

MAPLE tutorial

Nicola De Maio, 3rd June 2025

In this tutorial we will use MAPLE to estimate a phylogenetic tree from a sample of 3,600 SARS-CoV-2 genomes of the lineage B.1.429.

First, we will perform a simple phylogenetic inference as (details of MAPLE's algorithm and performance can be found in [De Maio et al. 2023 Nature Genetics](#)).

Then, we will perform the same inference under a more complex model of rate variation and error rate variation (see [De Maio et al 2024a bioRxiv](#)).

Finally, we will use SPRTA (see [De Maio et al. 2024b bioRxiv](#)) to assess and represent uncertainty in the phylogenetic inference.

1 Installation and data preparation

The latest version of the MAPLE python script is available from <https://github.com/NicolaDM/MAPLE> ("MAPLEv*.py").

No installation of the script is needed, however you will need python3 to be installed on your machine (see <https://www.python.org/downloads/>) to run the script.

Even better, running the script with pypy3 (see <https://pypy.org/>) is around 10x faster, and is highly recommended.

Here we assume that python3 and pypy3 have been installed on the machine being used. In case you could not install pypy3, simply replace the "pypy3" commands below with "python3".

We will also assume that the MAPLE script [MAPLEv0.7.5.py](#) has been downloaded and placed into a folder in which we will run all analyses and store all our data - for simplicity in the following we will call this folder (including path) "/MAPLE_example".

Next step, let's download the example B.1.429 fasta alignment that we will analyse. This can be downloaded [here](#).

(Note that this file contains ALIGNED genomes. If your genomes are not aligned, you can align them with MAFFT, for example following [these instructions](#).)

Unzip this file and note that this is >100MB.

Let's create a MAPLE format alignment which is more concise, readable, and which we can use as input for MAPLE.

First, download the script [createMapleFile.py](#) and move it to our /MAPLE_example folder.

Now let's convert the fasta file into a MAPLE file; open a terminal and execute:

```
cd /MAPLE_example
```

```
pypy3 createMapleFile.py --path /MAPLE_example/ --fasta B.1.429_partial.fa
```

--output B.1.429_partial.maple --overwrite

Now file “B.1.429_partial.maple” contains the same alignment in MAPLE format. You can see this file is much more concise (around 400 times smaller than the fasta file). This file starts with a consensus genome created by the script. Then all the samples are listed, each with a list of differences with respect to the consensus. You can see that some samples are identical to the consensus, others have 1 or more SNPs, and others are potentially identical but with some ambiguous characters.

```
gccaaactgtcactaagaatctgctgctgaggtcttcaagaagcctcgcaaaacgtactgccaataagcatatacacaagcttctggcagagctgtccagaacaacccaagga
aattttggggaccaggaactaatcagacaaggaactgattacaacattggccgcaattgcacaatttgccccagcgcttctcggaatgtcgcgattggcatggaagtcaacc
ttcgggaacgtgtgtgacctacacaggtgcatcaaatggatgacaagaatccaatttcaagatcaagtcatttctgtaataagcatattgacgatacaaaacatttcaccaacagagc
ctaaaaggacaaaagaagaaggctgtgaaactcaagccttacgcagagacagaagaacagcaactgtgactcttcttctgctgcagatttggatatttctccaacaaattgcaaca
tcctagcagcagtgctgactcaactcaggtctaaactcatgcagaccacacaaggcagatgggctatataaacgttttgcgtttacgatatagtctactcttgtgcagaatgaattc
tgcatacatagcaccaagtagatgtagtttaactttaactcacatagcaatctttaatcagtggttaacattagggaggacttgaagagccaccacatttcaccagagccacgcggaatc
gatcgagtgtagtgaaacatgctagggagagctgcctatatggaagagccctaatgtgtaaaataatttttagtagtgcattcccatgtgactgttagctgcctaggagaacagccata
tggaaagagaaaaa
>SRR16791746
>SRR17363680
>SRR16791620
>SRR13527035
>SRR16942814
>SRR16791919
>SRR16649320
>SRR16791636
y 29362
>SRR13689667
c 2395
t 2597
c 8947
c 9738
t 11345
c 12100
a 12878
t 16394
g 17615 274
- 21717
t 24349
a 27626
g 27890
g 28860
>SRR17362437
a 19419
>SRR17362100
t 23191
>SRR17361965
t 10641
>SRR20694822
a 19419
>SRR16943000
a 8785
>SRR17364956
t 2086
```

2 Running MAPLE

Let's run a simple MAPLE tree inference run using as input the .maple file we just created:

```
python3 MAPLEv0.7.5.py --input B.1.429_partial.maple --output
/MAPLE_example/B.1.429_partial_maple --model UNREST
```

This should run for less than a minute.

