SARSCoV2Analysis

Biology3579

2025-03-18

R-set-up

Important note: *Comprehensive instructions, including all prerequisites and detailed guidelines for running this analysis and relevant code, are available in the README file of my GitHub repository for this project. Please refer to the README to ensure proper setup and usage.*

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

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

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

Introduction

The emergence and rapid spread of SARS-CoV-2 variants have been central to the evolution of the COVID-19 pandemic. As the virus acquires mutations, its transmissibility, immune escape potential, and overall fitness can shift — driving waves of infection led by increasingly competitive lineages such as Alpha, Delta, and Omicron [1]. Tracking how these variants emerge, spread, and replace one another provides key insight into viral adaptation and informs timely public health responses, including vaccine updates and intervention strategies.

Our understanding of these dynamics relies heavily on surveillance systems, each offering distinct advantages and limitations. While genomic sequencing reveals lineage-specific trends, population-based surveys provide broader, less biased snapshots of community transmission. By integrating data from both the Sanger Institute's genomic surveillance and the ONS-CIS infection survey, this analysis examines the rise and spread of major variants in England. Through comparative modelling, it explores transmission dynamics, regional spread, and true infection levels — building a clearer picture of how SARS-CoV-2 evolves and spreads

---

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

To investigate the dynamics of major SARS-CoV-2 variant emergence and replacement in England, this section uses data from the Sanger Institute's COVID-19 Genomics UK (COG-UK) sequencing programme. This dataset — representing one of the largest and most comprehensive sources of genomic surveillance data in the UK during the COVID-19 pandemic — provides weekly counts of sequenced SARS-CoV-2 genomes, stratified by Pango lineage [2].

1.1 Load Sanger Data

The dataset is sourced from an open-access GitHub repository, ensuring accessibility and facilitating reproducibility. The file Genomes_per_week_in_England.csv provides weekly counts of SARS-CoV-2 genomic sequences from England, categorised by Pango lineage.

```
# 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 proceeding with the analysis, the raw dataset must first be processed to ensure it is appropriately structured for statistical modelling and meaningful interpretation.

The key processing steps involve:

- Formatting the dataset to ensure compatibility with downstream analytical and visualisation workflows.
- Classifying dominant SARS-CoV-2 lineages.

In particular, this analysis focuses on the set of key variants that have played significant roles in shaping the trajectory of the pandemic in the UK: Alpha (B.1.1.7), Delta (B.1.617.2), and several Omicron subvariants (BA.1, BA.2, BA.2.75, BA.4, BA.5, BA.5.3/BQ.1, and XBB) [1]. All remaining lineages — typically circulating at low prevalence with limited epidemiological impact — are grouped into a single 'Other' category. This reduces statistical noise and improves the interpretability of trends among the dominant variants.

A full breakdown of these data processing steps is implemented in the cleaning_and_curating.R script, which is sourced as part of the analysis pipeline.

```
# 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 how SARS-CoV-2 lineages have changed over time, two stacked area plots are produced to display:

- Weekly lineage counts – showing the absolute number of sequenced samples assigned to each lineage over time.
- Weekly lineage proportions – illustrating the relative frequency of each lineage as a share of the total genomes sequenced in a given week.

Together, these complementary insights into both the scale of sequencing and the shifting dominance of different variants. All visualisation functions are sourced from the plotting.R script.

```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)
```

1.3.1 Weekly Total Counts of Major Lineages

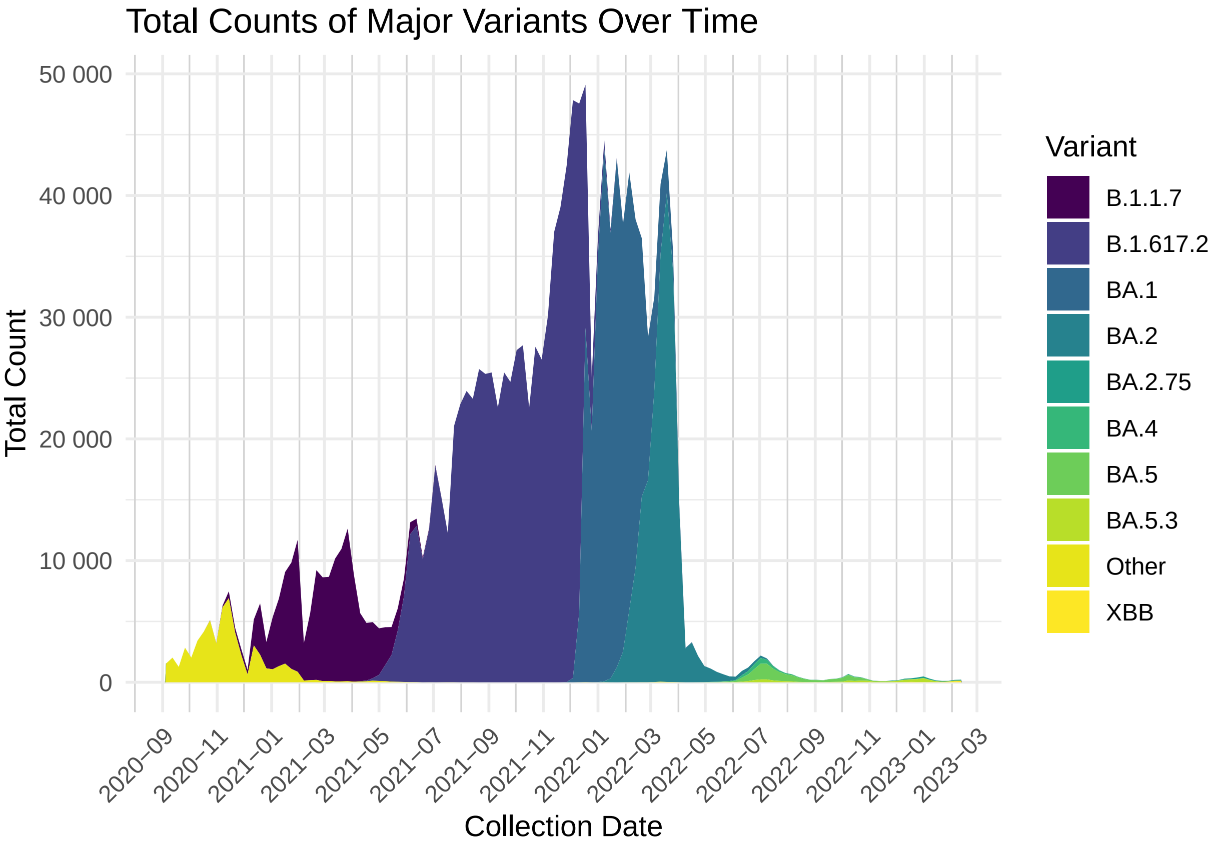

Figure 1: Stacked area plot of weekly counts of major SARS-CoV-2 variants in England over time, with each variant represented by a distinct coloured band.

