasen, Ola Flesland (2019). Nordic Centre of Excellence Network in Fishmeal and Fish oil.

Mente, E., et al. (2006). "Effect of feed and feeding in the culture of salmonids on the marine aquatic environment: a synthesis for European aquaculture." Aquaculture International **14**(5): 499-522.

Merriman, B., et al. (2012). "Progress in ion torrent semiconductor chip based sequencing." Electrophoresis **33**(23): 3397-3417.

Meyer, F., et al. (2019). Assessing taxonomic metagenome profilers with OPAL. Genome Biology. **20**.

Miles, R. D. and F. A. Chapman (2006). "The benefits of fish meal in aquaculture diets." EDIS **2006**(12).

Noble, W. S. (2006). "What is a support vector machine?" Nature Biotechnology **24**(12): 1565-1567.

Ondov, B. D., et al. (2016). "Mash: fast genome and metagenome distance estimation using MinHash." Genome Biology(1474-760X (Electronic)).

Quail, M. A., et al. (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." BMC Genomics **13**(1): 341.

Quail, M. A., et al. (2009). "Improved protocols for the illumina genome analyzer sequencing system." Current protocols in human genetics **Chapter 18**: 10.1002/0471142905.hg0471141802s0471142962-0471142918.0471142902.

Salter, A. M. (2017). "Improving the sustainability of global meat and milk production." Proceedings of the Nutrition Society **76**(1): 22-27.

Team, R. C. (2021) R: A Language and Environment for Statistical Computing.

Tille, A. (2014). Manual for art_illumina at Ubuntu manuals. Ubuntu manuals.

Torsten Bauer, P. W., Walter Hammes, Christian Hertel (2003). The effect of processing parameters on DNA degradation in food. European Food Research and Technology**:** 338-343.

Weitschek, E., et al. (2014). "Supervised DNA Barcodes species classification: analysis, comparisons and results." BioData Mining **7**(1): 4.

Wood, D. E., et al. (2019). "Improved metagenomic analysis with Kraken 2." Genome Biology **20**(1): 257.

10 Appendix

10.1 Scripts

All scripts used in this study is available at the GitHub repository:

[halvor-ekeland/Master: Master i bioinformatikk/anvendt statistikk (github.com)]

