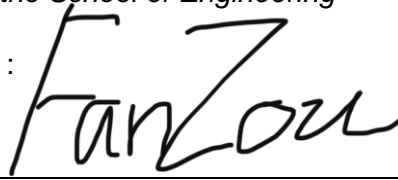
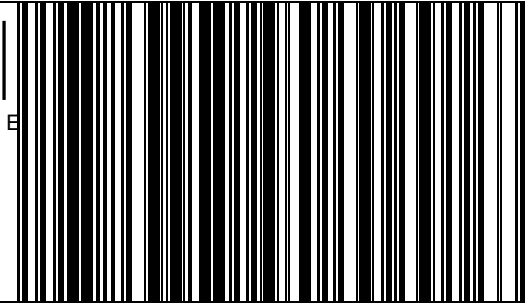


Coursework Declaration and Feedback Form

The Student should complete and sign this part

Student Number: 2596373Z	Student Name: Fan Zou
Programme of Study (e.g. MSc in Electronics and Electrical Engineering): Master Of Science in Computer System Engineering	
Course Code: ENG5059P	Course Name: MSc Project
Name of First Supervisor: Dr Umer Zeeshan Ijaz	Name of Second Supervisor: Professor Barabara Mable
Title of Project: Building a plant sequence reference database	
Declaration of Originality and Submission Information	
<i>I affirm that this submission is all my own work in accordance with the University of Glasgow Regulations and the School of Engineering requirements</i> Signed (Student) : 	
Date of Submission : 8/20/2021	

<i>Feedback from Lecturer to Student – to be completed by Lecturer or Demonstrator</i>	
Grade Awarded: Feedback (as appropriate to the coursework which was assessed):	
Lecturer/Demonstrator:	Date returned to the Teaching Office:



University
of Glasgow

Building a plant sequence reference database

Fan Zou

2596373Z

Supervised by Dr Umer Zeeshan Ijaz

A thesis submitted in partial fulfilment of the requirements for the
degree of

MASTER OF SCIENCE IN COMPUTER SYSTEM ENGINEERING

Contents

Abstract.....	1
Acknowledgements.....	2
1 Background and Aims.....	3
1.1 Background of DNA barcode	3
1.2 Background of rbcL and trnL genes in plants.....	4
1.3 Background of The Naive Bayes classifier.....	5
1.4 Background of Qiime2.....	5
1.5 Amplicon Sequence Variants Overview	6
1.6 Aims and Objectives	7
2 Methods.....	7
2.1 Creation of database.....	7
2.1.1 rbcL database	7
2.1.2 trnL database	8
2.2 Taxonomy file and train classifier	8
2.3 Verification of database	10
2.3.1 rbcL – study and amplicon sequences workflow	10
2.3.2 trnL – study and amplicon sequences workflow	11
2.3.3 Database classification workflow for ASVs	11
3 Results	11
3.1 Summarise of database and taxonomy.....	11
3.1.1 rbcL database	12
3.1.2 Taxonomy of rbcL databas.....	12
3.1.3 trnL database	12
3.1.4 Taxonomy of trnL database	12

3.2	Verification of database results.....	12
3.2.1	rbcl amplicon sequence variants study result	13
3.2.2	rbcl taxonomy result	17
3.2.3	trnL amplicon sequence variants study result	19
3.2.4	trnL taxonomy result	23
4	Discussion	25
4.1	Performance of rbcl as a barcode	25
4.2	Performance of trnL as a barcode	25
4.3	Weaknesses and Improvements	26
5	Conclusions.....	27
	Reference.....	28
	Appendix I	31
	Appendix II.....	33
	Appendix III	34

Abstract

DNA barcode refers to a standard, easy to amplify, relatively short DNA fragment, which can represent the species in the organism by containing enough variability to allow for variations across phyla. When an unknown species or part of a species is found, researchers track the DNA barcodes of its tissues and compare them with other barcodes in an international database of DNA sequences. If it matches one of them, researchers can therefore identify the species. Around the world, scientists are taking action to study DNA barcodes for different biological communities and to make these results of barcodes available to help understand species. For different species of plants, they have different chloroplast genes to choose from, with no consensus on which genes best identify species. The core DNA barcode of plant is the coding region in chloroplast. The common coding genes in chloroplast genome are *rbcl*, *matK*, *ndhF*, *ATPB*, *rps16*, *rpl16*, *ITS2*, *trnL-F*, *trnt-l*, etc. In my research, in order to build a high-quality plant genes database, I used *rbcl* and *trnL* DNA sequence as barcodes, download all sequence data from NCBI Nucleotide of these regions to set up my database. Once the database is set up, I can build classifiers from the *rbcl* or *trnL* DNA database sequence data I have created to quickly identify species using taxonomic methods. To confirm the utility of the created databases and classifiers (for both *rbcl* and *trnL* regions), I searched the literature for studies that had used these regions and had the data publically available on NCBI. I then re-analysed this data using my classifier using the QIIME2 pipeline. Finally, the results obtained from the analysis were used to verify my database.

Keywords: DNA barcoding, Plants, Metabarcoding, *rbcl*, *trnL*, Bioinformatics, Database

Acknowledgements

I would like to express my gratitude to Dr Umer Zeeshan Ijaz and Dr Ciara Keating for providing me with the opportunity to research this project. I really appreciate the time Dr Umer spent time meeting with me every week to discuss progress, he clearly pointed out the main points I should focus on in the project and gave me advice on project planning. Dr Ciara Keating always responded to my messages on Microsoft Teams as soon as possible, provided guidance on all aspects and taught me tutorials on handling data through practical demonstrations, and she shared relevant literature to broaden my She also shared relevant literature to broaden my knowledge in this area.

1 Background and Aims

1.1 Background of DNA barcode

When we go to the supermarket to check out, the salesperson will use a scanner to scan the barcode on the back of the product and the name, number and price of the product will appear on the computer screen. A barcode is a graphical representation of a set of information, consisting of a number of black bars and blanks of varying widths arranged according to certain coding rules. In addition to the product name and price, barcodes can also indicate the manufacturer of the item, the date of production, the classification number of the book, the origin and destination of the mail and many other information, and are widely used in the circulation of goods, books, postal services, banks and other fields, it can be said that the use of barcodes greatly facilitate our life.

So is the DNA barcode a similar graphic identifier? What instrument is used to scan it?

DNA barcoding is a powerful tool for specimen identification (Jianping Xu et al 2016).

Deoxyribonucleic acid (DNA) is the genetic material of living things and consists of four different deoxyribonucleotides. DNA of different lengths and sequences can be stably transcribed into different RNAs and translated into different proteins, just as a barcode consisting of black bars and blanks can represent a commodity. In simple terms, a DNA barcode is a DNA sequence that is unique to a species and stable within the species, and requires molecular biology (PCR amplification, etc.) to "read" the sequence.

DNA barcoding is a tool for species identification that uses internationally agreed protocols and regions of DNA to create a global database of living organisms (Hebert PDN et al 2003). International initiatives are taking place across hundreds of countries to DNA barcode the world's biodiversity and make these data publicly available to all users (Hebert PDN et al 2005).

I did some searches on the research on DNA barcodes. Many researchers use this technology to identify unknown genes and perform data analysis on them. In addition, if we want to study some biological information that cannot be directly obtained, we DNA barcode technology can also be used for analysis, such as the pollination network of bees. For humans, it is very difficult to directly observe the pollination of bees all the time, because there are many pollination sites for bees and a branch of the colony is located in for pollination in different places, it is difficult for us to know all the locations directly. At this time, we only need to analyze the pollen DNA in the honey, and the results will be easily obtained by using DNA barcodes here.

The use of DNA barcoding data can help us to reconstruct a clear molecular phylogeny for the study of species, a technique that offers great value to the community of ecologists and evolutionary biologists in terms of biodiversity, and to those who use phylogenetic data to

address the ecological and evolutionary mechanisms that promote and maintain species diversity (Harvey PH et al 2006).

1.2 Background of *rbcl* and *trnL* genes in plants

It is estimated that there may be about 380,000 land plant species in the world, including about 352,000 angiosperms, 1,300 gymnosperms and 13,000 ruderals, ferns and ferns, and if we are to build a plant database? the DNA barcoded genes we tag must effectively use existing research findings and knowledge to successfully classify such a huge number of species effectively (Fazekas A et al 2012). Fortunately, over the years the world's herbaria have accumulated a wealth of important plant material that has been identified and preserved over the years into the taxonomic expertise we need (de Vere N et al 2012).

Researchers have found that because the standard animal DNA barcodes that comprise part of the mitochondrial gene CO1 have evolved very slowly in plants, animal DNA barcodes cannot be used as useful DNA barcodes for plants (Fazekas AJ et al. 2008). However, the plastids and ribosomes inside plant genes, when used as DNA barcodes, only exhibit low discriminatory power (Hollingsworth PM et al., 2011).

The Consortium for the Barcode of Life (CBOL) Plant Working Group proposed the chloroplast gene *rbcl* and *matK* as the core barcodes of plant species, as well as intergenic sequence *trnH-psbA* and nuclear gene ITS as the supplement barcodes (CBOL Plant Working Group at 2009).

No particularly prominent gene fragment was used for barcoding. Selecting standard plant DNA barcodes is difficult because there are many studies illustrating that all the different gene regions have different strengths and weaknesses (Fazekas AJ et al 2008, Chase MW et al 2005, Kress WJ et al 2005, Newmaster SG et al 2006, Cowan RS et al 2006, Kress WJ et al 2007, Chase MW et al 2007, Newmaster SG et al 2008). From the characteristics of DNA barcodes it is easy to see that DNA barcodes need to be standardised, minimalist and scalable. Typically, studies select one or a few standard loci as DNA barcode regions, and these loci need to be routinely and reliably sequenced in a very large and diverse sample set to produce easily comparable data that allows species to be distinguished from each other. So we decided to use *rbcl* and *trnL* separately to build the barcode database.

The *rbcl* gene is part of a DNA sequence located in chloroplast DNA that researchers believe has characteristics that can be used as a DNA barcode (Les D H et al 1991) because the *rbcl* gene region contains these universal, easily amplified and analysed features (Newmaster S G et al 2006). Researchers have found that this gene can provide many features to better allow them to study plant phylogeny, and it is approximately 1400 bp in length (CBOL Plant Working Group, 2009). The *rbcl* gene sequence has a low level of mutation compared to other barcodes in chloroplast DNA, and researchers have also found that the *rbcl* gene sequence has a high level of similarity between species (Kellogg E A et al 1997). *rbcl* also has the

advantage of having a low level of mutation (Papuangan et al 2019).

What is trnL gene? TrnL-F of chloroplast genome of terrestrial plants is composed of transfer RNA genes trnluua and trnfgaa arranged in series, which are separated by non-coding spacer. (KUSUMADEWI SRI YULITA et al 2013).

1.3 Background of The Naive Bayes classifier

A classifier is a tool that uses some training data to understand the relationship between a given input variable and a category. According to the research report, the naive Bayes classification method performs well on problems similar to the classification of sequence data (Karavaiko et al 2000).

A naive Bayesian classifier is a classification algorithm that introduces the assumption of conditional independence of attributes in a probabilistic framework. Let the sample data set $D = \{x_1, x_2, \dots, x_N\}$ each sample $x_i = \{x_{i1}, x_{i2}, \dots, x_{iN}\}$ be an n-dimensional attribute vector and the category label $y = \{c_1, c_2, \dots, c_N\}$. That is to say D can be classified into N classes, Based on Bayes' theorem, the posterior probability $P(c|x)$ can be expressed as

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

where $P(c)$ is the class prior probability; $P(x|c)$ is the class conditional probability of the sample x with respect to the label c , also known as the likelihood; and $P(x)$ is the evidence factor.

Estimating the posterior probability $P(c|x)$ in a Bayesian classifier translates into estimating the prior probability $P(c)$ and the likelihood $P(x|c)$, $P(c)$ is estimated by the frequency of occurrence of each class of samples, but $P(x|c)$ is the joint probability over all attributes of x , which is very difficult to estimate.