As shown in Figure 1, the total counts of major SARS-CoV-2 variants in England exhibit distinct, non-overlapping peaks over time, each corresponding to a dominant variant during a particular wave of the pandemic. At the end of 2020, 'Other' variants dominate with weekly counts peaking around 8,000, before B.1.1.7 (Alpha) rapidly rises, reaching over 25,000 weekly cases by January 2021. From May 2021, B.1.617.2 (Delta) accelerates sharply, peaking at nearly 45,000 weekly cases in July 2021. Around December 2021, BA.1 (Omicron) begins to surge, followed closely by BA.2, which reaches the highest observed weekly count — just over 50,000 cases — around March 2022. Later Omicron subvariants such as BA.4, BA.5, and XBB emerge but with much smaller peaks (generally under 5,000 per week), reflecting reduced transmission, competition, or sequencing intensity. The plot shows largely non-overlapping waves, each dominated by a single variant.

1.3.2 Proportional Frequency of Major Lineages

```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)
```

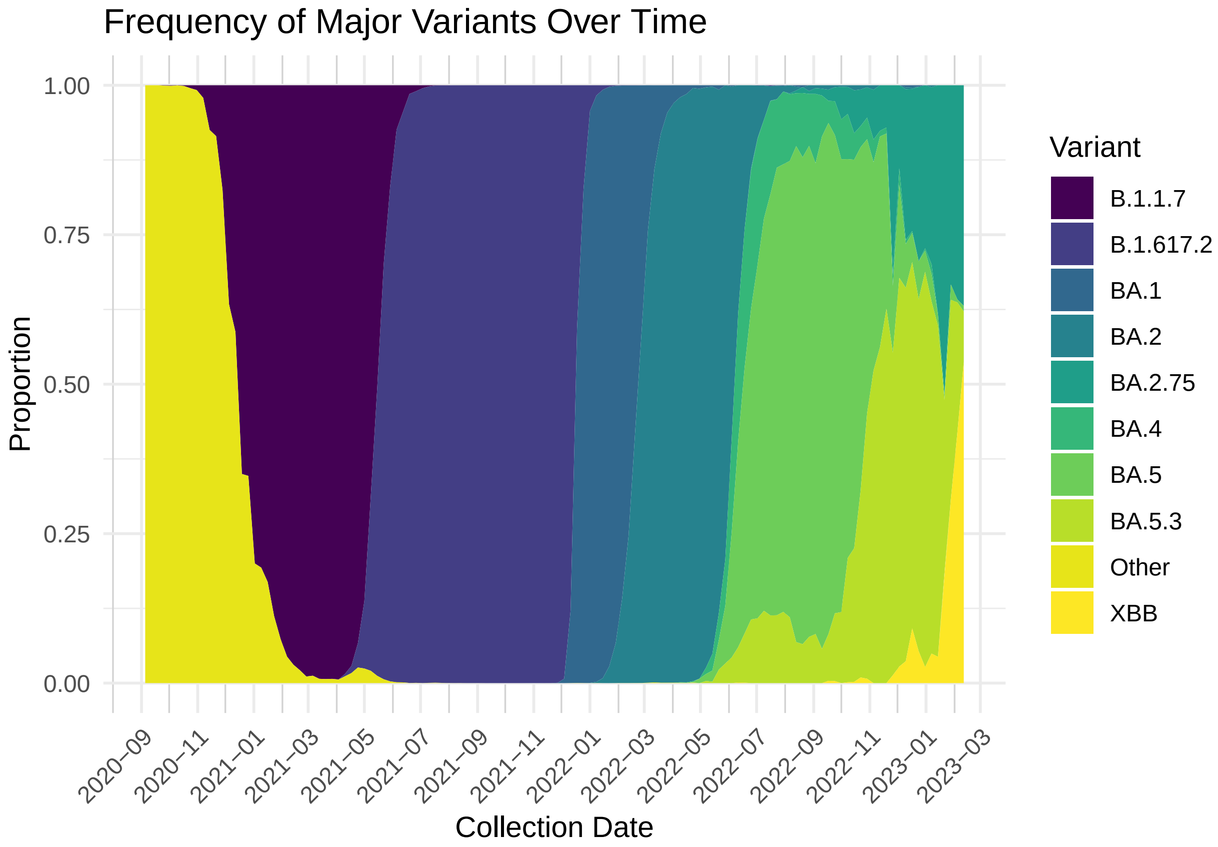

Figure 2: Stacked area plot depicting weekly shifts in the proportional distribution of major SARS-CoV-2 variants in England, with each major variant differentiated by colour.

Figure 2 illustrates the dynamic replacement of major SARS-CoV-2 variants in England over time. Each coloured band represents the relative frequency of a variant among sequenced genomes in a given week. Initially, B.1.1.7 (Alpha), which rose to dominance in late 2020 and early 2021 is shown to have displaced earlier variants, before being rapidly overtaken by B.1.617.2 (Delta) during the summer of 2021. Delta itself was displaced in late 2021 by the emergence of Omicron subvariants, beginning with BA.1, which rose sharply before being swiftly replaced by BA.2 in early 2022. This was followed by further turnover, with BA.4 and BA.5 gaining prevalence, and eventually giving way to BA.5.3 and XBB by late 2022 [1]. These transitions appear abrupt, underscoring the strong selective advantages conferred by increased transmissibility, immune escape, or both. After mid-2022, although the overall sequencing volume decreased markedly, the proportion plot still reveals ongoing shifts in variant composition, with newer Omicron sublineages continuing to circulate and occasionally dominate for short periods.

Question 2: BA.2 Variant Trajectory

To gain a clearer understanding of the dynamics behind variant emergence and dominance, this section focuses on the frequency trajectory of the SARS-CoV-2 BA.2 variant.

Two complementary genomic surveillance datasets are used: the Sanger Institute's COG-UK sequencing dataset and the Office for National Statistics COVID-19 Infection Survey (ONS-CIS). Unlike the Sanger dataset, the ONS-CIS dataset is derived from a structured and randomly sampled population survey. This makes ONS-CIS less susceptible to biases associated with testing behaviour and sequencing selection, offering a more representative snapshot of variant prevalence in the community [1].

By comparing these two sources, we can assess differences in the observed emergence, growth rate, and eventual fixation of BA.2, and evaluate how data collection methods influence our understanding of variant dynamics.