10.2 PCA and scree plots

10.2.1 PCA score plots

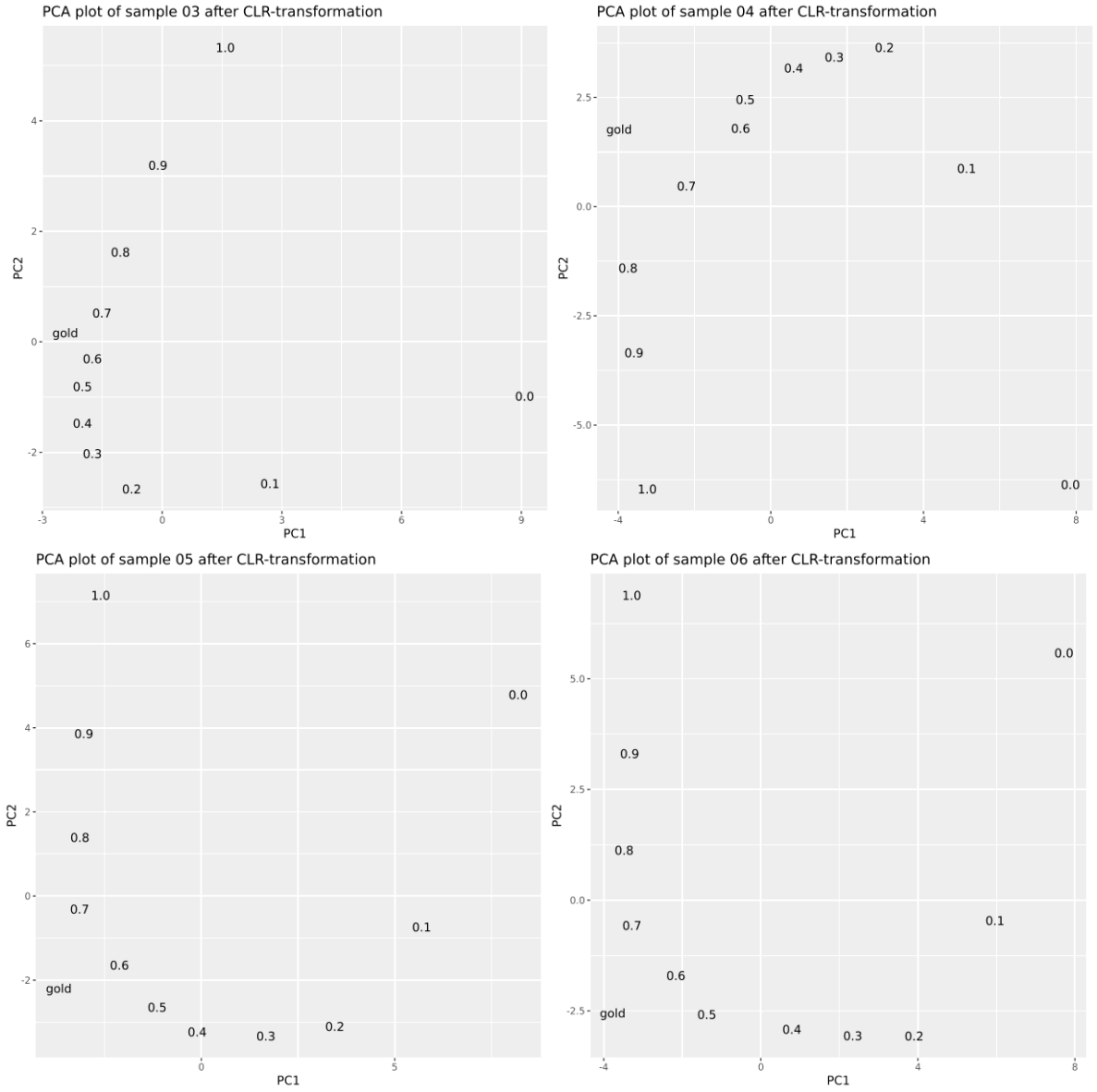

Figure 24: Score plots for PCA of sample 3 (top left), sample 4 (top right), sample 5 (bottom left) and sample 6 (bottom right) simulated with Art. Each potin represents the results for the sample at the confidence denoted by the label of the point. The gold standard is denoted by the text "gold"

