CODECHECK certificate 2025-006

Item	Value
Title	Emergence of novel SARS-CoV-2 variants in the Netherlands
Authors	Aysun Urhan (0000-0001-8584-4736) and Thomas Abeel (0000-0002-7205-7431)
Publication	https://doi.org/10.1038/s41598-021-85363-7
Codecheckers	Elviss Dvinskis (0000-0003-2873-0139), Leila Iñigo de la Cruz (0000-0003-0852-9219) and Yasel Quintero (0009-0005-9240-7370)
Date of check	2025-05-01
Summary	Partial reproduction.
Codecheck repository	https://github.com/codecheckers/sarscov2-variants

Table 1: CODECHECK summary

Summary

This project provides clear instructions on generating figures 1 – 9, described in the README.md document located in the GitHub repository linked to the publication. However, to run figures 8 and 9, the author recommends the use of an HPC cluster, which is beyond the scope of this CODECHECK. Figures 1 – 7 could be reproduced on certain systems and configurations.

Output	Comment
fig_1_Distribution-of-SARS-CoV-2-genomes-across-five-continents.png	manuscript Figure 1
fig_2_Distribution-of-SARS-CoV-2-clades-in-a-selection-among-the-12-most-sampled-countries.png	manuscript Figure 2
fig_3_Number-of-mutations-per-sample-per-day-over-the-course-of-pandemic-in-the-Netherlands.png	manuscript Figure 3
fig_4_Total-number-of-nucleotide-mutations.png	manuscript Figure 4
fig_5_Total-frequency-of-the-top-15-mutations-in-the-most-sampled-countries-in-our-dataset.png	manuscript Figure 5
fig_6_Change-in-frequency-of-the-top-15-mutations-in-the-most-sampled-countries-in-our-dataset.png	manuscript Figure 6
fig_7_Change-in-frequency-of-the-top-15-mutations-in-the-Netherlands.png	manuscript Figure 7

Table 2: Summary of output files generated

CODECHECKER notes

The author provided all the necessary code, data and documentation to easily reproduce the figures in the publication. When using a setup similar to what was used by the author, reproducing the figures was a straightforward process.

Installation prerequisites and computational environment

The code requires a conda package manager (Anaconda or miniconda) for the computational environment.

To process the data to generate figures 8 and 9, use of an HPC cluster is required. Due to the runtime needed, this was out-of-scope for this CODECHECK.

The required dependencies are listed in the `environment.yml` file. While most dependencies are not fixed to exact versions, Python is specifically fixed to version 3.7. The dependencies are also stated in the README file.

Data preparation

The data can be downloaded from https://data.4tu.nl/datasets/11bff1ea-4784-463e-90d0-eb2e2b64fe96. It takes ~1GB of space. The access link to the data can be found in the README.

Running the code

The `environment.yml` file specifies that Python 3.7 must be installed in the user's environment to run the code. While this is not an issue for Linux, Windows, and Intel-based Macs, it cannot be installed natively on M series Macs. This is because Python 3.7 was released before Apple Silicon processors and lacks native ARM support, limiting reproducibility. The author Aysun Urhan was contacted regarding this issue. The `environment.yml` file was updated and the Python version was set to 3.8. However, this change introduced dependency issues[1].

The code checkers recommend to the author to either specify that Mac systems with an M series chip are not supported, or address this at a higher level in the codebase.

The plots can be reproduced seamlessly on a Linux machine within a Python 3.7 environment. However, on a Windows machine reproducing the figures is not possible because the mafft package required is not available for Windows platforms.

---

[1] ` TypeError: sum() got an unexpected keyword argument 'level'`. This dependency issue is likely raised because `pandas` version 1.5.3 was installed while setting up the environment, and the `level` parameter was removed from `DataFrame.sum()` in `pandas` 2.0 (released April 2023). The code appears to require `pandas` ≥2.0 syntax.