The software:

- 1) Builds an initial tree by placing samples one at a time.
- 2) Estimates model parameters.
- 3) Optimises branch lengths.
- 4) Re-roots the tree.
- 5) Improves the topology with shallow SPR moves.
- 6) Further improves the topology with a deeper SPR search.

Some of these steps, like inferring branch lengths and model parameters, are repeated multiple times.

We can instead run MAPLE with a more complex model of rate variation using option “--rateVariation”. Adding also option “--estimateMAT” MAPLE will also infer mutation events on the estimated tree:

```
pypy3 MAPLEv0.7.5.py --input B.1.429_partial.maple --output
/MAPLE_example/B.1.429_partial_maple --model UNREST --rateVariation
--estimateMAT --overwrite
```

When given enough genomes, if we want to further estimate site-specific error rates we can also add option “--estimateSiteSpecificErrorRate”, and to infer also individual sequence errors we can use option “--estimateErrors” the command line.

2.1 Running SPRTA

MAPLE, while inferring a tree, can also assess and summarise the uncertainty in the inferred tree using the approach called SPRTA. This usually comes at negligible additional computational cost. To do this, add option “--SPRTA” to the command line, and to allow visualization of alternative trees in the output add option “--networkOutput”:

```
pypy3 MAPLEv0.7.5.py --input B.1.429_partial.maple --output
/MAPLE_example/B.1.429_partial_maple --model UNREST --rateVariation
--estimateMAT --SPRTA --networkOutput --overwrite
```

3 Visualise the output in Taxonium

MAPLE outputs newick format trees that can be read by most phylogenetic software. For the command above the output tree will be the “B.1.429_partial_maple_tree.tree” file.

However, MAPLE is most useful for inferring large trees, and visualisation tools like FigTree can struggle to process these.

A phylogenetic visualisation tool that can manage trees with millions of samples is [Taxonium](#) ([Sanderson 2022](#)).


However, to use Taxonium we will need to convert the newick tree from MAPLE into a jsonl tree for Taxonium.

To do this, you can first [install TaxoniumTools](#) and then run the [newick_to_taxonium](#) script:

```
newick_to_taxonium -i B.1.429_partial_maple_tree.tree -m
B.1.429_partial_maple_metadata.tsv -o B.1.429_partial_maple_tree.jsonl -c
support,supportGroup,supportTo,mutationsInf,Ns
```

This script is combining the output newick tree from MAPLE with the metadata output from MAPLE to create a jsonl file for Taxonium containing a tree annotated with the metadata. The metadata contains information regarding the mutations on the branches of the tree (“mutationsInf”), incomplete parts of the genome sequences (“Ns”), and phylogenetic uncertainty (“support”, “supportGroup” and “supportTo”).

Let's open the jsonl tree in [Taxonium](#):

 Taxonium

Welcome to Taxonium, a tool for exploring large trees

Select, drag-and-drop, or enter the URL for tree or metadata files (jsonl, newick, nextstrain, tsv, etc.):


Choose files

No files chosen


Add

..or use text entry.

