Equity and efficiency in global respiratory virus genomic surveillance

Simon P.J. de Jong[1], Brooke E. Nichols[1,2,3], Menno D. de Jong[1], Alvin X. Han[1,*], Colin A. Russell[1,3,*]

[1]Department of Medical Microbiology & Infection Prevention, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands

[2]Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland

[3]Department of Global Health, Boston University, Boston, MA, USA

*Contributed equally

Correspondence to C.A.R. – c.a.russell@amsterdamumc.nl

Summary

Public health interventions for respiratory virus outbreaks increasingly rely on genomic sequencing for the rapid identification of new (variant) viruses[1–5]. However, global sequencing efforts are unevenly distributed[6–9], with some high-income countries sequencing at >100,000 times the rate of many low-income countries. Given the importance of virus genomic sequencing and substantial global disparities in sequencing capacities, there is a need for meaningful minimum sequencing targets and functional upper bounds that maximise resource efficiency[1,2,8,10,11]. Here, using mathematical models and analyses of data on global SARS-CoV-2 sequencing output in 2022, we show that increases in sequencing rates typical of low-income countries are >100-fold more effective at reducing time to detection of new variants than increases from rates typical of high-income countries. We find that relative to 2022 sequencing rates, establishing a minimum respiratory virus sequencing capacity of two sequences per million people per week (S/M/wk) with a two-week time from sample collection to sequence deposition in all countries, while simultaneously capping sequencing rates at 30 S/M/wk in all countries, could reduce mean time to first variant detection globally by weeks-to-months while also reducing global sequencing output by >60%. Our results show that investing in a minimum global respiratory virus sequencing capacity is far more effective at improving variant surveillance than expanding local sequencing efforts in countries with existing high-intensity respiratory virus surveillance programs and can guide rightsizing of global respiratory virus genomic surveillance infrastructure.

Main

Genomic surveillance of respiratory viruses has expanded substantially since the late 1990s and is now a critical component of public health preparedness and response, particularly for identifying and monitoring the spread of new virus variants of concern[1–5]. However, genomic surveillance infrastructure is unequally distributed globally[2,6,8,9]. For viruses collected in 2022 alone, ~7 million SARS-CoV-2 genomes have been submitted to GISAID (www.gisaid.org, the most commonly used repository for respiratory virus genome sequencing data), but country-level sequencing rates as estimated from GISAID submissions varied by over six orders of magnitude (Fig. 1a), with 17 countries not depositing any sequences. Half of all publicly shared SARS-CoV-2 genomes from samples collected in 2022 originated from countries that account for only 4.4% of the global human population, while countries comprising half of the global population deposited only 0.7% of available genomes (Fig. 1b). Additionally, the time from sample collection to sequence deposition (henceforth, turnaround time) ranged across countries from less than two weeks to hundreds of days (interquartile range 28-108 days; Fig. 1c). Sequencing rates and median turnaround times (Spearman's ρ = 0.79, $P = 6.3\times10^{-41}$; ρ = -0.54, $P = 7.1\times10^{-16}$, respectively) are strongly correlated with *per capita* GDP, indicating that the capacity of a country's genomic surveillance infrastructure correlates with its economic output (Fig. 1d, Extended Data Fig. 1). As new variants can potentially emerge in any country, this global variability in genomic surveillance capacity raises important questions about the amount of sequencing and associated turnaround time needed to effectively and efficiently detect new virus variants worldwide[1,2,8,10,12,13].

To address these questions, we deterministically simulated the emergence of a variant virus in the background of circulating wildtype virus with susceptible-infected-recovered (SIR) dynamics under different scenarios of variant emergence (i.e. initial R_e of wildtype and variant viruses and prevalence of wildtype virus; Extended Data Fig. 2). Using the simulations, we computed the expected day of variant detection with 95% confidence based on binomial sampling for different sequencing rates. We then derived a new mathematical model characterising the relationship between sequencing rates and time to detection of the new virus variant. For a variant virus, introduced in a population at an initial frequency f_0, of which the change in variant proportion through time can be described by a logistic growth rate s, the time since variant introduction after which the variant virus is expected to have

been detected with confidence level 1-q, when sequencing n samples per unit time, is equal to $(\log[(q^{-s/n}-1)/f_0]+1)/s$. This model is applicable to all respiratory viruses that can be described by SIR dynamics[14], including SARS-CoV-2, respiratory syncytial virus, and pandemic or seasonal influenza viruses.

