February 10, 2022

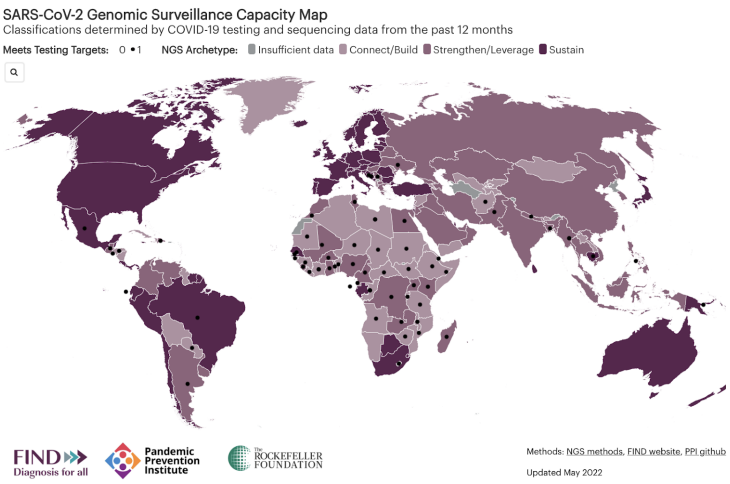
SARS-CoV-2 Sequencing Surveillance Capacity Map, Redefining Archetypes

Abstract

The global response to the COVID-19 pandemic has highlighted enormous inequities in all facets of disease surveillance and control. Surveillance tools – such as diagnostic testing and genomic sequencing – are critical to understand transmission dynamics and evolution of emerging pathogens, as well as to provide local situational awareness on circulating variants. Using real-time diagnostic testing and genomic sequencing data, we developed a tool to assess country-level COVID-19 diagnostic testing and SARS-CoV-2 sequencing surveillance capacity in real-time.

The SARS-CoV-2 Genomic Sequencing Surveillance Capacity Map is a monthly updated country level dashboard based on genomic sequencing data, NGS (Next Generation Sequencing) facility access data, and COVID-19 diagnostic testing data. Because its complexities are shared with other infectious agents – we intend for the framework presented here to be readily extendable for coordinating and relaying surveillance capacity for other infectious diseases. Our tool is useful for a wide range of audiences – from academic researchers looking to understand where we have gaps in our surveillance systems to global decision-makers aiming to reduce disparities by devising adaptive, effective control strategies of moving targets.

In this document, we will present the first version of our dashboard and the methods behind developing the targets and associated recommendations. This project will serve as a tool to coordinate the expansion of disease surveillance data collection to improve data transparency, coordinate efforts, and inform decision making.



SARS-CoV-2 Sequencing Surveillance Capacity Map, Redefining Archetypes

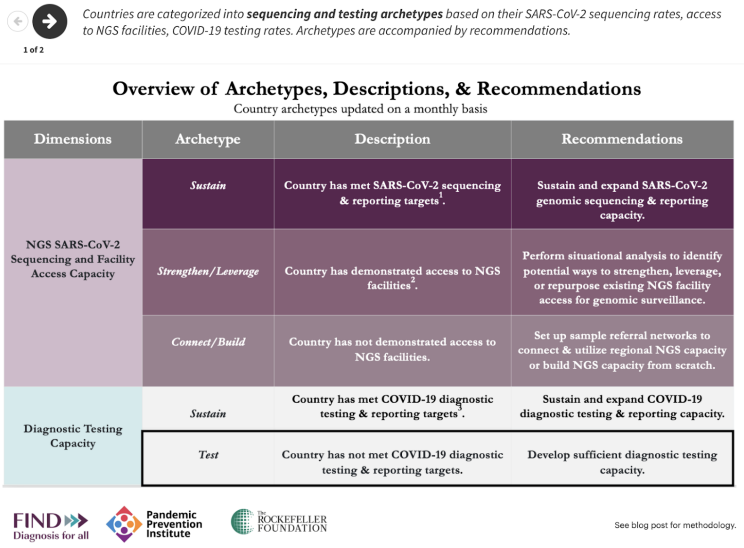


Figure 1: The SARS-CoV-2 Genomic Surveillance Capacity Map visualizing and describing intervention-based archetypes and recommendations, as of May 2022.

1.0. Introduction

Genomic surveillance is essential to uncovering pathogen threats, understanding the spread of disease, and detecting emerging variants (Amman & Markt, 2022, 22). By integrating multiple data streams– such as SARS-CoV-2 genomic surveillance, COVID-19 case data, and socio-economic data – The Pandemic Prevention Institute (PPI) & FIND Alliance for Diagnostics developed a dashboard aimed at providing actionable recommendations alongside supporting metrics that highlight disparities in diagnostic and surveillance capacity across countries.

