Norwegian University
of Life Sciences

Master's Thesis 2019 30 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Benchmarking of Metagenomic Classification Tools and Storage Stability of Bioaerosol Samples

Kristina S. R. Stenløkk
Chemistry and Biotechnology (M.Sc.) - Bioinformatics

Abstract

Air is a microbial habitat of crucial importance for public health. As such it is relevant for detection of potential epidemic or biothreat agents. The study of microbiological diversity in air through metagenomic analysis is a field under rapid development, and demands more knowledge. The work presented in this thesis investigated the current procedures used in metagenomic analysis of air samples, and consists of two parts. The first part assessed how long-time storage at low temperatures affects the stability of DNA concentration of filter-based air samples. Qubit and qPCR targeting the 16S rRNA gene were used to measure the DNA concentration. No evidence was found suggesting a detrimental effect of filter storage at -80°C. However, the findings may suggest negative effect of repeated freeze-thaw cycles on the yield of purified DNA. The second part assessed the performance of three metagenomic classification tools for creation of taxonomic profiles of air samples: Kraken 2, One Codex and Kaiju. The testing was conducted on various datasets. The results showed that Kraken 2 is the superior classification tool of well-studied species. However, Kraken 2 performed poorly on more complex datasets closer resembling the biological composition in air samples, due to inadequacies in the reference database. The classification of real air samples showed substantial variation between the profiles made by the tools. These findings further emphasise the need for improvements of the reference databases by adding more species specific for air, which should be a key objective for further work. There could also be improvements from altering the lowest common ancestor approach implemented in the classification algorithms, which seems to be a limiting factor for the taxonomic resolution.

Sammendrag

Luft er et mikrobielt habitat med stor betydning for folkehelsen, med relevans for deteksjon av potensielle smittestoffer som kan føre til epidemiske utbrudd. Studiet av mikrobiell diversitet i luft ved metagenomisk analyse er et forskningsfelt i rask endring, og det kreves mer kunnskap. Arbeidet som er presentert i denne oppgaven tok for seg prosedyrene brukt for metagenomanalyser av luft, og består av to deler. Den første delen tok for seg hvordan langtidslagring på lave temperaturer påvirker DNA-konsentrasjonen av filterbaserte luftprøver. Qubit og qPCR basert på 16S rRNA-genet ble brukt til å måle DNA-konsentrasjonen. Det ble ikke funnet bevis for at lagring på -80°C opp til syv måneder har negativ effekt på konsentrasjonen. Resultatene antyder imidlertid negativ effekt fra gjentatte tine-fryse-sykluser på DNA-konsentrasjonen av renset DNA. I den andre delen vurdertes prestasjonen av tre metagenomiske klassifiaksjonsverktøy brukt til å lage taksonomiske profiler av luft: Kraken 2, One Codex og Kaiju. Testingen ble gjennomført på ulike typer datasett. Resultatene viste at Kraken 2 gjør den beste klassifiseringen på velstuderte arter, men presterte dårligst på mer komplekse datasett som ligner mer på artene funnet i luft. Dette skyldes mangler i referansedatabasen. Klassifikasjonen av reelle luftprøver viste betydelige avvik mellom profilene fra de testede verktøyene. Disse funnene understreker at databasene må utbedres ved å legge til arter mer spesifikke for luft, som bør være et hovedpunkt for videre arbeid. «Least common ancestor»-tilnærmingen som brukes av verktøyene kan også forbedres, da det ser ut til å være en begrensende faktor for den taksonomiske oppløsningen.

Acknowledgments

The work presented in this thesis was performed for the Comprehensive Defence Division at the Norwegian Defence Research Establishment (FFI) at Kjeller as a part of the Master program in Chemistry and Biotechnology, at the Faculty of Chemistry, Biotechnology and Food Sciences (KBM) at the Norwegian University of Life Sciences (NMBU) the spring of 2019.

I would like to thank Marius Dybwad, my main supervisor at FFI, for helping me throughout this thesis. Your working capacity is astonishing, and you are a big inspiration for the scientist I one day hope to become. I also want to thank Kari Oline Bøifot and Tone Aarskaug at FFI for guiding me in the lab.

My main supervisor at NMBU, Lars-Gustav Snipen, has given excellent guidance and support. I am truly grateful for all your help. Further, my co-supervisor Knut Rudi deserves a note of gratitude for advising me. Also, a special thanks to Jon Olav Vik for introducing me to Tidyverse and for the enthusiasm for my project.

Lastly, a big thanks to my family and friends for the support through the five years at NMBU. In particular, I want to thank Pelle Mikkelsen for invaluable advice and encouragement.

Contents

List of Figures

List of Abbreviations

bp	Base pairs
BLAST	Basic local alignment search tool
BWT	Burrows wheeler transform
C_q	Quantification cycle
FFI	Forsvarets Forskningsinstitutt (Norwegian Defence Research Establishment)
LCA	Lowest common ancestor
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NMBU	Norges miljø- og biovitenskapelige universitet (Norwegian University of Life Sciences)
MAGs	Metagenome assembled genomes
MEMs	Maximum exact matches
MSA	Multiple sequence alignment
PCR	Polymerase chain reaction
qPCR	Quantitative polymerase chain reaction

1 Introduction

Air is a microbial habitat of crucial importance for public health. As such it is relevant for detection of potential epidemic or biothreat agents (Kuske, 2006; Be et al., 2014). Microbial studies of air was first of scientific interest due to disease transmission, but it has been shown that air is a habitat where microbes can grow and reproduce (Sattler et al., 2001; Dimmick et al., 1975; Amato et al., 2007). The acknowledgement of air as a relevant microbial environment combined with the rapidly decreasing cost of next generation sequencing (NGS) (Shokralla et al., 2012) has led to sequencing being the gold standard in the field. Specifically, shotgun metagenome sequencing is becoming a widely adapted method for determining the biological composition of air (Tringe et al., 2008; Behzad et al., 2015; Rosario et al., 2018). With increasing access to high quality sequence data (Lindgreen et al., 2016), adequate metagenomic classification tools are crucial for reliable taxonomic profiles.

1.1 Microbes in the air

The term bioaerosol is often used when addressing microbes in the air. Aerosols are liquid or solid particles suspended in a gaseous medium, typically air (Dybwad, 2014). Accordingly, bioaerosols are defined as aerosols of microbial, plant or animal origin (Heedrik et al., 2003). This includes bacteria, virus, fungi, toxins and pollen, animal and plant debris. There is a distinction between microbial compositions in indoor and outdoor air. While the largest contributor to microorganisms in indoor air are humans (Prussin and Marr, 2015), the origin of components in outdoor air varies greatly (Kuske, 2006). There are seasonal variations, variation between local climates, and due to changing weather conditions.

The density of biomass in air is extremely low compared to other environmental samples (Behzad et al., 2015). The density is estimated to be approximately 10^4 cells$/m^3$ (Burrows et al., 2009), compared to $10^{10} - 10^{11}$ cells per gram of soil, or $10^{13} - 10^{14}$

in the human gut (Gill et al., 2006). This puts high demands on sampling and sample processing to get the DNA yield required for sequencing.

1.2 Studying the biological diversity in environmental samples

Breitwieser et al. (2017) states that the hierarchical taxonomy being used to classify life is not optimal for microorganisms, as it was first intended for multi-cellular organisms. The microbial world has turned out to be a lot more complex than scientist first thought when they started creating taxonomic naming schemes. The term "species" originally refers to individuals that can interbreed and produce fertile off-springs in the next generation (Rosselló-Móra and Amann, 2015). This definition is not directly transferable to bacteria for instance, as they do not sexually reproduce, and there is a possibility for horizontal gene transfer. Despite these problems, scientist has to adhere to this taxonomy. It is hence important to keep in mind that the classification system is not ideal when assessing the microbial content in aerosol samples.

For a long time, the study of bacterial content in environmental samples was based on cultivation. This includes the study of bioaerosols (Fang et al., 2005; Cronholm, 1980; Dybwad, 2014). A major drawback of this technique is that the proportions of cultivable microorganisms may not necessarily reflect the true composition of the environment. Amann et al. (1995) suggested that less than 1 % of the bacteria in any environmental sample is cultivatable. Although there is increasing knowledge of how to cultivate bacteria in the laboratory, the representation of the biological composition is still biased (Stewart, 2012).

As NGS became widely available at the onset of the 21st century, the use of amplicon sequencing increased (Mardis and McCombie, 2017). Amplicon sequencing, also known as marker gene sequencing, is a cost and labour effective method for investigation of the biological diversity. A highly conserved gene is targeted, most commonly the 16S ribosomal gene for bacteria and 18S for eukaryotes (Yooseph et al., 2013). However, there are major drawbacks for studying the entire biological diversity when only parts of the genomic content is sequenced. As the targeted gene is conserved, the taxonomic resolution, which is a required for a species-level classification, is reduced (Zou et al., 2019). Further, amplicon sequencing suffers from a primer bias, because the primers of less known species are less likely to be amplified and represented in the final metagenomic profile (Schirmer et al., 2015).

Shotgun sequencing is an alternative to amplicon sequencing. Shotgun offers direct sequencing of the entire genomic content, instead of being limited to a marker gene (Breitwieser et al., 2017; Yooseph et al., 2013). The method randomly fragments the DNA into pieces. The fragments are then sequenced, which results in short nucleic acid sequences called reads.

A major benefit of shotgun sequencing is the increased taxonomic resolution compared to amplicon sequencing (Be et al., 2014). Shotgun sequencing allows both analysis of taxonomic composition and metabolic potential. It is also possible to screen for virulence factors and genes associated with antibiotic resistance with shotgun sequence data. Furthermore, as viral genomes do not contain an analogue gene to the 16S and 18S gene often used for amplicon sequencing, viral species will not be detected. Hence this is a major drawback of amplicon sequencing if the entire biological diversity is to be studied. While amplification based methods are less costly to perform (Breitwieser et al., 2017), the high taxonomic resolution from shotgun is beneficial when studying environmental samples at species-level.

1.3 Bioaerosol metagenomics

Despite the choice of amplicon or shotgun sequencing, metagenomic classification tools are used for the creation of taxonomic profiles. This is done by mapping the reads to sequences contained in a reference database (Gardner et al., 2019). The term metagenomics can be used for a variety of techniques and bioinformatic tools, ranging from taxonomic profiling to the study of expressed genes of an environment (Thomas et al., 2012). For the purpose of this work, metagenomics referrers to taxonomic profiling with the use of shotgun sequence data.

Metagenomics of bioaerosol samples is a novel field under rapid development. One of the greatest challenges is the lack of standardised protocols and methodologies, according to Behzad et al. (2015). This makes studies on the biological content of air difficult or even impossible to compare, due to a variety of methods being used for both sampling and quantitative or qualitative analysis (Dybwad, 2014). With a proliferation of bioinformatic tools for analysing metagenomic datasets, the selection of adequate tools are more important than ever (Gardner et al., 2019). Many benchmarkings of metagenomic classification tools are conducted (Lindgreen et al., 2016; Peabody et al., 2015; Sczyrba et al., 2017; Almeida et al., 2018), but none specifically for bioaerosol samples.

Another challenge in the field of metagenomic profiling is the substantial part of novel organisms not yet contained in any existing databases. As this used to be the case for other environmental samples, such as the gut microbiota, it has been shown that the development of more environment specific reference databases may solve this problem (Zou et al., 2019; Forster et al., 2019).

McIntyre et al. (2017) states that identification of microorganisms in clinical and environmental samples is one of the main challenges of metagenomics. Both the number of studies using metagenomics for analysis of environmental samples and new bioinformatic tools are rising with the decreasing cost of NGS (Breitwieser et al., 2017). This makes selection of appropriate tools more demanding, but also of a greater importance than before (Gardner et al., 2019).

1.4 Translating air into sequence

The creation of metagenomic profiles from bioaerosol samples is a process consisting of several steps. The first step is the collection of bioaerosol samples. There is a variety of sampling methods that can be used, ranging from collection into a liquid, to collection onto a dry filter (Dybwad et al., 2014). The next step is DNA isolation optimised for bioaerosol samples, followed by sequencing. The read data can then go through a quality control, removing adapters and reads of poor quality. Lastly, the files are used as input to a metagenomic classification tool, with the final result being a taxonomic profile of the bioaerosol sample.

Since the whole process from air to sequence takes several days, it is not one continuous workflow, and hence often require storage before further processing or after DNA isolation. How the stability of DNA samples are affected by storage by freezing is an unexplored subject. Ross et al. (1990) found that after freezing DNA extracted from blood once, the yield decreased by more than 25 %. This indicates a departmental effect of freezing DNA samples. However, the research was conducted using isolated DNA from one cell type. On the contrary, bioaerosol samples are often stored as filters and contains a complex mixture of cell types a variety of taxa, including both gram-positive and gram-negative bacteria, spores and fungi. As a result of this, differences in their cell wall composition could lead to different effects from freezing and storage. If that is the case, the resulting composition found in the taxonomic profiles may be biased. Additionally, the DNA concentration in bioaerosol samples are considerably lower than that used by Ross et al. (1990) which could impact the stability.

At FFI, bioaerosol samples are often stored at -80°C on order to preserve the samples. It is of interest to confirm that the storage is not disrupting the stability of the samples. Further, as a way of making the handling of samples more streamlined is to store the samples in a buffer, this storage method is also of interest to investigate.

1.5 Aims of the study

The overall goal of this study is to expand the knowledge of metagenomic analysis of filter-based bioaerosol samples to improve the current procedures. This is carried out through two subgoals:

1. Assess how filter-based bioaerosol samples are affected by long time storage at low temperatures.
2. Getting insight into how the selection of metagenomic classification tools influence the resulting taxonomic profiles of bioaerosol samples.

An outline of filter-based bioaerosol sample preparation, subjects of matter and methods used for investigation in this thesis are illustrated in Figure 1.1.

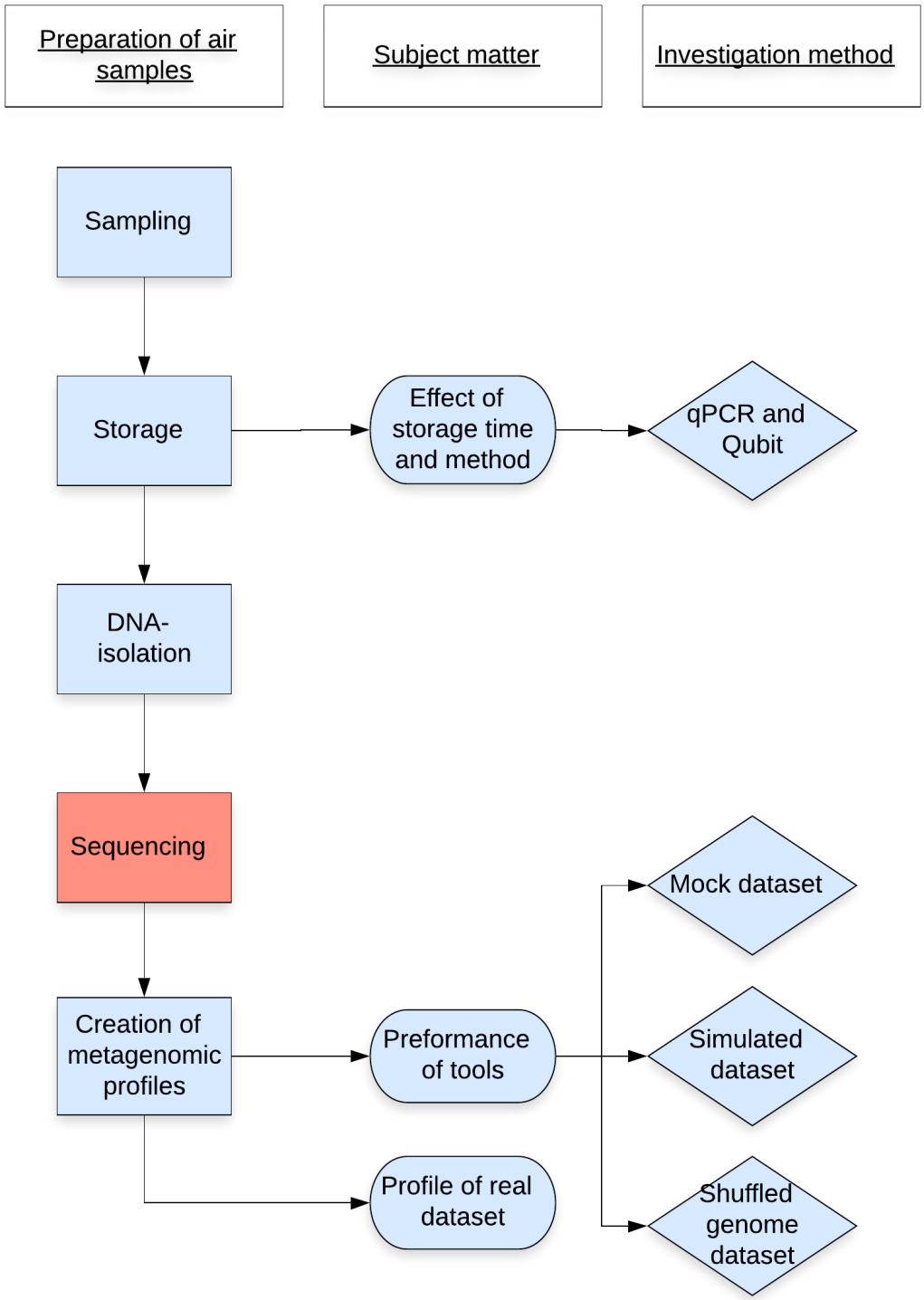


Figure 1.1: The flowchart shows the steps in filter-based bioaerosol sample preparation, with associated subjects of matter relevant in this thesis, and the method used for assessment. The sequencing (red colour) was conducted by FFI.

