SARSCoV2Analysis

Biology3579

2025-03-18

R-set-up

```
# Restoring project's environment
renv::restore()
```

```
## - The library is already synchronized with the lockfile.
```

```
# Source library loading function
source(here::here("functions", "libraries.R"))
# Load necessary libraries for analysis
load_libraries()
```

Introduction

The evolution of emerging pathogens, such as SARS-CoV-2, plays a critical role in shaping the trajectory of pandemics. Viral variants arise through genetic mutations, some of which confer increased transmissibility, immune escape, or enhanced replication fitness. These evolutionary changes impact the spread, dominance, and public health burden of different variants. The Alpha, Delta, and Omicron variants, for instance, demonstrated distinct epidemiological characteristics, leading to waves of infections with varying severity and control challenges.

Understanding the fitness advantage of viral variants—defined by their ability to outcompete predecessors in a given population—is essential for anticipating shifts in transmission dynamics. The basic and effective reproduction numbers, which quantify how efficiently a variant spreads, provide key epidemiological insights into variant dominance and the impact of interventions. By leveraging genomic surveillance data, it is possible to estimate these metrics and assess the selective advantage of emerging variants.

This analysis aims to:

Quantify the differences in fitness advantage and reproduction number among major SARS-CoV-2 variants during the COVID-19 pandemic in the UK. Estimate variant-specific growth rates using genomic sequencing data. Evaluate how changes in these metrics influence the effectiveness of epidemic control strategies.

Question 1: SARS-CoV-2 Major Lineages and Trends

This section analyses the major SARS-CoV-2 lineages circulating in England over time using data from the Sanger Institute's COG-UK programme. The dataset contains weekly counts of virus samples per lineage, providing insight into the evolution and spread of different variants.

The analysis consists of the following steps: 1. **Loading the dataset** – Importing the raw data from an online repository. 2. **Classifying major lineages** – Identifying key variants and grouping all others under a general category. 3. **Visualising lineage trends** – Generating stacked area plots to illustrate changes in lineage prevalence over time.

1.1 Load Sanger Data

The dataset is imported from a GitHub repository. The file `Genomes_per_week_in_England.csv` contains weekly counts of SARS-CoV-2 genomic sequences assigned to specific lineages.

```
# Load the raw dataset from an online repository (ensures reproducibility)
sanger_raw <- read.csv("https://raw.githubusercontent.com/Biology3579/SARSCoV2Assignment/main/data/Genomes_per_week

# Save a local copy of the dataset for future reference
write_csv(sanger_raw, here("data", "sanger_raw.csv"))
```

1.2 Process Sanger Data

Before conducting the analysis, the raw data must be processed to ensure that it is structured appropriately for statistical analysis and meaningful interpretation. The key processing steps include:

- Identification and classification of major SARS-CoV-2 lineages: Variants of concern and interest exhibit different transmission rates, immune evasion capabilities, and epidemiological impact. Classifying major lineages enables a clearer understanding of their relative dominance and replacement dynamics over time. This classification focuses on key variants such as Alpha (B.1.1.7), Delta (B.1.617.2), Omicron subvariants (BA.1, BA.2, BA.2.75, BA.4, BA.5, BA.5.3/BQ.1), and XBB, which have played significant roles in shaping the pandemic. By isolating these lineages, their individual growth trends and competitive advantages can be assessed.

All other other lineages - low-prevalence lineages with a much less significant epidemiological impact - are grouped as 'Other'. This approach helps avoid statistical noise and enhances the clarity of lineage trends.

- Formatting the dataset for further statistical analysis and visualisation: The raw dataset requires restructuring to facilitate robust statistical modeling and visualisation.

A more detailed breakdown of these processing steps is provided in the cleaning_and_curating.R script, from which the processing functions are sourced.

```
# Source the data processing functions
source(here("functions", "cleaning_and_curating.R"))

# Pipe to process data concisely
sanger_analysis_data <- sanger_raw %>%
  clean_sanger_data(variants = c("B.1.1.7", "B.1.617.2", "BA.1", "BA.2", "BA.2.75", "BA.4", "BA.5", "BA.5.3", "XBB
  counts_and_frequencies() # Compute total counts and frequencies

# Save the processed dataset for further analysis
write_csv(sanger_analysis_data, here("data", "sanger_analysis_data.csv"))
```

1.3 Stacked Area Plots

To visualise the temporal trends of SARS-CoV-2 lineages, stacked area plots are generated. These plots provide a clear representation of:

- Total lineage counts over time – The absolute number of samples belonging to each lineage in a given week.
- Relative lineage frequencies over time – The proportion of each lineage among the total sequenced genomes in a given week.
  All visualisation functions are sourced from the plotting.R script.

1.3.1 Total Counts of Major Lineages This plot illustrates the absolute number of samples assigned to each major lineage over time. It provides insight into which variants were most prevalent at different points in the pandemic.

```r
# Load the plotting function
source(here("functions", "plotting.R"))

# Generate a stacked area plot of total lineage counts over time
plot_total_variant_counts(sanger_analysis_data)
```

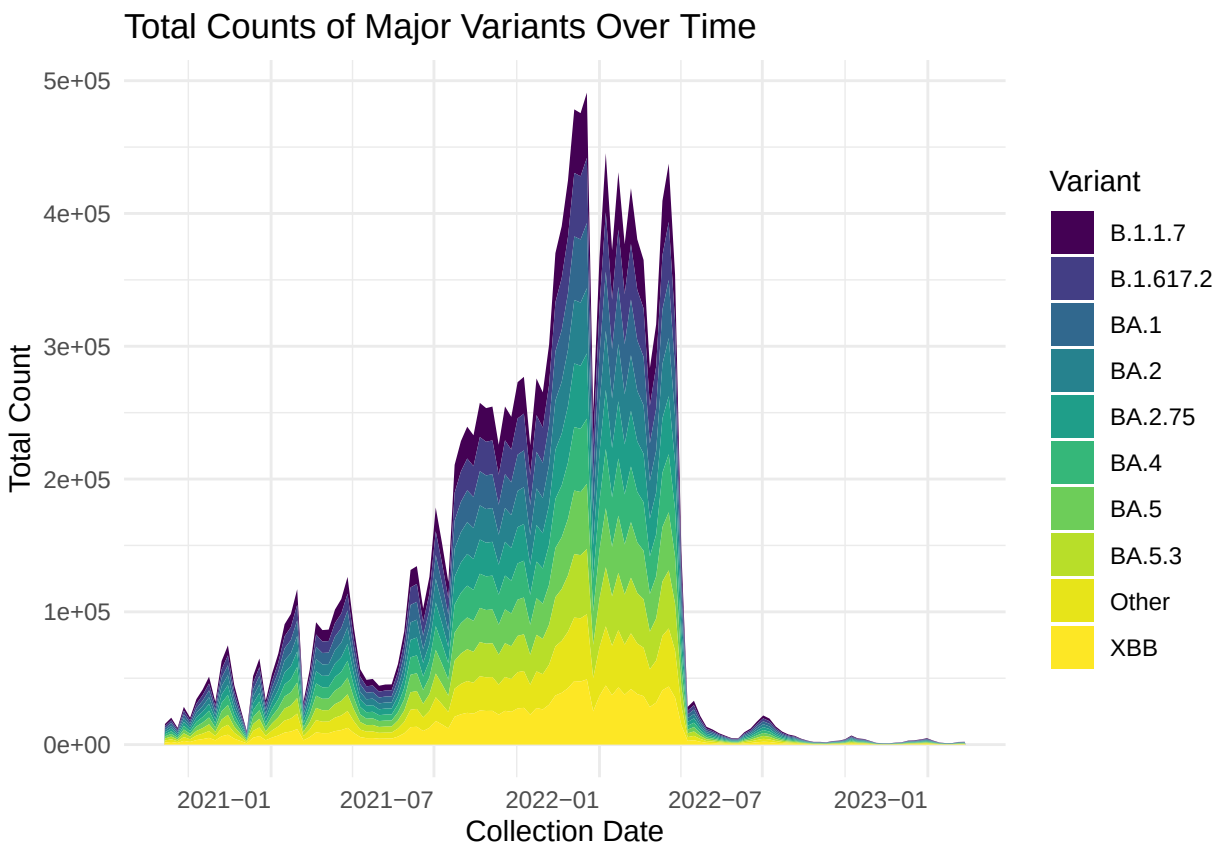

Figure 1: Stacked area plot of weekly counts of major SARS-CoV-2 variants in England over time.

