New England
Pathogen Genomics Center of Excellence

Instructions for applying the *phylosamp* sample size calculation framework for genomic surveillance of viral variants

Authors: Carolyn Casiello[1], Zachary Thompson[2]

I. Purpose and Scope

This document provides a workflow for assessing the number of sequences needed for viral variant detection using the *phylosamp* R package. This package incorporates methods that account for biases due to common biological and logistical factors (e.g., ability to generate a high-quality sample for sequencing, asymptomatic infection rates) into standard sample size calculations. Limitations of the methods described in this workflow are discussed in Section IV and include important underlying assumptions of the *phylosamp* calculations that may not be met depending on the sampling strategy.

This workflow aims to help answer three types of questions relevant to variant surveillance:

The first question relates to establishing sample size goals for variant detection: **1) what is the number of sequences we need to detect a variant in a population?** For example, a health department may be interested in answering this when heading into respiratory season to determine how many specimens should be requested from providers for sequencing.

Once sequencing activities are ongoing, *phylosamp* can be used to answer a second question: **2) given the actual number of sequences reported for our population, what is the probability of detecting a variant?** Both questions require making decisions about the desired parameters for variant detection, e.g., deciding the variant prevalence level you are interested in detecting (2%, 3%, 5%, etc.) and at what level of confidence (80%, 90%, 95%, etc.).

The third and final question that this workflow answers invovles variant-specific biases in detection: **3) do we suspect, or should we account for, variants that may have evolved to be easier or more difficult to detect?** The *phylosamp* package has parameters that can estimate the bias that may arise due to biological differences across variants that impact variant detection and characterization. Using SARS-CoV-2 as an example pathogen, this document includes guidance for selecting reasonable input parameters to apply the *phylosamp* package for variant surveillance and provides examples of visualizations for assessing confidence in variant detection over time.

See this 2023 article by Wohl et al. to review the peer-reviewed publication this work is based on. Additional examples on how to run various functions in the *phylosamp* package and instructions on downloading the R package can be found in these phylosamp vignettes. This method builds on the foundational work from the APHL Influenza Virologic Surveillance Right Size Sample Size Calculators and The University of Texas COVID-19 Modeling Consortium Sample Size Calculator.

[1] Massachusetts Department of Public Health, Division of Surveillance, Analytics, and Informatics
[2] Massachusetts Department of Public Health, Division of Sequencing, Bioinformatics, and Capacity Expansion

II. Workflow Overview

This sample size calculation workflow can be applied to any population of interest, such as a state, a city, a hospital population, etc. The section below gives an overview of the steps, with each step described in more detail (with example code snippets) on the following pages.

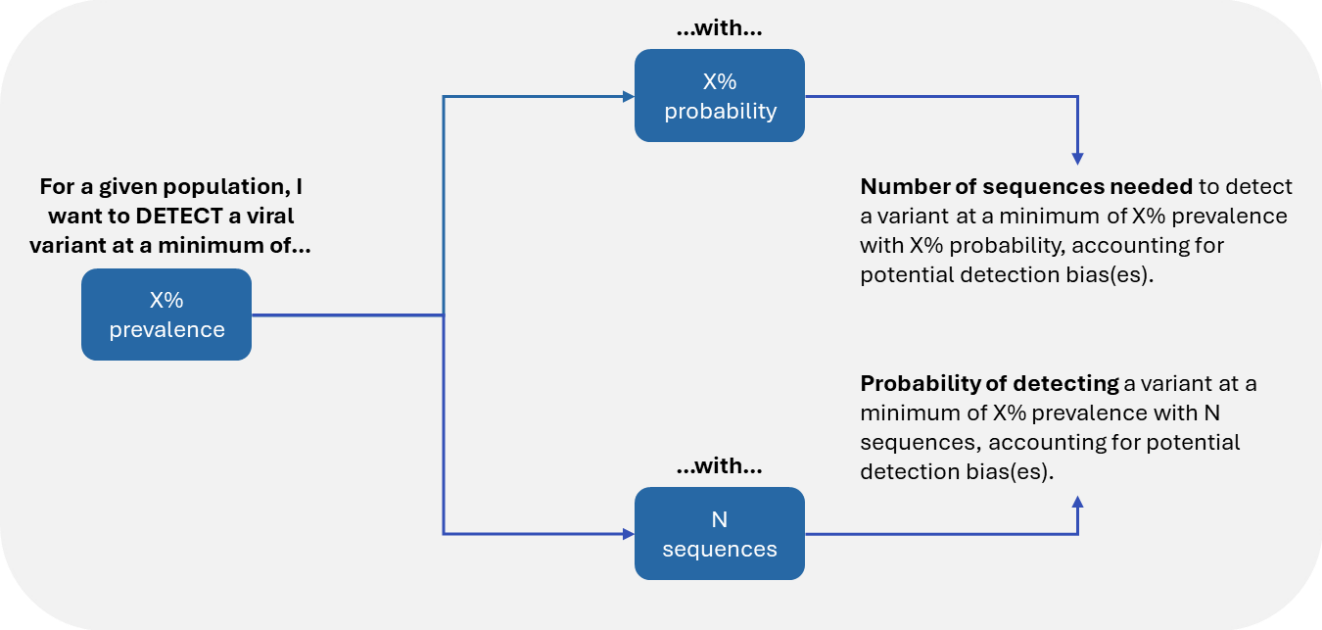

Figure 1. Workflow for power analyses for viral variant detection using the phylosamp R package.

