elephants_TAX_CONFIDENCE

Joe Gunn

8/12/2019

AFRICAN ELEPHANT MICROBIOME - TAXONOMIC CLASSIFICATION OF OTUS

Purpose: Here I am assessing the confidence with which all individual microbial OTUs were assigned to a given taxonomic level (DOMAIN, KINGDOM, PHYLUM, CLASS, ORDER, FAMILY, GENUS, SPECIES). We used this analysis to determine the most specific (lowest) taxonomic level with high confidence that would be appropriate to use to compare taxonomic composition between African Elephant species, diets, and habitats.

Data used:

```
  TAXONOMY table output from QIIME: all OTUs detected across African Elephant sampels and confidence with which t
hey were assigned to each taxonomic level <- 9,066 OTUs
```

Libraries Needed for Analysis

```
##
## *********************************************************
```

```
## Note: As of version 1.0.0, cowplot does not change the
```

```
##   default ggplot2 theme anymore. To recover the previous
```

```
##   behavior, execute:
##   theme_set(theme_cowplot())
```

```
## *********************************************************
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.2.0     ✔ purrr   0.3.2
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   0.8.3     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-4
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
## Loading required package: foreign
```

```
## Loading required package: survival
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: nnet
```

```
##
## Attaching package: 'epiDisplay'
```

```
## The following object is masked from 'package:lattice':
##
##     dotplot
```

```
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: TH.data
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
##
## Attaching package: 'taRifx'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, distinct, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     rep_along
```

Metadata

Taxonomy Table (OTUs present in the total data set) and Confidence Calculations

```
##Read in Taxonomy Table. This table contains all OTUs present in the total data set, but does not provide the re
lative abundance of each taxon per sample. It only contains the taxonomic level of each OTU present and the confi
dence at which that taxonomic unit was classified.
tax_all <- read_qza("../data/qiime_data/taxonomy.qza") #read in table

tax_all<- tax_all$data %>%
  as.tibble() %>%
  separate(Taxon, sep = "; ", c("Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species")) #clean data
and label columns by taxonomic level
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
## Warning: Expected 7 pieces. Missing pieces filled with `NA` in 1469 rows
## [1, 2, 3, 4, 5, 6, 7, 9, 11, 13, 15, 16, 18, 20, 22, 24, 25, 26, 33,
## 38, ...].
```

```
tax_all_df <- as.data.frame(tax_all)
  colnames(tax_all_df) <- c("otu_names", "Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species", "co
nf")

##Calculate mean and standard deviation of confidence
mean_conf <- mean(tax_all$Confidence) #mean confidence across all OTUs (mean = 0.9615)
sd_conf <- sd(tax_all$Confidence) #standard deviation of confidence across all OTUs (sd = 0.07256)

#change all unclassified taxa that aren't already called "NA" to "NA"
tax_all$Species[tax_all$Species == "s__"] <- NA
tax_all$Genus[tax_all$Genus == "g__"] <- NA
tax_all$Family[tax_all$Family == "f__"] <- NA
tax_all$Order[tax_all$Order == "o__"] <- NA
tax_all$Class[tax_all$Class == "c__"] <- NA
tax_all$Phylum[tax_all$Phylum == "p__"] <- NA
tax_all$Kingdom[tax_all$Kingdom == "k__"] <- NA

phy_order <- tax_all[,-c(1:2,4,6:9 )]

tax_all <- as.data.frame(tax_all) #change table to data frame

#Create separate data frames for each taxon and confidence

kingdom <- as.data.frame(cbind(tax_all$Kingdom, tax_all$Confidence)) #kingdom
  colnames(kingdom) <- c("Kingdom", "Confidence") #add column names

phylum <- as.data.frame(cbind(tax_all$Phylum, tax_all$Confidence)) #phylum
  colnames(phylum) <- c("Phylum", "Confidence")

class <- as.data.frame(cbind(tax_all$Class, tax_all$Confidence)) #class
  colnames(class) <- c("Class", "Confidence")

order <- as.data.frame(cbind(tax_all$Order, tax_all$Confidence)) #order
  colnames(order) <- c("Order", "Confidence")

family <- as.data.frame(cbind(tax_all$Family, tax_all$Confidence)) #family
  colnames(family) <- c("Family", "Confidence")

genus <- as.data.frame(cbind(tax_all$Genus, tax_all$Confidence)) #genus
  colnames(genus) <- c("Genus", "Confidence")

species <- as.data.frame(cbind(tax_all$Species, tax_all$Confidence)) #species
  colnames(species) <- c("Species", "Confidence")

##Create data frames for ONLY the individual OTUs that were successfully classified at each level. Kingdom had ma
ny individual OTUs successfully classified, while Species had substantially fewer OTUs successfully classified

known_species <- as.data.frame(species[complete.cases(species),]) #species
known_genus <-  as.data.frame(genus[complete.cases(genus),]) #genus
known_family <- as.data.frame(family[complete.cases(family),]) #family
known_order <- as.data.frame(order[complete.cases(order),]) #order
known_class <- as.data.frame(class[complete.cases(class),]) #class
known_phylum <- as.data.frame(phylum[complete.cases(phylum),]) #phylum
known_kingdom <- as.data.frame(kingdom[complete.cases(kingdom),]) #kingdom

##Raw number of unique OTUs per taxonomic level (classified successfully)
#nrow(known_species) #271
#nrow(known_genus) #2279
#nrow(known_family) #6138
#nrow(known_order) #8706
#nrow(known_class) #8840
#nrow(known_phylum) #8878
#nrow(known_kingdom) #9066

otu_prop_vector <- c(271/9066, 2279/9066, 6138/9066, 8706/9066, 8840/9066, 8878/9066, 9066/9066)
otu_prop_vector <- as.data.frame(otu_prop_vector)

##mean confidence at each taxonomic level

known_species <- known_species %>% mutate(Confidence = as.character(Confidence)) #need to change the confidence v
ariable from a factor to a numeric. NOTE: you MUST first convert from factor to character, then from character to
factor. Not sure why, but otherwise it doesn't work right

  known_species <- known_species %>% mutate(Confidence = as.numeric(Confidence))
  known_genus <- known_genus %>% mutate(Confidence = as.character(Confidence))
  known_genus <-  known_genus %>% mutate(Confidence = as.numeric(Confidence))
  known_family <- known_family %>% mutate(Confidence = as.character(Confidence))
  known_family <-  known_family %>% mutate(Confidence = as.numeric(Confidence))
  known_order <- known_order %>% mutate(Confidence = as.character(Confidence))
  known_order <-  known_order %>% mutate(Confidence = as.numeric(Confidence))
  known_class <- known_class %>% mutate(Confidence = as.character(Confidence))
  known_class <-  known_class %>% mutate(Confidence = as.numeric(Confidence))
  known_phylum <- known_phylum %>% mutate(Confidence = as.character(Confidence))
  known_phylum <-  known_phylum %>% mutate(Confidence = as.numeric(Confidence))
  known_kingdom <- known_kingdom%>% mutate(Confidence = as.character(Confidence))
  known_kingdom <-  known_kingdom %>% mutate(Confidence = as.numeric(Confidence))
```