1.3.2 Proportional Frequency of Major Lineages

This plot shows the relative frequencies of major SARS-CoV-2 variants over time, allowing for a better understanding of how different lineages competed and replaced one another.

```r
# Load the plotting function
source(here("functions", "plotting.R"))

# Generate a stacked area plot of lineage proportions over time
plot_variant_frequencies(sanger_analysis_data)
```

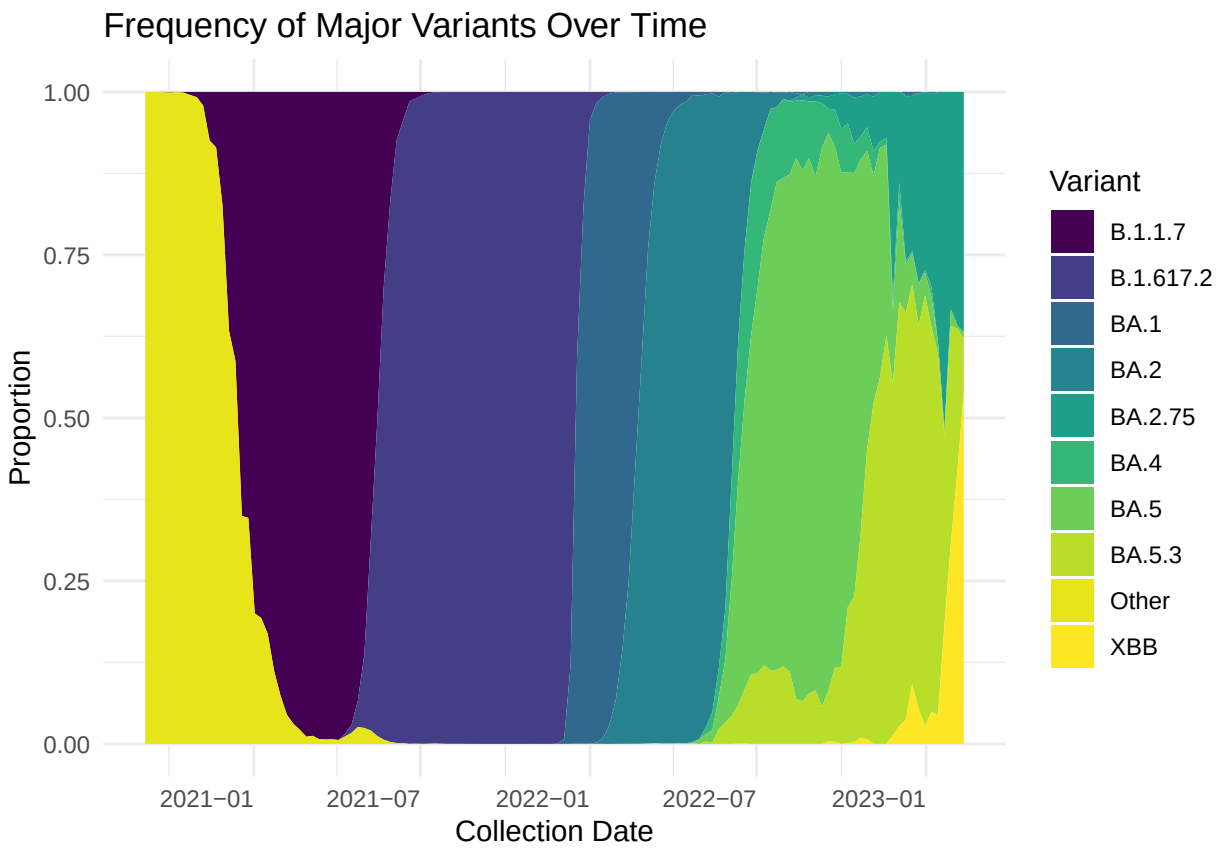

Figure 2: Stacked area plot of proportional representation of major SARS-CoV-2 variants in England over time.

Question 2: BA.2 Variant Trajectory

This section examines the frequency trajectory of the SARS-CoV-2 BA.2 variant using data from two different genomic surveillance sources: the **Sanger Institute's COG-UK dataset** and the **ONS-CIS dataset**. By comparing the trends observed in these datasets, differences in BA.2's emergence, growth, and fixation can be evaluated….

The analysis consists of: 1. **Loading and processing the data** – Importing raw sequencing data from the ONS-CIS and preparing it for analysis. 2. **Visualising BA.2 frequency trajectories** – Comparing weekly counts from the Sanger dataset against 10-day bin counts from ONS-CIS. 3. **Analysing trajectory differences** – Assessing the discrepancies in BA.2's rise and fixation timing between the two datasets.

2.1 BA.2 Trajectories Across Sanger and ONS-CIS Datasets

To investigate BA.2's frequency dynamics, genomic sequence data from ONS-CIS is loaded and then processed to facilitate its usage.

**** 2.1.1 Loading Raw Data**** The **ONS-CIS dataset** provides daily genomic sequence data …

The dataset is imported from a GitHub repository to ensure … The file `lineage_data.csv` contains daily counts of SARS-CoV-2 genomic sequences assigned to specific lineages.

```r
# Import ONS-CIS daily genomic sequence data
onscis_raw <- read_csv("https://raw.githubusercontent.com/mg878/variant_fitness_practical/main/lineage_data.csv", s

# Save the dataset locally for reproducibility
write_csv(onscis_raw, here("data", "onscis_raw.csv"))
```

**** 2.1.2 Processing and Cleaning ONS-CIS Data**** Before visualisation, the ONS-CIS dataset is processed to standardise lineage classification, bin the data into 10-day intervals and compute lineage frequencies within each of these 10-day bins.

```r
# Load data cleaning functions
source(here("functions", "cleaning_and_curating.R"))

# Pipe to process data concisely
onscis_analysis_data <- onscis_raw %>%
  clean_onscis_data() %>%   # Clean the raw ONS-CIS data
  bin_and_calculate_frequencies(bin_size = 10)  # Bin data into 10-day intervals and calculate frequencies

# Save the cleaned and processed dataset
write_csv(onscis_analysis_data, here("data", "onscis_analysis_data.csv"))
```

**** 2.1.3 Plotting BA.2 Frequency Trajectories**** To compare BA.2's trajectory across datasets, the variant's frequency is extracted from both the Sanger dataset (weekly counts) and the ONS-CIS dataset (10-day bin counts) and the results are plotted to illustrate their respective growth patterns

```r
# Load the plotting function
source(here("functions", "plotting.R"))

# Extract BA.2 data from Sanger dataset
ba2_sanger <- sanger_analysis_data %>%
  filter(variant == "BA.2") %>%
  mutate(source = "Sanger (Weekly)")

# Extract BA.2 data from ONS-CIS dataset
ba2_onscis <- onscis_analysis_data %>%
  filter(variant == "BA.2") %>%
  mutate(source = "ONS-CIS (10-day Binned)")

# Plot BA.2 frequency trajectory
plot_ba2_frequency_comparison(ba2_sanger, ba2_onscis)
```

