A bioinformatic analysis of the fungi growing on biodegradable plastics mulch in agricultural soil.

Writer: Tim de Boer Supervisor: Anna Heintz-Buschart Examiner: Franciska de Vries

Date: 4 july 2022 Institute: Swammerdam Institute for Life Sciences

Abstract

For seventy years, plastic pollution of marine and terrestrial ecosystems has steadily grown, with widely documented detrimental effects on ecosystems(Windsor et al., 2019). Plastics accumulate in the environment due to the fact that they generally are not biodegradable. This is why the FAO calls for a wider use of biodegradable plastic(*Assessment of agricultural plastics and their sustainability: A call for action,* 2021) despite that the effects of biodegradable plastic on the environment are not yet know. This study looked at the fungi found on poly(butylene succinate-co-adipate)(PBSA) mulch using bioinformatic methods and assessing how well these methods perform. This study found *Botrytis cinerea* to be the most prominent fungi on the mulch opposed the expected result of *Tetracladium*. Whilst these different results are promising the current bioinformatic are not accurate enough yet to give definite proof. As Eukrep and Whokaryote were able to find equal amounts of fungal sequence in the samples but half of their sequences were unique to one and other. Resulting in their alignment only reaching 0.2% coverage of the *Botrytis cinerea* genome. Which is not high enough to give a conclusion on which fungi grow on biodegradable plastics in agricultural soil.

Introduction

Problems with plastic

For seventy years, plastic pollution of marine and terrestrial ecosystems has steadily grown, with widely documented detrimental effects on ecosystems(Windsor et al., 2019). Leslie et al. (2022) demonstrated that plastic particles are bioavailable for uptake into the human bloodstream, thereby providing evidence that exposure to micro-plastics is not safe to humans.

Plastics accumulate in the environment due to the fact that they generally are not biodegradable. They only fall apart into smaller parts, by mechanical forces and weathering forming them into micro-plastic particles and releasing pollutants into the environment. The effects of these plastics and their pollutants on the environment and humans is not yet understood but they are found to have carcinogenic effects among other problems(Kumar et al., 2022). Still, plastics are widely used everywhere in the world. For example, plastic mulches are used in agriculture to improve crop quality and yield by reducing soil erosion, increasing water efficiency, regulating soil temperature, and exerting pathogen control(Pathan et al., 2020). Most plastics used for mulching are made of Polyvinyl Chloride(PVC)and low-density polystyrene due to their low cost. These plastics introduce polystyrenes and harmful additives like phthalates into the soil and the food chain.

Alternative plastics

These harmful additives, the carbon footprint and to limit accumulation of microplastics in the environment are reasons replace plastics with other compounds. One replacement of plastics in some applications is with biodegradable plastics. The FAO defines biodegradable

plastics as follows: *"Biodegradable-plastics are broken down by naturally occurring microorganisms – such as bacteria and fungi – into water, biomass, and gases such as carbon dioxide and methane. The rate of biodegradation depends on environmental conditions such as temperature, humidity, the consortia of microorganisms present and the presence or absence of oxygen. Biodegradable plastics can be made from bio-based and fossil-based precursors, and sometimes a mixture of the two."* (*Assessment of agricultural plastics and their sustainability: A call for action,* 2021). However, for most biodegradable plastics and conditions, it is not known how the biodegradation happens outside of lab conditions. Besides this biodegradable plastic have a lower environmental risk, but do not see a lot of use yet due to their higher price.

In a recent study, poly(butylene succinate-co-adipate)(PBSA) film, biodegradable plastic mulch, was buried in agricultural soil at an agricultural experimental field station in Bad Lauchstädt, Central Germany (51°22′60″ N, 11°50′60″ E, 118 m a.s.l.) and the microbial community development and degradation rates were described(Purahong et al., 2021; Tanunchai et al., 2021). In this study, an rRNA gene/region amplicon sequencing approach was employed, which allows for an overview of the microbial community, but does not give a functional view on the community members. Using a subset of samples from a mulch after 180 days of being buried in the soil, this study aims to functionally describe the microbial communities on the biodegradable-plastic samples, with a focus on the eukaryotes. After 180 day two fungi were found to be most prevalent *Dothideomycetes* with 30% and *Tetracladium* with 35% ("ID 853344 - BioProject - NCBI," 2022; Purahong et al., 2021). The fungal DNA was quantified using qPCR. So, these fungi or ones related to it are expected to be found in the samples. These fungi are not yet well described in the studied fields or as a potential degrader of biodegradable plastics. Literature suggests that *Penicillium* and *Aspergillus* species are the most commonly found plastic degraders(Moore-Kucera et al., 2014). However new research has found that between different types of plastic and even in between different types of degradable plastics mulches different microbial communities will establish themselves(Bandopadhyay et al., 2020). Bandopadhyay et al.(2020) found that the richness on biodegradable plastics is increased compared to normal plastics. On biodegradable plastics, they found *Dothideomycetes*, *Sordariomycetes* and T*remellomycetes* to be the most dominant classes of fungi to establish themselves.

Metagenomics

This study aims to test whether a metagenomic approach to the analysis of these microbiomes can lead to new and unique insights regarding these microbiomes. The approach is to use different types of machine learning to determine if a sequence is prokaryotic or eukaryotic. Doing this in the lab is difficult so using machine learning to determine it after sequencing will be practical. The expectation is the larger part of the sequences will be prokaryotic due to the fact that prokaryotes are significantly smaller than eukaryotes but have nearly the same amount of DNA as micro-eukaryotes, such as uni-cellular fungi. So, from equal volume of prokaryote and eukaryote sample one would expect to find several times more prokaryotic DNA. Acquiring high-quality gene predictions in eukaryotic DNA mixed with prokaryotic DNA is a challenge due to the fact that eukaryotes and prokaryotes differ greatly in the build-up of their sequences. Eukaryotes have more complex promotor regions, regulatory signals, and genes spliced into introns and exons, which also vary between species(West et al., 2018). This is why prokaryotic gene predictors will not always do well predicting eukaryotic gene sequences. On the other hand, the differences between eukaryotic and prokaryotic genome sequences can be employed to distinguish genomic sequences of the two. Which is doing this in the lab proves to be difficult

but using bioinformatics and machine learning it is possible. This should allow for their separation in metagenomics data and separate annotation of gene positions.

This study aims to find what microbial communities live on biodegradable plastic with a focus on the fungi. The study also aims to test different bioinformatics methods to differentiate between prokaryotic and eukaryotic DNA. The expectation is that the different bioinformatics methods will all be capable of yielding results but that the more conventional programs will yield better results as they have more often proven to be successful. The fungi that are expected to be found are *Tetracladium, Dothideomycetes, Penicillium* and *Asperigillus*, based on previous data(Purahong et al., 2021; Tanunchai et al., 2021) and literature(Moore-Kucera et al., 2014).

