Rscript ‘PIISA’

March 28, 2021

Type RScript

Title Pipeline for Interactive and Iterative Sequence Analysis

Version 1.0

Date 2021-04-13

Maintainer/Maintainer Daniel Grant <d.grant@uleth.ca>

License MIT License

Depends R(>= 4.0.3), RStudio(>=1.4)

Imports BiocManager(>=1.30.10), dada2(>=3.12), DECIPHER(>=2.18.1), ggplot2(>=2.1.0), phyloseq(>=1.34.0)

URL https://github.com/DanielSGrant/PIISA

Introduction

The pipeline for interactive and iterative sequence analysis (PIISA) is an interactively run pipeline for the analysis of Illumina sequence data. PIISA utilizes the DADA2 R package for its analysis and the phyloseq R package for creation of abundance and diversity plots. Through use of PIISA, it is possible to check quality scores, filter and trim, view error rates, and assign taxonomy to the input sequence data. Several steps in this analysis process are encapsulated in loops which allow the user to iteratively tune input parameters, allowing for increased accuracy. PIISA takes Illumina sequence data, filters, trims, and analyzes it, assigns taxonomy using a reference database, and generates diversity and abundance plots for the processed data.

Requirements

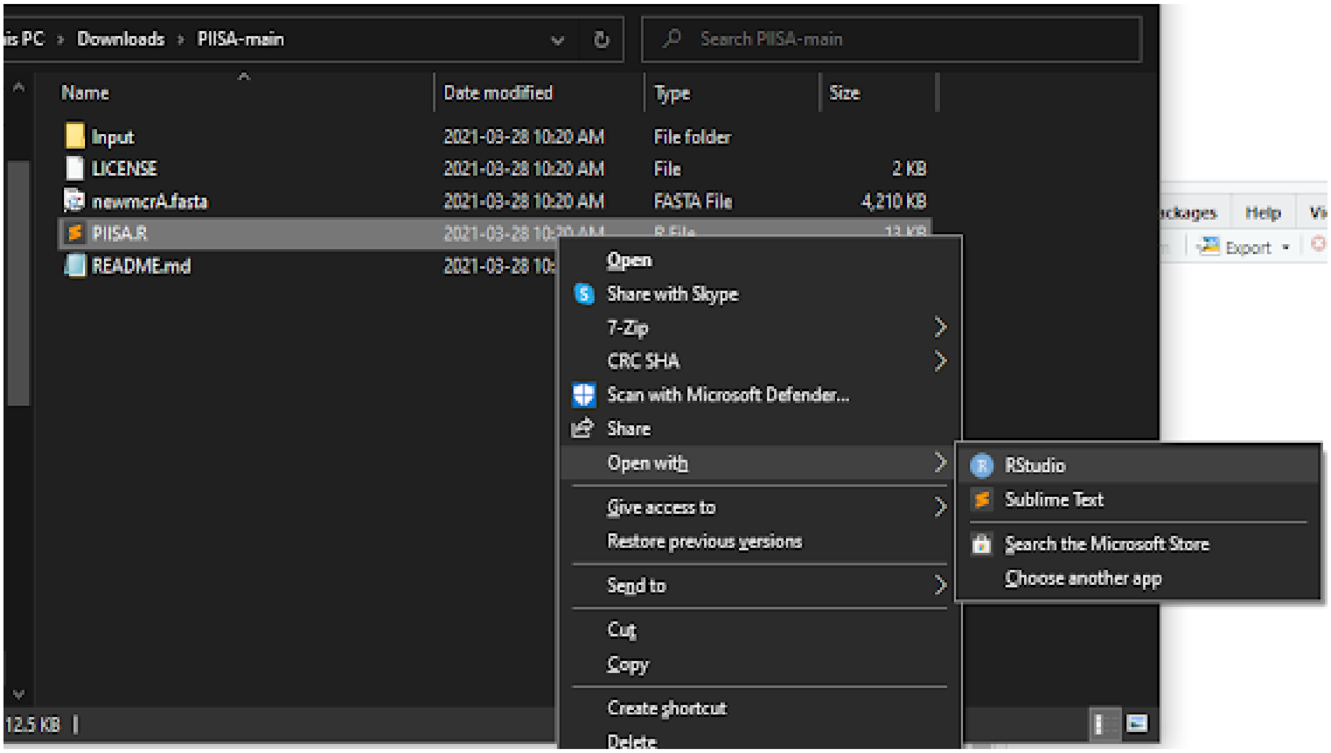
As PIISA is an RScript, running it requires the user to install both the R programming language and the RStudio integrated development environment (IDE) for its use. Furthermore, several R packages are required for its use including BiocManager, dada2, DECIPHER, ggplot2, and phyloseq. The first portion of the PIISA script will attempt to install these R packages for you automatically from within the RStudio IDE while running the pipeline. However, depending on your individual system settings, it may be necessary to install these packages manually. Please refer to the FAQ section of this document for more information if you receive errors regarding the installation of packages. PIISA can be used on Windows, Mac, and Linux operating systems. Due to the computational requirements of the program, it is recommended that the user run it on a modern, 64-bit computer with a minimum of 4 GB of RAM.