Calculate mean confidence of assigning an individual to a given taxon

```
##Species Confidence
ms <- mean(known_species$Confidence) #0.9107
ss <- sd(known_species$Confidence) #0.0976

##Genus Confidence
mg <- mean(known_genus$Confidence) #0.9577
sg <- sd(known_genus$Confidence) #0.0772

##Family Confidence
mf <- mean(known_family$Confidence) #0.9605
sf <- sd(known_family$Confidence) #0.0734

##Order Confidence
mo <- mean(known_order$Confidence) #0.9631
so <- sd(known_order$Confidence) #0.0708

##Class confidence
mc <- mean(known_class$Confidence) #0.9623
sc <- sd(known_class$Confidence) #0.0715

##Phylum confidence
mp <- mean(known_phylum$Confidence) #0.9619
sp <- sd(known_phylum$Confidence) #0.0721

##Kingdom confidence
mk <- mean(known_kingdom$Confidence) #0.9615
sk <- sd(known_kingdom$Confidence) #0.0725
```

Read in new Excel table with confidence intervals and means

```
conf_by_tax <- read_excel("../data/excel_data/tax_confidence/confidence.xlsx") #made a new dataframe using mean a
nd sd data calculated above. Reading in this data frame and cleaning the data below

conf_by_tax <- conf_by_tax %>%
  mutate(taxon = factor(taxon))

conf_by_tax$taxon <- factor(conf_by_tax$taxon, levels = c("kingdom", "phylum", "class", "order", "family", "genus
", "species")) #reorder levels - otherwise it will automatically list aplphabetically

conf_by_tax <- cbind(conf_by_tax, otu_prop_vector)


#Visualize ##Plot relative confidence of each taxonomic group classificaiton
pdf("/Users/joegunn/Desktop/Grad_School_Stuff/Research/Projects/Elephant_Microbiome/Attempt_2/tax_confidence.pdf"
, width=7, height=6)

ggplot(conf_by_tax, aes(x = taxon, y = mean)) +
  geom_point(size = 2) +
  geom_errorbar(aes(ymin = mean - sd, ymax = mean + sd), width=.3) +
  theme_set(theme_cowplot(12)) +
  scale_y_continuous(name = "Mean Confidence of Classification", sec.axis = sec_axis(~. /0.97, name = "Proportion
of Classified OTUs")) +
  geom_line(mapping = aes(x = conf_by_tax$taxon, y = conf_by_tax$otu_prop_vector, group = 1), size = 2, color = "
blue", alpha = 0.6) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  labs(x = "Taxonomic Group")

dev.off()
```

```
## quartz_off_screen
##                 2
```