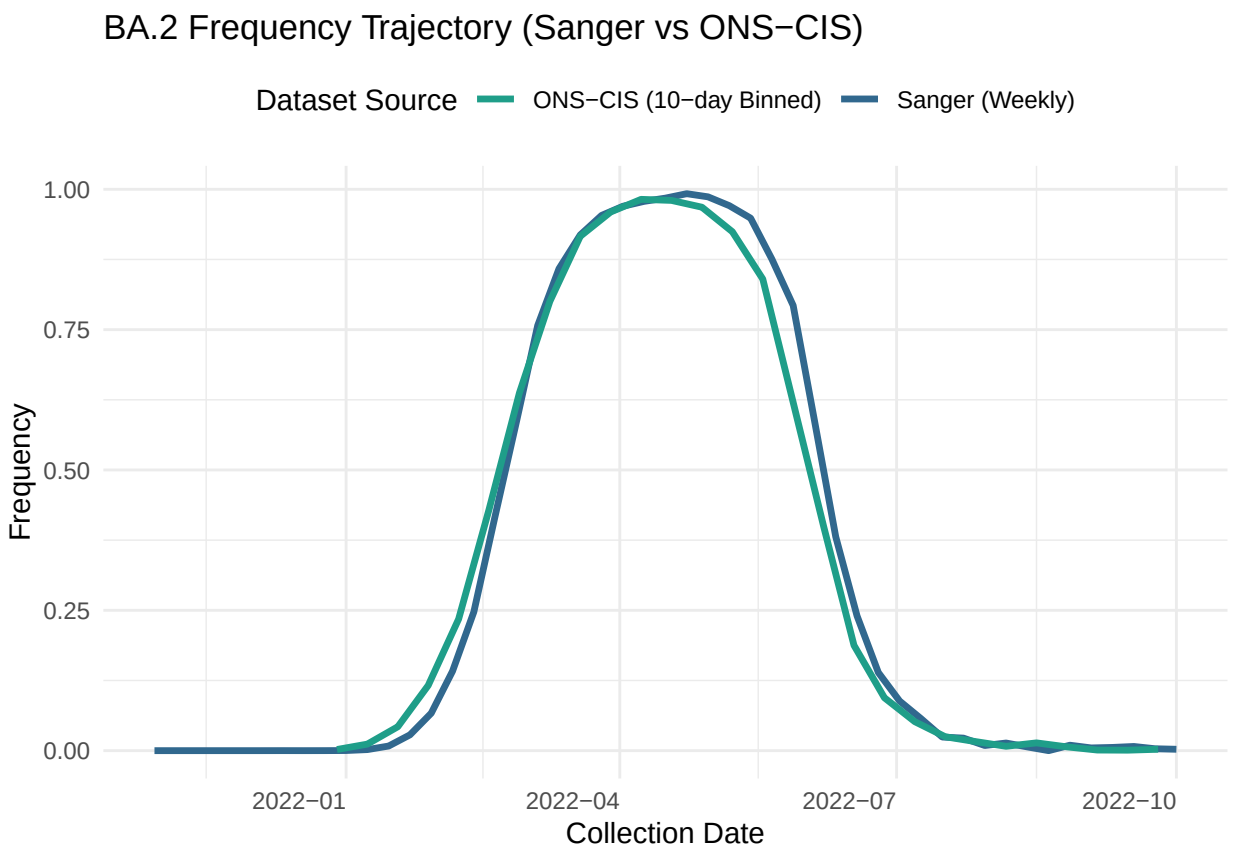

Figure 3: Frequency trajectories of BA.2 from the Sanger (weekly) and ONS-CIS (10-day binned) datasets.

###2.2 Comparing Trajectories

The frequency trajectories of the BA.2 variant from the Sanger (weekly) dataset and the ONS-CIS (10-day binned) dataset exhibit a similar overall pattern, but differences in the timing of emergence, growth rate, and fixation are observed. BA.2 appears to rise slightly earlier in the Sanger dataset, likely due to its reliance on community testing and targeted sequencing, which are more responsive to emerging trends. In contrast, the ONS-CIS dataset, which follows a structured longitudinal survey, may introduce a slight delay in detecting early exponential growth. Additionally, the 10-day binning in ONS-CIS smooths out fluctuations, making the trajectory appear more gradual.

BA.2 reaches near 100% prevalence at approximately the same time in both datasets, but the Sanger dataset shows a steeper increase, suggesting a faster observed expansion. This could be attributed to testing policies, where sequencing efforts may have prioritized BA.2 once it became dominant, whereas the ONS-CIS dataset, which samples individuals randomly, provides a less biased population-wide estimate. The decline phase also differs, with the Sanger dataset showing a sharper drop, likely due to sequencing efforts shifting to newer variants, whereas the ONS-CIS dataset suggests a more gradual decline, indicating that BA.2 may have persisted longer at low levels.

These differences underscore the importance of data collection strategies and surveillance methods when interpreting genomic epidemiology data. While both datasets confirm BA.2's rise, dominance, and decline, variations in sampling strategies, binning effects, and sequencing priorities can influence the observed trajectory. Using multiple surveillance approaches provides a more comprehensive and reliable picture of variant dynamics, which is essential for monitoring viral evolution and informing public health responses.

Question 3: variant Fixation Analysis

This section investigates the fixation dynamics of three SARS-CoV-2 variants—B.1.617.2 (Delta), BA.1 (Omicron), and BA.2 (Omicron)—using weekly counts from the Sanger dataset. The objective is to determine which variant reached fixation the fastest and which exhibited the highest selective advantage under a logistic growth model. The selective advantage (□) is estimated by fitting logistic growth curves to the frequency trajectories of each variant.

3.1 Selecting variant growth phases for logistic growth modelling

To model variant growth, it is necessary to identify the period during which each variant was actively expanding in the population. The first step is to visualise the frequency trajectories of B.1.617.2, BA.1, and BA.2 to determine their respective growth phases.

```
# Load the plotting function
source(here("functions", "plotting.R"))

# Plot variant frequency trajectories
plot_variant_frequency_trajectories(variants = c("B.1.617.2", "BA.1", "BA.2"))
```