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

2.1.1 Loading Raw Data To examine the frequency dynamics of BA.2, genomic sequence data from the ONS-CIS dataset is first imported and processed for analysis. As with the previous dataset, the file lineage_data.csv is sourced from an open-access GitHub repository, ensuring transparency and reproducibility. This contains daily counts of SARS-CoV-2 sequences classified by Pango lineage.

```
# 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 Prior to plotting, the ONS-CIS dataset is processed by standardising lineage labels, grouping the data into 10-day intervals, and calculating the frequency of each lineage within these time bins. This binning smooths short-term fluctuations and allows for clearer observation of variant trends, particularly during periods of rapid change. Standardisation ensures consistency with other datasets, enabling direct comparison.

```
# 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.

```
# 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)
```

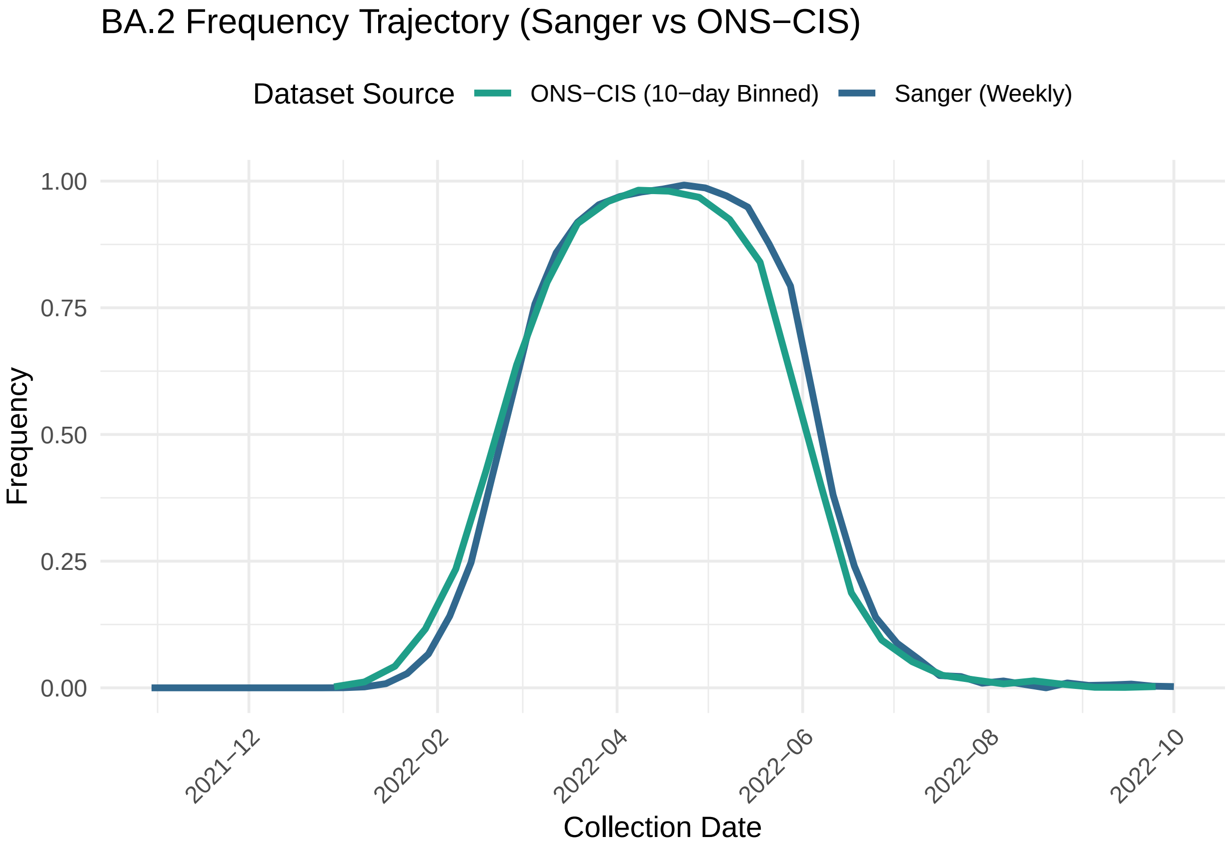

Figure 3: Line plot comparing the frequency trajectories of the BA.2 variant as observed in the Sanger (weekly) dataset and the ONS-CIS (10-day binned) dataset.

2.2 Comparing Trajectories

Figure 3 compares the frequency trajectories of the BA.2 variant using data from the Sanger (weekly) and ONS-CIS (10-day binned) datasets. While the overall patterns are highly consistent, with both datasets showing a characteristic rapid rise and decline in BA.2 frequency, there are subtle differences in timing and trajectory shape. In the Sanger dataset, BA.2 appears to begin its rise slightly earlier — around late January 2022 — compared to early February in the ONS-CIS data. This may reflect the fact that the Sanger dataset is based on community testing and may detect early signals more quickly, especially if sequencing efforts begin to focus on an emerging lineage. In contrast, the ONS-CIS dataset, which is based on a structured, randomly sampled household survey, may lag slightly in capturing the initial emergence of a variant due to its design.

The steepness of the increase also differs. The Sanger dataset shows a sharper curve, suggesting a more rapid observed expansion, whereas the ONS-CIS data presents a smoother and slightly delayed ascent. This is partly due to the 10-day binning in ONS-CIS, which reduces short-term variability but also dampens the perceived acceleration in variant spread. BA.2 reaches near-fixation (approaching 100% frequency) in both datasets at approximately the same time — around mid-April 2022. However, the decline phase differs notably. The Sanger data shows a faster and more abrupt drop, likely due to a redirection of sequencing capacity towards newer emerging variants, leading to an apparent reduction in BA.2 detection. In contrast, the ONS-CIS data shows a more gradual decline, suggesting that BA.2 continued to circulate at low levels in the population even as newer Omicron sublineages began to take over.

These differences observed between the datasets underscore the critical influence of surveillance design and data collection methods on the interpretation of genomic epidemiological trends. Although both datasets capture the overall pattern of BA.2's emergence, dominance, and eventual decline, discrepancies in timing and trajectory arise from differences in sampling strategies, data resolution, and sequencing priorities. Combining insights from both structured population surveys and routine community sequencing offers a more robust and nuanced understanding of variant dynamics—an essential foundation for tracking viral evolution and guiding effective public health interventions.

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 (s) 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.

