Low testing rates limit the ability of genomic surveillance programs to monitor SARS-CoV-2 variants: a mathematical modelling study

Alvin X. Han, Ph.D.[1,*], Amy Toporowski, M.D.[2], Jilian A. Sacks, Ph.D.[3], Mark Perkins, M.D.[3], Sylvie Briand, M.D.[3], Maria van Kerkhove, Ph.D.[3], Emma Hannay, Ph.D.[2], Sergio Carmona, M.D.[2], Bill Rodriguez, M.D.[2], Edyth Parker, Ph.D.[4], Brooke E. Nichols, Ph.D.[1,2,5,†], Colin A. Russell, Ph.D.[1,5,†,*]

[1]Department of Medical Microbiology & Infection Prevention, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands
[2]Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland
[3]Department of Epidemic and Pandemic Preparedness and Prevention, Emergency Preparedness Programme, World Health Organization, Geneva, Switzerland
[4]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA
[5]Department of Global Health, School of Public Health, Boston University, Boston, MA, USA

†Contributed equally
*Correspondence to Alvin X. Han (x.han@amsterdamumc.nl) and Colin A. Russell (c.a.russell@amsterdamumc.nl)

Summary (293/300 words)

Background

Genomic surveillance is essential for monitoring the emergence and spread of SARS-CoV-2 variants. SARS-CoV-2 diagnostic testing is the starting point for SARS-CoV-2 genomic sequencing. However, testing rates in many low- and middle-income countries (LMICs) are low (mean = 27 tests/100,000 people/day) and global testing rates are falling in the post-crisis phase of the pandemic, leading to spatiotemporal biases in sample collection. Various public health agencies and academic groups have produced recommendations on sample sizes and sequencing strategies for effective genomic surveillance. However, these recommendations assume very high volumes of diagnostic testing that are currently well beyond reach in most LMICs.

Methods

To investigate how testing rates, sequencing strategies and the degree of spatiotemporal bias in sample collection impact variant detection and monitoring outcomes, we used an individual-based model to simulate COVID-19 epidemics in a prototypical LMIC. Within the model, we simulated a range of testing rates, accounted for likely testing demand and applied various genomic surveillance strategies, including sentinel surveillance.

Findings

Diagnostic testing rates play a substantially larger role in monitoring the prevalence and emergence of new variants than the proportion of samples sequenced. To enable timely detection and monitoring of emerging variants, programs should achieve average testing rates of at least 100 tests/100,000 people/day and sequence 5-10% of test-positive specimens, which may be accomplished through sentinel or other routine surveillance systems. Under realistic assumptions, this averages to ~10 samples for sequencing/1,000,000 people/week.

Interpretation

For countries where testing capacities are low and sample collection is spatiotemporally biased, surveillance programs should prioritize investments in wider access to diagnostic testing to enable more representative sampling, ahead of simply increasing quantities of sequenced samples.

Funding

European Research Council, the Rockefeller Foundation, and the Governments of Germany, Canada, UK, Australia, Norway, Saudi Arabia, Kuwait, Netherlands and Portugal.

Research in context

Evidence before this study

Genomic sequencing has been an integral part of the COVID-19 pandemic response, critical to monitoring the evolution of SARS-CoV-2 and identifying novel variants of interest and variants of concern (VOCs). As of March 2022, more than 10 million unique sequences had been submitted to GISAID. However, SARS-CoV-2 sequences have been disproportionately submitted from high-income countries (HICs), with large surveillance gaps existing in most LMICs. To strengthen genomic surveillance of SARS-CoV-2, previous studies focused on estimating a minimal number of positive SARS-CoV-2 tests to reflex for sequencing for effective variant detection and monitoring. We searched PubMed and Google Scholar using combinations of search terms (i.e., "SARS-CoV-2", "COVID-19", "diagnostic", "genomic surveillance", "sequencing", "LIC", "LMIC") and critically considered published articles and preprints that studied or reviewed SARS-CoV-2 testing and genomic surveillance, especially in the LMIC context. We also reviewed SARS-CoV-2 sequencing recommendations published by the World Health Organization (WHO) and European Centre for Disease Prevention and Control (ECDC). We reviewed all studies and the latest recommendations published in English up to February 2022. We found that prevailing recommendations for estimating sequencing sample size to identify or monitor the prevalence of new variants assume that COVID-19 testing is performed at high rates per capita and in high absolute numbers, such that the sequenced samples are largely representative of the circulating SARS-CoV-2 viral diversity. This is, however, not the case in many countries, particularly in many LMICs, and can vary dramatically depending on the epidemiological situation.

Added value of this study

To our knowledge, this is the first study that quantitatively estimates the joint impact of COVID-19 testing rates and sequencing strategies on SARS-CoV-2 variant detection and monitoring. We developed an individual-based COVID-19 transmission model that was specifically designed to simulate VOC emergence in LMICs under a wide range of test availability and sampling strategies for sequencing. We showed that given the current average COVID-19 testing rate of 27 tests per 100,000 people per day across LMICs, the sequencing sample size recommendations for early variant detection from WHO/ECDC and other academic groups would likely result in delayed detection of a new VOC until it had spread through a substantial portion of the population. We quantitatively demonstrated that increasing COVID-19 testing rates to at least 100 tests per 100,000 people per day, including through sentinel surveillance sites, and sampling as broadly as possible, yields far earlier VOC detection and greater accuracy of variant prevalence estimates than simply increasing the proportion of samples to be sequenced.

Implications of the available evidence

Spatiotemporal representativeness of SARS-CoV-2 positive samples being sequenced, which can be accomplished by increasing diagnostic testing rates, and widening the geographic coverage from where samples are collected, as well as shortening sequencing turnaround time are the key features of an effective genomic surveillance program aimed at detection and monitoring of novel SARS-CoV-2 variants. Only once these areas have been strengthened does increasing the volume of sequenced samples have significant impact.

