Horse nemabiome R and PowerShell workflow

25/01/2022

## Separation of unique individual sequences

The individual sequences of each species and unclassified groups were separated using *get.lineage* command in Mothur. This produced separate fasta files for each species or unclassified group. These files were then put in a new folder and processed further using R loops.

The following R packages were used in this process:

library(microseq)  
library(tidyverse)  
library(seqRFLP)  
library(DECIPHER)  
library(readxl)

The working directory was set to the folder containing files, and a list of file names was generated. 3 subfolders (aggregated, trimmed, final) were also generated to save the files after each step of processing

setwd("C:/Horse")  
list <- sub(".fasta", "", list.files(pattern = ".fasta"))  
dir.create("aggregated")  
dir.create("trimmed")  
dir.create("final")

In the first step, the sequences were aggregated to merge the identical ones. The top 1000 most common sequences from each group/species present at least six times were then saved as new fasta files in the “aggregated” folder. Excel files were also generated for each group/species to show how many identical sequences were merged for each existing sequence.

for (i in list) {  
 loadfile <- file.path(paste0(i, ".fasta"))  
 Data <- subset(readFasta(loadfile),select = 2,)  
 Data2 <- aggregate(list(numdup=rep(1,nrow(Data))), Data, length)  
 Data3 <- Data2[order(-Data2$numdup),]  
 Data3$name <- i  
 Data3$number <- c(1:nrow(Data3))  
 Data3 <- Data3 %>%  
 unite("Header", name:number, remove = TRUE)  
 Data3 <- Data3[, c(3,1,2)]  
 Data3 <- Data3 %>% filter(numdup > 5) %>% head(1000)  
 excelfile <- file.path("aggregated", paste0(i, ".xlsx"))  
 write\_xlsx(Data3, excelfile)  
 fastafile <- file.path("aggregated", paste0(i, ".fasta"))  
 Data3.fasta = dataframe2fas(Data3[,c(1:2)], file=fastafile)}

The second step involved the alignment of these sequences and trimming of extra bases from both sides of the primers, thus making sure all the sequences were starting and ending precisely at the same point. For this purpose, parts of forward and reverse nemabiome primers (“ACGTCT”, “AACTAA”) were identified, and any bases outside of those were removed. These sequences were again saved as new fasta files in the “trimmed” folder.

for (i in list) {  
 loadfile <- file.path("aggregated", paste0(i, ".fasta"))  
 seqs <- RemoveGaps(readDNAStringSet(loadfile, format = "fasta"))  
 aligned <- AlignSeqs(seqs, iterations = 5, refinements = 5)  
 adjusted <- AdjustAlignment(aligned)  
 trimmed <- TrimDNA(adjusted, "ACGTCT", "AACTAA", type = "sequences")  
 trimmedfile <- file.path("trimmed", paste0(i, ".fasta"))  
 writeXStringSet(trimmed, file=trimmedfile)}

Finally, the aligned and trimmed sequences were aggregated again; unique sequences present at least twice (out of the top 1000 selected in step 1) were saved as new fasta files in the “final” folder and used for the following stages.

for (i in list) {  
 loadfile <- file.path("trimmed", paste0(i, ".fasta"))  
 Data <- subset(readFasta(loadfile),select = 2,)  
 Data2 <- aggregate(list(numdup=rep(1,nrow(Data))), Data, length)  
 Data3 <- Data2[order(-Data2$numdup),]  
 Data3$name <- i  
 Data3$number <- c(1:nrow(Data3))  
 Data3 <- Data3 %>%  
 unite("Header", name:number, remove = TRUE)  
 Data3 <- Data3[, c(3,1,2)]  
 excelfile <- file.path("final", paste0(i, " final", ".xlsx"))  
 write\_xlsx(Data3, excelfile)  
 Data3 <- Data3 %>% filter(numdup > 1)  
 fastafile <- file.path("final", paste0(i, ".fasta"))  
 Data3.fasta = dataframe2fas(Data3[,c(1:2)], file=fastafile)}

## 

## Identity Confirmation by blasting

Using Windows PowerShell and NCBI blast-2.12.0+, all the unique sequences were blasted to check their identities. This produced a table showing the top 5 results (including query cover, identity and description) for each sequence. The purpose of this was to make sure none of these sequences was misidentified. In the case of unclassified sequences, this helped ensure they did not belong to any known nematode species. The PowerShell script used for this is pasted below.

cd \\...\NCBI\blast-2.12.0+\bin  
  
.\blastn -remote -db nt -query input\_file\_name.fasta -outfmt "6 qseqid sscinames pident qcovs" -max\_target\_seqs 5 -out \\...\output\_file\_name.tab