2 | Materials and methods

All data analysis was carried out using Rstudio 3.5.0 (RStudio Team, 2015). The figures are made with the *ggplot2* (Wickham, 2016) and *ggtree* (Yu et al., 2017) packages.

2.1 Storage study

The storage study was conducted to investigate whether aerosol samples have a stable DNA concentration after storage at -80 °C up to seven months. To quantify this, 78 aerosol samples were collected and stored for different time periods (three and seven months) and by different storage conditions (filter storage and buffer storage). DNA was then isolated from the samples and concentration was measured by two methods; quantitative polymerase chain reaction (qPCR) targeting the 16S rRNA-gene, and by using Qubit which measures the total DNA concentration by fluorescence.

2.1.1 Sample collection

Collection of aerosol samples was done using SASS 3100 high-volume air samplers (Research International, Monroe, WA, USA). The air was sampled directly onto a dry electret filter, at a flow rate of 265 L/min for 3 hours, resulting in a filtration of $47.7\,\mathrm{m}^3$ of air per sample. The sample collection was carried out between 1st and 13th of June 2018 at Kjeller in a rural environment. Six identical SASS air samplers were used per sampling day, giving six parallels of filters from each of the 13 days, which in total resulted in 78 air filters.

2.1.2 Sample storage

The six parallel samples from each day were then divided into three groups:

1. DNA isolated directly after collection
2. DNA isolation after storage for three months at -80°C
3. DNA isolation after storage for seven months at -80°C

The two groups being stored (2 and 3) were further divided into two groups:

1. Half of the samples stored filters
2. Half of the samples stored as filter extracts

Separation into groups is illustrated in Figure 2.1. The filters were put directly into a 50 ml Falcon tubes and frozen. In order to prepare the filter extract, samples from 2-3 days were accumulated in the freezer, thawed and then filter extracted. This is the first step in the DNA isolation procedure, resulting in a supernatant and a pellet (see Figure 2.2). The supernatant (NucliSENS lysis buffer) was stored in 50 ml Falcon tubes, while the pellet was resuspended in 150 µl of the buffer solution Phosphate-buffered saline (PBS) in 1,5 ml Eppendorf tubes. This storage methods will be refereed to as filter storage and buffer storage.

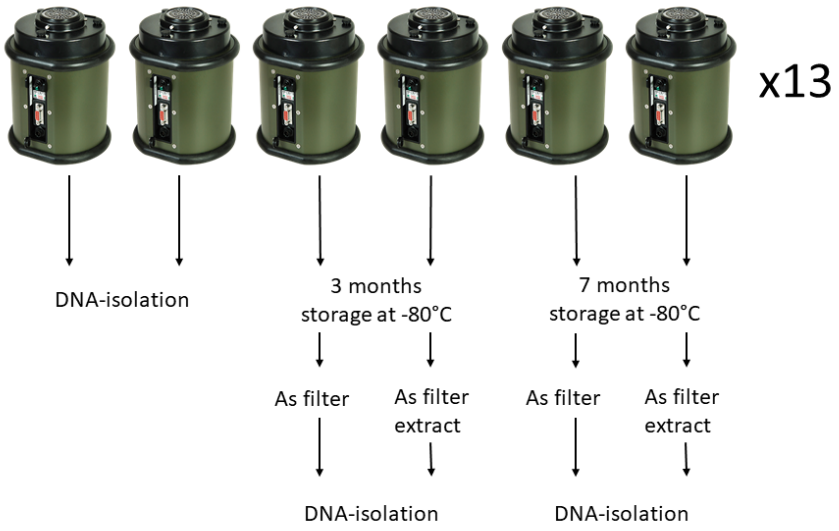


Figure 2.1: Sampling setup for 78 aerosol samples conducted in June 2018. Six SASS air samplers collects six aerosol samples per day of sampling. Two of the six samples are isolated directly, two stored for three months, and two stored for seven months, by either storage as filter or buffer.

2.1.3 DNA isolation

The DNA isolation was conducted according to the method optimised for bioaerosol samples by FFI (Bøifot et al., 2019). An overview of the method is shown in Figure 2.2.

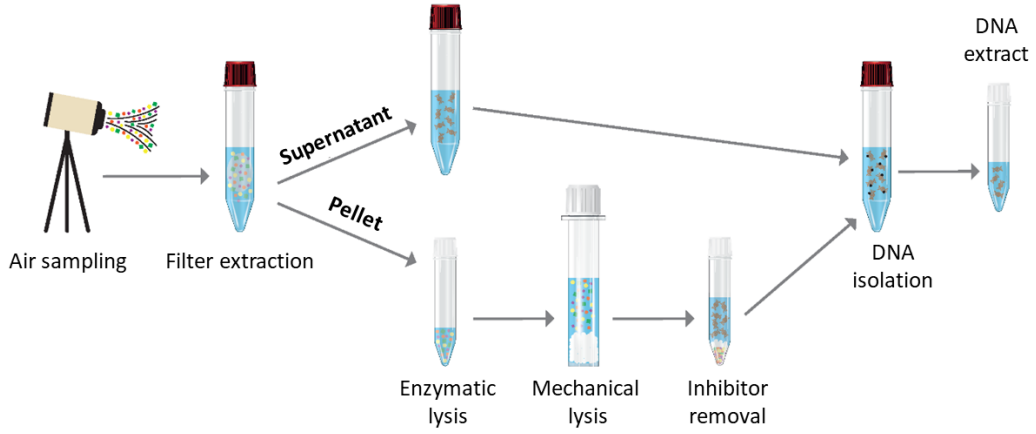


Figure 2.2: Illustration of DNA isolation procedure optimised for bioaerosol samples. After filter extraction, the sample is split into supernatant and pellet, which can be stored separately. Figure reprinted by permission of Bøifot et al. (2019)

The DNA isolation procedure starts by filter extraction. The filters are transferred into Falcon tubes with 10 ml NucliSens Lysis Buffer (BioMérieux, Marcy-l'Étoile, France), before vortexing for 20 seconds. The filters are then put into a syringe to extract the liquid.

Separation of supernatant and pellet is done by centrifuging the samples at 7.000 g for 30 minutes. The supernatant is transferred to a new tube. The pellet is dissolved in 150 ml PBS (pH 7.5, Sigma-Aldrich, St. Louis, MO, USA) and centrifuged at 17.000 g for 5 minutes. The liquid is added to the Falcon tubes containing the supernatant. 10 µL 5 mM MetaPolyzyme (Sigma-Aldrich, St. Louis, MO, USA) and 5 µL sodium acid (0.1 M, Sigma-Aldrich, St. Louis, MO, USA) was added to the pellet, before 60 minutes of incubation at 35 °C.

The samples were transferred to ZR BashingBead Lysis Tubes (0.1/0.5 mm beads, Zymo Research Corp) pre-filled with 550 µL PowerSoil Bead Solution (Qiagen, Hilden, Germany) and 60 µL PowerSoil Solution C1. Tubes were put on a Mini Bead Beater-8 (BioSpec Products) at 17.000 g for 3 minutes. The tubes were then centrifuged at 13.000

g for 2 minutes. The supernatant was treated with Solution C2 and C3 according to the DNeasy PowerSoil protocol (Qiagen, Hilden, Germany). The resulting supernatant was combined with the supernatant from the filter extraction.

DNA was purified according to the protocol of the NucliSENS Magnetic Extraction Reagents kit (BioMérieux, Marcy-l'Étoile, France). Instead of 60 µL silica suspension, the volume was increased to 90 µL, with a 20 minute incubation.

2.1.4 Measurements of DNA concentration

Qubit

A Qubit™ 3.0 Fluorometer (ThermoFisher Scientfic) was used to quantify the total DNA concentration in all 78 aerosol samples. 10 µL of each sample was measured immediately after isolation. The high-sensitivity dsDNA Qubit Kit has a detection range of 0.2–100 ng, and binds specifically to DNA (Mardis and McCombie, 2017). A house standard was also measured for each round Qubit measurements to ensure a stable signal.

16S rRNA qPCR

qPCR is a DNA quantification method that when used with probes is marker specific. It is widely used for quantifying microorganisms in environmental samples (Zhang and Fang, 2006). The amount of PCR product is measured during the course of the reaction by monitoring the fluorescence of dyes or probes. With an exponential growth, it is then possible to estimate the initial concentration (Kubista et al., 2006). The method is reproducible and highly sensitive, compared to older techniques, but only amplified DNA will produce a signal strong enough for detection. The output is a quantification cycle (C_q), that can be translated to a 16S rRNA gene copy number using a standard curve (Bustin et al., 2009). This is refereed to as absolute quantification.

A standard curve was constructed ranging from 10 to 100.000 genome copies of the 16S gene in *Escherichia coli*. The amplification and standard curve can be found in Figure 1 and in Figure 2 in the attachments.

The number of *E. coli* gene copy equivalents, or gene copies for short, of all 78 samples was measured with qPCR of the 16S gene. This was done using the BactQuant assay designed by Liu et al. (2012) on a LightCycler480 (Roche Diagnostics, Oslo, Norway) for universal bacterial quantification. Two parallels of each sample were

measured, in addition to three parallels of the standard solution with known concentration of *E. coli*. The mean value for each sample was used for further analysis.

2.1.5 Statistical analysis

The dataset of DNA concentration measured by Qubit and qPCR was analysed by subtracting the mean value of the two samples without any storage time per sampling day from the four other parallels. Hence, all values displays the change in DNA concentration relative to the samples not stored.

Model of DNA concentration by linear mixed models

Linear mixed models are linear models containing both fixed and random effects. Eisenhart (1947) found that there are two fundamentally different explanatory variables: fixed and random effects. Fixed effects affects the response variable in a non-random manner, for example if storage time is used as an predictor variable for DNA concentration. On the other hand, random effects affect the response variable in a random way that can not be manipulated in the experiment. For example, the effect of sampling day in this study, where the weather, temperature or other factors can give random effects to the DNA concentration in the sample. In general; fixed effects influence the mean of the response variable, while random effects influence the variance. This is variation from differences between the levels of the random effects. (Crawley, 2013). Therefore, a key point is to estimate how much of the total variation is from the random effects in the model.

Equation 2.1 shows a mixed model for DNA concentration (y_k) explained by the two fixed effects storage time (α) and method (β) and the random effect (D_k) of sampling day k = 1,2,..,13. μ is the change in concentration of a sample stored for three months as filter, hence α is change when the fixed effect of storage time is changed from three to seven months, while β is the effect of changing storage method from filter to buffer. As said, there are two types of variation in the a model with random effects:

- σ^2 is the overall variation in the model
- σ_D^2 is the variation from differences between the levels of the random effect

$$y_k = \mu + \alpha + \beta + D_k + e \tag{2.1}$$
$$D_k = N \sim (0, \sigma_D^2)$$
$$e = N \sim (0, \sigma^2)$$

A linear mixed effect model was stated for each of the two methods of measurements by the *lme4* package (Bates et al., 2015) in R, referred to as Qubit model and qPCR model. DNA concentration for the two models refers to change in Qubit concentration after storage in the Qubit model, and change in gene copy equivalents in the qPCR model.

95 % confidence intervals were calculated for the data to test for statistically significant differences between the groups.

2.1.6 Model simulation by bootstrapping

To get more robust model estimates, bootstrapping can be used to simulate alternative datasets by re-sampling from the original dataset (Rodgers, 1999). While non-parametric bootstrapping methods re-samples from the original data, semi-parametric methods instead re-samples from the residuals of the data. This gives a better view of how the parameter estimates are affected by the underlying dataset.

Semi-parametric bootstrapping for mixed models in the *bootMer* function in the *lme4* package was used in this work. 10.000 simulations were conducted, only including the fixed effects of storage time and storage method.

2.1.7 Freeze-thaw cycles

To test for a possible effect of freeze-thawing, five repeated cycles of freezing and thawing were conducted. Isolated DNA samples were thawed at room temperature and frozen at -20°C, extracting 15 µL between each cycle for measurement by Qubit.

2.2 Comparing tools for taxonomic profiling

The second main aim in this work was to better understand how the taxonomic profiles of bioaerosol samples are affected by the selection of metagenomic classification tools. The performance of the three tools Kraken 2, One Codex and Kaiju were evaluated on test sets, and on real aerosol samples.

2.2.1 Shotgun-sequencing

Since the first high throughput sequencing in the mid-2000s, the technology has evolved dramatically (Goodwin et al., 2016). There are currently two main paradigms in NGS; short and long read sequencing. The increased read length from long-read sequencing is preferable for genome assembly projects. On the contrary, short read sequencing introduces a more challenging assembly step, but has a lower cost and a higher throughput of reads. More reads increases the chances of detecting rare taxa, which is beneficial for the purpose of this study.

While there are many NGS technologies, Illumina tends to be the most widely used in the realm of short read sequencing (Goodwin et al., 2016). HiSeq2000 and MiSeq are leading options among these. As was stated by Caporaso et al. (2012), these platforms successfully recaptures known biological composition in the tested microbial communities. The two technologies use similar chemistry and produce similar data, but at different scales, making the applications differ. HiSeq produces more reads at a lower cost per read, and is better suited for large projects with larger time scales. MiSeq is better fitted for smaller projects, producing fewer reads with higher quality (Caporaso et al., 2012).

The fragments created in shotgun sequencing can either be sequenced from the 5'-end, called single-end, or from both sides (5'-end and 3'-end), called paired-end. The nucleic acid sequence of each fragment is called a read, with a length measured in numbers of base pairs (bp) (Goodwin et al., 2016). Quail et al. (2012) showed that using paired reads on the Illumina MiSeq gave a strong positive effect on the coverage of bp.

In this work both real and synthetic sequence read sets were used:

- Illumina X Ten HiSeq 150 bp paired-end sequence reads sets (2 samples) for metagenomic profiling of mock microbial community samples.

- Illumina MiSeq 150 bp paired-end sequence reads sets (6 samples) for metagenomic profiling of bioaerosol samples from a subway environment.

- Simulated MiSeq 250 bp paired-end sequence read set (1 test dataset and 2 shuffled test datasets) from downloaded genomes.

2.2.2 Metagenomic classification tools

The decision of which tools to test was based on the results of the newly published article by Gardner et al. (2019), assessing four former benchmarkings. Kraken and One Codex were amongst the tools consistently ranked on top. Kaiju was selected due to a proposed increased accuracy for metagenomic samples with a large proportion of novel organisms.

Lowest common ancestor

The lowest common ancestor (LCA) approach is used by many metagenomic classification tools. LCA determines at what taxonomic level a read should be assigned to when multiple taxa matches (Bender et al., 2005). This could be either on the same or different taxonomic levels. The taxonomic hierarchy can be viewed as a directed graph with taxa as the nodes. The LCA-approach will then output the deepest node connecting the taxon. For example, if a read matches both the species *E. coli* and the genus *Yersinia*, their least common ancestor will be the family *Enterobacteriaceae*. This is illustrated in Figure 2.3. All metagenomic classification tools in this work uses the LCA approach.

Kraken

Kraken is a metagenomic classification tool using exact matching of k-mers to gain a high accuracy, created by Wood and Salzberg (2014). K-mers are all possible subsequences of length K of the original sequence. A sequence of length L have L-k+1 possible k-mers. For example a sequence of length 8 would have 5 4-mers (8-4+1 = 5), as illustrated in Figure 2.4.

The algorithm used in Kraken created k-mers of a given read, and searches against a database also stored as k-mers with exact matching. Each k-mer is then assigned to a taxa with the LCA-approach already described. The number of times a k-mer is assigned to a LCA taxa for a given read is counted, and the entire read is classified as the most

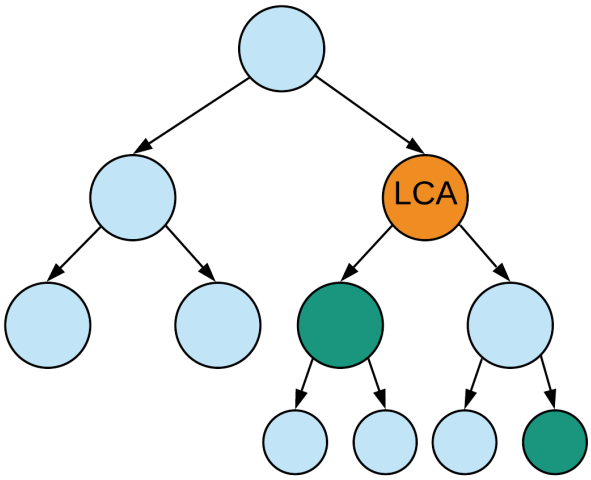


Figure 2.3: The lowest common ancestor-approach illustrated by a directed graph, with taxa as the nodes. The green nodes represents taxa on different taxonomic levels that both matches to the same read. The lowest common ancestor will be the classified taxa (orange node).

abundant taxa. If there is no exact match, the k-mer will be assigned 'unclassified'. The default k-mer length is 31-mers.