Outputs

A minor difference can be observed between figure 4 in the publication and the one reproduced by the code checkers. The values on the x-axis in the reproduced figure are multiplied by a factor of three, compared to the published figure. Apart from this discrepancy, all other aspects of the plots coincide. The author noted that the difference is due to the code plotting nucleotide positions whereas the figure in the publication represents amino acid positions.
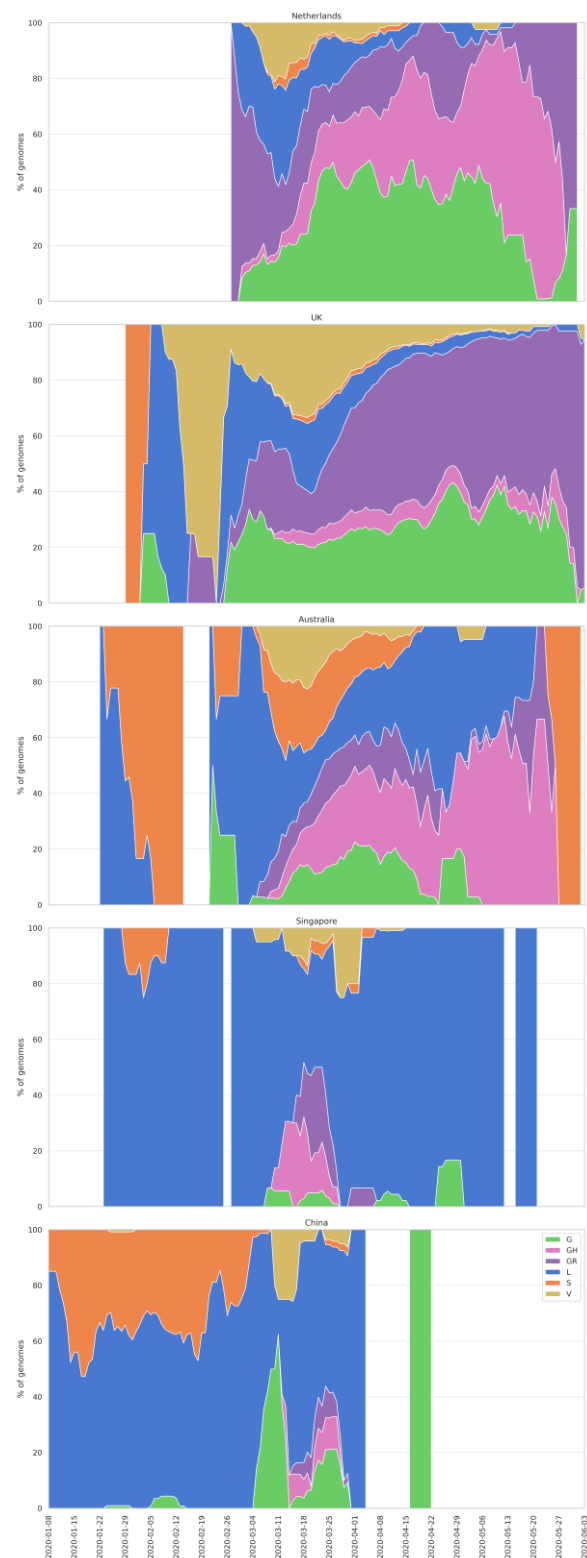
Manifest Files

