Computing Skills HT 2025: SARS-CoV-2 emergence in humans and its introduction into the UK

March 2025

Mahan Ghafari – mahan.ghafari@ndm.ox.ac.uk

Introduction

SARS-CoV-2, the virus responsible for COVID-19, emerged in late 2019 and rapidly spread worldwide, triggering a pandemic. Understanding its origins and early transmission dynamics is critical for reconstructing how the virus spread between countries and identifying key mutations that shaped its evolutionary trajectory. Phylogenetic methods allow researchers to estimate the time to the most recent common ancestor (TMRCA) of viral sequences, infer introduction events into specific regions, and track the emergence of key mutations that allow the virus to acquire increased transmissibility and immune evasion properties.

In this practical, you will analyse a dataset of SARS-CoV-2 genome sequences sampled between late December 2019 and early May 2020 from around the world. Using Bayesian phylogenetics, you will estimate the virus's emergence in humans, track its early introductions into the UK, and explore how certain mutations—including a Spike protein mutation S:D614G—became widespread. You will also examine the evolutionary divergence between SARS-CoV-2 and its closest known bat coronavirus relatives. This hands-on approach will introduce key bioinformatics tools and methods for studying pathogen evolution.

By the end of this practical, you will:

1. Build a phylogenetic tree using SARS-CoV-2 sequences and estimate the time to the most recent common ancestor.
2. Identify the earliest introductions of SARS-CoV-2 into the UK.
3. Analyse SARS-CoV-2 mutation dynamics.
4. Use Nextclade to verify lineage-defining mutations.
5. Identify closest ancestors of SARS-CoV-2 in animal hosts.

Part 1: Constructing a phylogenetic tree of SARS-CoV-2 collected from humans

You are provided with 248 SARS-CoV-2 genome sequences sampled between late December 2019 and early May 2020 (sarscov2_practical.fasta).

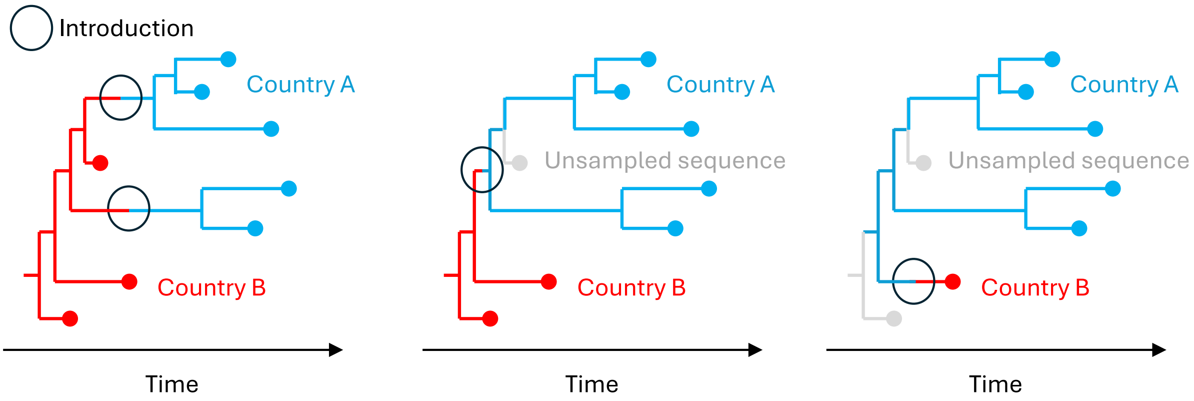
Task 1: Build a phylogenetic tree using Delphy

1. Navigate to Delphy and upload the provided FASTA file.
2. Use default settings for tree inference and wait for convergence to be achieved (minimum effective sample size >100). What is the estimated evolutionary rate of SARS-CoV-2? What is the expected number of mutations that SARS-CoV-2 accumulates in one month across its entire genome (~30,000 bps long)?
3. Navigate to the "Lineages" tab and report the TMRCA of the SARS-CoV-2 sequences. What does this suggest about the timing of the first human infections? When did the first two big clades of SARS-CoV-2 emerge?

Part 2: Identifying early introductions into the UK

This dataset includes sequences from countries around the world, with a subset from England, Northern Ireland, Scotland, and Wales.

When inferring introduction events, it's important to acknowledge that sampling can limit or even bias our inferences!



Task 2: Trace the introduction events

Navigate to the "Lineages" tab on Delphy.

1. Identify the four Welsh sequences in the tree.
2. Is there evidence of multiple independent introductions of SARS-CoV-2 into Wales?
3. Can you identify transmission clusters involving multiple nations from the UK?
4. What is the TMRCA of all four Welsh sequences?