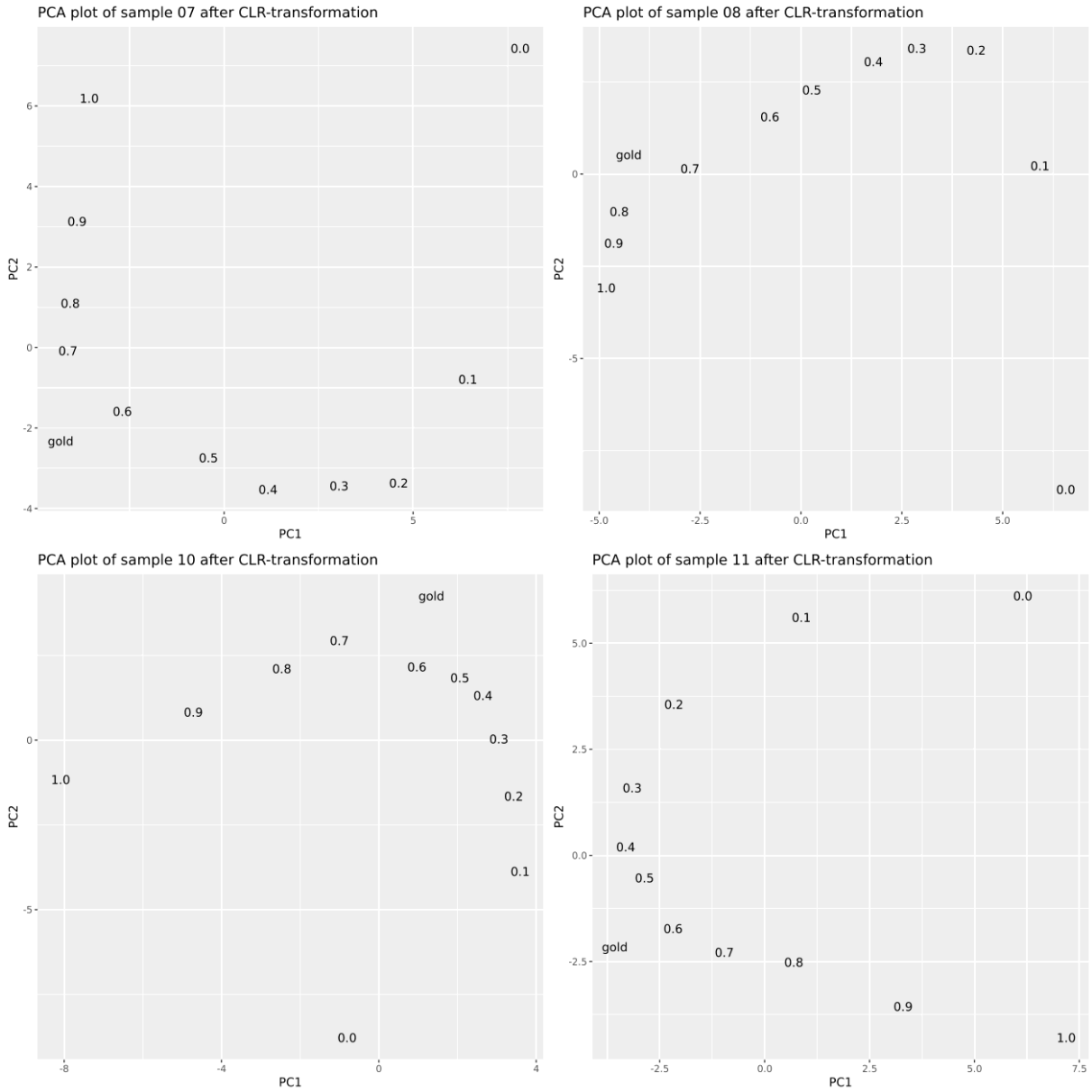

Figure 25: score plots for PCA of sample 7 (top left), sample 8 (top right), sample 10 (bottom left) and sample 11 (bottom right) simulated with Art. Each potin represents the results for the sample at the confidence denoted by the label of the point. The gold standard is denoted by the text "gold"

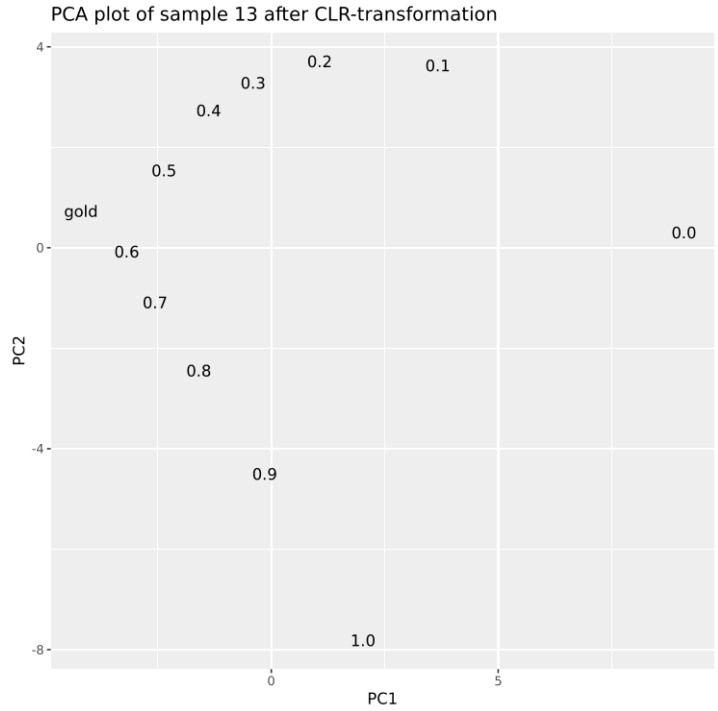

Figure 26: Score plot for PCA of sample 13. Each potin represents the results for the sample at the confidence denoted by the label of the point. The gold standard is denoted by the text "gold"