fig_1_Distribution-of-SARS-CoV-2-genomes-across-five-continents.png

Comment: Figure 1 in the manuscript.
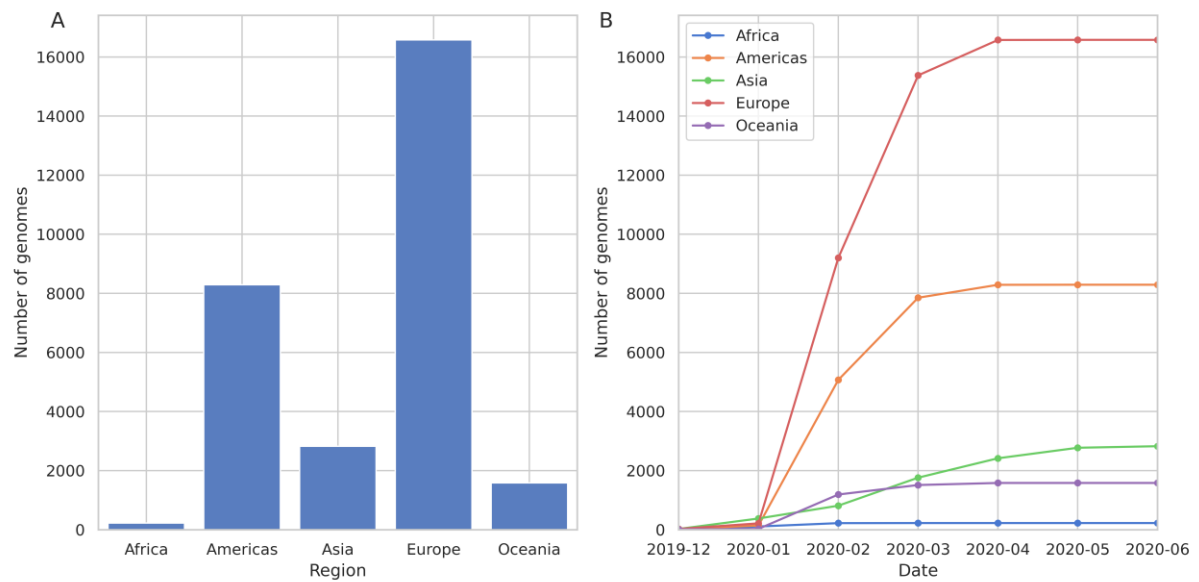
fig_2_Distribution-of-SARS-CoV-2-clades-in-a-selection-among-the-12-most-sampled-countries.png
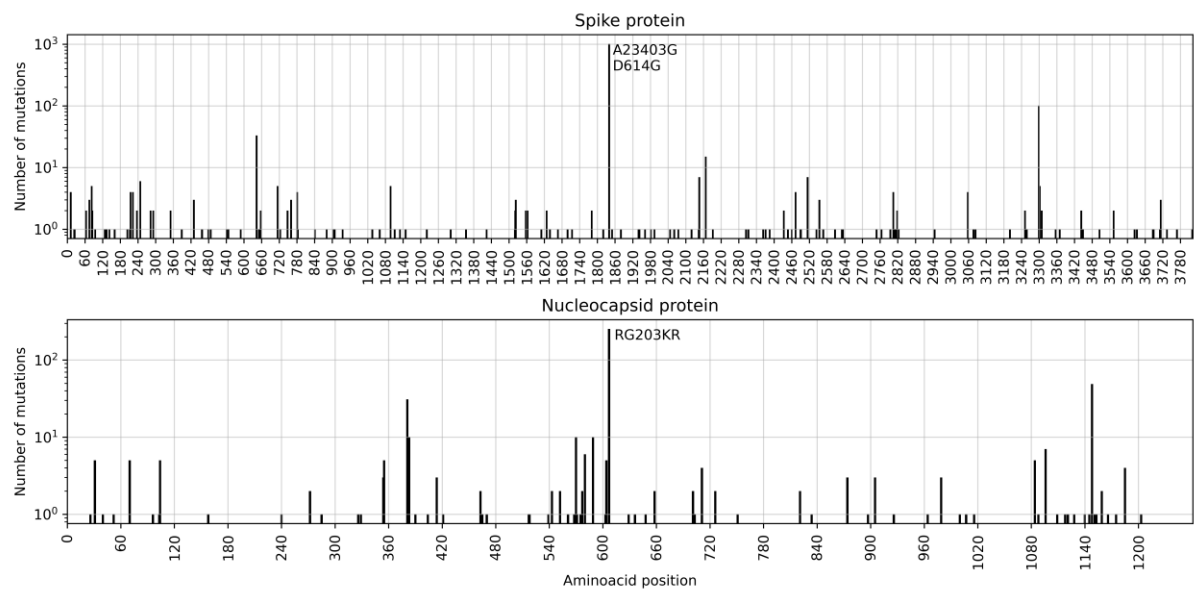
Comment: Figure 2 in the manuscript