Tutorial

The following tutorial contains instructions and specific example photos for use with a Windows 10 operating system due to Windows having the greatest proportion of users. Linux and Mac users should be able to follow the steps in this tutorial closely by completing the equivalent instructions on their respective devices. At many points in the process of running PIISA, prompts where the user must enter information will appear. At these points suggested values may be provided in round parentheses (), and all possible options may be provided in square brackets []. If you would like to use the suggested value, you can simply hit the "Enter" key without typing anything, otherwise type the desired response in and hit the "Enter" key. At any stage of the program that accepts input, type 'q' and hit enter to quit the program. Prompts for user input should appear in a different colour text, which will be purple if you are using the default RStudio settings.

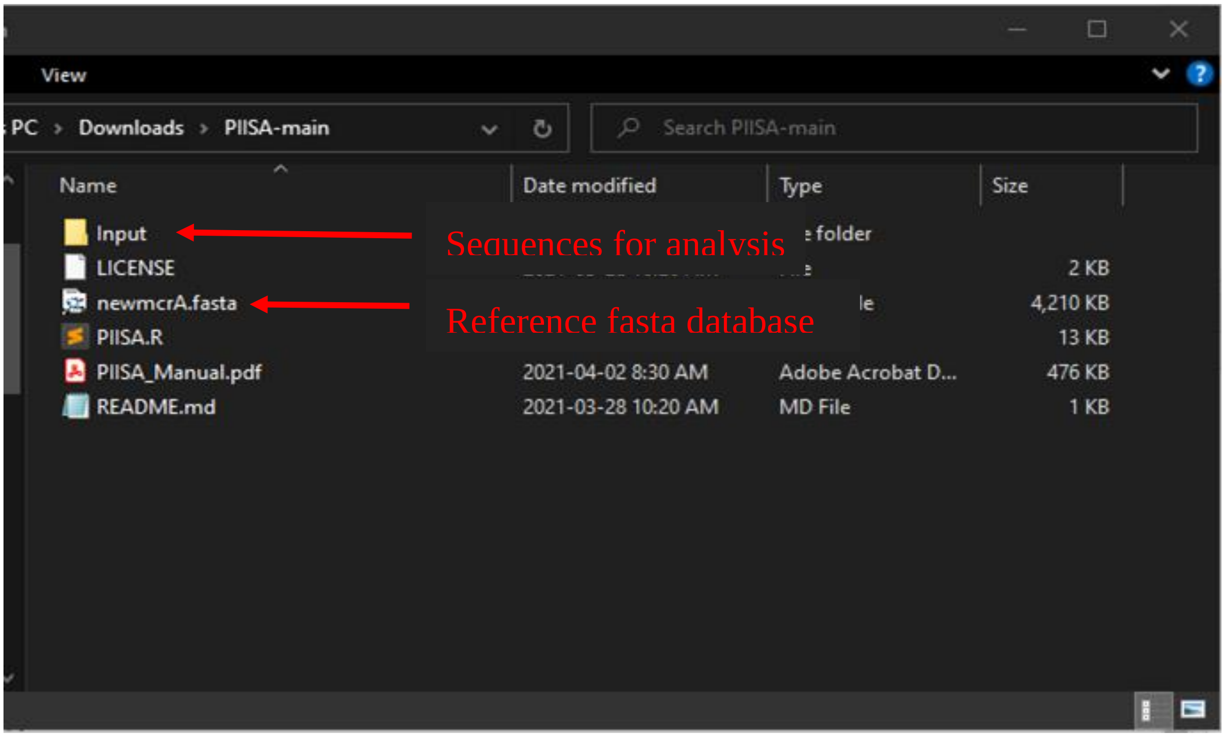
Getting Started

The first step to using PIISA is to open the PIISA.R file using the RStudio IDE. The simplest way to do so is to open the file with RStudio directly from a file explorer window as shown below. It is also possible to open the RStudio IDE first and then open the PIISA.R file from there.



