EDNA Data

Ashley Grinstead

4/4/2022

Libraries Used

```
#Load libraries

library(tidyverse)
library(here)
library(janitor)
library(readxl)
library(stringr)
```

Read Data

```
# Import EDNA data
edna_info <- read.csv("data/edna_info.csv") %>%
  # Clean names so that they are all lowercase and have underscores
  clean_names()

# edna_sites <- read.csv("data/edna_sites.csv") %>%
#   clean_names() %>%
#   subset(select=-date) %>%
#   subset(barcode != "") %>%
#   rename(id_number = barcode, site = i_site_name) %>%
#   select(id_number, site) %>%
#   pivot_wider(names_from = "id_number", values_from = "site") %>%
#   clean_names() %>%
#   pivot_longer(1:10, names_to = "id_number", values_to = "site")

# Import JVB data
jvb_12s <- read.csv("data/JVB1470-12S.csv") %>%
  clean_names()

jvb_16s <- read.csv("data/JVB1470-16S.csv") %>%
  clean_names()

jvb_23s <- read.csv("data/JVB1470-23S.csv") %>%
  clean_names()

jvb_arth_coi <- read.csv("data/JVB1470-ArthCOI.csv") %>%
  clean_names()
```

```
jvb_its <- read.csv("data/JVB1470-ITS.csv") %>%
  clean_names()

jvb_18s <- read.csv("data/JVB1470-18S.csv") %>%
  clean_names()

jvb_trnl <- read.csv("data/JVB1470-trnL.csv") %>%
  clean_names()

jvb_unicoi <- read.csv("data/JVB1470-UniCOI.csv") %>%
  clean_names()

# jvb_mifishu <- read.csv("data/JVB1470-MiFishU.csv") %>%
#   clean_names()
```

Data Organization

```
jvb_12s1 <- jvb_12s %>%
  #selecting columns of sample names with data in it
  select(test_id:s034168) %>%
  # Renaming columns that had symbols
  rename(percent_match = x_match) %>%
  rename(number_species = x_species)

jvb_16s1 <- jvb_16s %>%
  select(test_id:s034128) %>%
  rename(percent_match = x_match) %>%
  rename(number_species = x_species)

jvb_23s1 <- jvb_23s %>%
  select(test_id:s034210) %>%
  rename(percent_match = x_match) %>%
  rename(number_species = x_species)

jvb_arth_coi1 <- jvb_arth_coi %>%
  select(test_id:cvi9jr2k) %>%
  rename(percent_match = x_match) %>%
  rename(number_species = x_species)

jvb_its1 <- jvb_its %>%
  select(test_id:s034128) %>%
  rename(percent_match = x_match) %>%
  rename(number_species = x_species)

jvb_18s1 <- jvb_18s %>%
  select(test_id:s034148) %>%
  rename(percent_match = x_match) %>%
  rename(number_species = x_species)

jvb_trnl1 <- jvb_trnl %>%
  select(test_id:s034126) %>%
```

```
  rename(percent_match = x_match) %>%
  rename(number_species = x_species)

jvb_unicoi1 <- jvb_unicoi %>%
  select(test_id:s034168) %>%
  rename(percent_match = x_match) %>%
  rename(number_species = x_species)

# jvb_mifishu1 <- jvb_mifishu %>%
#   select(test_id:cvi9jr2k) %>%
#   rename(percent_match = x_match) %>%
#   rename(number_species = x_species)
```

Merging Columns

```
jvb_12s1 <- jvb_12s1 %>%
  # Merge kingdom, phylum, class, order, family, genus, and species columns
  # KPCOFGS = kingdom, phylum, class, order, family, genus, species
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_16s1 <- jvb_16s1 %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_23s1 <- jvb_23s1 %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_arth_coi1 <- jvb_arth_coi1 %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_its1 <- jvb_its1 %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_18s1 <- jvb_18s1 %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_trnl1 <- jvb_trnl1 %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_unicoi1 <- jvb_unicoi1 %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

# jvb_mifishu1 <- jvb_mifishu1 %>%
#   unite("kpcofgs", kingdom:species, sep = ", ")
```