fig_3_Number-of-mutations-per-sample-per-day-over-the-course-of-pandemic-in-the-Netherlands.png
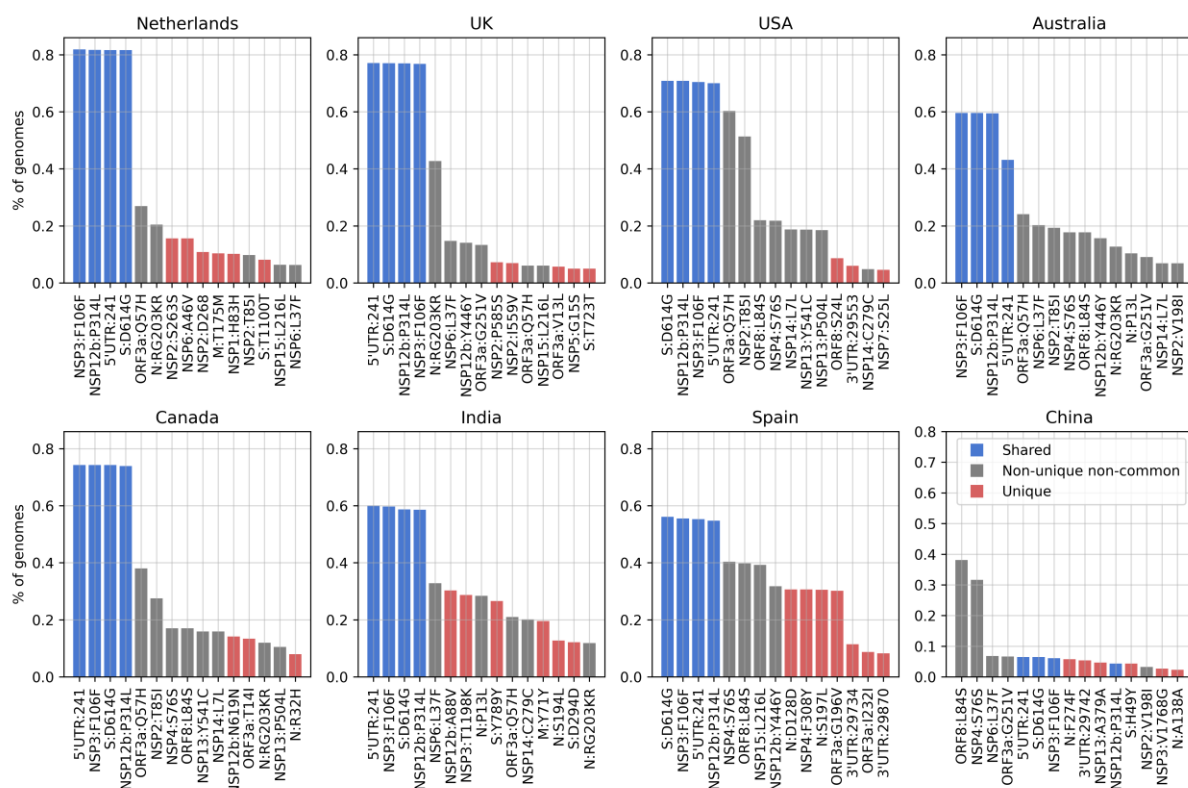
Comment: Figure 3 in the manuscript.
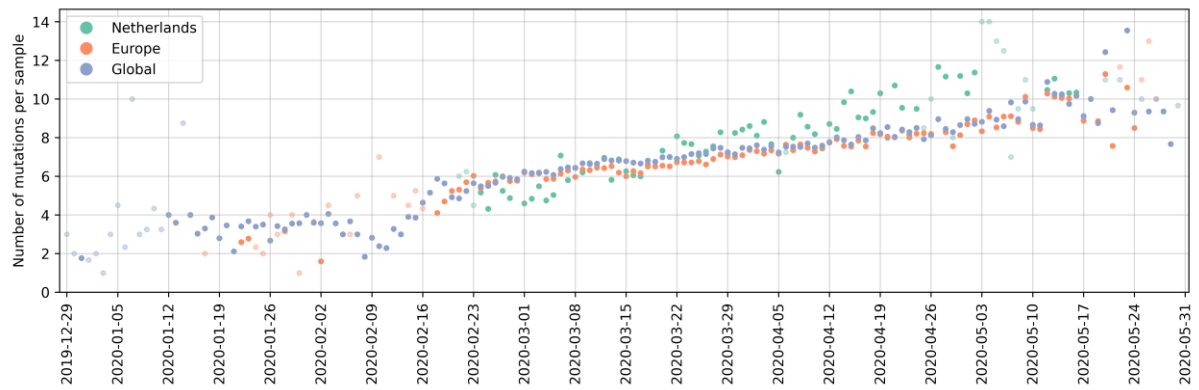
fig_4_Total-number-of-nucleotide-mutations.png