Name	Date modified	Type	Size
Input	2021-03-28 10:20 AM	File folder	
LICENSE	2021-03-28 10:20 AM	File	2 KB
newmcrA.fasta	2021-03-28 10:20 AM	FASTA File	4,210 KB
PIISA.R	2021-03-28 10:20 AM	R File	13 KB
README.md	2021-03-28 10:		

Prior to running the PIISA script, the files containing sequence data should be placed in the "Input" folder, and the Fasta database(s) for assigning taxonomy should be placed in the base directory as shown below. Fasta files for several reference databases can be found through the DADA2 GitHub site which is linked in the FAQ section at the end of the tutorial.

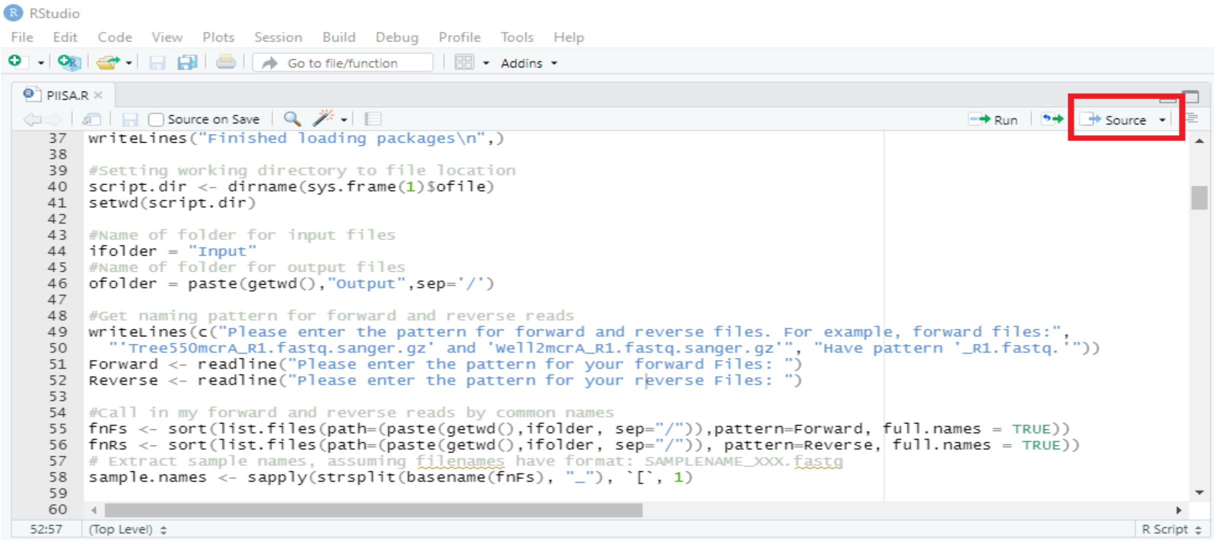


Name	Date modified	Type	Size
Input		folder	
LICENSE			2 KB
newmcrA.fasta		le	4,210 KB
PIISA.R			13 KB
PIISA_Manual.pdf	2021-04-02 8:30 AM	Adobe Acrobat D...	476 KB
README.md	2021-03-28 10:20 AM	MD File	1 KB