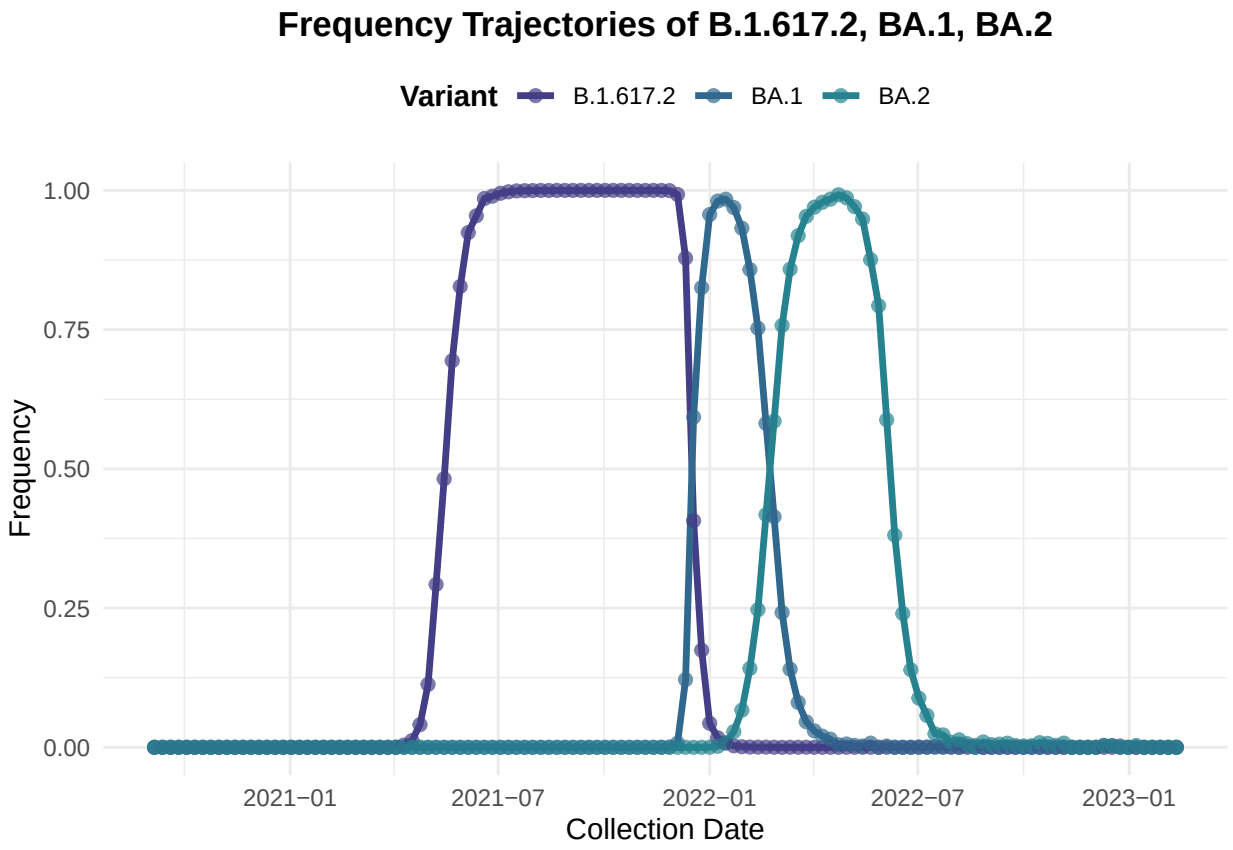

Figure 4: Frequency trajectories of B.1.617.2, BA.1, and BA.2 over time.

Then, a function was implemented to automatically extract the growth phase start and end dates for each variant, ensuring a consistent approach across lineages. Details on the selection criteria and methodology can be found in the variant_analyses script.

```
# Load the growth phase extraction function
source(here("functions", "variant_analyses.R"))

# Extract growth phase dates for each variant
growth_phases <- bind_rows(
  extract_growth_phase_dates(sanger_analysis_data, "B.1.617.2"),
  extract_growth_phase_dates(sanger_analysis_data, "BA.1"),
  extract_growth_phase_dates(sanger_analysis_data, "BA.2")
)

# Print outputs
print(growth_phases)
```

```
## # A tibble: 3 x 3
```

```
##   variant   start_date end_date
##   <chr>     <date>     <date>
## 1 B.1.617.2 2021-04-03 2021-09-11
## 2 BA.1      2021-10-30 2022-01-15
## 3 BA.2      2022-01-01 2022-04-23
```

Table 1: Estimated growth phase start and end dates for each variant.

Once the growth phases are identified, the dataset is filtered to include only data within these periods. This ensures that the logistic growth model is fitted to the exponential growth phase, excluding periods of stagnation or decline.

```
# Filter the Sanger analysis data for selected variants and their growth phases
selected_variant_data <- sanger_analysis_data %>%
  inner_join(growth_phases, by = "variant") %>%
  filter(collection_date >= start_date & collection_date <= end_date)
```

3.2 Logistic Growth Modelling and Selective Advantage Estimation

A logistic growth model is used to estimate the selective advantage (s) of each variant. This model assumes that a variant follows a sigmoidal trajectory, growing exponentially at first and then slowing as it approaches fixation. The logistic growth function is defined as:

$$f(t) = \frac{f(0)e^{st}}{1 + f(0)\left(e^{st} - 1\right)}$$

where:

- $f(t)$ is the variant frequency at time t.
- s is the selective advantage.
- $f(0)$ is the initial frequency.

The model is fitted separately to each variant's growth phase to estimate s, allowing for a comparison of the relative fitness of B.1.617.2, BA.1, and BA.2.

```
# Load the logistic growth model fitting function
source(here("functions", "variant_analyses.R"))

# Define Logistic Growth Function
logistic_growth <- function(t, s, f0) {
  1 / (1 + ((1 - f0) / f0) * exp(-s * t))
}

# Fit logistic growth models for each variant
logistic_predictions_variant <- selected_variant_data %>%
  group_by(variant) %>%
  group_split() %>%
  map_dfr(~fit_logistic_growth_general(.x,
                                       time_col = "collection_date",
                                       frequency_col = "variant_frequency",
                                       group_col = "variant"))
```

Then, to visually assess the model fits, the estimated logistic growth curves are plotted alongside the observed data.

```r
# Load the plotting function
source(here("functions", "plotting.R"))

# Plot logistic growths for varaints
plot_logistic_growth(
  data = selected_variant_data,
  growth_phases = growth_phases,
  variants = unique(selected_variant_data$variant)
)
```

