Basic Dataset Stats

Courtney Stepien

August 4, 2016

Contents

File status

Current

File goals

Goals of this file are to get basic statistics on the dataset that we used to search for sequence data. Initial goals include number and names of genera with only 1 representative in the dataset.

This information will help us ID taxa who lack DNA sequence but are monotypic genera in this dataset. These taxa are candidates for using congeneric DNA sequence to represent them at the genus level in our phylogeny.

Additionally:

- histogram of taxa by genus
- histogram of taxa by family
- histogram of taxa by order
- histogram of taxa by class

Class, Order and Family data will be reported using 3 different taxonomy sources: BOLD, NCBI and AlgaeBase.

Next steps

Next steps are:

1. wait on family, order, class data from taxize to calculate histogram by these data
2. Create a table of how many family, orders and classes are represented
3. Determine total number of families, orders and classes in Rhodophyta so I can see our coverage

Analysis

Data setup and packages

```
library(dplyr)
library(ggplot2)
library(grid)
data <- read.csv("../data/mean_13c_per_species.csv")
```

Genus Counts

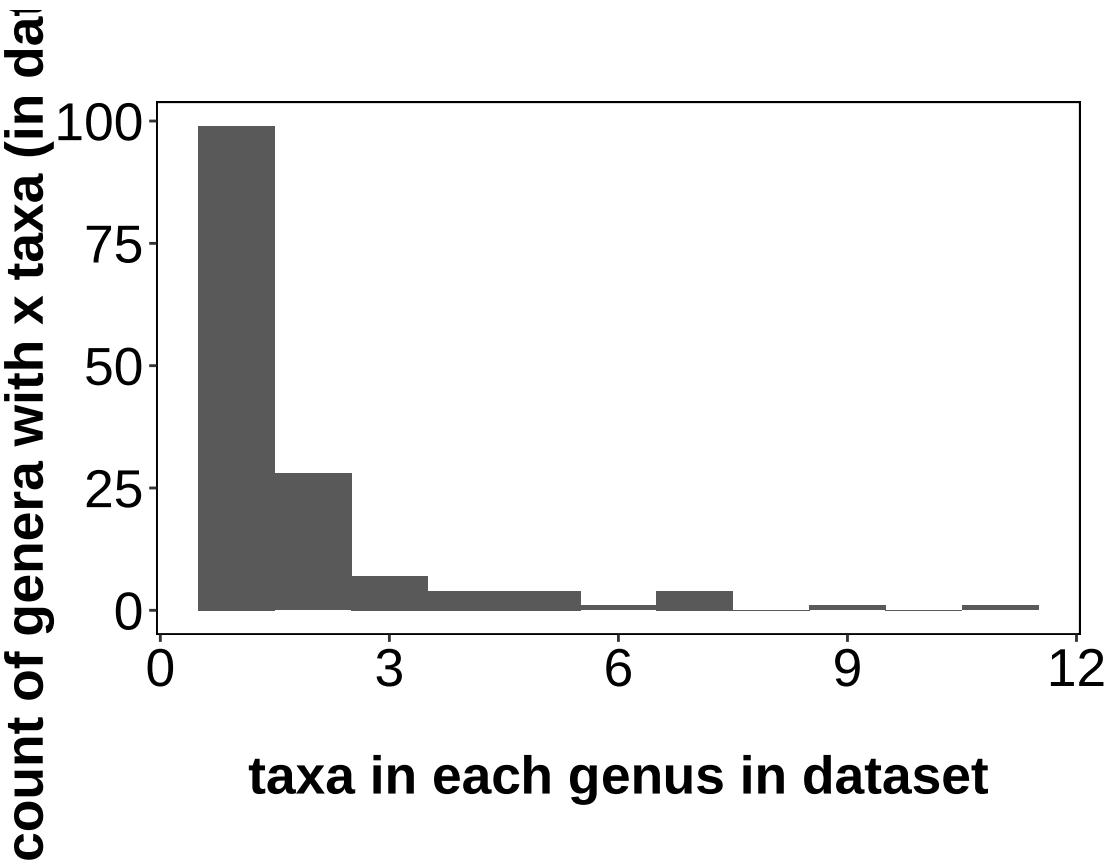
Histogram of taxa count by genus in the dataset

```
genus_count <- data %>% group_by(genus) %>% summarize(n())
genus_count <- data.frame(genus_count)
colnames(genus_count) <- c("genus", "n_taxa")
genus_hist <- genus_count %>% group_by(n_taxa) %>% summarize(n())
genus_hist <- data.frame(genus_hist)
colnames(genus_hist) <- c("n_taxa", "genera_with_n_taxa")
```