So the assumption of conditional independence of attributes introduced by naive Bayes can be expressed as

$$P(c|x) = \frac{P(c) \prod_{i=1}^N P(x_i|c)}{P(x)}$$

where n is the number of attributes, and since $P(x)$ is the same for all classes, the Bayesian decision criterion gives

$$h_b(x) = \arg \max_{c \in y} P(c) \prod_{i=1}^N P(x_i|c)$$

This is the expression for the naive Bayesian classifier.

1.4 Background of Qiime2

QIIME 2 (Bolyen et al., 2019) is a new version of the microbiome analysis process QIIME, written in python 3 and fully replaced in January 2018, and represents the standard for end-of-life (computationally reproducible) analysis methods. A powerful, scalable and decentralised microbiome analysis package with an emphasis on transparency in data analysis,

QIIME 2 enables researchers to start their analysis from raw DNA sequences and obtain publication-quality statistical and graphical results directly. It has these advantages: it is easier to install, once QIIME was difficult for many researchers to install, from QIIME 2 onwards it introduces its Miniconda package manager, which can be easily installed without administrator privileges, and a docker image is released, which can be downloaded and run. At the same time, it is versatile in its use, supporting command line mode (q2cli) as well as a graphical user interface, q2studio; and the Artifact API (similar to IPython notebook), which Python users love. It also supports standardisation of the analysis process, full procedural documentation and repeatability, so that users are not confused about what to do next and can easily view the records they have analysed. It also has enhanced visualisation, making it easier and more beautiful to visualise results than QIIME, and has new interactive graphical results that can be clicked on to view details and make analysis easier. In terms of collaboration, projects can rarely be completed by a single group, and the graphs of results can be easily shared between multiple people and locations, making it suitable for today's research collaboration needs. It is extensible, supporting custom functions and adding to the analysis process; masters can write their own packages and add to the QIIME2 process now. And with QIIME2 the analysis is repeatable, with a newly defined file system that includes the analysis data, as well as the analysis process and results, and the results of each step can be traced back through the entire analysis process, making it easy to check and repeat.

1.5 Amplicon Sequence Variants Overview

Amplicon sequencing variants (ASV) have been proposed as an alternative to operational taxonomy (OTU) for the analysis of biomes. ASVs are becoming more and more popular, partly because they want to reflect a finer level of classification, because they do not cluster sequences based on distance-based thresholds (Schloss et al 2021).

The ASV method will first determine which exact sequences have been read, and how many times each exact sequence has been read. These data will be combined with the error model of the sequencing operation so that similar readings can be compared to determine the probability that a given reading at a given frequency is not caused by sequencer error. This essentially creates a P value for each exact sequence, where the null hypothesis is equivalent to the exact sequence due to sequencing errors.

After this calculation, the sequence is filtered according to a certain confidence threshold, leaving a set of precise sequences with a defined statistical confidence. Since these are precise sequences that can be generated without clustering or reference databases, it is easy to compare ASV results between studies using the same target area. In addition, a given target gene sequence should always produce the same ASV, and the given ASV as an accurate

sequence can be compared with a higher-resolution reference database, so that the species level can be more accurately identified (Callahan BJ et al 2019).

1.6 Aims and Objectives

The goal of this paper is to develop a plant reference database using the rbcL and trnL regions, then train classifiers for each region, eventually validate these classifiers with other studies, and generate summary statistics on how many unique taxonomic groups (genera, families, species, etc.) were found. Sequence data for the rbcL and trnL regions were downloaded from inside the NCBI gene database, and corresponding gene sequence files and taxonomic files were generated for these sequences. These taxonomic files are useful for our study of biodiversity and evolutionary processes. After building the rbcL and trnL plant barcode databases, we will use the classifier trained on the data in this paper to classify the research metadata for the relevant gene regions on the NCBI, to obtain classification results and representative sequences, and we will verify the validity of our database.

2 Methods

2.1 Creation of database

In general, a gene database is like a library, each organism is like a book, and the barcode genes is their barcode. The most basic database needs to have a gene sequence file, which stores all the gene sequences. In this study, we need to build a plant gene barcode database. First, we need the barcode genes of all plants, and also the plant taxonomy data behind them, such as which phylum, which class, and so on. For the past years, the National Center for Biotechnology Information (NCBI) Nucleotide database has always been an important resource for genomic, genetic, and proteomic research. The nucleoside database is a collection of sequences from multiple sources such as GenBank, RefSeq, TPA and PDB. We decided to download the data we needed from here.

2.1.1 rbcL database

NCBI Nucleotide search terms used:

rbcL sequences: 'rbcL[All Fields] OR rubisco[All Fields] AND (plants[filter] AND ("0"[SLEN] : "10000"[SLEN]))'(Richardson et al 2020)

rbcL downloaded on 06/07/2021

The gene file we downloaded contains the accession number, such as AB003566.1 and its attribute description, such as tRNA-Phe of *Primula cuneifolia* chloroplast DNA, partial sequence and finally its gene sequence, data like this is stored one after another in files, all we need for our database is its sequence files and their taxonomy, now We have the sequence file here, but we don't need the attribute descriptions, because the goal of our database is not the attribute descriptions, even the attribute descriptions of our database can be said to be its rbcL genes, a segment of rbcL genes can represent the identity of this segment of genes, and

this is exactly what we need to do, all we need to do is to remove the attribute descriptions from the file and just All we need to do is to remove the attribute descriptions from the file and keep only its accession number and its rbcL gene. After removing the attribute descriptions from the downloaded raw file, we are left with the accession number we need and the rbcL gene it corresponds to. The complete procedure for creating the trnL database is shown in Appendix I.

2.1.2 trnL database

NCBI Nucleotide search terms used:

trnL sequences: 'trnL[All Fields] AND plants[filter] AND ("0"[SLEN] : "10000"[SLEN])'(Richardson et al 2020)

trnL downloaded on 06/07/2021

The trnL database creation process is similar to that of rbcL. The complete procedure for creating the trnL database is shown in Appendix I.

2.2 Taxonomy file and train classifier

As mentioned earlier in this article, the purpose of this article is to create a genetic barcode database. Our purpose is to make it easier for researchers to distinguish a piece of unknown DNA. It is impossible to complete these functions only with data. When we get the rbcL or trnL genes of some unknown plants, how do we distinguish them? It is very difficult to compare the gene pairs in the database section by section, and if it is completely unknown, a completely new species, we are very It is difficult to know the specific categories of these plants. All we need to get the taxonomy files of the current database, and get a classifier based on the data of our database based on their genes. In this way, if we get a piece of unknown gene and use the trained classifier generated by our database to classify, we will get its taxonomy with a high probability to identify its identity. The research in this chapter It is how to produce taxonomy files and how to train a classifier based on our database.

As for the taxonomy file. In the downloaded file, it is easy to see that each rbcL gene has its own accession number. each gene sequence in NCBI has an accession number, which is given when the database accepts the sequence, and is a unique number, just like a human ID number. The main use is to search the NCBI database. When we search the NCBI database for the accession number, we get the origin of the gene, as shown in Figure 1 below, we can easily see that the NCBI database contains the locus of the gene as ab003566395 bp dna linear pln 28-apr-2007, definition as *primula cuneifolia* chloroplast dna for trna-phe, partial sequence, and accession, version, keywords, source, and other things. Obviously, organism is the taxonomic data we need. We can see that the owner of this gene belongs to the kingdom of Eukaryota, phylum of Streptophyta, class of Magnoliopsida, order of Ericales, family of Primulaceae, Genus of *Primula*. In other words, we only need to use the accession number of genes to get its taxonomic data, not their specific genes, so what we do next is to extract the

accession number of all rbcL-gene plants, and then to get their organism, that is, taxonomic data, and to use them with ID to indexed, stored in a file. Here we decided to use R library rentrez to separate the accession number from the file and store it as a separate file, using a crawler technique to crawl the taxonomy of each accession number separately and finally store them as a taxonomy file based on the index of the accession number. The complete procedure for creating the taxonomy file is shown in Appendix II.

```

LOCUS      AB003566                395 bp    DNA     linear   PLN 28-APR-2007
DEFINITION Primula cuneifolia chloroplast DNA for tRNA-Phe, partial sequence.
ACCESSION  AB003566
VERSION    AB003566.1
KEYWORDS    tRNA-Phe; trnF; tRNA-Phe(GAA).
SOURCE      chloroplast Primula cuneifolia
  ORGANISM  Primula cuneifolia
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetales; asterids; Ericales; Primulaceae; Primula.
REFERENCE   1
  AUTHORS   Fujii, N., Ueda, K., Watano, Y. and Shimizu, T.
  TITLE     Intraspecific sequence variation in chloroplast DNA of Primula
            cuneifolia Ledeb. (Primulaceae)
  JOURNAL    J. Phytogeogr. Taxon. 43, 15-24 (1995)
REFERENCE   2
  AUTHORS   Fujii, N., Ueda, K., Watano, Y. and Shimizu, T.
  TITLE     Further analysis of intraspecific sequence variation of chloroplast
            DNA in Primula cuneifolia Ledeb. (Primulaceae): implications for
            biogeography of the Japanese alpine flora
  JOURNAL    J. Plant Res. 112, 87-96 (1999)
REFERENCE   3 (bases 1 to 395)
  AUTHORS   Fujii, N., Ueda, K., Watano, Y. and Shimizu, T.
  TITLE     Direct Submission
  JOURNAL    Submitted (03-MAY-1997) Noriyuki Fujii, Department of Biological
            Science, Graduate School of Science and Technology, Kumamoto
            University; 2-39-1, Kurokami, Kumamoto 860-8555, Japan
            (E-mail:nfujii@kumamoto-u.ac.jp, Tel:81-96-342-3474,
            Fax:81-96-342-3474)
FEATURES             Location/Qualifiers
     source            1..395
                       /organism="Primula cuneifolia"
                       /organelle="plastid:chloroplast"
                       /mol_type="genomic DNA"
                       /db_xref="taxon:49648"
                       /note="collection_site: Mt. Shozu"
     misc_feature       <1..368
                       /note="intergenic region between trnL(UAA) 3' exon and
                       trnF (GAA)"
     gene               369..>395
                       /gene="trnF"
     tRNA               369..>395
                       /gene="trnF"
                       /product="tRNA-Phe"
                       /note="anticodon: GAA"

ORIGIN
1  ctccctaact atgtatccta tctatactat atcttttttg ttagcagtta taaattaggt
61  atttttttga ttacttttac tatttgacaa atggatctga acagaaatgc tttctottat
121 tatcagtttt gcggtattat gctatttaac acatgtacaa atgaacatct ttgagcaaga
181 aatcccccgtg tgaatgattt tcggttcata ttattcgtac tgaaccttac atagttttcc
241 tttttttgaa aatccaagaa attacagggc cctcagaaga ctttaataat accotttttgt
301 tttttaattg acatagactc aatttattga catagactca ctagtaaaat gagtaggatg
361 tatcgggagt ggtcaggata gtcagctgg tagag

```

Figure 1: Search result of AB003566.1. Provided by NCBI

Once we have obtained the taxonomy file of our database, that is, now we have two files in our hands, one is the gene sequence data file that relies on the accession number and the other is the taxonomy file that relies on the accession number, both of them should correspond to the same database, now we need to use the relationship between the gene and the taxonomy file to In this paper, we use Qiime2 to generate our classifier with the naive bayes classifier approach. An introduction to the naive bayes classifier can be found in Section 1.3 of this paper.

The complete procedure for creating the classifier is shown in Appendix II.

2.3 Verification of database

After we finish creating our database, how do we go about analyzing it? How can we know if the functionality of our database is feasible? Will it be able to classify unknown rbcL genes or trnL genes and obtain their taxonomy? Here we use the meta-analysis method to perform a meta-analysis of several papers on rbcL or trnL using our database, and finally get the results for validation, so that we can verify whether our database is usable or not.

2.3.1 rbcL – study and amplicon sequences workflow

For a study on validating rbcL data I have chosen a study on large-scale monitoring of plants through soil environmental DNA metabarcoding (Nicole A. Fahner et al 2016). In this study, researchers found that environmental DNA (eDNA) that can be extracted from soil samples through DNA metabarcoding analysis can include taxa represented by active and dormant tissues, seeds, pollen, and detritus can provide a more complete picture of a site's plant diversity from a single assessment. The researchers used four DNA markers (matK, rbcL, ITS2 and trnL P6 loop) to build the database.