First of all, the frequency trajectories of B.1.617.2, BA.1, and BA.2 are plotted together to visualise 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"))
```

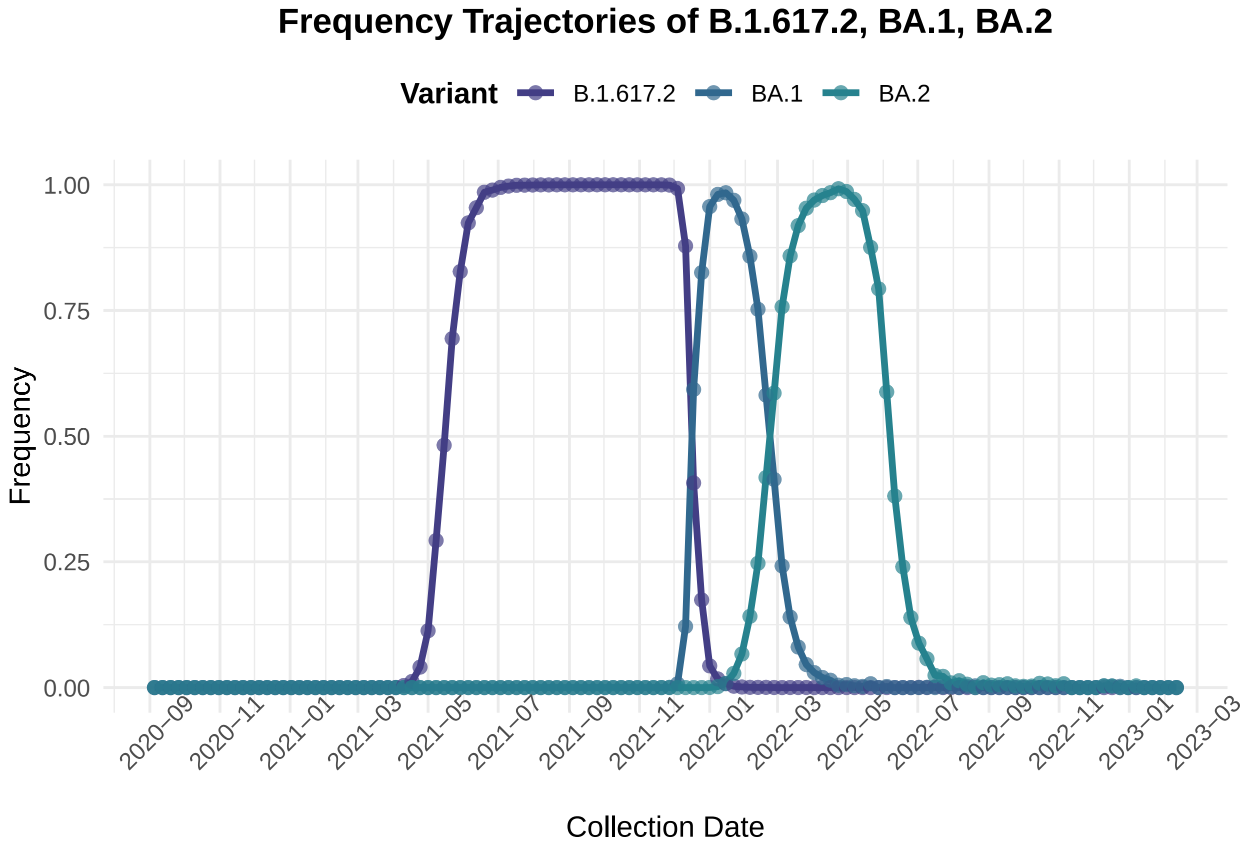

Figure 4: Line plot showing the frequency trajectories of SARS-CoV-2 variants B.1.617.2 (Delta), BA.1, and BA.2 in England over time.

Figure 4 shows that the major SARS-CoV-2 variants B.1.617.2 (Delta), BA.1, and BA.2 each followed a distinct and rapid trajectory of emergence, dominance, and replacement over time. Delta rose sharply in spring 2021, reaching near-complete dominance by June 2021 before being swiftly displaced by BA.1 (Omicron) in December 2021. BA.1 itself was quickly overtaken by BA.2 in early 2022, which became dominant by March. These transitions occurred over short time intervals and with minimal overlap, reflecting strong selective sweeps where more transmissible or immune-evasive variants rapidly outcompeted their predecessors.

A function is then 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.

```r
# 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.

```r
# 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.

```r
# 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.

```
# 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)
)
```