or load an ex

 SARS-CoV-2

All seven million public sequences of SARS-CoV-2 from the INSDC databases

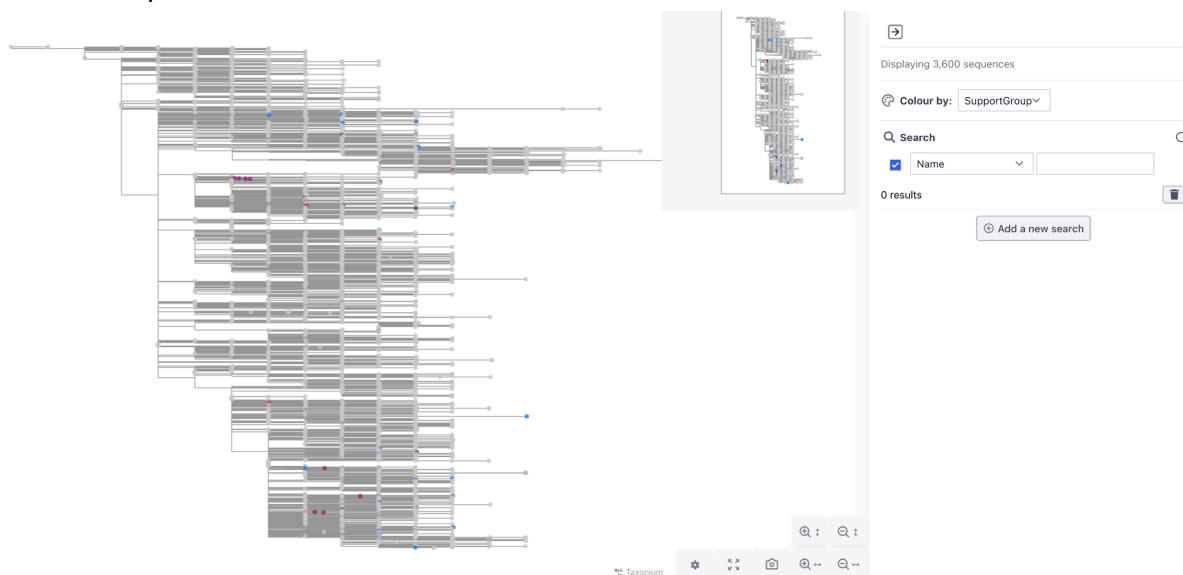
 NCBI Taxonomy (full)

Full 2.2M NCBI Taxonomy of species

Click “Choose files” and select the jsonl tree you just created.

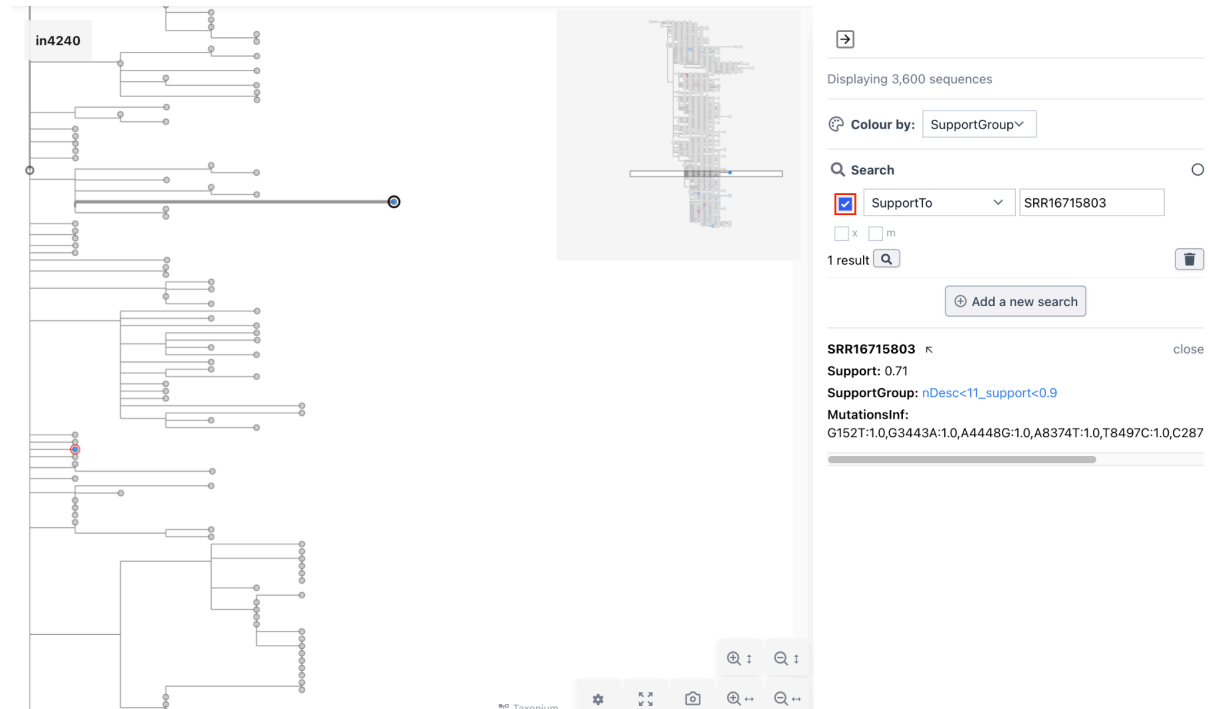
Taxonium will visualize the tree.