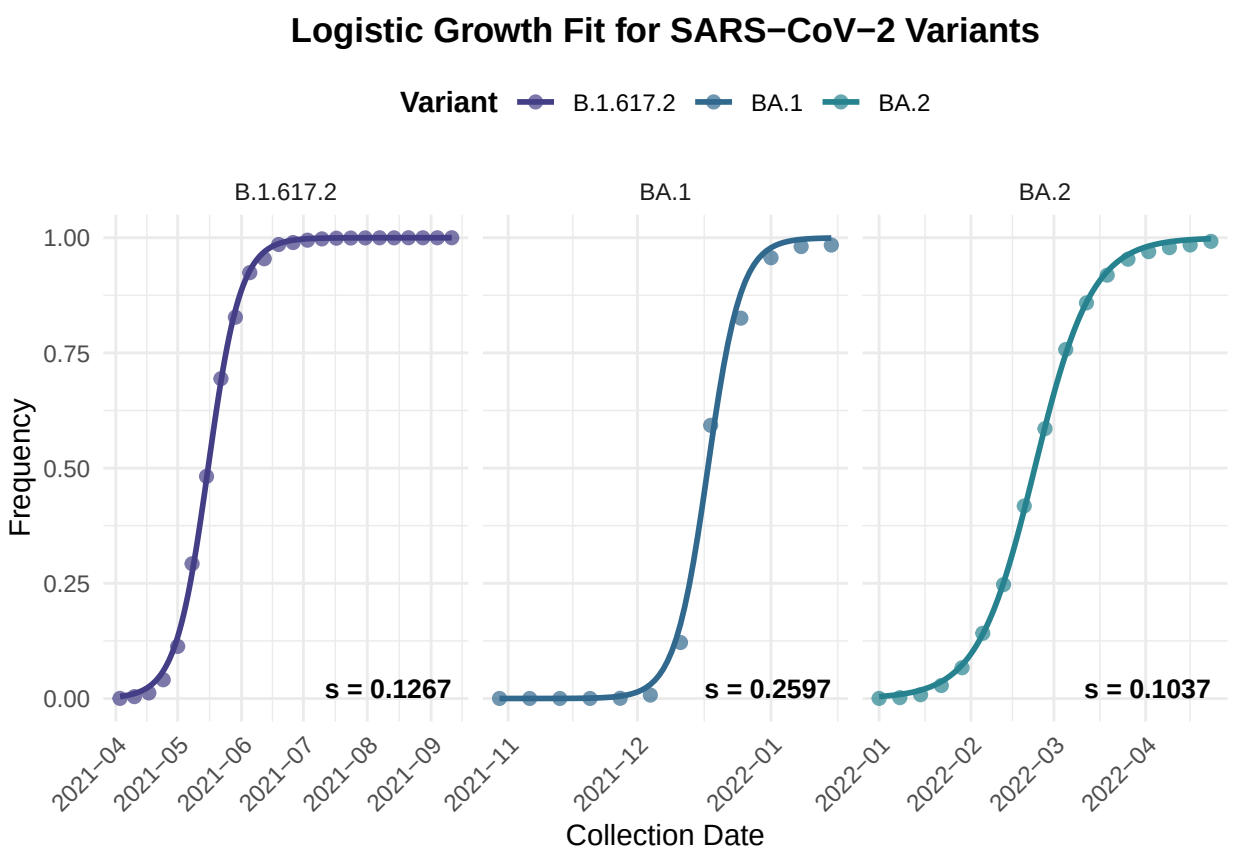

Figure 5: Figure 5: Logistic growth model fits for B.1.617.2, BA.1, and BA.2, with their respective s-values.

3.3 Interpretation of Variant Fixation and Selective Advantage

The logistic growth model allows for a direct comparison of the fixation speed and selective advantage of B.1.617.2, BA.1, and BA.2. The estimated □-values provide insight into how rapidly each variant displaced its predecessors.

Preliminary results indicate that BA.2 reached fixation the fastest, with a steeper growth curve compared to B.1.617.2 and BA.1. This suggests that BA.2 had the highest selective advantage, likely due to increased transmissibility or immune escape properties.

B.1.617.2 (Delta): Exhibited strong growth but was eventually displaced by BA.1. BA.1 (Omicron): Replaced Delta rapidly but showed a more gradual fixation curve compared to BA.2. BA.2 (Omicron subvariant): Displayed the steepest logistic trajectory, indicating the highest selective advantage among the three. The differences in selective advantage align with epidemiological observations, where BA.2 demonstrated enhanced transmissibility over BA.1, leading to its rapid dominance. These findings underscore the importance of logistic growth modeling in understanding variant competition and evolutionary dynamics.

Question 4: Regional Analysis of Delta Variant

This section examines the regional spread of the Delta (B.1.617.2) variant across England using an anonymised dataset from COG-UK. The analysis aims to visualise Delta's frequency trajectory by region, fit a logistic growth model to estimate its selective

advantage, and assess whether regional differences in Delta's emergence and growth could be attributed to a founder effect.

** 4.1 Load and process regional delta data **

The dataset contains sequenced Delta cases from various regions in England. It is first imported from a GitHub repository to ensure …, and processed to remove missing regional identifiers and structure the data for analysis.

```
# Read the RDS file from GitHub
url <- "https://raw.githubusercontent.com/Biology3579/SARSCoV2Assignment/main/data/delta-d2.rds"
regional_delta_raw <- readRDS(url(url, "rb"))  # "rb" ensures reading in binary mode

# Save the dataset locally
write_rds(regional_delta_raw, here("data", "regional_delta_raw.rds"))
```

```
# Load processing functions
source(here("functions", "cleaning_and_curating.R"))

# Pipe to process data concisely
delta_analysis_data <- regional_delta_raw %>%
  clean_delta_data() %>%  # Clean the dataset
  counts_and_frequencies_delta() %>%  # Calculate variant frequencies
  write_csv(here("data", "delta_analysis_data.csv"))  # Save the processed dataset
```

4.2 Delta Frequencies and Logistic Growth by Region

** 4.2.1 Delta Frequencies by Region **

To assess regional variation in Delta prevalence, weekly frequencies of Delta are computed and visualised.

```
# Load plotting function
source(here("functions", "plotting.R"))

# Aggregate data by week and region
delta_weekly <- delta_analysis_data %>%
  mutate(week = floor_date(date, unit = "week")) %>%
  group_by(week, phecname) %>%
  summarise(
    delta_frequency = mean(delta_frequency, na.rm = TRUE),
    .groups = "drop")

# Plot Delta frequency trajectories by region
plot_delta_frequencies(delta_weekly)
```

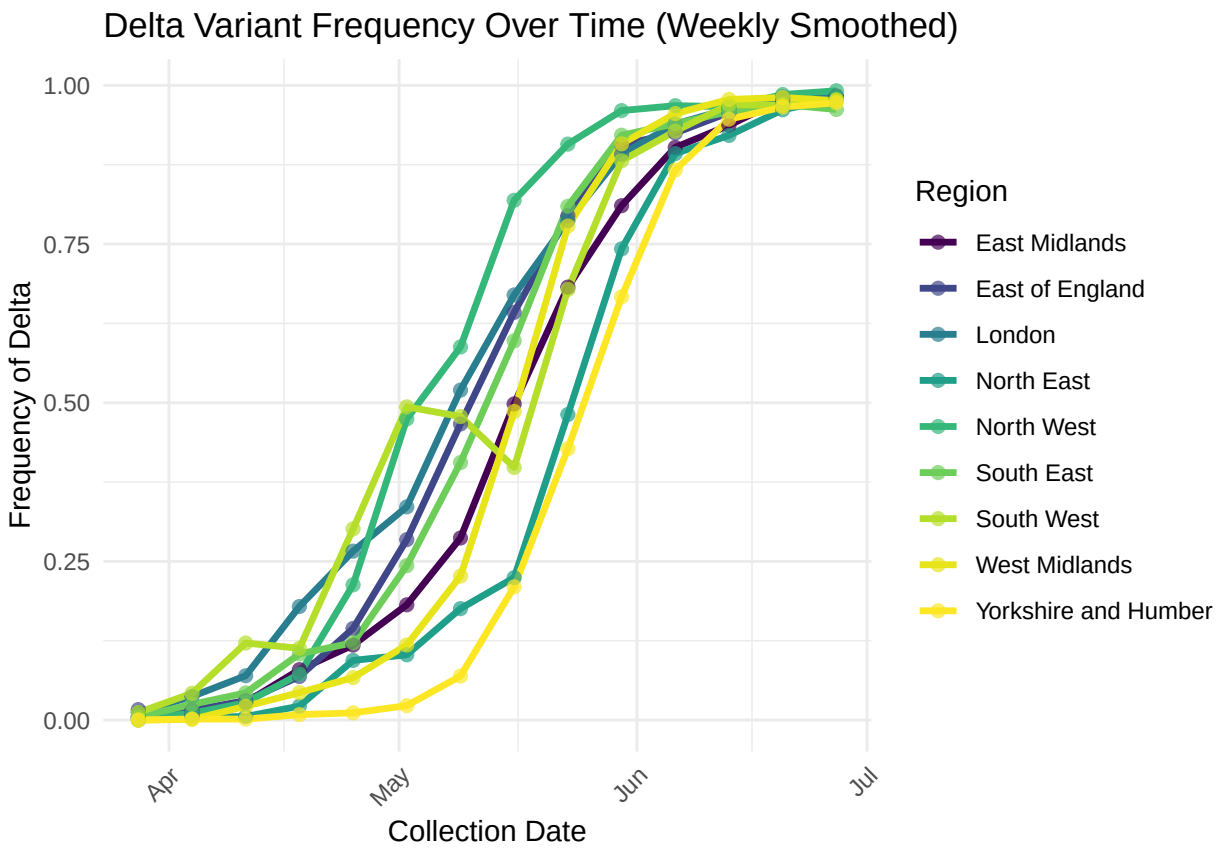

Figure 6: Weekly frequency trajectories of Delta across English regions.