This paper refers to the Qiime2 (Bolyen et al 2019) and DADA2 (Callahan et al 2016) workflows from https://github.com/umerijaz/tutorials/blob/master/qiime2_tutorial.md. The DADA2 method is an algorithm for inferring ASVs (amplicon sequence variants, i.e. true error-free sequences) in a sample from a database of noisy reads generated by amplicon sequencing. DADA2 was originally developed for short-read long amplicon sequencing (Callahan et al 2016). This workflow aims to create amplicon sequences, the abundance table of variants and ASV representative sequences, for the full workflow of amplicon sequences, please see Appendix III. In short, in the study data we meta-analyze, researchers add barcodes to each sequence, which is used to distinguish different samples, and generally cut the library based on the sequence barcode, that is, to distinguish which sample each sequence belongs to. We first need to extract the barcode, and then split the forward and reverse sequences, that is, put the forward and reverse sequences into different files, import them into Qiime2 and demultiplex them. The readings are then exported to the Qiime2 viewer (<https://view.qiime2.org/>) for visual quality assessment. The sequences are then denoised and clustered based on the quality of the forward

reads and the quality of the reverse reads using the DADA2 algorithm. here the produce table.qza will be produced as an abundance table and produce rep-seqs.qza will contain the ASV sequences. Then in qiime2 creates the phylogenetic tree using align-to-tree-mafft-fasttree and then exports it in NEWICK format.

2.3.2 trnL – study and amplicon sequences workflow

Regarding the study of verifying trnL data, I chose a pollen DNA macrobarcode to study the interaction between plants and pollinators (Bell KL et al 2017). In this study, the researchers wanted to study the pollination network in a constantly changing environment. By studying pollen genes collected from bees and using the DNA meta-barcoding method of these genes, the researchers sampled pollen from 38 species of bees collected in different forests in Florida. They isolated DNA from the pollen mixture. Sequenced the gene region, created a small database, and finally successfully verified that the DNA meta-barcode is superior to the microscopic identification of pollen in terms of efficiency and resolution.

The workflow for trnL to obtain abundance tables and ASV representative sequences for amplicon sequences is similar to that for rbcL, which is what is covered in section 2.3.1 of this paper, and will not be mentioned here. The full workflow of amplicon sequences, please see Appendix III.

2.3.3 Database classification workflow for ASVs

In general, the first step in amplicon analysis species composition analysis is to annotate the sequences of FeatureData[Sequence] with species. This was done using a pre-trained Naive Bayes classifier with the q2-feature-classifier plugin. This classifier was trained on Greengenes 13_8 99% OTU, where the sequence was trimmed to include only 250 bases from a 16S region that was amplified and sequenced in this analysis using primers 515F/806R from the V4 region. We will apply this classifier to the sequences and can generate visualizations of the association from sequence to species annotation results (Bolyen et al 2019). Whereas in this paper it is necessary to train these study genes with the new classifier created in this paper, in the classification process we only need to use the classifier produced by our database to classify the representative sequence files output in 2.3.1 and 2.3.2. The classify-sklearn plugin for qiime2 (Bolyen et al 2019) is used here.

The full workflow of the classification workflow, please see Appendix III.

3 Results

Regarding the present study, the results were generated in two main ways, on the one hand for the database results, regarding the taxonomic results within the database and the results of the gene reads included in the database. On the other hand is the result for the validation of the database, whether the database we have created is correct in its ability to classify unknown genes and with how much confidence.

3.1 Summarise of database and taxonomy

3.1.1 rbcL database

For the rbcL database, we ended up with 254,982 rbcL gene reads, where we can probably say that this is the number of all plant rbcL genes that we humans have studied so far (by 6 July 2021). Of these gene data, only 219676 gene taxonomies were read in their entirety, with the remainder being blank due to NCBI IP restrictions

(https://www.ncbi.nlm.nih.gov/books/NBK25497/#chapter2.Usage_Guidelines_and_Requiremen) or were not yet classified by the NCBI database resulting in blank taxonomic data.

3.1.2 Taxonomy of rbcL databas

Taxonomy	Number
Kingdom	1
Phylum	2
Class	24
Order	137
Family	646
Genus	11212

3.1.3 trnL database

For the trnL database, we ended up with 315,932 trnL gene reads, the same as the rbcL database, and this is probably the number of all plant trnL genes that we humans have studied to date (by 6 July 2021). Of these gene data, only 287,489 gene taxonomies were read in their entirety, with the remainder being blank due to NCBI IP restrictions

(https://www.ncbi.nlm.nih.gov/books/NBK25497/#chapter2.Usage_Guidelines_and_Requiremen) or were not yet classified by the NCBI database resulting in blank taxonomic data.

3.1.4 Taxonomy of trnL database

Taxonomy	Number
Kingdom	1
Phylum	3
Class	38
Order	172
Family	790
Genus	10100

3.2 Verification of database results

In section 2.3 of this article, we will go through the splitting of positive and negative sequences, removing barcodes, cutting libraries, splicing double-ended sequences, removing primers, quality control, sequence clustering into ASV, selecting ASV representative sequences, and obtaining ASV abundance (frequency) Table (called feature table in qiime2), remove chimera, and represent the steps of sequence comparison reference database to get the classification information of each ASV microbial species.

The DADA2 algorithm of amplicon sequence denoising and clustering of sequences. The workflow first needs to evaluate the visual quality of the sequence, which will be shown in this chapter. After that, this chapter will show how to run DADA2 algorithm to generate variation abundance table and ASV. Represents the result of the sequence, and the result of creating a phylogenetic tree. Finally, the classification of ASV generated by the database classifier we created will produce its classification results.

3.2.1 rbcL amplicon sequence variants study result

After putting the sequences together and demultiplexing the sequences, the qiime2 software produces demux.qzv, After opening the resulting file with the name demux.qzv on the Qiime2 viewer <https://view.qiime2.org>, we see the following results as shown in Figure 2, where we can see that the total number of gene sequences included in this study is 41025444 and also contains 140 sample sequences which is 140 representative sequences.

Demultiplexed sequence counts summary

Minimum:	67743
Median:	295215.5
Mean:	293038.8857142857
Maximum:	525755
Total:	41025444

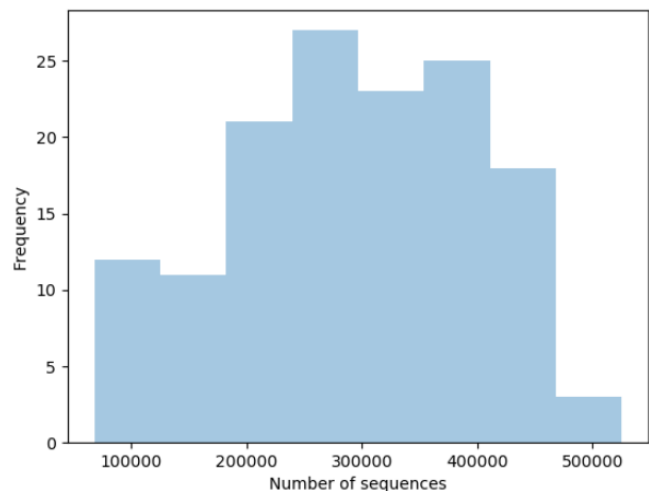


Figure 2: Demultiplexed sequence counts summary. Provided by qiime2

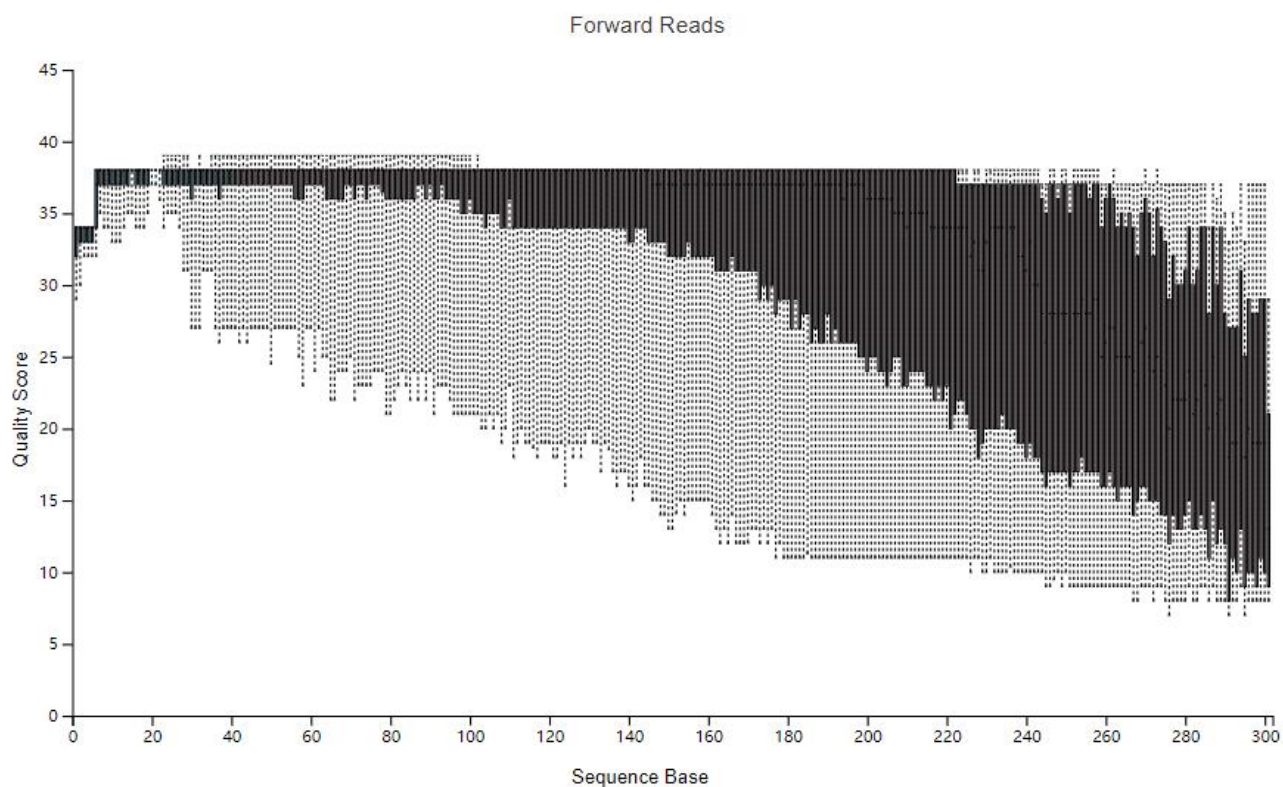


Figure 3: Bar plot of quality score and sequence base in rbCL study. Provided by qiime2