For all modelled scenarios of variant emergence (Extended Data Fig. 2), time to variant detection rapidly decreased as sequencing rate increased up to ~10 S/M/wk (Fig. 2a, Extended Data Fig. 3a). In comparison, the benefits of further increases in sequencing rate beyond 10 S/M/wk were much smaller (Fig 2a). In 2022, many high-income countries sequenced SARS-CoV-2 genomes at rates well in excess of 10 S/M/wk (some >>100 S/M/wk), whereas many lower-and-middle-income countries sequenced at rates (<<1 S/M/k in many countries) at which, in absolute terms, small increases in sequencing rates would substantially speed up variant detection (Fig. 1a, 2b, Extended Data Fig. 3b). For example, in a country of 100 million people sequencing at the median 2022 SARS-CoV-2 sequencing rate in low-income countries (0.035 S/M/wk), increasing the sequencing rate by 1 S/M/wk would reduce the time to detection of a variant with R_e = 1.6 at 95% confidence by ~28 days, given a wildtype prevalence of 0.5% and a wildtype R_e of 1.1 at time of variant emergence. In contrast, if the same country was sequencing at the 2022 median high-income country rate (58.7 S/M/wk), the reduction in time to detection resulting from the same 1 S/M/wk increase in sequencing rate would be only 3.5 hours (Fig. 2b).

Sequencing rates similarly impact the number of people that will have been infected by the variant when it is first detected (Fig. 2c, Extended Data Fig. 4a). In the same scenario of variant emergence described above (variant R_e = 1.6, wildtype R_e = 1.1, wildtype prevalence = 0.5% at variant emergence), given a sequencing turnaround time of two weeks, the expected number of variant infections by the day of first detection with 95% confidence amounted to ~4.7 million in a country of 100 million people sequencing at the median low-income country rate. Increasing the sequencing rate in this country by 1 S/M/wk would reduce the expected number of variant infections by the time of detection by ~4.5 million infections (Fig. 2d). In contrast, only ~3,400 variant infections would be expected by the day of first detection in a country sequencing at the median 2022 sequencing rate in high-income countries (Fig. 2c), and increasing the sequencing rate in this country by the same 1 S/M/wk would only reduce the expected number of variant infections by the day of first detection by ~60 infections (Fig. 2d). Hence, for reducing a variant's extent of spread through a

population by the time of first detection, the benefits of increases in sequencing rates are far more substantial at lower sequencing rates (Fig. 2d, Extended Data Fig. 4b).

In addition to sequencing rate, turnaround time is an essential component of effective genomic surveillance[1,7,8,15]. For reducing time to variant detection, any reduction in turnaround time is functionally equivalent to a fold increase in sequencing rate, and the magnitude of this equivalent fold increase depends on the scenario of variant emergence (Fig. 2e). For example, if the wildtype virus is circulating with R_e = 1 at 0.1% prevalence, reducing turnaround time by three weeks is equivalent to increasing the sequencing rate 2.4-fold for detecting a variant with R_e = 1.2. In contrast, for detecting a variant with R_e = 2, the same three-week reduction in turnaround time is equivalent to a ~55-fold increase in sequencing rate. As reductions in turnaround time might be more cost-efficient than increases in sequencing rate, the benefits of increasing sequencing output should be carefully weighed against the gains from strengthening the ancillary infrastructure necessary for rapid sequencing.