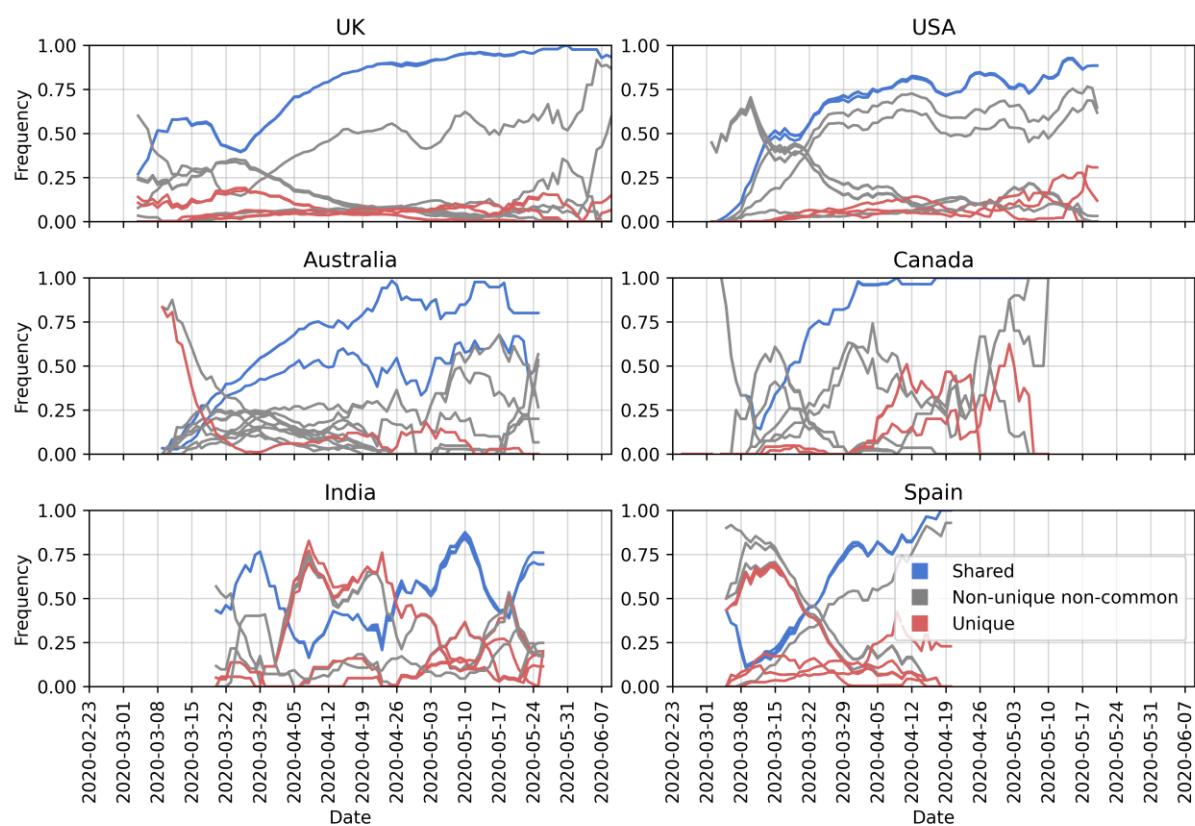Comment: Figure 4 in the manuscript.

