Independent Project

Anahei Lara

2025-03-10

Introduction

Borrelia burgdorferi is a spirochete that causes Lyme disease, and it is the most prevalent vector-borne pathogen in the United States (Eisen *et al.* 2017). While the prevalence of the pathogen is widespread, the ecological and environmental factors influencing its dynamics differ significantly between the eastern and western regions of the United States. In the eastern United States, particularly in the primary vector is the black-legged tick, *Ixodes scapularis* and the white-tailed deer, *Odocoileus virginianus* is its main host (Gern *et al.* 2000). This region has long been recognized as a Lyme disease hotspot and has been studied extensively. Conversely, in the western United States, the western black-legged tick, *Ixodes pacificus* is the primary vector for *Borrelia burgdorferi* but the pathogen can also be found in two other *Ixodes* vectors: *Ixodes angustus* and *Ixodes spinipalpis* (Xu *et al.* 2019).*Ixodes pacificus* hosts are varied and consist of gray squirrels (*Sciurus griseus*), deer mice (*Peromyscus maniculatus*), and the Western fence lizard (*Sceloporus occidentalis*) among many other species (Castro and Wright 2007, Furman and Loomis 1984, McVicar *et al.* 2012). However, more research is needed on the western United States to determine *B. burgdorferi's* genetic diversity which can illustrate the mechanism of how it's transmitted between vector and host.

There are at least 35 documented outer surface proteins with known roles (Pulzova and Bhide 2014). By observing four outer surface proteins, OspA, ospB, ospC, and BBA64 (p35), we can determine how variations in these proteins effectively transmit the spirochete. Outer surface protein A and ospB are lipoproteins that have the ability to persist and settle in the tick's midgut which is essential for the future transmission of the *Borrelia* spirochete (Caine *et al.* 2016). Additionally, ospC and p35 help the bacteria infect mammalian hosts after being transmitted through a tick bite that has been attached for at least 24 hours (Kenedy *et al.* 2012). In a tick that has not fed, ospA and ospB protein levels are elevated in the midgut while ospC and p35 levels are not present (Kenedy *et al.* 2012). In ticks that have fed on a host, ospC and p35 have elevated levels while ospA and ospB have low levels of expression (Tokarz *et al.* 2004). My research focuses on investigating these proteins genetic diversity to test for potential correlations in various tick and host species collected from the same geographic region in North Coastal California.

Research Question

How does the genetic diversity of *Borrelia burgdorferi* outer surface proteins (OspA, OspB, OspC, and p35) correlate with the diversity of tick and host species, and what implications does this have for pathogen transmission and persistence?

Aims

1. Characterize the genetic diversity of OspA and OspB in *Borrelia burgdorferi* across various *Ixodes* tick species to determine their role in tick-borne persistence.

2. Analyze the genetic variation of OspC and p35 in *B. burgdorferi* collected from different host species to assess their influence on host immune evasion and transmission dynamics.

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 species and collection sites.

Alternative

Number of samples vary significantly among 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 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 species. https://github.com/MeenaAnahei/Prevalence-of-Borrelia/tree/main

Dataset Description

The Borrelia burgdorferi outer surface protein dataset is a molecular ecology dataset that consists of DNA sequences and other metadata for four outer surface proteins A,B,C, and p35 (BBA64).This data is derived from Ixodes pacificus ticks and their vertebrate hosts collected in the Bay Area of California.This dataset helps support research on how the genetic diversity of outer surface proteins correlates with vector and host diversity which provides insights into the transmission dynamics of Borrelia in the western United States. https://github.com/MeenaAnahei/Prevalence-of-Borrelia/tree/main

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
```

```
df <- read.csv("osp_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
```