Sequence	A	T	T	C	G	A	A	T
1)	A	T	T	C				
2)		T	T	C	G			
3)			T	C	G	A		
4)				C	G	A	A	
5)					G	A	A	T

Figure 2.4: The five possible 4-mers for a DNA sequence of 8 bases. Exact matching with k-mers is used to increase speed for many metagenomic classification tools.

Kraken 2

Kraken 2, which is a newer version of Kraken, was used in this work. The major improvement from the original Kraken is that the database is stored as minimizers instead of entire k-mers. This further increases the classification speed (Wood and Salzberg, 2014). Minimizers are used to bin the k-mers without disrupting the contiguity of the sequence. A minimizer is the first subsequence of a k-mer after sorting the subsequences

alphabetically, as illustrated in Figure 2.5 (Hunt et al., 2004). A given sequence only have one minimizer, and hence will always go to the same bin.

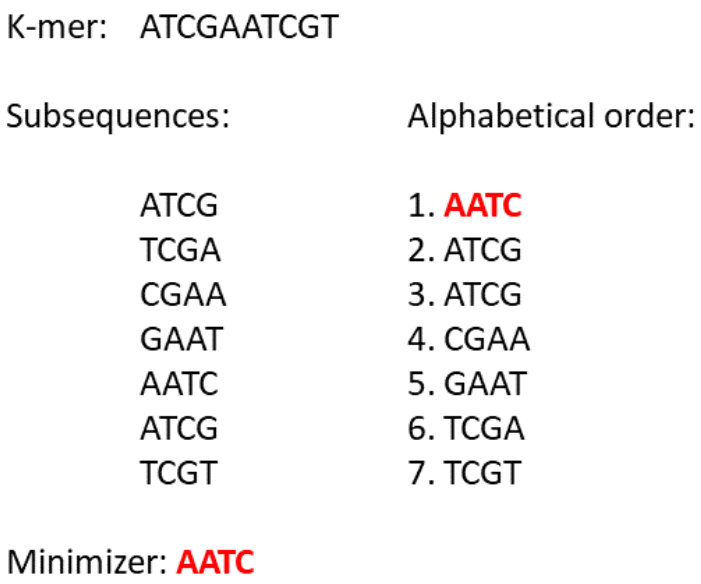


Figure 2.5: Illustration of minimizers, which is used by Kraken 2. The minimizer of a k-mer is the first of the subsequences after sorting all possible subsequences alphabetically.

The classification speed is increased because a given k-mer is first translated into a minimizer, and mapped to the matching bin. In the next step, the content of the bin can be loaded, and the query k-mer can be mapped to the k-mer with the exact match, without loading all other possible bins. In addition to minimizers, Kraken 2 has implemented a compact hash table for storing the database, further reducing the memory requirements (Wood and Salzberg, 2014).

Another benefit of Kraken 2 is the easily modifiable databases. It is possible to add new genomes, or remove parts of the existing one. The latter is sometimes used to serve as a negative test when assessing specificity.

To adjust the stringency of the classification, a confidence score can be set for Kraken 2 in the interval [0,1]. The score (equation 2.2) of 1 is the strictest possible setting. A sequence with a score lower than the confidence score will be labelled "unclassified".

$$\frac{C}{Q} = \frac{\text{k-mers mapped to the LCA values in the clade rooted at the label}}{\text{k-mers in the sequence that lack the ambiguous nucleotide}} \tag{2.2}$$

Kraken 2 was in this benchmarking assessed with the default setting of 31-mers and confidence level = 0.0. The standard databases for archaea (RefSeq), bacteria (RefSeq),

fungi (RefSeq), human (GRCh38) and viral (RefSeq) were used. In addition UniVec_core was used, which is a database containing vector, adapter, linker, and primer sequences that may be a source of contamination.

One Codex

Similarly to Kraken, One Codex uses a k-mer based algorithm. The main advantages of this taxonomic classification tool is the use of a large database of microbial reference genomes and being user friendly for non-expert users (Minot et al., 2015). The algorithm used by One Codex is the same as the first version of Kraken by Wood and Salzberg (2014), as well as the default of 31-mers. One Codex is fully Web-based. The tool is semi-commercial, in that up to 25 samples can be uploaded for free.

Kaiju

Kaiju is fundamentally different from Kraken and One Codex. While most super fast metagenomic classifiers use a k-mer based algorithm at a nucleotide-level, Kaiju instead finds maximum exact matches (MEMs) at the protein-level. It is possible to allow mismatches using a greedy mode. All six possible reading frames of a given sequence are compressed with the Burrows Wheeler Transform (BWT).

The BWT lists all possible rotations of a given string in the reference database, and sort them alphabetically. The last character of each of the sorted rotations make up the transform. An example of a transform is shown in Figure 2.6. A benefit of transforming the strings with BWT is that there is now a set ordering that can easily be used to search up similar sequences. The transform is reversible. The BWT is done for all six possible reading frames of an amino acid sequence from the original nucleotide sequence.

Protein-level is used to increase accuracy, as the degree of conservation at protein-level is higher (Menzel et al., 2016). If there are several matches to sequences contained in the database, Kaiju will output the taxonomic identifier, or determine the LCA.

Since k-mer based methods needs at least one k-mer per read to match to sequence contained in the database, these tools work best for samples where the majority of the diversity already are sequenced and stored in the databases. In environmental samples, and aerosol samples in particular, there are a large proportion of novel sequences not yet contained in any databases. Therefore, it is possible that Kaiju will be well suited to

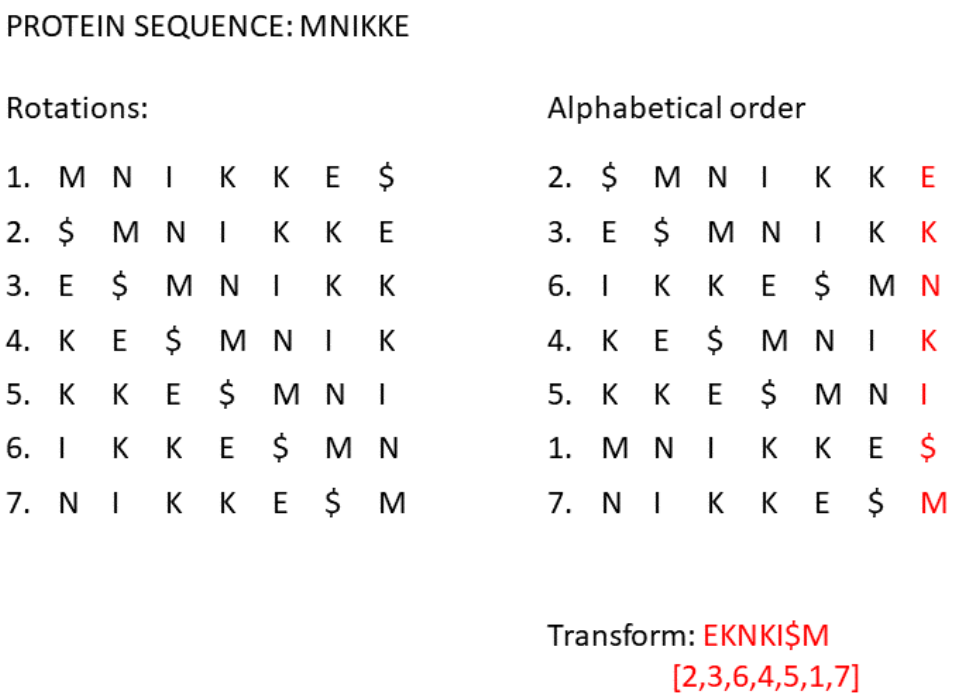


Figure 2.6: Example of a Burrows Wheeler Transform compressing the six first amino acids of a protein sequence. All possible rotations of the sequence are listed and sorted alphabetically. The last character for the rotations in this order is the Burrows Wheeler Transform.

aerosol samples. Menzel et al. (2016) claims that Kaiju can classify up to 10 times more reads in real metagenomes compared to other tools.

Kaiju can either be downloaded locally or be used through a Web Server. The latter was done in this work. The default greedy mode was used with the minimum match score = 75 and allowed mismatches = 5. The non-redundant National Center for Biotechnology Information (NCBI) Basic local alignment search tool (BLAST) database including fungi and microbial eukaryotes was used.

2.2.3 Selecting control datasets

According to Gardner et al. (2019), selection of a adequate control dataset for benchmarking of metagenomic tools is essential, and should be given considerable thought. A functional control dataset needs to have a known composition, and should also resemble real data as closely as possible. There are two main types of positive control datasets: In vitro datasets that are real samples with predetermined ratios of taxa that have been

sequenced, while in silico datasets can either be downloaded from publications, or they can be simulated using a variety of tools (Gardner et al., 2019). To truly represent an environmental sample, part of the dataset should consist of novel sequences that is not contained in the databases used. Both categories of control datasets were used for this work.

2.2.4 ZymoBIOMICS dataset

As in vitro datasets, the sequence data of two samples of ZymoBIOMICS Microbial Community Standard was obtained from FFI. The standard consist of eight species of bacteria and two species of yeasts in known proportions, shown in Table 2.1. All species in the ZymoBIOMICS dataset are well-studied species expected to be contained in the databases of metagenomic classification tools. A phylogenetic tree of the ten species is shown in Figure 2.7 as a visualisation of the evolutionary relationships between the species. The method used for creating the phylogenetic tree will be discussed in section 2.2.7.

Table 2.1: The microbial composition of the ZymoBIOMICS Microbial Community Standard used as a test dataset.

Organism name	**Proportion**
Listeria monocytogenes	12 %
Pseudomonas aeruginosa	12 %
Bacillus subtilis	12 %
Escherichia coli	12 %
Salmonella enterica	12 %
Lactobacillus fermentum	12 %
Enterococcus faecalis	12 %
Staphylococcus aureus	12 %
Saccharomyces cerevisiae	2 %
Cryptococcus neoformans	2 %

2.2.5 Simulating control datasets

Considering the novelty of the field of metagenomics on aerosol samples, finding an excising well-tested dataset reflecting the contents of air proved to be problematic. Instead, as an in silico dataset of synthetic metagenomes were created for this study by using the simulation tool ART on genomes known to be present in aerosol samples. ART is developed to improve testing and benchmarking of tools (Huang et al., 2012).

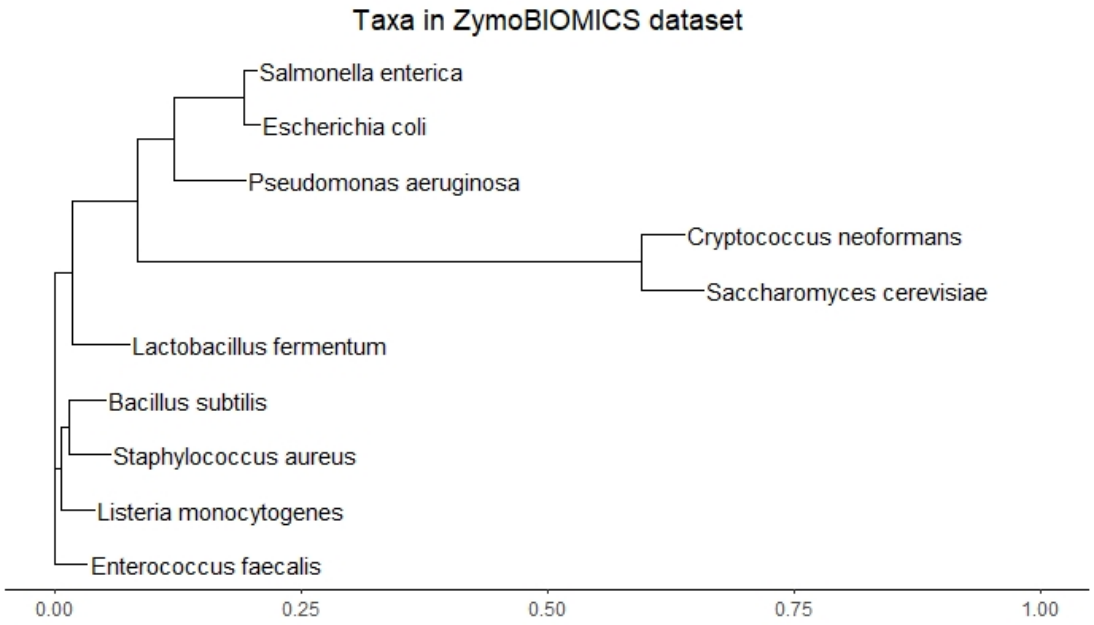


Figure 2.7: Phylogenetic tree of taxa in the ZymoBIOMICS community standard dataset used to visualise the evolutionary relationship between them. The x-axis shows Tamura and Nei-distances, and a higher distance compares to a greater evoltionary distance.

Mimicking of error profiles for several sequencing technologies are supported, amongst them are the Illumina MiSeq 250 bp paired-end sequencing. The same technology is used for sequencing of aerosol samples at FFI except from the read length of 150 bp.

The selection of genomes for the synthetic metagenome was based on the results from Tringe et al. (2008). The study collected aerosol samples at two shopping centres in Singapore, and the 16S ribosomal DNA was sequenced. The genomes of the 12 most abundant species, based on the number of phylogroups from the study was downloaded from the RefSeq database by NCBI (O'Leary et al., 2015). The chosen species are confirmed as species found in aerosol samples by FFI.

A total of 1.000.000 read pairs were simulated from the genome of the most abundant of the 12 species, and 500.000 for each of the remaining 11 genomes. The magnitude of this simulation was selected to ensure that the dataset was simulated from a large enough pool of reads. In addition, reads from the human genome was simulated as a contaminant control. The simulated metagenome was then created by randomly sampling simulated reads in the same proportions of species as the phylogroups found in Tringe et al. (2008). This is illustrated in Figure 2.8. In total this made up 1.000.000 read pairs. Table 2.2 shows an overview of the species and composition of simulated metagenome, and figure 2.9 shows a phylogenetic tree of the species in the dataset. The method used for creating

the phylogenetic tree will be explained in section 2.2.7. The simulated dataset was utilised as a part of the evaluation of Kraken 2, One Codex and Kaiju.

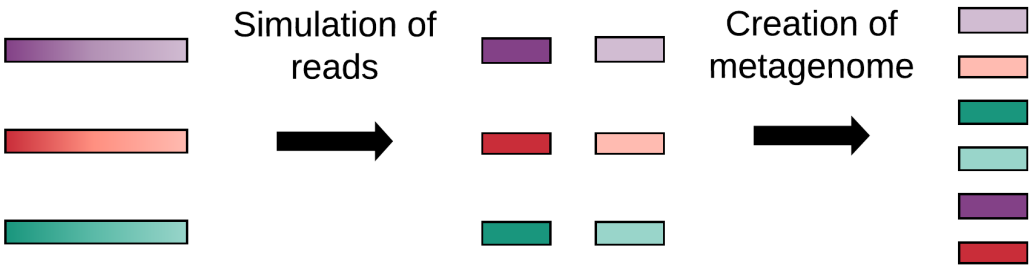


Figure 2.8: Illustration of creation of simulated in silico metagenome dataset. Paired-end reads are simulated from complete genomes obtained from the RefSeq database using ART simulation tool. Reads are then randomly sampled in a given proportion to make up a metagenome.

Table 2.2: The composition of the simulated metagenome made of genomes from the RefSeq database. There are in total 1.000.000 simulated read pairs.