The Access to COVID-19 Tools (ACT) Accelerator mission to accelerate equitable access to COVID-19 testing, vaccines, and treatments, is echoed in the SARS-CoV-2 Genomic Surveillance Capacity Map (referred to as the Capacity Map going forward), with a focus on illuminating inequities in genomic sequencing and diagnostic testing in lower and middle income countries (LMICs) (The WHO et al., 2020). The map is intended to be used to routinely monitor the state of diagnostic and genomic surveillance capacity to help target mitigation strategies in resource-limited settings. The metrics and targets used in the Capacity Map are designed to be robust to pandemic waves and reflect recent surveillance.

2.0. Methods

2.1. Study Population & Metrics

This updated version of the Capacity Map provides real-time surveillance data, to best inform the current state of the diagnostic and sequencing landscape globally (FIND & PPI, 2022). The Capacity Map is composed of country-level SARS-CoV-2 sequencing data, COVID-19 case data and COVID-19 testing data reported daily over the past 12 months. PPI and FIND plan to update data and subsequent intervention-based archetypes monthly. Due to collaboration, partnership, and reliability of the FIND Test Tracker platform (FIND, 2022), the countries and metrics represented in the FIND Test Tracker were used as the main study population for designing the targets and associated recommendations in the Capacity Map (Figure 2).

Figure 2: FIND COVID-19 Test Tracker assessing country's Positivity Rate correlation with the average daily testing rate per 1000 persons.

Data from January 2021 through January 2022 was used to determine the targets to evaluate countries' sequencing and testing levels. We narrowed our study population to the 199 countries and territories in the FIND test tracker database. Following deduplication, removal of countries with missing testing data (n = 31), and removal of countries who had not reported tests in six months (n = 20) the final study population yielded 147 countries/territories. Due to the product focus on resource-limited countries, LMIC countries' (n=95) surveillance were used to design targets (Figure 3).

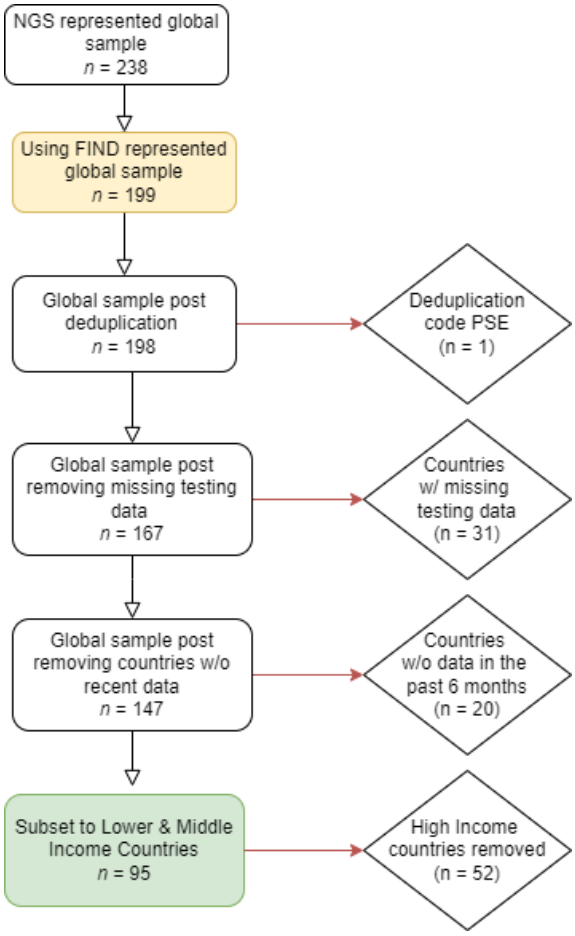


Figure 3: Flowchart illustrating the study population starting from n=199 and yielding a sample size of n=95 after data processing and subsetting to LMICs.

2.2 Data Sources

SARS-CoV-2 genomic sequencing was assessed using the global genomic sequencing repository GISAID (GISAID, 2022). GISAID provides transparent data sharing for genomic sequencing data for SARS-Cov-2 and influenza, in addition to establishing mandatory recognition of sequencing laboratory contributions. Sequencing metadata for SARS-CoV-2 is available publicly for registered users.

Real-time COVID-19 diagnostic surveillance was obtained from the FIND COVID-19 Test Tracker platform (FIND, 2022). COVID-19 case data was obtained from Our World in Data (OWID), a publicly available repository of COVID-19 data (Our World in Data, 2022).

NGS facility access data were extracted from WHO and proprietary manufacturer device data provided to FIND in spring of 2021 (The WHO et al., 2020). The World Bank (WB) socioeconomic data were used to assess income levels by country (The World Bank, 2022).

2.3. Data Processing

2.3.1. Country-level Aggregation

SARS-CoV-2 sequences submitted to GISAID were summed by country based on collection date, with sequences needing to be both collected and submitted within the 12 month period. Country iso3 codes were consistently used in the aggregation process, to avoid issues in case-sensitivity or misspelling of country names at the time of individual sequence submission.