More Data Organization

```
# Pivot wider to set ID numbers to the same format as the ID numbers in the
# JVB datasets
edna_info1 <- edna_info %>%
```

```
  pivot_wider(1:2, names_from = "dna_vial", values_from = "sample_type") %>%
  clean_names() %>%
  # Pivot longer to format into rows instead of columns
  pivot_longer(1:26, names_to = "id_number", values_to = "sample_type")

edna_info2 <- edna_info %>%
  select(dna_vial, ucsb_id) %>%
  pivot_wider(names_from = "dna_vial", values_from = "ucsb_id") %>%
  clean_names() %>%
  pivot_longer(1:26, names_to = "id_number", values_to = "ucsb_id")

edna_info3 <- inner_join(edna_info1, edna_info2, by = "id_number")

# Create new dataframe to not mess up old one for editing purposes
jvb_12s2 <- jvb_12s1 %>%
  # Selecting relevant data and leaving out test_id & sequence
  select(esv_id:s034168) %>%
  # Pivot longer to change from columns to rows
  pivot_longer(6:7, names_to = "id_number", values_to = "sample_number") %>%
  # Add in sample type to ID what type of sample we are looking at
  inner_join(edna_info3, by = "id_number") %>%
  # Reorder the columns so that ID numbers is first
  select(esv_id, id_number, kpcofgs, sample_type, ucsb_id, sample_number, number_species,
         percent_match) %>%
  # Group by columns without numbers
  group_by(esv_id, kpcofgs, sample_type, ucsb_id) %>%
  # Remove values of 0
  filter(sample_number != 0) %>%
  # Summarize the sample number, percent match, and number of species
  summarise(total_sample_number = sum(sample_number),
            total_percent_match = sum(percent_match),
            total_number_species = sum(number_species))

jvb_16s2 <- jvb_16s1 %>%
  select(esv_id:s034128) %>%
  pivot_longer(6:17, names_to = "id_number", values_to = "sample_number") %>%
  inner_join(edna_info3, by = "id_number") %>%
  select(esv_id, id_number, kpcofgs, sample_type, ucsb_id, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, sample_type, ucsb_id) %>%
  filter(sample_number != 0) %>%
  # There are duplicate esv_id rows so distinct() keeps only one of them to
  # keep the percentages at max 100%
  distinct(esv_id, .keep_all = TRUE) %>%
  summarise(total_sample_number = sum(sample_number),
            total_percent_match = sum(percent_match),
            total_number_species = sum(number_species))

jvb_23s2 <- jvb_23s1 %>%
  select(esv_id:s034210) %>%
  pivot_longer(6, names_to = "id_number", values_to = "sample_number") %>%
  inner_join(edna_info3, by = "id_number") %>%
  select(esv_id, id_number, kpcofgs, sample_type, ucsb_id, sample_number, number_species,
```

```
         percent_match) %>%
  group_by(esv_id, kpcofgs, sample_type, ucsb_id) %>%
  filter(sample_number != 0) %>%
  summarise(total_sample_number = sum(sample_number),
            total_percent_match = sum(percent_match),
            total_number_species = sum(number_species))

jvb_arth_coi2 <- jvb_arth_coi1 %>%
  select(esv_id:s034148) %>%
  pivot_longer(6:10, names_to = "id_number", values_to = "sample_number") %>%
  inner_join(edna_info3, by = "id_number") %>%
  select(esv_id, id_number, kpcofgs, sample_type, ucsb_id, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, sample_type, ucsb_id) %>%
  filter(sample_number != 0) %>%
  summarise(total_sample_number = sum(sample_number),
            total_percent_match = sum(percent_match),
            total_number_species = sum(number_species))

jvb_its2 <- jvb_its1 %>%
  select(esv_id:s034128) %>%
  pivot_longer(6:17, names_to = "id_number", values_to = "sample_number") %>%
  inner_join(edna_info3, by = "id_number") %>%