Organism name	**Number of simulated read pairs**	**Proportion of simulated metagenome**
Brevundimonas abyssalis	523459	52,3 %
Stenotrophomonas acidaminiphila	201840	20,2 %
Brachybacterium alimentarium	66421	6,6 %
Acinetobacter apis	60534	6,1 %
Methylobacterium aquaticum	27047	2,7 %
Microbacteriaceae bacterium	25575	2,6 %
Micrococcaceae bacterium C1-50	15087	1,5 %
Sphingomonas adhaesiva	10304	1,0 %
Sphingobacterium cellulitidis	9016	0,9 %
Massilia alkalitolerans	8648	0,9 %
Bacillaceae bacterium B16-10	8464	0,8 %
Janthinobacterium agaricidamnosum	6808	0,7 %
Homo sapiens	36799	3,7 %

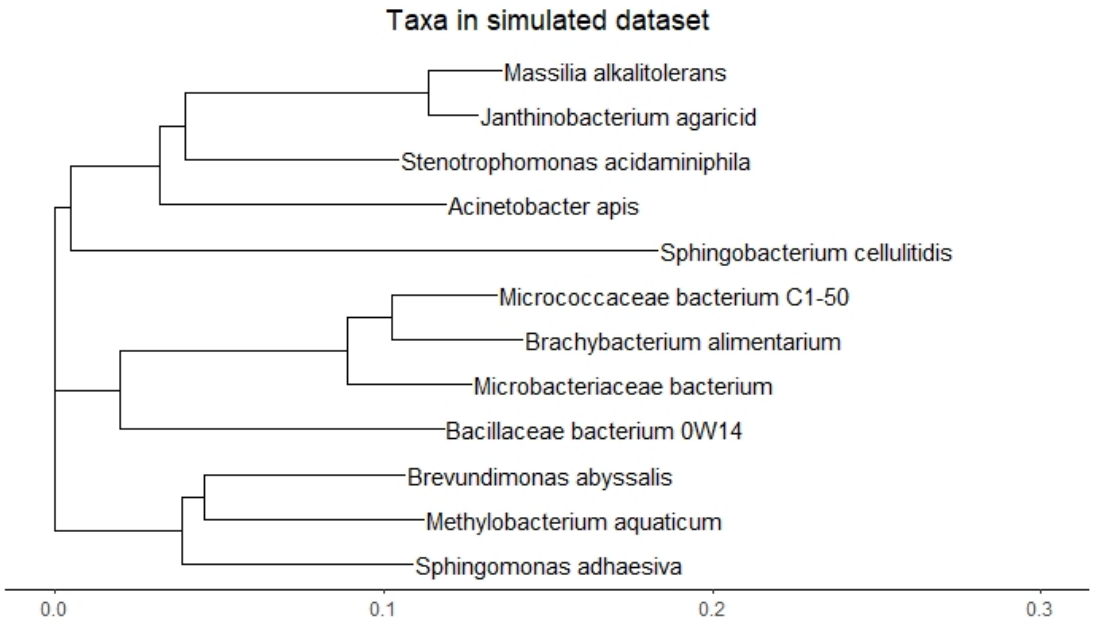


Figure 2.9: Phylogenetic tree of taxa in the simulated metagenomic dataset. The x-axis shows Tamura and Nei-distances. A greater distane makes a greater evolutionary distance.

2.2.6 Simulating negative control datasets

As emphasised by Gardner et al. (2019), an adequate test dataset should also mimic the proportion of unknown reads. There are several strategies for this. One approach is to remove sequences from the reference database used by the tools, as done by Peabody et al. (2015). Almeida et al. (2018) instead simulated random mutations in 2 % of the bases in the sequences, while Lindgreen et al. (2016) shuffled genomes using the shuffle program from HMMER (Eddy, 2011) to simulate unfamiliar DNA. For the purpose of this study, the latter approach was chosen for simulation of negative reads. This choice was made because both One Codex and the web server version of Kaiju has no option for excluding parts of the database.

The shuffling method used by Lindgreen et al. (2016) makes the original sequence totally disrupted, thus no longer resembling actual DNA. To maintain more of the sequence integrity, another shuffling method was made for the purpose of this thesis. By dividing the genome into sub-sequences of a given length, referred to as window length, these sub-sequences can be shuffled. The shuffled sequence will be increasingly disrupted with a decreasing window length.

To choose a window length appropriate for detection of false positives, a selection of lengths ranging from 20 to 250 bp were tested by shuffling the reference genome of *Acinetobacter apis* from the RefSeq database. Distance from the original genome was

evaluated by MinHash distances from the Mash program by Ondov et al. (2016). This dissimilarity tool is a convenient way of getting an indication of the effect of shuffling. The ideal window length should not completely dismantle the biological composition, but should still differ enough not to be recognised. The test results reviled two window lengths of particular interest; 25 and 50 bp. A negative dataset was created for each of the two window lengths, by sampling 1.000.000 simulated read pairs from shuffled genomes (listed in Table 2.2), as shown in Figure 2.10.

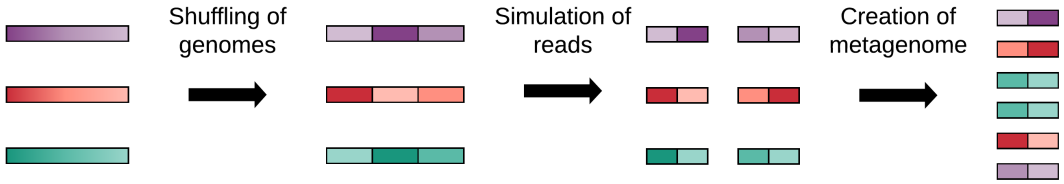


Figure 2.10: Creation of simulated negative dataset. Reference genomes from the RefSeq database are shuffled before read simulation. The shuffled reads are then sampled to make up a metagenome to act as a negative test.

2.2.7 Metrics for comparing metagenomic tools

The selection of metrics is a crucial step when evaluating taxonomic classification tools (Gardner et al., 2019). For classification, the terms negatives and positives are used. When using these terms for metagenomic classification tools, the number of true positives are equivalent to the number of reads classified to the correct taxa. False positives are the number of reads classified to anything else than the true taxa. False negatives are the number of reads remaining unclassified, while true negatives are only relevant when negative control reads are added to the test dataset intended to be unrecognisable by the tools. This can be illustrated with a confution matrix, as seen in Table 2.3.

Common measures are sensitivity and specificity. Sensitivity (also called recall) is the proportion of true positives that are identified as positives. Specificity measures the proportion of actual negatives that are identified as negatives (equation 2.3). Another similar and widely used measure is precision, which is the proportion of true positives amongst all positives.

Table 2.3: General confusion matrix. If both the predicted and actual class are positive or negative, the classification is correct.

		Actual	
		Positive	Negative
Predicted	Positive	True positive	False positive
	Negative	False negative	True negative

$$\text{Sensitivity} = \text{recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \tag{2.3}$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

When selecting metrics, the application the classification tool is intended for has to be taken into account. Some cases require a high sensitivity, which means that there is a high probability of the tool classifying the the given read, even though the classification has a higher chance of being false. Detection of rare species is an example of a situation when a high sensitivity is beneficial. On the other hand, a false positive could have unfortunate consequences in many cases, for example in medical diagnostics. In applications like this, an increased specificity at the expense of the sensitivity is beneficial.

Even though sensitivity, specificity and precision are essential metrics, combining these with more application specific measures could help getting a more profound evaluation of the metagenomic classification tools.

Abundance estimate metrics

Metrics considering abundance estimates for each taxa can be advantageous when evaluating metagenomic tools. A common output from these tools is abundance estimates listing the number of reads for each detected taxa. Metrics used for calculating the distances between metagenomic profiles include Bray-Curtis and UniFrac (Meyer et al., 2019).

Bray-Curtis dissimilarity has been widely used for assessing community similarity in microbial ecology (Beiko and Parks, 2012). The dissimilarity is calculated by dividing the sum of the absolute pairwise differences of taxa abundance i by the sum of all abundances at a given taxonomic level (equation 2.4) (Meyer et al., 2019). x_i is the true taxa abundance, while $x*_i$ is the estimated abundance.

$$\text{Bray-Curis distance} = \frac{\sum_i |(x)_i - (x*)_i|}{\sum_i (x)_i + \sum_i (x*)_i} \tag{2.4}$$

The Bray-Curtis dissimilarity can be viewed as the proportion of dissimilarity between two compositions, bounding the value between 0 and 1. 0 means that the two profiles have identical compositions, while 1 would indicate that the profiles do not share any taxa.

UniFrac is a distance measure that also includes the genetic distance in the calculation (Beiko and Parks, 2012). By adding a phylogenetic tree, the weighted UniFrac measures the total amount of predicted abundances that must be moved to overlap with the true abundances by branch length (Meyer et al., 2019). On the contrary, the unweighted UniFrac distance only measures the overlap of presence and absence of taxa in the true and estimated profiles, and is hence not used in this study.

Bray-Curtis dissimilarities and UniFrac distances are implemented in the *Vegan* (Oksanen et al., 2019) and *GUniFrac* (Chen, 2018) R-packages respectively, and were used to evaluate both the ZymoBIOMICS and simulated dataset.

Creating phylogenetic trees

Phylogenetic trees shows the evolutionary relationship between taxa, and is used for calculation of UniFrac-distances. 16S- and 18S-rRNA sequences were downloaded from the Silva database (Quast et al., 2012), and multiple sequence alignments (MSA) using Clustal Omega by Sie (2011) were created. Tamura-Nei distances (TN93) were then calculated, which is a model of DNA evolution accounting for the difference between transitions and transversions (Tamura and Nei, 1993). Neighbour joining was used to cluster the taxa. Phylogenetic trees were plotted with the *ggtree* (Yu et al., 2017) package for both the ZymoBIOMICS and the simulated dataset.

2.2.8 Metagenomic profiles of aerosol samples from Nationaltheatret subway station

Metagenomic profiles of aerosol samples collected at Nationaltheatret subway station was made by using all tested tools. The sampling was conducted by FFI at three locations (at the platform inside the subway station at day time, night time and outside the station building at day time). Two parallel samples were taken at each location, resulting in six bioaerosol samples. The samples were sequenced with Illumina Miseq 2x150 bp paired end sequencing. The sample names Di1 and Di2 refers to the parallel samples collected inside the station building at day time and Du1 and Du2 refers to the samples collected outside the station building at day time, while Ni1 and Ni2 refers to the samples taken inside at night time.

Mash distances (described in section 2.2.6) between the six metagenome samples were calculated to investigate their similarity. Sketches of genomes can be used to make the comparison less computationally demanding. For the purpose of this thesis, sketches were made with a k-mer length of 21, and sketch size of 10.000, as done by Ondov et al. (2016) for metagenomes. A dendrogram using average linkage was created.

3 | Results

3.1 Storage study

To investigate how the DNA concentration is affected by storage, quantitative measurements by Qubit and qPCR were conducted. The raw data visualised as boxplots in Figure 3.1 shows a larger decrease in DNA concentration for Qubit measurements of buffer storage relative to filter storage, seen by the group medians. This is true for both storage times, but with a larger spread of data for the samples with seven months storage. The qPCR-measurements (Figure 3.1 B), shows that samples with three months buffer storage has a marginally higher median than the filter stored samples. This contradicts the trend observed with Qubit measurements. However, the group of samples stored for seven months as filter shows the same relatively large spread as in Figure 3.1 A, as seen by the 75th and 25th percentile containing a wider interval.

Box plots grouped by the sampling day (Figure 3.2) shows that this variable contributes to variation in the data. Each of the groups consists of six parallel samples, which has a group medians in the range 0.1-0.3 ng/µL for Qubit and 2500-10.000 gene copies for qPCR, which is considered large in this context.

3.1.1 Statistical analysis

The parameter estimates for the two mixed linear models (Qubit and qPCR) in equation 3.1 both show negative β estimates, which means an estimated decrease in DNA concentration from buffer storage. However, the 95 % confidence intervals in Table 3.1 shows that there are no significant differences between the parameter estimates, since none of the intervals contain 0. Hence, the hypothesis of no difference between the groups can not be rejected. The original Rstudio output is presented in Listing A.1 and Listing A.2 in the attachments.

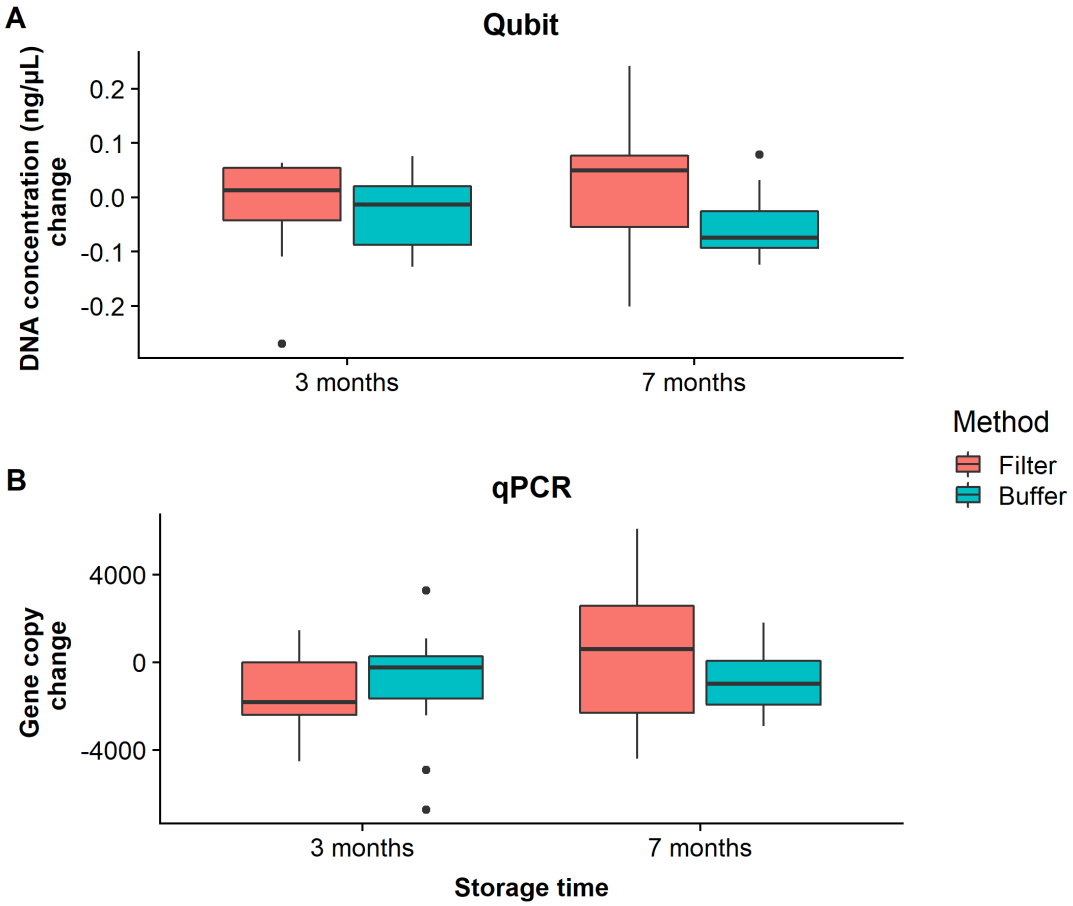


Figure 3.1: DNA concentration Measurements by Qubit and qPCR represented by box plots grouped by storage time (3 and 7 months) and storage method (filter and buffer). The values are difference compared to the parallel samples without storage.

$$y_k = \mu + \alpha + \beta + D_k + e \tag{3.1}$$
$$D_k = N \sim (0, \sigma_D^2)$$
$$e = N \sim (0, \sigma^2)$$

Qubit model estimates:	qPCR model estimates:
$\hat{\mu} = 0.0021$	$\hat{\mu} = -918.6$
$\hat{\alpha} = 0.0046$	$\hat{\alpha} = 841.7$
$\hat{\beta} = -0.0457$	$\hat{\beta} = -379.0$
$\hat{\sigma_D^2} = 0.054$	$\hat{\sigma_D^2} = 1506$
$\hat{\sigma^2} = 0.069$	$\hat{\sigma^2} = 1849$

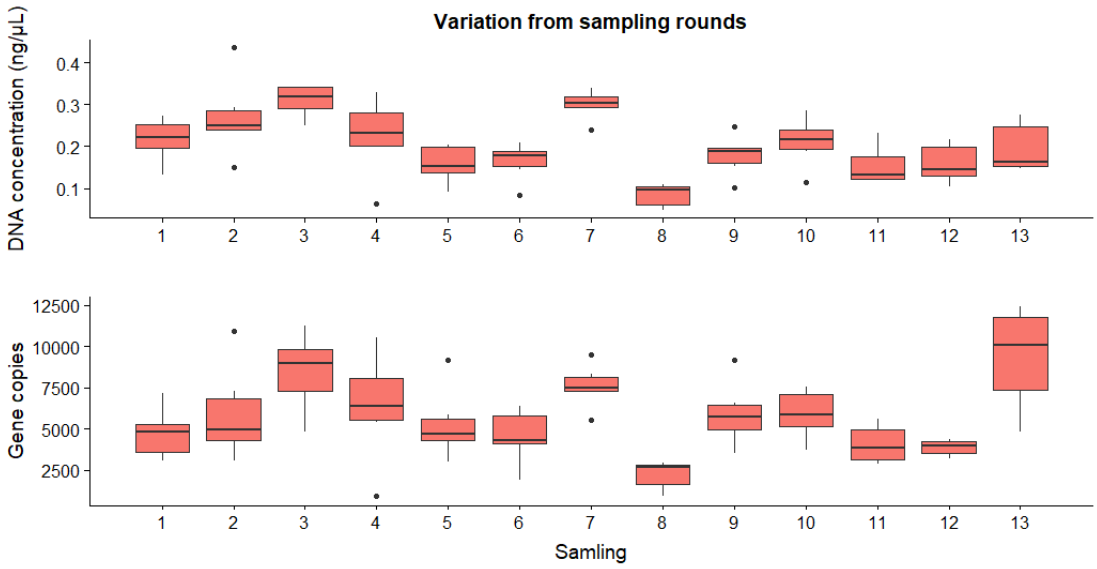


Figure 3.2: Variation in measured DNA concentration between 13 sampling days. Each day contains n = 6 measurements. The upper plot shows DNA concentration measured by Qubit, and the lower qPCR measurements in number of gene copy equivalents.