The FIND COVID-19 daily testing data were aggregated over the 12 month study period into country-level metrics, reflecting the available data for each country (FIND, 2022). The COVID-19 case data were truncated to the last date of reported testing data. For all days prior to the last day of test reporting, the FIND Test Tracker assumes that the tests reported are evenly distributed across all prior days without test reporting, so each day is assigned a number of tests in the "all_new_tests" column of the FIND dataset. The sum of truncated cases and average smoothed number of daily reported tests were aggregated per country, over the time window in which tests were reported within the last 12 months.

The WHO sequencing facility data and WB socio-economic data were already aggregated to the country-level. Therefore, the FIND testing, GISAID sequencing, WHO facility, and WB economic country-level dataset were joined into a single dataset by country iso3 code, to reflect metrics over the past 12 months. For further data information, The PPI NGS Capacity Map github contains repositories containing data with all necessary metrics, metric definitions, codebooks, and the scripts used to calculate the metrics from the raw data sources.

2.3.2 Metric Calculations

2.3.2.1 SARS-CoV-2 Sequencing Metrics

SARS-CoV-2 sequencing levels are evaluated by: 1) percent of cases sequenced, and 2) total number of sequences per 100,000 persons, over the past 12 months. The percent of cases

sequenced represents the proportion of COVID-19 cases that were selected for SARS-CoV-2 genomic sequencing and reported to GISAID within the timeframe. Effective genomic sequencing should both maintain a baseline level of surveillance and be able to surge during times of increased incidence. We intended these two metrics to capture both the surge capacity (can a country continue to sequence X% of cases as cases increase?) as well as the overall annual levels of surveillance.

The numerator – the number of COVID-19 cases sequenced and submitted per country in the past 12 months – is not exactly equivalent to the number of sequences submitted in that timeframe, as sequences submitted but collected outside (after) the timeframe were excluded. We consider only the sequences that were both collected and submitted within the 12 month period, in order to assess both volume and timeliness of sequencing.

Because of the lag-time from specimen collection to sequence submission, the percent of cases sequenced metric for a country is expected to decrease as the time window shortens, so we advise caution in applying the targets decided here to other timeframes.

The percent of cases sequenced, $\boldsymbol{\rho}$, as a percent, is calculated as follows:

$$\rho = 100 \times \frac{\sum_{i=1}^{y} s_i}{\sum_{i=1}^{y} c_i} \quad \text{(Equation 1)}$$

Where s_i is the number of sequences collected on the *i*th day, c_i is the number of COVID-19 cases reported on the *i*th day, and *y* is the number of days in the12 month period.

The number of sequences per 100,000 persons is the crude number of sequences collected and reported, by the population unit of per 100,000 persons over the12 month time period. We include both the percent of cases sequenced and the per capita sequencing metric so that countries who are conducting high rates of sequencing, but also are detecting high rates of cases due to sufficient testing, are not penalized for sequencing a low percent of their overall cases. The number of sequences per 100,000 persons, $\boldsymbol{\nu}$, is calculated as follows:

$$\nu = \frac{100,000}{n} \times \sum_{i=1}^{y} s_i \quad \text{(Equation 2)}$$

Where *n* is the population size of the country, s_i is the number of sequences collected on the *i*th day, and *y* is the number of days in the 12 month time period. Again, we only count sequences

that were both collected and submitted within the 12 month time period, to assess both volume and timeliness of sequencing.

2.3.2.2 Diagnostic Testing Surveillance Metrics

For each represented country, diagnostic testing capacity was assessed using the following metrics calculated over the past year: 1) test positivity rate (TPR), and 2) average daily testing per 1,000 persons. TPR represents the proportion of tests during a time period, in this case the past 12 months, that result in positive samples – or infected individuals. Ideally, TPR is low, indicating that there is sufficient testing to assess the amount of circulating virus. TPR, *p*, as a percent, is calculated as follows:

$$p = 100 \times \frac{\sum_{i=1}^{\tau} c_i}{\sum_{i=1}^{\tau} d_i} \quad \text{(Equation 3)}$$

Where c_i is the number of COVID-19 cases reported on the *i*th day, d_i is the number of COVID-19 diagnostic tests reported on the *i*th day (corresponding to the "all_new_tests" column in the FIND test tracker data set), and τ is the number of days of the past year that tests have been reported. This equation effectively averages cases per tests over the past τ days, as tests reported in a single bolus are distributed evenly across all previous days since the last date of test reporting, as is described above. The time periods can vary by country, as a number of countries have delayed test reporting. For example, if a country has not reported tests within the past 30 days, the *i*=1 day would correspond to a year ago, and τ would be 335, with the *i*=335 day corresponding to the reported tests and cases 30 days from the reference (current) date. If a country had reported tests every day then τ for that country would be a full 365 days.

Average daily testing represents a central point of daily diagnostic tests conducted per country, standardized by the population unit of per 1,000 persons. The average number of daily tests per 1,000 persons, μ, is calculated as follows:

$$\mu = \frac{1000}{n} \times \frac{\sum_{i=1}^{\tau} d_i}{\tau} \quad \text{(Equation 4)}$$

Where *n* is the population size of the country, d_i is the number of diagnostic tests reported on the *i*th day, and τ is the number of days of the past year that tests have been reported, with the same smoothing of test reporting delays as described above.