Main Text (~3,500/3,500 words)

Introduction

Since the start of the COVID-19 pandemic in 2019, unprecedented expansion of genomic surveillance efforts has led to the generation of more than 10 million SARS-CoV-2 sequences deposited in the publicly accessible GISAID database (https://www.gisaid.org/) as of May 2022. These efforts have been integral to understanding the COVID-19 pandemic,[1] including the identification of the Alpha variant in the United Kingdom during the fall 2020,[2] the Delta variant in India in late 2020,[3] and the Omicron variant in Southern Africa in November 2021.[4] Despite the value of these efforts for monitoring the evolution of SARS-CoV-2, the intensity of genomic surveillance is highly heterogenous across countries. High-income countries (HICs) on average produced 16 times more SARS-CoV-2 sequences per reported case than low- and middle-income countries (LMICs) as a result of longstanding socioeconomic inequalities and consequent underfunding of laboratory and surveillance infrastructures.[5] To strengthen global pandemic preparedness, initiatives such as the Access to COVID-19 Tools Accelerator Global Risk Monitoring Framework, the Pan American Health Organization COIVD-19 Genomic Surveillance Regional Network and Africa Pathogen Genomics Initiative, among others, have supported LMICs in developing pathogen genomic surveillance programs.

As resources are finite, it is critical that sequencing sample sizes, and the diagnostic testing needed to obtain samples for sequencing, are carefully set for genomic surveillance programs to detect and monitor variants as efficiently as possible. Current recommended sample sizes are based on sampling theory[5–8] and assume that the volume of diagnostic testing is large enough such that the diversity of sampled viruses is representative of the diversity of viruses circulating in the population. However, LMICs test at a mean rate of 27 tests per 100,000 persons per day (tests/100k/day) as opposed to >800 tests/100k/day across HICs between January 2020 and March 2022 (https://www.finddx.org/covid-19/test-tracker/), with even higher testing rates in some HICs (Figure 1). Low testing rates lead to spotty information and smaller virus specimen pools available for sequencing, resulting in strong sampling biases. These factors can render efforts to monitor the emergence of new variants or prevalence of existing variants highly unreliable.

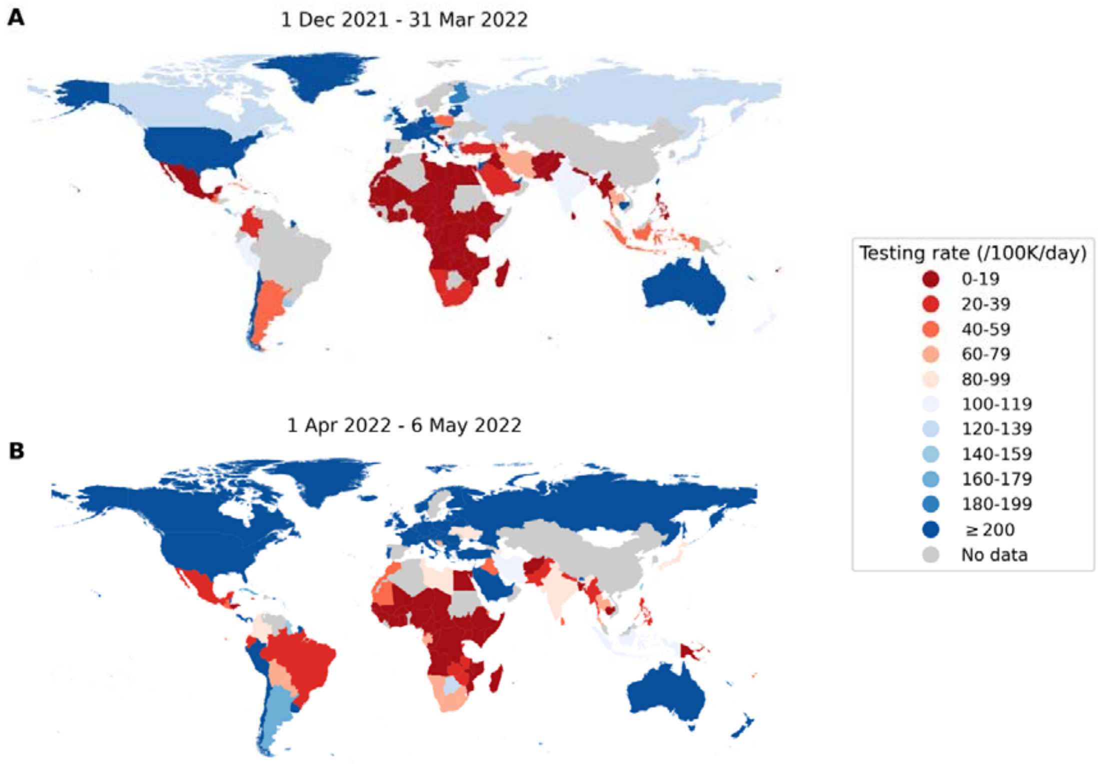

Figure 1: **Global disparities in SARS-CoV-2 testing rates**. Each country is colored by the average total number of SARS-CoV-2 tests performed per 100,000 persons per day (/100K/day) (**A**) between 1 December 2021 and 31 March 2022 when the Omicron variant-of-concern spread around the world; (**B**) between 1 April 2022 and 6 May 2022 when most countries were past peak Omicron wave of infections (https://www.finddx.org/covid-19/test-tracker/).

Here, we studied how different testing rates can impact genomic surveillance outcomes. Specifically, we used an individual-based modelling framework to simulate concurrently-circulating wild-type SARS-CoV-2/Alpha- as well as Delta-/Omicron-like epidemics in Zambia as a representative LMIC archetype. We then applied different genomic surveillance sampling strategies (i.e., sources of sample collection and varying proportion of specimens to sequence) to elucidate how testing, sequencing volumes and the degree of sampling bias arising from sources of specimens jointly impact the timeliness of variant detection and the accuracy of variant monitoring.

Methods

*Simulating SARS-CoV-2 epidemics with the **P**ropelling **A**ction for **T**esting **A**nd **T**reating (PATAT) model*

We used PATAT, a stochastic individual-based model to simulate SARS-CoV-2 epidemics in a community with demographic profiles, contact mixing patterns, and level of public health resources mirroring those typically observed in LMICs. Here, the model was based on Zambia. Briefly, PATAT creates an age-structured population, linking individuals within contact networks of multi-generational households, schools, workplaces, and churches (i.e., regular mass gatherings) (Table S1). Healthcare facilities (i.e., community clinics and tertiary hospitals) where individuals with mild symptoms seek symptomatic testing and have their virus specimens collected are simulated to approximate localized community structures based on an empirical clinic-to-population ratio. Households are proximally distributed around these facilities based on the given empirical distance distribution that correlates with probabilities of symptomatic individuals seeking testing at clinics (Table S1).