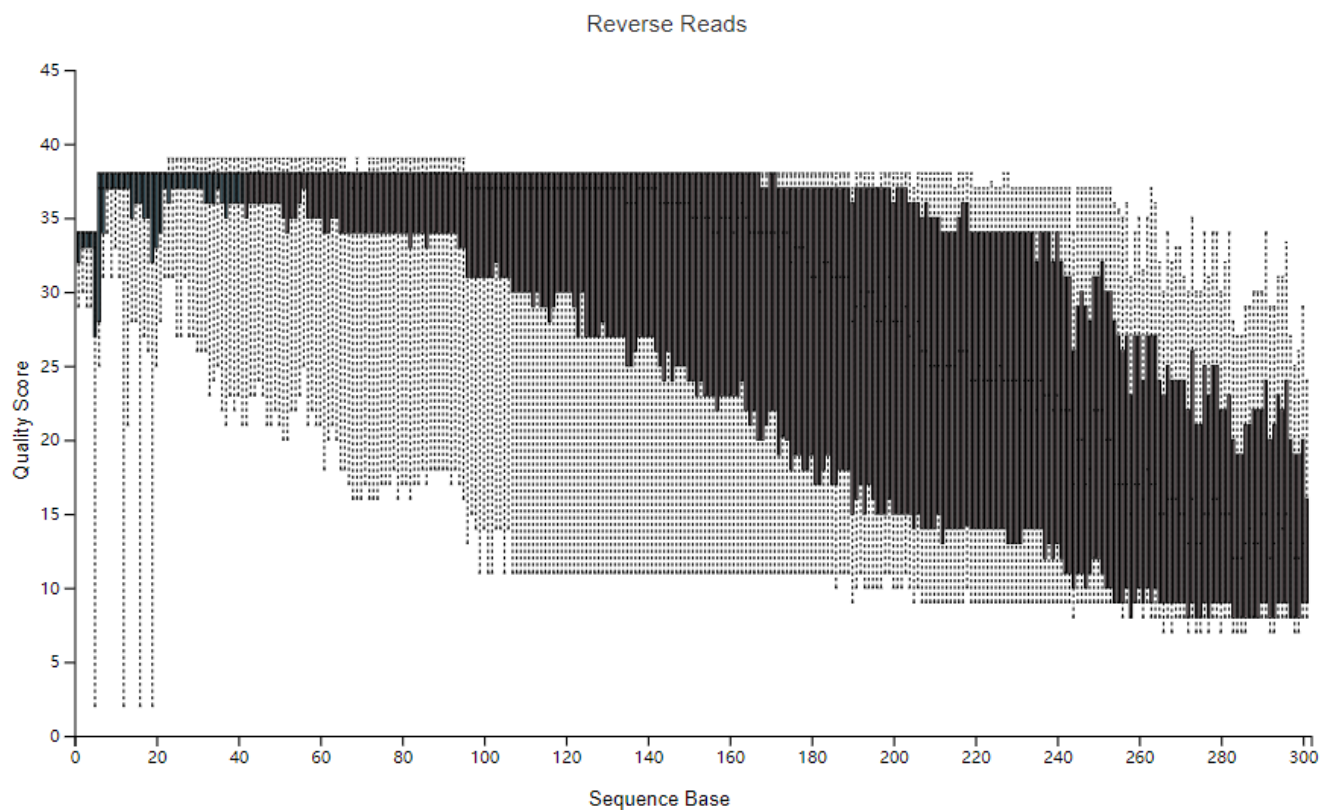


Figure 4: Bar plot of quality score and sequence base in rbCL study. Provided by qiime2

Forward Reads

Total Sequences Sampled	10000
2%	45 nts
9%	49 nts
25%	93 nts
50% (Median)	162 nts
75%	301 nts
91%	301 nts
98%	301 nts

Reverse Reads

Total Sequences Sampled	10000
2%	45 nts
9%	49 nts
25%	93 nts
50% (Median)	162 nts
75%	301 nts
91%	301 nts
98%	301 nts

Figure 5: Demultiplexed sequence length summary in rbCL study. Provided by qiime2

Based on the results in Figures 3, 4 and 5, we manually find the threshold for noise reduction, i.e. where the quality drops significantly. We can clearly see that in Figure 3 Forward Reads, somewhere near 160, the quality starts to drop significantly, so we choose 160 as the parameter, and in Figure 4 Reverse Reads, somewhere near 120, the quality starts to drop, so we choose 120 as the parameter.

After the denoising is completed, we are left with the representative sequence file rep-seqs.qza. Figures 6 and 7 show the visual analysis of the rep-seqs. Figure 8 captures some of the statistical results of the denoising process. Figure 9 is the resulting visual newick tree.

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
4867	160	268	207.98	108	33.33

Figure 6: Sequence Length Statistics of rep-seqs in rbCL study. Provided by qiime2

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	160	162	179	205	238	258	265

*Values rounded down to nearest whole number.

Figure 7: Seven-Number Summary of Sequence Lengths of rep-seqs in rbcL study. Provided by qiime2

sample-id fq2-types	input numeric	filtered numeric	denoised numeric	merged numeric	non-chimeric numeric
SRR3378038	282933	237904	237469	6562	6562
SRR3378039	352325	1416	1328	267	267
SRR3378040	434895	4308	4234	3485	3309
SRR3378041	350977	90102	89945	5	5
SRR3378042	391099	1295	1218	352	352
SRR3378043	445260	1330	1259	22	22
SRR3378044	192225	8474	8124	591	422
SRR3378045	249040	61077	60955	144	144
SRR3378046	462337	25449	25320	23766	23760
SRR3378047	382010	23371	23164	4286	3373
SRR3378048	209345	2002	889	556	527
SRR3378049	219936	51084	50980	15	15
SRR3378050	410308	23124	22980	16109	16109
SRR3378051	185252	4049	3418	2505	2283
SRR3378052	361940	28223	28019	5087	3977
SRR3378053	360714	2228	1631	1210	763

Figure 8: some statistical results of the denoising process of rep-seqs in rbcL study. Provided by qiime2

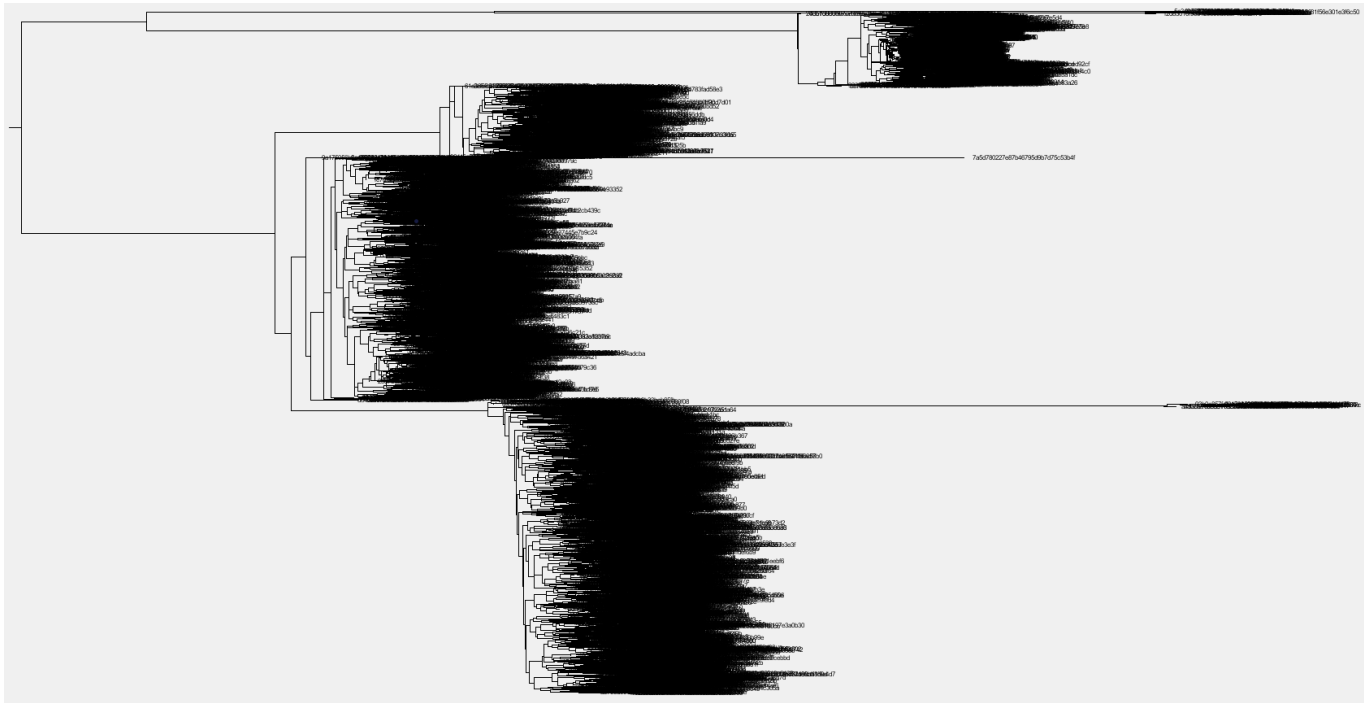


Figure 9: newick tree in rbcL study. Provided by figtree

3.2.2 rbcL taxonomy result

The classification of the representative sequences in section 3.2.1, i.e. ASV, using the classifier trained from our database gives a taxonomy file which is shown as Figure 10, and a bar plot consisting of Figure 11 and Figure 12.

Feature ID #q2types	Taxon categorical	Confidence categorical
0004bfc152565d6cf29d2a3257b8f73	k__Eukaryota;p__Chlorophyta;c__Chlorophyceae;o__Chlamydomonadales;f__Chlamydomonadaceae;g__Chloromonas;s__;	0.9098338594481487
00118ef49c6c47ab67ef93ce39c9b8c0	k__Eukaryota	0.9710019688925491
00154beb71c8afb15dc3ba1ed292448	k__Eukaryota;p__Chlorophyta;c__Chlorophyceae;o__Chlamydomonadales;f__Chlamydomonadaceae;g__Chloromonas;s__;	0.8634163872977013
0017485f08fa538a55fd41866bb4571b	k__Eukaryota;p__Chlorophyta;c__Chlorophyceae;o__Chlamydomonadales;f__Chlamydomonadaceae;g__Chloromonas;s__;	0.9193230900962691
001f233e8bf1f739aee746c88ef76e21	k__Eukaryota	0.9454161526481427
00305ed74b039ed9c98ef558ef2c364	k__Eukaryota;p__Chlorophyta;c__Chlorophyceae;o__Chlamydomonadales;f__Chlamydomonadaceae;g__Chloromonas;s__;	0.9120704727957512
0033cb37c94ed8c626bbcbf1e98b340d	Unassigned	0.4792108593555615
0037c79f23f7e2e84895164e72a75f01	k__Eukaryota;p__Chlorophyta	0.9548105081130939
0057c7a29538bb028fd98dcf1b041deb	k__Eukaryota	0.985491211886472
005b3351c537020411ec71efeaf6528a	k__Eukaryota;p__Chlorophyta;c__Chlorophyceae;o__Chlamydomonadales;f__Chlamydomonadaceae;g__Chloromonas;s__;	0.8848877182115683
0068a7cfeef92abd2845bca89c838f28	k__Eukaryota	0.9733572477854959
007ddca79ba7b35389e4895c8402aa3a	k__Eukaryota	0.9609547522997061
0093716bf637baa5240d74f035d95770	k__Eukaryota	0.9244119041156118
009db79bfd3dc326d84a096e90de2273	k__Eukaryota;p__Chlorophyta;c__Chlorophyceae;o__Chlamydomonadales;f__Chlamydomonadaceae;g__Chloromonas;s__;	0.8294001414761125
00a11fdac6d0dc889349fc3523dc676	k__Eukaryota;p__Chlorophyta;c__Chlorophyceae;o__Chlamydomonadales;f__Chlamydomonadaceae;g__Chloromonas;s__;	0.9160258838431868

Figure 10: taxonomy of rep-seqs in rbcL study. Provided by qiime2

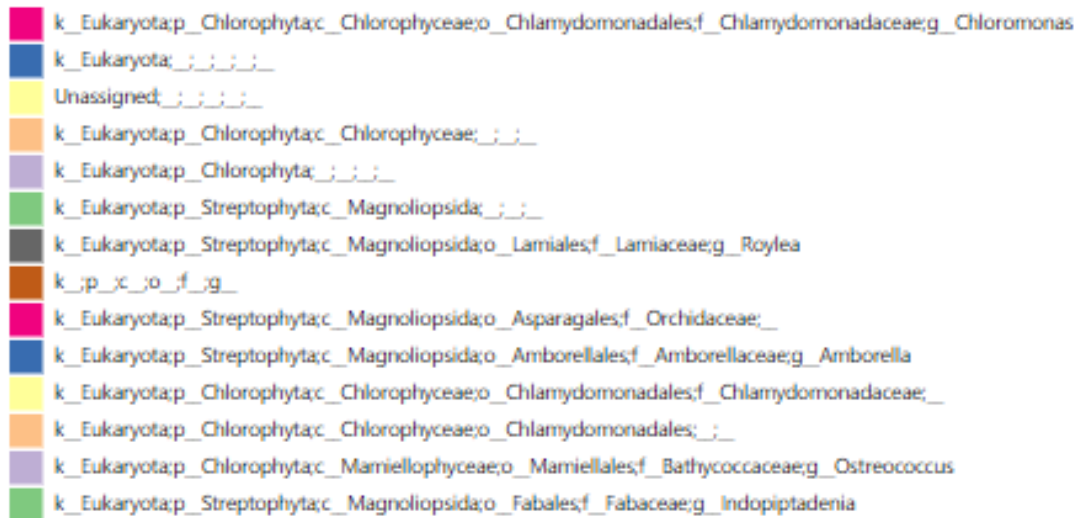


Figure 12: bar plot of taxonomy in rbcl study. Provided by qiime2

3.2.3 trnL amplicon sequence variants study result

After putting the sequences together and demultiplexing the sequences, the qiime2 software produces demux.qzv, After opening the resulting file with the name demux.qzv on the Qiime2 viewer <https://view.qiime2.org>, we see the following results as shown in Figure 13, where we can see that the total number of gene sequences included in this study is 24277123 and also contains 226 sample sequences which is 226 representative sequences.