**** 4.2.2 Logistic Growth Model by Region ****

A logistic growth model is fitted to each region's frequency trajectory to estimate Delta's selective advantage (s) and its initial frequency ($f(0)$).

```
# Load logistic growth fitting function
source(here("functions", "variant_analyses.R"))

# Fit logistic growth models to regional data
logistic_predictions_region <- delta_weekly %>%
  group_by(phecname) %>%
  group_split() %>%
  map_dfr(~fit_logistic_growth_general(.x, time_col = "week",
                                       frequency_col = "delta_frequency",
                                       group_col = "phecname"))
```

The fitted logistic growth curves are overlaid onto the observed frequency data to assess how well the model represents regional growth patterns.

```
# Load plotting function
source(here("functions", "plotting.R"))

# Plot Logistic Growths for each region
plot_logistic_growth_region(
  observed_data = delta_weekly,
  predicted_data = logistic_predictions_region
)
```

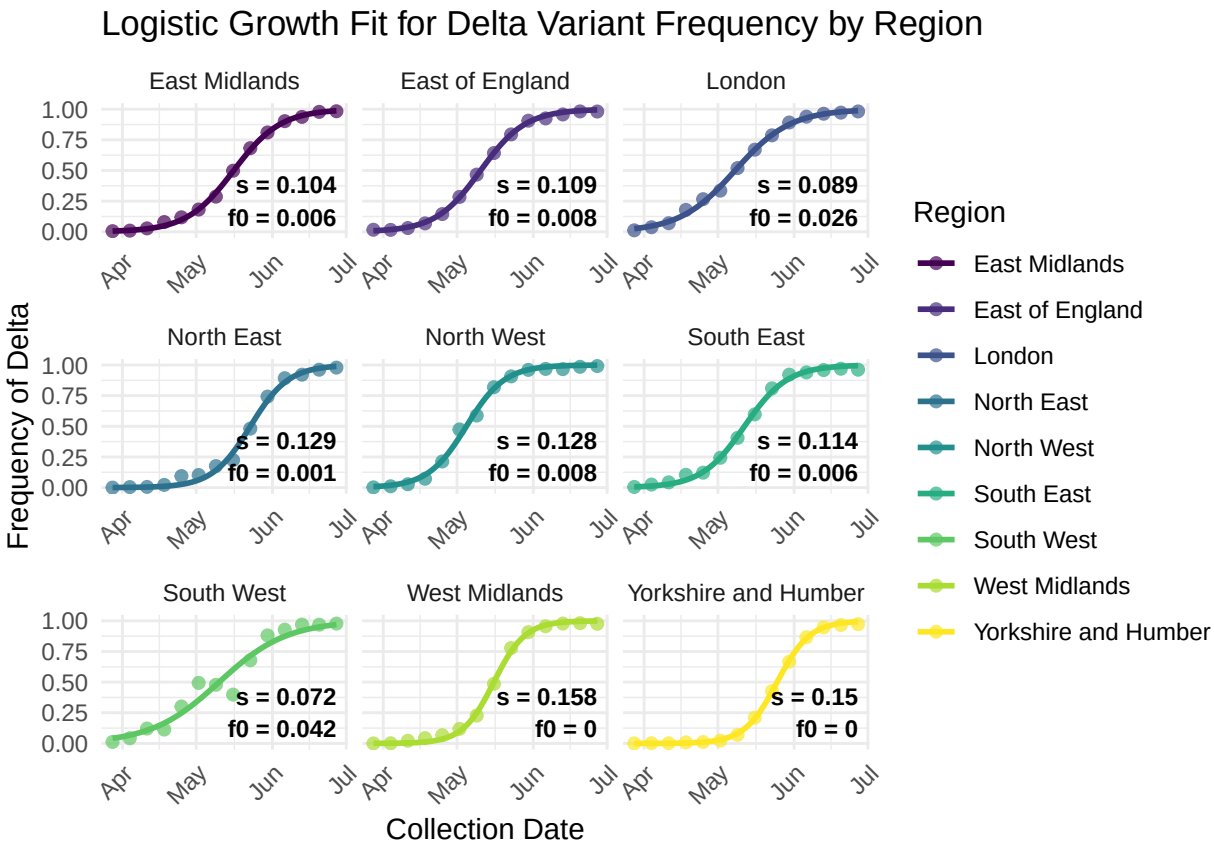

Figure 7: Logistic growth model fits for Delta frequency trajectories across regions.

**** 4.3 logistic growth for each region ****

The logistic growth model provides insights into regional differences in the emergence and spread of the Delta variant. The West Midlands exhibited the highest growth rate ($s = 0.1577$), indicating that once introduced, Delta spread rapidly within the region. In contrast, the South West had the earliest rise in frequencies ($f_0 = 0.0416$), suggesting that Delta was detected earlier in this region compared to others. These differences highlight the role of regional factors in shaping the trajectory of variant emergence and dominance.

Several factors may explain these regional variations. Population density and mobility likely influenced the spread, as seen in London, where a relatively high initial frequency (f_0) was observed, but the growth rate remained moderate. This suggests that high levels of mobility may have facilitated early introductions, but sustained transmission did not occur as rapidly as in the West Midlands. In contrast, the West Midlands had a lower initial frequency but the highest growth rate, suggesting that fewer early introductions led to rapid expansion once the variant became established.

Differences in testing and surveillance may have also contributed to variations in Delta's observed growth patterns. Regions with more intensive genomic surveillance may have detected Delta earlier, even before it reached a dominant presence. Additionally, travel patterns and introduction events played a role, as regions with major transport hubs, such as London and the South East, likely experienced earlier introductions via international travel. In contrast, regions with lower initial detection (low f_0) but high growth rates (high s), such as the West Midlands, suggest that while Delta was not introduced as early, it spread rapidly once present.

Sociodemographic and behavioral factors may have further influenced transmission rates. Variations in vaccination uptake, workplace exposure, and social interactions could have impacted the speed at which Delta spread in different regions. Areas with large populations of essential workers and high-contact professions may have experienced accelerated transmission compared to regions with lower occupational exposure.

The possibility of a founder effect in Delta's regional spread was also considered. A founder effect occurs when a new population is established by a small number of initial individuals, potentially leading to distinct growth dynamics and reduced genetic variation. The variation in f_0 across regions suggests that Delta was introduced at different times and locations, supporting the hypothesis of multiple independent introductions rather than a single uniform spread. In particular, regions with low f_0 but high s, such as the

West Midlands and Yorkshire and Humber, align with the characteristics of a local founder effect, where a few early cases triggered rapid outbreaks.

However, there is also evidence against a founder effect as the growth rates (s) across regions remain relatively consistent, which would not be expected if a strong founder effect were the primary driver. Additionally, in some regions, such as the South West, where f_0 was high, Delta spread widely early on, suggesting that multiple introduction events, rather than a single founder-driven emergence, shaped the overall transmission pattern.

The analysis suggests that multiple independent introductions, rather than a single founder event, were primarily responsible for the spread of Delta across England. However, regional founder effects may have played a role in specific locations, particularly where Delta was introduced by a small number of individuals and subsequently expanded rapidly. The regional differences in Delta's spread underscore the importance of considering population density, mobility, testing strategies, and travel patterns when assessing variant dynamics. While a founder effect may have contributed to localized outbreaks, the broader transmission trends indicate a complex interaction of epidemiological and demographic factors shaping Delta's spread across the country.