Sequences for analysis

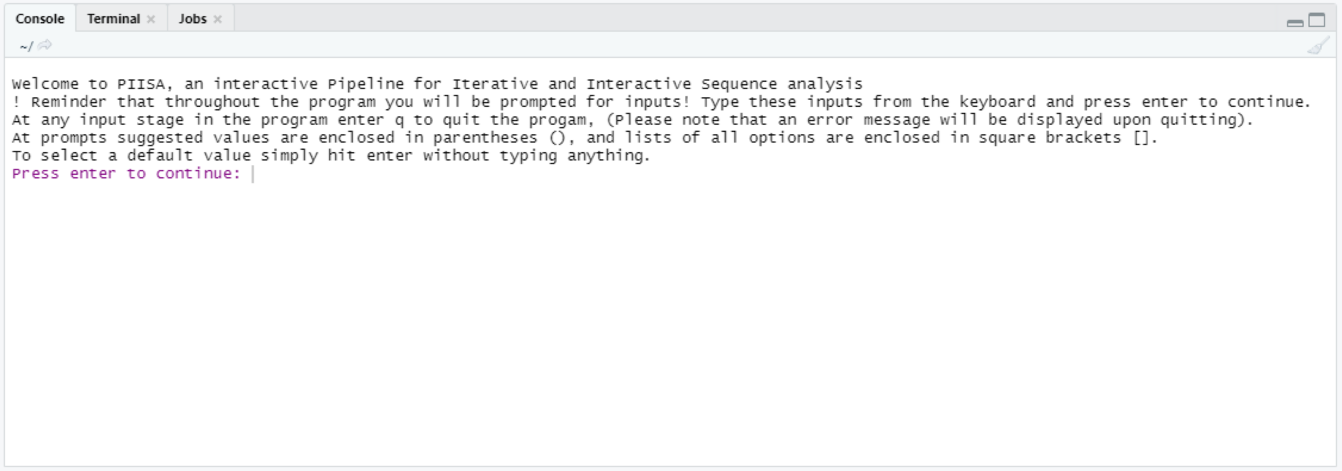
Reference fasta database

Once all required files have been placed in the appropriate locations, the program can be run by clicking the "Source" button in the RStudio IDE as shown below.



```r
writeLines("Finished loading packages\n",)

#Setting working directory to file location
script.dir <- dirname(sys.frame(1)$ofile)
setwd(script.dir)

#Name of folder for input files
ifolder = "Input"
#Name of folder for output files
ofolder = paste(getwd(),"Output",sep='/')

#Get naming pattern for forward and reverse reads
writeLines(c("Please enter the pattern for forward and reverse files. For example, forward files:",
  "'Tree550mcrA_R1.fastq.sanger.gz' and 'Well2mcrA_R1.fastq.sanger.gz'", "Have pattern '_R1.fastq.'"))
Forward <- readline("Please enter the pattern for your forward Files: ")
Reverse <- readline("Please enter the pattern for your reverse Files: ")

#Call in my forward and reverse reads by common names
fnFs <- sort(list.files(path=(paste(getwd(),ifolder, sep="/")),pattern=Forward, full.names = TRUE))
fnRs <- sort(list.files(path=(paste(getwd(),ifolder, sep="/")), pattern=Reverse, full.names = TRUE))
# Extract sample names, assuming filenames have format: SAMPLENAME_XXX.fastq
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
```

Once the source button is pressed, all further interaction with the program will occur through the console in the lower left-hand side of the RStudio IDE as shown below.



```
Welcome to PIISA, an interactive Pipeline for Iterative and Interactive Sequence analysis
! Reminder that throughout the program you will be prompted for inputs! Type these inputs from the keyboard and press enter to continue.
At any input stage in the program enter q to quit the progam, (Please note that an error message will be displayed upon quitting).
At prompts suggested values are enclosed in parentheses (), and lists of all options are enclosed in square brackets [].
To select a default value simply hit enter without typing anything.
Press enter to continue: 
```