Part 3: Investigating mutation dynamics

Delphy also reconstructs mutations along the SARS-CoV-2 phylogeny and estimates their prevalence over time. This allows researchers to track how certain mutations rise in frequency and potentially contribute to the virus's evolutionary success.

Mutations in SARS-CoV-2 are reported using a standardised notation that describes nucleotide changes in the genome. For example, a mutation written as A23403G means that at genomic position 23,403, the ancestral nucleotide A (adenine) has changed to G (guanine). Since SARS-CoV-2 is an RNA virus, these nucleotide changes can sometimes lead to amino acid substitutions in viral proteins. Protein-level mutations are described using a different notation: S:D614G refers to an amino acid change in the spike (S) protein, where the original aspartic acid (D) at position 614 has mutated to glycine (G). This type of notation follows the format: Protein: Original_AA Position Mutated_AA.

<u>Task 3: Identify key mutations that rose in frequency</u>

1. Identify mutations that rose to a frequency of >50% by May 2020.
2. Can you identify any mutation that rose in frequency in the same way as A23403G? What is the significance of having mutations rising in frequency? Why would some mutations be rising in frequency together?

Part 4: Identify lineage-defining mutations in Nextclade

As viruses evolve, researchers classify them into distinct lineages based on their genetic similarities and shared mutations. The Pango lineage classification system is a widely used framework for naming and tracking SARS-CoV-2 variants. Each Pango lineage is defined by a set of characteristic lineage-defining mutations, which distinguish it from other lineages and help researchers trace viral evolution and transmission patterns.

Nextclade is a bioinformatics tool that automatically assigns SARS-CoV-2 sequences to a Pango lineage, detects mutations, and assesses their potential effects.

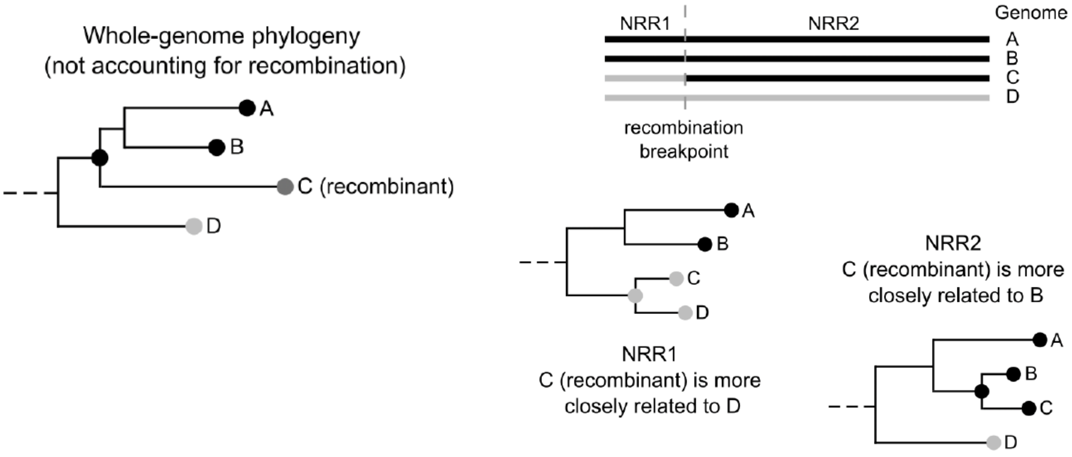
<u>Task 4: Verify mutations in Nextclade</u>

1. Upload the provided sequences into Nextclade.
2. Locate the A23403G mutation and confirm its effect on the spike protein. Does this mutation lead to an amino acid change, and if so, which codon position is affected?
3. Does the presence of S:D614G correlate with a specific Pango lineage (or clade)? Can you identify which lineages existed before those carrying the S:D614G mutation?