```
print(chisq_test)
```

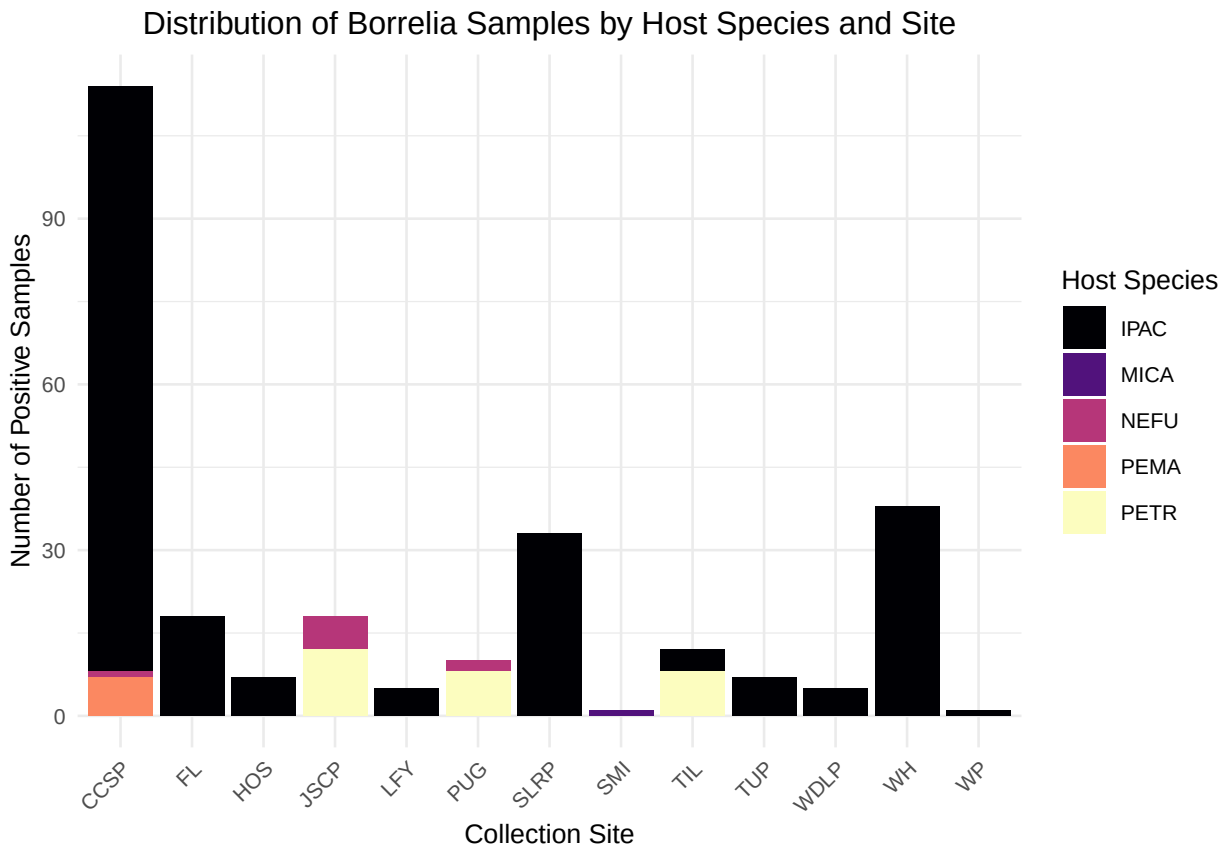
```
##
##  Pearson's Chi-squared test
##
## data:  species_site_table
## X-squared = 534.96, df = 48, p-value < 2.2e-16
```

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

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

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

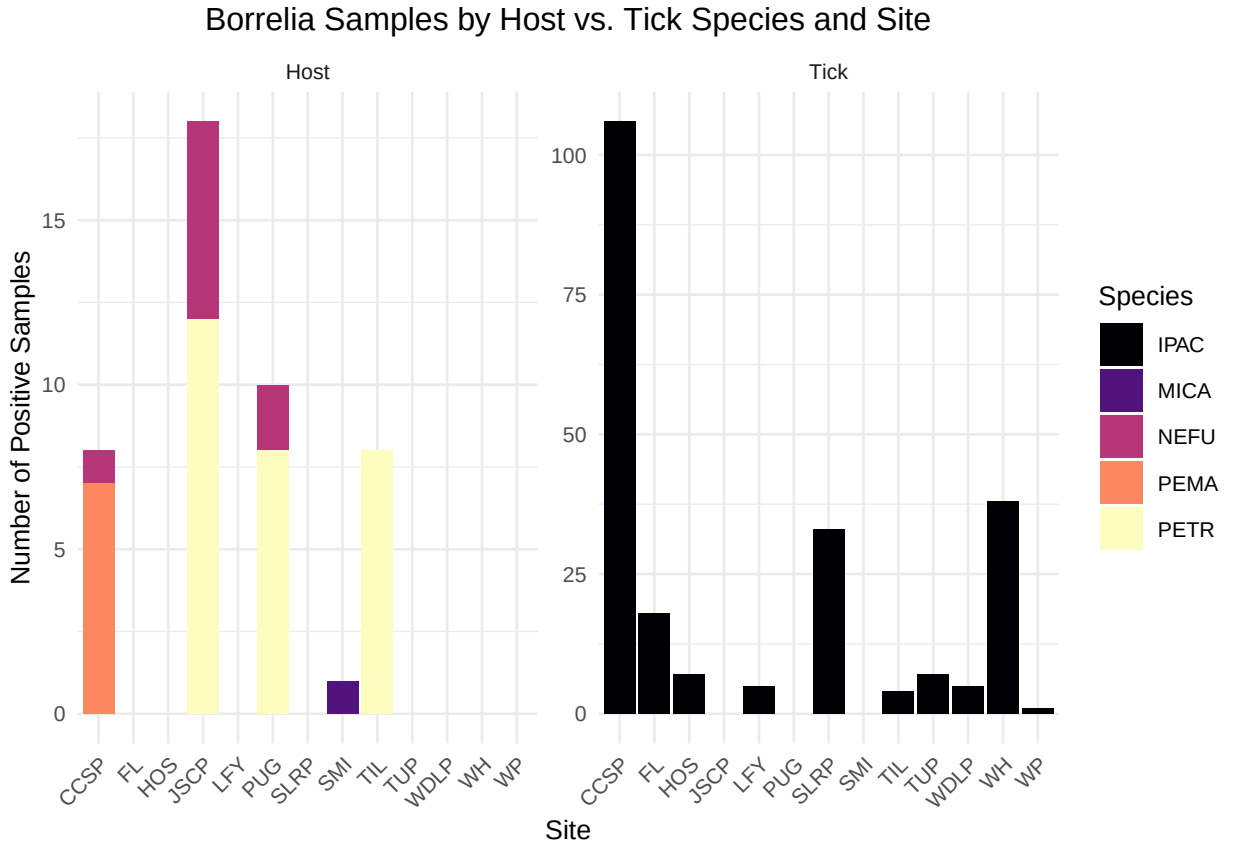
```
ggplot(plot_df, aes(x = Site, y = Count, fill = Species)) +
  geom_bar(stat = "identity") +
  labs(title = "Distribution of Borrelia Samples by Host Species and Site",x = "Collection Site",y = "Nu
  scale_fill_viridis_d(option = "magma", name = "Host Species") +
  theme_minimal(base_size = 10) +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




```r
species_df <- df %>%
  mutate(Species_Type = ifelse(Species == "IPAC", "Tick", "Host"))

plot_df <- species_df %>%
  mutate(Site_grouped = Site) %>%
  group_by(Site_grouped, Species, Species_Type) %>%
  summarise(Count = n(), .groups = "drop")

ggplot(plot_df, aes(x = Site_grouped, y = Count, fill = Species)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Species_Type, scales = "free_y") +
  labs(
    title = "Borrelia Samples by Host vs. Tick Species and Site",
    x = "Site",
    y = "Number of Positive Samples"
  ) +
  scale_fill_viridis_d(option = "magma", name = "Species") +
  theme_minimal(base_size = 10) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5))
```



