LINEAGE ASSIGNMENT OF SARS-COV-2

December 2024

- **State of the pangolin-data Art. How lineage assignment is performed in Netxclade, Pangolin and GISAID. Documentation.**

Three different methods can be used to perform the lineage assignment on SARS-CoV-2 sequences: Phylogenetic Assignment of Named Global Outbreak Lineages (PANGOLIN), Nextclade and Global initiative on Sharing All Influenza Data (GISAID).

PANGOLIN.

This tool is considered the gold standard for SARS-Cov-2 lineage classification. The information for the current document is based on detailed information obtained from the documentation (https://cov-lineages.org/resources/pangolin.html) and its github (https://github.com/cov-lineages). In order to simplify the document, the summary is only focused on the version 4 of Pangolin software.

Pangolin works with two main components: the *software* itself and the *database* used for the assignment of the lineages. The latest versions of each component are available in the repositories of Pangolin (https://github.com/cov-lineages/pangolin/releases) and the dataset (known as pangolin-data, https://github.com/cov-lineages/pangolin-data/releases). Although pangolin-data was introduced to replace the dependencies for pangoLEARN (now deprecated) and pango-designation, the latter is still used to collect the latest lineages assigned by the SARS-cov-2 Lineage Assignment Committee, which can be tracked in https://github.com/cov-lineages/pango-designation/commits/master/pango_designation and https://github.com/cov-lineages/pango-designation/blob/master/lineages.csv. Additional information about the new lineages can be found in https://github.com/cov-lineages/pango-designation/blob/master/lineage_notes.txt.

The Pangolin algorithm consists of three different phases:

1. Pre-processing. In a first step, the query FASTA sequences are aligned to a SARS-CoV-2 reference sequence using minimap2 and the non-coding regions of the resulting alignment are masked out with Ns using gofasta. The alignment then is transformed into hashes. A file called *lineages.hash.csv* (available in https://github.com/cov-lineages/pangolin-data/blob/main/pangolin_data/data/lineages.hash.csv), generated from an alignment of all the defined lineages (with multiple representative for each lineage), is used to evaluate whether the query sequences have previously designated as lineages. A second step, or SeqQC, evaluates the quality of the query sequences for ambiguous positions. A third step of this phase would be the Scorpio classification. In this process the software scorpio (https://github.com/cov-lineages/scorpio), in combination with constellations or collection of mutations with biological importance

(https://github.com/cov-lineages/constellations), is used to evaluate the presence of changes associated with variants of concern or interest. Briefly, there a several definitions for constellations of mutations (i.e. https://github.com/cov-lineages/constellations/blob/main/constellations/definitions/cBA.2.json) which contain the sites of the changes to study and certain rules (i.e. minimum number of alternative positions and/or máximun number of reference positions allowed out of the total of sites) to establish if a query sequence belongs to specific constellation. Additional details about the way scorpio works are explained in https://github.com/cov-lineages/scorpio/wiki, if needed. In a final step, the results from the previous processes are merged into a report in CSV format.

2. Inference. Pangolin uses UShER (https://usher-wiki.readthedocs.io/en/latest/UShER.html) to perform a phylogenetic classification. In order to assign a specific lineage to a query FASTA sequence, UShER carries out two different processes. In a first one, called *Preprocessing phase*, it assigns the mutations from a VCF file to specific nodes and tips in a Newick tree and stores this information as a mutation-annotated tree object. This output, in protocol buffer format, is available in each release of pangolin-data as lineageTree.pb (https://github.com/cov-lineages/pangolin-data/blob/main/pangolin_data/data/lineageTree.pb). The Newick tree and VCF file needed for building the tree object are generated by the Pangolin team from the global alignment with those reference sequences that the SARS-cov-2 Lineage Assignment Committee defines as representatives for lineages at a specific moment. In a second step, *Placement phase*, UShER would try to find the best location for the query sequences into the lineageTree.pb based on variants from their corresponding VCF files. Once a phylogenetic inference is reached, a report in CSV format is generated.
3. Final report or merging of the csv files from the previous phases in a Lineage report, in CSV format as well.

