Recommendations for SARS-CoV-2 Sequence Data Quality & Reporting

Version 1 • March 1, 2021

APHL™

www.aphl.org

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19), emerged in late 2019 in Wuhan, China.[1] Since being identified, SARS-CoV-2 has spread across the globe resulting in over 113 million cases of COVID-19 and 2.5 million deaths.[2] In the US, public health laboratories have played a critical role in the response to the COVID-19 pandemic. As the availability of molecular and serologic/antibody based testing diagnostics has expanded, so too has the work performed by public health laboratories.

Currently, efforts are underway to improve the coordination of genomic sequencing of SARS-CoV-2 in the US. Next generation sequencing (NGS) is a tool commonly used in public health surveillance and outbreak response. There are currently multiple efforts in public and private sectors to sequence SARS-CoV-2 genomes, but they are not yet well coordinated. This includes increasing the amount of sequencing performed in state and local public health laboratories. Currently, laboratories are using different sequencing strategies and protocols, including but not limited to differences in sample selection, library preparation and sequencing platforms, bioinformatics workflows, and data interpretation, which are leading to inconsistent data quality standards and biases among SARS-CoV-2 sequences generated by public health laboratories in public databases. While these issues are not unique to public health laboratories, providing basic guidance on SARS-CoV-2 sequencing and data sharing practices will improve coordination of laboratories that have been conducting sequencing and provide expectations for public health laboratories that are currently building capacity.

Recommendations

In order to provide clear expectations and a basic standardization across public health laboratories, APHL puts forward the following recommendations.

Methods

There are a number of methods and platforms currently available for SARS-CoV-2 sequencing. APHL has a compilation[3] of publicly available protocols, training resources and other information. The US Food and Drug Administration (FDA), in collaboration with other federal partners, has developed quality-controlled reference sequence data for the SARS-CoV-2 reference strain for the US.

Data Quality and Sharing Parameters

Submitting to multiple public databases ensures public health and the broader research community access to SARS-CoV-2 sequencing data. To support this effort, public health laboratories are encouraged to submit SARS-CoV-2 consensus assemblies to the GISAID EpiCoV™ repository, the National Library of Medicine's National Center for Biotechnology Information (NCBI) GenBank and the NCBI Sequence Read Archive (SRA). **If staffing resources are limited, upload to GISAID EpiCoV™ should be prioritized.** Read user agreements for public databases carefully. GISAID has strict policies on data usage and publication.

- Upload sequences with 90% or greater genome coverage to GISAID and GenBank.
- When uploading to NCBI, create a BioProject under the US Centers for Disease Control and Prevention (CDC) umbrella BioProject (Accession: PRJNA615625) and submit each sample's metadata to BioSample under the BioProject.

SARS-CoV-2 Reference Sequence Data:

- **Reference sequence** from shotgun data under GenBank accession MT233526.1* or the original sequence from Wuhan MN908947.3**
- **Reference sequence** from target-capture data under GenBank accession MT246667.1
- **Metadata** under BioSample accession SAMN143844141 and here.
- **Raw data and protocol** from shotgun sequencing under SRA accession SRX7972536
- **Raw data and protocol** from target-capture sequencing under SRA accession SRX7988130
- **Reference material** from BEI Resources catalog NR-52281 (lot 70033135)

**FDA-controlled reference sequence*

***Original sequence from Wuhan utilized by ARTIC protocol*

- Upload FastQ or Fast5 raw (filtered) reads to SRA. If available, also upload reference aligned BAM files to SRA.
    - Only data cleared of human reads should be uploaded to SRA. NCBI released their STAT tool for human read removal.

Metadata Recommendations

To ensure the utility of the sequencing data generated by public health laboratories, APHL recommends following the Public Health Alliance for Genomic Epidemiology (PHA4GE) comprehensive metadata recommendations.

- **At a minimum**, laboratories should report collection date, submitting laboratory and geolocation of the sample. Where possible, age and sex should also be included.
- For laboratories submitting data from multiple states or jurisdictions, isolate naming should reflect the original collection site of the sample.

