Pathogen Genomics Centers of Excellence Network

Standard Operating Procedure (SOP) for Computational Implementation of PGCoE Viral Sequencing Benchmarking Study

I. Purpose & Scope

This SOP outlines the standardized procedures for the computational (dry lab) component of the viral sequencing benchmarking study conducted by the Pathogen Genomics Centers of Excellence network. The study aims to compare currently implemented bioinformatics methods across participating sites for pathogen identification via whole genome sequencing analysis. This SOP applies to all participating laboratories conducting genome assembly and bioinformatics analyses as part of the benchmarking study. This study includes the following pathogens:

- SARS-CoV-2
- Respiratory Syncytial Virus (RSV) A
- Respiratory Syncytial Virus (RSV) B
- Influenza A (H1N1, H3N2)
- Influenza B
- Pathogen mixtures as described in the Network Proposal (available here)

The computational workflow is broken out into the following implementation phases:

- Phase I Computational Only: participating sites will download data from SRA (dataset defined here), analyze SRA data, and provide resultant outputs/metrics to the benchmarking team.

- Phase II Laboratory End-to-End: participating sites will first complete the wet-lab component to this study (SOP will be linked here when available). Participants will then analyze the sequencing data generated in the laboratory and provide outputs/metrics to the benchmarking team as described.

II. Computational Workflow

For both phases, follow the steps below to complete the benchmarking computational workflow:

1. Acquire raw sequencing data (FASTQ files)

- Phase I Computational Only: download the SRA dataset here to acquire FASTQ data.

- Phase II Laboratory End-to-End: utilize FASTQ files resulting from wet-lab prep/sequencing, generated on any sequencing platform.

2. Analyze sample data adhering to standard computational processes

- Phase I Computational Only:
  - First perform pathogen genome assembly with the provided reference sequence(s) (indicated here) using your standard workflow.
  - If your workflow has an option for automated reference selection, or if you have standardized practices for selecting an appropriate reference sequence, then ***repeat the analysis*** using the same workflow with the automated/selected reference.
  - If you are unable to complete assembly for any SRA samples due to workflow capability (e.g., workflow is appropriate for only data generated on the Illumina platform, and provided data was generated on the ONT platform), please note this and proceed to the next standard.
- Phase II Laboratory End-to-End:
  - Perform pathogen genome assembly following your standard workflows, selecting a reference sequence following internal SOPs.

3. Submit data to benchmarking working group

- Navigate to the benchmarking submission portal: https://benchmarking.nepgcoe.org and log in with the appropriate username and password. Within the Phase I (computational only, from SRA) and Phase II (laboratory end-to-end) folders, you will see a subfolder for your organization. Please upload all files (see below) to the appropriate Phase and Organization folders. If your organization does not have a subfolder, please create one in the appropriate Phase I or Phase II folder.
  - Note on mixtures: If your sample is a mixture, please document this one row per genome (i.e. you will end up duplicating "Sample IDs", in the case of mixtures)
  - Note on segmented assemblies: in the case of segmented genomes, these can be shared as a single multi-FASTA or as multiple files with single FASTA per segment
  - Note on missing/absent data: if you could not assemble a sample (e.g., lab sequencing didn't work), annotate the "Genome Completeness" column as "Failed"
  - Note on incompatible samples: If your lab does not have a process for a given sample/pathogen type, annotate the "Genome Completeness" column as "Incompatible"
- Please upload the following ***REQUIRED*** data:
  - Completed benchmarking data capture table (please download from here and upload a completed copy). Metrics should be computed as defined in the Network Proposal (available here) and in the Appendix below.
  - Genome assemblies (FASTA files) for all SRA or wet lab standards, as appropriate

    - [Phase II Laboratory End-to-End only] Completed laboratory data capture table (will be available here when ready)
    - [Phase II Laboratory End-to-End only] Raw reads (FASTQ files)

- Please upload, if possible, the following requested data:
    - All workflow outputs, including primary results, intermediate files, and log files describing the compute environment, software versions, and workflow parameters.

Please reach out to Andrew Lang (andrew.lang@theiagen.com) and Jared Johnson (jared.johnson@doh.wa.gov) with any questions or issues.

APPENDIX

Computational metrics to be reported in data capture table:

Genome completeness: Genome completeness will be measured as the proportion of nucleotide positions in the reference or “truth” sequence that are captured in the test sequence, as determined via sequence alignment allowing for up to 20% sequence divergence. Test sequences exceeding 20% sequence divergence from the reference will not be compared, resulting in 0% genome completeness. Suspected discrepancies in the reference sequence will be resolved by expert review.

$$Genome\ Completeness\ =\ 100\%\ *\ \frac{Test\ nucleotides\ aligned\ to\ reference}{Total\ reference\ nucleotides}$$

Genome accuracy: Genome accuracy will be measured as the proportion of nucleotide positions in the test sequence that match exactly the nucleotide identity of the reference or “truth” sequence, as determined via sequence alignment allowing for up to 20% sequence divergence. Test sequences with ≤ 80% genome completeness and/or ≥ 20% sequence divergence from the reference will not be compared, resulting in a null result.

$$Genome\ Accuracy\ =\ 100\%\ *\ \frac{Test\ nucleotides\ aligned\ to\ and\ matching\ the\ reference}{Test\ nucleotides\ aligned\ to\ the\ reference}$$

Percent Target Reads: The percent target reads will be measured as the proportion of reads that align to the reference or “truth” sequence.

$$\%\,Target\ Reads\ =\ 100\%\ *\ \frac{Reference\ aligned\ reads}{Total\ reads}$$

Average Depth of Coverage: The average depth of coverage will be measured as the mean number of reads aligned to each nucleotide in the reference or “truth” sequence. Only reference nucleotides with at least one aligned read will be considered.

$$Average\ Depth\ of\ Coverage\ =\ \frac{\sum\ \ Reads\ aligned\ to\ each\ reference\ nucleotide}{Reference\ nucleotides\ with\ at\ least\ one\ aligned\ read}$$

Average Read Phred Score (Q-score): The average read Phred score will be measured as the mean Phred score across all reads that align to the reference sequence.

$$Average\ Read\ Phred\ Score\ =\ \frac{\sum\ \ Phred\ score\ of\ reference\ aligned\ reads}{Reference\ aligned\ reads}$$