Table 3.1: Confidence intervals for Qubit and qPCR model.

	Qubit model		qPCR model	
	2.5 %	97.5 %	2.5 %	97.5 %
$\hat{\mu}$	-0.051	0.056	-2397	560
$\hat{\alpha}$ (Stored for seven months)	-0.035	0.042	-175	1859
$\hat{\beta}$ (Stored as buffer)	-0.115	0.024	-2291	1533

$$\text{Qubit model} = \frac{\hat{\sigma_D}}{\hat{\sigma_D} + \hat{\sigma}} = \frac{0.054^2}{0.054^2 + 0.069^2} = 0.38 = 38\% \tag{3.2}$$
$$\text{qPCR model} = \frac{\hat{\sigma_D}}{\hat{\sigma_D} + \hat{\sigma}} + \frac{1506.2^2}{1506.2^2 + 1849.2^2} = 0.40 = 40\%$$

A noteworthy result from the mixed model estimations are the estimated proportions of variation explained by sampling day, which are are 38 % and 40 % of the variation for Qubit and qPCR respectively, as shown in equation 3.2, suggesting that the sampling should be included as a random effect.

3.1.2 Model simulation by bootstrapping

The model estimates from bootstrapping, shown in Figure 3.3, shows that filters stored for seven months are estimated to have a change in DNA concentration close to 0 by both measurement methods. Further, the density plots indicate a difference in DNA concentration between storage method measured by Qubit. There is no detectable difference between groups of samples with different storage times (three and seven months) for the Qubit parameter estimates. The qPCR measurements however, show a separation of density curves for storage time, where samples that are stored for three moths has a decreased yield compared to samples that are stored for seven months.

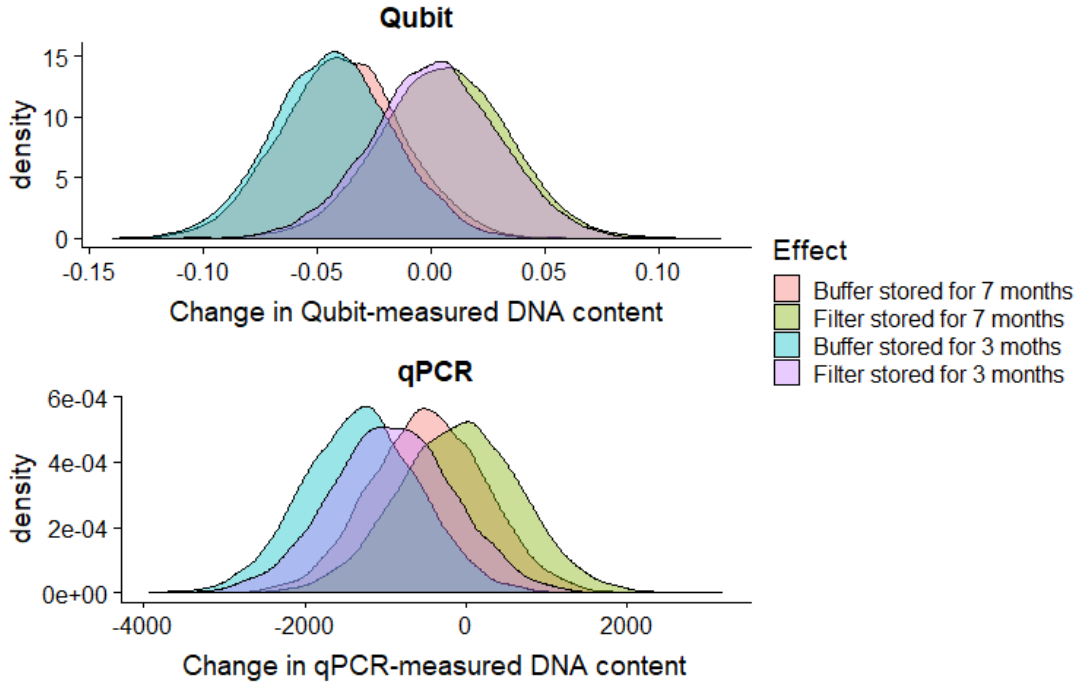


Figure 3.3: Density plots of group estimates from mixed models of DNA concentration change after storage by qPCR and Qubit. The estimates are made by bootstrapping with 10000 simulations, and hence creating 10.000 sets of group estimates.

3.1.3 Freeze-thaw cycles

The freeze-thawing included in the buffer storage preparation procedure was suspected as a cause for the observed difference in Qubit measurements. Therefore the effect of freeze-thaw cycles were exclusively tested on four isolated DNA samples. The concentrations measured by Qubit is shown in Figure 3.4. The values indicates a decrease after repeated cycles of freezing and thawing.

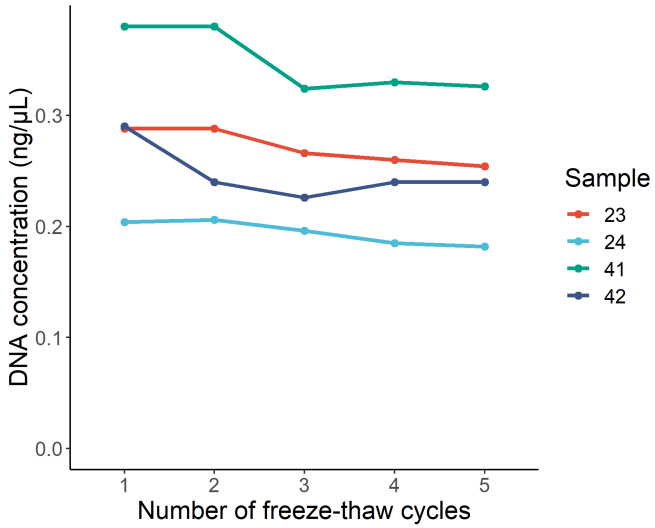


Figure 3.4: DNA concentration measured by Qubit of four DNA samples freeze-thawed from one to five times.

3.2 Comparing tools for taxonomic profiling

Kraken 2, One Codex and Kaiju were selected as metagenomic classification tools for assessment of performance on aerosol samples specifically. Some main differences found when comparing the features of the three tools include Kaiju searching against the reference database at protein-level with MEMs, and the Kraken 2 and One Codex implementing k-mers in their classification algorithm. One Codex has a larger database than Kraken 2, while they are both difficult to compare with the database of Kaiju, which stores the entries as protein sequences, and not as genomes. The major differences between the metagenomic tools Kraken 2, One Codex and Kaiju in terms of features are shown in table 3.2.

Table 3.2: The major differences between the taxonomic classification tools Kraken 2, One Codex and Kaiju. *As of April 2019.

	Kraken 2	One Codex	Kaiju
Sequence comparison level	DNA	DNA	Protein
Algorithm	K-mer with minimizer	K-mer	MEMs with BWT
Default K-mer length	31-mer	31-mer	Not fixed, using MEMs
Database size*	∼42.000 genomes	∼83.000 complete genomes	∼103.000.000 protein sequences
Special feature	Highly ranked in benchmarks	Large and manually curated database	Protein-level comparison giving increased sensitivity

3.2.1 ZymoBIOMICS dataset

The estimated taxa abundances in Figure 3.5 shows a difference between the species and genus-level for the three metagenomic classification tools. The abundance of reads classified to "other species" (referring to other than the taxa in the dataset) is larger at species-level compared to genus-level for all tools. Kraken 2 is prominent, with the abundance estimates resembling the real abundances at a higher degree than Kaiju and One Codex. For the exact abundance estimates, see Table 1 in the attachments.

Kraken 2 has the highest precision and recall (Figure 3.6 and Table 2 in the attachments) for both genus and species-level for sample A and B. Kaiju has a similar recall as One Codex, but a lower precision. The Bray-Curtis dissimilarities and UniFrac distances also shows that Kraken 2 is closest to the true composition of the dataset (3.3).

Table 3.3: Calculated Bray-Curtis dissimilarities and weighted UniFrac distances from true metagenomic profiles of ZymoBIOMICS standard to estimates derived from three different metagenomic classification tools.

		Genus-level		Species-level	
Tool	Sample	Bray-Curtis	Weighted UniFrac	Bray-Curtis	Weighted UniFrac
Kraken 2	A	0.108	0.0580	0.161	0.0766
	B	0.128	0.0560	0.175	0.105
One Codex	A	0.196	0.0830	0.553	0.161
	B	0.263	0.0865	0.557	0.143
Kaiju	A	0.284	0.0936	0.597	0.123
	B	0.276	0.104	0.599	0.102

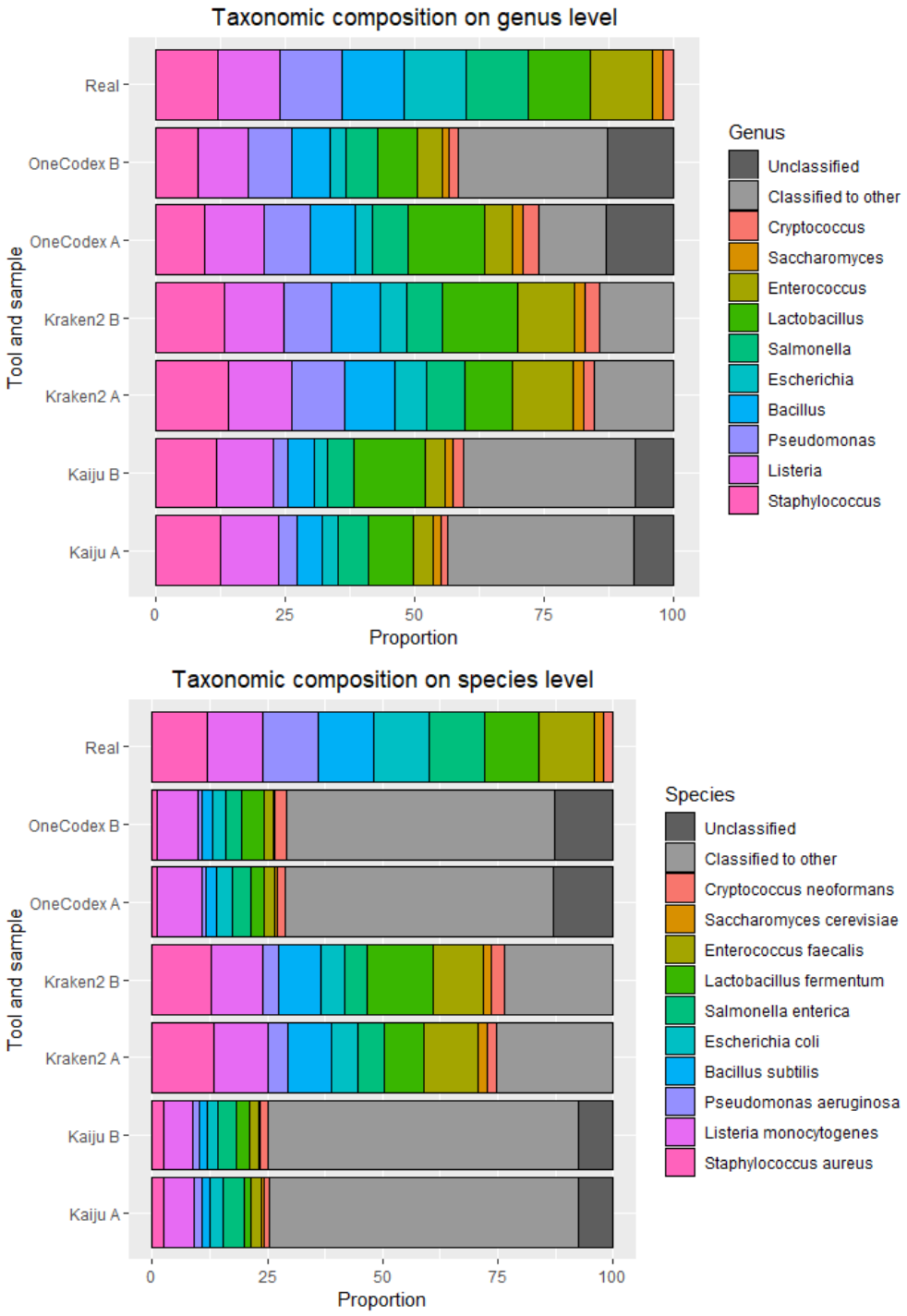


Figure 3.5: Real and estimated taxa abundances for profiles of sample A and B of ZymoBIOMICS standard by Kraken 2, One Codex and Kaiju. The estimates are presented for genus-level (on top) and species-level (on bottom).

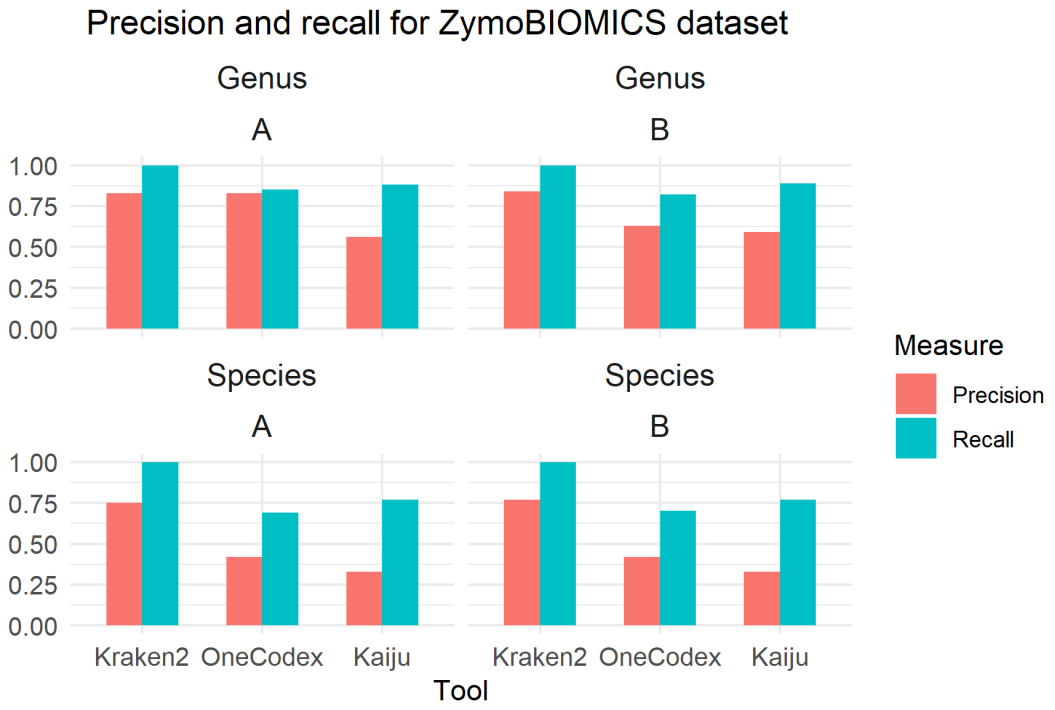


Figure 3.6: Precision and recall for Kraken 2, One Codex and Kaiju on ZymoBIOMICS dataset. On top is the measures on genus-level, while species-level is at the bottom. A and B referrers to two parallel samples.

3.2.2 Simulated metagenome dataset

The simulated dataset was included in the benchmarking to evaluate the performance on a dataset with a more realistic composition of species. The taxonomic profiles (Table 3.4) show that the most abundant species, *Brevundimonas abyssalis*, is underestimated by both Kraken 2 and One Codex which both estimates this taxa to be ~20 % of the total reads, when the actual proportion is ~52 %. Kaiju is the closest to the known composition with ~51 %. As this species makes up approximately half of the reads in the dataset, this error may impact the performance of the tools accordingly. Conversely, estimates of the second most abundant taxa *Stenotrophomonas acidaminiphila* by Kraken 2 and One Codex are closer to the real values than Kaiju. Due to the possibly extensive impact by the first taxa, distances were also calculated without the estimate of the first taxa.

One Codex is the only tool that classifies some reads to all of the species in the dataset. Kraken 2 does not classify any reads to half of the species.

For evaluation of the tools by the simulated dataset, precision and recall were calculated (barplots in Figure 3.7 and exact abundances in Table 3 in the attachments). Kraken 2 has the lowest precision and recall of the three tools at genus and species-level.

Table 3.4: Overview of the real and estimated proportions of taxa on the in silico simulated metagenome dataset. *Taxa classified to family-level.

Taxa	Proportion	Kraken 2		One Codex		Kaiju	
		Genus	Species	Genus	Species	Genus	Species
Brevundimonas abyssalis	52.35	19.43	0	22.26	21.86	50.57	50.39
Stenotrophomonas acidaminiphila	20.18	20	19.8	24.81	24.55	10.37	10.29
Brachybacterium alimentarium	6.64	3.76	0	13.06	13.06	2.76	0
Acinetobacter apis	6.05	2.37	0	12.09	12.09	3.54	0
Methylobacterium aquaticum	2.70	2.36	0.47	3.09	0.14	2.25	2.2
*Microbacteriaceae bacterium**	2.56	0.08	0.08	3.14	3.14	2.74	2.74
*Micrococcaceae bacterium C1-50**	1.51	0.03	0.03	0.54	0.54	1.14	1.14
Sphingomonas adhaesiva	1.03	0.81	0	1.81	1.78	0.75	0.7
Sphingobacterium cellulitidis	0.90	0.25	0	1.75	1.75	0.76	0
Massilia alkalitolerans	0.86	0.47	0	1.56	1.55	0.59	0.57
*Bacillaceae bacterium B16-10**	0.85	0.01	0.01	0.09	0.09	0.69	0.69
Janthinobacterium agaricidamnosum	0.68	0.73	0.7	1.36	1.36	0.65	0.65
Classified to other	0	28.54	57.75	7.61	11.26	18.89	26.33
Unclassified	0	21.16	21.16	6.83	6.83	4.30	4.30

Both Kaiju and One Codex has a high recall score, Kaiju has lower precision than One Codex.

