Independent Project Visualization

Anahei Lara

2025-05-05

Aim

I don't have enough data for my aims yet so I will be tweaking it in the meantime. My aim that I will be focusing on is to assess whether my samples are associated with certain species more frequently at different collection sites.

Null

The number of samples is equally distributed across host species and collection sites.

Alternative

Number of samples vary significantly among host species and collection sites.

Visualization

The best statistical approach would be to run a chi squared test of independence since all my variables are categorical and this test will let me know if there is a statistically significant association between the host species and the collection site such as if certain species are more common at certain sites. The best visualization for this data would be a grouped bar plot. The x axis would be the collection site, the y axis is the count of samples and the fill for the bars would be the host species.

Code

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
df <- read.csv("Database for oSp proteins - Master Data.csv")

species_site_table <- table(df$Species, df$Site)

chisq_test <- chisq.test(species_site_table)
```

```
## Warning in chisq.test(species_site_table): Chi-squared approximation may be
## incorrect
```

```r
print(chisq_test)
```

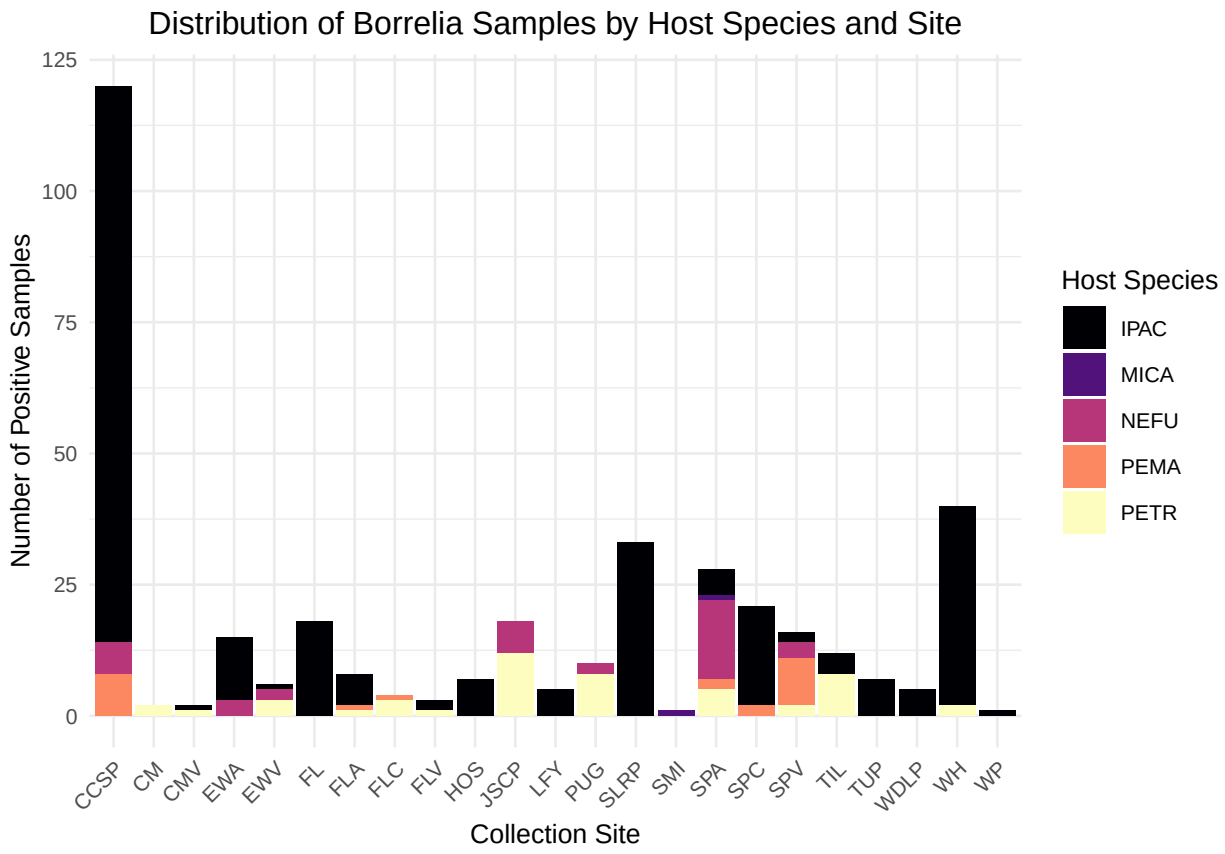
```
##
##  Pearson's Chi-squared test
##
## data:  species_site_table
## X-squared = 609.74, df = 88, p-value < 2.2e-16
```

The Chi-squared test reveals that the samples are not evenly distributed amongst host species at different sites as the p-value is less than 0.005 and we reject the null hypothesis.

```r
plot_df <- df %>%
  group_by(Site, Species) %>%
  summarise(Count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Site'. You can override using the
## `.groups` argument.
```

```r
ggplot(plot_df, aes(x = Site, y = Count, fill = Species)) +
  geom_bar(stat = "identity") +
  labs(title = "Distribution of Borrelia Samples by Host Species and Site",x = "Collection Site",y = "Nu
  scale_fill_viridis_d(option = "magma", name = "Host Species") +
  theme_minimal(base_size = 10) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),plot.title = element_text(hjust = 0.5))
```



Visualization with grouped sites

```
fl_sites <- c("FL", "FLA", "FLV", "FLC")
cm_sites <- c("CM", "CMV")
ew_sites <- c("EWA", "EWV")
sp_sites <- c("SPA", "SPC", "SPV")

df <- df %>%
  mutate(Site_grouped = ifelse(Site %in% fl_sites, "FL",
                        ifelse(Site %in% cm_sites, "CM",
                        ifelse(Site %in% ew_sites, "EW",
                        ifelse(Site %in% sp_sites, "SP", Site)))))
```

```
plot_df <- df %>%
  group_by(Site_grouped, Species) %>%
  summarise(Count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Site_grouped'. You can override using the
## `.groups` argument.
```

```
ggplot(plot_df, aes(x = Site_grouped, y = Count, fill = Species)) +
  geom_bar(stat = "identity") +
  labs(title = "Borrelia Samples by Host Species and Grouped Site",x = "Grouped Collection Site",y = "Nu
  scale_fill_viridis_d(option = "magma", name = "Host Species") +
```

```
  theme_minimal(base_size = 10) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),plot.title = element_text(hjust = 0.5))
```