Demultiplexed sequence counts summary

Minimum:	149
Median:	72819.0
Mean:	107420.8982300885
Maximum:	638071
Total:	24277123

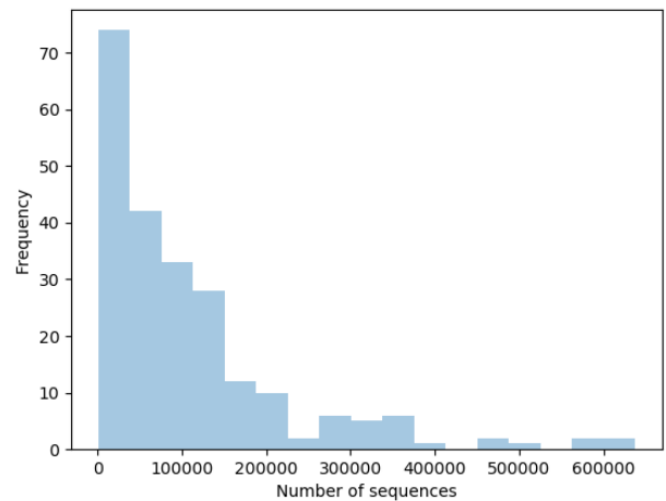


Figure 13: Demultiplexed sequence counts summary. Provided by qiime2

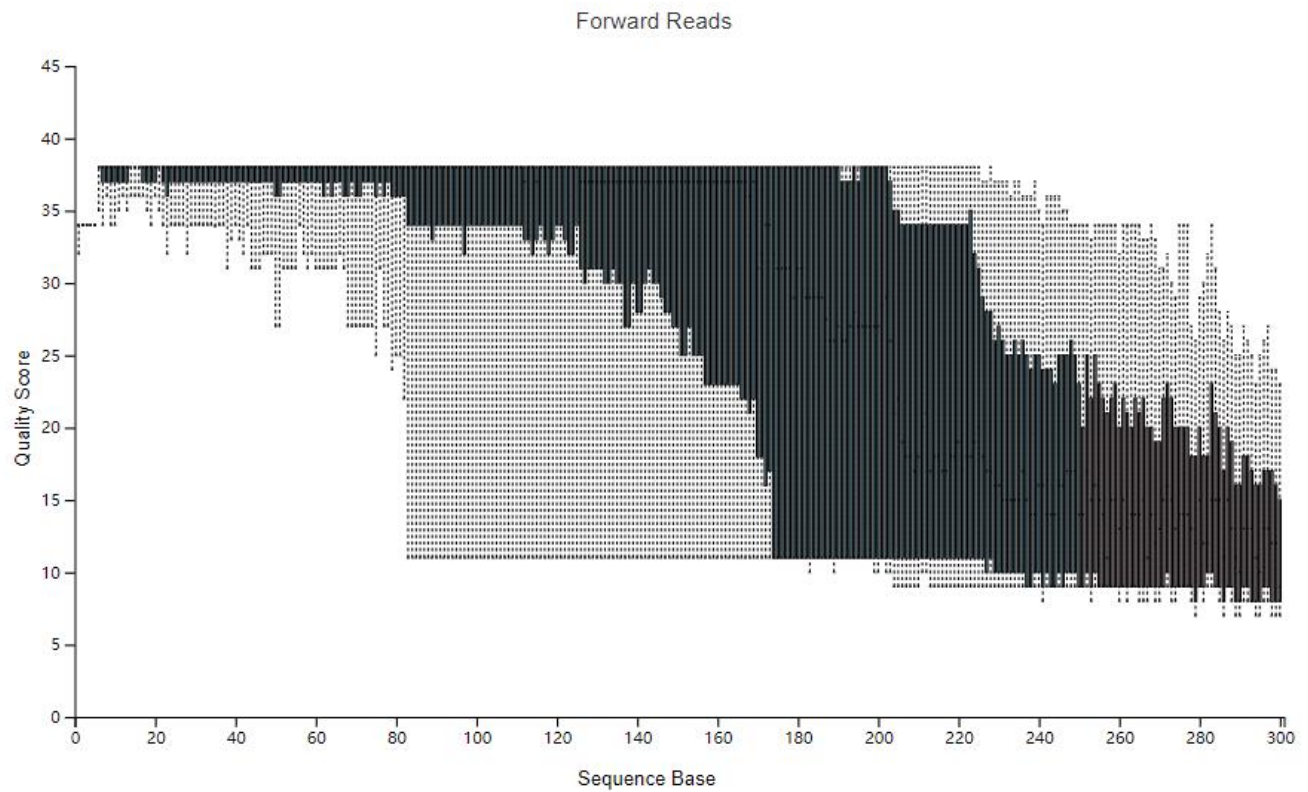


Figure 14: Bar plot of quality score and sequence base in trnL study. Provided by qiime2

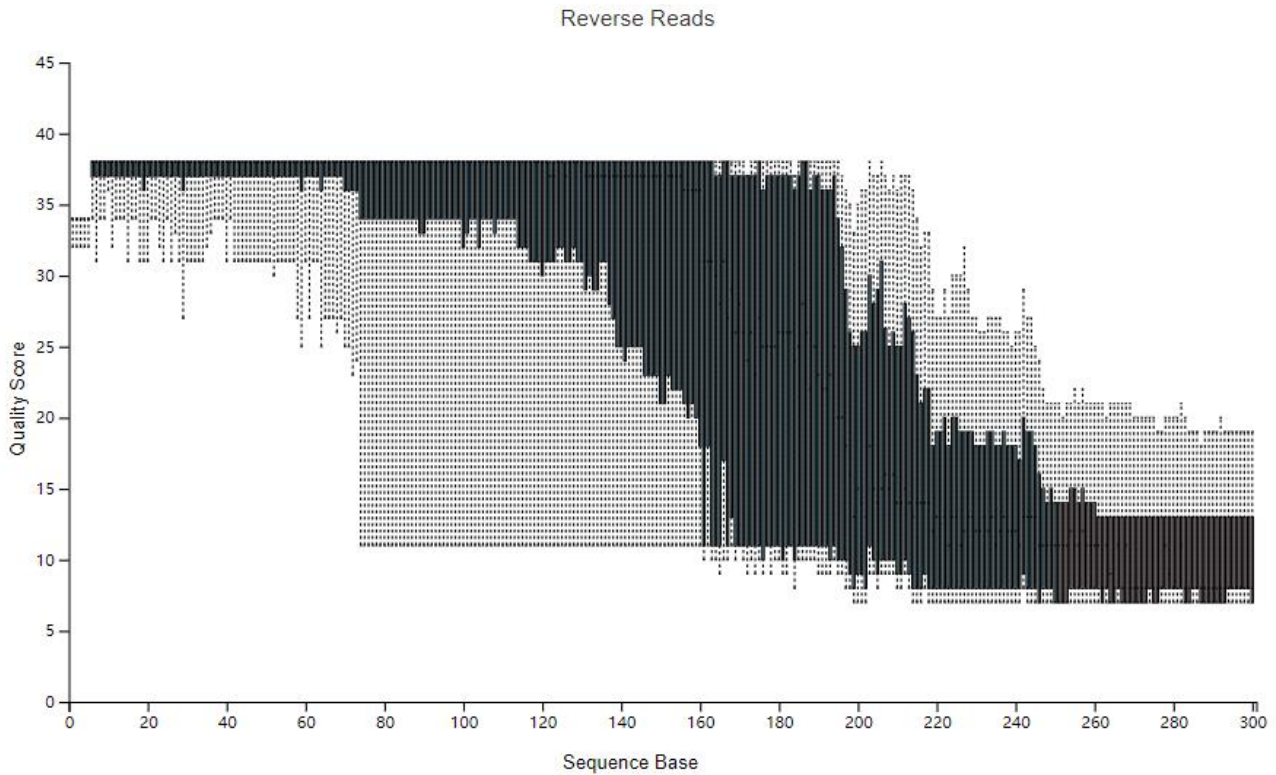


Figure 15: Bar plot of quality score and sequence base in trnL study. Provided by qiime2

Demultiplexed sequence length summary

Forward Reads

Total Sequences Sampled	10000
2%	250 nts
9%	300 nts
25%	300 nts
50% (Median)	300 nts
75%	300 nts
91%	300 nts
98%	300 nts

Reverse Reads

Total Sequences Sampled	10000
2%	250 nts
9%	300 nts
25%	300 nts
50% (Median)	300 nts
75%	300 nts
91%	300 nts
98%	300 nts

Figure 16: Demultiplexed sequence length summary in trnL study. Provided by qiime2

Based on the results in Figures 14, 15 and 16, we manually find the threshold for noise reduction, i.e. where the quality drops significantly. We can clearly see that in Figure 3 Forward Reads, somewhere near 170, the quality starts to drop significantly, so we choose 170 as the parameter, and in Figure 4 Reverse Reads, somewhere near 150, the quality starts to drop, so

we choose 150 as the parameter.

After the denoising is completed, we are left with the representative sequence file rep-seqs.qza. Figures 17 and 18 show the visual analysis of the rep-seqs. Figure 19 captures some of the statistical results of the denoising process. Figure 20 is the resulting visual newick tree.

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
3092	170	306	224.95	136	47.65

Figure 17: Sequence Length Statistics of rep-seqs in trnL study. Provided by qiime2

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	170	170	184	220	241	305	305

*Values rounded down to nearest whole number.

