

# BIOL 710- Final Project

Gillyan Valencia

## Introduction

Thermophilic microorganisms, capable of living at extreme temperatures in habitats like geothermal springs, offer unique insights into the evolution of microbial life and life's potential in harsh environments. (Damer and Deamer, 2020). While microbial evolution has been studied across different environments (Nguyen et al., 2021), the evolution of population genomics and community-level dynamics within geothermal springs remain poorly understood. The Great Boiling Spring (GBS) in Gerlach, Nevada, provides an ideal system to explore these dynamics, as its microbial communities are structured by temperature gradients and micro-environments such as sediment and the water column (Cole et al., 2013). *Thermoflexus hugenholtzii*, a dominant thermophile across sites in the Great Boiling Spring (GBS) system in Gerlach, Nevada, provides an ideal location to investigate fine-scale genetic variation across environmental gradients. This study uses metagenomic sequencing and population genomics tools to examine the microevolutionary dynamics of *T. hugenholtzii*, focusing on single codon variations (SCVs) within genes critical to core cellular functions.

Given that temperature is a key factor shaping microbial communities in GBS (Cole et al., 2013; Kees et al., 2022), this study aims to examine how microbial community composition and population genomics evolve in response to seasonal temperature variations. A high-quality metagenome-assembled genome (MAG) generated from the most complete, least contaminated sample (June Site B) served as the reference genome. SCV frequencies across five sample sites were analyzed in relation to site-specific and seasonal temperature fluctuations. This approach allows for the exploration of how environmental variables influence genetic variation within microbial populations, contributing to a broader understanding of microbial genome plasticity and ecological adaptation in extreme ecosystems.

## Research Question

How do temperature and site-specific environmental conditions influence single codon variation (SCV) in the genes *rpoB*, *gyrB*, and *recA* of *Thermoflexus hugenholtzii* across spatial and seasonal gradients in Great Boiling Spring?

## Specific Aims

**Aim 1:** To quantify and compare SCV frequencies in three conserved genes of *T. hugenholtzii* across five metagenomic samples and assess their correlation with temperature and sampling site.

## Dataset Description

This dataset originates from sediment samples collected from Great Boiling Spring in Gerlach, Nevada, with DNA extracted and sequenced using Illumina short-read technology. Metagenomic data were assembled using SPAdes and binned into MAGs with MetaBat2, followed by quality filtering and taxonomic classification. A high-quality *Thermoflexus hugenholtzii* MAG was selected for functional and evolutionary analysis using COG-based annotations in Anvi'o. Single-codon variant (SCV) frequencies were calculated per gene to assess codon-level variation across sites.

The dataset and all scripts can be found at the following link: <https://github.com/gillyv8/GBSanaly.git>