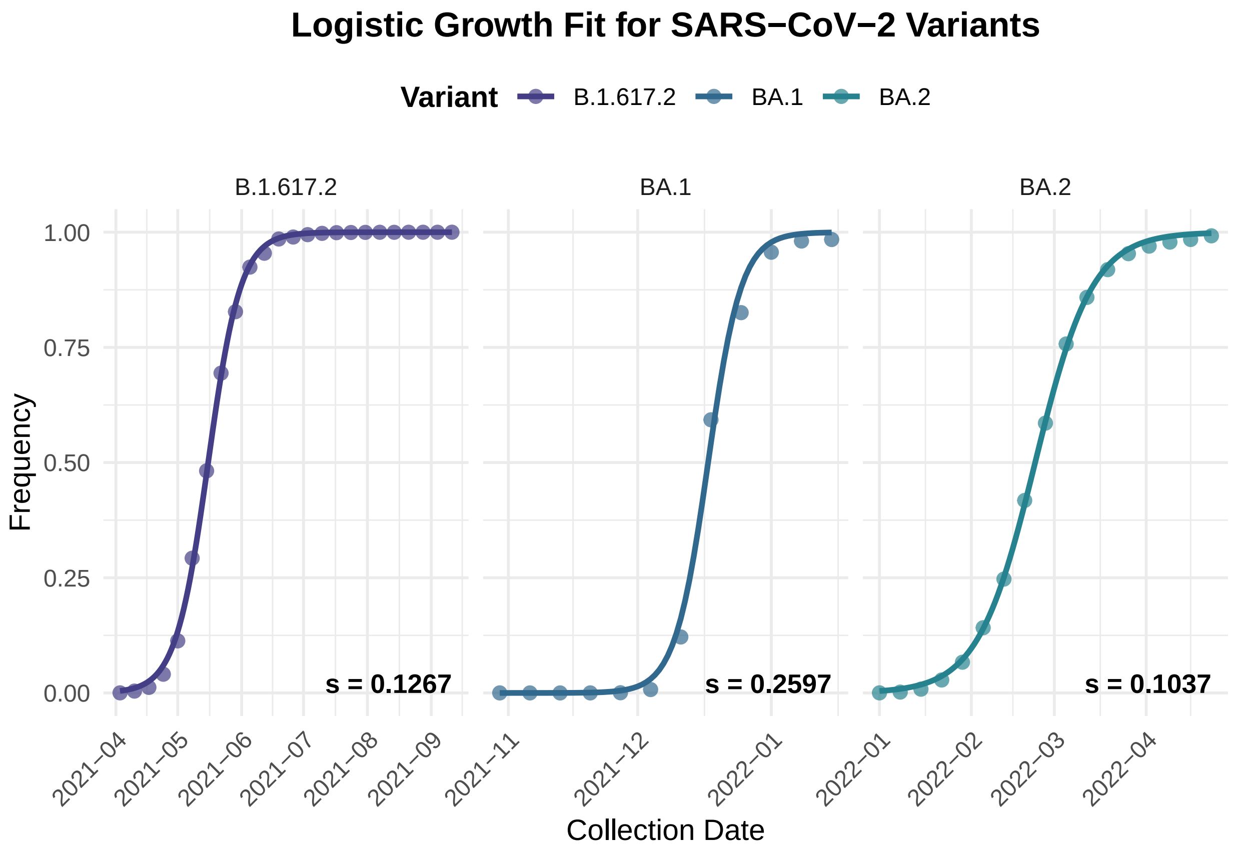

Figure 5: Logistic growth curves fitted to the frequency data of SARS-CoV-2 variants B.1.617.2 (Delta), BA.1, and BA.2 over time.

3.3 Interpretation of Variant Fixation and Selective Advantage

Figure 5 presents logistic growth model fits for the frequency trajectories of B.1.617.2 (Delta), BA.1, and BA.2, allowing for a direct comparison of how rapidly each variant rose to dominance in England. The estimated growth rate parameter (s) for each variant provides a proxy for its selective advantage—how quickly it was able to outcompete existing lineages.

Among the three, BA.1 exhibited the steepest logistic curve, with an estimated s of 0.2597, indicating the fastest rise in frequency. This suggests that BA.1 had the strongest selective advantage during its emergence, likely driven by a combination of high transmissibility and significant immune escape. It rapidly displaced Delta in late 2021 and early 2022. B.1.617.2 (Delta), with an s of 0.1267, showed a more gradual but still substantial increase during mid-2021, overtaking Alpha as the dominant lineage. Its growth reflects a significant, though comparatively lower, fitness advantage that was sufficient to drive a major wave of infections during that period. BA.2, although it ultimately replaced BA.1, had the lowest estimated growth rate of the three ($s = 0.1037$). This suggests a slower overall fixation. However, this does not necessarily imply lower transmissibility—it may reflect more complex dynamics, such as population-level immunity following the BA.1 wave, or differences in the timing and scale of seeding events.

Overall, these results reveal that while all three variants successfully outcompeted their predecessors, they did so with differing speeds and likely under different epidemiological conditions. The rapid ascent of BA.1 in particular underscores how immune escape can strongly influence variant success. Logistic modelling offers a useful framework for quantifying these dynamics and understanding how new variants gain a foothold and spread within a population.

Question 4: Regional Analysis of Delta Variant

To examine the regional spread of the Delta (B.1.617.2) variant across England, an anonymised dataset from the COG-UK sequencing programme is used. This includes individual-level sequencing results, allowing for the analysis of spatial and temporal trends in Delta variant frequency across different parts of the country.

4.1 Load and process regional delta data

The delta-ds.rds dataset is imported from an open-access GitHub repository and processed for analysis.

```r
# 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"))
```

```r
# 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 plotted.

```r
# 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)
```

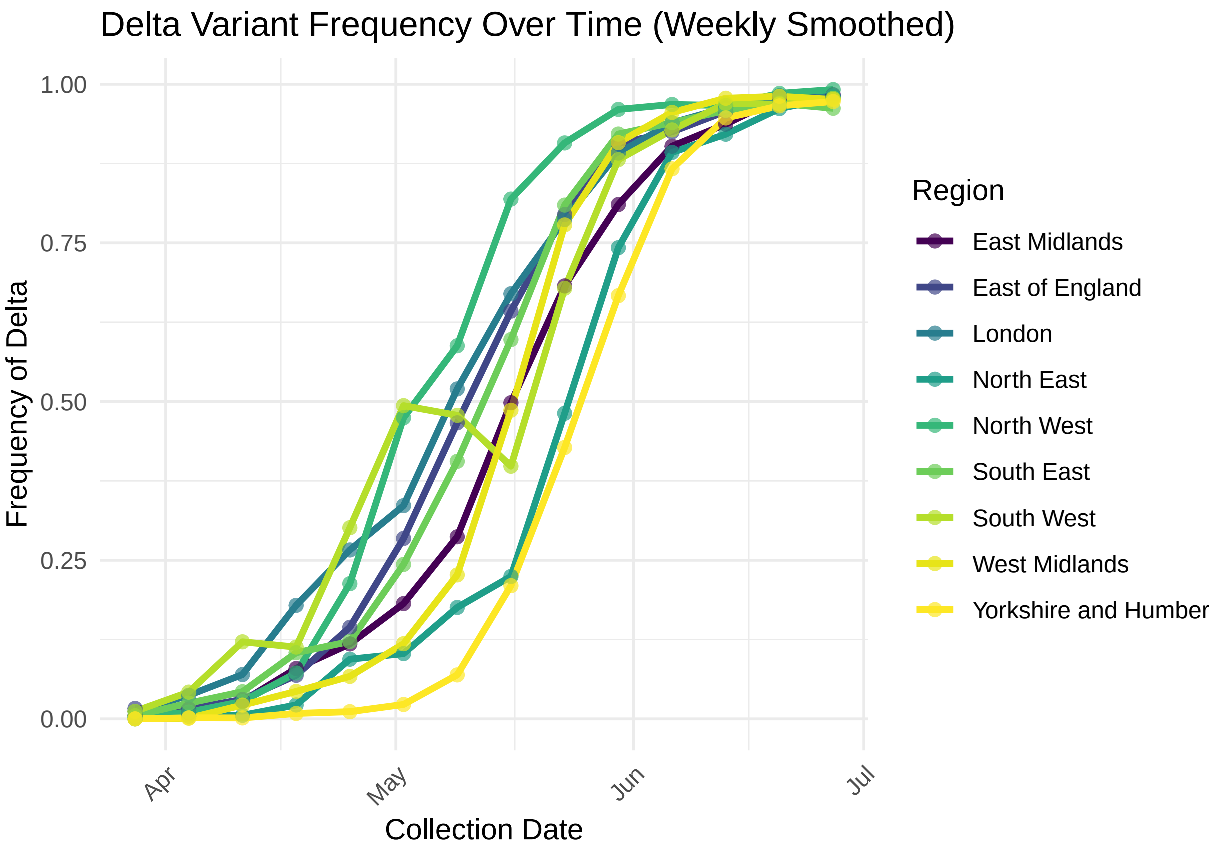

Figure 6: Weekly smoothed frequency trajectories of the Delta (B.1.617.2) variant across different regions of England.

Figure 6 displays the weekly smoothed frequency trajectories of the Delta (B.1.617.2) variant across English regions between April and July 2021. While all regions eventually reached high Delta frequencies, the timing and speed of Delta's rise varied. For instance, the North West, London, and South East saw earlier and steeper increases in Delta prevalence, indicating faster local expansion. In contrast, regions like Yorkshire and the Humber and the South West experienced later onsets and more gradual rises. These differences suggest region-specific dynamics in Delta's establishment, potentially influenced by factors such as population density, mobility, and timing of introductions.

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.

