Pathogen Genomics Centers of Excellence Network

Network Proposal for Benchmarking of Viral Respiratory Genomic Sequencing and Analysis

---

I. Scope and Objectives

The primary goal of this benchmarking plan is to develop a standardized set of metrics that enable comparison of sequencing library preparation and bioinformatic processing methods for a particular set of pathogens. This is crucial for increasing the availability of approaches for generating pathogen whole genome sequence data while maintaining the consistency and reliability required by public health. This study and any outcomes are applicable only for surveillance purposes (not clinical).

The benchmarking study will consist of a planning phase and two implementation phases. In the planning phase, laboratory and computational standards will be selected for testing all pathogens of interest. In the first implementation phase, participating laboratories will be given the set of computational standards (described below) and asked to assemble pathogen genome sequences from the provided raw data. In the second implementation phase, participating laboratories will be given viral RNA samples and asked to perform both the laboratory processing (library construction and sequencing) and bioinformatics parts of the genome generation process. The provided standards will be consistent across all laboratories, though each laboratory will be at liberty to use any library preparation protocol, sequencing approach, or bioinformatic pipeline for analysis. Bioinformatic outputs will be compared across all laboratory and computational methods, resulting in a set of protocols that produce consistent results.

Pathogens of interest include: SARS-CoV-2, Respiratory Syncytial Virus (RSV) A, Respiratory Syncytial Virus (RSV) B, Influenza A, and Influenza B.

II. Selection of Standards

Selection of standards is critical to ensure rigorous comparison across laboratories. Standards will be selected for the computational phase (implementation phase I) as well as for the laboratory phase (implementation phase II) and made available to all participating sites.

Standards will include all pathogens of interest as well as pathogen mixtures, which can be a useful sample type for testing the ability to detect viruses in co-infected or contaminated samples. Mixed samples will be selected/designed to test a range of plausible genetic distances between mixtures; it is not expected that all methods will be able to segregate mixed infections.

A. Computational Standards

To avoid challenges associated with data sharing (including practical challenges around sharing the data and delays associated with establishing data sharing agreements), all computational standards will be derived from previously published data—specifically raw sequencing data (FASTQ files) submitted to the NCBI Sequence Read Archive (SRA).

A working list of SRA standards to include in the benchmarking study is available here. This list was selected to include all *pathogens of interest*, key variants or strains of each pathogen, mixtures of pathogens, both long- and short-read data, and datasets with varying quality.

B. Laboratory Standards

All standards used for development or benchmarking must have a fully sequenced genome associated with them to serve as a gold standard for comparison. Benchmarking participants will be provided with RNA from a panel of viral pathogens and mixtures (courtesy of Springer Laboratory at Harvard Medical School) with pre-established Ct values and viral copy estimates.

RNA samples will include all *pathogens of interest* (to be sequenced at various dilutions) as well as the following mixtures, which are ordered below by decreasing genetic distance:

- SARS-CoV-2
- RSVA
- RSVB
- Influenza A H1N1
- Influenza A H3N2
- Influenza B

- H13: H1N1 (ATCC VR-1469) + H3N2 (ATCC VR-777)
- RV2: RSVA2 (ATCC VR-1540) + RSVB1 (ATCC VR-3381)
- RVH1: RSVA2 (ATCC VR-1540) + H1N1 (ATCC VR-1469)

Each mixture pair will be tested in ratios of 4:1, 1:1, and 1:4.

Note on decision to start with inactivated viral RNA:

Sample background is a major factor affecting both sequence and data quality and should be considered when validating sequencing protocols. However, starting with inactivated RNA ensures standard uniformity while providing minimal barriers to shipping and sample access. As an extension of this benchmarking study, sites are encouraged to create spike-in samples by mixing inactivated viruses with background matrices and to perform their own nucleic acid extraction from a variety of typical backgrounds (nasopharyngeal swab, saliva, wastewater, etc.).