## Statistical Approach

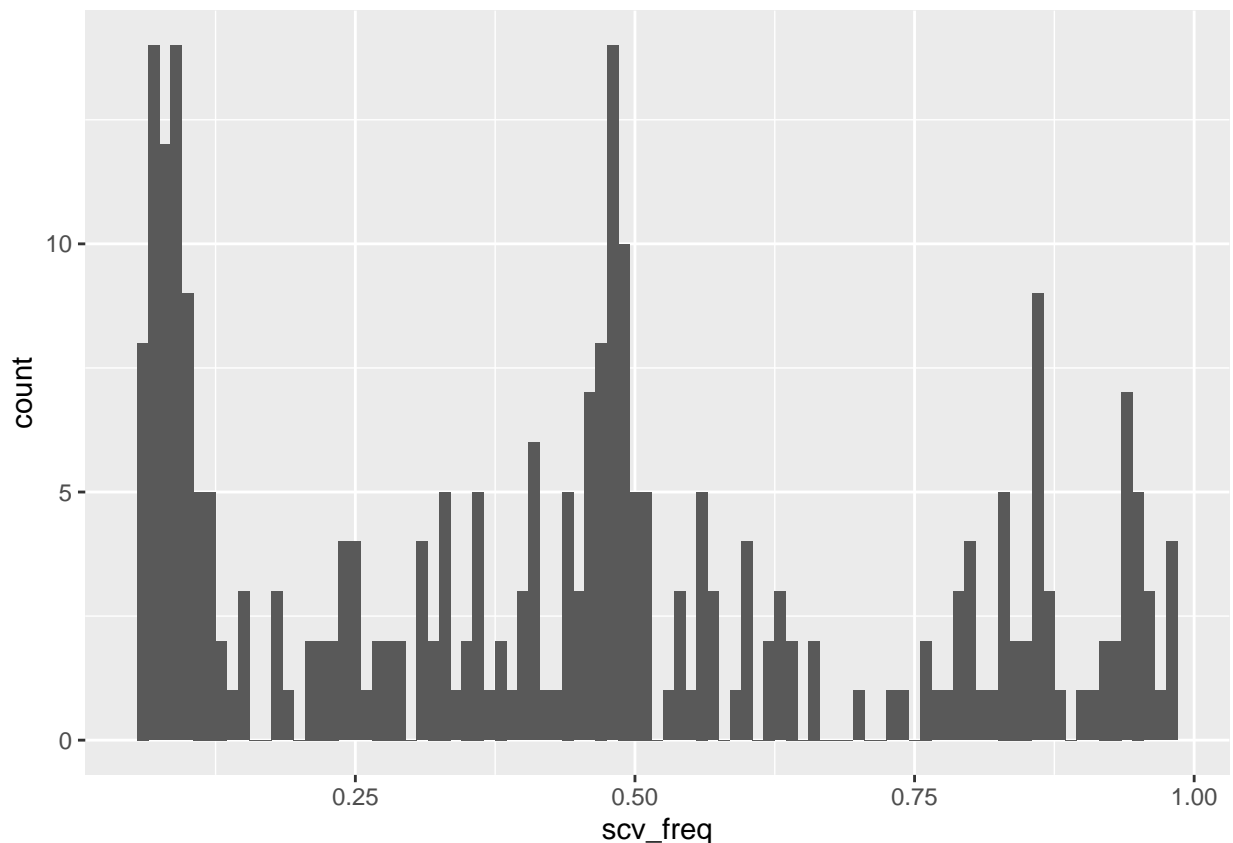
I first checked the distribution of the data for normality to figure out which statistical method was best. All three genes showed non-normality on their distributions. I used a generalized linear mixed model (GLMM) with a beta distribution to analyze single codon variants (SCVs) frequency, where the frequencies all fell between 0 and 1. The model included fixed effects for Site and Month and their interactions on SCV frequencies within the rpoB, recA, and gyrB genes. This allowed for me to assess both main effects and gene-specific environmental interactions. The model was fit using the glmmTMB package in R.

```
library(ggplot2)