For individual countries, the above results inform how resources can be efficiently allocated to detect new virus variants locally. However, new (variant) viruses can emerge anywhere globally[1,16]. The global time to variant detection is shaped by (1) the global human mobility network, which determines how the virus spreads internationally[17–22] and (2) the global genomic surveillance network, which determines how rapidly it can be detected in individual countries where it is present. To investigate how global variation in respiratory virus genomic surveillance infrastructure impacts the speed of new variant detection, we simulated global variant spread using a global metapopulation model, validated against GLEAM[23,24] (Extended Data Fig. 5). For each value of variant R_e, ranging from 1.2 to 2, we performed 10,000 independent simulations. In each simulation, the country where the variant emerged was randomly selected based on a country population size-weighted probability. We then estimated the expected global time to variant detection for each simulation given empirical country-specific SARS-CoV-2 sequencing rates and turnaround times in 2022 as estimated from submissions to GISAID[25].

The mean time to first variant detection globally, averaged across all simulated variant R_e values, was 82.1 days (95% CI 17 – 193) with substantial variability at lower values of variant R_e (Fig. 3a). The global number of variant infections by the day of first global

detection varied widely (mean 566,413 infections, 95% CI 73 – 4,999,369) and spanned up to five orders of magnitude for all values of variant R_e (Fig. 3b). The continent in which the variant emerged strongly shaped the time to variant detection (Fig. 3c) and the number of global variant infections by the day of first detection (Fig. 3d), the latter ranging from a mean of 20,264 infections (95% CI 26 – 235,022) when emerging in Europe to 1,559,748 infections (95% CI 950 – 12,213,845) in case of emergence in Africa (Fig. 3d). The differences in time to detection (Fig. 3e) and the number of variant infections by the day of detection (Fig. 3f) were strongly associated with the sequencing rate in the variant's country of emergence. In 29.1% of all simulations, new variants were first detected outside of their continent of origin, driven especially by variants emerging in Africa (detected outside origin continent in 74.4% of simulations), Asia (23.8%) and South America (20.0%), meaning that the variant would have frequently spread widely within and between continents prior to initial detection (Fig. 3g, 3h).

Since reductions in time to detection resulting from increases in sequencing rate beyond ~10 S/M/wk (Fig. 2b) are limited, we hypothesised that reducing sequencing output in countries that strongly exceeded this rate would have little effect on speed of variant detection while substantially reducing global sequencing output. We re-simulated the genomic surveillance process using the same metapopulation epidemic simulations and found that relative to the 2022 baseline (henceforth, strategy 1), the expected time to variant detection (Fig. 4a, Extended Data Fig. 6a) and the expected number of variant infections by the day of detection (Fig. 4b, Extended Data Fig. 6b) remained largely unchanged if sequencing rates were capped at 30 S/M/wk in all countries (henceforth, strategy 2): mean time to variant detection would increase by only 4.5 days, from mean 82.1 days to 86.6 days (95% CI 24-199) (Fig. 4a), while global sequencing output would be reduced by 67.0% (Fig. 4c).

Because the largest reductions in time to detection are attained at relatively low sequencing rates (Fig. 2b), we further hypothesised that establishing basic sequencing infrastructure globally, even at a limited sequencing rate but with a low turnaround time, could substantially reduce the global time to variant detection relative to the 2022 baseline. Ensuring a global minimum sequencing capacity of 2 S/M/wk with a turnaround time of 14 days, while maintaining sequencing output in countries that already satisfied this capacity in 2022 (henceforth, strategy 3), reduced mean time to global variant detection by 26.1 days to 56.0 days (95% CI 16 – 118) (Fig. 4a). The mean number of global variant infections by the day of