The five steps of this workflow are:

1. **Define your detection targets** for a variant (e.g., desire to detect a variant circulating in the statewide population at 3% prevalence or higher with at least 80% confidence each week).

2. **Account for biological differences** in pathogen variants that could bias the detection of a variant compared to other variants, either by using the ***vartrack_cod_ratio*** function to calculate a coefficient of detection ratio or using an estimated ratio or range.

3. **Determine the sample size needed** for variant detection using the ***vartrack_samplesize_detect*** function.

4. **Get the probability of variant detection** based on actual sample size, targets, and bias(es) using the ***vartrack_prob_detect*** function.

5. **Calculate confidence intervals** for variant prevalence values based on actual sample sizes and variant data (note that we present an example method outside of the *phylosamp* package).

See the phylosamp vignettes for additional examples on how to run various functions in the *phylosamp* package and for instructions on downloading the R package.

III. Workflow Instructions

1 **Define your detection targets** for a variant (e.g., desire to detect a variant circulating in the statewide population at 3% prevalence or higher with at least 80% confidence each week).

- Sample size calculations require selecting a desired confidence and minimum variant prevalence to detect within a pre-defined surveillance time period. These parameters should be selected in reference to the population of interest (e.g., town, county, state, country).
- **EXAMPLE:** The Massachusetts Department of Public Health (DPH) uses the following detection targets for SARS-CoV-2 variant surveillance: **80% confidence** in detecting a variant at **3% prevalence** statewide each week. This is equivalent to approximately 96% confidence in detecting a variant at 3% prevalence bi-weekly, and was selected to align with CDC's 2024 National SARS-CoV-2 Strain Surveillance (NS3) guidance for bi-weekly jurisdictional surveillance targets for novel variant detection.

2 **Account for biological differences** in pathogen variants that could bias the detection of a variant compared to other variants, either by using the ***vartrack_cod_ratio*** function to calculate a coefficient of detection ratio or using an estimated ratio or range.

- Changes in viral variant genomics can impact detection and can be accounted for via three variant-specific parameters in the *phylosamp* R package: asymptomatic rate, test sensitivity, and probability of generating a high-quality sample for sequencing. As these variant properties change, the observed prevalence of a variant is affected (e.g., lower test sensitivity results in decreased variant detection, giving a lower observed prevalence estimate than the true variant frequency). *Phylosamp* can generate a correction factor called the **coefficient of detection ratio** that allows users to adjust variant prevalence estimates in power analyses to account for shifts in asymptomatic rate, test sensitivity, and/or probability of generating a high-quality sample for sequencing by using the ***vartrack_cod_ratio*** function.
- A coefficient of detection ratio of **1 represents a variant that is just as likely to be detected** as the comparative variant(s). A value of **<1 is a variant less likely to be detected** and results in variant prevalence underestimation, and a value **>1 represents a variant more likely to be detected** and results in variant prevalence overestimation.
- Users can also input different probabilities of testing asymptomatic and symptomatic infections (not variant specific), if known, to better understand how changes in a variant's asymptomatic rate can bias variant detection in the context of observed asymptomatic infection detection. If these parameters are not known, set them to 1.
- If asymptomatic rate, test sensitivity, and/or probability of generating a high-quality sample for sequencing cannot be estimated, we recommend using a range of coefficient of detection ratios, such as 0.5 and 1 (half as likely to detect and just as likely to detect, respectively) to represent a range of potential scenarios. For reference, coefficient of detection ratio=0.5 equates to an approximately 15-20% increase in asymptomatic rate, 15-20% reduction in test sensitivity, and a 15-20% reduction in the probability of generating a high-quality sample for sequencing for a variant of interest compared to all other circulating variants.

- **EXAMPLE:** For ***vartrack_cod_ratio*** details, see the **Estimating bias in observed variant prevalence** vignette.

3 **Determine the sample size needed** for variant detection using the ***vartrack_samplesize_detect*** function.