library(tidyverse)
library(glmmTMB)
```

```
#SCV dataset
scv <- read.csv("Thermo_Hugo_SCV.csv")
#str(scv)
```

```
# Checking for normality

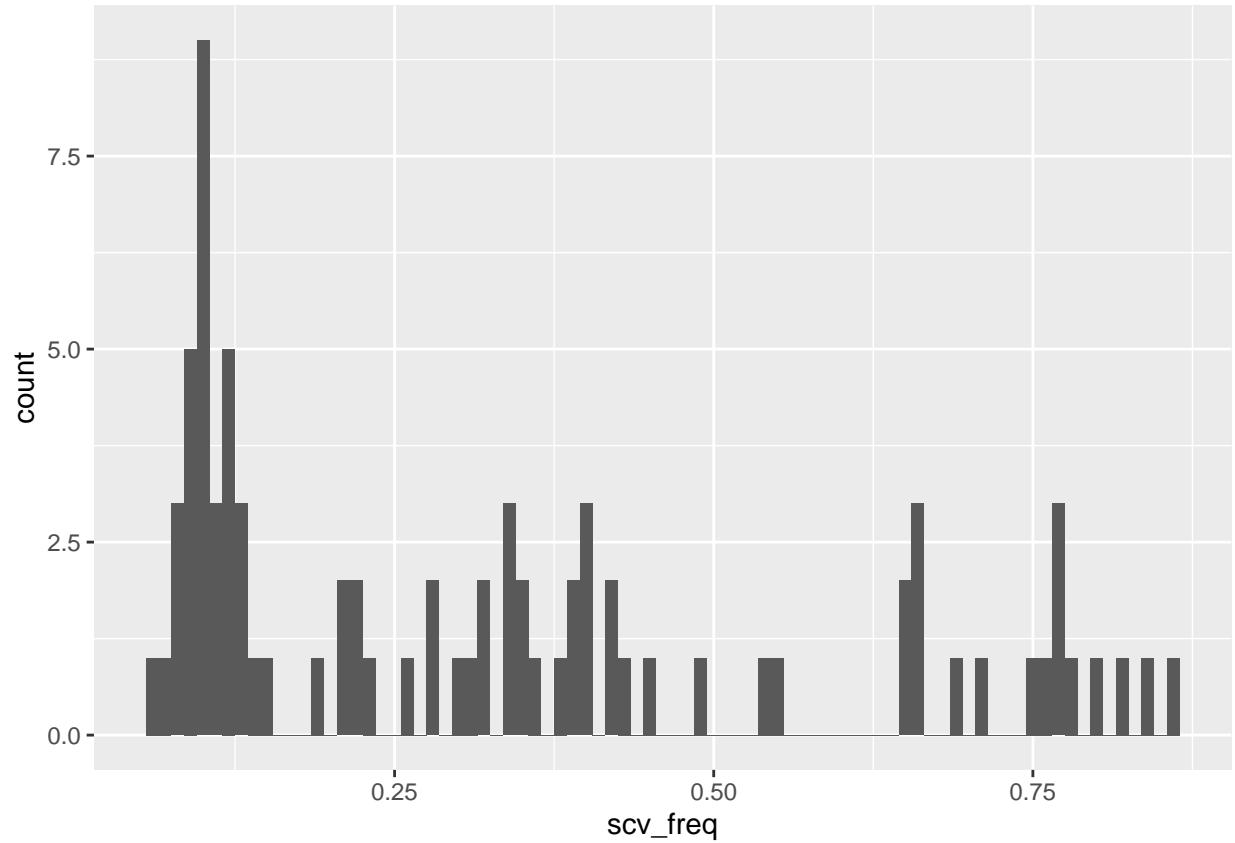
rpoB_scv_data <- subset(scv, Gene_ID == "rpoB")
rpoB_scv_hist <- ggplot(rpoB_scv_data, aes(x = scv_freq)) +
  geom_histogram(binwidth = 0.01)
rpoB_scv_hist
```



```

recA_scv_data <- subset(scv, Gene_ID == "recA")
recA_scv_hist <- ggplot(recA_scv_data, aes(x = scv_freq)) +
  geom_histogram(binwidth = 0.01)
recA_scv_hist

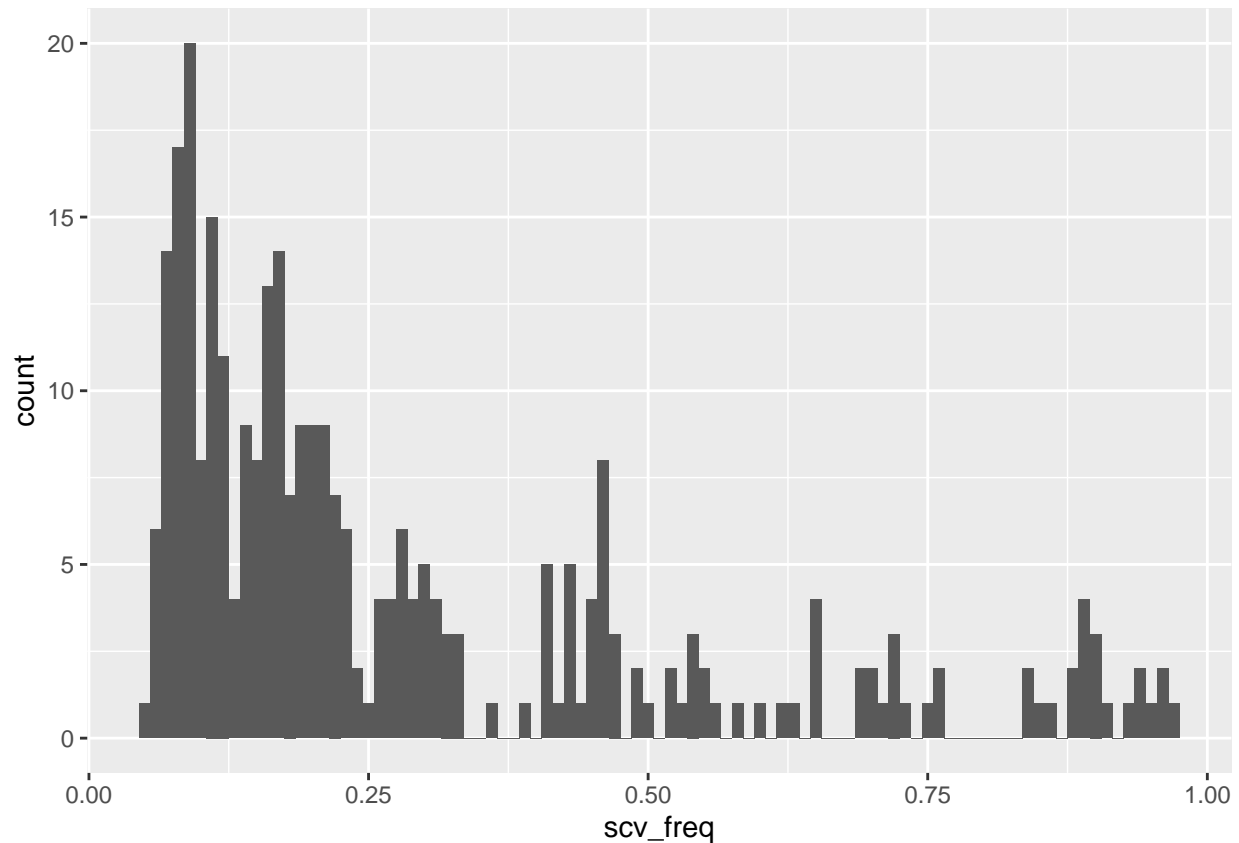
```



```

gyrB_scv_data <- subset(scv, Gene_ID == "gyrB")
gyrB_scv_hist <- ggplot(gyrB_scv_data, aes(x = scv_freq)) +
  geom_histogram(binwidth = 0.01)
gyrB_scv_hist