Figure 18: Seven-Number Summary of Sequence Lengths of rep-seqs in trnL study. Provided by qiime2

sample-id #q2-types	input numeric	filtered numeric	denoised numeric	merged numeric	non-chimeric numeric
SRR2748706	208484	123079	122633	95642	87483
SRR2748707	150328	82319	81944	18	18
SRR2748708	196806	123505	123362	53172	49646
SRR2748720	197499	133231	132959	108309	104429
SRR2748721	576937	466975	466116	398762	358035
SRR2749159	301478	235256	234497	159819	154291
SRR2749321	472793	377825	376894	315905	301295
SRR2749469	638071	549885	548912	481606	372732
SRR2749666	586289	483589	483103	401406	313022
SRR2749712	192000	130085	129574	58281	57172
SRR2749718	286289	212131	211793	139078	136499
SRR2749724	266674	206717	206262	133067	123180
SRR2749752	353887	276267	275804	202424	178147
SRR2749753	194971	134912	134685	65080	59664
SRR2749761	475308	362015	361484	286166	201872
SRR2749762	221188	97383	95200	20707	20526

Figure 19: some statistical results of the denoising process of rep-seqs in trnL study. Provided by qiime2

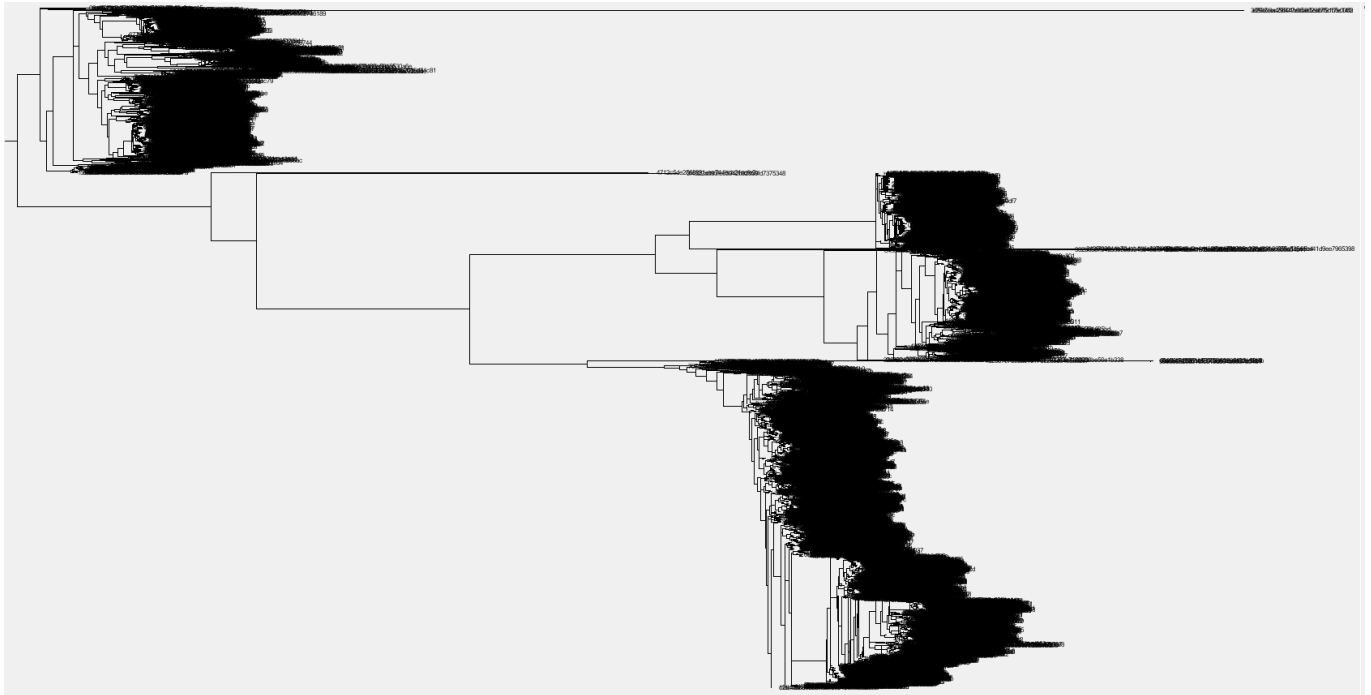


Figure 20: newick tree in trnL study. Provided by figtree

3.2.4 trnL taxonomy result

The classification of the representative sequences in section 3.2.1, i.e. ASV, using the classifier trained from our database gives a taxonomy file which is shown as Figure 21, and a bar plot consisting of Figure 22 and Figure 23.

Feature ID #q2-types	Taxon categorical	Confidence categorical
0011632d6896d5c629d65d90b06c354d	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.9908722877922959
002b5f73148e362bcdafc17d6d9d9589	Unassigned	0.6178313015797532
0034a24f61b237e9dc4194f51775b27	k__Eukaryota;p__Streptophyta;c__Magnoliopsida;o__Malpighiales;f__Euphorbiaceae	0.9123087578194237
00597513d2c72eb079fc4722eed7eed	Unassigned	0.5689499818366075
0061730e180c9e2262402085a15b0ccc	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.9461522849934476
006692bfe97b60ada1448d88488288c8	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.9942538688439667
0073650110a0b474325a6e5d2c924798	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.937440295957738
009a7b9d76486338fd4cb2539c68244	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.9011403513021622
00aa75e37c9f5fd2cf96a594f9e8e0af	Unassigned	0.5555459847636459
00bef100f13eac97e61a26d1c6258da7	Unassigned	0.498866574378221
00d2bb28e4701e1da5518e47af217ccf	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.8836800761617482
00e9a8735046509c2850e37ca27009fe	k__Eukaryota;p__Streptophyta;c__Magnoliopsida;o__Malpighiales;f__Euphorbiaceae	0.7124752442492769
00f50b10e431a38b435577f1d31cd5c7	Unassigned	0.5681446943551086
0147bca59ac31f8c178f24ed12725b21	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.896295676056023
0148c017bca81a7af8a0b5fc2ddc9d	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.910787001648996
0151d751f9e9651b0f8cae364b01eb5	k__Eukaryota;p__Streptophyta;c__Magnoliopsida	0.9297175166670356
01695adc4fc27e9ed9a445e4da8de1f2	Unassigned	0.5359710854915771

Figure 21: taxonomy of rep-seqs in trnL study. Provided by qiime2

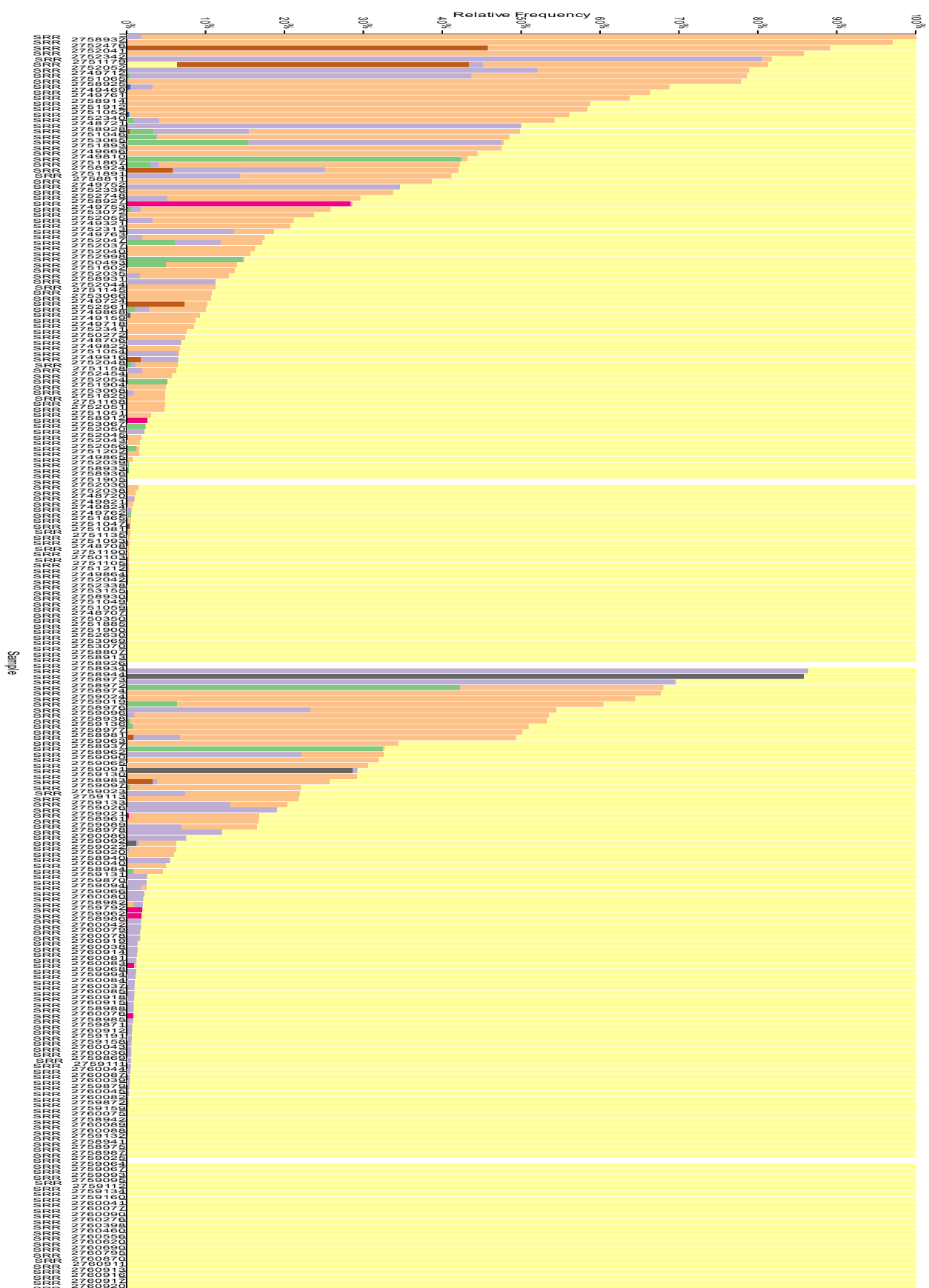


Figure 22: bar plot of taxonomy in trnL study. Provided by qiime2

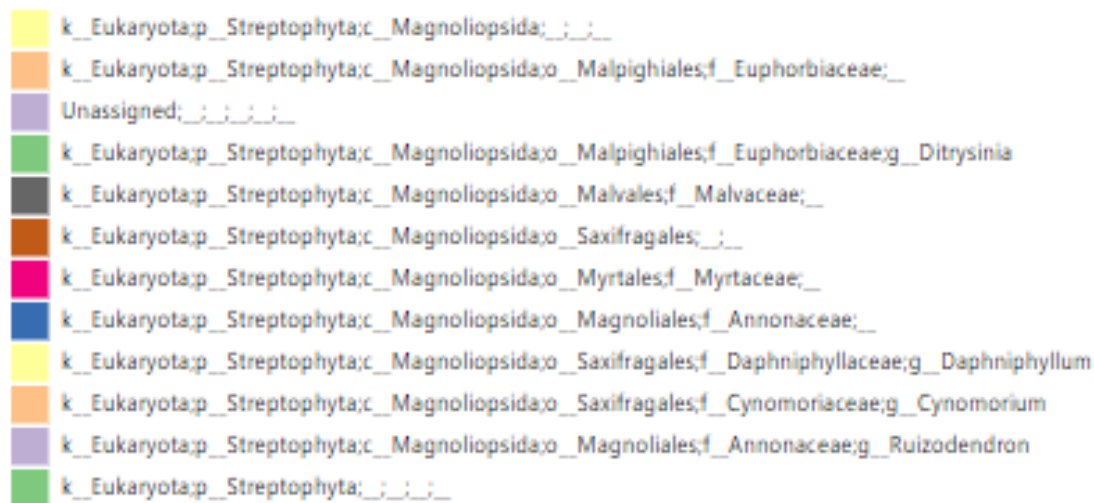


Figure 23: bar plot of taxonomy in trnL study. Provided by qiime2

4 Discussion

4.1 Performance of rbcL as a barcode

The total number of ASVs in my study was 4867 and 4565 ASVs were assigned, with all classifications having a confidence level of over 70 percent confidence. Here is a table of the results of 3420 ASVs with a confidence level of over 80 per cent.

Taxonomy	Number
k_Eukaryota	1158
k_Eukaryota;p_Chlorophyta	35
k_Eukaryota;p_Chlorophyta;c_Chlorophyceae;o_Chlamydomonadales;f_Chlamydomonadaceae;g_Chloromonas;s_;	2212
k_Eukaryota;p_Chlorophyta;c_Mamiellophyceae;o_Mamiellales;f_Bathycoccaceae;g_Ostreococcus;s_;	1
k_Eukaryota;p_Streptophyta;c_Magnoliopsida	12
k_Eukaryota;p_Streptophyta;c_Magnoliopsida;o_Fabales;f_Fabaceae;g_Indopiptadenia;s_;	1

4.2 Performance of trnL as a barcode

The total number of ASVs in my study was 3092 and 2310 ASVs were assigned, with all classifications having a confidence level of over 70 percent. Here is a table of the results of 2171 ASVs with a confidence level of over 80 per cent.