Method

The data used in this study comes from two different plots (1-2A and 2-4F) of biodegradable plastic mulch incubated in the soil for 180 days (Figure 1). The 1-2A plot was under the

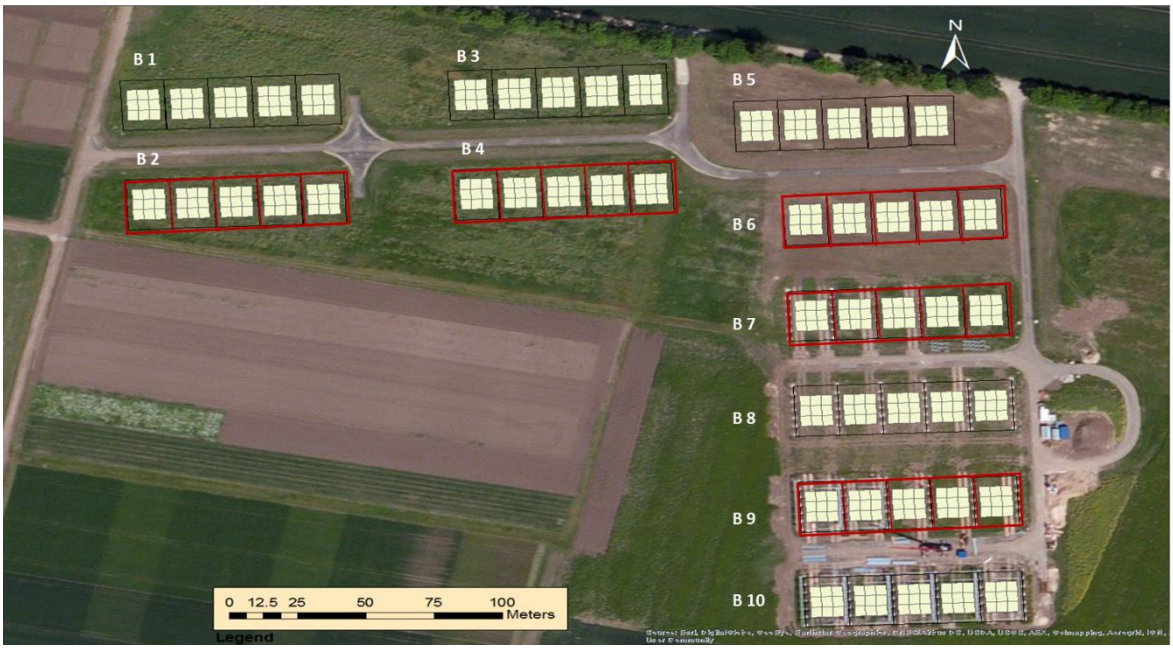

Figure 1: *Conventional farming treatment plots of the Global Change ExperimentalFacility39(GCEF) located at the field research station of the Helmholtz-Centre for Environmental Research in Bad Lauchstädt, Central Germany (51°22'60''N,11°50'60''E, 118 m a.s.l.). The fields in the red squares were held under a climate which the earth will have in the Future(F) the other fields are held at current climate(A).*

ambient environment. The 2-4F plot was grown under the environment expected to be in the future so dryer and hotter. After 180 days, this foil was dug up and DNA samples were taken. Using shotgun short-read (Illumina) sequencing, metagenomics was performed to reconstruct the functional repertoire of the bioplastic-degrading microbial community with a focus on the fungal community members. The steps that were performed started with the assembly followed by the classification then a taxonomic profile was created from the assembly and lastly an alignment was run in between the fungal DNA found in the assembly and the species selected from the taxonomic profile(Figure 2).

Assembly

First, the short and long reads were assembled using the IMP3 pipeline (https://imp3.readthedocs.io/en/latest/, Narayanasamy et al., 2016). The first step of the pipeline is preprocessing, which starts with trimming the samples using trimmomatic(Bolger et al., 2014). Trimming means removing the inaccurate bases at the beginnings and ends of the sequences, so only the reliable parts are left. Then reference filtering is done to remove the unwanted DNA in the sample using a Burrows-Wheeler aligner (BWA)(Narayanasamy et al., 2016). Unwanted DNA could include host or human genomes, but in this case, only the sequencing spike-in, the genome of phage PhiX174, is removed.

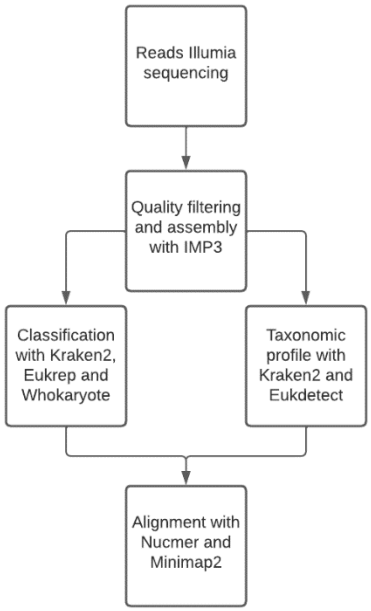

Figure 2: Bioinformatic approach

Then the cleaned-up files were ready for step two in the IMP3 pipeline: assembly. This is done using Megahit assembler (https://github.com/voutcn/megahit, Li et al., 2015). Megahit assembles the contigs using succinct de Bruijn Graphs(Bowe et al., 2012). By using these succinct de Bruijn Graphs, Megahit is faster and uses less storage. This is because it uses a more efficient way of representing every part of the de Bruijn Graph in a succinct way. In a 4m + o(m) bits method in which "m" is the number of strings or k-mers here, and "o" is the size of the alphabet, in this case all types of nucleotides (4). The sequences that could not be mapped in the first run of Megahit were put through the megahit program a second time Following this, the overlapping contigs were merged together using Cap3. This results in longer contigs without reducing their quality. Lastly, the short metagenomic reads were mapped against the assembly using BWA.

The length of the assembly is quantified using the N50. The N50 is the shortest contig length that needs to be included for the contigs to cover of 50% of the genome. Meaning the sum of all the contigs of length N50 or longer is 50% or more than the total genome sequence. This tells you how successful the assembly is in building larger contigs. Larger contigs are preferable because when they correspond with a marker or genome it gives more complete information about this marker or genome.

Classifying the eukaryotes

The second step is using and comparing classification programs, Kraken2, EUKrep and Whokaryote, to find the eukaryotic sequences in the assembly. The first program used to get a rough grasp of the dataset is Kraken2 (Wood et al., 2019). Kraken2 is a software that matches sequences to a database, in this case containing bacterial and eukaryotic genomes. Kraken2's k-mer-based approach provides a fast taxonomic classification of metagenomic sequence data(https://github.com/DerrickWood/kraken2, Wood et al., 2019). A *k*-mer is a string of length *k*(Compeau and Pevzner, 2015). Kraken2 is fast due to the fact that it reduces the amount of memory necessary by using a compact hash table, which is a data structure that gives all the data a key referencing to the data point. So, it can work with the smaller memory key to test the data and not the full datapoint. By finding the lowest common ancestor (LCA) of the taxa that contain a *k*-mer, Kraken2 determines what taxa are in the assembly.

