**Pipeline for weekly update of new CovidTrackerCT website**Author: Kien Pham  
Date created: 24/06/2022  
Date updated: 13/07/2022

1. Data needed:

|  |  |
| --- | --- |
| File | Description |
| metadata\_CT\_all.csv | File containing curated/cleaned metadata for all genomes in Connecticut, obtained on GISAID. The current file should have all Connecticut submissions up to the date of website update and is a template should one wants to create a manual metadata file. |
| gisaid\_hcov-19\*.csv | Latest metadata of sequences uploaded since last reporting date, obtained on GISAID. |
| Lab metadata file | Lab metadata file, for Ct value graph, template TBD or contact [grubaughlab@gmail.com](mailto:grubaughlab@gmail.com)  *Note: template and codes refer to this file as the original Grubaugh Lab metadata sheet (Glab…). Should one wishes to use own lab data, change csv file name in server.R before reading.* |
| Table1.csv | Table 1, 1 week old, as backup, as template. |
| Table1\_new.csv | Table 1, latest update, as template. |
| time\_series\_covid19\_confirmed\_US.csv | File containing weekly COVID-19 case counts for all US states and territories, obtained from John Hopkins [github](https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv) page. |
| colors.csv | Color code for variant frequency plot |
| variant\_names.csv | Table for matching Pango lineage names to WHO variant names |

1. Revisions to dataset
2. First, download the latest version of the lab metadata file as a csv file.
3. Second, download the cumulative SARS-CoV-2 cases file from John Hopkins [github](https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv) page.
4. Third, log into GISAID and download the week's submission of Connecticut sequences:
   * + gisaid.org → 'Login' → Enter log in credentials → 'Search'
       - * Email GISAID if you do not yet have an account
     + In the Search menu:
       - Type "Connecticut" in **Location** box to filter for Connecticut sequences only
       - Type the **last date** of web update in the **left box** of **Collection Date**, so as to obtain the newly uploaded sequences for this week.
         * Reference date\_github variable in server.R file.
       - Check the box to select all samples
       - 'Download' → Select "Patient status metadata" → Download
     + Save downloaded file (gisaid\_hcov-19\*.tsv) into data folder
5. Update list of new Pangolin lineages in variant\_names.csv

* Refer to Pangolin [github](https://github.com/cov-lineages/pango-designation/) or double check after test run.

1. Finally, delete Table1.csv and rename Table1\_new.csv into the new Table1.csv
   * + Note to include any new variants into the csv file before running script

III. Revisions to R script

It is advisable to perform these changes and do at least one test run in gisaid.R scripts to check for errors prior to updating server.R and ui.R scripts.

1. server.R

* Update the dates variables:
  1. Date\_3\_months for lineage plot
  2. Date\_3\_weeks for variant plot and Table 1
  3. Date\_github for reference during metadata download
* Update total sequences variable:
  1. Gisaid.org → Search → Type “Yale” in Virus name (or other title identifiers for different labs)
  2. Copy number of search result

1. ui.R

* Change the date of reporting in the title variable.

Once all the steps are completed, run the Shiny App and check for any errors in the codes or the graphics.

If none spotted, publish the files onto the Shiny App server.

Remember to back up codes and data files after every successful run.

IV. Additional note

1. System:

* REMEMBER to save data and script files to the backup folder regularly.

1. Codes:

* It is always advisable to run the server scripts on the separate gisaid.R file first to check for errors in the individual code blocks before running the Shiny App.
* Table 1 scripts include merging 3 different datasets and exclusion of old data. The scripts themselves are very hands on and might bug given any formatting changes.
  + NOTE TO DOUBLE CHECK AND IMPROVE GIVEN TIME
* Be careful with formatting of Collection Date for metadata files, advisable to use as.Date() instead of mdy() but may change.
  + Hence backing up files is so important as should Collection Date bug to NAs, one would have to redownload and filter the full metadata file from GISAID.

1. Documents:

* Make sure to regularly check for new Pango lineage assignments to add to variant\_names.csv
  + One way to efficiently add new lineages as they appear is to run the app (DO NOT PUBLISH YET) and view Table 1 to see how many sequences are being classified as Other, then double check for actual lineage assignment in the metadata file for inclusion in variant\_names.csv.
* If new WHO variant appears, update colors.csv as well

IV. Debugging guide

* Lost or corrupted metadata file: If lost, the master file (metadata\_nextstrain.tsv) can be downloaded from GISAID (‘Downloads’ -> ‘Genomic epidemiology’ -> ‘metadata’) for initial run.