Ran a Poisson glmm for fun to test if the number of positive samples vary significantly across both collection site and species type. Found that species and site significantly influence the number of positive detections. Most species and sites like SMI,WP,LFY had few positive samples comapared to the reference group. From this we get that infection risk is uneven across the sites and certain species are less likely to carry the Borrelia samples compared to the reference.

```
library(glmmTMB)
# Summarize counts of positive samples
host_plot_df <- species_df %>%
  group_by(Site, Species) %>%
  summarise(Count = n()) %>%
  ungroup()
```

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

```r
# Run Poisson regression
host_glmm <- glmmTMB(Count ~ Site + Species, family = poisson(), data = host_plot_df)
```

```
## dropping columns from rank-deficient conditional model: SpeciesMICA
```

```r
summary(host_glmm)
```

```
##  Family: poisson  ( log )
## Formula:          Count ~ Site + Species
## Data: host_plot_df
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##     119.6     133.9     -43.8      87.6         2
##
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.61919    0.09878   46.76  < 2e-16 ***
## SiteFL      -1.72881    0.25556   -6.76 1.34e-11 ***
## SiteHOS     -2.67328    0.39066   -6.84 7.75e-12 ***
## SiteJSCP    -0.93543    0.50284   -1.86  0.06284 .
## SiteLFY     -3.00975    0.45799   -6.57 4.98e-11 ***
## SitePUG     -1.52322    0.54524   -2.79  0.00521 **
## SiteSLRP    -1.12268    0.20015   -5.61 2.03e-08 ***
## SiteSMI     -4.61918    1.00487   -4.60 4.29e-06 ***
## SiteTIL     -2.46875    0.33043   -7.47 7.94e-14 ***
## SiteTUP     -2.67328    0.39066   -6.84 7.75e-12 ***
## SiteWDLP    -3.00975    0.45799   -6.57 4.98e-11 ***
## SiteWH      -0.98160    0.18993   -5.17 2.36e-07 ***
## SiteWP      -4.61918    1.00487   -4.60 4.29e-06 ***
## SpeciesMICA       NA         NA      NA       NA
## SpeciesNEFU -2.89846    0.38802   -7.47 8.03e-14 ***
## SpeciesPEMA -2.67328    0.39066   -6.84 7.75e-12 ***
## SpeciesPETR -0.92331    0.47918   -1.93  0.05400 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```