In parallel, the specific programs EUKrep and Whokaryote were used to search for general hallmarks of eukaryotic genomes. EUKrep uses a *k*-mer-based approach for identification of assembled eukaryotic sequences in data sets from diverse environmental samples (West et al., 2018). With the use of *k*-mers, patterns in the sequences can be brought to light and using these patterns EUKrep can classify whether a sequence is

eukaryotic or prokaryotic. In EUKrep specifically, the frequencies of 5-mers are counted for the training and classification. This length compromises between speed and accuracy for classifying eukaryotic scaffolds(West et al., 2018). EUKrep was trained on a library of complete prokaryotic and eukaryotic genomes taken from NCBI. Then the contigs were split into 5kb chunks and only chunks larger then 3kb were used. The frequencies of the 5-mers for each contig were calculated(West et al., 2018). These 5-mer frequencies were used to train a linear support vector machine (linear-SVM). A linear-SVM is a model that draws hyperplanes, a plane in many dimensions, through a data set to most accurately divide the data set and create groups. The groups the model selected on are archaeal, bacterial, opisthokonta or protist origin. In classification, once the linear-SVM has classified the 5kb chunks they are stitched back together so the taxonomy of the complete contig can be determined.

On the other hand, Whokaryotic is trained on random forest classifiers. The classifier uses intergenic distance, gene density and other 92 genomic features to predict whether a given metagenomic contig belongs to a eukaryote or a 93 prokaryote(Pronk and Medema, 2021). Then it is improved by having it learn from the Tiara, a machine learning similar to EUKrep(Pronk and Medema, 2021).

As Whokaryote is a recently developed program its use is not yet widely tested. To see how it holds up to the already established EUKrep will be worthwhile and their results together will give a more complete dataset.

Taxonomic profiling

For taxonomic profiling, Eukdetect will use the reads and Kraken2 the assembly to give a first estimation of what species can be found in a sample. Kraken2 utilizes spaced seeds in the storage and querying of minimizers to improve classification accuracy (https://github.com/DerrickWood/kraken2, Wood et al., 2019). Using this, Kraken2 is able to find all types of species in the assembly, including both prokaryotes and eukaryotes. It will relate every contig to somewhere in the taxonomic tree.

Eukdetect uses a database of 521,824 universal marker genes from 241 conserved gene families including 3,713 fungal species(https://github.com/allind/EukDetect, Lind and Pollard, 2021). The Eukdetect pipeline aligns reads to these markers and filters the alignment on mapping quality. This mapping quality needs to be greater than 30(Lind and Pollard, 2021). The mapping quality entails the confidence of how the reads are mapped to the sequence. Eukdetect also uses a minimum alignment of 80% of the marker genes. Then the aligned reads to a marker are counted and their percent sequence identity is calculated. This percent sequence identity describes how much of the sequence is described by the reads per species.

Together these two programs create a reference library with species expected to be found in the sample with which the classification programs run their alignment. Using the taxonomic profiles, the most likely species will be chosen in this paper for further analysis. For kraken2 profiles, only the contigs it can determine down to species level will be used. Their genome will be retrieved from National Centre for Biotechnology Information database(NCBI, https://www.ncbi.nlm.nih.gov/genome/). These species will be selected on the basis of literature, Eukdetect and previous results.

Classification

Then in third step the classification programs Nucmer aligner from Mummer4 (Marçais et al., 2018) and MiniMap2 (Li, 2021, 2018) were used to determine whether the classified fungal sequences closely resembled the reference genomes of organisms detected by the taxonomic profilers. Nucmer is the genome wide sequence aligner from the Mummer4 (https://github.com/mummer4/mummer) package. Nucmer uses a suffix array with a lock free first-in-first-out (FIFO) queue. The suffix array functions similarly to BWA, in that it orders all the parts of the string alphabetically with their corresponding part in the string as number with it. Following this, each worker thread computes the exact alignments, cluster them, and runs the banded Smith-Waterman alignment routine for its single query sequence(Marçais et al., 2018). The Smith-Waterman alignment is a basic dynamic programming algorithm that works as follows: it puts the two sequences that need to be compared in as the row- and column names of matrix: every time two nucleotides in the alignment are the same, the value is increased by +2 - when it is different the value is decreased by -1. The values in the matrix are not allowed to go below 0. In this way, the Smith-Waterman aligner will look at the route over the matrix in which it ends on the highest number on the matrix. This route will then be equal to the longest aligning part of the sequences. From this it can output how much your sequence aligned with sequences from the genome library.

		A	C	A	C	A	C	T	A
	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

A - C A C A C T A

A G C A C A C - A

Figure 1: **The Smith-Waterman alignment.**(Razmyslovich et al., 2010)

MiniMap2(https://github.com/lh3/minimap2 Li, 2021) uses a seed-chain aligning procedure to align the sequences. This works by taking a seed, the k-mer query that is looked for and every time it overlaps with the sequence that it is aligned too it is chained to the previous one across the sequence giving it an alignment score(Compeau and Pevzner, 2015). Additionally it uses a hash table to decrease the memory load of the alignment. Minimap2 also uses the Smith-Waterman aligner to determine how much the sampling sequence aligns with sequences form the genome library(Li, 2021). Then it pairs your sample data with the given genome library.

Together, Nucmer and Minimap2 show what species can be found on the PBSA mulch. Their comparison brings perspective to the reliability of the programs but also when species were found by both programs to be very apparent in the samples, resulting in a more trustworthy analysis.

Results

The following sections present the results from assembly, classification, taxonomic profiling, and comparison to known genomes. The aims are to evaluate the assembly to reconstruct metagenomes, to separate the eukaryotic and fungi from the assembly and analyse which eukaryotic taxa are found and which fraction of these known genomes of these organisms are detected.

Assembly

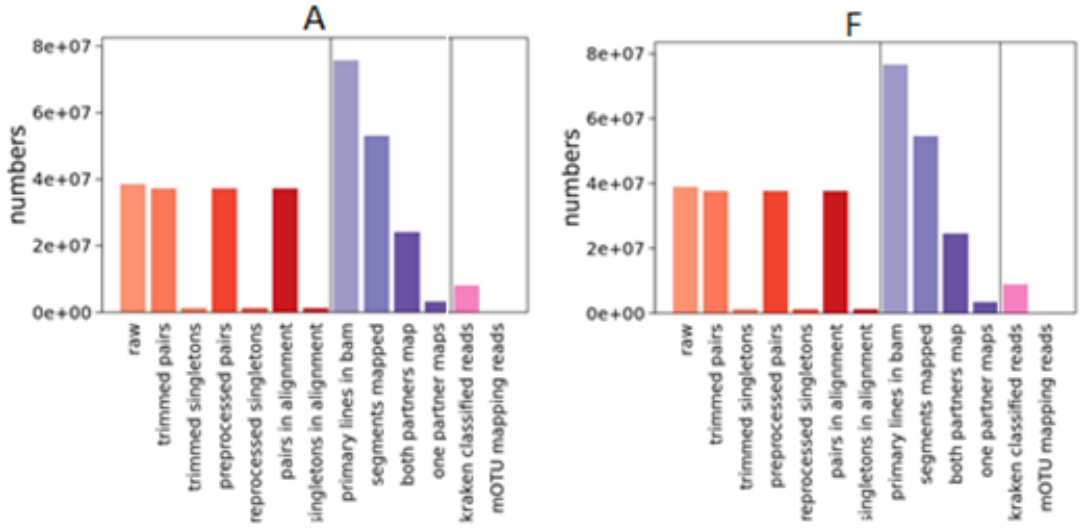

Figure 3: *The IMP3 pipeline shows how the number of sequencing reads is transformed during the pipeline. Starting with the pre-processing in red, followed by mapping to the assembly in blue and a taxonomic estimation in yellow and pink. A (left) and F (right) reference two samples.*