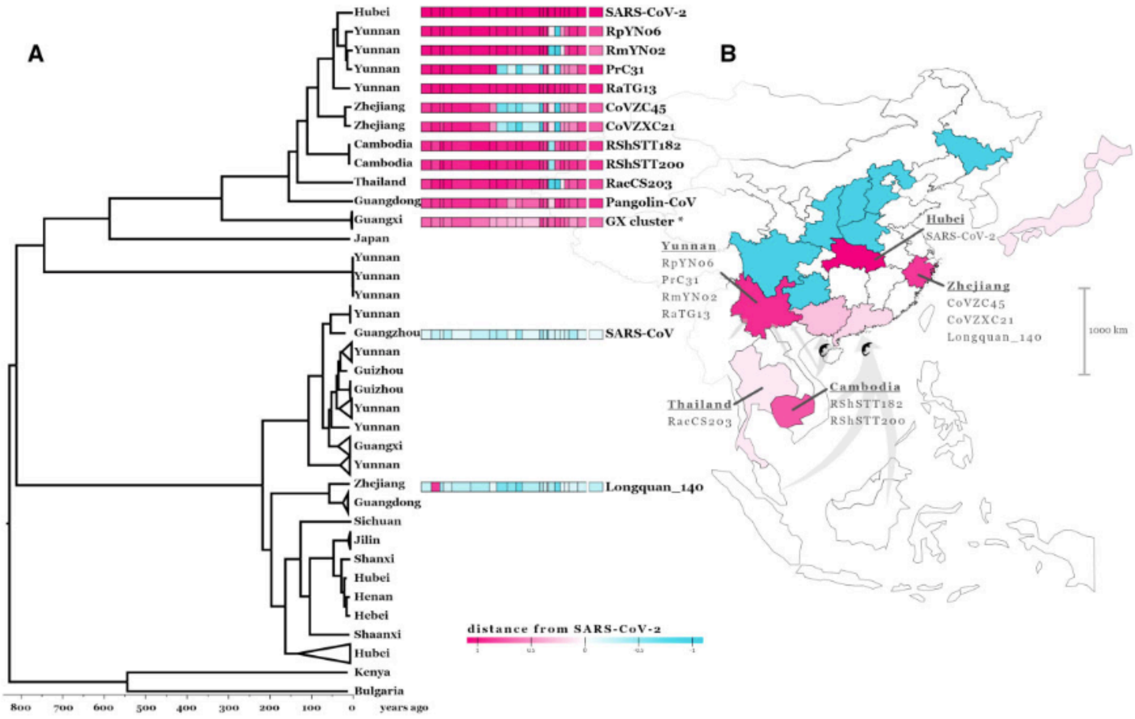
Part 5: Closest ancestors of SARS-CoV-2 in animals

Bats are considered the primary reservoir hosts of coronaviruses, with SARS-related coronaviruses (SARSr-CoVs) circulating in multiple bat species. However, other animals, such as raccoon dogs and badgers, have also been suggested as potential intermediate hosts that may have facilitated the spillover of SARS-CoV-2 into humans. Understanding the evolutionary history of SARS-CoV-2 requires comparing its genome to closely related viruses found in wildlife.

A key challenge in tracing viral ancestry is that coronaviruses frequently undergo recombination, meaning that different regions of the genome may have distinct evolutionary histories. As a result, a single genome-wide phylogeny may not fully capture the complex origins of SARS-CoV-2. By analysing specific non-recombinant genomic regions, we can estimate when SARS-CoV-2 and its closest known bat relatives last diverged.



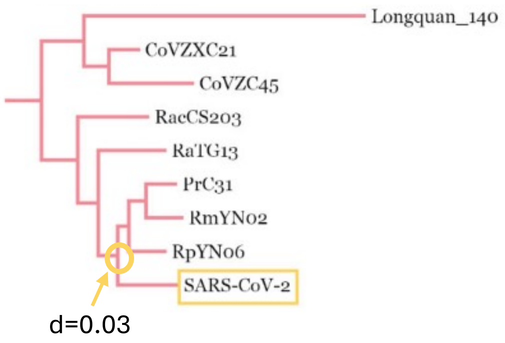
Sarbecoviruses (the subgenus that includes SARS-CoV-2 and SARS-CoV) have at least 21 identifiable recombination breakpoints, dividing the genome into 22 non-recombinant regions (Lytras et al., GBE 2022). The closest known viral relatives of SARS-CoV-2 have been found in horseshoe bats in Yunnan and Zhejiang, China.



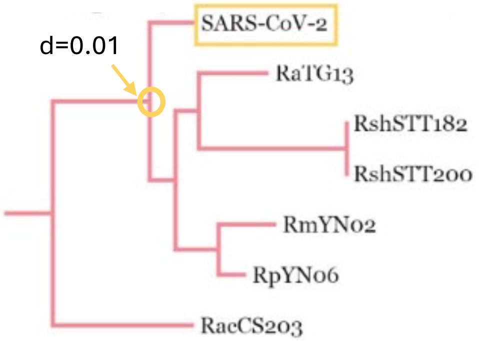
Task 5: Estimating TMRCA between SARS-CoV-2 and bat coronaviruses

1. Taking phylogenetic distance trees from two non-recombinant regions shown below (identified in Lytras et al 2022), what is the estimated TMRCA of SARS-CoV-2 and bat coronaviruses? (hint: assume an evolutionary rate of $5.5x10^{-4}$ s/s/y for all coronaviruses; genetic distance = evolutionary rate x time from root to tip)

non-recombinant region 2



non-recombinant region 9



2. What is the implication of having two different TMRCA estimates for SARS-CoV-2 closest bat ancestors?