We then simulated SARS-CoV-2 infection waves in a population of 1,000,000 individuals over a 90-day period that begins with an initial 1% prevalence of an extant SARS-CoV-2 variant and the introduction of a mutant variant at 0·01%. We assumed that clinic-based professional-use Antigen Rapid Diagnostic Tests (Ag-RDTs) form the basis of testing given persistent reports that polymerase chain reaction (PCR) tests are poorly accessible for detection of symptomatic cases in most LMICs.[9] As Ag-RDT sensitivity depends on within-host viral loads,[10] PATAT generates viral load trajectories, measured in cycle threshold (Ct) values, for infected individuals by randomly sampling from known viral load distributions of different SARS-CoV-2 variants.[11,12] We performed simulations for two variant replacement scenarios – Alpha variant introduction while the wild-type virus was circulating (wild-type/Alpha) and Omicron (BA.1) variant introduction while Delta was circulating (Delta/Omicron), applying known distributions of their peak viral load, incubation, and virus clearance periods[11,13] (Table S1). Before simulating the two-variant epidemic, we first calibrated the transmission probability parameter for the extant variant such that it would spread in a completely susceptible population at R_0 = 2·5-3·0. We then assumed Alpha and Omicron (BA.1) were more transmissible that the respective extant virus to achieve growth rates of ~0·15/day and ~0·35/day respectively.[2,14]

For both sets of simulations, we assumed that 10% of the population had infection-acquired immunity against the extant strain initially with some level of protection against infection by the mutant virus (wild-type SARS-CoV-2: 80% protection against Alpha;[15] Delta: 20% protection against Omicron[14]). We also investigated the scenario where 40% of the population had infection-acquired immunity as part of sensitivity analyses (see below). We did not investigate scenarios involving vaccine-acquired immunity due to low vaccine uptake in most LMICs.[16]

PATAT uses the SEIRD (Susceptible-Exposed-Infected-Recovered/Death) epidemic model for disease progression and stratifies infected individuals based on their symptom presentation (asymptomatic, mild, or severe). After an assumed random delay post-symptom onset (mean = 1 day; s.d. = 0.5 day), symptomatic individuals who seek testing would do so at their nearest healthcare facility, where test-positive samples may be reflexively collected for sequencing. We assumed that symptomatic individuals sought testing based on a probability distribution of health services-seeking behaviour that inversely correlates with the distance between the individual's household and the nearest healthcare facility (Table S1)[17]. We varied levels of Ag-RDT stocks per day (i.e., 27, 100, and 200-1,000 (in increments of 200) tests/100k/day), running 10 bootstrap simulations for each testing rate. Given the start of a week on Monday, we assumed that a week's worth of tests are delivered to healthcare facilities every Monday and unused Ag-RDTs in the previous week are carried forward into the next week. Due to overlapping symptoms between COVID-19 and other respiratory diseases, a proportion of available Ag-RDTs would be used by individuals who are not infected with SARS-CoV-2. Based on test positivity rates reported by various countries in the second half of 2021,[18] we assumed 10% test positivity rate at the start and end of the simulated epidemic, and 20% test positivity at its peak, linearly interpolating the rates between these timepoints. We also assumed that false positive specimens could be sampled based on reported Ag-RDT specificity of 98·9%.[10]

We assumed that any specimens collected for genomic surveillance after positive detection through Ag-RDT would be reflexively confirmed with PCR. We also assumed that all symptomatic individuals who have severe symptoms require hospitalization, and are tested separately from mild symptomatic persons who sought testing. Given that likely only ~10-20% of COVID-19 deaths in Zambia were tested for the disease in life,[19,20] we assumed that only 20% of individuals with severe disease would be tested by Ag-RDT or PCR upon presenting severe symptoms and have specimens collected for sequencing.

The list of parameters and full technical details on PATAT can be found in the Supplementary Appendix. The model source code is available at https://github.com/AMC-LAEB/PATAT-sim.

Genomic surveillance strategies

Twenty percent of healthcare facilities were assumed to be tertiary facilities based on empirical data collected from Zambia.[21,22] We assumed that tertiary facilities provide testing for mild symptomatic individuals as well as hospitalized patients with severe symptoms. We also assumed that a proportion of tertiary facilities serve as sentinel surveillance sites that reflexively collect SARS-CoV-2 positive samples for sequencing. We then simulated six strategies with varying degrees of sampling coverage where positive specimens collected from testing sites would be consolidated and sampled for sequencing: (1) all samples from community clinics and tertiary hospitals are sent to a centralized facility and further sampled for sequencing (i.e. *population-wide* strategy); (2) only *one* tertiary sentinel facility for the population of 1,000,000 simulated people would sequence a portion of positive specimens it has collected, both from mild individuals seeking symptomatic testing and severe patients who sought tertiary care at the facility; or only (3) 10%, (4) 25%, (5) 50%, and (6) 100% of all tertiary sentinel facilities would sample and sequence a proportion of the specimens they have collected.

For all strategies, we assumed that a proportion (1%-100%; in 2% increments between 1% and 5%, in 5% increments between 5% and 100%) of positive specimens are collected daily for sequencing. We also assumed that positive specimens sampled within each week for sequencing are consolidated into a batch before they are referred for sequencing. Turnaround time refers to the time between collection of each weekly consolidated batch of positive specimens to the acquisition of its corresponding sequencing data. Since the within-host viral loads of infected individuals were simulated, we assumed that only high-quality samples where Ct values < 30 could be sequenced and that sequencing success rate is 80% as assumed in other studies.[6]

For each strategy and sequencing proportion, we performed 100 bootstrap simulations for each epidemic simulation with a given test stock availability, thus totaling to 1,000 random simulations for each set of variables (i.e., testing rate, sequencing proportion, and strategy).

Results

Performance of current guidance

We first applied current guidance from different stakeholder and academic groups on the number of positive specimens to sequence to detect SARS-CoV-2 variants at low prevalence (Table 1) for simulated wild-type/Alpha and Delta/Omicron epidemics in Zambia with a mean testing rate of 27 tests/100k/day (Figure 2). Even when assuming negligible turnaround time (i.e. time from specimen collection to acquisition of sequencing data), the current recommended approaches were insufficient to detect the variant on their respective target detection day when testing rates were low, regardless of the genomic surveillance sampling strategy. The first strategy of sampling specimens collected from the whole population that were sent to one sequencing facility (i.e. population-wide strategy) led to the best performance (closest to target detection day) for all recommendations, as it involves random uniform sampling of all available samples, a fundamental assumption made by all current guidance. However, if the specimen pools available for sequencing are restricted to those collected from a subset of sentinel tertiary facilities only, the non-uniformity in sampling coverage results in spatiotemporal bias within the sequenced samples, and leads to delayed detection of VOCs, which gets progressively worse as the proportion of tertiary facilities performing sequencing decreases to one facility.

