



From: Trevor Bedford [REDACTED]@bedford.io
Subject: Re: Json file for your global phylogenetic tree
Date: October 8, 2025 at 3:25 PM
To: GISAID Service Support [REDACTED]@gisaid.org

Dear GISAID Service Support Team,

Thanks for getting in touch and I'm glad you've been able to run our Nextstrain software to provide trees for many other users. The pattern of generating and sharing Augur JSONs for viewing in auspice.us is something we also recommend to public health labs. Additionally some groups choose to share these JSONs either publicly or privately via nextstrain.org/groups. We generally just hope the software is being useful. Please let me know if there are specific things that we could do to make it more useful for you.

And I hugely appreciate the specific offer of help. However, we're currently running <https://github.com/nextstrain/ncov> that produces a number of JSONs that live under nextstrain.org/ncov/gisaid/, including for example:

- <https://nextstrain.org/ncov/gisaid/global/all-time>
- <https://nextstrain.org/ncov/gisaid/north-america/all-time>
- <https://nextstrain.org/ncov/gisaid/asia/2m>
- <https://nextstrain.org/ncov/gisaid/21L/europe/6m>

This is a matrix across 7 different regions, across 4 different temporal resolutions and across 2 rootings (Wuhan and 21L). This totals 56 different JSON files that we produce weekly via the `ncov` workflow.

In addition we run Multinomial Logistic Regression (MLR) on clades and lineages from sequence counts as visible here: <https://nextstrain.org/sars-cov-2/forecasts>

I suspect there's also specifics in our `ncov` workflow that are not in your current Augur-based workflow, including things like coloring the tree from MLR fitness (https://nextstrain.org/ncov/gisaid/global/6m?c=mlr_lineage_fitness) and coloring the tree based on immune escape from DMS assay data (https://nextstrain.org/ncov/gisaid/21L/global/all-time?c=immune_escape).

For these reasons, I'd prefer flat file access to keep things updated on our end. I believe that for historical reasons you were creating a Nextstrain specific file here:

https://www.epicov.org/epi3/3p/nextflu/export/corona2020_fulldump.json.bz2

But we shouldn't need a Nextstrain specific file and should be able to use the same file as provided to other users like <https://cov-spectrum.org> and <https://cov.lanl.gov/content/index>. I also had previously written to you about sharding these files to decrease load on everyone's end (please see below email from July 2024).

What would be the process to get access to these same files that are already created for CoVSpectrum, LANL, etc...?

If at all helpful I'd be very happy to talk directly over Zoom. I can send you a Zoom link if you have some preferred times.

Thank you and take care,
- Trevor

Begin forwarded message:

From: Trevor Bedford [REDACTED]@bedford.io>
Subject: Sharded EpiCoV data
Date: July 2, 2024 at 2:16:25 PM PDT
To: TechDev GISAID [REDACTED]@gisaid.org>
Cc: [REDACTED]@gisaid.org, [REDACTED]@gisaid.org, [REDACTED]@gisaid.org

Dear GISAID,

Thank you so much for your continued shepherding of SARS-CoV-2 genomic data. I hugely appreciate the efforts of data curation and promoting data sharing. Thank you as well for continuing to share a flat file of the entire SARS-CoV-2 dataset to:

https://www.epicov.org/epi3/3p/nextflu/export/corona2020_fulldump.json.bz2

This has allowed us at Nextstrain to continue to keep nextstrain.org/ncov up to date over these past 4 years. I do have a request here however (along the same lines as the emails I sent below in fall 2022).

Currently, there's just the 43 Gb single file of everything. We check this file every day to see if there are updates and if there are updates we need to process the file. If instead we had something like the following:

- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2020.json.xz
- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2021.json.xz
- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2022.json.xz
- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2023.json.xz
- https://www.epicov.org/epi3/3p/nextflu/export/epicov_2024.json.xz

each with sequences submitted in 2020, in 2021, etc... we'd only need to check the 2024 file daily. This would save us substantial download and processing overhead compared to the current 43 Gb file and I imagine it would also save you substantial overhead in terms of creating a database dump and in terms of bandwidth. This way you'd only need to update the much smaller `epicov_2024.json.xz` file every ~3 days when updates are made.

Would this sort of "sharded" database dump be something you might consider? Again, I'd hope this would improve things for both organizations. Though of course I'd be open to other suggestions as well. But generally, it would be welcome to have ways to more easily deal with the very large dataset, especially considering that it only grows every day.

Thank you again for everything you continue to do for SARS-CoV-2 genomic surveillance.

Best,
- Trevor

On Oct 6, 2025, at 1:37 PM, GISAID Service Support [REDACTED]@gisaid.org> wrote:

Dear Trevor

Apologies. We've just been alerted that we may have used an outdated email address when we invited you to send us the parameters for your global ncov/gisaid tree (see below). We are now using the email provided to us by the Secretariat.

Again, we are happy to generate in Augur your json of your customized subsampled phylogenetic tree, which should result in a relatively small (50-70 MB) file size, even uncompressed.

GISAID Service Support Team

From: GISAID Database Support [REDACTED]@gisaid.org>
Date: Saturday, October 4, 2025 at 16:55
To: Trevor Bedford [REDACTED]@fredhutch.org>
Subject: Json file for your global phylogenetic tree

Hi Trevor,

You have asked 'about a strategy to lighten the load for provisioning data' and ways to keep the global phylogenetic tree up-to-date that you publish as "ncov/gisaid." We are happy to share with you specifics on how we support trees from other users that receive GISAID's support.

We have been providing many of our users with a json file that is generated for them on a weekly basis, ie. basically, with the same cadence you are updating your global tree with data shared via GISAID. This approach has turned out to be a most efficient process. Users' parameters are run on Augur which takes full advantage of EpiCoV's inventory to generate their customized subsampled phylogenetic tree for use with a variety of applications, including Auspice.

Simply provide us the parameters for your global phylogenetic tree in a timely manner, and we'll ensure that you too can take advantage of this service to keep your tree updated as well.

Your GISAID Service Support Team