After the IMP3 pipeline ran to completion, most data was kept during the trimming step (trimmed pairs Figure 3). Around half the sequences could not be used in the making of the contigs (both partners map bar in Figure 3) and only a quarter of the reads could be classified (kraken classified reads bar in Figure 3). In more detail, the assembly by megahit of sample A used 38 million reads with 850 million base pairs forming 1.58 million contigs with an N50 of 596 base pairs and the longest contig being 50,265 base pairs long. The assembly by megahit of sample F used 39 million reads with 924 million base pairs forming 1.81 million contigs with a N50 of 562 base pairs and the longest contig being 58,877 base pairs long. So, the total assembly reduced the total data volume by 20 fold and resulted in 3.5 million contigs to be classified and to be analysed for a taxonomical profile by Eukdetect and Kraken2.

Classifiers

The different classfiers, Kraken2, Eukrep and Whokaryote, will all isolate the eukaryotic sequences from the assembly. The amount of eukaryotic contigs are compared to assess the capabilities of the different programs and their reliability.

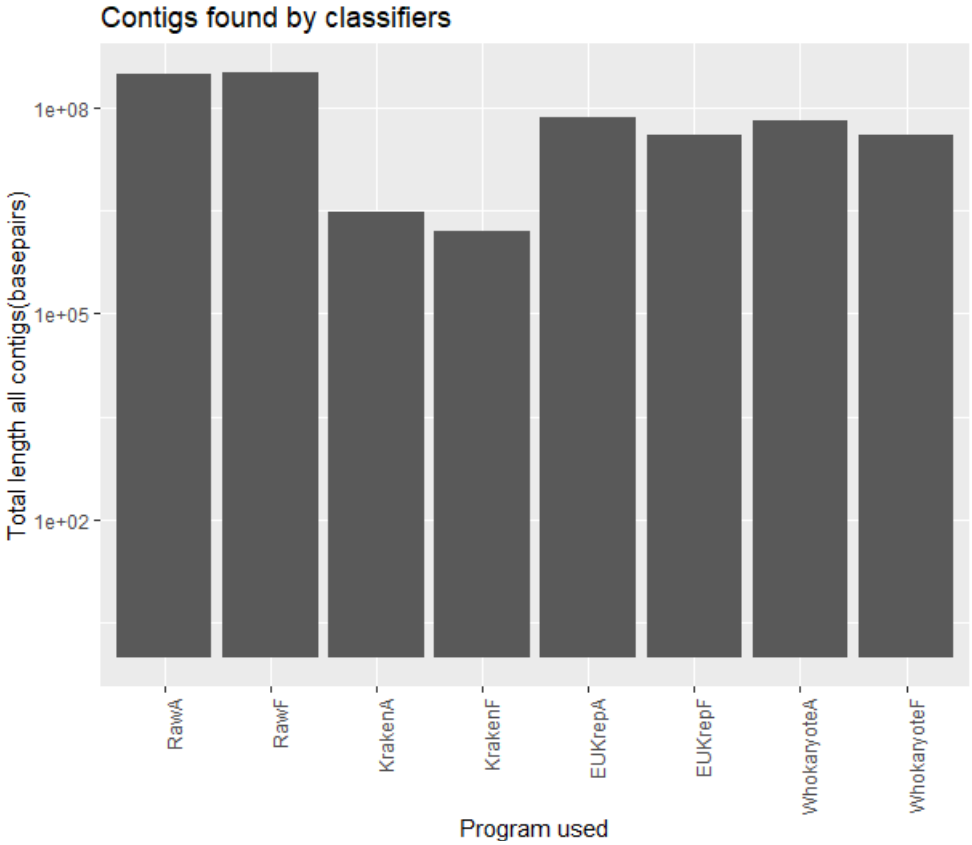

Figure 2: *This bar graph shows a comparison between the amount of eukaryotic sequences found by the different classifier programs. EUKrep and Whokaryote only looked at contigs larger then 800 base pairs. So, the bar graph only shows the contigs larger than 800bp.*

The classifiers are assessed on their capability to recognise contigs of 800 base pairs or longer as eukaryotic. The minimum is set at 800 base pairs, due to the fact that this is the minimum length a contigs needs to be reliably classified. Shorter contigs also tend to have lower coverage, so the information loss in later steps is limited.

All classifiers found twice the amount of eukaryotic sequences in sample A in comparison to sample F(Figure 4). The general taxonomic classifier Kraken2 was found ten times less eukaryotic sequences in comparison to the eukaryotic sequence detection classifiers EUKrep and Whokaryote. EUKrep and Whokaryote had similar levels of recognised eukaryotic sequences in the assembly. The overlap in contigs found was quiet low, around half of the contigs found per sample were unique to the program (Table 1). In summary, they found around 20% of the base pairs in sample A were eukaryotic and around 10% of the base pairs in sample F.

Table 1: *Unique and the same eukaryotic contigs found in the assembly by Eukrep and Whokaryote.*

Sample	Program	Unique contigs	Overlapping contigs
A	EUKrep	19343	16303
F	EUKrep	11448	8516
A	Whokaryote	8702	16303
F	Whokaryote	8543	8516

Taxonomic profiling

In this section Kraken2's and Eukdetect's estimates about the species of origin for each contig in the assembly are analysed. These estimates will be used to pick genomes to compare the assembly to, in order to see which species were found in the samples.

