Technical Assistance and Instructions for Public Health Laboratories on Categorizing Sequence Data as "Baseline Surveillance" for Inclusion in CDC's National SARS-CoV-2 Genomic Surveillance

CDC initiated the National SARS-CoV-2 Strain Surveillance (NS3) program in November 2020 to establish a nationally representative system for baseline genomic surveillance. Through NS3, CDC has built a collection of representative SARS-CoV-2 specimens and sequences that laboratories use to characterize viruses and better inform public health decisions.

By monitoring the evolution of circulating SARS-CoV-2 variants through sequencing and comprehensive analyses, the NS3 program is able to assess their potential impact on the effectiveness of current vaccines, therapeutics, diagnostics, and public health activities. The emergence of concerning SARS-CoV-2 variants in late 2020 necessitated rapidly increasing the number of available U.S. SARS-CoV-2 sequences. The resulting volume of genomic surveillance data has been substantial, that are important to provide estimates of circulating SARS-CoV-2 variants at the national, regional, and jurisdictional levels.

NS3 is just one important advance. Significant efforts and contributions continue to be made by U.S. public health laboratories as part of the ongoing global public health response to the COVID-19 pandemic. The SARS-CoV-2 sequencing capacity at state and local public health laboratories has expanded rapidly, with immediate benefits for public health. In addition to conducting virus surveillance and SARS-CoV-2 variant detection, public health laboratories have implemented genomic sequencing for epidemiologic, scientific, and public health purposes and have collectively published more than 100,000 SARS-CoV-2 sequences to public databases. However, sequence data generated at U.S. public health laboratories are not currently included in CDC's estimates of circulating SARS-CoV-2 variants.

To expand data included in CDC's estimates, CDC would like to include SARS-CoV-2 sequences generated by your laboratory that meet the baseline surveillance criteria.

The goals of including these data are to:

1. Allow for more robust state-level estimates of circulating SARS-CoV-2 lineages, including variants of interest and concern[1]
2. Monitor viral evolution within jurisdictions at a more granular level
3. Provide comprehensive data for public health decision makers at the jurisdictional and national levels
4. More accurately represent sequencing efforts and contributions made by jurisdictions to the broader scientific community

Baseline surveillance is achieved by sequencing specimens that represent geographic, demographic (e.g., age), and clinical (e.g., disease severity or outcome) diversity across a jurisdiction through a random selection of SARS-CoV-2-positive, diagnostic specimens.

Sequences that **meet the criteria** for baseline surveillance analyses include those:

- Sampled randomly for genomic surveillance
- Not identified in a targeted sampling effort (targeted efforts defined below)

[1]https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html

- Sampled across targeted sequencing efforts to be representative of the community

Sequences from targeted efforts include, but are not limited to, those:

- Sampled based on cluster/outbreak investigations
- Longitudinally or repeatedly sampled from the same individual
- Sampled based on pre-screening for a particular variant (e.g., S-gene target failure)
- Sampled for the purpose of vaccine escape studies
- Sampled based on travel history
- Sampled based on disease severity (i.e., targeted sequencing of cases resulting in hospitalization or death)

Identifying a representative subset of sequences from targeted sequencing efforts:
Inclusion of all sequences from targeted sequencing efforts in baseline surveillance could bias estimates of circulating SARS-CoV-2 lineages by overrepresenting lineages. However, sampling sequences from targeted sequencing efforts that are representative of the community should be included in baseline surveillance.

To achieve a representative sample:

- Sample a similar proportion of sequences from a targeted sequencing effort as what is sampled for general surveillance efforts.
- For targeted efforts involving longitudinal or repeated sampling of the same individual, tag only one sequence per individual as baseline surveillance.

For CDC to correctly identify and ingest SARS-CoV-2 sequences generated by your laboratory in the baseline surveillance analysis, the sequences need to be tagged as such in online databases:

- For NCBI submissions, this is done by including a keyword: "purposeofsampling:baselinesurveillance".
- For GISAID submissions, this is accomplished by selecting "Baseline surveillance" in the **sampling strategy** field.
- See further instructions below for how to tag new and former submissions as baseline surveillance in both NCBI and GISAID EpiCov.
- Use the standard file formats available from each data source to improve the timeliness of data ingestion and analyses, which CDC performs daily.
- Where possible, use **the database tag instead of directly emailing sequences/accession to CDC.**

Instructions for NCBI Baseline Surveillance

Below we outline two methods for tagging sequence data in NCBI. The BioSample method of tagging is preferred; however, if your lab does not wish to submit BioSamples, the GenBank method is acceptable. If you or your sequencing partner(s) are already marking baseline surveillance biosamples using the purpose of sequencing field with the "Baseline surveillance (random sampling)" option outlined in the PHA4GE metadata specification, you do not have to make any changes to your sequence data tagging. PHA4GE compliant instructions for marking baseline surveillance samples through BioSample appear below.

Submitting SARS-CoV-2 metadata to BioSample (Preferred)

BioSample is a central location in which to store normalized, descriptive information about biological source materials used to generate experimental data. Metadata included in the archival BioSample database are reciprocally linked with BioProjects as well as with derived experimental data in NCBI's primary archives, including the Sequence Read Archive (SRA) and GenBank.