```r
# 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
)
```

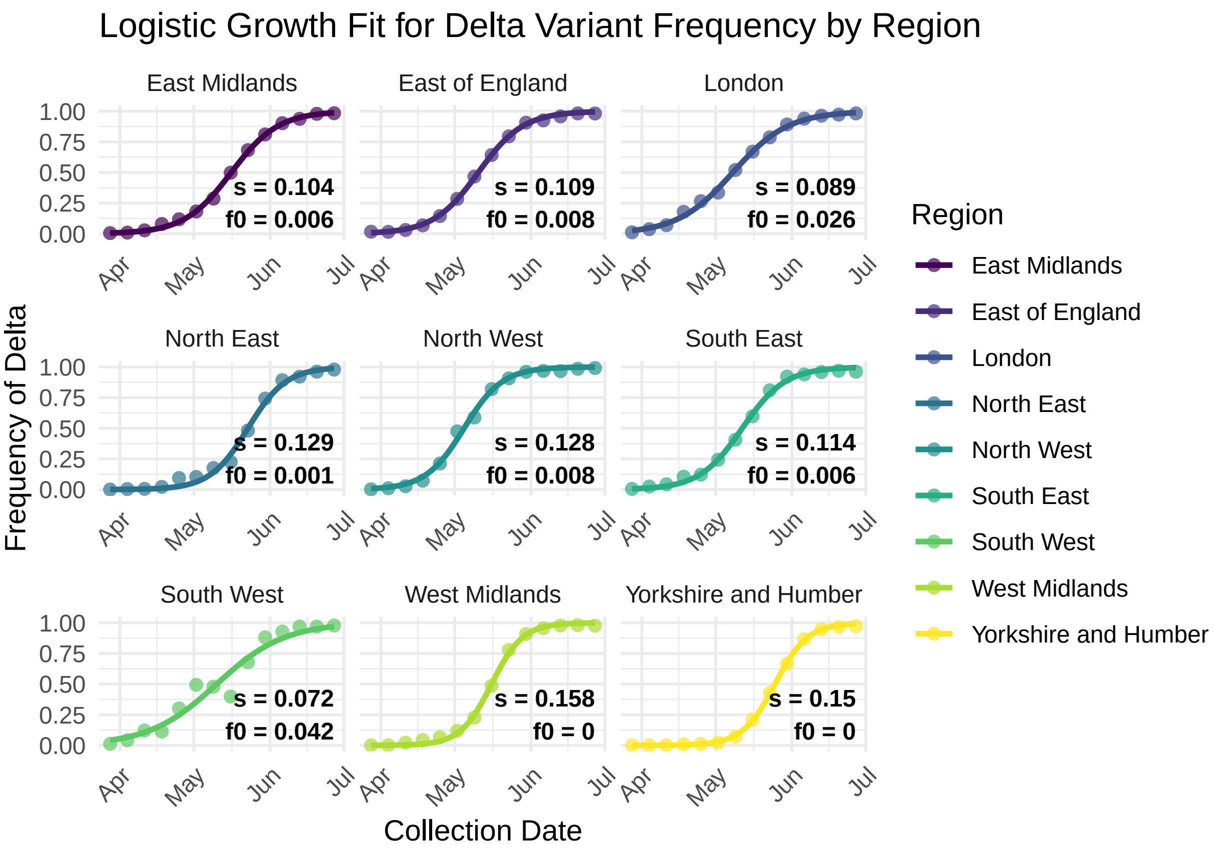

Figure 7: Logistic growth model fits for the Delta (B.1.617.2) variant frequency trajectories across English regions.

4.3 Logistic growth for each region

The logistic growth model fitted to Delta variant frequencies across English regions provides insight into how the variant spread geographically. As shown in Figure 7, the West Midlands exhibited the highest growth rate ($s = 0.158$), indicating the fastest-growing Delta outbreak. In contrast, the South West had the highest initial frequency ($f_0 = 0.042$), suggesting earlier detection or seeding of Delta in that region. These regional differences likely reflect underlying variations in population movement, exposure settings, and surveillance intensity. For example, London displayed a relatively high f_0 but a more moderate s, implying early introduction — likely facilitated by international travel and dense transport networks — but less explosive growth. The elevated f_0 may also reflect enhanced testing and genomic surveillance capacity in the capital, allowing for earlier detection of circulating variants even before widespread transmission occurred. On the other hand, the slower initial expansion may be attributed to higher early vaccination coverage, better access to testing, or heterogeneous mixing patterns that limited sustained transmission in the initial stages.

In contrast, the West Midlands, despite a low initial frequency, showed the steepest rise in Delta cases. This could reflect fewer early importation events, followed by rapid community transmission, potentially driven by lower vaccine uptake, occupational exposure in essential industries, or socioeconomic factors that limited the ability to reduce contact rates. Similarly, the South West's early detection of Delta may not have translated into rapid spread due to its lower population density, fewer transport connections, and delayed introduction into high-transmission settings, all of which could moderate the variant's initial growth despite a higher f_0.

A founder effect occurs when a new population is seeded by a small number of individuals, often leading to reduced genetic diversity and distinct, localised transmission dynamics. In the context of viral spread, this might occur if one or two early introductions spark rapid local outbreaks [3] In this analysis, some regions show patterns consistent with a founder effect—for example, Yorkshire and the Humber, where Delta exhibited a low initial frequency ($f_0 = 0$) but a high growth rate ($s = 0.15$), suggesting few early cases followed by rapid expansion. However, when considering the similarity in growth rates (s) across most regions, the broader picture points to multiple independent introductions of Delta rather than a single national founder event. While localised founder effects may have shaped early dynamics in specific areas, Delta's nationwide spread was more likely driven by a complex interplay of factors, including mobility, testing intensity, and local epidemiological conditions.

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

While the variant frequencies processed by the Sanger Institute are valuable for analysing the relative growth and replacement dynamics of Delta compared to other SARS-CoV-2 lineages, they do not provide a direct estimate of true incidence. This is because sequencing data captures only a selected subset of confirmed infections, often influenced by testing policies, logistical constraints, and regional sampling differences. In contrast, true incidence refers to the actual number of individuals infected with Delta at a given time, regardless of whether they were tested or sequenced. Many infections — particularly those that are mild or asymptomatic — may go undetected. 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 [1]. 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-c

# 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)
```