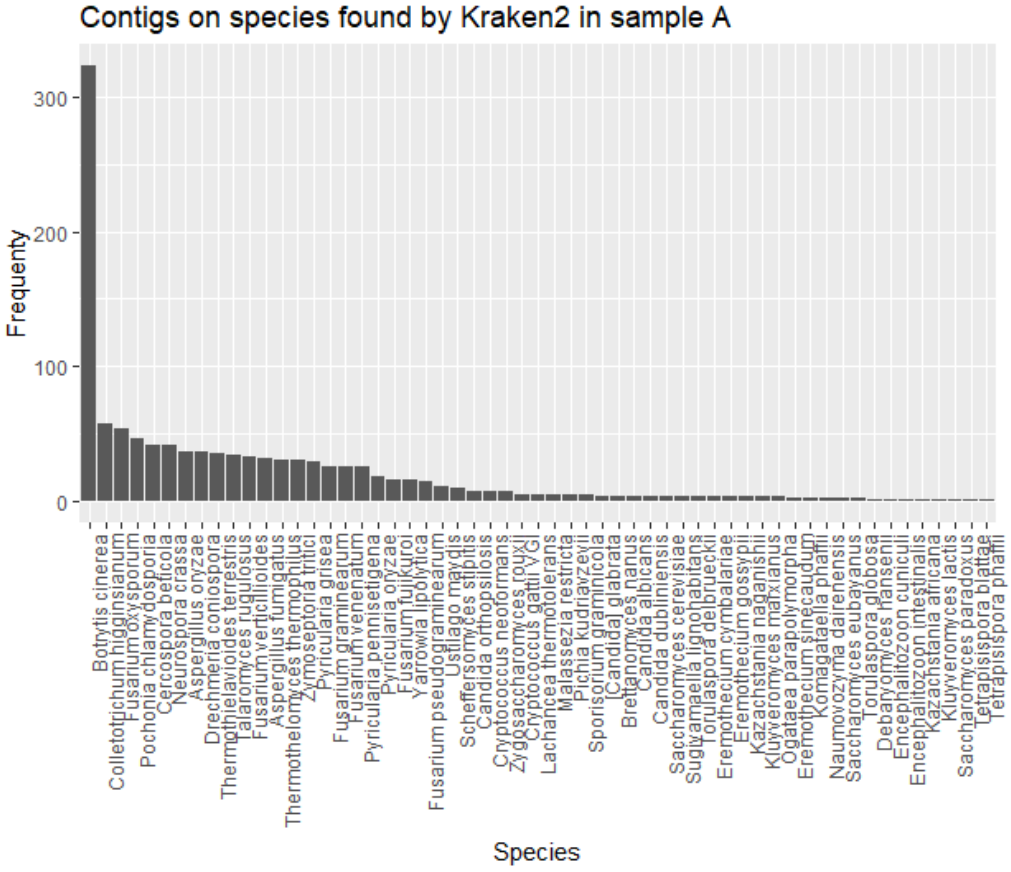

Figure 3: *Bar graph shows the amount of sequences found by Kraken2 to be belonging to fungal genomes in sample A.*

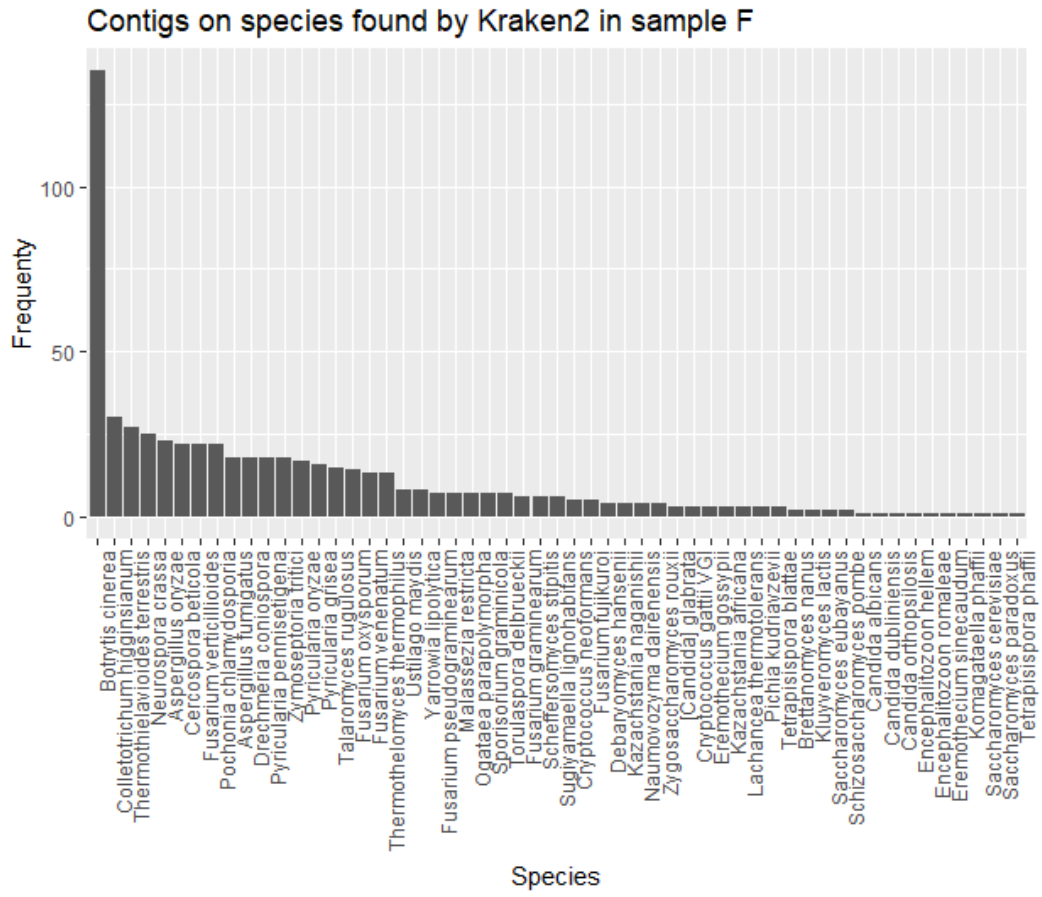

Figure 4: *Bar graph shows the amount of sequences found by Kraken2 to be belonging to fungal genomes in sample F.*

Both Kraken2 and Eukdetect found numerous *Fusarium* species. Surprisingly, Kraken2 found in both sample A and sample F an overwhelming amount of sequences relating to *Botrytis cinerea*(Figure 5 and 6) whilst Eukdetect did not relate any sequence to *Botrytis cinerea* (Figure 7 and 8). Looking at all the species found by Kraken2 a literature search for known plastic- or bio-degradable plastic-inhabiting or degrading species was performed. *Aspergillus oryzae* and *Fusarium oxysporum* were chosen for further analysis, as they were shown to thrive near plastic(Spina et al., 2021).The plastic in that study was polyethylene (PE), which is a non-biodegradable plastic but they might also thrive near PBSA. *Botrys cinerea* and *Aspergillus oryzae* are also found to be able to degrade large carbons like crude oils (Olukunle and Oyegoke, 2016), which in essence are similar to biodegradable plastics. *Pyricularia oryzae* was found to be able to help with the degradation of difficult to degrade organic waste (Awais et al., 2021), which shows its promise as a degrader of biodegradable plastics.

The species selected from the Eukdetect results were the top 6 most prominent fungi in both samples that were not yet selected using Kraken2. The following fungi were selected (Figure 7 and 8): *Alternaria alternata, Aureobasidium pullulans, Cladosporium sphaerosphermum, Filobasidium wieringae, Rhodotorula toruloides,* and *Stachybotrys chloromata.*

Lastly, *Tetracladium* was added because this was a prominent fungus found in the dataset by the previous researchers(Purahong et al., 2021).