10.2.2 Scree plots

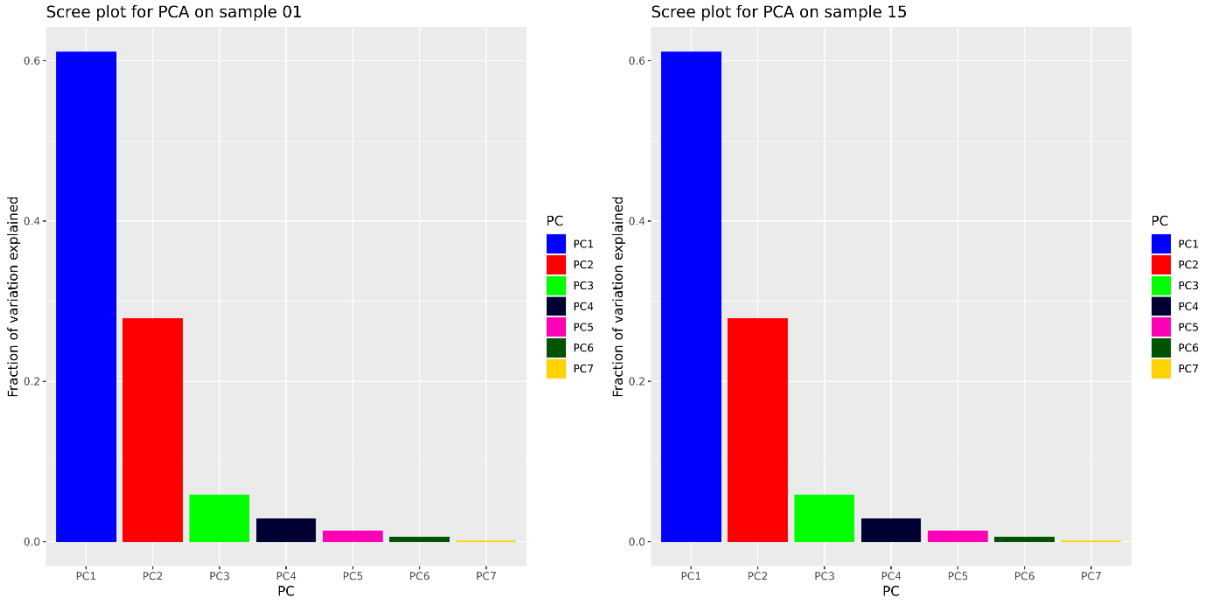

Figure 27: Scree plots for sample 1(left) and sample 15 (right) simulated with Art. columns represent each principal component; the height of the column being determined by the share of variance explained by that component. y-axis denotes the share of variance

