User’s guide to variant detection calculators

Meyers Lab[*]

The University of Texas at Austin

March 2, 2021

On our website[1], we provide a set of tools to support genomic surveillance efforts for public health agencies to detect novel variants of SARS-CoV-2. In this setting, we assume that there is a pool of SARS-CoV-2 positive test specimens from which a random sample is chosen for genomic sequencing to detect an new emerging variant. We present these tools as a set of two calculators.

The main distinction between these two calculators is that calculator 1 is *prospective*—it computes how many samples must be sequenced moving forward to detect a variant—while calculator 2 is *retrospective*—that is, it addresses sensitivity of variant detection with a number of samples already sequenced.

Calculator 1. Sample Size for Detecting COVID-19 Variants

This calculator is designed to answer the question: How many SARS-CoV-2 positive samples must be collected and sequenced to detect a novel variant of the virus? For this calculator, the user supplies two values. First is the detection threshold, or the prevalence at which you want to first detect the variant. Second is the confidence level, or the desired probability of observing the variant at the detection threshold. For instance, to observe a variant at 1/400 (0.25%) prevalence with 95% confidence, there must be 1197 total specimens sequenced.

A screenshot of the interactive calculator is shown in the top panel of the figure on the next page. The x-axis corresponds to the detection threshold, and the color of the curves represents the confidence level. As the desired detection threshold becomes more rare, there are more sequences needed. Similarly, greater confidence levels require a greater number of sequences.

Calculator 2. Speed of Variant Detection

This calculator is designed to answer the question: Given that there have already been a number of sequences performed, how early can an emerging variant be detected? For example, suppose that 1197 total specimens have been sequenced and there has not been a variant detected in all of these. In this case, with a confidence level of 95% we can say that there is *not* a variant present with greater than 1/400 (0.25%) prevalence in the total SARS-CoV-2 population.

[*] Email to spencer.woody@utexas.edu

[1] https://covid-19.tacc.utexas.edu/dashboards/variants/

A screenshot of calculator 2 is shown in the bottom panel of the figure. The x-axis corresponds to the number of specimens already sequenced, and again the color of the curves represents the confidence level. With a greater number of sequences we have increased sensitivity to detect a rare variant. However, a higher confidence level leads to less sensitivity of variant detection with a fixed number of sequences.

Appendix: Assumptions

The main assumption for our calculators is that the pool of SARS-CoV-2 positive test specimens is representative of the entire SARS-CoV-2 positive population. This assumption may not hold depending on the testing regime and the characteristics of the variant.

For example, imagine a scenario where testing is performed only on symptomatic patients and there is a novel variant which is more likely to produce symptoms (i.e., has a lower asymptomatic rate compared to the wildtype). In this case, the pool of test specimens is skewed toward those which are positive for the variant, and our estimate of prevalence of the variant will be biased upward.

Importantly, there is no assumption made here concerning geography—the calculations hold regardless of geographic scale. For example, as stated previously, to observe a variant at 1/400 (0.25%) prevalence with 95% confidence, there must be 1197 total specimens sequenced. This means that these 1197 sequenced can be collected nationwide and across the states to detect a variant at 1/400 prevalence within the U.S. as a whole, or collected within one state to detect the variant which is at this prevalence within the state only.

Appendix: Derivation of calculator equations

Let N be the number of SARS-CoV-2 samples sequenced, ϕ be the presumed prevalence of the novel variant of SARS-CoV-2 (e.g. 0.5%), and p be the desired confidence level (e.g., 95%). Equivalently, p is the probability of observing the variant at least once among N random samples of the population. Using laws of probability, p is then given by