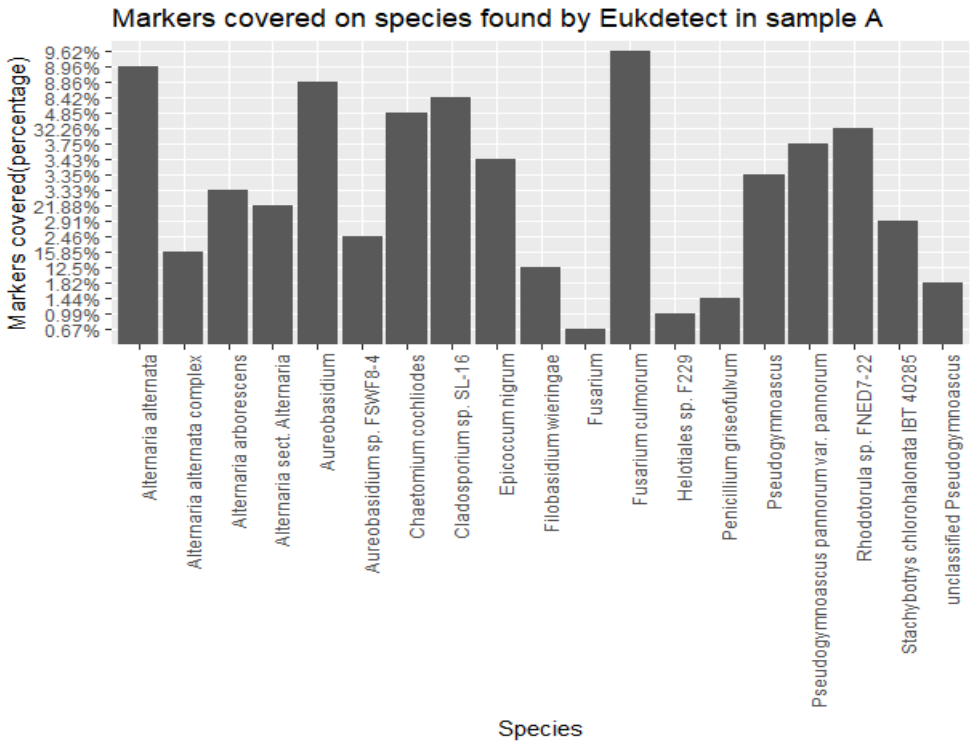

Figure 5: *Bar graph with percentage of marker genes covered by the sequences found by Eukdetect in sample A.*

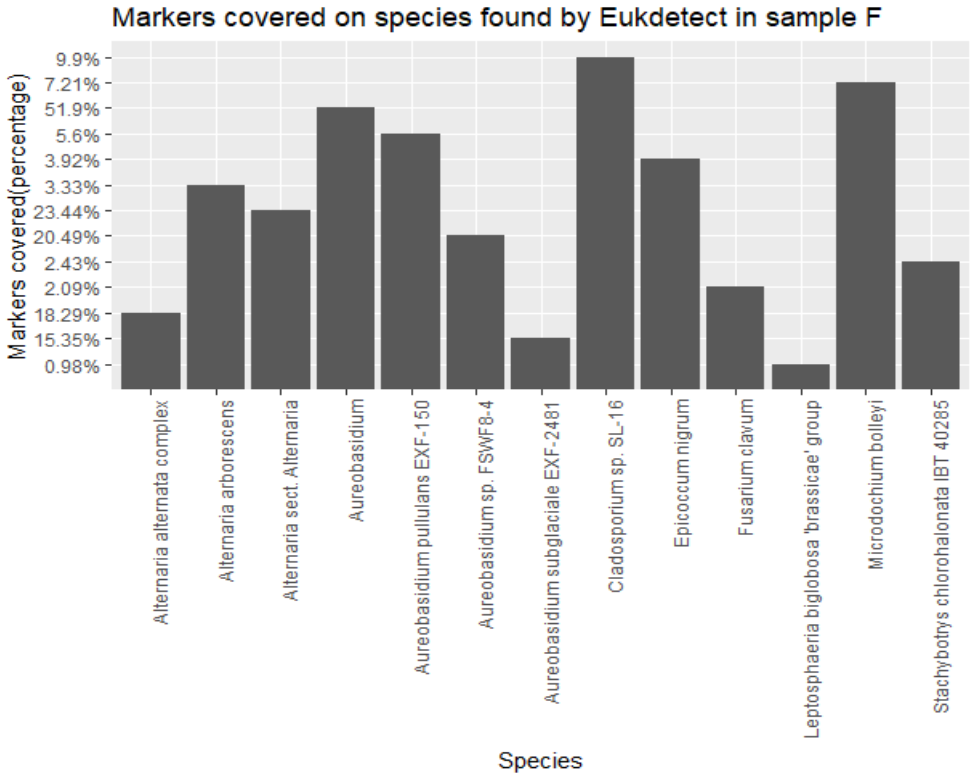

Figure 6: *Bar graph with percentage of marker genes covered by the sequences found by Eukdetect in sample F.*

In total the taxonomic profiling resulted in eleven different fungal species of interest (Table 2). Their genomes were downloaded from NCBI to form our reference library which will be aligned to the assembly.

Table 2: Species selected through taxonomic profiling and previous sequencing results.

Species	Method	Motivation
Tetracladium	ITS amplicon sequencing	Previous results
Aspergillus oryzae	Kraken2	Literature
Botritys cinerea	Kraken2	Literature
Fusarium oxysporum	Kraken2	Literature
Pyricularia oryzae	Kraken2	Literature
Alternaria alternata	Eukdetect	Statistics
Aureobasidium pullulans	Eukdetect	Statistics
Cladosporium sphaerospermum	Eukdetect	Statistics
Filobasidium wieringae	Eukdetect	Statistics
Rhodotorula toruloides	Eukdetect	Statistics
Stachobotrys chloronata	Eukdetect	Statistics

Alignment

The alignment will show with which species out of the reference library(table 2) our assembly aligns. The alignment was done with the eukaryotic sequences found by Whokarote and

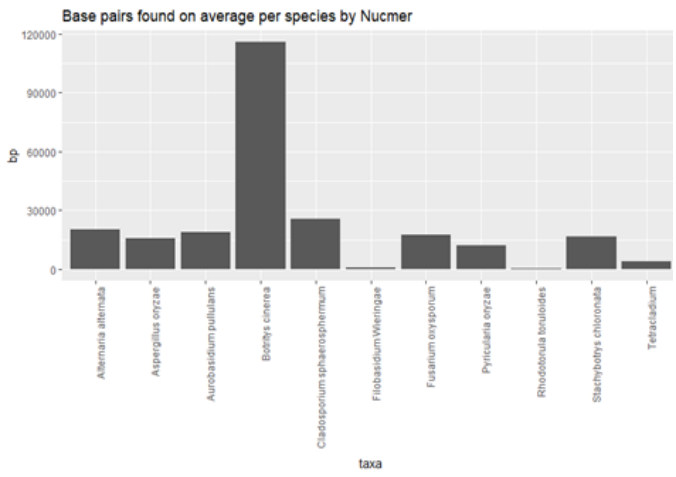

Figure 9: Bar graph showing the amount of base pairs aligning by using Nucmer to the most prominent species found by Kraken2 and Eukdetect. The average of the samples are presented.