Plot Generation

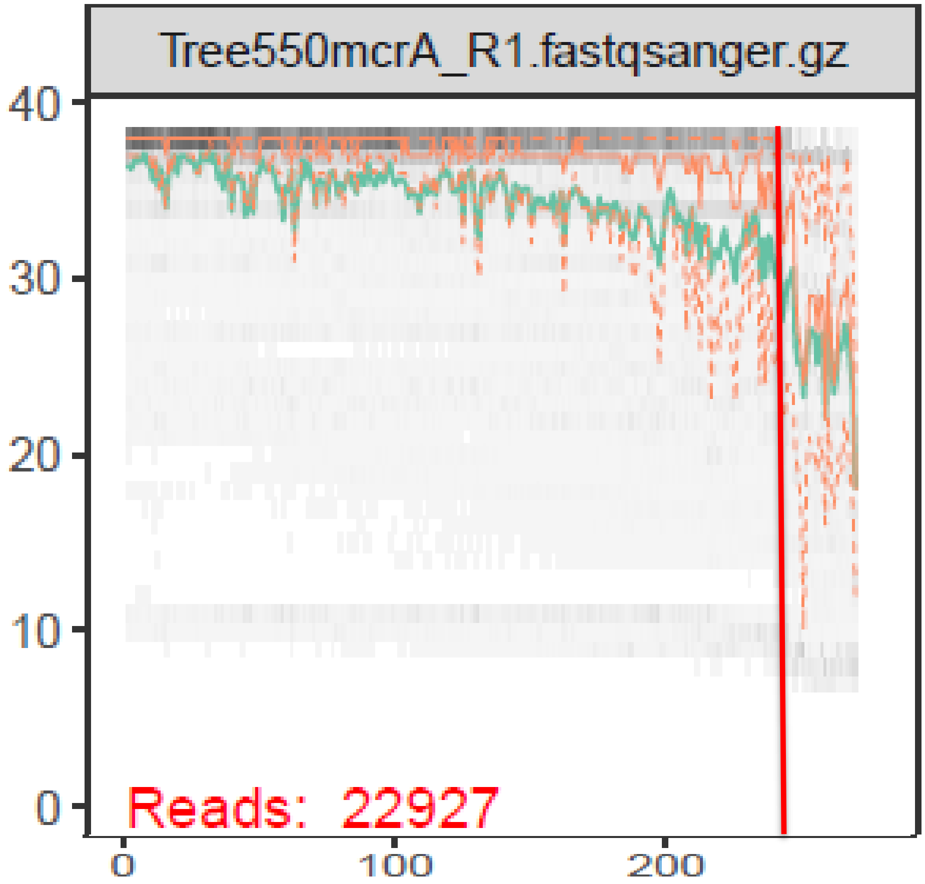
At various stages in the program, the user will be prompted for input regarding the sizes, fonts, and pdf size of created plots. For each input value, defaults are provided that the user can select if they do not need to change the plot sizes. As these steps are common for each plot generated, they will be summarized once here. The first two prompts ask the user for the width and height of the plot, please note that this is the size of the plot itself, not the paper size of the pdf generated. Following that, the user will be prompted for the type of font. Please note that the available fonts are limited compared to a typical word processing program. The selected font must be an exact match to one of the available fonts. If an incorrect font is entered, a prompt will appear that allows the user to see all available fonts. The final prompt allows the user to select the paper type for the generated pdf. The default value is always 8.5”x11” legal paper, although for abundance plots the default is in landscape configuration rather than portrait. If the user selects “n” to indicate they do not want standard sized paper, the plots will automatically be generated with a paper size that matches the input plot width and height.

Opening Files and Quality Score Plots

When the program starts, some text will be displayed displaying information about installing R packages. If any errors are encountered at this point refer to the FAQs regarding package installation. Following package installation and loading, a prompt in the console will appear asking the user to enter the pattern for forward and reverse files. This refers to a common substring of text present in all forward and reverse filenames for the input sequences. This will likely be something along the lines of "_R1.fastq." or "_R1_001.fastq." for forward reads, and "_R2.fastq." or "_R2_001.fastq.". Ensure that the patterns for forward and reverse read files are an exact match for a portion of the filename for each input file, or the input files will fail to open. Once the files are opened the program will plot the quality scores in pdf documents for the forward and reverse reads and place them in the "Output" folder. If there is no "Output" folder present, one will be automatically created. It should be noted that each time the program is run, any items in the "Output" folder from previous runs will be overwritten. As such, any results you wish to keep should be removed from the "Output" folder and stored elsewhere prior to running the program again. Alternatively, the "Output" folder can simply be renamed, which will prevent it from being overwritten with future runs. The quality score plots generated in this stage should be reviewed and used to inform decisions about selection of filtering and trimming parameters for the next phase of analysis.