Question 5: Delta Incidence and Rt Estimation

This section examines the true incidence of Delta infections and estimates the time-varying reproduction number (R_t) using both sequencing data from the Sanger dataset and daily COVID-19 case counts from the ONS-CIS dataset. The goal is to assess how sequencing-based estimates compare to broader epidemiological trends and to evaluate the reliability of R_t estimates derived from different data sources.

5.1 Estimating the True Incidence of Delta

The analysis of the Delta variant thus far has been based on sequencing data from samples processed by the Sanger Institute. While this method is valuable for monitoring the relative growth and spread of Delta compared to other variants, it does not provide a direct measure of true incidence. True incidence refers to the actual number of individuals infected with Delta at a given time, including those who were never tested or whose infections were not sequenced.

There are several reasons why sequencing-based estimates may differ from the true number of Delta infections. Firstly, not all PCR-positive samples undergo sequencing, meaning that the available sequencing data represents only a fraction of total infections. This can introduce bias, as the sampled subset may not be fully representative of the broader infected population. Additionally, geographic disparities in sequencing rates may further affect the accuracy of estimates, as some regions may contribute proportionally more sequenced samples than others.

Another factor influencing incidence estimates is testing bias. Many individuals infected with Delta may never be tested, either due to mild or asymptomatic infections or limited access to testing facilities. As sequencing relies on PCR-confirmed cases, infections that are undiagnosed remain unaccounted for, leading to an underestimation of true incidence. Furthermore, delays in sequencing and reporting introduce additional uncertainty. While PCR test results are typically available within one to two days, sequencing requires more time. By the time sequencing data is processed and reported, the actual number of Delta cases in the population may have already changed, making real-time incidence estimation challenging.

To overcome these limitations, alternative approaches are required to estimate the true incidence of Delta. One method involves integrating sequencing data with population-level COVID-19 case counts, such as those from the ONS-CIS survey, which captures infections beyond those that were sequenced. Other data sources, including wastewater surveillance and hospitalisation records, can also contribute to a more comprehensive understanding of Delta's prevalence. These complementary methods help address biases in sequencing data and provide a more reliable estimate of the true number of Delta infections.e population may have already changed, making it difficult to use this data for real-time incidence tracking.

Thus, to approximate the true incidence of Delta, the proportion of Delta sequences in England (as reported in the Sanger dataset) is applied to the 7-day averaged daily case counts from the ONS-CIS dataset. This approach assumes that the proportion of Delta observed in sequenced samples reflects its relative frequency among all infections in the community, allowing for an indirect estimation of the total number of Delta cases over time.

```
# Load daily COVID-19 case data
delta_daily_raw <- read.csv("https://raw.githubusercontent.com/Biology3579/SARSCoV2Assignment/main/data/daily-new-
```

```r
# Save the dataset locally
write_csv(delta_daily_raw, here("data", "delta_daily_raw.csv"))

# Clean and format the dataset
delta_daily_clean  <- delta_daily_raw %>%
    mutate(
      date = as.Date(date),  # Convert date column to Date format
      cases_sevendayaveraged = as.numeric(cases_sevendayaveraged)  # Ensure numerical format
    )
```

Since weekly sequencing data begins later than daily case data, early months are discarded to align timeframes. The daily case counts are aggregated into 7-day bins to match the weekly proportions of Delta from the Sanger dataset.

```r
# Load delta case estimation function
source(here("functions", "variant_analyses.R"))

# Estimate Delta cases using Sanger variant proportions
daily_delta_estimates <- estimate_delta_cases(daily_cases = delta_daily_clean,
                                              sanger_data = sanger_analysis_data)
```

A comparison is made between estimated daily Delta cases and sequencing-based Delta cases to evaluate discrepancies.

```r
# Load plotting function
source(here("functions", "plotting.R"))

# Plot estimated vs sequenced Delta cases
plot_estimated_vs_sequenced_delta_cases(
  delta_estimates = daily_delta_estimates,
  sanger_data = sanger_analysis_data)
```

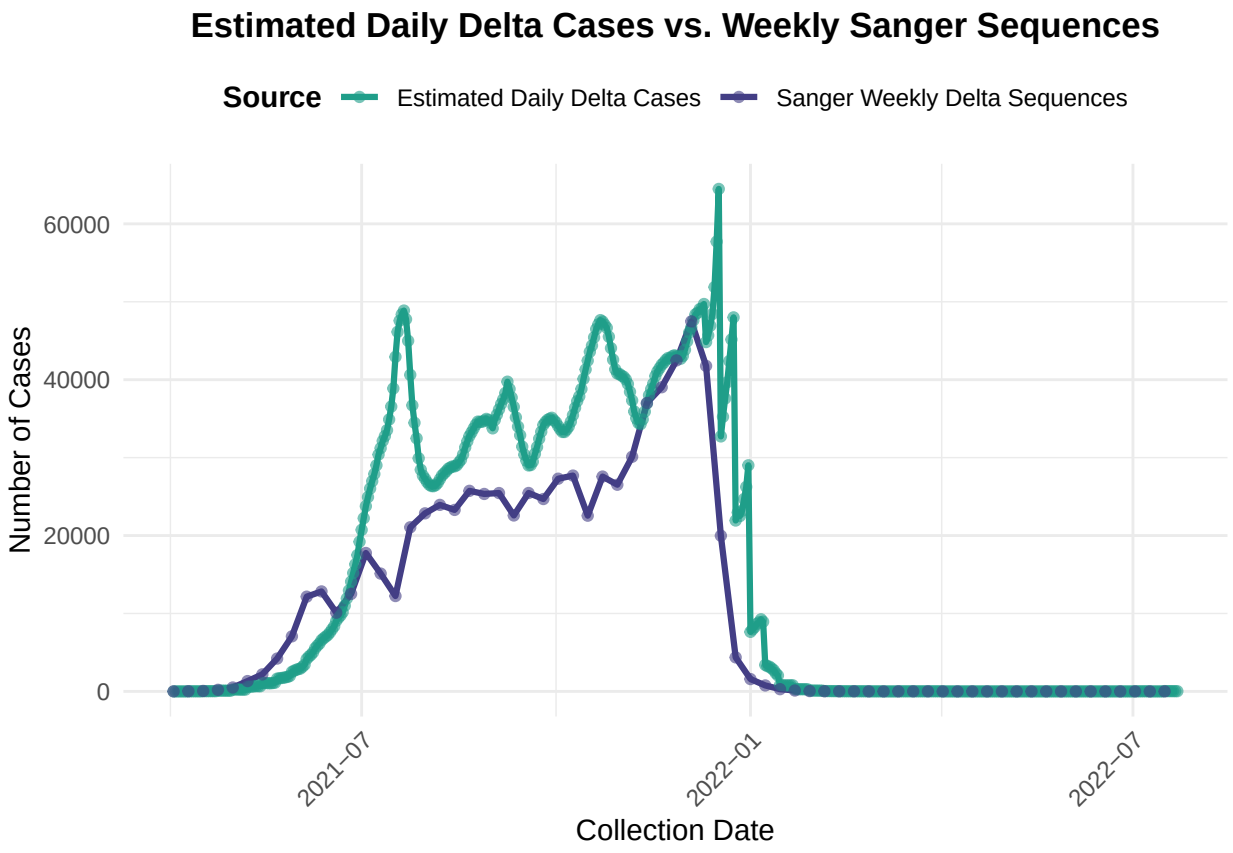

Figure 8: Estimated daily Delta cases vs. Delta cases from sequencing data.