Further the Bray-Curtis dissimilarities and UniFrac distances (Figure 3.8) for both the full and reduced profile without *B. abyssalis* were made. Since not all tools listed *Homo sapiens* in their results, the taxa were removed from the profiles. The taxa abundance metrics show that Kaiju has the most correct abundance estimates at both taxonomic levels on the full dataset. Kraken 2 is by far the least correct tool at species-level due to estimating 0 % on six of 12 taxa. The results from the reduced profile however, show that One Codex has the best estimates at species-level, but ambiguity at genus-level, with no tool standing out in terms of good abundance estimations.

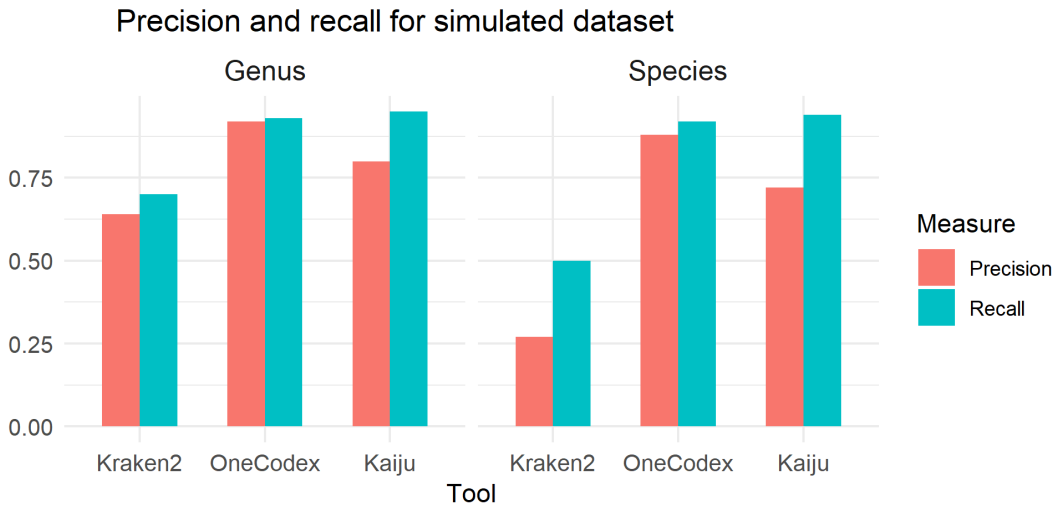


Figure 3.7: Precision and recall measurements for profiles by Kraken 2, One Codex and Kaiju on a simulated dataset containing 12 species found in air on genus and species-level.

Figure 3.8: Performance of Kraken 2, One Codex and Kaiju on a simulated metagenomic dataset. The performance is measured by two distance measures: Bray-Curtis and Weighted UniFrac. Full (left) refers to the full profiles of 12 taxa, while reduced (right) to the most abundant species *B. abyssalis* being removed.

3.2.3 Simulated negative datasets

The two window lengths 25 and 50 bp were chosen for creation of shuffled genomes acting as negative control datasets, by the distances plotted in Figure 3.9. The graph shows a steep increase between 50 and 25 (the two red points) for the mash distances to the original *A. apis* genome. The rapid change indicates that the shuffled genome of length 25 has lost much of the original integrity, compared to results when shuffled with a window length of 50. To see whether this is reflected in the classification results, both lengths were used to create negative datasets.

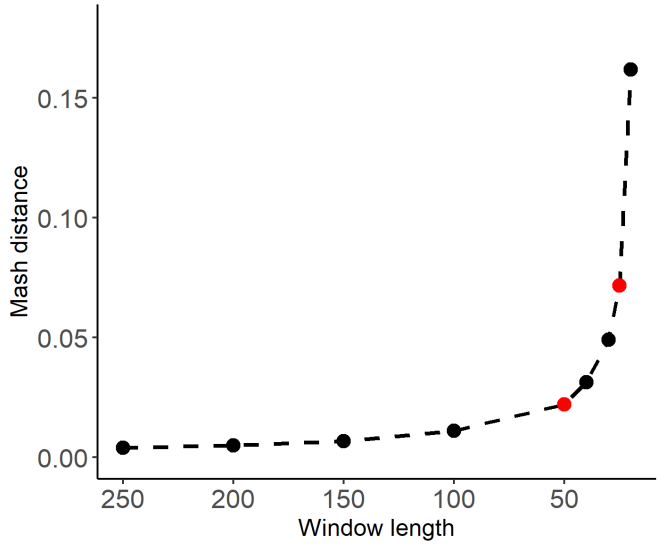


Figure 3.9: Mash distance from the original genome to the shuffled genome of *A. apis* shuffled with different window lengths. The two red points are the window lengths lengths 25 and 50.

The shuffled genomes represents novel taxa in aerosol samples, and is included in the benchmarking for detection of false positives. The results (Table 3.5) shows that Kraken 2 has a relatively low precision of 0.81 on the dataset shuffled with window length 25, compared to 0.99 by the other tools. Amongst the reads classified by Kraken 2, there are none classified to the most abundant species *B. abyssalis*, while about 1 % of the reads are classified to *Brevundimonas* (genus-level), indicating that the reads are mainly falsely classified. The confidence level of Kraken 2 was increased from 0.00 (default setting) to 0.05, which drastically increased to the specificity to 1, indication few false positives.

Table 3.5: Specificity of classification by metagenomic classification tools of datasets simulated from shuffled genomes.

Tool	Confidence level	Window length	
		50	25
Kraken	0	0.54	0.81
	0.05	0.71	1.00
One Codex		0.50	0.99
Kaiju		0.51	0.99

3.2.4 Aerosol samples from Nationaltheatret subway station

When assessing the performance of the metagenomic classification tools on real data from Nationaltheatret subway station, the results show substantial variation between the tools. Figure 3.10 lists the 10 most abundant taxa by read count found by the three tools on both genus and species-level. None of the species listed are found at all three lists, and only four of the genera are consistently found at all three lists. The top 10 most abundant taxa at family-level is listed in Table 3.6, and shows four corresponding taxa, which is the same at at genus-level.

The fraction of unclassified reads are similar between the tools: Kraken 2: 47 % unclassified, One Codex: 58.4 % unclassified and Kaiju: 57.9 % unclassified.

Table 3.6: The ten most abundant families found in samples from Nationaltheatret subway station by Kraken 2, One Codex and Kaiju. Taxa consistently found at the top 10 list by all tools are marked with bold text.

	Kraken 2	One Codex	Kaiju
1	**Micrococcaceae**	Sphingomonadaceae	Enterococcaceae
2	**Propionibacteriaceae**	**Microbacteriaceae**	**Nocardioidaceae**
3	Streptomycetaceae	**Nocardioidaceae**	**Micrococcaceae**
4	**Nocardioidaceae**	**Micrococcaceae**	Mycobacteriaceae
5	**Microbacteriaceae**	Hymenobacteraceae	**Microbacteriaceae**
6	Sphingomonadaceae	Comamonadaceae	Plasmodiidae
7	Corynebacteriaceae	Intrasporangiaceae	**Propionibacteriaceae**
8	Pseudomonadaceae	Geodermatophilaceae	Streptococcaceae
9	Moraxellaceae	**Propionibacteriaceae**	Pseudonocardiaceae
10	Staphylococcaceae	Pseudomonadaceae	Geodermatophilaceae

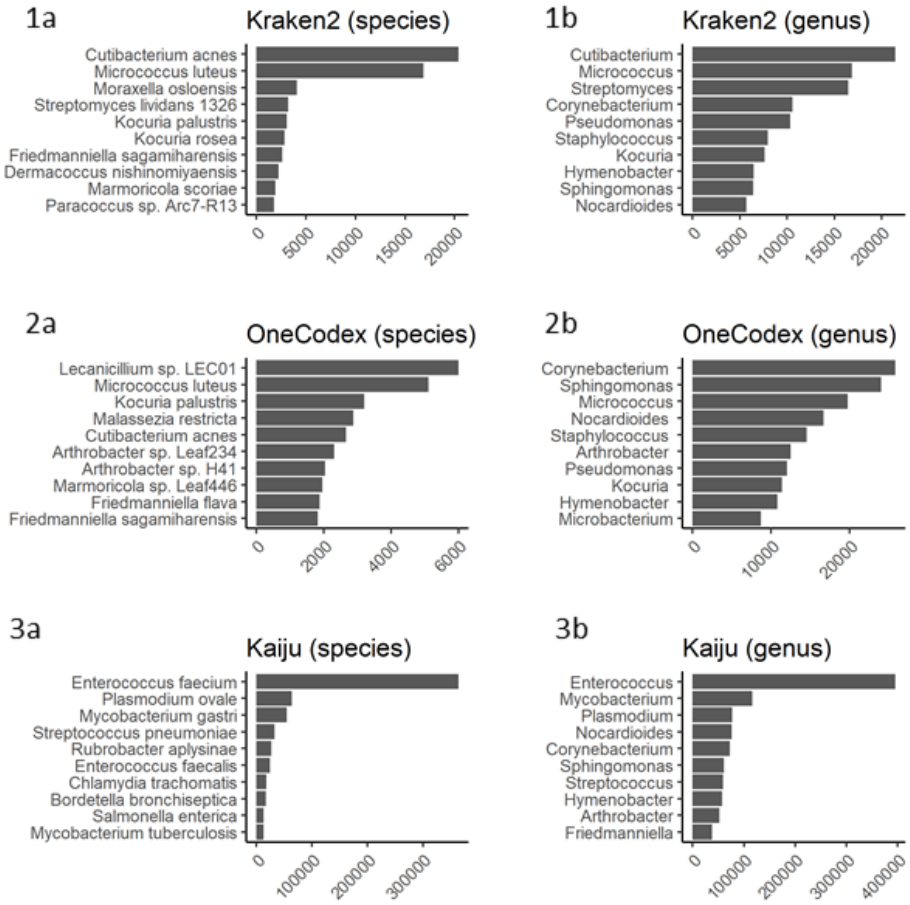


Figure 3.10: The ten most abundant species (a) and genera (b) by read count for profiles made by 1) Kraken, 2) One Codex, and 3) Kaiju in bioaerosol sample from Nationaltheatret subway station.

A distinct feature seen in Figure 3.10 is the abundance of *Enterococcus faecium* classified by Kaiju, with more than than 300.000 classified reads. This is not listed in the top 10 lists of neither Kraken 2 nor One Codex, despite all tools having entries for this species in their databases. Furthermore, the reads classified to *E. faecium* by Kaiju was shown to be mainly classified as "other sequences" and *H. sapiens* by Kraken 2.

The Mash distances between the six bioaerosol samples are overall similar, as seen in Figure 3.11. The parallel samples (Di1 and Di2 etc.) are clustered together, proving that they are most similar. Samples taken at night time are most dissimilar from the samples taken inside at day time. Still, the distances are all in the range 0.149-0.18, which can be considered a low spread in this context.

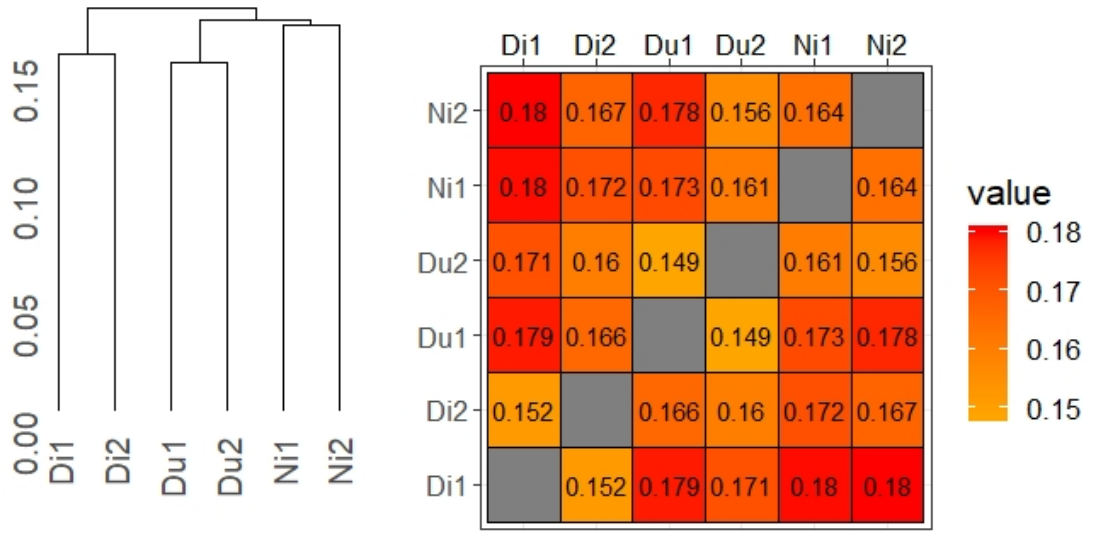


Figure 3.11: Mash distances between six bioaerosol samples from Nationaltheatret subway station represented by a dendrogram (left) and a tile plot (right). The dendrogram is clustered using average linkage.

4 | Discussion

4.1 Storage study

The aim of the storage study was to determine whether current storage procedure at FFI of bioaerosol samples affect the DNA concentration. No evidence was found that suggested a negative effect on DNA concentration from storage time up to seven months. However, freeze-thawing was expected as a cause for the observed difference between the storage methods.

4.1.1 Model simulation by bootstrapping

The density plots from bootstrapping (Figure 3.3) shows that there is a measured difference between storage method by Qubit, but none for storage time. This suggests that the samples are not affected by the storage time itself, but rather the buffer storage. Filter storage has a mean value for change in concentration close to 0, indicating no departmental effect of storage of filters. These robust group estimates are possible interpret because small deviations caused by random variation are removed by sub-setting 10.000 times.

The separation between the density curves by qPCR measurements is not as conclusive. Estimates for both buffer and filter storage for three months indicate a decrease in DNA concentration, but the same is not true for seven months storage. It would not make sense to consider the samples stable for seven months, but not for three months, and for that reason one should not put too much emphasis on the qPCR measurements. Considering the distributions, there seems to be a horizontal offset by the samples stored for three months. A likely cause is a deficiency in the qPCR measurement for those samples since the qPCR measurements are done in three batches: samples without storage, three months storage and seven months storage. If the standard solution used to convert (C_q) to gene copy equivalents by the standard curve has a too low concentration, the estimated

concentration of all the samples will be underestimated. Yet, there are no evidence to underpin this hypothesis.

4.1.2 Freeze-thaw cycles

Judging by the density plot of Qubit parameter estimates, storage method have an effect on DNA concentration. The handling of samples stored in buffer includes one extra freeze-thaw cycle compared to samples stored in filter storage. This may to some degree explain the detectable difference. Ross et al. (1990) concluded that the DNA concentration of blood samples were affected by freeze-thaw cycles. Too our knowledge, there is no other existing studies for bioaerosol samples or other low-concentration environmental samples. However, Figure 3.4 shows that there could be a detectable effect on bioaerosol samples, where all four samples measured by Qubit had a decrease in DNA concentration after repeated freeze-thaw cycles. This is not conclusive evidence, since there are no statistical test conducted as a result of there not being enough data points. The realisation of that there could be an effect of freeze-thawing was unfortunately detected too late in the study for further investigation of the matter.

4.1.3 Sources of variation

There is substantial variation in the raw data. By visual inspection of the box plots in Figure 3.1, one can see that the spread of data points are generally large which make the uncertainty is high. As expected by the large spread of data in the box plots, no significant differences could be found between groups by the confidence intervals in Table 3.1. This, however, could as well be caused by the uncertainty being too large, instead of there truly being no difference between groups.

There was found substantial variation between samples collected on different days, as it accounts for 38 % and 40 % of the total variation in the dataset (equation 3.2). Both the median and the spread of data is variable (Figure 3.2), but including sampling day as a random effect eliminates a considerable amount of the variation seen in the raw data.

Similarly, the DNA isolation procedure also contributes to variation. The isolation protocol contains many steps, and several possibilities for loss of sample. Even though the potential loss is small for each step, this could sum up to substantial variation. This is also true for the qPCR measurements, requiring more steps and more reagents than the Qubit, and also requiring parallel measurements of the same sample.

4.2 Comparing tools for taxonomic profiling

The benchmarking of Kraken 2, One Codex and Kaiju included testing on four types of datasets; the in vitro dataset ZymoBIOMICS, an in silico simulated dataset, a negative test dataset and real bioaerosol samples. The measured performance on the various datasets gave ambiguous results. An overview of the overall performance of the tools can be found in Table 4.1.