There are 149 genera in the dataset, with 1 to 11 taxa in each genus. 99 genera have only 1 species in the dataset, while the remaining 50 genera have 2 to 11 species per genus (see distribution in the table below).

```
##   n_taxa genera_with_n_taxa
## 1      1                 99
## 2      2                 28
## 3      3                  7
## 4      4                  4
## 5      5                  4
## 6      6                  1
## 7      7                  4
## 8      9                  1
## 9     11                  1
```

```
ggplot(data=genus_count, aes(genus_count$n_taxa)) + geom_histogram(bins = max(genus_count$n_taxa), binwi
  xlab("\ntaxa in each genus in dataset") + ylab("count of genera with x taxa (in dataset)") +
  theme(axis.title.y = element_text(size=20, face = "bold"), panel.background = element_blank(),
        plot.title=element_text(size=20), axis.title.x = element_text(size=20, face = "bold"),
        axis.text.x = element_text(size=20, color="black"),
        plot.margin = unit(c(1.2,1.2,1,1),"cm"),
        axis.text.y=element_text(size=20, color="black"), strip.text.x = element_text(size = 20),
        axis.line=element_line(), panel.border = element_rect(fill = NA, color = "black"))
```



List of genera with only 1 representative in the dataset

```
congener_candidates <- filter(genus_count, n_taxa == 1) %>% select(genus)
write.csv(congener_candidates, file = "../data/taxa names for sequence search/congener_candidates.csv",
```

Below is the list of the 99 genera in the dataset with only 1 species:

Acanthophora, Acrosorium, Acrotylus, Actinotrichia, Aglaothamnion, Ahnfeltia, Amansia, Anotrichium, Apophlaea, Audouinella, Ballia, Bonnemaisonia, Bornetia, Botryocladia, Catenella, Champia, Chylocladia, Craspedocarpus, Cumagloia, Cystoclonium, Dichotomaria, Dictyomenia, Digenea, Dumontia, Endocladia, Furcellaria, Georgiella, Gibsmithia, Gloiocladia, Gloiopeltis, Gloiosaccion, Gloiosiphonia, Gracilariopsis, Gymnogongrus, Halicnide, Halopeltis, Halopithys, Halopitys, Haloplegma, Halosaccion, Halurus, Hemineura, Hildenbrandia, Hydrolithon, Hymenocladia, Iridaea, Jeannerettia, Kallymenia, Lemanea, Lenormandia, Liagora, Liagoropsis, Lophocladia, Martensia, Mastophoropsis, Membranoptera, Metamastophora, Microcladia, Mychodea, Neogoniolithon, Neorhodomela, Neuroglossum, Odonthalia, Opuntiella, Osmundaria, Pachymenia, Palisada, Parviphycus, Peyssonnelia, Phacelocarpus, Phymatolithon, Plumaria, Pollexfenia, Polycoelia, Polyides, Polyopes, Portieria, Pterocladiella, Ptilophora, Rhodothamniella, Rytiphlaea, Schizoseris, Schottera, Scinaia, Solieria, Sonderopelta, Sphaerococcus, Spongoclonium, Stenogramma, Symphyocladia, Thamnoclonium, Trematocarpus, Tricleocarpa, Tylotus, Vertebrata, Weeksia, Wrightiella, Yamadaella, Yuzurua