NEXTCLADE.

Similar to pangolin, Nextclade tool has two main components, the software and the databases associated with the lineages, which versions are available in https://github.com/nextstrain/nextclade/releases and https://github.com/nextstrain/nextclade_data/releases, respectively. A consideration to take into account is that Nexlcade can be used for the analysis of different viruses (influenza, measles, SARS-CoV-2, dengue…). As a result, there are several datasets to use, which can contain information for different viruses or being virus-specific. It is strongly recommended to check the data that nextclade_data contains to evaluate whether it is suitable or not for the analysis.

The algorithm that Nextclade uses, summarized from Nexclade documentation (https://docs.nextstrain.org/projects/nextclade/en/stable/index.html), follows the next steps:

1. Sequence alignment. The query sequences are aligned to the reference sequence (known as root) using seed matching approach. The process tries to find exact

matches ignoring the third position of a codon and, consequently, ignoring most of the synonymous mutations. Once the alignment is properly done, insertions are recorded and removed from the alignment, so that the coordinates remain similar to the reference sequence used in the process. A genome annotation can also be used to enable the alignment to deal in a more efficient way by introducing a lower gap-open penalty when deletion or insertion are present.

2. Translation. The provided genome annotation also contains the range of positions to be covered for each coding region (CDS). Using this information, Nextclade extracts the nucleotides for each CDS, translates them into amino acids and generates a protein alignment.
3. Mutation calling. Both alignments, nucleotide and amino acid, are used to find the mutations by comparing the positions to the reference sequences and a report is generated covering the substitutions, insertions, missing information (Ns) as well as ambiguous nucleotides (noted as non-ACGTN) found in the query sequences.
4. Detection of PCR primer changes. If a table with details of the PCR primers is provided, Nextclade identifies the regions where those primers fall and evaluate if any of mutations listed in the previous step is included in those regions. If so, it would report a potential change of PCR primers.
5. Phylogenetic placement. In order to place the query sequences in the reference tree, the root of the tree must be the same as the reference sequence used for the alignment. The reference tree (i.e. https://github.com/nextstrain/nextclade_data/blob/master/data/nextstrain/sars-cov-2/BA.2.86/tree.json) defines reference nodes, branches and tips with their specific mutations. The mutations from the query sequences, listed in the step Mutation calling, can be compared to the ones defining every node and tip in the reference tree so that the query sequences can be associated to the nearest reference node. At this step, the mutations found in the query sequences can be divided into two groups; *shared mutations* and *private mutations* or changes of the query sequences that differ from the nearest node. After doing the placement of a query sequence in an initial node, Nextclade will evaluate if the mutations of the query sequences are shared with other nodes and will refine the placement of the query sequences, if needed, to get a final tree.
6. Clade assignment. In the previous step, the query sequences are placed in a final tree based on their mutations and those ones found in the tree.json (reference tree). The tree also contains the annotation for the clade assigned by Nextclade that each node belongs to at a specific time. As the query sequences are resolved in a specific node of the reference tree, their clades are equally inferred from that placement.
7. Quality control. Nextclade checks for quality of the sequences evaluating the content of missing data (Ns), ambiguous nucleotides, private mutations (due to either unusual mutations or sequencing errors), stop codons and frame shifts.

An important consideration to make is that, rather than using the most complete reference sequence dataset that Pangolin uses for the lineage assignment, Nextclade is focused on using recent sequences (obtained within the last 12 months), which makes the analysis easier and quicker.

GISAID

There are some tools, such as CoVsurver, Audacity, Audacity Instant and Emerging Variants, that GISAID uses for the analysis of SARS-CoV-2 FASTA sequences, but the information and documentation on the algorithm and the process they follow to assign lineages is scarce. The developers have been contacted for detailed information several times but no additional information has been provided.

- **Pre-existing issues**

The main issue concerning the SARS-CoV-2 lineage assignment seems to be the discrepancy among the results obtained depending on the tool used for the classification at a specific time.

Although every tool has its own dataset and nomenclature, Nextcalde and GISAID also include the Pangolin designation, as Pangolin is the gold standard for assignment. For instance, Nextclade would use its own convention for a specific clade under the term “clade_nextstrain” but also other nomenclatures such as “Nextclade_pango” or “clade_who”, to refer Pangolin and WHO classification, respectively. So does GISAID, by using the terms “Clade” and “Pango Lineage” to define their own classification and Pangolin one, respectively. In spite of the existence of different genetic classifications, WHO is currently using Pangolin and Nextclade for SARS-CoV-2 surveillance (https://www.who.int/activities/tracking-SARS-CoV-2-variants)

Every time there is a new lineage designated, Pangolin updates it in pango-designation (https://github.com/cov-lineages/pango-designation/commits/master/pango_designation) by means of commits with all the changes, easily tracked in time. Those new lineages would be eventually included in the latest dataset and released in https://github.com/cov-lineages/pangolin-data/releases. However the inclusion of those new lineages in the pangolin-data may not happen immediately and the new dataset can take a time to be released.

Nextclade can use all the changes included in the pangolin-designation repository to update and release its latest dataset. Nextclade always describes the range of dates when those new Pangolin lineages were populated into its dataset (for instance, see https://github.com/nextstrain/nextclade_data/releases/tag/2024-06-13--23-42-47Z), so the information Nextclade incorporates from Pangolin can be properly tracked. However, this could create a scenario in which Nextclade could be already updated with some new lineages that Pangolin has not released on a new pangolin-data yet. As a consequence, Nextclade could be the most updated and accurate tool to perform SARS-CoV-2 lineage assignment at specific times, which usually corresponds to those periods between pangolin-data version releases.

GISAID also seems to update the information on new lineages in its dataset before Pangolin releases the latest dataset. They may capture the information from the pangolin-designation repository as Nextclade does, but the way that GISAID updates their records with new lineages is not clear yet. GISAID has been contacted for additional information on that scope. Some speculations have been suggested about the way GISAID performs the assignment (maybe using Nextclade) and it seems that the information available about their algorithm is scarce /https://github.com/cov-lineages/pangolin-data/issues/58 https://github.com/cov-lineages/pangolin-data/issues/46 so is the documentation.

- **Recommendations for SARS-CoV-2 lineage assignment for Relecov members**

Discrepancies between Pangolin and Nextclade results of the SARS-CoV-2 lineage assignment have been frequently reported and usually happen in periods between pangolin-data version releases. As mentioned above, it seems Nextclade updates its dataset with new lineages more often than pangolin-data does, which also has been confirmed by Pangolin developers (https://github.com/cov-lineages/pangolin-data/issues/58). The following section aims to offer some guidelines to follow if discrepancies between Pangolin and Nextclade SARS-CoV-2 lineage assignments are found before reporting the results:

1. Confirm the versions of Pangolin and Nextclade software and their dataset version (pangolin-data and nextclade-data) used for the assignment. This information is also available for users of the web version tools (https://pangolin.cog-uk.io/ and https://clades.nextstrain.org/). Please share the details of both components (software as well as dataset) when asking for help. It is strongly recommended to use the latest versions of the software packages and the datasets, as they will be the most updated tools for the assignment. Additional information about how to keep the software and database updated can be found in **Annex 1**.
2. Evaluate which one of the datasets (pangolin-data or nextclade data) is the most updated dataset at a specific time. This can be done easily by checking the release dates and other details included with the dataset. The information about the latest version of pangolin-data and nextclade_data are available in the repository https://github.com/cov-lineages/pangolin-data/releases and https://github.com/nextstrain/nextclade_data/releases, respectively. Both tools always offer details about the changes the dataset includes. Nextclade gives a time range that covers the incorporation of new Pangolin lineages to its dataset. If this time range is more recent than the release date of latest pangolin-data, Nextclade assignment is more updated and more accurate than Pangolin one until a new pangolin-data is available.
3. Check the repository https://github.com/cov-lineages/pango-designation/commits/master/pango_designation for changes on the SARS-CoV-2 assignment. Use the time range mentioned above to identify those new Pangolin lineages that Nextclade may have already included in its dataset but not added to the latest pangolin-data.
4. Once the analysis using a specific dataset has been performed, it is extremely recommended checking the quality of the analysis and the individual results for each sequence before giving a SARS-CoV-2 lineage assignment as valid. Details on the expected outputs for pangolin and nextclade analyses can be found in https://cov-lineages.org/resources/pangolin/output.html and https://docs.nextstrain.org/projects/nextclade/en/stable/user/output-files/04-results-tsv.html. It has been reported that sequences not having good quality (i.e. high contain of Ns, missing information about specific mutations) or even new lineages never

reported before have been assigned to the nearest common ancestor, as the software would assign the sequences to the closest phylogenetic strain available in its dataset.

We do not discourage from using GISAID for the SARS-CoV-2 assignment at all. However, more feedback from the developers is needed to understand the tool and give recommendations for its use.

It has been suggested the creation of a potential dataset for validation of result assignment when discrepancies are reported. The best way of creating this dataset is still under consideration. The SARS-CoV-2 lineage assignment is a process very dynamic and a sequence belonging to a specific Pangolin lineage could be reassigned to a different one in a very short period, which always depends on how long the pangolin-data takes to be released. This fact can make the selection of data a very complex process and the dataset being ephemeral.

- **Final conclusion**

Considering that the most commonly used nomenclature is the assignment of lineages, which is given by pangolin, the most recommended software to use when doing the analysis of SARS-CoV-2 genome sequences should be pangolin software in combination with the latest version of pangolin data.

On the other hand, the assigned lineages must be carefully reviewed to avoid the communication of results that could be inconsistent with the circulation of the viruses at the time of the analysis.

ANNEX 1

PANGOLIN

To install Pangolin you can follow the instructions in the following link. https://cov-lineages.org/resources/pangolin/installation.html.

As an alternative to Bioconda, a Pangolin singularity image can be used, which is available in the repository https://depot.galaxyproject.org/singularity/.

Before using Pangolin, it is recommended to ensure that the latest version of the **software** is being used (https://cov-lineages.org/resources/pangolin/updating.html). If you are using the singularity image, go to the repository (https://depot.galaxyproject.org/singularity) and verify that the latest version is currently downloaded.

Besides, it is necessary to update the pango-data **database**. To do so, use the command *pangolin --update-data*, which downloads the latest version of pango-data available in a default folder. For better tracking of the versions used, we recommend adding the *--datadir* parameter. This will specify the path to the folder where the database is going to be downloaded (this folder must have been previously created).

An example command would be as followed:

`pangolin --update-data --datadir /path/to/folder/`.

In case you have specified a folder for pango-data download, its path should be provided when running Pangolin by adding the *--datadir* parameter again. An example of use would be the following:

```
pangolin secuences.fasta --outfile results.csv --datadir /path/to/folder/ --threads 4
```