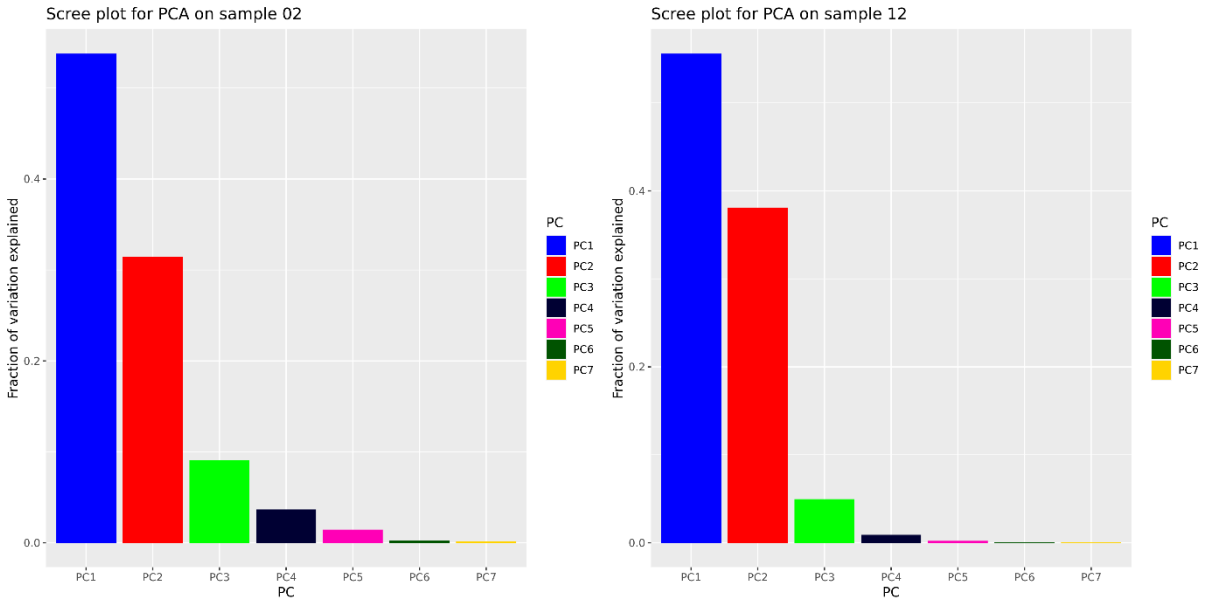

Figure 28: Scree plots for sample 2 (left) and sample 12 (right) simulated with Art. columns represent each principal component; the height of the column being determined by the share of variance explained by that component. y-axis denotes the share of variance

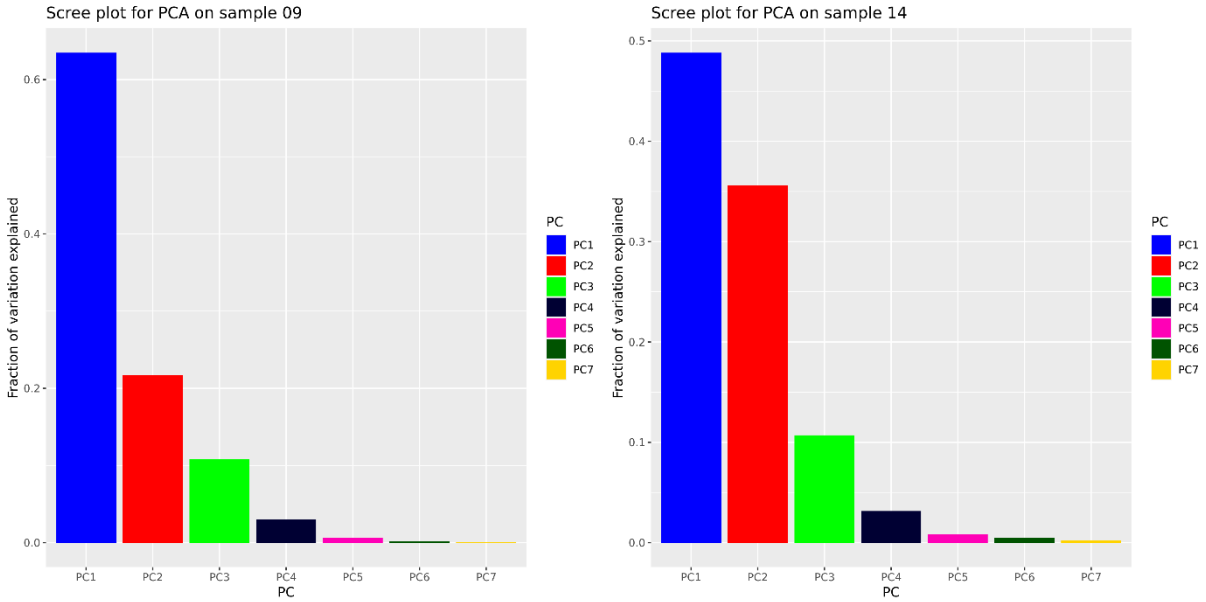

Figure 29: Scree plots for sample 9 (left) and sample 14 (right) simulated with Art. columns represent each principal component; the height of the column being determined by the share of variance explained by that component. y-axis denotes the share of variance