1. Start your BioSample submission here.
   Submission of BioSamples can be done in batches using a tab-delimited text file that describes each of the samples and attributes. You can download template files from the attributes tab within the submission portal wizard. Please use the following template for clinical SARS-CoV-2 sequence data: SARS-CoV-2: clinical or host-associated.

2. Once you choose the correct attribute package, you will have the option of using a built-in table editor or uploading a spreadsheet that includes the attributes for each of your BioSamples. Required attributes are marked with an asterisk within the built-in table editor and spreadsheet.

   The value for the following optional "purpose of sequencing" attribute should be filled in to specify "baseline surveillance (random sampling)".

3. Once you have finished registering your BioSamples, they will be assigned BioSample accession numbers that you can include within your FASTA file. These have the following format: SAMNXXXXXXXX.

Submitting SARS-CoV-2 metadata to GenBank

Your submission is important to NCBI and the global research community! Get GenBank accessions in 2 hours (average) when you submit assembled SARS-CoV-2 reads with FASTA files and source metadata. NCBI annotates all SARS-CoV-2 submissions on your behalf to ease submission.

Important: To tag your submission as part of CDC's baseline surveillance effort and make it more readily searchable once released, please follow these steps:

Updating existing submissions

If you need to add or edit the keyword in submissions, follow these instructions:

- If all records are to have the same keyword, provide a list of the relevant accessions and the keyword (purposeofsampling:baselinesurveillance) to be added to gb-admin@ncbi.nlm.nih.gov.
- If adding a mix of keywords to your records, send a two-column table where the first column is the GenBank accession number and the second column is the keyword text.

Example:

MZxxxxx1	purposeofsampling:baselinesurveillance
MZxxxxx2	purposeofsampling:environmentalsurvey

Submitting new SARS-CoV-2 data to Genbank using the web-based Submission Portal

1. Begin your submission with all required materials. Click the "New submission" button under the Submission Portal header to get started.
2. Select SARS-CoV-2 on the "Submission Type" tab.
3. Once you reach the "Sequences" tab, you will upload your FASTA file and include the following CDC-requested keyword **in this exact format in the location for keyword**: purposeofsampling:baselinesurveillance.
    - Your FASTA file should contain the following in the FASTA definition line, separated from the Sequence ID by a space
      [keyword=purposeofsampling:baselinesurveillance]

      Note: this tag should appear in ***each*** FASTA definition line. Examples:

      >Seq1 [keyword=purposeofsampling:baselinesurveillance]
      CTAGCTAGCTAGCTAGCTAGCTAGCTAG
      >Seq2 [keyword=purposeofsampling:baselinesurveillance]
      CTAGCTAGCTAGCTAGCTAGCTAGCTAG

4. Before you complete the submission, go to the "Review & Submit" tab which will show your submission's details including a preview of the GenBank record on the right side of the screen where you can see the keyword.
5. Please see the section for "Updating existing submissions" above to update your records after submission.

Submitting SARS-CoV-2 data to GenBank using FTP

1. Contact gb-admin@ncbi.nlm.nih.gov to receive an account and brief instructions from the NCBI team. They will help you get started.
2. Once you can submit via FTP, ensure that the files you place on FTP with your FASTA sequences include the following CDC-requested keyword **in this exact format in the location for *keyword*:** purposeofsampling:baselinesurveillance.
3. Your FASTA file should contain the following in the FASTA definition line, separated from the Sequence ID by a space [keyword=purposeofsampling:baselinesurveillance] Note: this tag should appear in ***each*** FASTA definition line.

    >Seq1 [keyword=purposeofsampling:baselinesurveillance]
    CTAGCTAGCTAGCTAGCTAGCTAGCTAG
    >Seq2 [keyword=purposeofsampling:baselinesurveillance]
    CTAGCTAGCTAGCTAGCTAGCTAGCTAG

4. You will receive an error report in your FTP folder if there is a problem with your submission. You do not need to resubmit previous SARS-CoV-2 data to add this keyword but can contact gb-admin@ncbi.nlm.nih.gov with any updates required for your data.

Your keyword will appear in the GenBank record and be indexed for searching.

```
LOCUS       EU865993               29903 bp    RNA     linear   VRL 03-MAY-2021
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate
            SARS-CoV-2/human/USA/CDC-xyz/2020, complete genome.
ACCESSION   EU865993
VERSION     EU865993
KEYWORDS    purposeofsampling:usbaselinesurveillance.
```

5. Please see the section for "Updating existing submissions" above to update your records after submission.

Instructions for GISAID EpiCov™ Baseline Surveillance

Single Upload

Enter and upload genetic sequence and metadata, available clinical and epidemiological data, as well as geographic data. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

- In the *Sample Information* section select "Baseline surveillance" in the *Sampling Strategy* filed

Bulk Upload

Follow the instructions for the Bulk Upload XLS. Always use the newest bulk-upload-XLS-Template.

- In the bulk-upload-XLS-Template column *covv_sampling_strategy/Sampling Strategy* enter "Baseline surveillance"

Upload your completed Excel sheet together with the FASTA-File through the Batch Upload interface. EpiCoV Curators across different time zones will be alerted and review your data. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Command Line Interface ("CLI")

Follow the instructions for the CLI bulk upload.

- Include *covv_sampling_strategy* in your .csv file and enter "Baseline surveillance"

In the event you experience any difficulties with your upload or have additional questions, please contact us for assistance at hCoV-19@gisaid.org