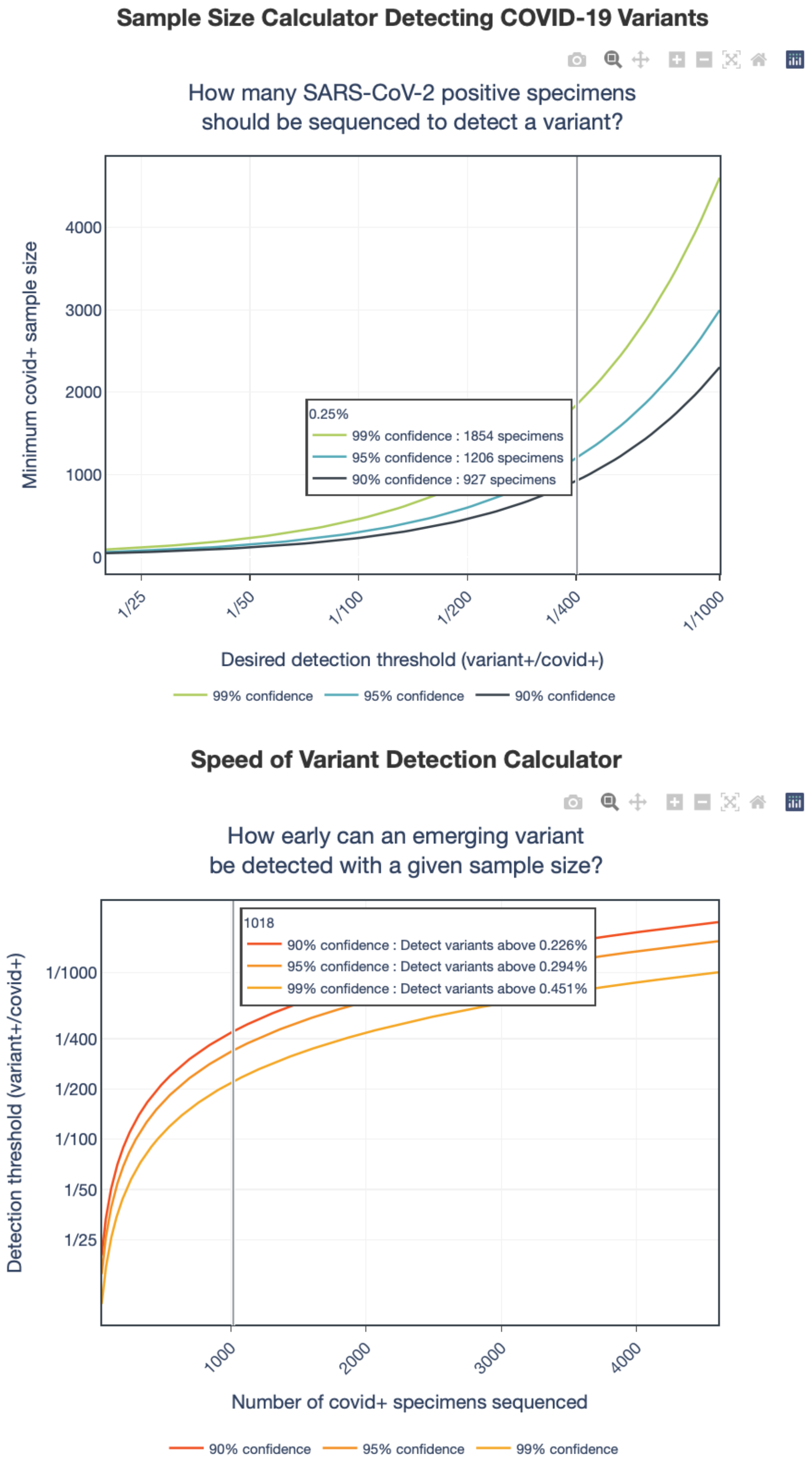
$$p = 1 - (1 - \phi)^N. \tag{1}$$

Algebraic rearranging of Eq. (1) gives an equation for the first calculator, that is, the number of sequences N needed to observe the variant at least once with confidence level p,

$$N = \frac{\log(1 - p)}{\log(1 - \phi)}.$$

Further rearranging of Eq. (1) also gives an equation for the second calculator, that is, the maximum prevalence of a novel variant which may be detected at least once with probability p given that N samples have already been collected and sequenced,

$$\phi = 1 - (1 - p)^{1/N}.$$



Sample Size Calculator Detecting COVID-19 Variants

How many SARS-CoV-2 positive specimens should be sequenced to detect a variant?

0.25%
99% confidence : 1854 specimens
95% confidence : 1206 specimens
90% confidence : 927 specimens

Speed of Variant Detection Calculator

How early can an emerging variant be detected with a given sample size?

1018
90% confidence : Detect variants above 0.226%
95% confidence : Detect variants above 0.294%
99% confidence : Detect variants above 0.451%

Figure: Screenshots of calculator 1 for sequence sample size to detect novel variants (top), and calculator 2 for the speed of variant detection with a fixed number of sequences (bottom).