Filtering, Trimming, and DADA Analysis

Once the sequence files are open and quality scores have been plotted, you will enter the filtering and trimming phase. The first prompt will ask you whether you are using a windows machine, simply enter 'y' if you are using a windows pc, or 'n' if you are using mac or Linux. Following that, you will be prompted to enter truncation values for the forward and reverse reads. Prior to entering these you will need to review the quality scores generated previously and placed in the "Output" folder. These truncation values dictate where the reads will be trimmed on the right side. Looking at the example quality score below, we can see that the quality scores are considerably worse at the end of the read, which is typical for Illumina sequencing. In this case we might select a truncation value of approximately 230 which is indicated by the vertical red line in the image below. It is quite common for the reverse reads to be of worse quality than the forward reads, so they may need to be trimmed more. To allow for merging of forward and reverse reads it is important to ensure the trimmed reads still maintain overlap. If you find that too many reads are being removed in the merging step, you may have set truncation value too low, and the truncation value should be raised. If you get a warning that none of the reads are making it through the filter, you may have set the truncation value too high, and should lower it. In general, a good place to start for truncation values is around 240, and the adjustment from there can be done based on the result seen in following sections.



Following right truncation of the read, you will be prompted for left trim values. This can be used if you have not removed the primers from your sequences. In this case you will simply enter the length (in nucleotides) of the forward and reverse primers for the forward and reverse reads, respectively. For best results, you may want to use a specific tool for removing primers such as the Cutadapt tool available for free through Galaxy Project. If your primers have already been removed, you can simply use the default value of 0 for both forward and reverse reads, as no trimming will be required. After trimming parameters have been entered, you will be prompted to enter maximum expected error (maxEE) values for the forward and reverse reads. Reducing maxEE may speed up later processing of data, however, if your maxEE is set too low the filtering step may remove too many sequences. If too many sequences are being removed, the maxEE can be relaxed by increasing it. If the reverse reads are of significantly lower quality, you may want to increase the maxEE for the reverse reads only. It is recommended to start with a maxEE of 2 for both forward and reverse reads and increase or decrease from there if needed based on the results of the following steps. Once filtering and trimming is complete, the filtered sequences will be places in the “Filtered” folder.

Following filtering and trimming, error plots will be generated for the filtered and trimmed sequences. These will be generated as a pdf document in the “Output” folder. These generated error plots should be observed to ensure there are no large or unexpected errors in the data. In general, the plotted black lines should be an approximate fit to the plotted red line which shows expected error rates. In addition, error frequency should generally decrease as quality score increases (moving along the positive x-axis).

The next stage of analysis is running the DADA2 algorithm for denoising data. This step required only one input from the user, which is to select the type of data pooling you would like to use. Selecting ‘n’ when prompted about data pooling will cause all samples to be run independently, which is the default for DADA2. Selecting ‘y’ will cause samples to be pooled for analysis, which can increase sensitivity to uncommon sequence variants such as singletons. This will also increase the computation time and memory requirements, which may be difficult to run on lower end computers. Pseudo-pooling can be selected by entering ‘p’ and offers similar benefits to pooling data without the same increase in computation requirements for the system.

The final step of this phase of analysis involves merging forward and reverse reads and removing chimeric sequences. You will be prompted to enter the minimum overlap for merging reads (12 by default), which can be increased for more stringent matching requirements. You will also be prompted to enter a maximum allowed mismatch within the overlap (0 by default, requiring matches to be identical). The maximum allowed mismatch can be increased if quality of reads is low and too many reads are being discarded during merging. Following merging, chimeras will be removed and the fraction on non-chimeric sequences remaining will be displayed. The majority of sequences should still be present after removing chimeras, and a number of ~0.85 or more is desirable. If too many chimeras are being removed, check previous processing steps, and consider removing primers with Cutadapt if present.

This part of the analysis is now completed, and all summary outputs will be created in the "Output" folder. At this point a prompt will appear asking if you would like to re-run this step of the analysis. If you are happy with these results you can select 'n' to move on to the next step of analysis. If not, select 'y' to rerun this step, allowing you to tweak input parameters to the model to achieve better results. If you select 'y' to run again, all the same results will be created in the "Output" folder, with the number 2 appended to the end. This will allow for comparison of each run to select the best model parameters. It should be noted that if this step is run multiple times, the data and values from the last run will be used in previous steps, as such, if optimal results were obtained in a previous run, re-run this stage one more time using the parameters from the previous run.