Similarly, no standard can provide a perfect amalgam to the complexity of clinical samples, and in the context of an evolving pathogen population, it is neither feasible nor desirable to maintain a set of clinical samples for use as benchmarking standards due to inadequate sample volume, variable viral

load and stochastic hindrances posed by clinical matrices, and ongoing evolution of viral variants. Our benchmarking study will not seek to directly compare results generated from clinical samples between network partners, nor to share clinical samples. However, each site will be encouraged to repeat studies on representative sets of clinical samples sourced from within their own laboratory partnerships, as available.

III. Performance Metrics

A. Computational Metrics

Computational metrics will be used to evaluate genomes generated during both implementation phases i.e., genomes assembled from publicly available data ***AND*** genomes assembled from sequence data generated in participating laboratories. These metrics will be generated from the following outputs, which should be provided alongside raw data by all participating laboratories as part of their results: total reads per sample, number of reference aligned reads, reference length, and link to bioinformatics pipeline used.

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.

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

B. Laboratory Metrics

Laboratory metrics will be used to evaluate possible sources of discrepancies between generated genomes. Each metric should be provided alongside details of the instrument used for measurement. Laboratory metrics will include: optional qPCR check, cDNA concentration, pre-pooling library concentration, pooled library concentration, fragment length distribution, read length distribution (for long-read approaches), and run metrics.

IV. Methodology

A. Computational Methodology

A computational protocol will be made available to participating labs that includes the list of SRA standards to assemble, desired output variables and format, and intermediate files requested.

Planning phase:

1. Identify a set of publicly available data (SRA), subsampled to varying levels (100x, 50x, 25x, etc.) that include various edge cases (high/mid/low quality, “dirty” data, mixtures, etc.).
2. Assemble genomes from computational standards dataset using a variety of established bioinformatics methods currently in use by the Benchmarking Working Group. Compare results (see Section III above) to determine disparity between methods and identify variability between algorithms, ultimately settling on acceptability criteria and summary output format for the metrics described in Section III.

Implementation phase:

3. Participating sites will be provided with SRA data and information on pathogen type, along with a set of instructions for the computational portion of the benchmarking study.
4. Participating sites perform genome assembly and analyses on a standard set of publicly available data to generate agreed-upon bioinformatics metrics.
5. Participating sites share requested outputs along with FASTA and BAM files. Benchmarking working group will compile outputs and make them publicly available, along with notes on which pipelines are best suited for particular purposes (e.g., reference-based assembly pipelines may perform less well on mixed samples).
6. To the extent possible, bioinformatics pipelines will be made publicly available.

While the intent of this study is not establish pass/fail criteria from expected values, it will provide users with known values of various metrics across various pipelines currently available at the time of the study. The interpretation/assignment of what is acceptable (pass/fail) should be performed individually, as labs are assessing current and/or future analytical methodologies (i.e., onboarding a new pipeline or updating existing).

B. Laboratory Methodology

A laboratory protocol will be made available to participating labs that includes dilution instructions, negative/contamination controls to include in sequencing runs, and quantification checkpoints to complete during library preparation and sequencing.

Planning phase:

1. Identify a set of RNA standards of the desired pathogens and mixtures, and develop a plan for sample preparation and shipment to participating sites.

Implementation phase:

2. Participating sites will be provided with inactivated RNA samples and mixtures, along with a set of instructions for the laboratory portion of the benchmarking study.
3. Participating sites are encouraged to make serial tenfold dilutions of received samples to establish the limit of detection for each sequencing method they employ.
4. Participating sites perform library construction and sequencing, ensuring to record quantification checkpoints during the laboratory process.
5. **Participating sites proceed to computational implementation instructions**, ensuring they share raw sequencing data including FASTQs (if permissible), laboratory metrics (quantifications, fragment length, read length distribution), key bioinformatic outputs (BAM and FASTA),qualifying sequencing run metrics, and details about their methodology.
6. To the extent possible, laboratory protocols will be made publicly available (guidance on dissemination to be provided).