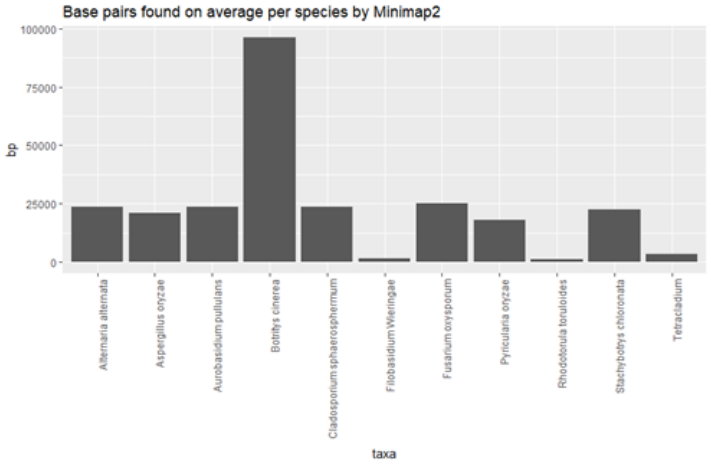

***Figure 10:** Bar graph showing the amount of base pairs aligning by using Minimap2 to the most prominent species found by Kraken2 and Eukdetect. The average of the samples are presented.*

EUKrep in sample A and F seperately using Mummer's Nucmer and Minimap2, the average combined results are shown(Figure 9 and 10). Minimap2 is less strict than Nucmer. However, similar results were found with both methods (Figure 9 and 10). Minimap2 found 5 times more contigs per species in comparison with Nucmer. But Minimap2 did not align 5 times the amount of contigs. So Minimap2 was better able to align smaller contigs. Otherwise, both Nucmer and Minimap2 found the same ratio of species, both aligning most sequences to *Botritys cinerae*(Figure 9 and 10). The highest coverage Nucmer reached on *Botritys cineraa* genome

***Table 3:** Base pair coverage by aligners, Nucmer and Minimap2, of Botritys cinerae genome in percentage.*

Sample	Nucmer (%)	Minimap2 (%)
Who A	0,2175	0,2589
Who F	0,2272	0,1605
EUKr A	0,2634	0,3292
EUKr F	0,1959	0,1657

was 0.26% (Table 3). For Minimap2 the highest coverage reached of the *Botritys cinerae* genome was 0.33% (Table 3). The other genomes had a lower coverage as they had less sequence aligned to them. Almost no sequences were aligned with *Filobasidium wieringae, Rhodorula toruloides*, and *Tetracladium*.

Discussion

This study successfully shows that it is possible for machine learning programs to differentiate in between prokaryotic and eukaryotic DNA and were able to give an estimation of which fungal species were in the samples.

Assembly and classification

The assembly of the reads for both samples ran successfully. Only a fragment of the reads ended up as singletons or were mapped without a partner. The partner-mapped contigs have higher information content than contigs without a partner, because they tend to be more contiguous/longer. The contigs were good enough for Kraken2 to discern fungal contigs (Figure 4).

EUKrep and Whokaryote found an almost equal amount of eukaryotes in the sample. EUKrep the more widely program, had on average a better alignment of the genome and was able to do more with sample F. This was the dataset half the size of that of sample A. So Eukrep was able to give a better estimation. Whokaryote did out perform on sample A the larger sample. This due to the fact that it is less strict so is capable to do more with smaller sequences. Additionally Whokaryote gives broad diagnostics on how it performed with and without Tiara, giving a more broad insight in how the results were achieved. In the end EUKrep is the stronger programmer giving more trustworthy results with a similar alignment coverage. Whokaryote would only preferably be used over EUKrep when there is a large dataset with smaller sequences.

EUKrep and Whokaryote both found twice the amount of fungi in sample A compared to sample F. So, either the futuristic circumstances in which sample F was different from sample A had a large impact on the amount of fungal growth on the biodegradable plastic mulch, but this does not seem to be the case(Purahong et al., 2021). Or there is a large variation between samples, either naturally or incurred by sampling. In which case sampling seems the most likely. As metagenomics and other bioinformatics methods scale better to larger datasets in comparison to in vitro methods that still rely heavily on the computational power of the human brain and manual labour, a larger dataset with more replicates or conditions would have benefited this project. The chance of finding species or having more confidence in the found species would have increased. Although the greater sampling and sequencing effort would have increased costs and lab work, this would only cost maximum of a day longer assembly time.

Profiling

The taxonomic profiles were surprising (Figure 5-8) as both Kraken2 and Eukdetect found mostly species not yet described in literature(Bandopadhyay et al., 2020; Moore-Kucera et al., 2014) as plastic degraders. The profilers also had vastly different results to one another. Kraken2 primarily found sequences for *Botrys cinerea*, which is a necrotrophic smut mold, primarily infecting grapes and other smaller fruit(Dean et al., 2012). It has also been found to be able to degrade crude oils(Olukunle and Oyegoke, 2016). In contrast, Kraken2 only found a couple of sequences aligning to *Aspergillus*, which was previously found to be a more prominent degrader of plastic(Moore-Kucera et al., 2014). Eukdetect did not find any sequences that aligned with *Botrytis cinerea.* Eukdetect's results also did not detect one

dominating fungi, but mostly molds and yeast-like fungi. All were likely to be found in soil and none are described in literature as plastic degraders.

Alignment

The results of aligning contigs to the candidate taxa's genomes found by the profilers were more comparable to the Kraken2 results than to the Eukdetect results. This was surprising, as Eukdetect was benchmarked as very reliable (Lind and Pollard, 2021). Eukdetect is also made to find these eukaryotes and Kraken2 has a more broader functionality(Wood et al., 2019). Both Nucmer and Minimap2 found that most sequences aligned with *Botrytis cinerea.* Nucmer and Minimap2 had similar results with respect to the ratio of sequences aligning to the different reference genomes (Figure 9 and 10) and in coverage of the *Botrytis cinerea* genome(Table 3). Minimap2 was less strict than Nucmer and hence was able to align more sequences with the reference library(Table 2). The more stringent Nucmer was not able to align these sequences. Regarding whether to use Nucmer or Minimap2, Minimap2 was found to be able to align more sequences and its results are easier to handle. But Nucmer is more strict and gives a more full analysis of the alignment which makes interpretation of the results easier.

Also, the species suggested by previous ITS region amplicon sequencing results, *Tetracladium,* was not found to be very prevalent in the samples. Several species out of the reference library got more alignments than *Tetracladium*. This means that our metagenomics method produced different results than the amplicon sequencing(Purahong et al., 2021; Tanunchai et al., 2021).

Conclusion

In conclusion, this bioinformatics pipeline gives new unique results with regards to the analysis of fungal species to be found on biodegradable plastics. As of right now it's methods are not strong enough yet to give definite results. The research field could benefit from using this pipeline more or having all of their results assessed in a similar manner to find data on fungi that might be hidden behind the copious amounts of bacterial DNA and doing so bettering the functionality of this pipeline. No clear microbial communities can be concluded to live on this biodegradable mulch. This leaves a lot to be researched into the field of biodegradable plastics and its degradation in nature as its use becomes more widespread and the material and degradation products become more abundant in our ecosystems.