2.5. Exploratory Data Analysis and Surveillance Targets

2.5.1 SARS-CoV-2 Sequencing Surveillance

Following the calculation of key sequencing metrics, distribution plots and interquartile ranges were analyzed to explore surveillance capacity. Quartile-Quartile (QQ) plots and interquartile ranges were used in combination to identify natural clustering and break-points in the data, with the goal being to set targets that helped distinguish countries that were doing sufficient amounts of sequencing from those who were sub-par. We combined this approach with qualitative validation of global health experts to develop surveillance targets that accurately reflected the landscape of genomic sequencing among LMICs.

Because case and sequencing data is not uniformly reported at more granular within country-levels, we were not able to take into consideration the representativeness of the samples sequenced, but we acknowledge that the thresholds defined here, based on quantity alone, may not truly reflect the robustness of a national sequencing program. Representative sampling of a small percent of cases is likely more effective than sampling large amounts of certain subgroups (Dean, 2022).

We identified sequencing targets of greater than 0.5% of cases sequenced and greater than 10 total sequences per 100,000 persons over the 12 month period (Figure 4). Countries are classified as having met sequencing targets if they reach both targets.

SARS-CoV-2 Sequencing Surveillance Capacity Map, Redefining Archetypes

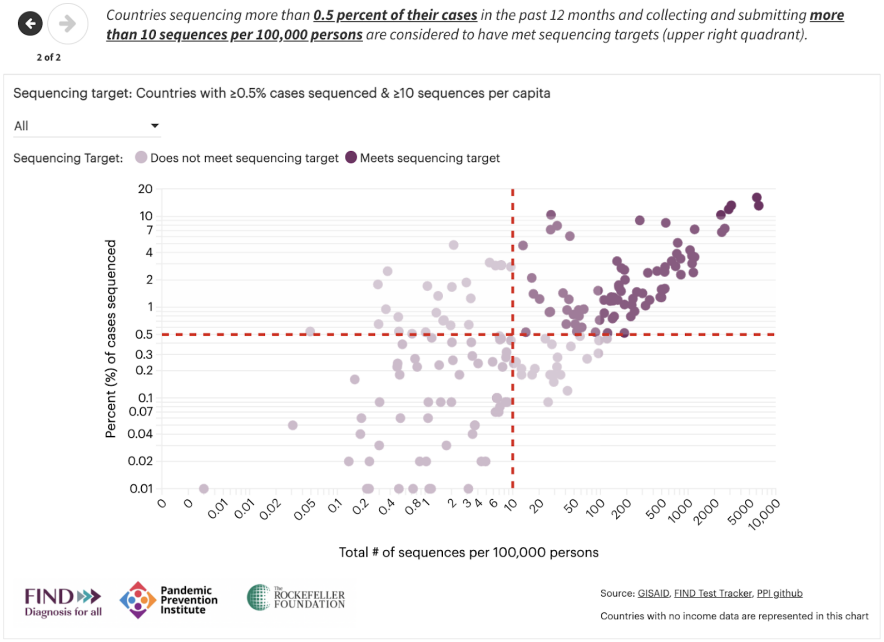


Figure 4: SARS-CoV-2 sequencing targets. Countries reporting at least 0.5% of cases sequenced and at least 10 sequences submitted and reported per 100,000 persons in the last 12 months are considered to have met sequencing targets (upper right quadrant). The red dotted lines indicate targets for both metrics.

2.5.2. COVID-19 Diagnostic Testing Surveillance

Similarly to SARS-CoV-2 sequencing, COVID-19 diagnostic testing was informed by scatter plots, interquartile ranges, descriptive statistics, natural clustering & breaking in the QQ-plots, and global health expertise.

The combination of TPR and average daily testing per 1,000 persons metrics were used to develop a definition of the COVID-19 diagnostic testing surveillance target. We identified testing targets of less than 20% TPR and an average daily testing rate greater than 0.5 average daily test per 1,000 persons. Countries with a TPR higher than 20% or an average daily testing rate less than 0.5 per 1,000 persons are considered to not have met COVID-19 testing targets (Figure 5).