Assigning Taxonomy

The next step of analysis is assigning taxonomy to the processed sequences using a reference database. You will be prompted to enter the name of the database you are using for comparison, which must be entered as an exact match for the database file name. Following this, a prompt asking if you would like to allow reverse complement classification. By default, 'n' should be selected to perform taxonomic assignment normally. However, if you find your sequences are not being assigned as expected, it is possible that the reference database sequences are in the opposite orientation from your sequences. In this case you can select 'y' to see if assigning taxonomy to the sequences in reverse gives more expected results. Finally, you will be prompted to enter a minimum bootstrap value, which is the minimum confidence value for assigning a taxonomic level using the RDP Naive Bayesian Classifier. The default value for this is 50, however for more confident taxonomic assignment a value of 80 or more is recommended. Following taxonomic assignment, a csv file containing the results for your sequences will be created in the "Output" folder. Similar to the previous step, you will then be asked if you would like to repeat this step. Selecting 'y' will allow you to repeat taxonomic assignment, allowing you to use a different reference database, try reverse complement classification, or change the bootstrap values. Similar to the above step, the data from the most recent run will be used in the next step.

Abundance and Diversity Plots

Once taxonomic assignment is completed, the final stage of PIISA is to generate some basic abundance and diversity plots. These include Simpson, Shannon, Observed, and Chao1 diversity plots. Abundance plots for each level of taxonomy will also be provided. It should be noted that read count is not normalized, so abundance plots represent the proportion of reads in the ESV table generated by assigning taxonomy. These plots will be generated without any input required from the user. It should be noted that if you have very few samples, these plots may fail to generate, or warning messages stating you have few data points may appear. All generated plots will be place automatically within the "Output" folder.

Frequently Asked Questions

Q. Automatic installation for one of the required packages failed, what should I do now?

The package that failed to install will need to be installed manually in RStudio. Instructions for installation of each required package can be found at the following links.

https://www.bioconductor.org/install/

https://benjjneb.github.io/dada2/dada-installation.html

https://www.bioconductor.org/packages/release/bioc/html/DECIPHER.html https://ggplot2.tidyverse.org/

https://bioconductor.org/packages/release/bioc/html/phyloseq.html

Q. Somewhere in the pipeline I received an error message saying I cannot open a certain file.

A. In this case if the file that can't be opened is an input file ensure you correctly enter the filename at the prompt to do so. Any misspellings including capitalization will cause an error in opening files. If the file that fails to open is one of the output files the program creates, ensure that any previously generated output files are closed prior to running the program.

Q. Some steps of the program take a long time to run, sometimes up to a few minutes.

A. Depending on your data set and computer hardware this is completely normal. Several steps including filtering and trimming and learning error rates may take several minutes if you have a large data set or a lower end or older computer.

Q. There are some warning messages that are displayed after I run the program.

A. Generally warning messages are not a problem, but error messages are. Several warning messages regarding infinite transformation in the y axis during error plotting are normal to see and not of concern. In addition, if you do not have many samples, several warning messages will be displayed related to the generation of diversity and abundance plots. It is also normal to receive certain warning messages about dependencies of the packages installed for this program, as long as the data has been processed normally, these are not cause for concern.

Most of my sequences are being discarded during chimera removal.

A. Often many sequences will appear to be discarded, but these are often singletons or low frequency sequences that make up a very small amount of your actual reads. If the fraction of remaining sequences is high, then there is no cause for concern. If the fraction is also low rerun the earlier stages of the pipeline and tweak input parameters for filtering and trimming, ensuring sequences aren't trimmed so short that they no longer overlap. Finally, if you have not removed the primers from your sequences, consider using a tool like Cutadapt to do so, especially if you are analyzing sequences where PCR was performed with more than one primer set.

Q. Something caused an error in the pipeline which forced it to exit partway through, and now I can't open on of the output pdf files.

A. In this case R likely still has the file opened for editing. This can be fixed by closing and restarting RStudio or typing "dev.off()" into the console and pressing enter. You may need to do this more than once depending on where in the process the interruption occurred.

Q. Where can I find a reference database for assigning taxonomy?

A. Several reference databases for different uses can be found at the link below.

https://benjjneb.github.io/dada2/training.html

References

Callahan, B. J. (2020). DADA2 pipeline tutorial (1.8). Retrieved from https://benjjneb.github.io/dada2/tutorial_1_8.html

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/NMETH.3869

Callahan, B. J., McMurdie, P. & Holmes, S. (2021). DADA2:1.18.0 Retrieved from https://www.bioconductor.org/packages/release/bioc/manuals/dada2/man/dada2.pdf

McMurdie, P., Holmes, S. (2019). phyloseq:1.34.0 Retrieved from https://bioconductor.org/packages/release/bioc/html/phyloseq.html