Frequency of Submissions to Public Repositories

To ensure timeliness of data and a complete national public health dataset, laboratories should submit data according to the following recommendations:

- Submit consensus sequences to GISAID and GenBank **at least once per week**.
- Submit data to GISAID by Friday at 5:00pm ET for inclusion in CDC's SARS-CoV-2 Genomic Sequencing Dashboard. Use a consistent institution name in the "submitting laboratory" field; this is critical for CDC to be able to accurately compile data.
- Submit data to SRA within two weeks of submission to GISAID and GenBank.
- Establish a routine that allows for convenient, timely dissemination of information to state epidemiology partners.

Data Submission Troubleshooting

Occasionally sequences will be rejected from public databases, and it can be difficult to determine the cause. Laboratories should attempt to resolve potential issues as time and resources allow. To reduce the chance of rejection, build pipelines within the laboratory to flag commonly seen problems.

- Build pipelines to flag common problems.
    - For example, the Florida Department of Health built their pipeline to include (1) an indel flag for any sample that has a deletion >50% or any insertion and (2) a stop codon flag for any sample that has a SNP that causes a stop codon at >50% frequency. Samples without flags can be immediately uploaded to GISAID and GenBank. Staff should examine flagged samples by (1) reviewing the iVar variants file to see which variant caused the flag; (2) run dnadiff to see that the indel or SNP in question was actually incorporated into the consensus assembly; and (3) review the BAM file in Integrative Genomics Viewer (IGV) if necessary depending on the situation. If your laboratory does not run iVar, you can reference the vcf file.
- **Do not manually remove the stop codon or frameshifts from your sequences and replace them with reference sequence in order to upload without flags**. Once you have confirmed these sequences using either the approach described above or through other quality control checks, upload the name of the sample, frequency of mutation and read depth at that site and submit along with an email to GISAID or Genbank with this information and they will remove the flags.

Quick Reference: Public Database Minimum Submission Criteria

The following table outlines minimum data submission requires for GISAID EpiCoV™ and NCBI Genbank. This table is provided as a quick reference to help you prepare for uploads and is subject to changes.

Minimum Criteria	
GISAID	10kb uninterrupted unambiguous bases >90% breadth-of-coverage or > 50% with warnings No evidence of contamination A confident call (unambiguous) is 10x Illumina or 20x Nanopore
GenBank	GenBank will accept any SARS-CoV-2 sequence with a min length of 50 nt. A “complete genome” record must meet the following criteria: • Minimum sequence length of 29400 • No stretches of >99 Ns in a row • 12 CDS features contained in sequence • all annotated features are complete. For example if a stretch of Ns occurs over the start codon of a protein, the CDS feature would be marked as partial and the genome would not be complete.

References

1. Burki, T. The Origin of SARS-CoV-2. The Lancet. September 2020. Available from https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30641-1/fulltext
2. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. Accessed February 27, 2021. Available from https://covid19.who.int/
3. APHL. Responding to the Coronavirus Disease (COVID-19) Pandemic. Available from https://www.aphl.org/programs/preparedness/Crisis-Management/COVID-19-Response/Pages/About.aspx

Linked Resources

Consolidated list of hyperlinks to the sites and resources referenced in this document are included below.

1. APHL SARS-CoV-2 Sequencing Resources: Find wet lab protocols, bioinformatics resources and National SARS-CoV-2 Strain Surveillance (NS3) specimen submission guidance.
2. GISAID EpiCoV™ repository
3. FDA-ARGOS SARS-CoV-2 Reference Grade Sequence Data
4. NCBI GenBank
5. NCBI Sequence Read Archive (SRA) and NCBI STAT tool for human read removal
6. Public Health Alliance for Genomic Epidemiology (PHA4GE)
7. PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology manuscript and GitHub with submission protocols
8. CDC SARS-CoV-2 Genomic Sequencing Dashboard