EDNA Data

Ashley Grinstead

06/07/2022

Libraries Used

```r
# Load libraries

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

Read Data

```r
# Import EDNA data

edna_info <- read.csv("data/edna_info.csv") %>%

  # Clean names so that all column titles are lowercase and have underscores

  clean_names()
```

```r
# Read in batch JVB1554 data sites

jvb1554_sites <- read.csv("data/JVB1554_sample.csv") %>%
  clean_names() %>%

  # Select columns of interest

  select(sample_id, site_name)
```

```r
# Import batch JVB1470 data

jvb_12s1470 <- read.csv("data/JVB1470-12S.csv") %>%
  clean_names()

jvb_16s1470 <- read.csv("data/JVB1470-16S.csv") %>%
  clean_names()

jvb_23s1470 <- read.csv("data/JVB1470-23S.csv") %>%
  clean_names()

jvb_arthcoi1470 <- read.csv("data/JVB1470-ArthCOI.csv") %>%
  clean_names()

jvb_its1470 <- read.csv("data/JVB1470-ITS.csv") %>%
  clean_names()

jvb_18s1470 <- read.csv("data/JVB1470-18S.csv") %>%
  clean_names()

jvb_trnl1470 <- read.csv("data/JVB1470-trnL.csv") %>%
  clean_names()

jvb_unicoi1470 <- read.csv("data/JVB1470-UniCOI.csv") %>%
  clean_names()

jvb_mifishu1470 <- read.csv("data/JVB1470-MiFishU.csv") %>%
  clean_names()
```

```
# Import batch JVB1554 data

jvb_mifishu1554 <- read.csv("data/JVB1554-MiFishU.csv") %>%
  clean_names()

jvb_arthcoi1554 <- read.csv("data/JVB1554-ArthCOI.csv") %>%
  clean_names()

jvb_23s1554 <- read.csv("data/JVB1554-23S.csv") %>%
  clean_names()
```

Data Organization

```
# Create a new variable so original variable isn't messed with; I added a "1"
# after the original naming convention indicating the new variable.

jvb_12s14701 <- jvb_12s1470 %>%

  # Look at the table and select the columns with data in it

  select(test_id:s034168) %>%

  # Rename columns that had symbols (%,#) instead of characters

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

  # Merge kingdom, phylum, class, order, family, genus, and species columns
  # KPCOFGS = kingdom, phylum, class, order, family, genus, species

  unite("kpcofgs", kingdom:species, sep = ", ")
```

```
jvb_16s14701 <- jvb_16s1470 %>%
  select(test_id:s034128) %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_23s14701 <- jvb_23s1470 %>%
  select(test_id:s034210) %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_arthcoi14701 <- jvb_arthcoi1470 %>%
  select(test_id:cvi9jr2k) %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_its14701 <- jvb_its1470 %>%
  select(test_id:s034128) %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_18s14701 <- jvb_18s1470 %>%
  select(test_id:s034148) %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_trnl14701 <- jvb_trnl1470 %>%
  select(test_id:s034126) %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_unicoi14701 <- jvb_unicoi1470 %>%
  select(test_id:s034168) %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_mifishu14701 <- jvb_mifishu1470 %>%
  select(test_id:cvi9jr2k) %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")
```

```
jvb_mifishu15541 <- jvb_mifishu1554 %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")
```

```
jvb_arthcoi15541 <- jvb_arthcoi1554 %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")

jvb_23s15541 <- jvb_23s1554 %>%
  rename(percent_match = x_match,
         number_species = x_species) %>%
  unite("kpcofgs", kingdom:species, sep = ", ")
```

More Data Organization

```
# Create first variable to match ID numbers with sample types

edna_info1 <- edna_info %>%

  # Pivot wider to set ID numbers to the same format as the ID numbers in the
  # JVB datasets

  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")


# create second variable to match ID numbers to UCSB IDs

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")


# Create third variable to join first two variables together by ID number

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

```
# Do the same for batch JVB1554 (this one is by site name, not sample type)

jvb1554_info <- jvb1554_sites %>%
  pivot_wider(1:2, names_from = "sample_id", values_from = "site_name") %>%
  clean_names() %>%
  pivot_longer(1:12, names_to = "sample_id", values_to = "site_name")
```

```
# Create new dataframe to again for next data organization wave

jvb_12s14702 <- jvb_12s14701 %>%
```

```r
  # 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_16s14702 <- jvb_16s14701 %>%
  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_23s14702 <- jvb_23s14701 %>%
  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_arthcoi14702 <- jvb_arthcoi14701 %>%
  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_its14702 <- jvb_its14701 %>%
  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_18s14702 <- jvb_18s14701 %>%
  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_trnl14702 <- jvb_trnl14701 %>%
  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) %>%
```

```r
  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_unicoi14702 <- jvb_unicoi14701 %>%
  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_mifishu14702 <- jvb_mifishu14701 %>%
  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))
```

```
## Error in is.data.frame(y): object ’edna_sites’ not found
```

```r
jvb_23s15542 <- jvb_23s15541 %>%
  select(esv_id:x0cel3no8) %>%
  pivot_longer(6:17, names_to = "sample_id", values_to = "sample_number") %>%
  inner_join(jvb1554_info, by = "sample_id") %>%
  select(esv_id, sample_id, kpcofgs, site_name, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, site_name) %>%
  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_mifishu15542 <- jvb_mifishu15541 %>%
  select(esv_id:x0cel3no8) %>%
  pivot_longer(6:15, names_to = "sample_id", values_to = "sample_number") %>%
  inner_join(jvb1554_info, by = "sample_id") %>%
  select(esv_id, sample_id, kpcofgs, site_name, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, site_name) %>%
  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_arthcoi15542 <- jvb_arthcoi15541 %>%
  select(esv_id:x0cel3no8) %>%
  pivot_longer(6:17, names_to = "sample_id", values_to = "sample_number") %>%
  inner_join(jvb1554_info, by = "sample_id") %>%
  select(esv_id, sample_id, kpcofgs, site_name, sample_number, number_species,
         percent_match) %>%
  group_by(esv_id, kpcofgs, site_name) %>%
  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 from batch 1 into one

jvb1 <- rbind(jvb_12s2, jvb_16s2, jvb_23s2, jvb_arth_coi2,
             jvb_its2, jvb_18s2, jvb_trnl2, jvb_unicoi2)
```

```
## Error in rbind(jvb_12s2, jvb_16s2, jvb_23s2, jvb_arth_coi2, jvb_its2, : object 'jvb_12s2' not found
```

```
# Merge all JVB datasets from batch 2 into one

jvb2 <- rbind(jvb_23s15542, jvb_mifishu15542, jvb_arthcoi15542)
```

Create CSV Files for Different Sample Types

```
# CSV files for batch 1

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")
```

```
## Error in filter(., grepl("owl", sample_type)): object 'jvb' not found
```

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

```
## Error in filter(., grepl("MO1|NVBR", sample_type)): object ’jvb’ not found
```

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

```
## Error in filter(., !grepl("owl|dropping|MO1|NVBR", sample_type)): object ’jvb’ not found
```

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

```
## Error in filter(., grepl("dropping", sample_type)): object ’jvb’ not found
```

```
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")
```

```
## Error in filter(., grepl("Pollen", sample_type)): object ’jvb’ not found
```

```
edna <- jvb %>%
  write.csv("output/edna.csv")
```

```
## Error in is.data.frame(x): object ’jvb’ not found
```

```
# CSV files for batch 2

batch_2 <- jvb2 %>%
  write.csv("output/batch_2.csv")
```