Table 1: Current guidance by various stakeholder and academic groups on the number of specimens to sequence for detection of novel variants at low prevalence.

	Recommendation on number/proportion of positive cases to sequence		**Critical considerations**
World Health Organization (WHO)/European Centre for Disease Prevention and Control (ECDC)[7,8]	No. of positive cases	No. of sequences to detect at 1% with 95% confidence	• Agnostic to variant properties • Assumes specimen pool to be sampled for sequencing is representative of circulating diversity but acknowledges that unless testing coverage is evenly distributed this will be a biased sample • Notes that in countries with limited sequencing capacity monitoring relative prevalence of variants should be prioritized
	<1000 cases	141	
	1001 – 2,500	196	
	2,500 – 5,000	243	
	5,001 – 10,000	270	
	>10,000	285	
Brito et al.[5]	At least 0·5% of all cases with a turnaround time of 21 days to detect novel lineage before it reaches 100 cases at 20% probability		• Based on sequencing data from Denmark which is testing at >2,000 tests per 100,000 persons per day (https://www.finddx.org/covid-19/test-tracker/)
Wohl et al.[6]	1-29 sequences per day to detect an Alpha-like variant based on 0·03% initial introduction for a population of 10,000 (assuming growth rate of 0·1/day) at 1% with 95% confidence. We used the spreadsheet (https://github.com/HopkinsIDD/VOCsamplesize) provided and input appropriate parameters to obtain the recommendation relevant to the simulated epidemics.		• Assumes that the *observed* variant proportion in the positive specimens collected is representative of the *circulating* variant proportions among the infected population • Assumes asymptomatic patients are tested as well which may not be applicable for many LMICs where self-testing and asymptomatic community testing programs are currently rare. Widespread asymptomatic testing has also been substantially reduced in most HICs in the post-crisis phase of the pandemic. • Requires a large number of specimens that are randomly collected for assumption to hold true at low circulating variant proportions.

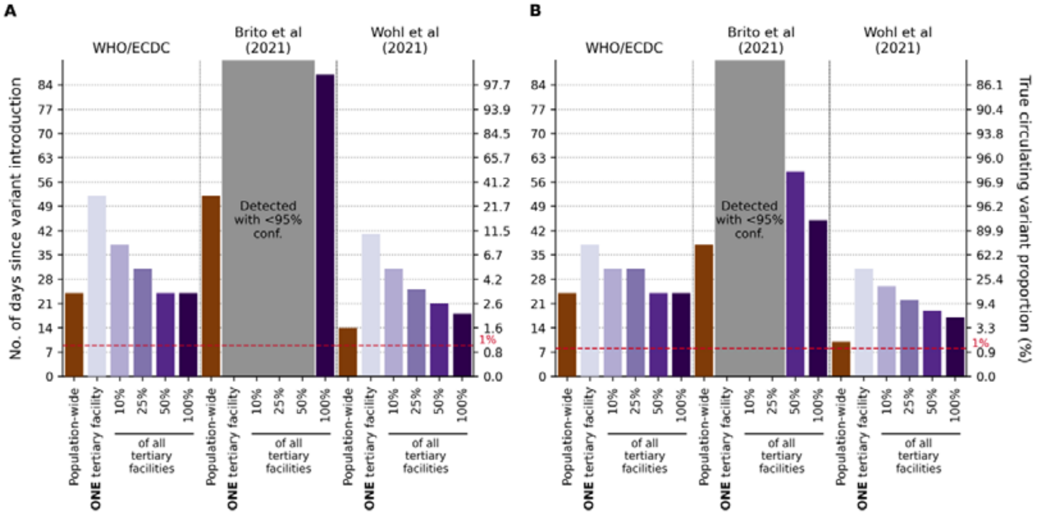

Figure 2: **Performance of current guidance on number of positive specimens to sequence for variant detection with testing rate at 27 tests per 100,000 persons per day**. First day of detection since variant introduction at 95% confidence and the corresponding circulating variant proportion using guidance from the World Health Organization (WHO)/European Centre for Disease Prevention and Control (ECDC)[7,8], Brito et al.[5], and Wohl et al.[6] (Table 1) under different genomic surveillance strategies with varying sampling coverage (i.e. all collected specimens from all healthcare facilities are sent to one facility to be sampled for sequencing (*population-wide* strategy); only *one*, 10%, 25%, 50%, or 100% of tertiary sentinel facilities would sample the specimens they collected for sequencing). Turnaround time (i.e. time from specimen collection to acquisition of sequencing data) was assumed to be negligible. 1,000 random bootstrap simulations were performed for each guidance/surveillance strategy. We simulated epidemics for (**A**) Wild-type SARS-CoV-2/Alpha. (**B**) Delta/Omicron. Grey regions denote that we could not reliably detect the variant virus with 95% confidence using the guidance in question under the assumed genomic surveillance strategy.

Variant detection

To elucidate how SARS-CoV-2 testing rates and the proportion of positive specimens sequenced impact the speed of variant detection, we simulated wild-type SARS-CoV-2/Alpha and Delta/Omicron epidemics at different Ag-RDT availability ranging from 27 tests/100k/day to 1,000 tests/100k/day (Figure 3). We assumed that specimens to be sequenced are sampled on their collection day, and varied the proportion of positive specimens to sample for sequencing each day between 1% and 100%. We analyzed the impact of testing rates and sequencing proportions on the expected day when the first specimen sampled for sequencing containing the variant was collected as a measure of variant detection speed. In Figure 3, we did not consider the time between sample collection and sequencing nor the turnaround time to obtaining sequencing results as they would only delay the actual day of variant detection by the assumed turnaround time.