4.2.1 ZymoBIOMICS dataset

Kraken 2 seems to be the most optimal metagenomic classification tool for datasets containing only well studied organisms. This is an important result, as the ZymoBIOMICS dataset is derived from real samples with known proportions. Kraken 2 is most similar to the true taxa composition valuating by both Bray-Curtis dissimilarity and UniFrac distances at both genus and species-level as seen in Figure 3.3, and is superior at species-level compared to the other tools as measured by Bray-Curtis. However, the performance is more similar between the tools when measured by UniFrac, indicating that a proportion of the falsely classified reads may be classified to taxonomic neighbours. The fact that Kraken 2 gives the best abundance estimates is also visible in Figure 3.5.

Figure 3.6 shows that Kraken 2 has the highest precision and recall. For this application, a high precision is equivalent to few reads classified to the wrong taxa among all the classified reads. A high recall or sensitivity means that a large proportion the reads are classified.

One Codex and Kaiju has a drastically reduced precision at species-level compared to genus-level (Figure 3.6). The LCA algorithm will classify reads to genus-level if k-mers in the read matches several species. If this happens often, the precision at species-level will be reduced, which could explain what is observed here. Still, Kraken 2 has by far the smallest decrease in precision form genus to species-level. The explanation could be that Kraken 2 has a smaller reference database than One Codex and is hence less likely to get several hits for a read within a genus.

Table 4.1: Main aspects of performance of the metagenomic classification tools Kraken 2, One Codex and Kaiju on the four datasets used in the benchmarking. The tool(s) with the overall best performance on a given dataset is marked grey.

	ZymoBIOMICS dataset	Simulated dataset	Negative dataset	Real air dataset
Kraken 2	The best abundance estimates	Half of the species not contained in the reference database	Too sensitive Good performance when confidence level is increased to 0.05	No shared species on top 10 most abundant species list
One Codex	Low precision on species-level	Lowest Bray-Curtis distance when most abundant taxon is removed from the profile. Highest precision	Few false positives with default settings	No shared species on top 10 most abundant species list
Kaiju	Low precision on species-level	Lowest Bray-Curtis distance when all taxa in the profiles are included. Highest recall	Few false positives with default settings	300.000 reads classified to E. faecium, possibly due to high sensitivity from protein-level comparison

4.2.2 Simulated metagenome dataset

The purpose of the simulated dataset was to test the performance on data containing species similar to real bioaerosol samples. The three tools use different reference databases that do not necessarily contain entries for taxa in the dataset, in contrast to the ZymoBIOMICS dataset. This had a major impact on the results, as Kraken 2 had the poorest performance on this dataset.

The reference database has a big influence on the result. Table 3.4 shows that Kraken 2 is not able to classify any reads to six of the twelve species in the dataset, as the species are not contained in the database. Kaiju lacks three species, while One Codex classifies reads to all species in the simulated data. This reflects that One Codex has a large and well curated database, which is one of the main strengths of the tool according to the creators Minot et al. (2015). It also underlines a crucial point when testing of metagenomic classification tools: A tool will never be better than the reference database. Even with the best possible algorithm, the classification results will be poor if the taxa in the dataset are not contained in the database. With a inadequate database, the tool will only perform well on datasets with well known species, as the ZymoBIOMICS dataset is an example of. This could give a exaggerated impression of how well the tool will perform on real environmental metagenome datasets. With this problem in mind, it is not surprising that Kraken 2 gets a low precision and recall (Figure 3.7) and high distance measures by Bray-Curtis and UniFrac (Figure 3.8) relative to the other tools.

It can be argued that the Bray-Curtis distance is the most meaningful metric to use for evaluation of the tools on this dataset. The abundance estimation metrics not only considers whether taxa was predicted as present or absent in the sample, but also considers their per taxa abundances (Meyer et al., 2019). A metric evaluating abundance estimations can be considered more informative than recall and precision on its own. If there are many unclassified reads, usually detected by a low recall, the proportions will be underestimated resulting in a high Bray-Curtis distance. Many reads classified to the wrong taxa, detected by a low precision, will also be detected by the abundance estimation metrics. Further, a distance measure that includes a phylogenetic tree, such as UniFrac, may not be optimal either. To UniFrac, a read classified to a neighbouring species of the true species is not as bad as a read classified to a species in another genus. Yet in practise, both wrongly classified reads could be just as wrong. As one of the desired applications for the metagenomic classification tools is to be able to detect pathogens, it does not matter how wrong the classification is, merely whether it is right or wrong.

Which tool is the most optimal for this dataset is ambiguous. Judging by the abundance estimate metrics, Kaiju performs better than One Codex at species-level, but only when *B. abyssalis* is present (see Figure 3.8). Kaiju and One Codex both have a high recall (Figure 3.7). About 92-95 % of the reads were classified correctly by these tools. One Codex has a higher precision than Kaiju, meaning that a larger proportion of the classified reads are classified to the correct taxa. Nevertheless, it is inadequate to conclude with one tool as the most optimal when the differences are marginal and the results are easily affected by the choice of taxonomic level or metric. It is more relevant to highlight strengths and weaknesses of the tools at different types of datasets.

4.2.3 Simulated negative datasets

The negative dataset was created to see how the tools perform on data representing novel species. One may expect Kaiju to perform differently from the others because of the increased sensitivity at protein-level. Surprisingly, it was Kraken 2 that classified the most reads after shuffling with a window length of 25. A plausible reason for Kraken 2 classifying the negative control reads is the confidence level being set too low. Kraken 2 and Kaiju both has adjustable mismatch-allowances, in contrast to One Codex. It was decided to run all the tools on default settings, which turned out to be too sensitive for Kraken 2. As seen in Table 3.5, the specificity was increased to 1 when the confidence level was adjusted to 0.05. This result suggests that the confidence level of 0 is too low, and should possibly have been increased for all the runs on other datasets as well.

The selection of window length was based on Mash distances that by default compares the genomes with a k-mer length of 21, which is shorter than the 31-mers Kraken 2 and One Codex are searching with. When the window length is shorter than the k-mer length, it is unlikely that parts long enough to match an entire k-mer is intact. If that were to happen, it would be because at least two neighbouring subsequences would randomly be placed next to each other after shuffling. As the probability for this occurring is extremely low for a long sequence, any shuffling with window length shorter than the k-mer length used by the tools should be sufficient to act as negative control as long as not too many mismatches are allowed when classifying the reads.

4.2.4 Aerosol samples from Nationaltheatret subway station

The dissimilarities between the profiles of the same sample are striking. None of the species are present in the top 10 list of all three profiles from the tools tested, and only

four taxa are consistently ranked at the top for both genus- (Figure 3.10) and family-level (Table 3.6). Though not an accurate measure of profile similarity, the lists suggests considerable variation.

This could in part be caused by the tools operating with different levels of stringency associated with assigning taxa to the reads. Even if the differences in stringency are small, there could be substantial accumulative effect if there is ambiguity between the tools for many reads. The settings for One Codex are not accessible, and Kraken 2 and Kaiju both have adjustable but different measures of stringency, by confidence level and number of allowed mismatches respectively. It is therefore impossible to make a good comparison. This factor alone does not explain why the most abundant species are that dissimilar between the tools.

Further exacerbating the ambiguity between the profiles is the differences in databases. In section 4.2.2 it was shown that this can severely influence the performance of the tools. Also, this could further complicate how the LCA approach affect the classification. For example if one of the tools has a many entries strains of a species, the classification is likely to be better at species-level compared to a tool with only one entry for the species, which could be either unclassified or classified to a higher taxonomic level. Bioaerosol samples has a high degree of biological diversity, that may increase the magnitude of this problem. Still, the expected effect of a higher similarity at a higher taxonomic level was only detected when comparing number of similar taxa between the tools at species to genus-level, and not at family-level.

These are daunting findings, suggesting that the profiles from different classification tools are incomparable. Hence, the listed species are highly uncertain, emphasising that one has to be critical of the taxonomic profiles of bioaerosol samples, as it is impossible to say what tool has the most correct classification. The differences in performance also indicates that the test datasets are not good representatives of real bioaerosol samples.

Another conspicuous finding is that Kaiju classifies more than 300.000 reads to *E. faecium*, which is not detected by any of the other tools. By Kraken 2, these reads were mainly classified to "other sequences" and *H. sapiens*. A proposed benefit of Kaiju is that searches at the protein-level increase the sensitivity (Menzel et al., 2016). With a higher sensitivity, more reads can be classified. If the increased sensitivity is of such a magnitude that Kaiju correctly classifies 300.000 reads the two other tools can not detect, this is a striking result. However, these findings could merely be due to contamination in the reference database. For instance, if the *E. faecium* entries in the NCBI BLAST non-redundant database used by Kaiju in fact contains a contamination from *H. sapiens*,

it is possible for human DNA in the sample to match this entry and falsely be classified to *E. faecium*, and explain why there are 20 times more of reads classified to this taxa than the second most abundant species.

The results of the variation between samples (Figure 3.11) point to the differences being small. The distances in the dendrogram shows that even though the parallel samples are grouped together and most similar, the distances are overall short. Still, Mash distances are only a rough estimate of distance, and not a precise measure.

4.2.5 The curse of the lowest common ancestor approach

The use of the LCA approach has some disadvantages. If there are many entries for very closely related species in the reference data base, it is more likely that some reads will also match to the neighbouring species. This is a disadvantage, because the LCA algorithm will then classify the entire read to genus-level. Using the same logic, it will be an advantage to have many entries of strains of a species if species-level is the desired classification level.

Schaeffer et al. (2017) states that the usage of the LCA algorithm is a step in the wrong direction for quantification at species-level. MEGAN was one of the first reference based read assignment programs with direct strain level taxonomic classification (Huson et al., 2007). GASiC by Lindner and Renard (2013) took the classification a step further, and implemented statistical methods for classifying ambiguous reads. Unfortunately, this method requires read alignment, which is computationally demanding (Schaeffer et al., 2017). One year after the publication of GASiC, Wood and Salzberg (2014) released Kraken. Kraken had to discard the direct classification in benefit of a super fast k-mer hashing. While the increased speed and accuracy was considered an enormous breakthrough, Schaeffer et al. (2017) claims that Kraken is unsuitable for quantification due to the LCA algorithm at low taxonomic levels. This also extends to One Codex and Kaiju. Yet, the abundance estimates by Kraken 2 on the ZymoBIOMICS dataset in section 4.2.1 proved that the quantification can be good with an adequate database for the given taxa.

4.2.6 Notes on the the test datasets

The ZymoBIOMICS dataset consists exclusively of well-studied species, which is far from the biological content of a real bioaerosol sample. The species are also all present

in high abundances, making this dataset excessively simplified. Real bioaerosol samples are complex, with a combination of a high biological diversity with most taxa lowly abundant. Hence, there is a large amount of organisms only visible by few reads. Also, there are considerable pars of novel species not contained in reference databases, as seen by the simulated dataset discussed in section 4.2.4.

Even though the ZymoBIOMICS community standard used in this work is a gross under-representation of the complexity in air, it still serves a purpose. As this is an in vitro dataset, it is real data with an identical DNA isolation procedure to the bioaerosol samples. On the other side, the in silico simulated dataset is more realistic in regards to the biological composition, but still only represents a fraction of the real biological complexity. According to Almeida et al. (2018) "in silico datasets are better for highlighting the computational pipelines—independently of experimental variation and technical biases, but may require further validation in real-world datasets". They further explain that the combination of in silico and in vitro datasets are essential for understanding the entire analysis.

When creating a simulated metagenome dataset, the selection of taxa will have a huge impact on the results. As seen by the results of the simulated dataset (section 4.2.2), deficiencies in the reference databases can affect the performance dramatically. In this study, the selection of taxa was based on a paper from 2008 (Tringe et al., 2008), in which the abundances were expressed in number of phylogroups. As the species selection should be as close to reality as possible, a newer publication based on shotgun metagenome sequencing data should be used. Still, there is evidence suggesting that the microbiome in outdoor air is extremely variable which makes the creation of one true dataset non-trivial (Kuske, 2006; Behzad et al., 2015). Further, more species should be included in the simulated dataset, as 12 species are only a fraction of the diversity found in air.

The read length simulated by ART is fixed at 2x250 bp, but the real dataset from Nationaltheatret is sequenced with 2x150 bp. Longer reads in the test dataset could mean that they are more easily recognised by the metagenomic tools compared to real reads, which could contribute to an overly optimistic estimated performance on the test datasets compared to real data.

4.3 Concluding remarks and further perspectives

In this work, two subjects related to the creation of metagenomic profiles of aerosol samples were investigated.

The first aim was to confirm that the currently used storage procedure by FFI does not affect the DNA concentration. There could not be found any evidence suggesting a departmental effect of filter storage up to seven months at -80°C. The samples with buffer storage showed a decrease in DNA concentration, which could be caused by the samples being freeze-thawed in the preparation. However, there was not enough data to further investigate this effect. The impact of repeated freeze-thaw cycles on both isolated DNA, filter extracts and filters should be examined. Also, the effect on mock community with known abundances should be studied.

The second aim was to benchmark three metagenomic classification tools for aerosol samples specifically. As this was done by four different categories of datasets which gave ambiguous results, there was no tool with the overall best performance. Instead, strengths and weaknesses for the different types of datasets are pointed out. Kraken 2 is superior for the purpose of abundance estimates of well-known species, but the negative dataset suggested that Kraken 2 was run on a too low confidence level, resulting in an overly sensitive classification. The simulated dataset with a more realistic selection of species highlighted the importance of an adequate reference database, as One Codex was the only tool classifying reads to all species in the dataset. The real dataset underlined how vastly simplistic the test datasets were compared to the complexity of real aerosol. A peculiar finding suggested that Kaiju may have an advantageous increased sensitivity from protein-level classification, but whether this was in fact due to contamination in the reference database should be investigated further.

The extensive disagreement between the metagenomic classification tools of real bioaerosol samples proves that there is an imminent need for more research and improvements in the field. A key objective should be to improve the reference databases specifically for the species in the environment, as this has shown to be essential for similar research fields, such as the gut microbiome (Zou et al., 2019; Forster et al., 2019). As stated by Behzad et al. (2015): "The solution is more metagenomic study. Metagenomic studies lead to larger genomic databases, whereas larger databases make metagenomic analysis easier".

The rapid development in the field of sequencing techniques could help improve metagenomic classification tools. Currently, third generation sequencing such as Nanopore

are increasing in popularity. These are long read technologies, better suited for assembly. This could lead to better metagenome assembled genomes (MAGs), that can be included in the reference databases to decrease the proportion of novel reads in bioaerosol samples. MAGs has proven to increase the insight into the microbial diversity of novel environments (Wilkins et al., 2019).

It was further concluded that the LCA approach has some drawbacks for quantification. A future solution could be to use approaches that utilises a better statistical framework that allows for probabilistic assignment of reads, for example Metakallisto as suggested by Schaeffer et al. (2017) to improve the metagenomic classification tools.

Bibliography

(2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539.

Almeida, A., Mitchell, A. L., Tarkowska, A., and Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience*, 7(5).

Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, 59(1):143 LP – 169.

Amato, P., Parazols, M., Sancelme, M., Laj, P., Mailhot, G., and Delort, A.-M. (2007). Microorganisms isolated from the water phase of tropospheric clouds at the Puy de Dome: major groups and growth abilities at low temperatures. *FEMS Microbiology Ecology*, 59(2):242–254.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Be, N., Thissen, J., Fofanov, V., Allen, J., Rojas, M., Golovko, G., Fofanov, Y., Koshinsky, H., and Jaing, C. (2014). Metagenomic analysis of the airborne environment in urban spaces. *Microbial Ecology*, 69(2):346–355.

Behzad, H., Gojobori, T., and Mineta, K. (2015). Challenges and opportunities of airborne metagenomics. *Genome biology and evolution*, 7(5):1216–1226.

Beiko, R. G. and Parks, D. H. (2012). Measuring Community Similarity with Phylogenetic Networks. *Molecular Biology and Evolution*, 29(12):3947–3958.

Bender, M. A., Farach-Colton, M., Pemmasani, G., Skiena, S., and Sumazin, P. (2005). Lowest common ancestors in trees and directed acyclic graphs. *J. Algorithms*, 57:75–94.

Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*.

Burrows, S. M., Elbert, W., Lawrence, M. G., and Pöschl, U. (2009). Bacteria in the global atmosphere - part 1: Review and synthesis of literature data for different ecosystems. *Atmospheric Chemistry and Physics*, 9(23):9263–9280.

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., and Wittwer, C. T. (2009). The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry*, 55(4):611 LP – 622.

Bøifot, K. O., Gohlo, J., Moen, L. V., and Dybwad, M. (2019). Dna isolation method optimized for aerosol microbiome studies. Submitted.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J. A., Smith, G., and Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The Isme Journal*, 6:1621.

Chen, J. (2018). *GUniFrac: Generalized UniFrac Distances*. R package version 1.1.

Crawley, M. J. (2013). *The R book*. Second edition. Chichester, West Sussex, United Kingdom : Wiley, 2013.