```



```
# All non-normal
```

## Hypotheses

**Null hypothesis (H0):** Site and Month have no effect on SCV frequency in any of the genes tested.

**Alternative hypothesis (HA):** SCV frequency varies significantly by Site and/or Month for at least one gene.

```
# SCV

# SCV Frequency GLM
SCV_GLMM <- glmmTMB(
  scv_freq ~ Gene_ID * Site + Gene_ID * Month,
  data = scv,
  family = beta_family()
)

# Add Model Predictions
scv$predicted <- predict(SCV_GLMM, type = "response")

# 3. Clean data
scv_clean <- scv %>%
  filter(!is.na(Month)) %>%
  mutate(
```

```

    Month = factor(trimws(as.character(Month))), # Clean text and convert to factor
    SiteMonth = interaction(Site, Month, sep = "_") # Create grouping variable
  )

# Order of x-axis
scv_clean$SiteMonth <- factor(scv_clean$SiteMonth,
  levels = c("A_February", "B_February", "A_June", "B_June", "C_June")
)

# Plot
SCV_Plot <-ggplot(scv_clean, aes(x = SiteMonth, y = scv_freq)) +
  geom_jitter(aes(color = Site), width = 0.2, alpha = 0.6) +
  stat_summary(aes(y = predicted), fun = mean, geom = "point",
    shape = 18, size = 3, color = "black") +
  facet_wrap(~ Gene_ID) +
  scale_color_viridis_d(option = "C") +
  labs(
    title = "SCV Frequency Variance in Thermoflexus hugenholtzii",
    x = "Site and Month",
    y = "SCV Frequency",
    color = "Site"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "right")
# Summary Statistics
SCV_Plot