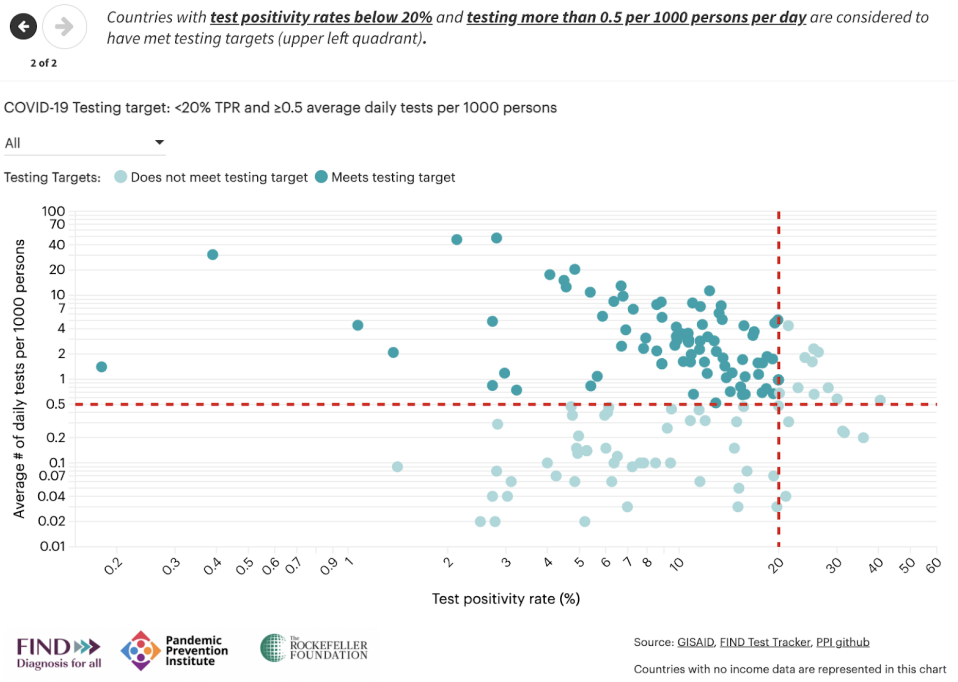


Figure 5: COVID-19 diagnostic testing targets. Countries reporting less than a 20% TPR and greater than0.5 average daily test reported per 1,000 persons in the last 12 months are considered to have met diagnostic testing targets (upper left quadrant). The red dotted lines indicate targets for both metrics.

2.5.3 NGS Facility Access Surveillance

A country's access to NGS sequencing facilities was determined by the WHO NGS facility and proprietary manufacturer device data. The data assigns a categorical value of zero; indicating the presence of no NGS facilities domestically, 1-3: indicating that at least one to three NGS facilities are based domestically, or 4+: indicating four or more NGS facilities are available. The Capacity Map consolidates this information for the countries represented in our study, and assigns a binary value to represent countries with in-country access to NGS sequencing laboratories, or country's without access. Refer to Figure 6 for an overview of the dimensions, metrics, and target definitions for the Capacity Map.

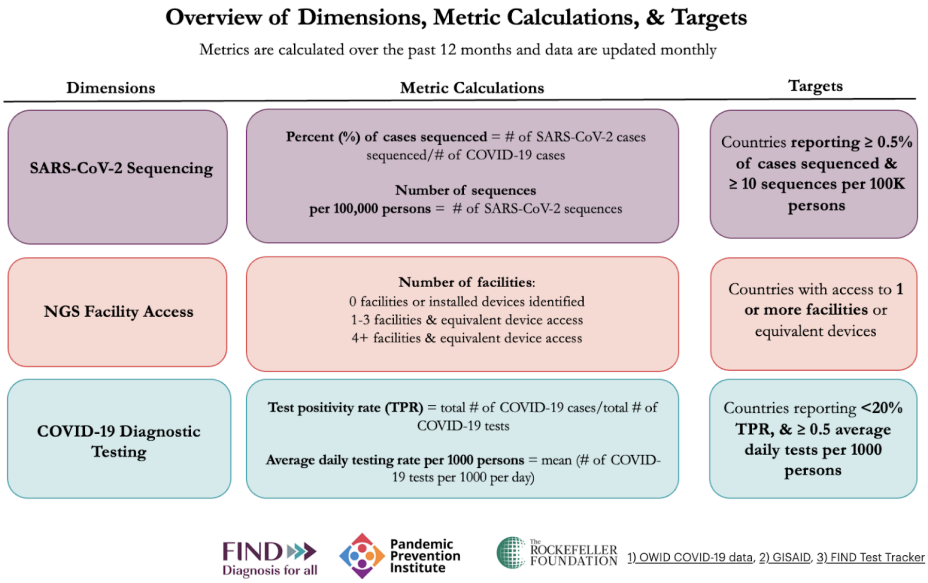


Figure 6: Overview of dimensions, metrics, and targets used in defining surveillance capacity.

2.6. Archetype Classification Scheme

Our team independently assessed SARS-CoV-2 sequencing and COVID-19 testing surveillance to assign evidence-based archetypes and recommendations (Figure 7). Using the classification framework below, countries are evaluated against targets, and classified into one of three sequencing archetypes to represent a countries' current state of SARS-CoV-2 genomic surveillance: 1) *Sustain*, 2) *Strengthen/Leverage*, or 3) *Connect/Build*. Independently, they are evaluated and classified into one of the two COVID-19 testing archetypes to reflect the current state of COVID-19 diagnostic testing surveillance: 1) *Sustain* or 2) *Test*. The full recommendations corresponding to these archetypes are described below.