5.2 Reflection on Differences Between the Two Estimates

The estimated Delta case counts, derived from population-level data, differ from sequencing-based counts due to several key factors. Testing bias plays a significant role, as the estimated cases include all reported positive tests, whereas the Sanger dataset represents only a subset of samples selected for sequencing. This introduces sampling bias, as not all infections are sequenced, potentially underrepresenting the true number of Delta cases.

Differences in temporal resolution also contribute to discrepancies. The estimated cases, based on daily reports, provide a smoother trend, whereas sequencing data is collected weekly, making it more discrete and subject to fluctuations due to sample collection and processing times. Additionally, a scaling effect exists—sequencing data captures only a fraction of all infections, leading to lower absolute numbers, though relative trends between the two datasets should broadly align.

Reporting delays further distinguish the datasets. Sequencing takes additional time, meaning the Sanger dataset may lag behind real-time incidence, whereas daily case estimates provide a more immediate measure of transmission dynamics.

Despite these differences, both estimates show similar overall trends, capturing Delta's rise, peak, and subsequent decline. These findings highlight the need to integrate multiple data sources for a more accurate assessment of variant spread and epidemiological trends.

5.3 Estimating the Reproduction Number (R_t)

To assess the transmission dynamics of Delta, the time-varying reproduction number (R_t) is calculated using the estimated daily Delta case counts.

```
# Load rt estimation function
source(here("functions", "variant_analyses.R"))

# Generate Rt estimates
rt_estimates <- generate_rt_estimates(daily_delta_estimates)
```

```
# Load plotting function
source(here("functions", "plotting.R"))

# Plot Rt estimates
plot_rt_estimates(rt_estimates)
```

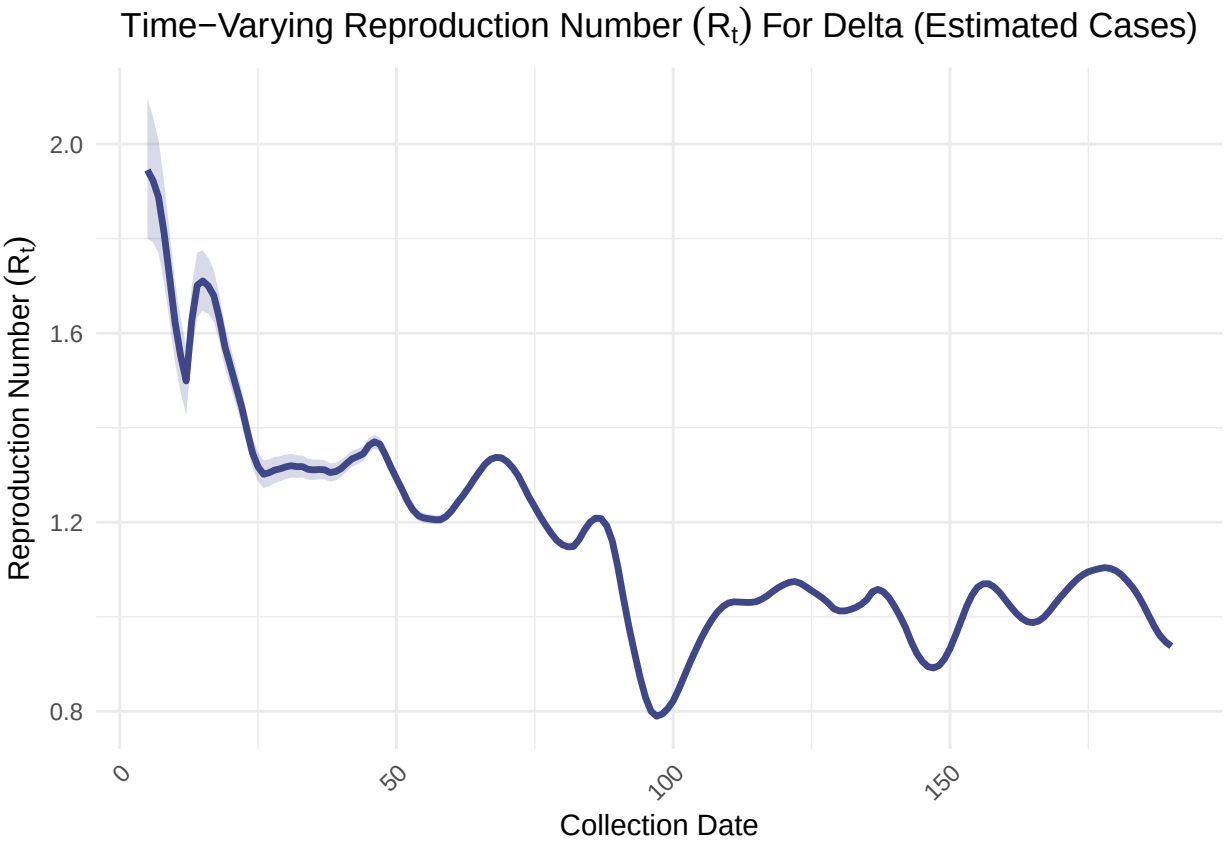

Figure 9: Estimated R_t values for Delta transmission.

5.4 Comparison of R_t Estimates

The estimated R_t values derived from Sanger sequencing data and ONS-CIS population data were compared to assess their reliability. The mean R_t values were similar across the two datasets, with the Sanger-based estimate at $R_t = 1.944 \pm 0.075$ and the ONS-CIS estimate at $R_t = 1.915 \pm 0.279$. The primary difference between these estimates lies in their spread of values. The Sanger-based estimate is more precise, with a smaller standard deviation (± 0.075), while the ONS-CIS estimate has a greater degree of uncertainty, as indicated by its larger standard deviation (± 0.279). However, the two estimates do not differ significantly from a statistical perspective, as the Sanger estimate falls within the confidence interval of the ONS-CIS estimate (1.636–2.194).

Each dataset has strengths and limitations when estimating R_t. The Sanger-based estimate ($R_t = 1.944 \pm 0.075$) offers higher precision due to its lower variability and directly reflects variant-specific transmission dynamics. However, it is derived only from sequenced cases, which means it is likely biased toward tested and sequenced individuals, potentially underestimating infections in the broader population. In contrast, the ONS-CIS estimate ($R_t = 1.915 \pm 0.279$) is more representative of real-world transmission, as it captures both symptomatic and asymptomatic cases, reducing testing bias. Since it is based on a structured random sampling approach, it is likely to provide a more accurate measure of total infections. However, the higher uncertainty associated with this estimate results in a wider confidence interval, making it less precise than the Sanger-based calculation.

Both estimates suggest that Delta had a high initial transmission rate ($R_t \approx 1.9$), meaning that each infected individual was, on average, spreading the virus to nearly two others, leading to rapid exponential growth. However, methodological differences in data collection and sampling influence the reliability of each estimate. The Sanger dataset provides a more precise estimate but is limited to sequenced cases, which may not fully capture the true scale of infections. On the other hand, the ONS-CIS estimate is likely a more accurate reflection of community-wide transmission, as it includes a broader and more representative sample. Ultimately, both approaches offer valuable insights into Delta's epidemiology. The Sanger dataset is useful for tracking variant-specific transmission, while the ONS-CIS dataset provides a more comprehensive understanding of overall spread. Integrating multiple data sources is essential for accurate public health decision-making, ensuring a well-rounded assessment of variant dynamics.

References