166 detection decreased from 566,413 infections (95% CI 73 – 4,999,369) to 26,415 infections
167 (95% CI 61 –196,092) (Fig. 4b). A sequencing rate of 2 S/M/wk corresponds to 0.18% of the
168 maximum country-specific SARS-CoV-2 sequencing rate in 2022 and its establishment
169 globally would increase global sequencing output by 6.0% relative to the 2022 baseline (Fig.
170 4c).
171
172 Combining the insights above, we hypothesised that reducing the inequity in the global
173 genomic surveillance could strongly improve its efficiency and effectiveness. In our
174 simulations, combining strategies 2 and 3 (i.e. capping individual countries' sequencing
175 output at 30 S/M/wk while also ensuring the minimum global capacity of 2 S/M/wk with
176 turnaround time of 14 days; henceforth, strategy 4) reduced mean time to detection and the
177 mean number of variant infections by the day of first detection to 57.7 days (95% CI 20 –
178 120) and 27,717 infections (95% CI 90 – 201,028) respectively (Fig. 4a, 4b, Extended Data
179 Fig. 6). The performance of strategy 4 is effectively identical to that of strategy 3, which
180 establishes the global minimum capacity but without capping individual countries' output,
181 but strategy 4 still reduces global sequencing output by 61.0% relative to the 2022 baseline
182 (Fig. 4c). While the initial costs of establishing the infrastructure necessary to achieve the
183 minimum respiratory virus sequencing capacity globally are likely to be high, our results
184 show that opportunities exist for redistribution of existing resources or investments in new
185 ones in order to achieve a more equal distribution of sequencing infrastructure across the
186 globe, yielding a global surveillance system that is more effective while more than halving
187 overall global sequencing output.
188
189 The establishment of global minimum respiratory virus sequencing capacity would also
190 increase the probability that a variant is first detected in the continent where it emerged from
191 70.9% (strategy 1) to 98.4% (strategy 4) (Fig. 4d). Consequently, in all countries, including
192 those that were to reduce their national sequencing output, the lead time between the variant's
193 first global detection and its first local case would increase (Fig. 4e). This would allow for
194 more time for potential local public health measures in preparation for variant outbreaks in all
195 countries. These benefits are particularly valuable for countries that are extensively connected
196 in the global mobility network, located largely in Europe, Asia and North America, as these
197 countries tend to experience especially early invasion and hence typically have a shorter lead
198 time *a priori* (Fig. 4e). Importantly, the relative performance of the different strategies for the
199 global distribution of genomic surveillance infrastructure is robust to biases in the estimates

of turnaround time resulting from delays in sequence deposition in GISAID[26,27] (Extended Data Fig. 7) and deviations from the assumed global mobility rates (Extended Data Fig. 8).

This study primarily focuses on the simple detection of a variant virus, but ascertaining the public health risk posed by a variant requires information such as its virulence and transmissibility[28], that at best can only partially be inferred from genomic sequencing[1,29–32]. Accruing such information and translating it to public health policy likely occurs on timescales beyond the reductions in time to variant detection that increases in sequencing rate beyond the order of 10-30 S/M/wk can yield, thus limiting the public health impact of further increases. For example, the SARS-CoV-2 Alpha variant was first detected in the UK in a sample collected on 20 September 2020, likely within days of its initial emergence[33]. At the time, the sequencing rate in the UK was ~100 S/M/wk. However, it was not until December 2020 that epidemiological evidence of the variant's transmission advantage relative to pre-existing viruses began to accumulate[5,33]. Hence, sequencing at rates much lower than ~100 S/M/wk would likely have had similar public health impact. Because of the importance of complementary clinical and epidemiological data, the value of investments in global genomic surveillance capacity can be enhanced through clinical and public health infrastructure development[27,34].

Our results are broadly applicable to respiratory viruses in both endemic and epidemic scenarios, including potential future pandemics similar to the 2009 influenza A/H1N1pdm09 and COVID-19 pandemics. The global heterogeneity in genomic surveillance of SARS-CoV-2 is also apparent for other respiratory viruses, including those with extensive global public health surveillance histories: seasonal influenza virus genomic sequencing output in the pre-pandemic era was similarly unequally distributed (Extended Data Fig. 9).