Taxonomy	Number
k_Eukaryota;p_Streptophyta;c_Magnoliopsida	1986

k__Eukaryota;p__Streptophyta;c__Magnoliopsida;o__Magnoliales;f__Annonaceae;g__Ruizodendron;s__;	1
k__Eukaryota;p__Streptophyta;c__Magnoliopsida;o__Malpighiales;f__Euphorbiaceae	172
k__Eukaryota;p__Streptophyta;c__Magnoliopsida;o__Malpighiales;f__Euphorbiaceae;g__Ditrysinia;s__;	8
k__Eukaryota;p__Streptophyta;c__Magnoliopsida;o__Myrtales;f__Myrtaceae	3
k__Eukaryota;p__Streptophyta;c__Magnoliopsida;o__Saxifragales;f__Cynomoriaceae;g__Cynomorium;s__;	1

4.3 Weaknesses and Improvements

Firstly I would like to clarify that some data loss problems occurred during the creation of the database due to NCBI's IP restrictions

(https://www.ncbi.nlm.nih.gov/books/NBK25497/#chapter2.Usage_Guidelines_and_Requiremen) or were not yet classified by the NCBI database resulting in blank taxonomic data. I think here we can still go ahead and re-add the missing data, in this thesis we did not go ahead and fix these missing data due to time constraints, as a In the process, I have figured out a way to split the data and crawl it with different IPs, but even so, there is still the problem of missing data, and if time permits, I think I will analyse and crawl the data again, removing the data that is not classified by NCBI, and This will help us to train the classifier later.

Secondly, it is clear that our database only stops at the Genus level for gene identification, because much of the plant data obtained in the NCBI database is only classified up to this point, and if possible, I will try to obtain a new database that will increase the classification level to the species level

Thirdly, in my literature research, I have found that more and more researchers are trying to combine two different genes as barcodes to analyse genes, and I think that our database classifier could be improved to a greater extent in this way.

5 Conclusions

We successfully created the database with rbcL and trnL respectively, and successfully completed the verification of them, proving that our database is capable of species identification. I think it is a very good genetic barcode database test. Even though it has many shortcomings, it finally achieves its own functions. At the same time, compared with traditional gene classification methods, our database can better identify some unknown genes. It shows that when it is difficult to directly obtain some research materials, we can identify species by studying the genes of the materials. For example, for the diet of animals, it is difficult for us to directly observe and count what they have been eating. We only need to obtain the plant genes in their metabolites, and classify and identify them through a database like this article. Knowing their diet, in summary, although it is still in its infancy, DNA meta-barcoding technology can be used today to build a new generation of databases, and with continuous improvement, there is great hope in the future.

Reference

- Jianping Xu. Fungal DNA barcoding. *Genome*. 59(11): 913-932.
- Adamowicz, S.J., Scoles, G.J. International Barcode of Life: Evolution of a global research. *Community (2015) Genome*, 58 (5), pp. 151-162. DOI: 10.1139/gen-2015-0094.
- Fazekas A, Kuzmina ML, Newmaster SG et al (2012) DNA barcoding methods for land plants In: Kress WJ, Erickson DL (eds). *Springer protocols methods in molecular biology 858 DNA barcodes methods and protocols*. Springer, New York, pp 223–252
- Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 12:499–510
- Lynch M, Milligan BG (1994) Analysis of population genetic-structure with RAPD markers. *Mol Ecol* 3:91–99
- de Vere N, Rich TCG, Ford CR et al (2012) DNA barcoding the native flowering plants and conifers of Wales. *PLoS One* 7:e37945
- Fazekas AJ, Burgess KS, Kesanakurti PR et al (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS One* 3:e2802
- Webb CO (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *Am Nat* 156:145–155
- Harvey PH, Leigh Brown AJ, Maynard SJ, Nees (2006) *New uses for new phylogenies*. Oxford University Press, Oxford
- Hebert PDN, Cywinska A, Ball SL et al (2003) Biological identifications through DNA barcodes. *Proc R Soc Lond B Biol Sci* 270:313–321
- Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Syst Biol* 54:852–859
- Fazekas AJ, Burgess KS, Kesanakurti PR et al (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS One* 3:e2802
- Chase MW, Salamin N, Wilkinson M et al (2005) Land plants and DNA barcodes: short-term and long-term goals. *Phil Trans Lond B* 360:1889–1895
- Kress WJ, Wurdack KJ, Zimmer EA et al (2005) Use of DNA barcodes to identify flowering plants. *Proc Natl Acad Sci USA* 102:8369–8374
- Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in the land plants: an evaluation of *rbcL* in a multigene tiered approach. *Can J Bot* 84:335–341
- Cowan RS, Chase MW, Kress WJ, Savolainen V (2006) 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon* 55:611–616
- Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS One* 2:e508
- Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madriñan S, Petersen G, Seberg O,

Jørgensen T, Cameron KM, Carine M, Pedersen N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TE, Kelly L, Wilkinson M (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon* 56:295–299

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2008) Testing candidate plant barcode regions with species of recent origin in the Myristicaceae. *Mol Ecol Notes* 8:480–490

KUSUMADEWI SRI YULITA, Secondary Structures of Chloroplast trnL Intron in Dipterocarpaceae and its Implication for the Phylogenetic Reconstruction

HAYATI Journal of Biosciences, Volume 20, Issue 1, 2013, Pages 31–39, ISSN 1978–3019

CBOL Plant Working Group, A. DNA barcode for land plants. *PNAS* 106, 12794–12797 (2009)

Hollingsworth PM, Graham SW, Little DP. 2011 Choosing and using a plant DNA barcode. *PLoS ONE* 6, e19254. (doi:10.1371/journal.pone.0019254) Crossref, PubMed, ISI, Google Scholar

Les DH, Garvin DK and Wimpee CF 1991 Molecular evolutionary history of ancient aquatic angiosperms. *Proceedings of the National Academy of Sciences* 88 10119–23

Newmaster SG, Fazekas AJ and Ragupathy S 2006 DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach *Canadian Journal of Botany* 84 335–41

CBOL Plant Working Group, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim K-J, Kress WJ, Schneider H, van Alphen Stahl J, Barrett SC, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M, Chacon J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TAJ, Hollingsworth ML, Husband BC, Kelly LJ, Kesanakurti PR, Kim JS, Kim Y-D, Lahaye R, Lee H-L, Long DG, Madrinan S, Maurin O, Meusnier I, Newmaster SG, Park C-W, Percy DM, Petersen G, Richardson JE, Salazar GA, Savolainen V, Seberg O, Wilkinson MJ, Yi D-K and Little DP 2009 A DNA barcode for land plants *Proceedings of the National Academy of Sciences* 106 12794–7

Kellogg EA and Juliano ND 1997 The structure and function of RuBisCO and their Implications for Systematic Studies *American Journal of Botany* 84 413–28

Papuangan, N. (2019, May). Amplification and analysis of RbcL gene (Ribulose-1, 5-Bisphosphate Carboxylase) of clove in Ternate Island. In IOP conference series: earth and environmental science (Vol. 276, No. 1, p. 012061). IOP Publishing.

Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E.

Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, Kim D. Pruitt, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, *Nucleic Acids Research*, Volume 44, Issue D1, 4 January 2016, Pages D733–D745, <https://doi.org/10.1093/nar/gkv1189>

Fahner, N. et al. "Large-Scale Monitoring of Plants through Environmental DNA Metabarcoding of Soil: Recovery, Resolution, and Annotation of Four DNA Markers." *PLoS ONE* 11 (2016): n. pag.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C (2019) Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *nature biotechnology* 37:852-857

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13: 581–583

Bell KL, Fowler J, Burgess KS, et al. Applying pollen DNA metabarcoding to the study of plant-pollinator interactions. *Appl Plant Sci.* 2017;5(6):apps.1600124. Published 2017 Jun 12. doi:10.3732/apps.1600124

Schloss, Patrick D. "Amplicon sequence variants artificially split bacterial genomes into separate clusters." *bioRxiv* (2021)

Callahan BJ, Wong J, Heiner C, et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Research.* 2019;47(18):e103-e103. doi:10.1093/nar/gkz569

Richardson, R. T., Sponsler, D. B., McMinin-Sauder, H., & Johnson, R. M. (2020). MetaCurator: A hidden Markov model-based toolkit for extracting and curating sequences from taxonomically-informative genetic markers. *Methods in Ecology and Evolution*, 11(1), 181-186.

Karavaiko, G. I., Turova, T. P., Tsaplina, I. A., & Bogdanova, T. I. (2000). The phylogenetic position of aerobic, moderately-thermophilic bacteria of the *Sulfobacillus* species, oxidizing Fe²⁺, S (0) and sulfide minerals. *Mikrobiologiya*, 69(6), 857-860.

Appendix I

```
//Download db.fasta (RBCL) from RBCL
```

```
//Then remove everything from the name except accession IDs.
```

```
bioawk -cfastx '{print ">"$1"\n"$2}' db.fasta > db_accession.fasta
```

```
bioawk -cfastx '{print $1}' db_accession.fasta > IDs_accession.txt
```

```
//Enable R-environment
```

```
export PATH=/home/opt/miniconda2/bin:$PATH
```

```
source activate r-environment
```

```
//Then i use R to get a taxonomy file
```

```
export PATH=/home/opt/miniconda2/bin:$PATH
```

```
source activate r-environment
```

```
R
```

```
library(rentrez)
```

```
#Load the mapping table up
```

```
mapping_table<-read.csv("IDs_accession.txt ",header=FALSE)
```

```
#extract gids
```

```
gids<-as.character(mapping_table$V1)
```

```
taxa_levels<-NULL
```

```
for(i in seq(1:length(gids))){
```

```
  print(paste("Processing",i,"/",length(gids)))
```

```
  tmp<-
```

```
tryCatch(paste(XML::xpathSApply(entrez_fetch(db="taxonomy",id=entrez_summary(db="nucleot  
ide", id=gids[i])$taxid,rettype="xml", parsed=TRUE), "//LineageEx/Taxon/ScientificName",  
XML::xmlValue),collapse=";"),error=function(e) "")
```

```
  tmp2<-
```

```
tryCatch(paste(XML::xpathSApply(entrez_fetch(db="taxonomy",id=entrez_summary(db="nucleot  
ide", id=gids[i])$taxid,rettype="xml", parsed=TRUE), "//LineageEx/Taxon/Rank",  
XML::xmlValue),collapse=";"),error=function(e) "")
```

```
#From the XML returned extract the taxonomy
```

```
tmp1_df<-strsplit(tmp,";")[[1]]
```

```
#From the XML returned extract the levels
```

```
tmp2_df<-strsplit(tmp2,";")[[1]]
```

```

#Now assemble the whole taxonomy
tmp<-paste(paste("k_",tmp1_df[tmp2_df=="superkingdom"],sep=""),";",
paste("p_",tmp1_df[tmp2_df=="phylum"],sep=""),";",
paste("c_",tmp1_df[tmp2_df=="class"],sep=""),";",
paste("o_",tmp1_df[tmp2_df=="order"],sep=""),";",
paste("f_",tmp1_df[tmp2_df=="family"],sep=""),";",
paste("g_",tmp1_df[tmp2_df=="genus"],sep=""),";",
paste("s_",tmp1_df[tmp2_df=="species"],sep=""),";",sep="")

if(is.null(taxa_levels)){taxa_levels<-tmp}else{taxa_levels<-c(taxa_levels,tmp)}
}
data_to_write<-data.frame(ID=mapping_table[,1],Taxa=taxa_levels)
write.table(data_to_write,"19.tax",sep="\t",row.names=F,col.names=F,quote=F)
quit()

```

Appendix II

```
// enable qiime2
```

```
//Then import the sequences in qiime2 format
```

```
export PATH=/home/opt/miniconda2/bin:$PATH
```

```
source activate qiime2-2019.7
```

```
qiime tools import --type 'FeatureData[Sequence]' --input-path db_accession.fasta --output-path db_accession.qza
```

```
//Imported db_accession.fasta as DNASequencesDirectoryFormat to db_accession.qza
```

```
qiime tools import --type 'FeatureData[Taxonomy]' --input-format
```

```
HeaderlessTSVTaxonomyFormat --input-path db_accession.tax --output-path db_accession-taxonomy.qza
```