SARS-CoV-2 Sequencing Surveillance Capacity Map, Redefining Archetypes

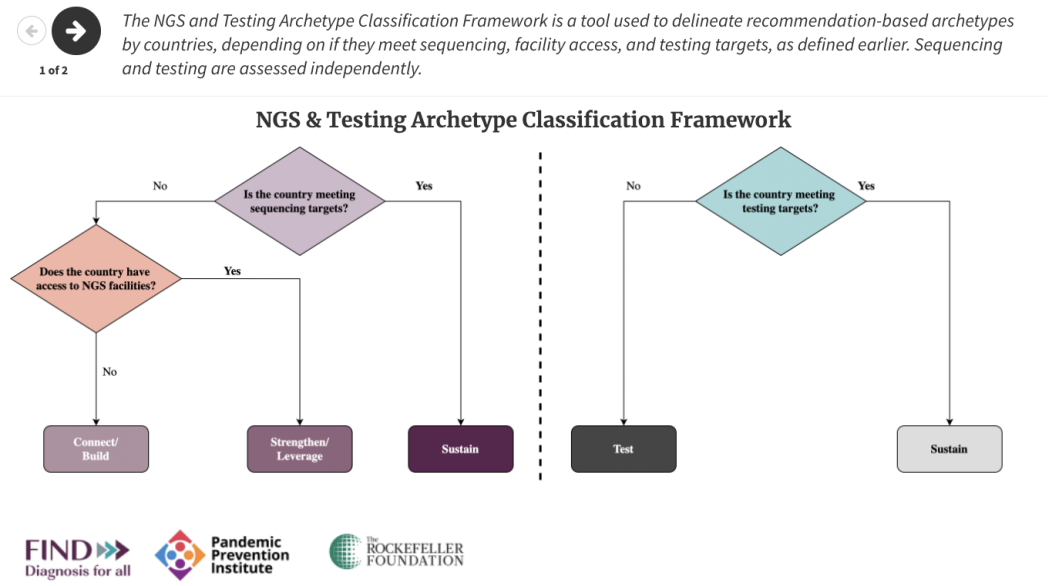


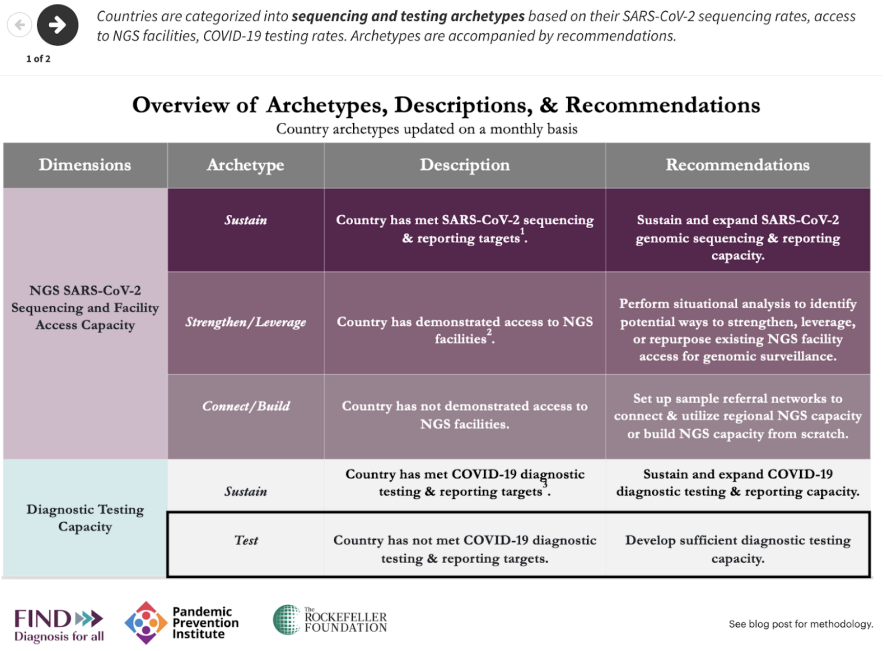
Figure 7: SARS-CoV-2 sequencing and COVID-19 diagnostic testing archetype classification framework

For both SARS-CoV-2 sequencing and COVID-19 testing archetypes, countries with missing COVID-19 case data are classified as having *Insufficient data.* Countries' who have not submitted sequencing to the GISAID repository in the past year were assumed to have zero sequences.

In the SARS-CoV-2 sequencing classification scheme, we first evaluate "Is the country meeting sequencing targets?". For the purposes of this tool, the SARS-CoV-2 sequencing surveillance target is defined as a country reporting both at least 0.5% cases sequenced and at least 10 total sequences per 100,000 persons in the previous 12 month period. If a country has reported at or above the stated targets for sequencing surveillance, then they are classified into the *Sustain* archetype. The *Sustain* archetype conveys to decision-makers that the country has met sequencing surveillance targets. These countries are then encouraged to sustain surveillance efforts.