```

```
summary(SCV_GLMM)
```

```

## Family: beta ( logit )
## Formula:          scv_freq ~ Gene_ID * Site + Gene_ID * Month
## Data: scv
##
##      AIC      BIC    logLik -2*log(L)  df.resid
##   -343.2   -284.7    184.6   -369.2      653
##
## Dispersion parameter for beta family (): 3.68
##
## Conditional model:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.72817    0.16719  -4.355 1.33e-05 ***
## Gene_IDrecA     0.09674    0.44582   0.217 0.828217
## Gene_IDrpoB     0.20359    0.25466   0.799 0.424039
## SiteB          -0.29012    0.13985  -2.074 0.038038 *
## SiteC           0.31696    0.13668   2.319 0.020394 *
## MonthJune      -0.06668    0.15327  -0.435 0.663545
## Gene_IDrecA:SiteB -0.50180    0.34921  -1.437 0.150732
## Gene_IDrpoB:SiteB -0.56341    0.23096  -2.439 0.014712 *
## Gene_IDrecA:SiteC  0.41433    0.28700   1.444 0.148836
## Gene_IDrpoB:SiteC  0.62460    0.18619   3.355 0.000794 ***

```

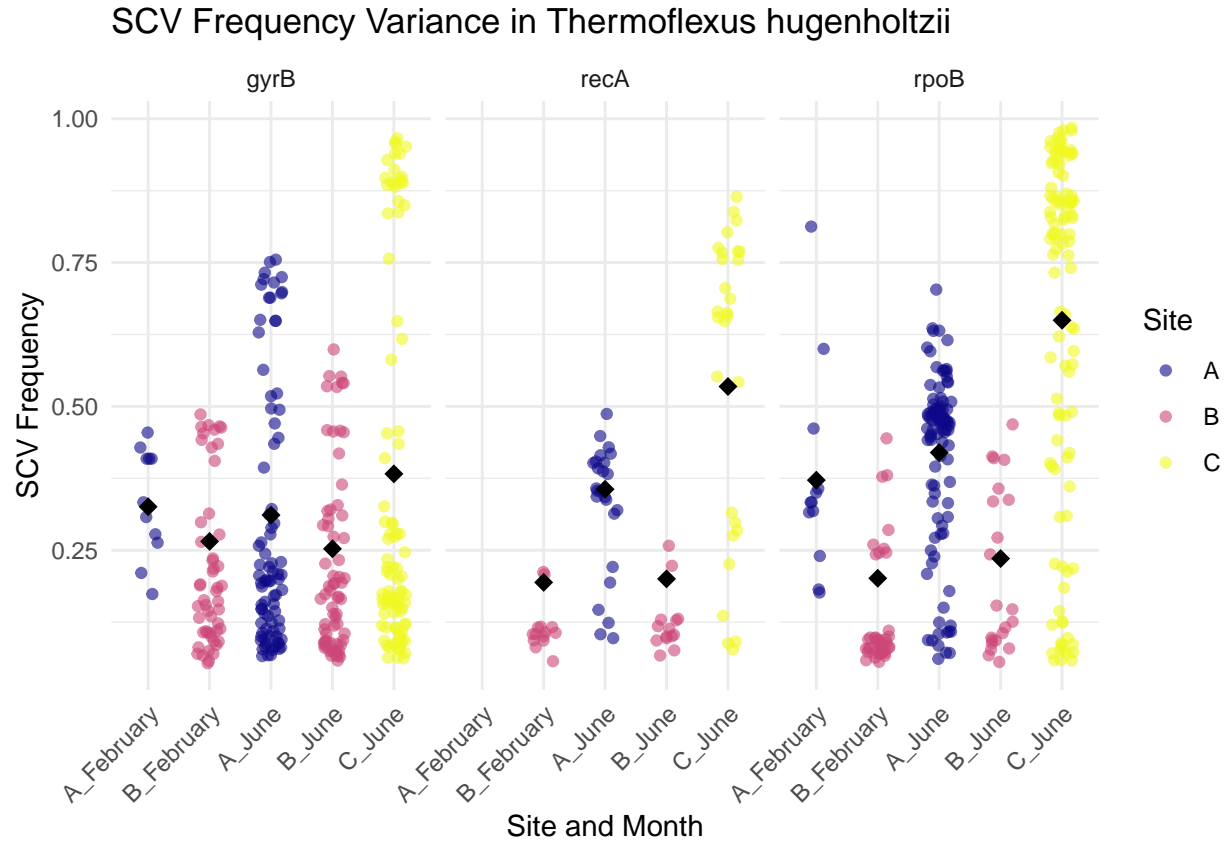


Figure 1: Observed and predicted SCV frequencies by site and sampling month across three genes (*gyrB*, *recA*, and *rpoB*) in *Thermoflexus hugenholtzii*. Points represent individual observations. Black diamonds show model-predicted means from a GLMM with a beta distribution. Data are grouped by site and month and faceted by gene.

```
## Gene_IDrecA:MonthJune 0.10537 0.40029 0.263 0.792368
## Gene_IDrpoB:MonthJune 0.26766 0.24530 1.091 0.275215
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Results

SCV frequency in *Thermoflexus hugenholtzii* varied significantly by gene and sampling site, with limited influence from sampling month. For the reference gene *gyrB* in February, SCV frequency was significantly lower at Site B compared to Site A ( $p = 0.033$ ) and significantly higher at Site C ( $p = 0.017$ ). The gene *rpoB* exhibited strong site-dependent patterns, showing a further significant decrease in SCV frequency at Site B ( $p = 0.012$ ) and a significant increase at Site C ( $p < 0.001$ ), relative to *gyrB*. In contrast, *recA* did not display significant variation across sites or months. No significant effects were observed for sampling month or any gene-by-month interactions ( $p > 0.26$ ). These results indicate that SCV frequency is strongly influenced by site and gene identity, with *rpoB* being particularly sensitive to spatial context.