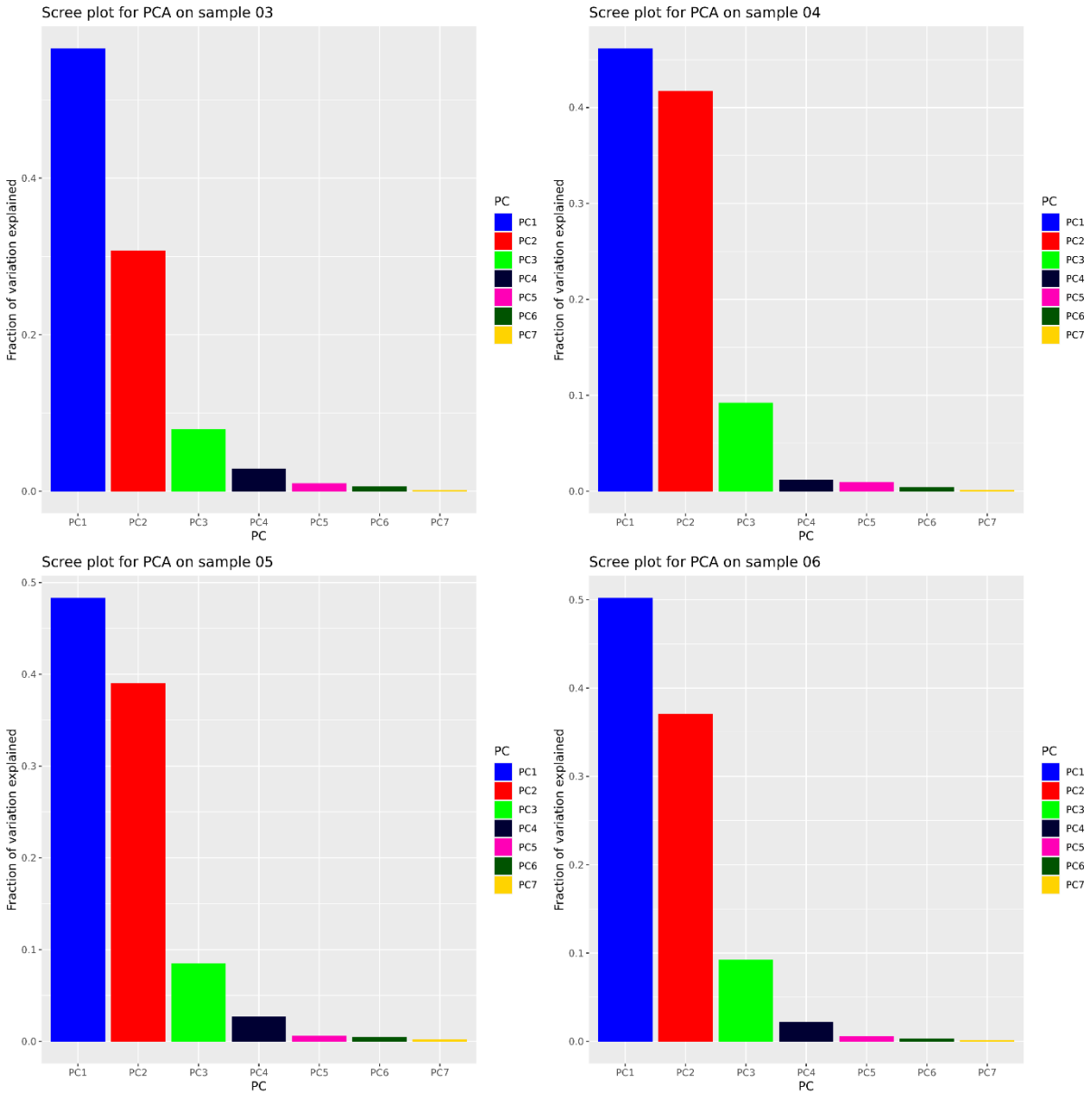

Figure 30: Scree plots for sample 3 (top left), sample 4 (top right), sample 5 (bottom left) and sample 6 (bottom right) simulated with Art. columns represent each principal component; the height of the column being determined by the share of variance explained by that component. y-axis denotes the share of variance