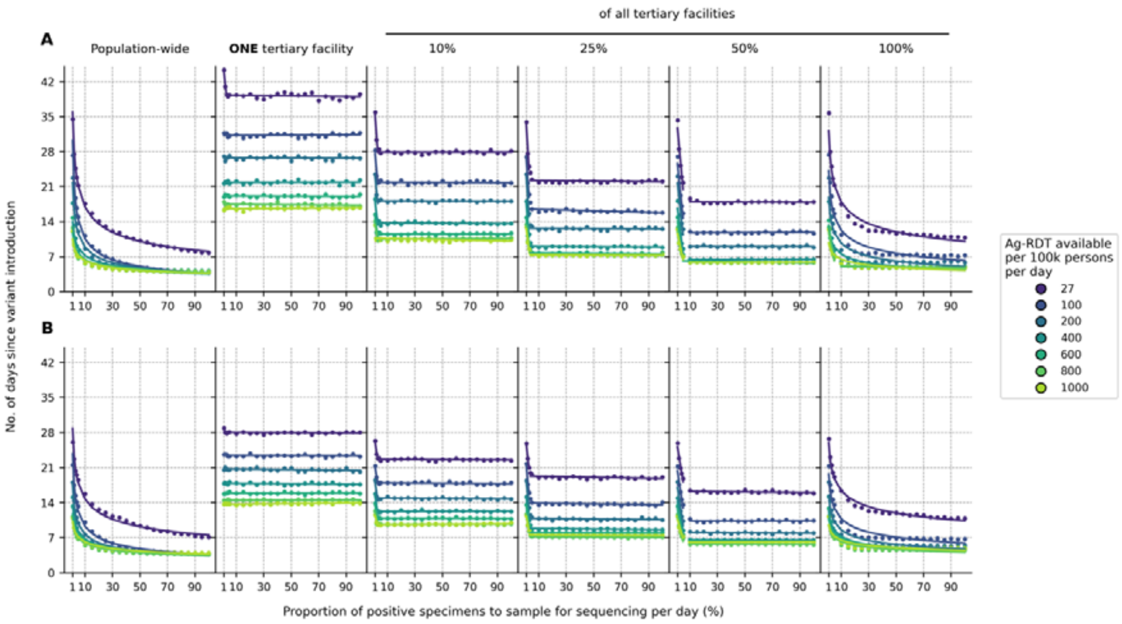

Figure 3: **Impact of SARS-CoV-2 testing rates and proportion of positive specimens to sequence on variant detection**. For each hypothetical daily test availability, the expected day when the first variant specimen is sampled for sequencing since its introduction is plotted against the proportion of positive specimens to be sampled for sequencing daily. Different genomic surveillance strategies with varying sampling coverage (i.e. all specimens collected from all healthcare facilities sent to one facility to be sampled for sequencing (*population-wide* strategy); only *one*, 10%, 25%, 50%, or 100% of tertiary sentinel facilities would sample the specimens they collected for sequencing) were simulated. (**A**) Wild-type SARS-CoV-2/Alpha. (**B**) Delta/Omicron.

For all testing rates, the relationship between the expected day when the first sample containing the variant was collected and the proportion of positive specimens sequenced per day can be described by a convex operating curve, reflecting rapidly diminishing returns in the speed of variant detection as more specimens are sampled for sequencing. Across all genomic surveillance sampling strategies, relatively larger marginal improvements to the speed of variant detection are generally made when the sequencing proportion is increased from 5 to 20% of all samples collected. Further sequencing only minimally shortens the expected time to variant detection, as the operating curve asymptotically approaches the earliest possible day of detection.

Importantly, increasing SARS-CoV-2 testing allows smaller sequencing proportions to attain similar detection day targets, and higher testing rates lower the earliest possible detection day. For both the Alpha and Omicron variants, increasing testing rates from 27 tests/100k/day to 100 tests/100k/day brings forward the expected day of sampling the first variant sequence by at least one week (Figure 3).

For the same level of testing and sequencing proportion, the population-wide strategy led to the earliest initial detection of a variant sequence. If sequencing were restricted to samples collected at a subset of tertiary sentinel facilities only, increasing the number of facilities sending samples for sequencing reduced the spatiotemporal bias in the specimen pool, thereby shaping the operating curves closer to the ones observed for the population-wide strategy. Interestingly, results similar to the population-wide strategy could be attained if all tertiary facilities acted as sentinel sites and sent the samples they collected for sequencing to increase the representativeness of sampling.

Observed variant proportion

Test availability and sampling coverage also affect the accuracy of the observed variant proportion (Figures 4 and S1). At a testing rate of 27 tests/100k/day, the observed variant proportion maximally differs from the true circulating proportion by >30% for both the Alpha and Omicron variants and for more than 15% of the time, the proportional difference between the observed and true variation was greater than 20%. Both the maximum absolute difference and percentage of timepoints where the difference is >20% can be lowered to <20% and <5% respectively if testing rate is increased to 100 or more tests/100k/day.

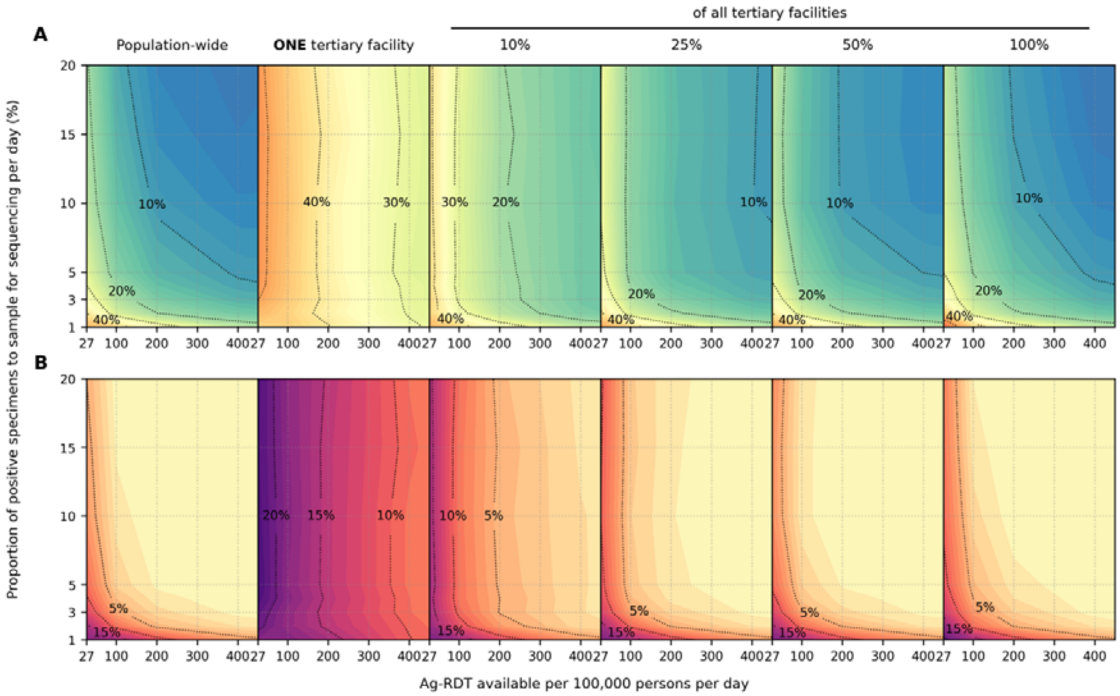

Figure 4: **Impact of SARS-CoV-2 testing rates on the capacity to monitor changes in variant prevalence based on diagnostic test availability and proportion of test-positive samples sequenced**. Different genomic surveillance strategies (i.e. all specimens collected from all healthcare facilities sent to one facility to be sampled for sequencing (*population-wide* strategy); only *one*, 10%, 25%, 50%, or 100% of tertiary sentinel facilities would sample the specimens they collected for sequencing) were simulated. (**A**) Maximum absolute difference between observed and circulating variant proportions. (**B**) Proportion of timepoints when sequencing was performed that the absolute difference between observed and circulating variant proportions is greater than 20%.

Critically, when the representativeness of the specimen pool is spatiotemporally biased by sequencing samples collected at tertiary sentinel facilities only, increasing the proportion of specimens to be sequenced only marginally lowers the maximum absolute difference or lessens the number of times where observed variant proportion deviates less than 20% from true circulating proportions (Figure 4, near vertical isoclines at low daily rates of testing). Increasing testing rates at sentinel surveillance sites provides more accurate detection in changes to circulating prevalence than sequencing more samples in the context of low testing rates.

Sensitivity analyses

We repeated our analyses using virus properties (i.e., incubation period, maximum viral load, protection against infection by the mutant virus after extant virus infection) of the Omicron variant but varied different relative transmissibility to the Delta variant (1·0 to 4·0) as well as the initial proportion of individuals who had been infected by the Delta variant (10% and 40%). The variant growth rates simulated for these hypothetical Delta/Omicron epidemics ranged from 0·17/day to 0·42/day.

Under these varied conditions, the expected day when the specimen of the first variant sequence is collected still follows a convex-shaped operating curve against the daily proportion of positive specimens to sequence. For all curves, the larger marginal improvements in shortening variant detection are still in sequencing proportions of up to 5-20% (Figure S2). In terms of the accuracy of observed variant to true circulating proportions, the maximum absolute difference and percentage of timepoints where difference is >20% are both substantially lowered if testing rate is increased to at least 100 tests/100k/day (Figures S3-4).

We also varied the prevalence of extant Delta infections when the Omicron variant was introduced (Figure 5). We found that lower test availability causes a delay in sampling the first variant specimen if the variant is introduced when pre-existing extant variant circulation is high. At 27 tests/100k/day, regardless of specimen proportions sequenced, detection could be delayed by ~1 week if Omicron was introduced when Delta was circulating at 10% prevalence as opposed to 1%. This is because a greater share of tests would be used to diagnose the more prevalent extant virus infections which in turn decreases the likelihood of detecting the newly introduced variant at low proportions.

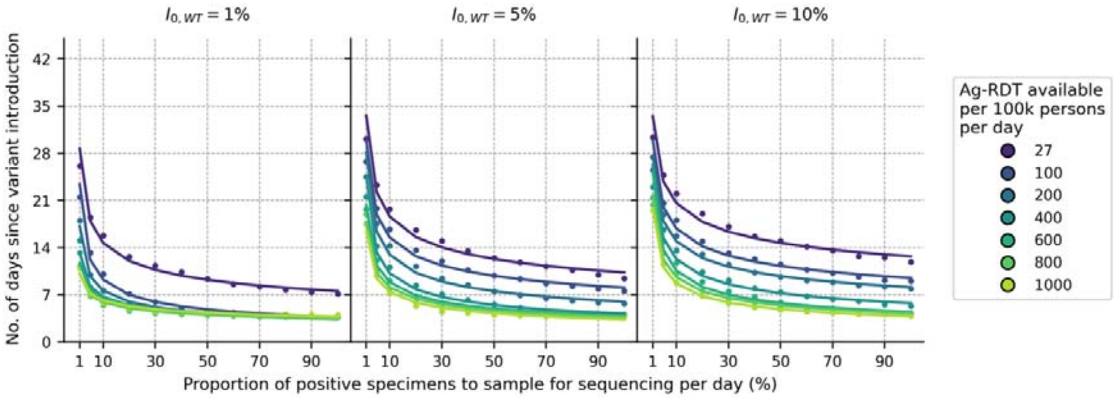

Figure 5: **Impact of prevalence of extant variant of concern (   ) at the time of new variant introduction**. For each Ag-RDT availability, the expected day when the first Omicron variant specimen (in the background of Delta) is sampled for sequencing since its introduction is plotted against the proportion of positive specimens to be sampled for sequencing daily. Each panel shows a different prevalence of the Delta variant (   ) at the point of Omicron introduction. Sampling for sequencing was drawn from the *population-wide* scenario.

Discussion

Our findings show that the emphasis on the size of the sample referred for genomic surveillance is misplaced if testing capacity is insufficient and sample sources are highly spatiotemporally biased. As such, at the current mean rate of testing in LMICs (27 tests/100k/day), current guidance[5–8] on sequencing sample size estimation could likely lead to later-than-predicted detection of novel variants at best or, at worst, leave new variants undetected until they have infected a majority of a population.

Based on our work, we identified three major areas of improvement that should be prioritized to enhance the robustness of genomic surveillance programs (Figure 6). First, the most substantial improvements are likely to come from increasing the mean testing rate in LMICs from 27 tests/100k/day (Figure 6A) to at least 100 tests/100k/day (Figure 6B). Even if one were to only doing sentinel surveillance at one tertiary facility, this increase in testing rate for the catchment area of the facility would speed up variant detection by 1-2 weeks.

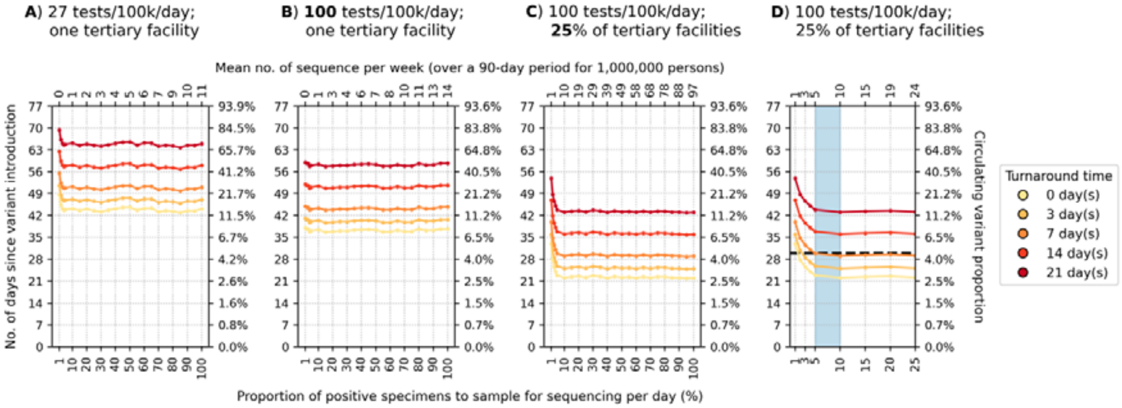

Figure 6: **Recommended approach to enhance genomic surveillance robustness**. In each plot, the operating curves of the expected detection day of the Alpha variant with wild-type SARS-CoV-2 in the background circulating at 1% are plotted for different proportion of specimens to sample for sequencing per day and turnaround times. The vertical axes denote the number of days passed since the introduction of the Alpha variant (left) and its corresponding circulating proportion (right). The horizontal axes denote the proportion of positive specimens to sample for sequencing per day (bottom) and the corresponding mean number of sequences to be generated per week per 1,000,000 people over a 90-day epidemic period. (**A**) Specimen pools for sequencing from *one* tertiary sentinel facility with testing rate at 27 tests per 100,000 persons per day (tests/100k/day). (**B**) Specimen pools for sequencing from *one* tertiary sentinel facility with testing rate at 100 tests/100k/day. (**C**) Specimen pools for sequencing from 25% of all tertiary sentinel facilities with testing rate at 100 tests/100k/day. (**D**) Zoomed-in plot of (C) to highlight sequencing proportions varying between 1-25%. Sequencing 5-10% of positive specimens (blue shaded region) would ensure that we would expectedly detect Alpha in 30 days if turnaround time is kept within one week.

Second, the representativeness of a specimen pool for sequencing can be further improved by expanding sampling coverage. In our model, variant detection was further sped up by 1-3 weeks by increasing the percentage of tertiary sentinel facilities sending the samples they had collected for sequencing to 25% of facilities (Figure 6C). Additionally, in terms of prevalence monitoring, if 25% of tertiary facilities sequenced 5% of all positive specimens they had collected to detect and monitor an Alpha-like variant, the maximum absolute difference to true circulating proportion is expected to decrease from >50% (assuming a single sentinel facility) to no more than 20%.

Third, reducing turnaround time of sample referred for sequencing results in a 1:1 decrease in time to new variant detection regardless of the proportion of sequenced samples, test availability or sampling coverage (Figure 6). These gains require scale up in sample transport networks, access to sequencing machinery, trained personnel, and/or increases in numbers of sequenced samples to make the most efficient use of each sequencing run.

After reducing spatiotemporal bias in the specimen pool through increased testing and sampling coverage, sequencing up to 5-10% of the positive specimens collected would return the greatest information gains while minimizing resource wastage. For an Alpha-like variant, at 100 tests/100k/day with sampling from 25% of tertiary sentinel facilities for sequencing, this amounts to an estimated 5-10 sequences per week averaged over a 90-day period per 1,000,000 people. If turnaround time is kept within one week, the variant would likely be detected within one month at ~4% circulating proportion (Figure 6D). Similarly, at the same testing rate, sampling coverage and turnaround time (i.e. average 5-11 sequences per week per 1,000,000 people), an Omicron-like variant would be detected before the first month since its introduction but at ~23% circulating proportion owing to its faster transmission (Figure S5).

Our findings here serve to inform expectations as genomic surveillance programs are being developed and should be interpreted according to the public health objectives of each program. If the objective is to serve as an early warning system for the emergence of new variants of concern before they are likely to have spread widely, then all factors above are essential and will likely require substantially more than 100 tests/100k/day. Critically, determining that a new variant is a threat requires not only detection of the variant itself but also the capacity to reliably monitor changes in its prevalence and potential clinical impact on short timescales. The results presented here also inform the design of programs for the sensitive and reliable detection of changes in variant prevalence.

The emergence and detection of each VOC to date represents interesting case studies for the work described here (Supplementary Appendix). For example, at the time of first detection of the Omicron variant, in South Africa in November 2021, the daily SARS-CoV-2 testing rate was 51 tests/100,000 people/day (https://www.finddx.org/covid-19/test-tracker/), which was among the highest testing rates in Africa. The Omicron variant was however detected 6-8 weeks after its likely emergence.[4] While this is commendable, Omicron had already infected a substantial portion of the population in Gauteng, South Africa (i.e., the estimated circulating variant proportion was >80% by mid-November).[4] Not only had the variant already spread across the rest of South Africa and neighboring Botswana,[4] Omicron samples were also collected in multiple other countries, including Hong Kong,[23] Denmark,[24] and the Netherlands[25] before the initial reports of the existence of the Omicron variant. This situation is consistent with the modelling findings, where novel variant detection is possible with <100 tests/100k/day but only after the new variant has spread widely across the population,[5] abrogating any possibility of containment.

Expanding genomic sequencing capabilities, especially in LMICs, is a global priority[26] and current investments in sequencing must continue.[27,28] Simultaneously, sustained investments in public health systems are required to expand access to, and availability of, diagnostic testing to underpin SARS-CoV-2 surveillance programs. Here, we primarily focused on LMICs but our findings on the impact of testing rates and representativeness on genomic surveillance programs are equally important for HICs as they consider dismantling parts of their testing and surveillance infrastructure in the post-crisis phase of the pandemic. While we find that routine representative sampling is vital for monitoring SARS-CoV-2 evolution, additional surveillance systems, including targeted surveillance of particular populations and settings (such as immunocompromised individuals or unusual events) could enable increased sensitivity.[29] Ultimately, detecting the next SARS-CoV-2 variant or pathogen that causes the next pandemic requires fundamental clinical diagnostic capacity to detect infections in the first place.

Data sharing

All data relevant to the study are included in the Article, the Supplementary Appendix and the github repository (https://github.com/AMC-LAEB/PATAT-sim). The PATAT model source code is also available at https://github.com/AMC-LAEB/PATAT-sim.

Declaration of interests

A.T., E.H., S.C., B.R. and B.E.N. declare that they are employed by FIND, the global alliance for diagnostics.