Countries who have not met sequencing targets (including those with zero sequences reported to GISAID) proceed to the next evaluation question: "Does the country have access to NGS facilities?". A country is considered to have access to NGS facilities if the WHO facility data reports that they have at least one NGS facility domestically. If a country is missing from the WHO NGS dataset or reports zero NGS facilities, then it is assumed that they do not have access to domestic NGS facilities. Countries that have domestic NGS facilities, yet are not meeting

sequencing targets are classified as *Strengthen/Leverage*. Countries classified as *Strengthen/Leverage* are recommended to leverage or repurpose existing NGS resources to improve sequencing capacity. Countries with no access to domestic sequencing facilities are classified as *Connect/Build*, and are recommended to build NGS facilities or connect to referral networks, i.e. neighboring countries with NGS facilities, to improve sequencing levels.



*Countries are categorized into **sequencing and testing archetypes** based on their SARS-CoV-2 sequencing rates, access to NGS facilities, COVID-19 testing rates. Archetypes are accompanied by recommendations.*

1 of 2

Overview of Archetypes, Descriptions, & Recommendations

Country archetypes updated on a monthly basis

Dimensions	Archetype	Description	Recommendations
NGS SARS-CoV-2 Sequencing and Facility Access Capacity	*Sustain*	Country has met SARS-CoV-2 sequencing & reporting targets[1].	Sustain and expand SARS-CoV-2 genomic sequencing & reporting capacity.
	Strengthen/Leverage	Country has demonstrated access to NGS facilities[2].	Perform situational analysis to identify potential ways to strengthen, leverage, or repurpose existing NGS facility access for genomic surveillance.
	Connect/Build	Country has not demonstrated access to NGS facilities.	Set up sample referral networks to connect & utilize regional NGS capacity or build NGS capacity from scratch.
Diagnostic Testing Capacity	*Sustain*	Country has met COVID-19 diagnostic testing & reporting targets[3].	Sustain and expand COVID-19 diagnostic testing & reporting capacity.
	Test	Country has not met COVID-19 diagnostic testing & reporting targets.	Develop sufficient diagnostic testing capacity.

FIND Diagnosis for all · Pandemic Prevention Institute · The Rockefeller Foundation

See blog post for methodology.

Figure 8: Overview of NGS Capacity archetypes, descriptions, and recommendations.

Independently, each country is evaluated against the COVID-19 testing targets to identify countries that are not meeting testing targets. Testing targets are defined as countries reporting less than 20% TPR and greater than 0.5 average daily tests per 1,000 persons. Countries meeting both of these targets are classified as *Sustain* and recommended to sustain testing capacity. Countries reporting a TPR greater than 20% or less than 0.5 average daily tests per 1,000 persons are classified into the *Test* archetype, and recommended to improve COVID-19 testing and reporting capacity.

3.0 Limitations & Interpretations

This study is limited by the reliability and breadth of the country-level data reported to The FIND Test Tracker, OWID, GISAID, the WHO, and the WB. There are reporting biases that are not

accounted for, even though these data sources are our best source for global COVID-19 data. The fidelity of reporting data to sources that are hosted and maintained by western countries may be influenced by cultural research ideologies. For instance, LMICs may not have access or ample support to report data weekly to the FIND Test Tracker, due to various personnel or resource barriers.

Additional limitations of the data include the static and potentially incomplete nature of the NGS facility data. The WHO and proprietary data on NGS facilities was assessed in February of 2021, and therefore is not a real-time representation of the NGS facilities present as of today, nor does it fully represent the functional utilization of the existing facilities. It is also important to note that many resource limited countries use more affordable techniques for genomic sequencing, such as Oxford nanopore technologies. Unfortunately, our team does not have access to other forms of sequencing datasets outside of GISAID, however we intend to update the facilities data with real-time lab-level data on the number of unique submitting labs by country from GISAID. This will provide a real-time, up-to-date, and functional assessment of the number of submitting labs per country to be used to tailor recommendations further.

Furthermore, this map assesses the sequencing and diagnostic surveillance of a country, not accounting for the representativeness of that sequencing and testing data. Without awareness of the sampling strategy by country, our team is not able to assess the representativeness of sequencing in relation to disease prevalence. Alternative epidemiological sampling techniques – such as wastewater surveillance or targeted sampling– better ensure representative sampling of diverse populations prior to sequencing.

The Capacity Map displays descriptive statistics of SARS-CoV-2 sequencing, NGS facility access, and COVID-19 testing by country, over the past year. These descriptive statistics may be used only for situational awareness, and to share SARS-CoV-2 sequencing and COVID-19 testing global disparities. This tool should not be interpreted as a predictive or forecasting assessment of SARS-CoV-2 sequencing and COVID-19 testing surveillance levels into the future.

4.0 Call to Action

Collectively, The SARS-CoV-2 Sequencing Surveillance Capacity Map is a decision-making tool used to easily identify areas of sequencing and testing disparities and suggest evidence-based mitigation actions needed to address them (Figure 8). The archetypes correspond to strategic recommendations based on near real-time data. A “call to action” recommendation is provided per archetype, to give policy-makers, funders, interventionists, and other stakeholders an opportunity to conceptualize how the NGS data may fuel strategic actions to mitigate the observed inequities of genomic sequencing and testing world-wide.

SARS-CoV-2 Sequencing Surveillance Capacity Map, Redefining Archetypes

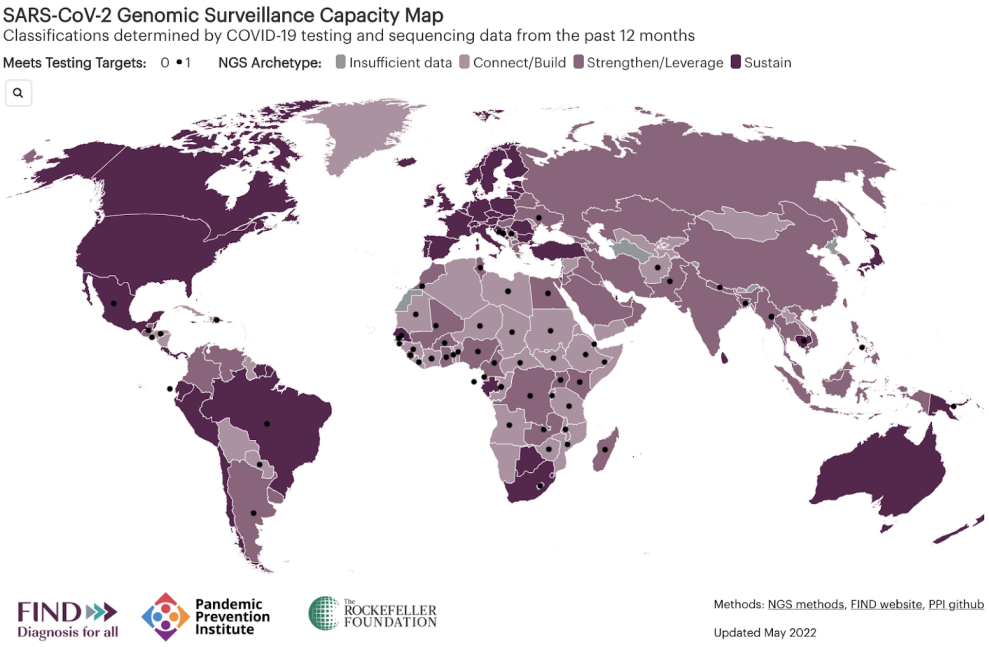


Figure 8: The SARS-CoV-2 Sequencing Surveillance Capacity Map visualizing intervention-based archetypes globally, as of May 2022.

References

Amman, F., & Markt, R. (2022). National-scale surveillance of emerging SARS-CoV-2 variants in wastewater. *medRxiV: The preprint service for Health Sciences*, 12. Retrieved January 15, 2022, from https://www.medrxiv.org/content/10.1101/2022.01.14.21267633v1.full-text

Dean, N. (2022, February 8). *Tracking COVID-19 infections: time for change*. Tracking COVID-19 infections: time for change. Retrieved March 8, 2022, from https://www.nature.com/articles/d41586-022-00336-8

FIND. (2022). *FINDCov19TrackerData/data_all.csv at master · dsbbfinddx/FINDCov19TrackerData*. GitHub. Retrieved February 14, 2022, from https://github.com/dsbbfinddx/FINDCov19TrackerData/blob/master/processed/data_all.csv

FIND. (2022, January 1). *Test tracker - FIND*. FindDx.org. Retrieved February 14, 2022, from https://www.finddx.org/covid-19/test-tracker/

FIND & Pandemic Prevention Institute (PPI). (2021, July 22). *NGS capacity mapping*. FIND. Retrieved February 14, 2022, from https://www.finddx.org/sequencing/ngs-capacity-mapping/

GISAID. (2022, March 2). *GISAID*. GISAID - Initiative. Retrieved March 2, 2022, from https://www.gisaid.org/

Our World in Data. (2022, March 2). *Coronavirus Pandemic (COVID-19)*. Our World in Data. Retrieved March 2, 2022, from https://ourworldindata.org/coronavirus

The WHO, The European Commission, France, & The Bill & Melinda Gates Foundation. (2020, April). *The Access to COVID-19 Tools Accelerator*. FindDx.org. Retrieved February 14, 2022, from https://www.finddx.org/covid-19/act-accelerator/

Wohl, S., Lee, E. C., DiPrete, B. L., & Lessler, J. (2022, January 1). Sample Size Calculations for Variant Surveillance in the Presence of Biological and Systematic Biases. *medRxiv*, 25. https://doi.org/10.1101/2021.12.30.21268453

The World Bank. (2022, March 2). *World Bank*. World Bank Group - International Development, Poverty, & Sustainability. Retrieved March 2, 2022, from https://www.worldbank.org/en/home