The observed variation in SCV frequency suggests that codon variation from mutations in *Thermoflexus hugenholtzii* are both gene-specific and influenced by local environmental conditions. The strong site-dependent differences in SCV frequency for *rpoB*, a core gene involved in transcription, imply that certain genomic regions may be more susceptible to mutation or subject to different selective pressures depending on the environment. Higher SCV frequencies at Site C, particularly in *rpoB* and *gyrB*, may indicate relaxed selection or ecological conditions that promote microdiversity. In contrast, the consistently lower SCV frequencies at Site B could indicate stronger purifying selection or environmental constraints that favor genomic stability. The lack of significant effects associated with sampling month suggests that these patterns are relatively stable over time and driven primarily by spatial, rather than temporal, factors. Together, these results point to localized evolutionary pressures shaping intra-species variation and highlight the potential for certain genes to act as indicators of environmental selection within microbial populations.

## References

- Cole, J. K., Peacock, J. P., Dodsworth, J. A., Williams, A. J., Thompson, D. B., Dong, H., Wu, G., & Hedlund, B. P. (2013). Sediment microbial communities in Great Boiling Spring are controlled by temperature and distinct from water communities. *The ISME Journal*, 7(4), 718–729. <https://doi.org/10.1038/ismej.2012.157>
- Damer, B., & Deamer, D. (2020). The Hot Spring Hypothesis for an Origin of Life. *Astrobiology*, 20(4), 429–452. <https://doi.org/10.1089/ast.2019.2045>
- Kees, E. D., Murugapiran, S. K., Bennett, A. C., & Hamilton, T. L. (2022). Distribution and Genomic Variation of Thermophilic Cyanobacteria in Diverse Microbial Mats at the Upper Temperature Limits of Photosynthesis. *mSystems*, 7(5), e00317-22. <https://doi.org/10.1128/msystems.00317-22>
- Nguyen, J., Lara-Gutiérrez, J., & Stocker, R. (2021). Environmental fluctuations and their effects on microbial communities, populations and individuals. *FEMS Microbiology Reviews*, 45(4), fuaa068. <https://doi.org/10.1093/femsre/fuaa068>