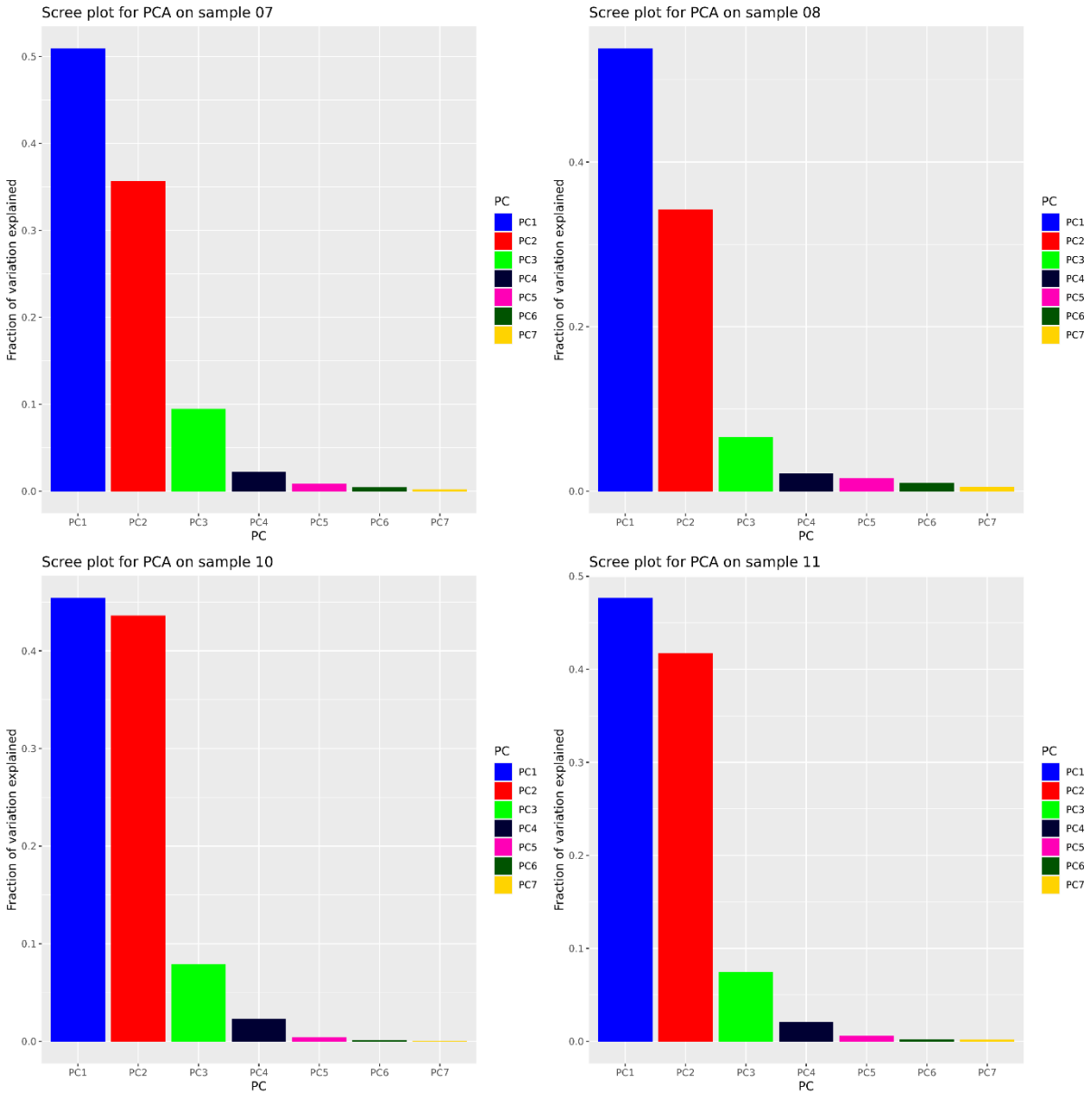

Figure 31: Scree plots for sample 7 (top left), sample 8 (top right), sample 10 (bottom left) and sample 11 (bottom right) simulated with Art. columns represent each principal component; the height of the column being determined by the share of variance explained by that component. y-axis denotes the share of variance