NEXTCLADE

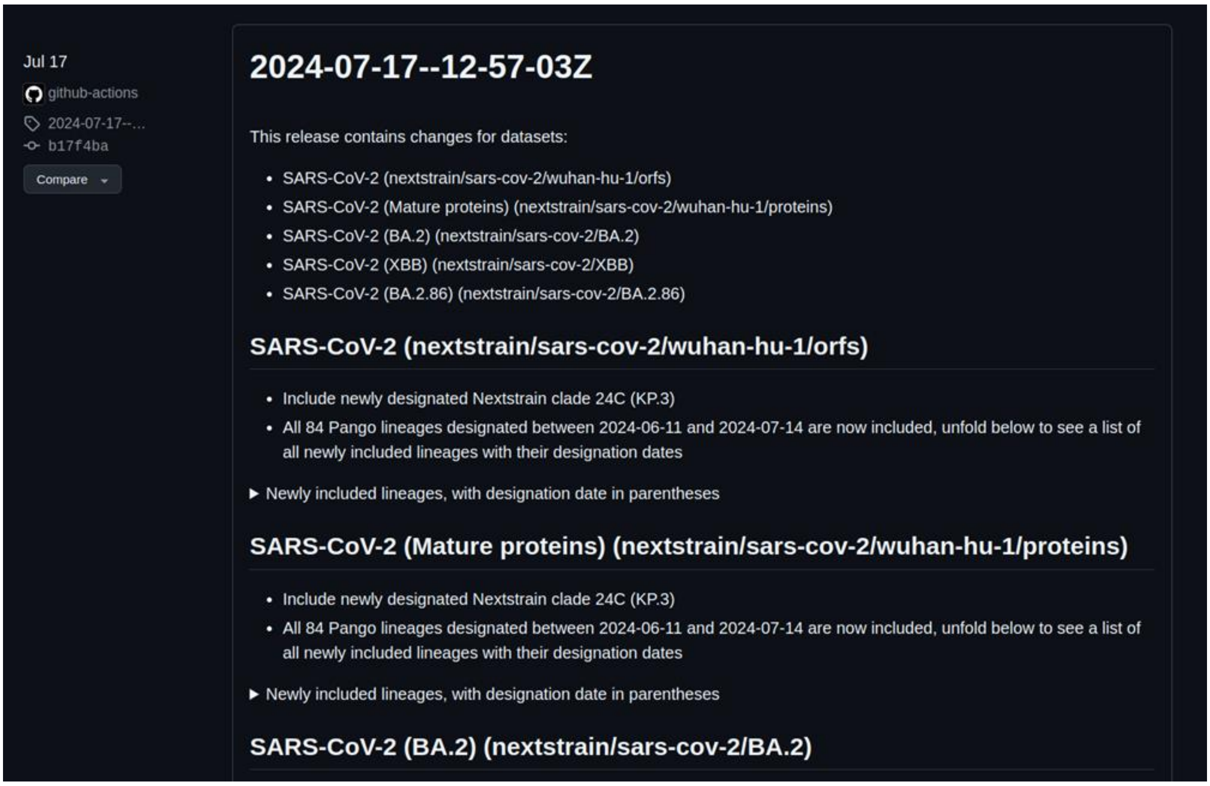
To install nextclade you can follow the instructions in the following link: https://docs.nextstrain.org/projects/nextclade/en/stable/user/nextclade-cli/installation/index.html. Similar to pangolin, the singularity image from nextclade can be used (https://depot.galaxyproject.org/singularity/) and it is strongly recommended to check that the most updated version of the **software** is in use.

For nextclade, the **databases** are separated into different "datasets" according to the virus under study (the dataset we use for SARS-CoV-2 assignment is called "*sars-cov-2*". Within each dataset, different versions are released, identified by a tag with the following format *YYYY-MM-DD--HH:MM:SSZ*.

Nextclade has an option that allows running the analysis without having a dataset locally downloaded. The software will temporarily download the latest available version of the dataset and, when the analysis is finished, delete it. However, for tracking purposes, it is recommended to download the dataset locally.

To download the latest version of the SARS-CoV-2 dataset for nexclade we can check the latest tag that includes the most recents changes at https://github.com/nextstrain/nextclade_data/releases. Be aware it is essential to look for the latest tag containing changes for SARS-CoV-2. As the dataset can be virus-specific, using tags that contain changes for other viruses than SARS-CoV-2 may cause failure of the analysis.

For instance, ***2024-07-17--12-57-03Z*** will be the tag to be used at a specific time.



2024-07-17--12-57-03Z

This release contains changes for datasets:

- SARS-CoV-2 (nextstrain/sars-cov-2/wuhan-hu-1/orfs)
- SARS-CoV-2 (Mature proteins) (nextstrain/sars-cov-2/wuhan-hu-1/proteins)
- SARS-CoV-2 (BA.2) (nextstrain/sars-cov-2/BA.2)
- SARS-CoV-2 (XBB) (nextstrain/sars-cov-2/XBB)
- SARS-CoV-2 (BA.2.86) (nextstrain/sars-cov-2/BA.2.86)

SARS-CoV-2 (nextstrain/sars-cov-2/wuhan-hu-1/orfs)

- Include newly designated Nextstrain clade 24C (KP.3)
- All 84 Pango lineages designated between 2024-06-11 and 2024-07-14 are now included, unfold below to see a list of all newly included lineages with their designation dates

▶ Newly included lineages, with designation date in parentheses

SARS-CoV-2 (Mature proteins) (nextstrain/sars-cov-2/wuhan-hu-1/proteins)

- Include newly designated Nextstrain clade 24C (KP.3)
- All 84 Pango lineages designated between 2024-06-11 and 2024-07-14 are now included, unfold below to see a list of all newly included lineages with their designation dates

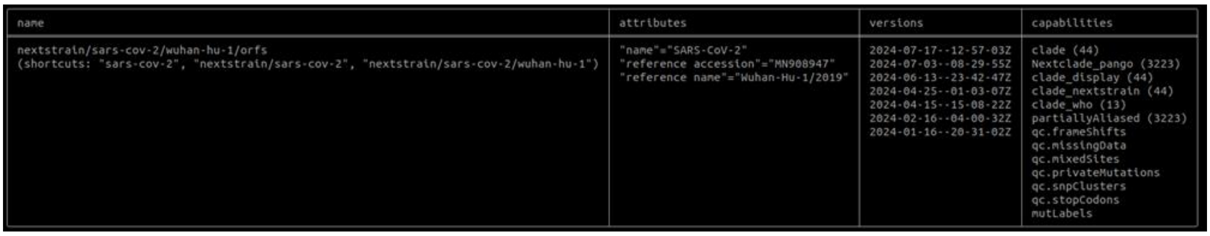
▶ Newly included lineages, with designation date in parentheses

SARS-CoV-2 (BA.2) (nextstrain/sars-cov-2/BA.2)

You can directly query the last available tag using the nexclade software itself, with the following command:

```
nextclade dataset list -n sars-cov-2
```

In this case we also see that the latest tag available for the SARS-CoV-2 dataset is ***2024-07-17--12-57-03Z***.



```
name                                                                     attributes                            versions                capabilities
nextstrain/sars-cov-2/wuhan-hu-1/orfs                                    "name"="SARS-CoV-2"                   2024-07-17--12-57-03Z   clade (44)
(shortcuts: "sars-cov-2", "nextstrain/sars-cov-2", "nextstrain/sars-cov-2/wuhan-hu-1")  "reference accession"="MN908947"   2024-07-03--08-29-55Z   Nextclade_pango (3223)
                                                                         "reference name"="Wuhan-Hu-1/2019"    2024-06-13--23-42-47Z   clade_display (44)
                                                                                                               2024-04-25--01-03-07Z   clade_nextstrain (44)
                                                                                                               2024-04-15--15-08-22Z   clade_who (13)
                                                                                                               2024-02-16--04-00-32Z   partiallyAliased (3223)
                                                                                                               2024-01-16--20-31-02Z   qc.frameShifts
                                                                                                                                       qc.missingData
                                                                                                                                       qc.mixedSites
                                                                                                                                       qc.privateMutations
                                                                                                                                       qc.snpClusters
                                                                                                                                       qc.stopCodons
                                                                                                                                       mutLabels
```

In order to download the dataset we will use the following command (replacing the tag with the most recent version):

```
nextclade dataset get --name 'sars-cov-2' --tag 'YYYY-MM-DD--HH-MM-SSZ' --output-dir '/path/to/folder/'
```

Once downloaded, when nextclade is run, the path to the folder containing the dataset will be specified:

```
nextclade run --input-dataset '/path/to/folder/' --output-tsv 'results.tsv' sequences.fasta
```