Appendix III

```
[studentprojects@howe /shared5/studentprojects/Fan/test]$ export
PATH=/home/opt/sratoolkit.2.9.0-centos_linux64/bin:$PATH
[studentprojects@howe /shared5/studentprojects/Fan/test]$ export
PATH=/home/opt/edirect:$PATH
[studentprojects@howe /shared5/studentprojects/Fan/test]$ esearch -db sra -query
PRJNA344894 | efetch --format runinfo |cut -d "," -f 1 > SRR.numbers
esearch -db sra -query PRJNA318025 | efetch --format runinfo |cut -d "," -f 1 > SRR.numbers
[studentprojects@howe /shared5/studentprojects/Fan/test]$ awk '/SRR|ERR/' SRR.numbers >
SRR.numbers.filtered
[studentprojects@howe /shared5/studentprojects/Fan/test]$ for i in $(cat
SRR.numbers.filtered); do echo Processing $i; fastq-dump --split-files --origfmt --gzip $i ;
done
```

```
[studentprojects@howe /shared5/studentprojects/Fan/test]$ ls
```

```
[studentprojects@howe /shared5/studentprojects/Fan/test]$ mkdir sequences
[studentprojects@howe /shared5/studentprojects/Fan/test]$ mv *.fastq.gz sequences/.
```

```
[studentprojects@howe /shared5/studentprojects/Fan/test/sequences]$ ls
[studentprojects@howe /shared5/studentprojects/Fan/test/sequences]$ gunzip *
[studentprojects@howe /shared5/studentprojects/Fan/test/sequences]$ for i in $(awk -F"_"
'{print $1}' <(ls *.fastq) | sort | uniq); do mkdir $i; mkdir $i/Raw; mv $i*.fastq $i/Raw/.; done
```

Step 2: Create a qiime2_tutorial folder

```
[studentprojects@howe /shared5/studentprojects/Fan/test]$ mkdir qiime2_tutorial
[studentprojects@howe /shared5/studentprojects/Fan/test]$ cd qiime2_tutorial
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ ls
```

Step 3: Get the path of sequences folder assigned to a variable d

```
[studentprojects@howe /shared5/studentprojects/Fan/test]$ cd sequences
[studentprojects@howe /shared5/studentprojects/Fan/test/sequences]$ pwd
/shared5/studentprojects/Fan/test/sequences
[studentprojects@howe
```

```

/shared5/studentprojects/Fan/test/sequences]$ d="/shared5/studentprojects/Fan/test/sequences/";
[studentprojects@howe /shared5/studentprojects/Fan/test/sequences]$
[studentprojects@howe /shared5/studentprojects/Fan/test/sequences]$ t=$(ls $d | wc -l);
[studentprojects@howe /shared5/studentprojects/Fan/test/sequences]$ cd ../qiime2_tutorial/
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ paste <(ls $d)
<(perl -le 'sub p{my $l=pop @_;unless(@_){return map [$_,@$l];}return map { my $l=$_; map
[@$l,$_],@$l} p(@_);} @a=[A,C,G,T]; print join("", @$_) for p(@a,@a,@a,@a,@a,@a,@a,@a);' |
awk -v k=$t 'NR<=k{print}') | awk 'BEGIN{print "sample-id\tbarcode-
sequence\n#q2:types\tcategorical"}1' > sample_metadata.tsv
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ cat
sample_metadata.tsv

```

Step 4: Generate barcodes for each read using the file as above

```

[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ (for i in $(ls $d);
do bc=$(awk -v k=$i '$1==k{print $2}' sample_metadata.tsv); bioawk -cfastx -v k=$bc '{print
"@ "$1" "$4"\n"k"\n+";for(i=0;i< length(k);i++){printf "#";printf "\n"}' $d/$i/Raw/*_1.fastq ;
done) > barcodes.fastq

```

Step 5: Collate all the forward reads from all the folders together in a single forward.fastq file

```

[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ (for i in $(ls $d);
do cat $d/$i/Raw/*_1.fastq ; done) > forward.fastq

```

Step 6: Collate all the reverse reads from all the folders together in a single reverse.fastq file

```

[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ (for i in $(ls $d);
do cat $d/$i/Raw/*_2.fastq ; done) > reverse.fastq

```

Sanity check: see if all the numbers match

```

[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ bioawk -cfastx
'END{print NR}' forward.fastq
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ bioawk -cfastx
'END{print NR}' reverse.fastq
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ bioawk -cfastx
'END{print NR}' barcodes.fastq

```

all match

Step 7: Zip all the FASTQ files and move them to emp-paired-end-sequences folder

```
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ gzip *.fastq
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ ls
barcodes.fastq.gz forward.fastq.gz reverse.fastq.gz sample_metadata.tsv
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ mkdir emp-
paired-end-sequences; mv *.gz emp-paired-end-sequences/
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ ls
emp-paired-end-sequences sample_metadata.tsv
```

Next, we enable Qiime2 on the Orion cluster

```
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ export
PATH=/home/opt/miniconda2/bin:$PATH
[studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$ source activate
qiime2-2019.7
```

Step 8: Import the sequences to qiime2

```
(qiime2-2019.7) [studentprojects@howe
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ qiime tools import --type
EMPPairedEndSequences --input-path emp-paired-end-sequences --output-path emp-
paired-end-sequences.qza
```

```
(qiime2-2019.7) [studentprojects@howe
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ ls
emp-paired-end-sequences emp-paired-end-sequences.qza sample_metadata.tsv
Imported emp-paired-end-sequences as EMPPairedEndDirFmt to emp-paired-end-
sequences.qza
```

Step 9: Demultiplex the sequences in Qiime2


```
(qiime2-2019.7) [studentprojects@howe
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ qiime demux emp-paired --p-no-golay-
error-correction --i-seqs emp-paired-end-sequences.qza --m-barcodes-file
sample_metadata.tsv --m-barcodes-column barcode-sequence --o-per-sample-sequences
demux.qza --o-error-correction-details demux-details.qza
```

Saved SampleData[PairedEndSequencesWithQuality] to: demux.qza

Saved ErrorCorrectionDetails to: demux-details.qza

```
(qiime2-2019.7) [studentprojects@howe /shared5/studentprojects/Fan/test/qiime2_tutorial]$
```

```
(qiime2-2019.7) [studentprojects@howe
```

```
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ ls
```

```
demux-details.qza  emp-paired-end-sequences  sample_metadata.tsv
```

```
demux.qza          emp-paired-end-sequences.qza
```

Step 10: Depends on the quality of your run, we want to fine tune Dada2 algorithm by specifying the thresholds

```
(qiime2-2019.7) [studentprojects@howe
```

```
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ qiime demux summarize --i-
```

```
data ./demux.qza --o-visualization ./demux.qzv
```

Next drag and drop the file on the Qiime2 viewer <https://view.qiime2.org> and manually figure out the thresholds, i.e., where the quality drops down significantly

Step 11: Run DADA2 algorithm which will produce table.qza as an abundance table and rep-seqs.qza will contain the ASV sequences

```
(qiime2-2019.7) [studentprojects@howe
```

```
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ qiime dada2 denoise-paired --i-
```

```
demultiplexed-seqs demux.qza --p-trim-left-f 0 --p-trim-left-r 0 --p-trunc-len-f 240 --p-trunc-
len-r 200 --p-n-threads 0 --o-table table.qza --o-representative-sequences rep-seqs.qza --o-
denoising-stats denoising-stats.qza --verbose
```

Step 12: Create a phylogenetic tree

```
(qiime2-2019.7) [studentprojects@howe
```

```
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ unset MAFFT_BINARIES
(qiime2-2019.7) [studentprojects@howe
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ qiime phylogeny align-to-tree-mafft-
fasttree --i-sequences rep-seqs.qza --o-alignment aligned-rep-seqs.qza --o-masked-
alignment masked-aligned-rep-seqs.qza --p-n-threads 0 --o-tree unrooted-tree.qza --o-
rooted-tree rooted-tree.qza
```

```
(qiime2-2019.7) [studentprojects@howe
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ $ qiime tools export --input-path
table.qza --output-path output
```

The table is exported as BIOM file (<https://biom-format.org/>)

```
(qiime2-2019.7) [studentprojects@howe
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ qiime tools export --input-path rep-
seqs.qza --output-path output
Exported rep-seqs.qza as DNASequencesDirectoryFormat to directory output
```

The above will produce dna-sequences.fasta in the output folder

```
(qiime2-2019.7) [studentprojects@howe
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ qiime tools export --input-path rooted-
tree.qza --output-path output
```

```
=====
=====
```

```
(qiime2-2019.7) [studentprojects@becker
/shared5/studentprojects/Fan/test/qiime2_tutorial]$ qiime feature-classifier classify-sklearn --
i-classifier /shared5/studentprojects/Fan/database_rbcl/db_accession_classifier.qza --i-reads
rep-seqs.qza --o-classification taxonomy.qza
```



University
of Glasgow

Declaration of Originality Form

This form **must** be completed and signed and submitted with all assignments.

Please complete the information below (using BLOCK CAPITALS).

Name: Fan Zou.....

Student Number: 2596373Z

Course Name: MSc Project

Assignment Number/Name: Building a plant sequence reference database

**An extract from the University's Statement on Plagiarism is provided overleaf.
Please read carefully THEN read and sign the declaration below.**

I confirm that this assignment is my own work and that I have:

Read and understood the guidance on plagiarism in the Student Handbook, including the University of Glasgow Statement on Plagiarism ☒

Clearly referenced, in both the text and the bibliography or references, **all sources** used in the work ☒

Fully referenced (including page numbers) and used inverted commas for **all text quoted** from books, journals, web etc. (Please check with the Department which referencing style is to be used) ☒

Provided the sources for all tables, figures, data etc. that are not my own work ☒

Not made use of the work of any other student(s) past or present without acknowledgement. This includes any of my own work, that has been previously, or concurrently, submitted for assessment, either at this or any other educational institution, including school (see overleaf at 31.2) ☒

Not sought or used the services of any professional agencies to produce this work ☒

In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations ☒

DECLARATION:

I am aware of and understand the University's policy on plagiarism and I certify that this assignment is my own work, except where indicated by referencing, and that I have followed the good academic practices noted above

Signed.....

Fan Zou

The University of Glasgow Plagiarism Statement

The following is an extract from the University of Glasgow Plagiarism Statement. The full statement can be found in the *University Regulations* at <https://www.gla.ac.uk/myglasgow/senateoffice/policies/uniregs/regulations2021-22/feesandgeneral/studentsupportandconductmatters/reg32/>

This should be read in conjunction with the discipline specific guidance provided by the School at [Insert link](#).

31.1 The University's degrees and other academic awards are given in recognition of a student's **personal achievement**. All work submitted by students for assessment is accepted on the understanding that it is the student's own effort.

31.2 Plagiarism is defined as the submission or presentation of work, in any form, which is not one's own, without **acknowledgement of the sources**. Plagiarism includes inappropriate collaboration with others. Special cases of plagiarism can arise from a student using his or her own previous work (termed auto-plagiarism or self-plagiarism). Auto-plagiarism includes using work that has already been submitted for assessment at this University or for any other academic award.

31.3 The incorporation of material without formal and proper acknowledgement (even with no deliberate intent to cheat) can constitute plagiarism.

Work may be considered to be plagiarised if it consists of:

- a direct quotation;
- a close paraphrase;
- an unacknowledged summary of a source;
- direct copying or transcription.

With regard to essays, reports and dissertations, the rule is: if information **or ideas** are obtained from any source, that source must be acknowledged according to the appropriate convention in that discipline; and **any direct quotation must be placed in quotation marks** and the source cited immediately. Any failure to acknowledge adequately or to cite properly other sources in submitted work is plagiarism. Under examination conditions, material learnt by rote or close paraphrase will be expected to follow the usual rules of reference citation otherwise it will be considered as plagiarism. Departments should provide guidance on other appropriate use of references in examination conditions.

31.4 Plagiarism is considered to be an act of fraudulence and an offence against University discipline. Alleged plagiarism, at whatever stage of a student's studies, whether before or after graduation, will be investigated and dealt with appropriately by the University.

31.5 The University reserves the right to use plagiarism detection systems, which may be externally based, in the interests of improving academic standards when assessing student work.

If you are still unsure or unclear about what plagiarism is or need advice on how to avoid it,

SEEK HELP NOW!

You can contact any one of the following for assistance:

Lecturer
Course Leader
Dissertation Supervisor
Adviser of Studies
Student Learning Service