Computing Assignment – measuring growth rate and reproduction number for SARS-CoV-2 variants

January 2025

Mahan Ghafari – [mahan.ghafari@ndm.ox.ac.uk](mailto:mahan.ghafari@ndm.ox.ac.uk)

Demonstrators: Punya Alahakoon – [punya.alahakoon@ndm.ox.ac.uk](mailto:punya.alahakoon@ndm.ox.ac.uk)  
 Charlie Hamilton – [charlie.hamilton@linacre.ox.ac.uk](mailto:charlie.hamilton@linacre.ox.ac.uk)

**Introduction**

Following on from the practical, this assignment will involve investigating the COVID-19 Genomics UK Consortium (COG-UK) genomic sequence data for England, which is [collected by the Wellcome Sanger Institute](https://covid19.sanger.ac.uk/downloads).

Unlike the ONS-CIS dataset used in the practical, this dataset includes samples from Lighthouse Labs, covering most of the Pillar 2 testing in England. These genomes aim to reflect the population undergoing PCR tests in England, which may vary in both **coverage** and **geographical representation**.

In this assignment, you will analyse the growth rate and reproduction number of SARS-CoV-2 variants using COG-UK data and reflect on potential reasons for differences in results compared to the ONS-CIS analysis from the practical. This script makes no assumptions about your level of R programming expertise, nor biological knowledge on specialised courses in year 3 in your curriculum.

You will also need to download the following files from Canvas to complete this assignment:

* Genomes\_per\_week\_in\_England.csv
* delta-d2.rds
* daily-new-confirmed-covid-19-cases.csv

**Assignment**

Question 1

* 1. Download the dataset: download Genomes\_per\_week\_in\_England.csv from Canvas. This includes *weekly counts* of virus samples per lineage over time across England collected as part of Sanger Institute COG-UK.
  2. Classify major lineages: identify the following variants as major lineages: B.1.1.7 (Alpha), B.1.617.2 (Delta), BA.1, BA.2, BA.2.75, BA.4, BA.5, BA.5.3 (BQ.1), and XBB. Group all other lineages into a single category labelled as Other.
  3. Visualise the data: generate a **stacked area plot** showing the **total counts** of each major lineage over time. Generate another **stacked area plot** showing the **frequencies** (proportions) of each major lineage over time.

Question 2

1. Visualise the COG-UK and ONS-CIS data for BA.2: plot the frequency trajectory for the BA.2 variant using both the **Sanger dataset** (weekly counts) and the **ONS-CIS dataset** (10-day bin counts from the practical).
2. Analysis: compare the two trajectories. Is there a difference in the timing of BA.2’s rise and when it reaches fixation? Reflect on potential reasons for these differences (sampling strategies and geographical or temporal biases in data collection)?

Question 3

Using the Sanger dataset, determine which variant—B.1.617.2, BA.1, or BA.2—reached fixation the fastest and exhibited the highest selective advantage under a logistic growth model. Use weekly counts to measure the selective advantage (𝑠).

Question 4

1. **Load the dataset:** download and load delta-d2.rds from Canvas. This dataset contains an anonymised line list of individuals with a sequenced Delta sample from various regions in England, collected as part of the COG-UK. (hint: remove rows for which there is no associated phecname subset(data, phecname != ""))
2. Analyse and visualise the data: (i) **plot Delta frequencies by region**; for each region, plot the frequency of Delta over time. Use distinct colours or facets to differentiate between regions. (ii) **Fit logistic growth for each region**; for each region, fit a logistic growth model to the frequency data. Overlay the logistic growth curves onto the frequency trajectories for each region.
3. Interpretation: (i) **Identify the region with the fastest Delta outbreak**; based on your analysis, determine which region had the most rapidly growing outbreak of Delta (highest 𝑠). **Identify the region where Delta had the earliest rise in frequencies** (highest 𝑓0). Discuss why the timing of Delta’s emergence differed between regions. (ii) Could Delta’s growth across regions be **associated with a founder effect**? Explain what a founder effect is and evaluate whether the observed data supports or refutes this hypothesis.

Question 5

1. **Estimate the true incidence of Delta**: Up to this point, we have relied on the number of PCR-positive tests sent to the Sanger Institute for sequencing to estimate the growth rate of variants. However, does this approach accurately reflect the true incidence of Delta infections in England? Explain how it is different from the incidence.  
     
   To obtain a more representative estimate of the true number of Delta infections, multiply the proportion of Delta sequences in England (from the Sanger dataset) by the daily (7-day averaged) COVID-19 case counts in England provided in the daily-new-confirmed-covid-19-cases.csv dataset on Canvas.  
     
   Note that while the daily COVID-19 case counts are reported on a daily basis, the proportion of Delta sequences from the Sanger dataset is calculated weekly. For this task, use the same weekly proportion of Delta for every day within each 7-day interval of the daily case counts.  
     
   Plot estimated daily cases of Delta and weekly number of Delta sequences from Sanger. Reflect on why the two counts are different from each other.
2. **Measure 𝑅𝑡:** Using the estimated daily Delta case counts, calculate the time-varying reproduction number (𝑅𝑡) using the same method as in the practical (hint: use this time range: “2021-04-23” to “2021-11-01” for measurement).   
     
   Compare your 𝑅𝑡 estimate to the one calculated during the practical using the ONS-CIS dataset.  
     
   Reflect on whether the two estimates differ significantly. Which estimate do you consider more reliable, and why? (hint: consider factors such as the sampling strategy used by the different datasets and the representativeness of the sequencing data for actual infections.)

**Submission requirements**

Write your report using R Markdown and submit both your .Rmd file containing the code and the .html produced by knitting the document. Anyone unsure about how to produce a report using R Markdown should run through the material on canvas from last term.

Aim to write your code so that it would run for a new pair of species simply by changing the input data at the start, i.e. using general rather than specific coding.

*Please note:* you should not be collaborating on your assignments, and the examiners will check code for evidence of plagiarism.

The marking scheme for computer assignments will be weighted equally between how well your code runs, how explicitly commented it is, the quality of the presentation of the report, and the understanding shown in your answers.