- Using the ***vartrack_samplesize_detect*** function:
    - Input the desired variant prevalence to detect (from Step 1), desired confidence in variant detection (from Step 1), and coefficient of detection ratio (from Step 2; either estimated from ***vartrack_cod_ratio*** or use one or more estimated values, such as 0.5 and 1).
    - Input your sequencing success rate, if known; if unknown, using a value of 0.8 inflates the necessary sample size to account for an 80% sequencing success rate. Alternatively, use 1 to get the exact number of successfully sequenced samples needed.
    - Finally, use "xsect" for the sampling framework since we are assuming this is a one-time, cross sectional sampling event (for details on how to extend the workflow provided here for a periodic sampling scenario, see the phlyosamp vignette).
    - The resulting value will be the sample size needed (either successfully sequenced samples if success rate=1 or total samples accounting for sequencing failures if success rate≠1) to detect a variant in the presence of detection bias(es), if any, at the desired confidence and prevalence.
- **EXAMPLE:** For ***vartrack_samplesize_detect*** details, see the **Estimating the sample size needed for variant monitoring** vignette.
- **CODE:** For a working ***vartrack_samplesize_detect*** example, see **Example Code Step 1: Get sample size targets**.
- **EXAMPLE:** Massachusetts DPH uses the following approach for SARS-CoV-2 variant surveillance: along with our desired level of 80% confidence in detecting a variant at 3% prevalence, we applied estimated coefficient of detection ratios of 0.5 and 1 to determine sample size targets under two bias scenarios, resulting in sample size targets of 53-105 successfully sequenced specimens per week statewide. We also evaluated our sequencing laboratory's average success rate during a period of relatively frequent testing, and factored this information into requests for specimens.

4 **Get the probability of variant detection** based on actual sample size, targets, and bias(es) using the ***vartrack_prob_detect*** function.

- Using the ***vartrack_prob_detect*** function:
    - Input your number of successfully sequenced samples for the timeframe of interest, a sequencing success rate of 1, desired variant prevalence to detect, desired confidence, and detection ratio (either estimated from ***vartrack_cod_ratio*** or use one or more estimated values, such as 0.5 and 1).

- Alternatively, you can input your number of samples attempted to be sequenced and account for sample dropout due to sequencing failure by inputting your lab's sequencing success rate (e.g., 80%).
- Finally, use "xsect" for the sampling framework.
- The resulting value will be the probability of detecting a variant at the target prevalence given your sequencing sample size (either successfully sequenced samples if success rate=1 or total samples accounting for sequencing failures if success rate≠1) in the presence of estimated detection bias(es).

- **EXAMPLE:** For ***vartrack_prob_detect*** details, see the **Estimating the probability of detecting a variant** vignette.
- **CODE:** For a working ***vartrack_prob_detect*** example, see **Example Code Step 2: Get probability of detection for each week and compare to target n**.
- **EXAMPLE:** For examples of tracking the probability of SARS-CoV-2 variant detection based on the number of successfully sequenced clinical surveillance samples reported weekly to the Massachusetts DPH, see **Figure 1 in Table 1 in example output**.

5 **Calculate confidence intervals** for variant prevalence values based on actual sample sizes and variant data (note that we present an example method outside of the *phylosamp* package).

- 95% Wald confidence intervals (CIs) can be calculated for an exploratory understanding of the precision of variant prevalence point estimates (see limitations of this method in Section IV).
- **CODE:** For a working 95% CI calculation example, see **Example Code Step 3: Get Wald 95% CIs**.
- **EXAMPLE:** For an example displaying 95% CIs around weekly variant prevalence point estimates alongside variant proportions graphs, see **Figure 2 in example output**.

IV. Limitations to keep in mind

The calculations described in this workflow provide sample size guidance, but there are several limitations of the methods used that may affect the true sample size needed or the true probability of variant detection. For example, the method assumes that sampling is representative of the infected population, except in ways that are explicitly accounted for by the selected parameters (e.g., asymptomatic rate). When sampling is not representative, the results will not always be generalizable to the population of interest. We recommend describing this limitation when providing interpretations of your data (**see example output, Figure 1** for an example) and pairing these analyses with additional evaluations of representativeness, if possible.

Another limitation is the use of Wald confidence intervals to convey precision of variant prevalence estimates. These intervals are designed to be exploratory and should not be used to assess statistically significant differences between sampling events for a specific variant or between variants within a sampling event. Doing so would result in an increase in false discovery rate (i.e., saying there is a difference between comparative groups when there is no difference). These intervals should solely be used as an estimate of precision for an individual variant prevalence estimate. Increased sampling (enough for a binomial distribution to approximate a normal distribution) and potential

multiple comparison corrections could expand the use of the Wald intervals to overcome these shortcomings.

Finally, these methods are not designed to assess changes in variant prevalence over time and are meant for snapshot descriptions of variants at a chosen point in time. While this is a drawback of cross-sectional sampling as it is portrayed in these methods, the *phylosamp* package allows for periodic sampling power analyses if the user has access to or can reasonably estimate the additional parameters required (e.g., daily logistic growth rate, starting variant prevalence).

It should be noted that while this workflow does not cover power analyses for variant prevalence point estimation, the *phylosamp* package does contain functions that allow users to generate sample sizes for estimating a minimum variant prevalence with a desired precision level (or, conversely, generating confidence in prevalence point estimates given a sample size and desired precision level).