Cronholm, L. S. (1980). Potential health hazards from microbial aerosols in densely populated urban regions. *Applied and environmental microbiology*, 39(1):6–12.

Dimmick, R. L., Straat, P. A., Wolochow, H., Levin, G. V., Chatigny, M. A., and Schrot, J. R. (1975). Evidence for metabolic activity of airborne bacteria. *Journal of Aerosol Science*, 6(6):387–393.

Dybwad, M. (2014). *Characterization of Airborne Bacteria at a Subway Station: Implications for Testing and Evaluation of Biological Detection, Identification, and Monitoring Systems*. PhD thesis, Norwegian University of Science and Technology.

Dybwad, M., Skogan, G., and Blatny, J. M. (2014). Comparative Testing and Evaluation of Nine Different Air Samplers: End-to-End Sampling Efficiencies as Specific Performance Measurements for Bioaerosol Applications. *Aerosol Science and Technology*, 48(3):282–295.

Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance.

Fang, Z., Ouyang, Z., Hu, L., Wang, X., Zheng, H., and Lin, X. (2005). Culturable airborne fungi in outdoor environments in Beijing, China. *Science of The Total Environment*, 350(1):47–58.

Forster, S. C., Kumar, N., Anonye, B. O., Almeida, A., Viciani, E., Stares, M. D., Dunn, M., Mkandawire, T. T., Zhu, A., Shao, Y., Pike, L. J., Louie, T., Browne, H. P., Mitchell, A. L., Neville, B. A., Finn, R. D., and Lawley, T. D. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology*, 37(2):186–192.

Gardner, P. P., Watson, R. J., Morgan, X. C., Draper, J. L., Finn, R. D., Morales, S. E., and Stott, M. B. (2019). Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ*, 7:e6160.

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science (New York, N.Y.)*, 312(5778):1355–1359.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17:333.

Heedrik, D., Douwes, J., Pearce, N., and Thorne, P. (2003). Bioaerosol Health Effects and Exposure Assessment: Progress and Prospects. *The Annals of Occupational Hygiene*, 47(3):187–200.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4):593–594.

Hunt, B. R., Yorke, J. A., Roberts, M., Mount, S. M., and Hayes, W. (2004). Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20(18):3363–3369.

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome research*, 17(3):377–386.

Kubista, M., Andrade, J. M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., Sindelka, R., Sjöback, R., Sjögreen, B., Strömbom, L., Ståhlberg, A., and Zoric, N. (2006). The real-time polymerase chain reaction. *Molecular Aspects of Medicine*, 27(2):95–125.

Kuske, C. R. (2006). Current and emerging technologies for the study of bacteria in the outdoor air. *Current Opinion in Biotechnology*, 17(3):291–296.

Lindgreen, S., Adair, K., and Gardner, P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports*, 6:19233.

Lindner, M. S. and Renard, B. Y. (2013). Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic acids research*, 41(1):e10.

Liu, Cindy M.and Aziz, M., Kachur, S., Hsueh, P.-R., Huang, Y.-T., Keim, P., and Price, L. B. (2012). Bactquant: An enhanced broad-coverage bacterial quantitative real-time pcr assay. *BMC Microbiology*, 12(1):56.

Mardis, E. and McCombie, W. R. (2017). Library Quantification: Fluorometric Quantitation of Double-Stranded or Single-Stranded DNA Samples Using the Qubit System. *Cold Spring Harbor Protocols*, 2017(6):pdb.prot094730.

McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., Minot, S. S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N. A., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R. R., Rosen, G. L., and Mason, C. E. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, 18(1):182.

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7:11257.

Meyer, F., Bremges, A., Belmann, P., Janssen, S., McHardy, A. C., and Koslicki, D. (2019). Assessing taxonomic metagenome profilers with OPAL. *Genome Biology*, 20(1):51.

Minot, S. S., Krumm, N., and Greenfield, N. B. (2015). One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. *bioRxiv*, page 27607.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2019). *vegan: Community Ecology Package*. R package version 2.5-4.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill,

K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132.

Peabody, M. A., Van Rossum, T., Lo, R., and Brinkman, F. S. L. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, 16(1):362.

Prussin, A. J. and Marr, L. C. (2015). Sources of airborne microorganisms in the built environment. *Microbiome*, 3(1):78.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1):341.

Quast, C., Pruesse, E., Gerken, J., Peplies, J., Yarza, P., Yilmaz, P., Schweer, T., and Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596.

Rodgers, J. L. (1999). The Bootstrap, the Jackknife, and the Randomization Test: A Sampling Taxonomy. *Multivariate Behavioral Research*, 34(4):441–456.

Rosario, K., Fierer, N., Miller, S., Luongo, J., and Breitbart, M. (2018). Diversity of DNA and RNA Viruses in Indoor Air As Assessed via Metagenomic Sequencing. *Environmental Science & Technology*, 52(3):1014–1027.

Ross, K. S., Haites, N. E., and Kelly, K. F. (1990). Repeated freezing and thawing of peripheral blood and DNA in suspension: effects on DNA yield and integrity. *Journal of Medical Genetics*, 27(9):569 LP – 570.

Rosselló-Móra, R. and Amann, R. (2015). Past and future species definitions for Bacteria and Archaea. *Systematic and Applied Microbiology*, 38(4):209–216.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Sattler, B., Puxbaum, H., and Psenner, R. (2001). Bacterial growth in supercooled cloud droplets. *Geophysical Research Letters*, 28(2):239–242.

Schaeffer, L., Pachter, L., Pimentel, H., Bray, N., and Melsted, P. (2017). Pseudoalignment for metagenomic read assignment. *Bioinformatics*, 33(14):2082–2088.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6):e37–e37.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiūtė, M., Hansen, L. H., Sørensen, S. J., Chia, B. K. H., Denis, B., Froula, J. L., Wang, Z., Egan, R., Don Kang, D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M. D., Lingner, T., Lin, H.-H., Liao, Y.-C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard, B. Y., Pop, M., Klenk, H.-P., Göker, M., Kyrpides, N. C., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T., and McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods*, 14:1063.

Shokralla, S., Spall, J. L., Gibson, J. F., and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, 21(8):1794–1805.

Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of bacteriology*, 194(16):4151–4160.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526.

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):3.

Tringe, S. G., Zhang, T., Liu, X., Yu, Y., Lee, W. H., Yap, J., Yao, F., Suan, S. T., Ing, S. K., Haynes, M., Rohwer, F., Wei, C. L., Tan, P., Bristow, J., Rubin, E. M., and Ruan, Y. (2008). The airborne metagenome in an indoor urban environment. *PLOS ONE*, 3(4):1–10.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkins, L. G. E., Ettinger, C. L., Jospin, G., and Eisen, J. A. (2019). Metagenome-assembled genomes provide new insight into the microbial diversity of two thermal pools in Kamchatka, Russia. *Scientific Reports*, 9(1):3059.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46–R46.

Yooseph, S., Andrews-Pfannkoch, C., Tenney, A., McQuaid, J., Williamson, S., Thiagarajan, M., Brami, D., Zeigler-Allen, L., Hoffman, J., Goll, J. B., Fadrosh, D., Glass, J., Adams, M. D., Friedman, R., and Venter, J. C. (2013). A Metagenomic Framework for the Study of Airborne Microbial Communities. *PLOS ONE*, 8(12):e81862.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36.

Zhang, T. and Fang, H. H. P. (2006). Applications of real-time polymerase chain reaction for quantification of microorganisms in environmental samples. *Applied Microbiology and Biotechnology*, 70(3):281–289.

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., Wan, D., Jiang, R., Su, L., Feng, Q., Jie, Z., Guo, T., Xia, Z., Liu, C., Yu, J., Lin, Y., Tang, S., Huo, G., Xu, X., Hou, Y., Liu, X., Wang, J., Yang, H., Kristiansen, K., Li, J., Jia, H., and Xiao, L. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*, 37(2):179–185.

Attachments

Listing 1: R output of Qubit-model.

```
> summary(modQb)
Formula:
Qubit ~ Storage + Method + (1 | Sampling_nr)
   Data: DNA2

REML criterion at convergence: -99.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6897 -0.3618  0.0803  0.5086  2.2760

Random effects:
 Groups      Name        Variance Std.Dev.
 Sampling_nr (Intercept) 0.002946 0.05428
 Residual                0.004800 0.06928
Number of obs: 52, groups:  Sampling_nr, 13

Fixed effects:
             Estimate Std. Error t value
(Intercept)   0.002108   0.027987   0.075
StorageLong   0.004200   0.019215   0.219
MethodBuffer -0.045715   0.035822  -1.276

Correlation of Fixed Effects:
            (Intr) StrgLn
StorageLong -0.343
MethodBuffr -0.689  0.000
```

Listing 2: R output of qPCR-model.

```
Formula:
qPCR_16S ~ Storage + Method + (1 | Sampling_nr)
   Data: DNA3

REML criterion at convergence: 899.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.98026 -0.48453 -0.04173  0.56739  2.49477

Random effects:
 Groups      Name        Variance Std.Dev.
 Sampling_nr (Intercept) 2268756  1506
 Residual                3419479  1849
Number of obs: 52, groups:  Sampling_nr, 13

Fixed effects:
             Estimate Std. Error t value
(Intercept)    -918.6      765.7  -1.200
StorageLong     841.7      512.9   1.641
MethodBuffer   -379.0      983.3  -0.385

Correlation of Fixed Effects:
            (Intr) StrgLn
StorageLong -0.335
MethodBuffr -0.691  0.000
```

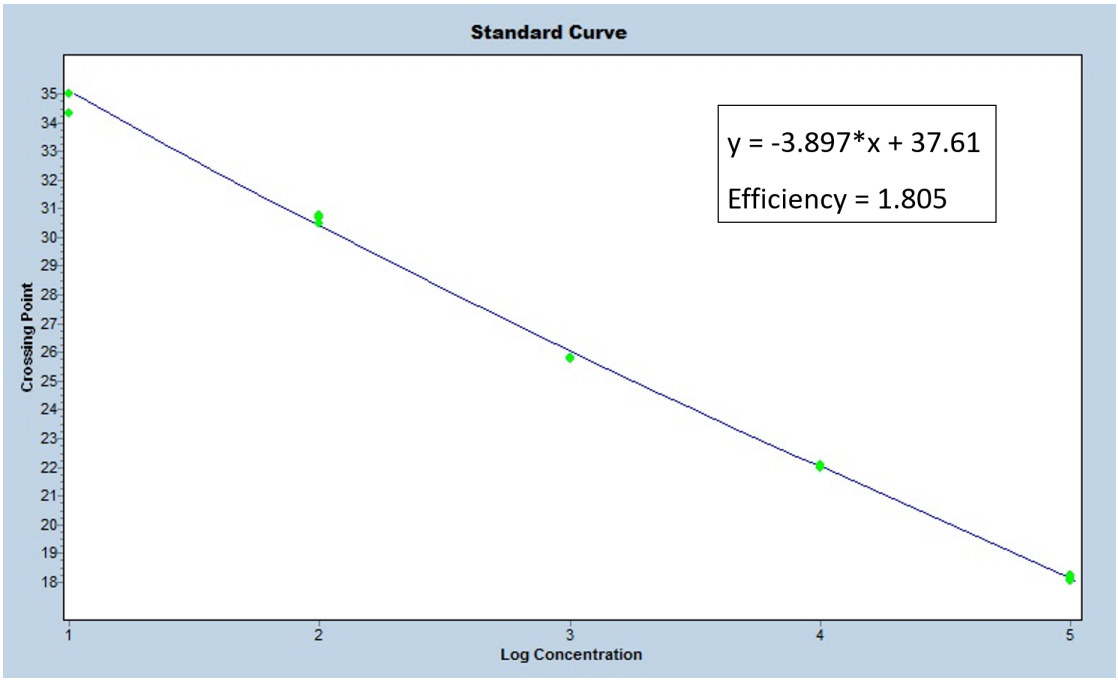


Figure 1: qPCR standard curve from *Escherichia coli* genomes

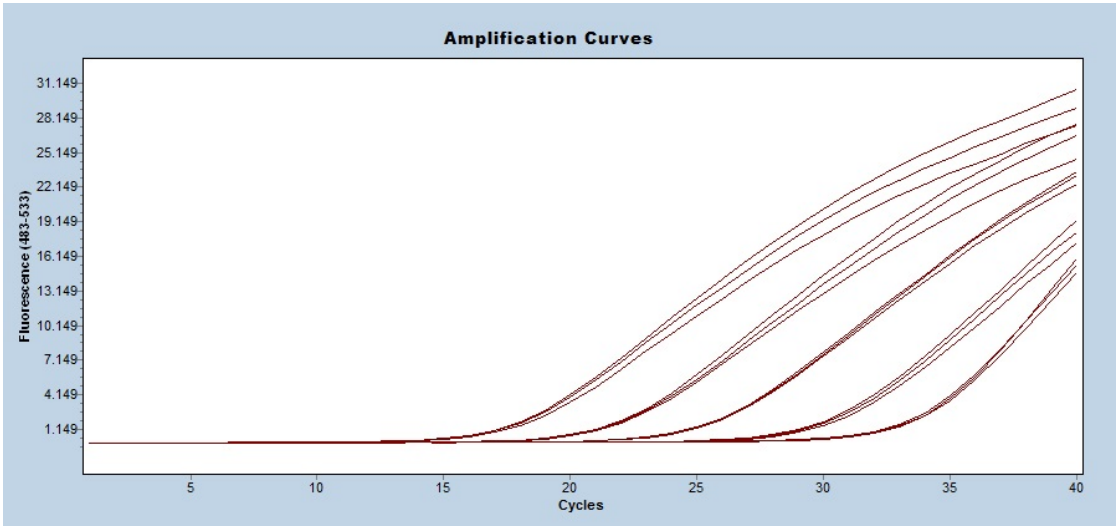


Figure 2: Amplification curve used to make qPCR standard curve

Table 1: Abundance estimated for ZymoBIOMICS community standard dataset made by Kraken2, OneCodex and Kaiju. All abundances are listed as estimated percent of all reads. Abundances are listed for sample A and B, and at genus and species level.

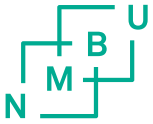
	Real	Kraken2		OneCodex		Kaiju	
		A	B	A	B	A	B
Genus							
Listeria	12	12.08	11.49	11.56	9.90	11.40	10.86
Pseudomonas	12	10.32	9.02	9.04	8.39	3.45	3.00
Bacillus	12	9.55	9.39	8.61	7.20	4.98	4.89
Escherichia	12	6.14	5.24	3.28	3.20	3.04	2.59
Salmonella	12	7.52	6.85	6.75	6.06	5.82	5.29
Lactobacillus	12	9.02	14.51	14.99	7.56	8.53	13.71
Enterococcus	12	11.81	11.00	5.39	4.83	4.00	3.68
Staphylococcus	12	14.17	13.39	9.41	8.13	12.46	11.79
Saccharomyces	2	2.05	1.98	1.88	1.47	1.49	1.61
Cryptococcus	2	1.99	2.90	2.98	1.66	1.38	2.01
Unclassified	0	0.04	0.04	13.02	12.64	7.62	7.40
Species							
Listeria monocytogenes	12	11.83	11.25	9.55	8.96	6.65	6.34
Pseudomonas aeruginosa	12	4.18	3.53	0.95	0.80	1.67	1.41
Bacillus subtilis	12	9.34	9.16	2.42	2.35	1.64	1.61
Escherichia coli	12	5.87	5.00	3.21	2.65	2.81	2.40
Salmonella enterica	12	5.64	5.09	4.12	3.68	4.57	4.16
Lactobacillus fermentum	12	8.80	14.17	2.92	4.70	1.66	2.64
Enterococcus faecalis	12	11.66	10.86	2.25	1.99	2.31	2.10
Staphylococcus aureus	12	13.41	12.68	1.12	1.06	2.50	2.41
Saccharomyces cerevisiae	2	2.04	1.96	0.42	0.41	0.37	0.38
Cryptococcus neoformans	2	1.97	2.87	1.83	2.64	1.11	1.63
Unclassified	0	0.04	0.04	13.02	12.64	7.62	7.62

Table 2: Recall and precision for Kraken 2, One Codex and Kaiju on the ZymoBIOMICS dataset. The recall and precision is calculated at both species and genus level, and for both sample A and B.

		Kraken		OneCodex		Kaiju	
		A	B	A	B	A	B
Precision	Species	0.75	0.77	0.42	0.42	0.33	0.33
	Genus	0.83	0.84	0.83	0.63	0.56	0.59
Recall	Species	1.00	1.00	0.69	0.70	0.77	0.77
	Genus	1.00	1.00	0.85	0.82	0.88	0.89

Table 3: Recall and precision for Kraken 2, One Codex and Kaiju on the simulated dataset containing 12 species found in air samples. The recall and precision is calculated at both species and genus level.

	Kraken 2		One Codex		Kaiju	
	Genus	Species	Genus	Species	Genus	Species
Recall	0.70	0.50	0.93	0.92	0.95	0.94
Precision	0.64	0.27	0.92	0.88	0.80	0.72



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

Postboks 5003
NO-1432 Ås
Norway