Acknowledgments

First and foremost I'd like to thank my supervisor Anna Heintz-Buschart for her willingness and availability to answer even the simplest question and the whole team at Swammerdam Institute for Life Sciences for hosting me for my bachelor thesis. Secondly the Helmholtz Centre for Environmental Research for acquiring the data for this study and allowing me to use it. Thirdly all the creators of the programs used in this study for publishing them and allowing me to use them. Fourthly Franciska de Vries for examining my thesis. Finally all my friends, family and especially my fellow students who were there to give me additional motivation or talk about my problems with whenever I needed it.

Literature

Assessment of agricultural plastics and their sustainability: A call for action, 2021. , Assessment of agricultural plastics and their sustainability: A call for action. FAO. https://doi.org/10.4060/cb7856en

Awais, M., Fatma, S., Naveed, A., Batool, U., Shehzad, Q., Hameed, A., 2021. Enhanced biodegradation of organic waste treated by environmental fungal isolates with higher cellulolytic potential. Biomass Conversion and Biorefinery 1, 1–16. https://doi.org/10.1007/S13399-021-01932-W/FIGURES/6

Bandopadhyay, S., Liquet y González, J.E., Henderson, K.B., Anunciado, M.B., Hayes, D.G., DeBruyn, J.M., 2020. Soil Microbial Communities Associated With Biodegradable Plastic Mulch Films. Frontiers in Microbiology 11, 2840. https://doi.org/10.3389/FMICB.2020.587074/BIBTEX

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Bowe, A., Onodera, T., Sadakane, K., Shibuya, T., 2012. Succinct de Bruijn Graphs. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7534 LNBI, 225–235. https://doi.org/10.1007/978-3-642-33122-0_18

Compeau, P., Pevzner, P., 2015. Bioinformatics algorithms an active learning approach, 3rd edition. ed. Active Learning Publishers, LLC, San Diego.

Dean, R., van Kan, J.A.L., Pretorius, Z.A., Hammond-Kosack, K.E., di Pietro, A., Spanu, P.D., Rudd, J.J., Dickman, M., Kahmann, R., Ellis, J., Foster, G.D., 2012. The Top 10 fungal pathogens in molecular plant pathology. Molecular Plant Pathology. https://doi.org/10.1111/j.1364-3703.2011.00783.x

ID 853344 - BioProject - NCBI [WWW Document], 2022. URL https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA853344 (accessed 7.4.22).

Kumar, R., Manna, C., Padha, S., Verma, A., Sharma, P., Dhar, A., Ghosh, A., Bhattacharya, P., 2022. Micro(nano)plastics pollution and human health: How plastics can induce carcinogenesis to humans? Chemosphere 298. https://doi.org/10.1016/J.CHEMOSPHERE.2022.134267

Li, D., Liu, C.M., Luo, R., Sadakane, K., Lam, T.W., 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676. https://doi.org/10.1093/BIOINFORMATICS/BTV033

Li, H., 2021. New strategies to improve minimap2 alignment accuracy. https://doi.org/10.1093/bioinformatics/btab705

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. https://doi.org/10.1093/bioinformatics/bty191

Lind, A.L., Pollard, K.S., 2021. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. https://doi.org/10.1186/s40168-021-01015-y

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018. MUMmer4: A fast and versatile genome alignment system. https://doi.org/10.1371/journal.pcbi.1005944

Moore-Kucera, J., Cox, S.B., Peyron, M., Bailes, G., Kinloch, K., Karich, K., Miles, C., 2014. ENVIRONMENTAL BIOTECHNOLOGY Native soil fungi associated with compostable plastics in three contrasting agricultural settings. https://doi.org/10.1007/s00253-014-5711-x

Narayanasamy, S., Jarosz, Y., Muller, E.E.L., Laczny, C.C., Herold, M., Kaysen, A., Heintz-Buschart, A., Pinel, N., May, P., Wilmes, P., 2016. IMP: a pipeline for reproducible integrated metagenomic and metatranscriptomic analyses. bioRxiv 039263. https://doi.org/10.1101/039263

Olukunle, O.F., Oyegoke, T.S., 2016. Biodegradation of Crude-oil by Fungi Isolated from Cow Dungcontaminated soils. Nigerian Journal of Biotechnology 31, 46–58. https://doi.org/10.4314/njb.v31i1.7

Pathan, S.I., Arfaioli, P., Bardelli, T., Ceccherini, M.T., Nannipieri, P., Pietramellara, G., 2020. Soil pollution from micro-and nanoplastic debris: A hidden and unknown biohazard. Sustainability (Switzerland) 12, 1–31. https://doi.org/10.3390/SU12187255

Pronk, L.J.U., Medema, M.H., 2021. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. bioRxiv 2021.11.15.468626. https://doi.org/10.1101/2021.11.15.468626

Purahong, W., Wahdan, S.F.M., Heinz, D., Jariyavidyanont, K., Sungkapreecha, C., Tanunchai, B., Sansupa, C., Sadubsarn, D., Alaneed, R., Heintz-Buschart, A., Schädler, M., Geissler, A., Kressler, J., Buscot, F., 2021. Back to the Future: Decomposability of a Biobased and Biodegradable Plastic in Field Soil Environments and Its Microbiome under Ambient and Future Climates. Environmental Science and Technology 55, 12337–12351. https://doi.org/10.1021/acs.est.1c02695

Razmyslovich, D., Marcus, G., Gipp, M., Zapatka, M., Szillus, A., 2010. Implementation of Smith-Waterman algorithm in OpenCL for GPUs. Proceedings of the 9th Int. Workshop on Parallel and Distributed Methods in Verification, PDMC 2010 - Joint with the 2nd Int. Workshop on High Performance Computational Systems Biology, HiBi 2010 48–56. https://doi.org/10.1109/PDMC-HIBI.2010.16

Spina, F., Tummino, M.L., Poli, A., Prigione, V., Ilieva, V., Cocconcelli, P., Puglisi, E., Bracco, P., Zanetti, M., Varese, G.C., 2021. Low density polyethylene degradation by filamentous fungi. Environmental Pollution 274, 116548. https://doi.org/10.1016/J.ENVPOL.2021.116548

Tanunchai, B., Juncheed, K., Wahdan, S.F.M., Guliyev, V., Udovenko, M., Lehnert, A.S., Alves, E.G., Glaser, B., Noll, M., Buscot, F., Blagodatskaya, E., Purahong, W., 2021. Analysis of microbial populations in plastic–soil systems after exposure to high poly(butylene succinate-co-adipate) load using high-resolution molecular technique. Environmental Sciences Europe 33. https://doi.org/10.1186/s12302-021-00528-5

West, P.T., Probst, A.J., Grigoriev, I. v., Thomas, B.C., Banfield, J.F., 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome Research 28, 569–580. https://doi.org/10.1101/gr.228429.117

Windsor, F.M., Durance, I., Horton, A.A., Thompson, R.C., Tyler, C.R., Ormerod, S.J., 2019. A catchment-scale perspective of plastic pollution. Global Change Biology 25, 1207–1221. https://doi.org/10.1111/GCB.14572

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. https://doi.org/10.1186/s13059-019-1891-0