Frequency Distributions

Jeff Oliver

September 20, 2018

Compare the frequency of species' occurence across sites in the two collection methods

```
library("ggplot2") # plotting
library("dplyr")
library("tidyr")   # for long-data formatting
source(file = "bioscan-functions.R")
bioscan <- CompleteBioscan()
```

Want to create a histogram with:

- X-axis: the number of sites
- Y-axis: number of species

Shows the frequency distribution of species, to see if there were lots of singlets/doublets, or if there were many species seen at many sites.

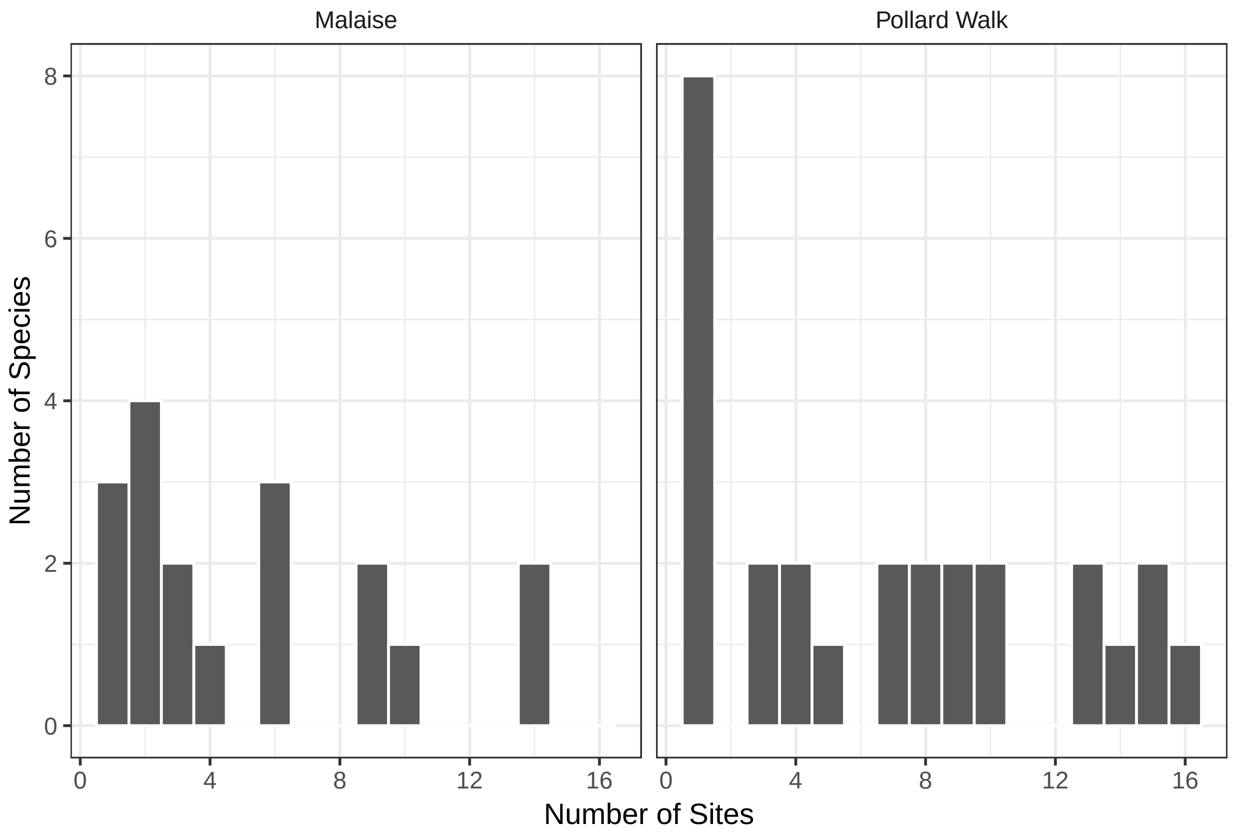
We'll want a data frame like:

Collection Method	Species	# Sites
Malaise	*Agraulis_vanillae*	6
Malaise	*Brephidium_exilis*	1
…	…	…
Pollard Walk	*Vanessa_cardui*	13

```
# Identify those columns with species data
species.cols <- c(5:33)

# Count just whether or not a species occurred for a particular site/method
bioscan.freqs <- as.data.frame(bioscan[, species.cols] > 0)
bioscan.freqs$Collection.Method <- bioscan$Collection.Method

# There has to be a tidyverse way of doing this, but for now, brute force
# Separate the two collection methods
malaise.freqs <- bioscan.freqs[bioscan.freqs$Collection.Method == "Malaise", ]
pollard.freqs <- bioscan.freqs[bioscan.freqs$Collection.Method == "Pollard Walk", ]

# Use boolean math to count the number of sites each species occurred in
malaise.sums <- colSums(malaise.freqs[, -(which(colnames(malaise.freqs) == "Collection.Method"))])
pollard.sums <- colSums(pollard.freqs[, -(which(colnames(pollard.freqs) == "Collection.Method"))])

# Convert to long format and add method as new column
malaise.long <- data.frame(species = names(malaise.sums),
                           num.sites = malaise.sums)
malaise.long$Collection.Method <- "Malaise"
pollard.long <- data.frame(species = names(pollard.sums),
                           num.sites = pollard.sums)
pollard.long$Collection.Method <- "Pollard Walk"

# Combine data back together for plot
bioscan.long <- rbind(malaise.long, pollard.long)
# Drop any zeros, since we're not interested in plotting those
bioscan.long <- bioscan.long[bioscan.long$num.sites > 0, ]
rownames(bioscan.long) <- NULL
```

We now have a table, so we can plot the histogram.

```
freq.plot <- ggplot(data = bioscan.long, mapping = aes(x = num.sites)) +
  geom_histogram(bins = max(bioscan.long$num.sites), color = "white") +
  facet_wrap(~ Collection.Method) +
  theme_bw() +
  xlab(label = "Number of Sites") +
  ylab(label = "Number of Species") +
  scale_x_continuous(breaks = c(0, 4, 8, 12, 16)) +
  theme_bw() +
  theme(strip.background = element_blank())
print(freq.plot)
```



```
ggsave(filename = "output/frequency-distribution.png", plot = freq.plot)
```

```
## Saving 6.5 x 4.5 in image
```