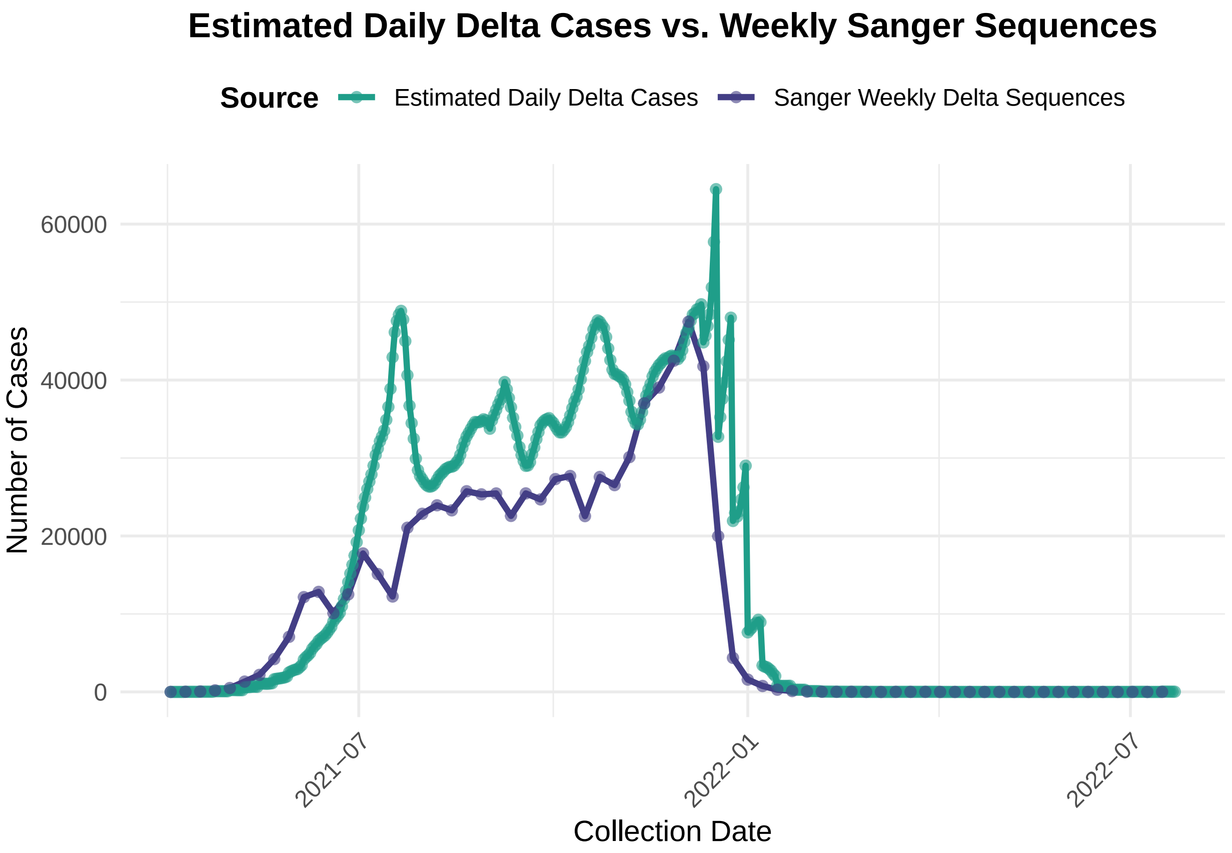

Figure 8: Line plot comparing estimated daily Delta case counts with the number of Delta sequences reported weekly by the Sanger Institute.

5.2 Reflection on Differences Between the Two Estimates

Figure 8 compares two approaches to estimating Delta incidence: the estimated daily Delta cases, derived by scaling population-level case data using variant frequencies, and the weekly Delta sequences obtained from the Sanger Institute's genomic surveillance. While both capture the broad trajectory of Delta's spread — including its rapid rise in mid-2021, sustained high prevalence into late 2021, and subsequent decline — there are noticeable differences in their magnitude and smoothness. The estimated daily Delta cases show higher counts throughout the peak period, often exceeding 60,000 cases per day, whereas the Sanger-based sequencing data peaks around 40,000. This discrepancy likely arises due to sampling bias since sequencing only covers a subset of PCR-positive cases, whereas the estimated data reflect all reported infections. As such, the sequencing data likely underrepresents the total burden of infections.

Another major difference lies in the temporal resolution. The estimated cases follow a smoother, more continuous curve due to their daily granularity and 7-day averaging, while the Sanger sequences exhibit more fluctuation, reflecting the weekly resolution and variability in sequencing throughput. The stepped nature of the Sanger line also illustrates processing and batching effects, where weekly data may obscure more subtle daily fluctuations. Reporting delays also play a role. Sequencing results lag behind testing, so the Sanger curve may be shifted slightly in time compared to real-world transmission dynamics, whereas the estimated cases offer a timelier approximation.

Despite these differences, the overall trend alignment between the two datasets supports the validity of the scaling approach and illustrates the value of combining genomic surveillance with broader testing data. Together, these methods provide complementary insights into Delta's spread and reinforce the importance of integrating multiple data streams for accurate epidemic monitoring.

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.

This analysis uses the EpiEstim R package [4], which is specifically designed to estimate the instantaneous reproduction number during an epidemic. It applies a Bayesian framework to infer R_t from incidence data and the distribution of the serial interval (the time between successive cases in a transmission chain). In this context, EpiEstim is used to compute R_t values for Delta over time, based on the estimated incidence curve. By specifying the serial interval parameters and appropriate time windows, the package provides both point estimates and credible intervals for R_t. These estimates help determine whether the epidemic was growing ($R_t > 1$) or declining ($R_t < 1$) at any given time, offering crucial insights into transmission dynamics and the effectiveness of interventions.