fig_5_Total-frequency-of-the-top-15-mutations-in-the-most-sampled-countries-in-our-dataset.png
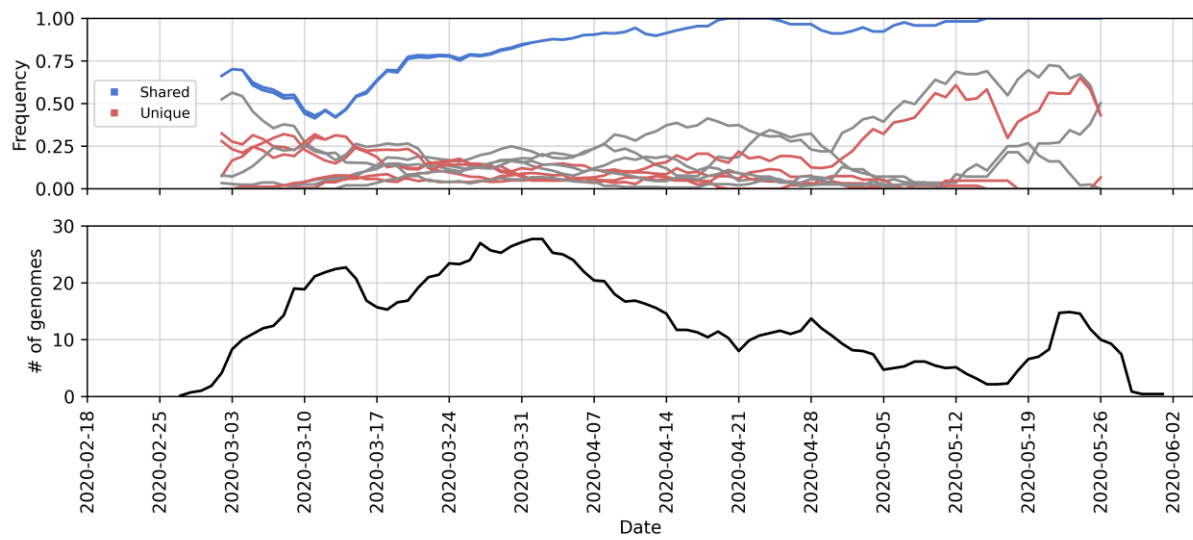
Comment: Figure 5 in the manuscript.

fig_6_Change-in-frequency-of-the-top-15-mutations-in-the-most-sampled-countries-in-our-dataset.png

Comment: Figure 6 in the manuscript.

fig_7_Change-in-frequency-of-the-top-15-mutations-in-the-Netherlands.png

Comment: Figure 7 in the manuscript.

Acknowledgements

Citing this document

Dvinskis, E., Iñigo de la Cruz, L. & Quintero Y. (2025). CODECHECK Certificate 2025-006. CODECHECK. https://doi.org/10.5281/zenodo.15520471

About CODECHECK

This certificate confirms that the code checkers could independently reproduce the results of a computational analysis given the data and code from a third party. A CODECHECK does not check whether the original computation analysis is correct. However, as all materials required for the reproduction are freely available by following the links in this document, the reader can then study for themselves the code and data.