In “Colour by” you can select “SupportGroup” so that phylogenetically uncertain samples will be colored up:



If you click the blue sample sticking out at the lower part of the tree (sample SRR16715803) you can see it has a SPRTA support of 0.71. If Now you search the field “SupportTo” and enter the sample name “SRR16715803”, you can see that the alternative placement for this sample becomes highlighted with a red circle (it's the branch leading to sample

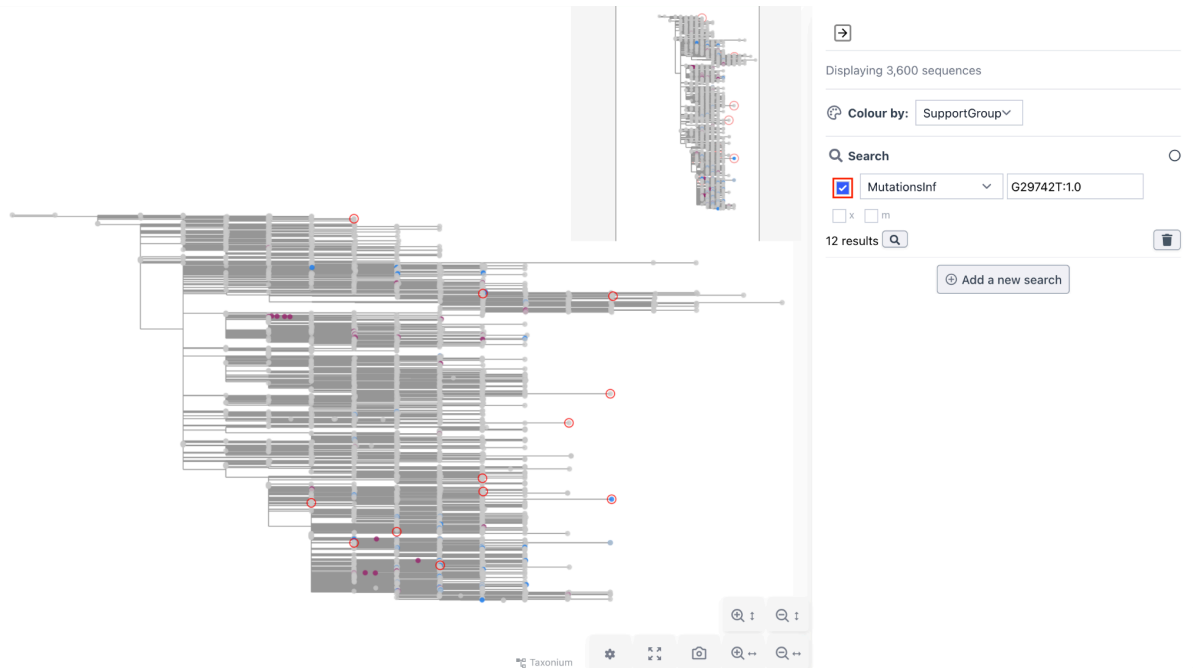
SRR17365533):



The support for this alternative placement is 0.284. Why is there uncertainty? In this case the reason is that both SRR16715803 and SRR17365533 contain mutation G29742T, as you can see in their “mutationsInf” fields. This means that in the alternative evolutionary history in which SRR16715803 descends from SRR17365533 we save the G29742T mutation on the branch leading to SRR16715803, but we replace it with a reversion of the C10702T mutation that leads to the subtree currently containing SRR16715803. This alternative evolutionary history is equally parsimonious to the one currently in the tree, but the one in the tree as slightly higher likelihood - in this case the difference is due to the fact that position 29742 has a higher inferred mutation rate than position 10702 (55.4 vs 6.8), although the background CT rate is higher than the GT rate. These rates can be found in the “B.1.429_partial_maple_subs.txt” file created by MAPLE.

Why is the G29742T rate so high compared to the average substitution rate? You can search for occurrences of the specific G29742T mutation in the search field “mutationsInf” in

Taxonium:



As you can see there are in total 12 mutations of this type in this small lineage considered (and you can see that a few other occurrences of this mutation also cause phylogenetic uncertainty), while only 2 mutations of type C10702T are present in the tree.