  select(esv_id, id_number, kpcofgs, sample_type, ucsb_id, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, sample_type, ucsb_id) %>%
  filter(sample_number != 0) %>%
  distinct(esv_id, .keep_all = TRUE) %>%
  summarise(total_sample_number = sum(sample_number),
            total_percent_match = sum(percent_match),
            total_number_species = sum(number_species))

jvb_18s2 <- jvb_18s1 %>%
  select(esv_id:s034148) %>%
  pivot_longer(6:9, names_to = "id_number", values_to = "sample_number") %>%
  inner_join(edna_info3, by = "id_number") %>%
  select(esv_id, id_number, kpcofgs, sample_type, ucsb_id, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, sample_type, ucsb_id) %>%
  filter(sample_number != 0) %>%
  distinct(esv_id, .keep_all = TRUE) %>%
  summarise(total_sample_number = sum(sample_number),
            total_percent_match = sum(percent_match),
            total_number_species = sum(number_species))

jvb_trnl2 <- jvb_trnl1 %>%
  select(esv_id:s034126) %>%
  pivot_longer(6:8, names_to = "id_number", values_to = "sample_number") %>%
  inner_join(edna_info3, by = "id_number") %>%
  select(esv_id, id_number, kpcofgs, sample_type, ucsb_id, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, sample_type, ucsb_id) %>%
  filter(sample_number != 0) %>%
```

```
  distinct(esv_id, .keep_all = TRUE) %>%
  summarise(total_sample_number = sum(sample_number),
            total_percent_match = sum(percent_match),
            total_number_species = sum(number_species))

jvb_unicoi2 <- jvb_unicoi1 %>%
  select(esv_id:s034168) %>%
  pivot_longer(6, names_to = "id_number", values_to = "sample_number") %>%
  inner_join(edna_info3, by = "id_number") %>%
  select(esv_id, id_number, kpcofgs, sample_type, ucsb_id, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, sample_type, ucsb_id) %>%
  filter(sample_number != 0) %>%
  distinct(esv_id, .keep_all = TRUE) %>%
  summarise(total_sample_number = sum(sample_number),
            total_percent_match = sum(percent_match),
            total_number_species = sum(number_species))

# jvb_mifishu2 <- jvb_mifishu1 %>%
#   select(esv_id:pskoyg86) %>%
#   pivot_longer(5:7, names_to = "id_number", values_to = "sample_number") %>%
#   inner_join(edna_sites, by = "id_number") %>%
#   select(esv_id, id_number, kpcofgs, sample_type, sample_number, number_species,
#          percent_match) %>%
#   group_by(esv_id, kpcofgs, sample_type) %>%
#   filter(sample_number != 0) %>%
#   distinct(esv_id, .keep_all = TRUE) %>%
#   summarise(total_sample_number = sum(sample_number),
#             total_percent_match = sum(percent_match),
#             total_number_species = sum(number_species))
```

Merge Datasets

```
# Merge all JVB datasets into one
jvb <- rbind(jvb_12s2, jvb_16s2, jvb_23s2, jvb_arth_coi2, jvb_its2, jvb_18s2, jvb_trnl2, jvb_unicoi2)
```

Create CSV Files for Different Sample Types

```
owl_pellet <- jvb %>%
  filter(grepl("owl", sample_type)) %>%
  select(esv_id, kpcofgs, ucsb_id, total_sample_number, total_percent_match, total_number_species) %>%
  write.csv("output/owl_pellet.csv")

aquatic <- jvb %>%
  filter(grepl("MO1|NVBR", sample_type)) %>%
  write.csv("output/aquatic.csv")

soil <- jvb %>%
  filter(!grepl("owl|dropping|MO1|NVBR", sample_type)) %>%
```

```
  write.csv("output/soil.csv")

dropping <- jvb %>%
  filter(grepl("dropping", sample_type)) %>%
  write.csv("output/dropping.csv")

pollen <- jvb %>%
  filter(grepl("Pollen", sample_type)) %>%
  select(esv_id, kpcofgs, ucsb_id, total_sample_number, total_percent_match, total_number_species) %>%
  write.csv("output/pollen.csv")
```