Acknowledgements

A.X.H. and C.A.R. were supported by ERC NaviFlu (No. 818353). C.A.R. was also supported by NIH R01 (5R01AI132362-04) and an NWO Vici Award (09150182010027). The authors are pleased to acknowledge that all computational work reported in this paper was performed on the Shared Computing Cluster which is administered by Boston University’s Research Computing Services (www.bu.edu/tech/support/research/).

Authors' contributions

A.X.H. contributed to the conceptualization, data curation, formal analysis, investigation, methodology, software, validation and visualization of the study. B.E.N. and C.A.R. contributed to the conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, validation and supervision of the study. J.A.S., M.P., S.B. and M.V.K. contributed to the conceptualization and interpretation of the study. A.T., E.H., S.C., and B.R. contributed to the conceptualization, funding acquisition, project administration, and resources of the study. E.P. contributed to the validation and visualization of the study. A.X.H. and C.A.R. wrote the original draft of the manuscript. All authors are involved in the review and editing of the manuscript. All authors had full access to all data of the study and the final responsibility for the decision to submit for publication.

References

1 Robishaw JD, Alter SM, Solano JJ, *et al*. Genomic surveillance to combat COVID-19: challenges and opportunities. *The Lancet Microbe* 2021; **2**: e481–4.

2 Davies NG, Abbott S, Barnard RC, *et al*. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science (1979)* 2021; **372**: eabg3055.

3 Cherian S, Potdar V, Jadhav S, *et al*. SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganisms 2021, Vol 9, Page 1542* 2021; **9**: 1542.

4 Viana R, Moyo S, Amoako DG, *et al*. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature 2022* 2022; : 1–10.

5 Brito AF, Semenova E, Dudas G, *et al*. Global disparities in SARS-CoV-2 genomic surveillance. *medRxiv* 2021; : 2021.08.21.21262393.

6 Wohl S, Lee EC, DiPrete BL, Lessler J. Sample Size Calculations for Variant Surveillance in the Presence of Biological and Systematic Biases. *medRxiv* 2022; : 2021.12.30.21268453.

7 Sequencing of SARS-CoV-2 - first update. https://www.ecdc.europa.eu/en/publications-data/sequencing-sars-cov-2 (accessed April 27, 2022).

8 Guidance for surveillance of SARS-CoV-2 variants: interim guidance, 9 August 2021. https://apps.who.int/iris/handle/10665/343775 (accessed Feb 25, 2022).

9 Adepoju P. Closing Africa's wide COVID-19 testing and vaccination gaps. *The Lancet Microbe* 2021; **2**: e573.

10 Brümmer LE, Katzenschlager S, Gaeddert M, *et al.* Accuracy of novel antigen rapid diagnostics for SARS-CoV-2: A living systematic review and meta-analysis. *PLOS Medicine* 2021; **18**: e1003735-.

11 Linton NM, Kobayashi T, Yang Y, *et al.* Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. *Journal of Clinical Medicine 2020, Vol 9, Page 538* 2020; **9**: 538.

12 Kissler SM, Fauver JR, Mack C, *et al.* Viral dynamics of acute SARS-CoV-2 infection and applications to diagnostic and public health strategies. *PLOS Biology* 2021; **19**: e3001333-.

13 Hay JA, Kissler SM, Fauver JR, *et al.* Viral dynamics and duration of PCR positivity of the SARS-CoV-2 Omicron variant. *medRxiv* 2022; : 2022.01.13.22269257.

14 Report 49 - Growth, population distribution and immune escape of Omicron in England | Faculty of Medicine | Imperial College London. https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-49-Omicron/ (accessed Feb 25, 2022).

15 Pouwels KB, Pritchard E, Matthews PC, *et al.* Impact of Delta on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. *medRxiv* 2021; : 2021.08.18.21262237.

16 Mathieu E, Ritchie H, Ortiz-Ospina E, *et al.* A global database of COVID-19 vaccinations. *Nature Human Behaviour 2021 5:7* 2021; **5**: 947–53.

17 Dovel K, Balakasi K, Gupta S, *et al.* Frequency of visits to health facilities and HIV services offered to men, Malawi. *Bull World Health Organ* 2021; **99**: 618–26.

18 Hasell J, Mathieu E, Beltekian D, *et al.* A cross-country database of COVID-19 testing. *Scientific Data 2020 7:1* 2020; **7**: 1–7.

19 Mwananyanda L, Gill CJ, Macleod W, *et al.* Covid-19 deaths in Africa: prospective systematic postmortem surveillance study. *BMJ* 2021; **372**. DOI:10.1136/BMJ.N334.

20 Gill CJ, Mwananyanda L, MacLeod W, *et al.* Sustained high prevalence of COVID-19 deaths from a systematic post-mortem study in Lusaka, Zambia: one year later. *medRxiv* 2022; : 2022.03.08.22272087.

21 Nichols BE, Girdwood SJ, Crompton T, *et al.* Monitoring viral load for the last mile: what will it cost? *J Int AIDS Soc* 2019; **22**: e25337.

22 Girdwood SJ, Nichols BE, Moyo C, Crompton T, Chimhamhiwa D, Rosen S. Optimizing viral load testing access for the last mile: Geospatial cost model for point of care instrument placement. *PLOS ONE* 2019; **14**: e0221586.

23 Gu H, Krishnan P, Ng DYM, *et al.* Probable Transmission of SARS-CoV-2 Omicron Variant in Quarantine Hotel, Hong Kong, China, November 2021. *Emerging Infectious Diseases* 2022; **28**: 460.

24 Espenhain L, Funk T, Overvad M, *et al.* Epidemiological characterisation of the first 785 SARS-CoV-2 Omicron variant cases in Denmark, December 2021. *Eurosurveillance* 2021; **26**: 2101146.

25 Omicron variant found in two previous test samples | RIVM. https://www.rivm.nl/en/news/omicron-variant-found-in-two-previous-test-samples (accessed March 17, 2022).

26 Adepoju P. Challenges of SARS-CoV-2 genomic surveillance in Africa. *The Lancet Microbe* 2021; **2**: e139.

27 Tegally H, San JE, Cotten M, *et al.* The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance. *medRxiv* 2022; published online April 22. DOI:https://doi.org/10.1101/2022.04.17.22273906.

28 Leite JA, Vicari A, Perez E, *et al.* Implementation of a COVID-19 Genomic Surveillance Regional Network for Latin America and Caribbean region. *PLOS ONE* 2022; **17**: e0252526.

29 Knyazev S, Chhugani K, Sarwal V, *et al.* Unlocking capacities of genomics for the COVID-19 response and future pandemics. *Nature Methods 2022 19:4* 2022; **19**: 374–80.