For more details and examples of use, the official documentation and vignettes are available at: https://mrc-ide.github.io/EpiEstim/

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

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

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

# Plot Rt estimates
plot_rt_estimates(rt_estimates)
```

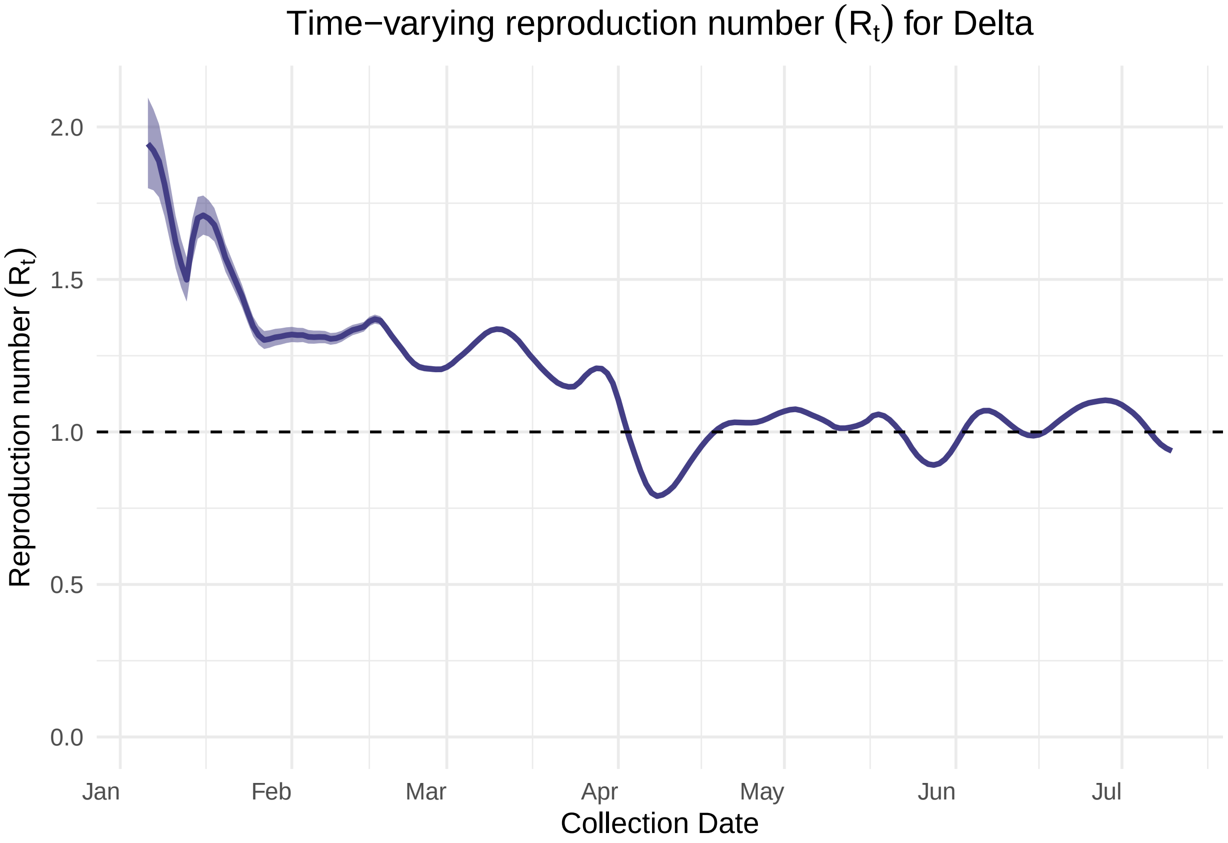

Figure 9: Time series plot of the estimated time-varying reproduction number (R_t) for Delta, calculated using inferred daily case counts.

5.4 Comparison of R_t Estimates

The estimated time-varying reproduction number (R_t) values derived from Sanger sequencing data and ONS-CIS population data are broadly consistent, with both indicating a high transmission rate during Delta's early spread. The Sanger-based estimate is calculated at $R_t = 1.944 \pm 0.075$, while the ONS-CIS estimate is $R_t = 1.915 \pm 0.279$. Although their mean values are nearly identical, the key difference lies in precision: the Sanger estimate has a much narrower standard deviation, while the ONS-CIS estimate shows greater uncertainty due to its broader sampling approach. This difference is clearly reflected in the plots. The Sanger-derived R_t (Figure 9) displays a smooth curve with tight confidence intervals, suggesting precise, high-resolution insight into the dynamics of Delta transmission. In contrast, the ONS-CIS plot (not shown here) exhibits a wider shaded region, reflecting greater variability and uncertainty in its estimates. However, the Sanger R_t lies entirely within the confidence interval of the ONS-CIS estimate (1.636–2.194), suggesting that the two estimates are not significantly different from a statistical perspective.

Each data source carries its own strengths and limitations. The Sanger dataset offers variant-specific precision, making it particularly valuable for tracking the behaviour of individual lineages like Delta. However, it is limited to PCR-positive cases selected for sequencing, introducing testing and selection biases — particularly underrepresenting asymptomatic or untested infections. In contrast, the ONS-CIS dataset is designed for population-level surveillance through random sampling, capturing both symptomatic and asymptomatic cases. This makes it more representative of overall transmission, albeit with reduced precision due to smaller sample sizes and less frequent data collection. Both approaches estimate an initial R_t close to 1.9, highlighting Delta's strong transmissibility and potential for exponential spread. Together, they tell a more complete story: while the Sanger data delivers granular, lineage-specific trends, the ONS-CIS survey offers broader, unbiased population-level insights. Integrating these complementary sources enhances the reliability of R_t estimation and supports more informed public health decisions.

References

[1] Lythgoe, K.A., Golubchik, T., Hall, M., House, T., Cahuantzi, R., MacIntyre-Cockett, G., Fryer, H., Thomson, L., Nurtay, A., Ghafani, M., Buck, D., Green, A., Trebes, A., Piazza, P., Lonie, L.J., Studley, R., Rourke, E., Smith, D., Bashton, M. and Nelson, A.

(2023). Lineage replacement and evolution captured by 3 years of the United Kingdom Coronavirus (COVID-19) Infection Survey. Proceedings. Biological Sciences, [online] 290(2009), p.20231284. doi:https://doi.org/10.1098/rspb.2023.1284.

[2] The COVID-19 Genomics UK (COG-UK) consortium (2020). An integrated national scale SARS-CoV-2 genomic surveillance network. The Lancet Microbe, 1(3), pp.e99–e100. doi:https://doi.org/10.1016/s2666-5247(20)30054-9.

[3] Ruan, Y., Luo, Z., Tang, X., Li, G., Wen, H., He, X., Lu, X., Lu, J. and Wu, C.-I. (2020). On the founder effect in COVID-19 outbreaks – How many infected travelers may have started them all? National Science Review. doi:https://doi.org/10.1093/nsr/nwaa246.

[4] mrc-ide (2020). GitHub - mrc-ide/EpiEstim: A tool to estimate time varying instantaneous reproduction number during epidemics. [online] GitHub. Available at: https://github.com/mrc-ide/EpiEstim [Accessed 12 Mar. 2025].