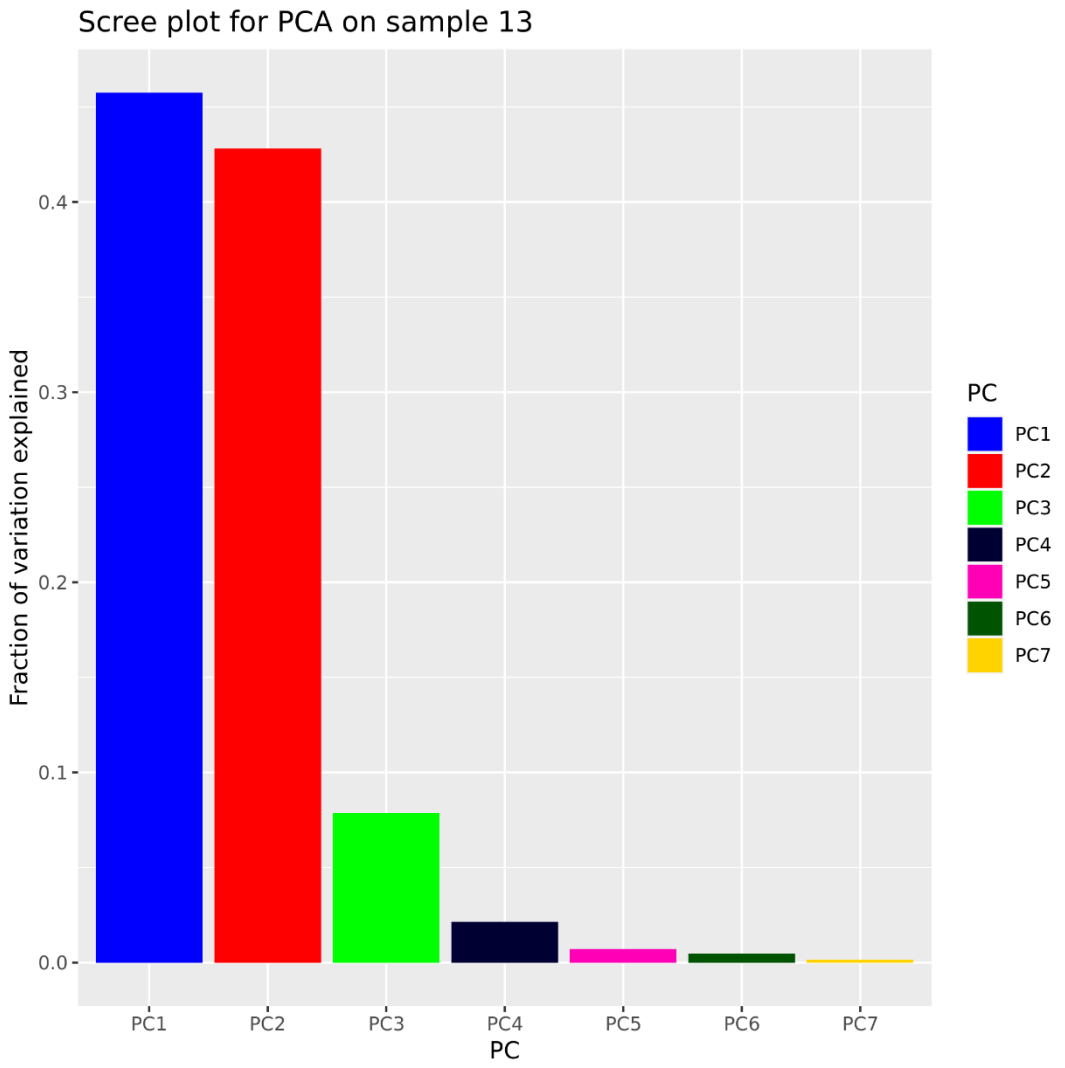

Figure 32: Scree plots for sample 13 simulated with Art. columns represent each principal component; the height of the column being determined by the share of variance explained by that component. y-axis denotes the share of variance

10.3 Time and CPU usage for creating the database

Table 15: CPU usage when creating the database.

JobID	JobName	AveRSS	MaxRSS	Elapsed	CPUTime	AveCPU
14340191	KBdatabase	05:31:50	2-07:18:20			
14340191.ba+	batch	178863724K	178863724K	05:31:50	2-07:18:20	09:39:35

10.4 Overview of figures in the document

10.5 Overview of tables in the document

10.6 Overview of equations in the document

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway