Nathan D. Grubaugh, PhD
Assistant Professor of Epidemiology
Yale School of Public Health

Expert testimony *presented to the House Committee on Science, Space, and Technology, Subcommittee on Investigations & Oversight during the remote hearing titled "COVID-19 Variants and Evolving Research Needs" on Wednesday, May 12, 2021 at 10:00 a.m. EDT.*

Questions presented by the Subcommittee

> 1. What is the state of data sharing regarding variants developing and spreading across the globe? How are new variants detected and, once their genomes are sequenced, how is that information proliferated?

SARS-CoV-2 genomes from COVID-19 cases have been sequenced from around the globe since the beginning of pandemic. This process, however, is expensive and technical, and thus there are significant inequities in SARS-CoV-2 genomic data generation. **Figure 1** summarizes the percent of COVID-19 cases per week that have been sequenced and shared on a public repository across regions and countries. Australia (AUS), Japan (JPN), Denmark (DNK), and Great Britain (GBR) are some of the only counties that have been able to consistently sequence >5% of the COVID-19 cases, while there is little to no SARS-CoV-2 genomic data from many countries in Asia, Africa, and the Caribbean. The United States has so far sequenced 0.5-1% of the total COVID-19 cases, though sequencing has significantly increased in recent months. These global and national genomic surveillance gaps severely limit our ability to detect new and emerging SARS-CoV-2 variants, and should be considered as a threat to US public health.

SARS-CoV-2 genomic data is primarily shared via GISAID (gisaid.org), and to a lesser extent, other repositories like the National Center for Biotechnology Information (NCBI) Genbank (https://www.ncbi.nlm.nih.gov/sars-cov-2/). As of May 6, 2021, there were 1,432,306 SARS-CoV-2 genome submissions on GISAID, compared to 386,022 on Genbank. It is unclear, however, the percent of the total SARS-CoV-2 genomes that have been sequenced that these databases represent. There are some disincentives for laboratories to not publicly share their SARS-CoV-2 genomic data. This list is not exhaustive, but it includes:

- Technical barriers to data transfers to online repositories.
- Lack of complete metadata (collection date, location, patient information).
- Lack of incentives to make expensive-to-generate genomic data available to the public.
- Lack of protection for the researchers to have first rights to publishing their data.

- Inappropriate international responses to publicly submitted data, such as naming a variant after a location or the implementation of travel restrictions.

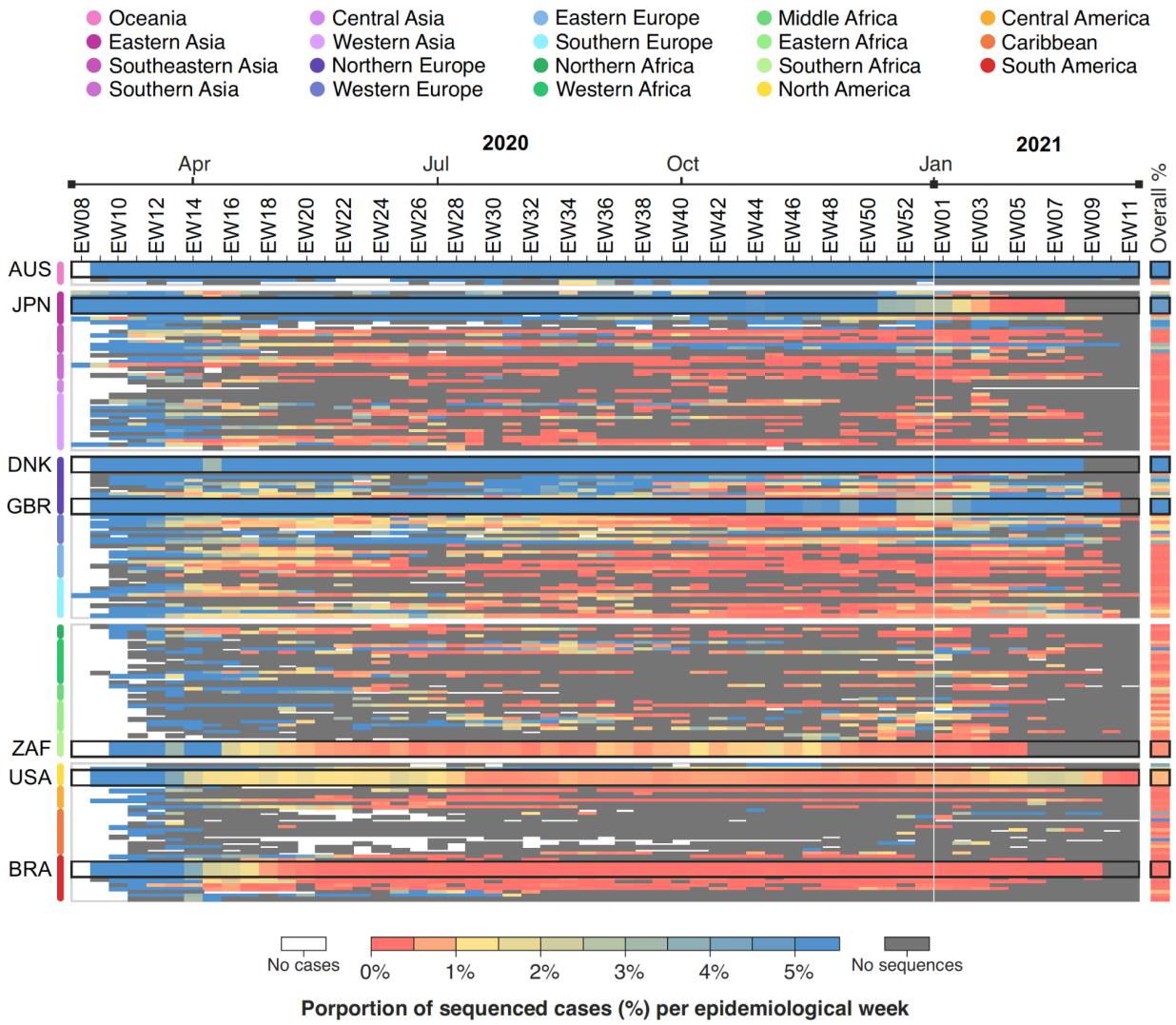

Figure 1. Proportion of sequenced cases per country per epidemiological week, 2020-2021 (up to April 16th, 2021). Few countries have capacity to sequence more than 5% of reported cases with genome coverage >= 70%, especially when COVID-19 incidence is high. When incidence is low, as in early phases of the pandemic, most countries were able to sequence high proportions of cases (3-5%, green and blue shades). However, with the aggravation of the pandemic, few countries were able to keep up, and in poor countries, despite cases being reported, many weeks had few (red) or no sequences (grey). Figure created by Anderson Brito, PhD (postdoctoral associate in the Grubaugh Laboratory at the Yale School of Public Health).

Most variants are initially detected by local laboratories or public health agencies. The SARS-CoV-2 genomic data are processed through open software like Pangolin (https://pangolin.cog-uk.io/) or Nextclade (https://clades.nextstrain.org/) that assign each sequenced to a specific lineage or clade based on the specific mutations in each sequence. This

provides an output such as “B.1.1.7” and a list of mutations. Many local laboratories or public health agencies are consistently monitoring the lineage assignments to (***1***) detect novel lineages that contain one or more mutations of interest, (***2***) detect the outside introduction of a known variant of concern or interest, and (***3***) track the frequencies of locally circulating variants.

There are also efforts to monitor for variants on national and global scales. There are now several programs, such as Outbreak.info (https://outbreak.info/situation-reports) and Nextstrain (https://nextstrain.org/), that pull data from GISAID daily to allow the user to generate custom variant tracking reports. Routine GISAID data retrievals are also used for many state and national surveillance programs to provide updates on the number of specific variants of concern or interest (e.g. https://covid.cdc.gov/covid-data-tracker/#variant-proportions). The outputs of these reports are presented on various platforms, including press releases, traditional media, and social media.

2. Are existing COVID-19 tests effective at diagnosing infections of known variants? How are variant-specific tests use to bolster public health decision-making?

To my knowledge, all known SARS-CoV-2 variants can still be detected by the common clinical diagnostic assays. While some deletions or mutations can impact individual diagnostic assay targets, most clinical diagnostic assays target multiple parts of the genome to overcome this issue. **Thus there is not currently a significant issue with variants causing inconclusive or false negative results**. However, this is an area to continuously monitor, and there are several ongoing and parallel efforts to track mutations in diagnostic targets.

The primary issue is that **standard diagnostics cannot differentiate between SARS-CoV-2 variants**. While whole genome sequencing is the gold standard for variant identification, the additional time, expense, and laboratory equipment make sequencing not practical in all circumstances. PCR and other less complicated assays have the ability to detect a limited number of virus mutations, which can be indicative of a limited set of variants. These assays, which can be faster, cheaper, and less complicated, have an advantage of being able to generate information about variant frequencies with shorter turnaround times and at a larger scale than whole genome sequencing.

For example, the SARS-CoV-2 variant B.1.1.7 has a 6 nucleotide deletion in its spike protein, which causes a spike gene target failure (SGTF) result in one of the three targets with the ThermoFisher TaqPath COVID-19 Combo Kit. The result is still valid, but by comparing the

number of positive results with and without SGTF, we can get a relative picture of B.1.1.7 prevalence. This was valuable in tracking the increasing frequency of B.1.1.7 in the UK, and it is now being used in the US. National data about B.1.1.7 provenance based on TaqPath SGTF results are provided by Helix (https://www.helix.com/pages/helix-covid-19-surveillance-dashboard). My group also uses SGTF results to help track the frequency of B.1.1.7 in Connecticut (**Figure 2**; https://covidtrackerct.com/variant-surveillance/).

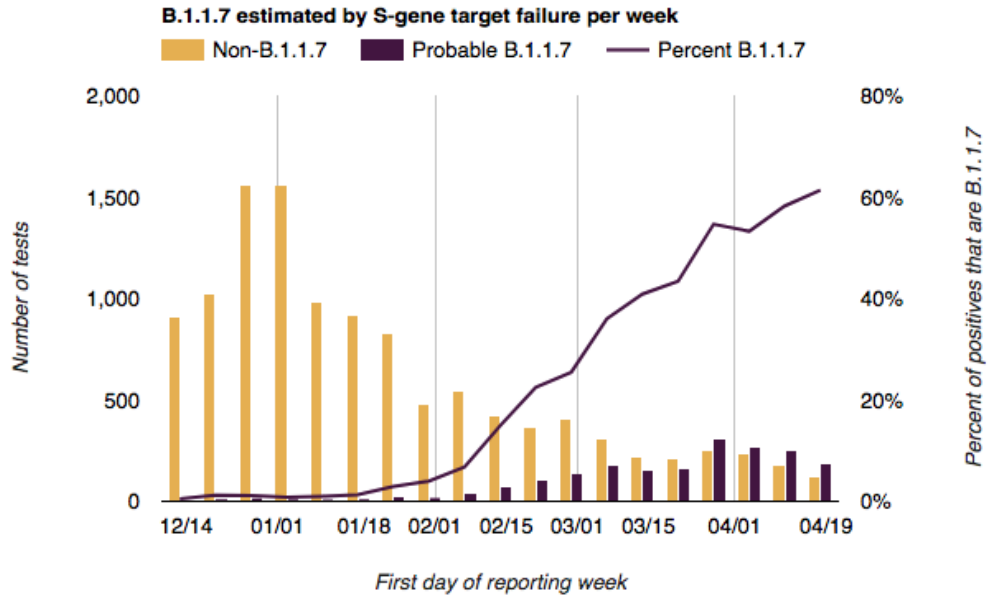

Figure 2. Presumed B.1.1.7 positivity (%). Tests performed by Yale-New Haven Hospital (primary catchment = New Haven and Fairfield Counties, CT) and Jackson Labs (primary catchment = New Haven and Hartford Counties, CT). Probable B.1.1.7 positivity defined as "spike gene target failure" (SGTF) frequency on the TaqPath SARS-CoV-2 diagnostic test. Figure from Covid Tracker CT (https://covidtrackerct.com/variant-surveillance/).

PCR assays specific to other variants have been developed, which can provide similar results to the above for B.1.1.7. These assays can be the most beneficial when they are used as the primary diagnostic test to immediately provide a SARS-CoV-2 test result and some information about the variant, rather than an add-on test. Variant-specific assays, however, cannot detect novel variants, and thus should only compliment whole genome sequencing, and not replace.

3. How can the federal government serve as a resource during and between pandemics when it comes to information aggregation and accessibility?

In my opinion, there are **three** primary ways that the federal government can facilitate data aggregation and accessibility during pandemics: policy, standards, and support.

The first is policy based. In my response to question 1, I outlined some barriers to pathogen data being shared on public repositories. It is not mandatory for data generated during pandemics that can benefit public health to be shared publicly. Furthermore, there are no policies in place to protect the rights of the data generators to have the first rights to publish the data. My group openly shares the genomic data that we generate in hopes that public health agencies can use it for decision making but also in hope that other academic labs will not scoop our data in their publication. Because some data (like sequencing data) can be very expensive to generate and publications are the "currency" for academic advancement, many groups are not open to sharing their data out of self preservation. Thus we often find data released upon a paper's acceptance in a journal. While data sharing during the COVID-19 pandemic has been exceptional, these problems continue to exist. Thus finding resolutions around the legality of data sharing and usage to create an equitable framework would enhance data sharing during future pandemics.

Second, many forms of data useful for public health, including pathogen genomic sequencing, can be generated, processed, and analyzed by applying a variety of controls and methods. Then compiling data generated among different laboratories can create biases and inaccurate findings because they may represent different populations, include different intrinsic errors, or have different definitions/classifications of data fields. Thus standardization is critical, and is only likely to come from the national level. The federal government could create panels of field-specific experts to provide standards for sample selection, data generation, computational processing, and associated metadata.

Most importantly, public health surveillance - including all aspects from data collection, generation, storage, and dissemination - needs to be fully supported outside of outbreak/epidemic/pandemic times. We have seen many "pop-up" efforts created to fill critical needs, and some of this can be alleviated with consistent support. For example, many of the online tools mentioned above (e.g., outbreak.info) were created to assist with the pandemic response, and some of them may not be supported for long after the pandemic. If the national agencies can learn from the openness and innovation of the private and academic initiatives, they may be able to help preserve these tools and expand their use beyond SARS-CoV-2. As another example, the generation of and consistent support for NCBI means that there is a database to obtain access to records and data, which is fundamental for research and public health. Expanding these programs to include surveillance data which is notoriously difficult to

obtain would help to ensure that we have systems in place for when there are public health emergencies.