bonus_question_practical_1

Mahan Ghafari

2025-01-09

Hard vs. Soft sweeps: Understanding genetic diversity of SARS-CoV-2 using Shannon Entropy and Simpson's Diversity Index

Introduction

Adaptation has traditionally been thought of as a slow and gradual process, driven by the sequential rise and fixation of beneficial mutations over time. This paradigm often assumes that adaptive mutations arise one at a time, with population adaptation strongly dependent on the waiting time for such mutations to occur. However, many examples from nature suggest that adaptation can also be **rapid** and **repeatable**, where multiple adaptive mutations sweep through a population almost simultaneously.

The distinction between **hard sweeps** and **soft sweeps** was systematically investigated by Hermisson and Pennings in their series of works on evolutionary adaptation (https://www.nature.com/articles/s41576-023-00585-x).

While the concept of hard and soft sweeps has primarily been studied for the emergence drug resistance evolution in HIV (https://academic.oup.com/g3journal/article/10/4/1213/6026186), it has broader implications for understanding virus evolution, including SARS-CoV-2. During the SARS-CoV-2 pandemic, we saw examples of both **hard sweep** where a single variant rapidly replaced all others, leading to near-zero genetic diversity, and **soft sweep** where multiple variants co-circulated during the transition, preserving genetic diversity.

In this bonus question, we want to see if we can identify episodes of hard vs soft sweep using two measurements of diversity: **Shannon Entropy** and **Simpson's Diversity Index** and to explore these concepts in the context of SARS-CoV-2 evolution.

Data and setup

We will use a dataset containing daily frequencies of SARS-CoV-2 variants, collected as part of the ONS COVID Infection Survey (https://royalsocietypublishing.org/doi/10.1098/rspb.2023.1284). Let's load the data first.

```
# Import ONS-CIS daily genomic sequence data
url <- "https://raw.githubusercontent.com/mg878/variant_fitness_practical/main/lineage_data.csv"
lineage_data <- read.csv(url)
# Ensure collection_date is in Date format
lineage_data$collection_date <- as.Date(lineage_data$collection_date)
# Check the input format and display the first few rows of the data
head(lineage_data)
```

```
##   collection_date pangolin_call major_lineage
## 1      2020-10-15    B.1.177.54         Other
## 2      2020-10-17       B.1.177         Other
## 3      2020-10-17     B.1.1.279         Other
## 4      2020-10-16          AD.2         Other
## 5      2020-10-17       B.1.177         Other
## 6      2020-10-18    B.1.177.81         Other
```

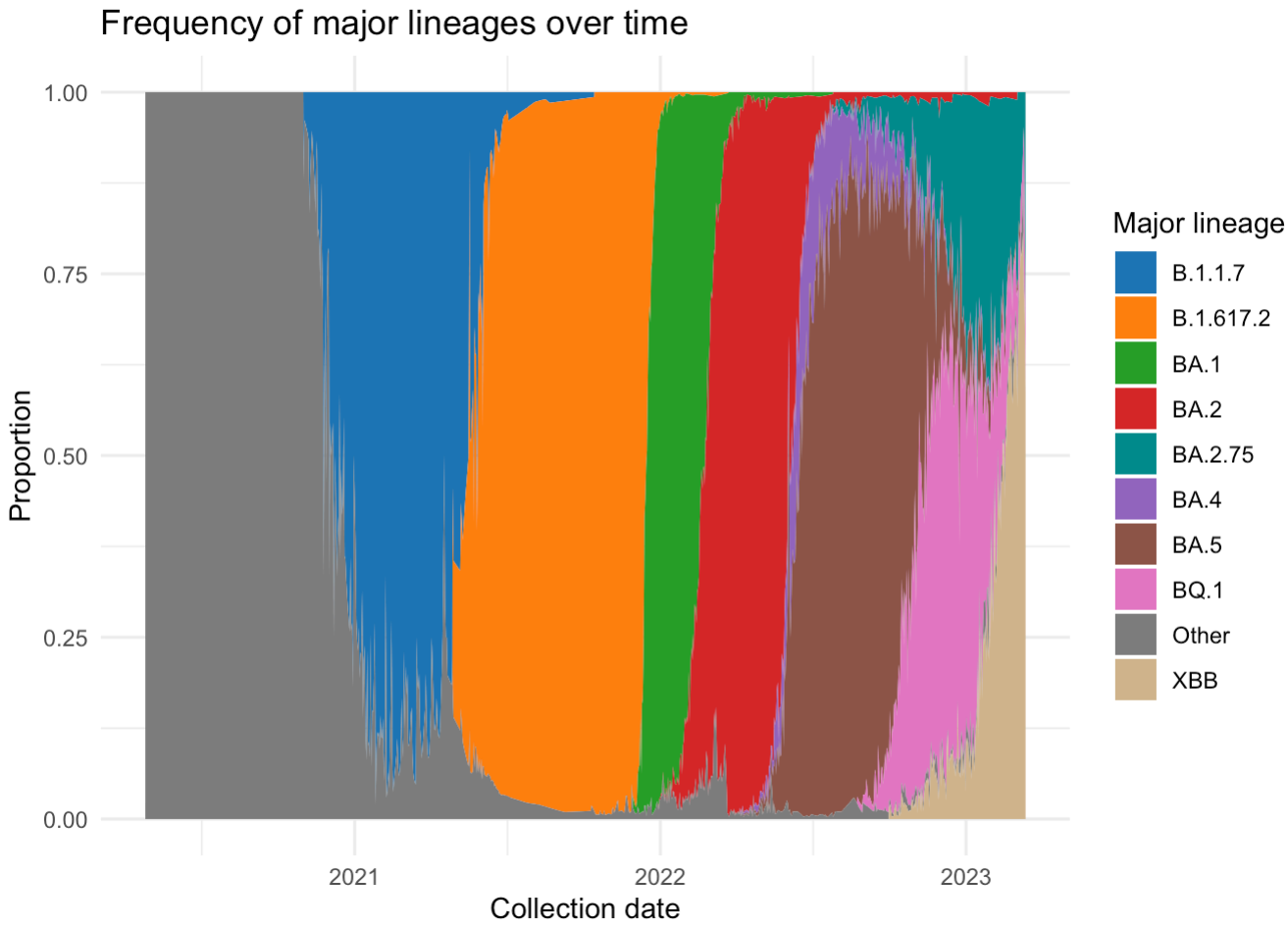
To simplify the long list of lineage names assigned by the Pango nomenclature, we group them into broader ‘major lineages’. For this part of the practical, we focus on variants that caused significant waves in the UK since late 2020. These include: Alpha (B.1.1.7), Delta (B.1.617.2), and various Omicron sublineages, including BA.1, BA.2, BA.4, BA.2.75, BA.5, BQ.1, and XBB. We put variants from other lineages into the ‘Other’ category.

```
lineage_summary <- aggregate(
  lineage_data$major_lineage,
  by = list(collection_date = lineage_data$collection_date, major_lineage = lineage_d
ata$major_lineage),
  FUN = length
)

# Rename columns for clarity
colnames(lineage_summary) <- c("collection_date", "major_lineage", "lineage_count")

# Calculate total counts per date
total_counts <- aggregate(lineage_summary$lineage_count, by = list(collection_date =
lineage_summary$collection_date), FUN = sum)
colnames(total_counts) <- c("collection_date", "total_count")

# Merge total counts back into the lineage summary
lineage_summary <- merge(lineage_summary, total_counts, by = "collection_date")

# Calculate frequencies
lineage_summary$lineage_frequency <- lineage_summary$lineage_count / lineage_summary
$total_count

# Display the first few rows of the new data frame
head(lineage_summary)
```

```
##   collection_date major_lineage lineage_count total_count lineage_frequency
## 1      2020-04-26         Other             1           1                 1
## 2      2020-04-27         Other             1           1                 1
## 3      2020-04-28         Other             2           2                 1
## 4      2020-04-29         Other             1           1                 1
## 5      2020-04-30         Other             1           1                 1
## 6      2020-05-01         Other             1           1                 1
```

Now, let’s look at the changes in the dynamics of major lineages of SARS-CoV-2 over time in the UK.

```
library(ggplot2)
# Define a custom color palette for the major SARS-CoV-2 lineages
custom_palette <- c(
  "B.1.1.7" = "#1f77b4", # Blue
  "B.1.617.2" = "#ff7f0e", # Orange
  "BA.1" = "#2ca02c",  # Green
  "BA.2" = "#d62728",  # Red
  "BA.2.75" = "#008B8B",  # Teal
  "BA.4" = "#9467bd",  # Purple
  "BA.5" = "#8c564b",  # Brown
  "BQ.1" = "#e377c2",  # Pink
  "XBB" = "#D2B48C",  # Light brown
  "Other" = "#7f7f7f"  # Gray
)

# Plot the daily frequencies
ggplot(lineage_summary, aes(x = collection_date, y = lineage_frequency, fill = major_
lineage)) +
    geom_area(position = "fill") +  # Create a stacked area plot
    scale_fill_manual(values = custom_palette) +  # Apply the custom palette
    labs(
      title = "Frequency of major lineages over time",
      x = "Collection date",
      y = "Proportion",
      fill = "Major lineage"
    ) +
    theme_minimal()
```



Two very common ways of measuring genetic diversity of a population are **Shannon Enotropy** and **Simpson's Diversity Index**.

Shannon Entropy

Shannon Entropy (H) is a widely used measure of diversity that accounts for both the number of variants and their relative frequencies. It is defined as:

$$H = -\sum_{i=1}^{n} p_i \log(p_i)$$

where:

- n is the total number of samples per major lineage.
- p_i is the frequency of the i^{th} major lineage.

Higher values of H indicate greater diversity, while $H = 0$ occurs when only a single lineage is present.

Simpson's Diversity Index

Simpson's Diversity Index (D) measures the probability that two randomly chosen individuals belong to *different* lineages. It is defined as:

$$D = 1 - \sum_{i=1}^{n} p_i^2$$

where:

- p_i is the frequency of the i^{th} lineage.

Higher values of D indicate greater diversity, with $D = 0$ when one lineage dominates completely.

Note that you might see alternative definitions of Simpson's Diversity Index in the literature, where D is defined as the probability that two randomly chosen individuals belong to the *same* lineage, for example.

Question

Using the formulas for Shannon Entropy and Simpson's Diversity Index:

1. **Calculate diversity metrics**:
    - Write code to calculate H and D for each collection date in the ONS-CIS dataset.
2. **Visualise diversity over time**:
    - Plot both H and D over time.
    - Use the plots to identify episodes of hard sweeps (where diversity is very small and close to 0) and soft sweeps (where diversity remains elevated). Can you identify the major lineages that result in hard vs soft sweeps in the UK?

Answer: Hard sweeps are observed when diversity is very small and close to 0, indicating that a single variant rapidly dominates the population. In the UK, this occurred during the emergence and establishment of major lineages like Alpha, Delta, and the original Omicron variants BA.1 and BA.2. These variants most likely arose from accelerated evolution during chronic infections, leading to distinct "step changes" in fitness compared to earlier lineages. This fitness advantage allowed them to replace previously circulating variants rapidly, resulting in a sharp drop in diversity.

Soft sweeps, on the other hand, are characterised by elevated diversity as multiple variants co-circulate and compete. This pattern was seen with variants like BA.2.75 and BQ.1 which emerged and spread concurrently until they eventually got replaced by XBB (fixation of XBB happened a few months after COG-UK terminated large-scale viral sequencing efforts). These variants share convergent mutations (at key spike protein sites) that provide similar adaptive advantages, such as immune escape, without any single variant fully dominating. Their coexistence reflects soft sweep dynamics, where multiple lineages with comparable fitness advantages increase in frequency simultaneously, maintaining higher diversity.

3. **Interpretation**:
    - What do you observe about the dynamics of the two diversity metrics over time? Do they show similar trends, or are there differences in how they capture the genetic diversity of SARS-CoV-2?

Answer: Both Shannon Entropy (red) and Simpson's Diversity Index (blue) generally follow the same overall trend, with peaks and troughs aligned over time. This indicates that both metrics capture fluctuations in the genetic diversity of SARS-CoV-2 major lineages. Peaks correspond to periods of higher diversity when multiple major lineages coexisted, while troughs align with periods when one lineage dominated (e.g., Delta in late 2021).

Shannon Entropy shows greater sensitivity to increases in diversity, with more pronounced peaks compared to Simpson's Diversity Index. For example, the red line shows a threefold increase in Shannon Entropy from 2021 to 2023, while Simpson's Diversity increases only marginally over the same period. Shannon Entropy remains higher even in periods when Simpson's Diversity is relatively low. This suggests that Shannon Entropy emphasises evenness in the distribution of lineage frequencies more than Simpson's Diversity.

- Why Shannon Entropy increased more dramatically in early 2023, while Simpson's Diversity Index only increased marginally relative to earlier peaks?

Answer: The sensitivity of each metric depends on how it weights evenness versus dominance: Shannon Entropy and Simpson's Diversity Index weigh diversity differently, which explains the more dramatic increase in Shannon Entropy between 2021 and 2023. From 2021 to 2023, the SARS-CoV-2 diversity landscape shifted from dominance by a single lineage (e.g., Delta in 2021) to coexistence of several lineages (BA.2.75, BQ.1, and XBB in 2023). Shannon Entropy increased more dramatically because it is more sensitive to the emergence of additional lineages and their evenness (how evenly the probabilities are spread across categories). Simpson's Diversity, being less sensitive to evenness, increased only marginally because it still weighs the dominance of the most frequent lineages more heavily.

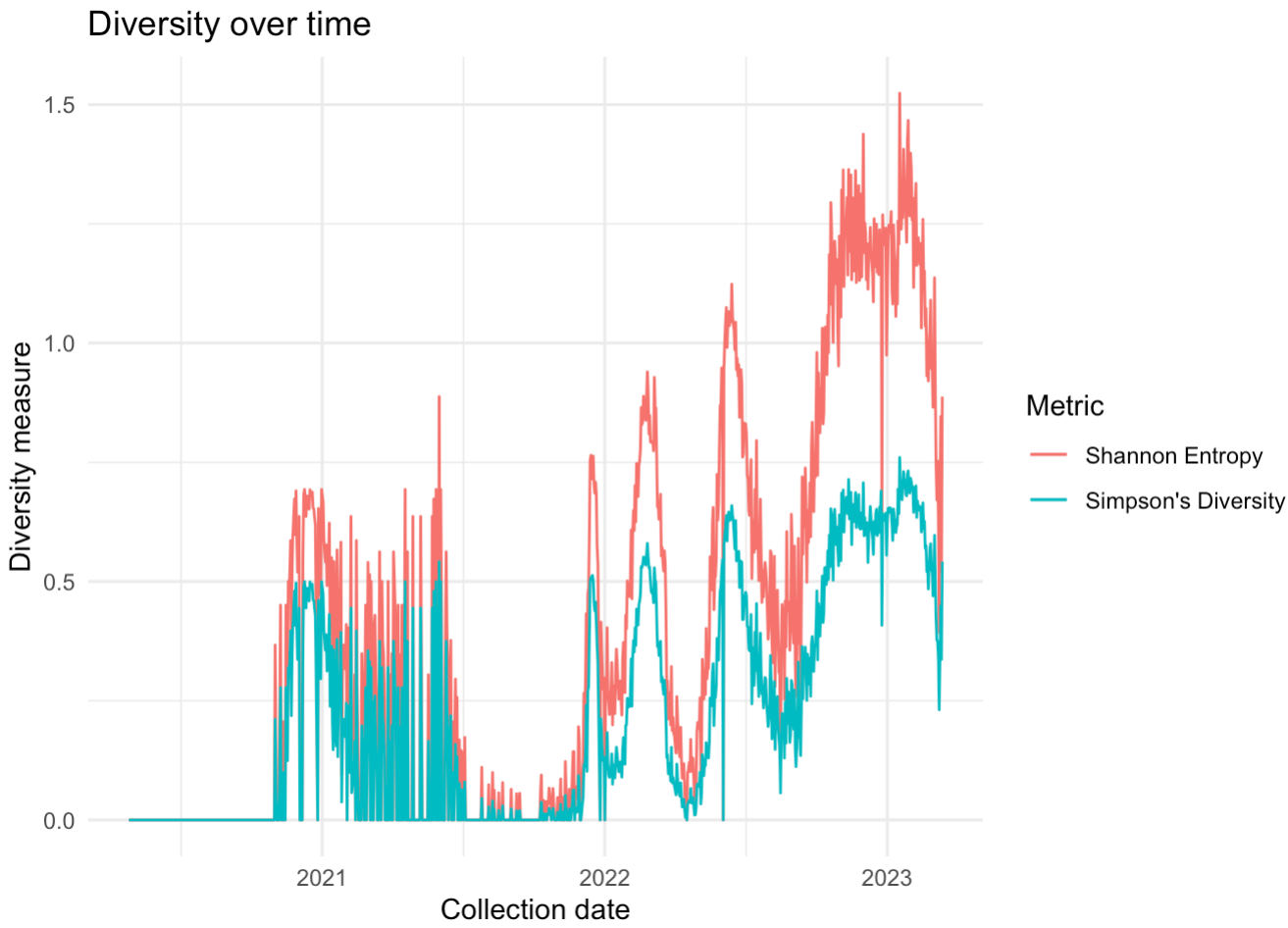
```r
# Calculate Shannon Entropy using base R
calculate_shannon_entropy <- function(data) {
  unique_dates <- unique(data$collection_date)
  results <- data.frame(
    collection_date = unique_dates,
    Shannon_Entropy = NA
  )

  for (i in seq_along(unique_dates)) {
    date_data <- subset(data, collection_date == unique_dates[i])
    frequencies <- date_data$lineage_frequency

    # Shannon Entropy
    shannon_entropy <- -sum(frequencies * log(frequencies), na.rm = TRUE)

    # Store the result
    results$Shannon_Entropy[i] <- shannon_entropy
  }

  return(results)
}

# Use the function on the dataset
shannon_entropy_results <- calculate_shannon_entropy(lineage_summary)

# View the results
head(shannon_entropy_results)
```

```
##   collection_date Shannon_Entropy
## 1      2020-04-26               0
## 2      2020-04-27               0
## 3      2020-04-28               0
## 4      2020-04-29               0
## 5      2020-04-30               0
## 6      2020-05-01               0
```

```
# Calculate Simpson's Diversity Index using base R
calculate_simpsons_diversity <- function(data) {
  unique_dates <- unique(data$collection_date)
  results <- data.frame(
    collection_date = unique_dates,
    Simpsons_Diversity = NA
  )

  for (i in seq_along(unique_dates)) {
    date_data <- subset(data, collection_date == unique_dates[i])
    frequencies <- date_data$lineage_frequency

    # Simpson's Diversity Index
    simpsons_diversity <- 1 - sum(frequencies^2, na.rm = TRUE)

    # Store the result
    results$Simpsons_Diversity[i] <- simpsons_diversity
  }

  return(results)
}

# Use the function on the dataset
simpsons_diversity_results <- calculate_simpsons_diversity(lineage_summary)

# View the results
head(simpsons_diversity_results)
```

```
##   collection_date Simpsons_Diversity
## 1      2020-04-26                  0
## 2      2020-04-27                  0
## 3      2020-04-28                  0
## 4      2020-04-29                  0
## 5      2020-04-30                  0
## 6      2020-05-01                  0
```

```
# Merge the two metrics
diversity_metrics <- merge(
  shannon_entropy_results,
  simpsons_diversity_results,
  by = "collection_date"
)

ggplot(diversity_metrics, aes(x = collection_date)) +
  geom_line(aes(y = Shannon_Entropy, color = "Shannon Entropy")) +
  geom_line(aes(y = Simpsons_Diversity, color = "Simpson's Diversity")) +
  labs(
    title = "Diversity over time",
    x = "Collection date",
    y = "Diversity measure",
    color = "Metric"
  ) +
  theme_minimal()
```


Diversity over time
Diversity measure
1.5
1.0
0.5
0.0
2021
2022
2023
Collection date
Metric
Shannon Entropy
Simpson's Diversity