Our results are limited by the assumption of representative sampling in the genomic surveillance process, including the ready availability and access to diagnostic tools, which does not always hold in reality[35,36]. As the departure from this assumption is especially strong in resource-constrained settings[15,35], the reported reductions in time to variant detection resulting from the establishment of a global minimum sequencing capacity are likely underestimates. Our results primarily apply to variant detection (and monitoring variant prevalence, see Extended Data Fig. 10, Supplementary Text), and not to other analyses such as reconstructing geographical spread[19] or targeted outbreak investigations[37], for which the

shape of the relationship between sequencing rate, turnaround time, and performance is likely different[12]. However, initial detection is the necessary starting point for all analyses that make use of genomic data. The proposed global minimum respiratory virus sequencing capacity offers increased and faster information for public health actions targeting the identification and monitoring of new variants, as well as tracking viruses through space and time[38,39]. Additionally, the optimal sequencing rate depends on the characteristics of the pathogen and the required timeliness of sequencing data for public health action, but a minimum capacity of 2 S/M/wk at 14 days turnaround time will even allow for relatively rapid detection when a highly transmissibly variant emerges in a background of high wildtype incidence. The balance of sequencing rate and turnaround time in our proposed minimum capacity serves as a potential target, but the most resource-efficient balance of sequencing rate and turnaround time could differ among countries. See Supplementary Text for further discussion of other approaches and comparisons to other guidance.

Our analyses use empirical SARS-CoV-2 sequencing rates based on submissions to GISAID from 2022, when COVID-19 was still a public health emergency of international concern. Although sequencing outputs have since declined in many countries[1,2], the fundamental notion persists that relatively small increases in global sequencing output in the right places can profoundly improve global respiratory virus genomic surveillance in ways that even large increases in places with established surveillance infrastructure cannot. Establishing the necessary infrastructure for robust global genomic surveillance will require substantial investments in countries that often have other competing public health priorities[2,40]. Our results suggest that, because the establishment of such infrastructure benefits the world at large, filling this investment gap will provide a strong return on investment for well-resourced countries that already possess strong genomic surveillance infrastructure locally. For these countries, such investments likely represent a more efficient use of public health resources than investments in increasing local sequencing output and should be a public health priority in the post-pandemic period.

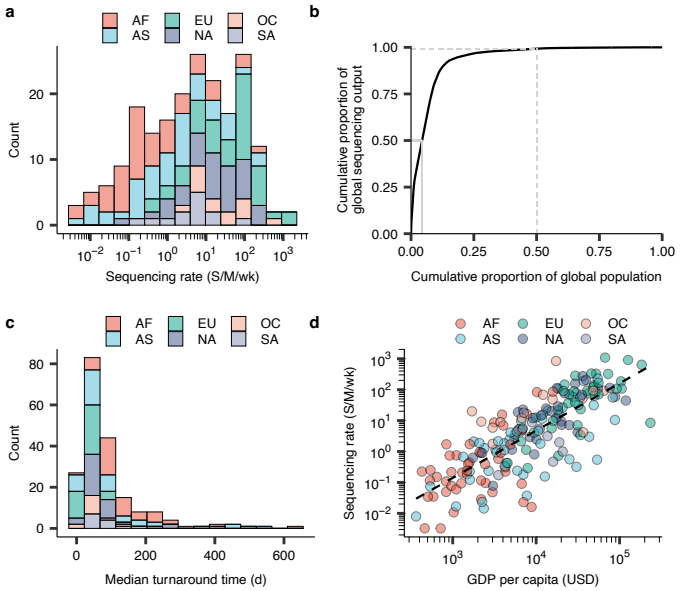


Fig. 1. The global landscape of SARS-CoV-2 genomic surveillance infrastructure in 2022. a) The distribution of non-zero weekly sequencing rates per million people, for individual countries (*n* = 198), coloured by continent (AF: Africa, EU: Europe, OC: Oceania, AS: Asia, NA: North America, SA: South America). **b)** The cumulative proportion of the global population that accounts for a cumulative proportion of global sequence output. Solid grey lines show the smallest proportion of the population that accounts for 50% of sequencing output. Dashed grey lines show the smallest proportion of sequencing output that is accounted for by 50% of the global population. **c)** The distribution of median country-specific turnaround times *(n* = 198), coloured by continent. **d)** Correlation between per capita GDP and weekly sequencing rate per million people by country (*n* = 188) (each circle represents one country, coloured by continent).

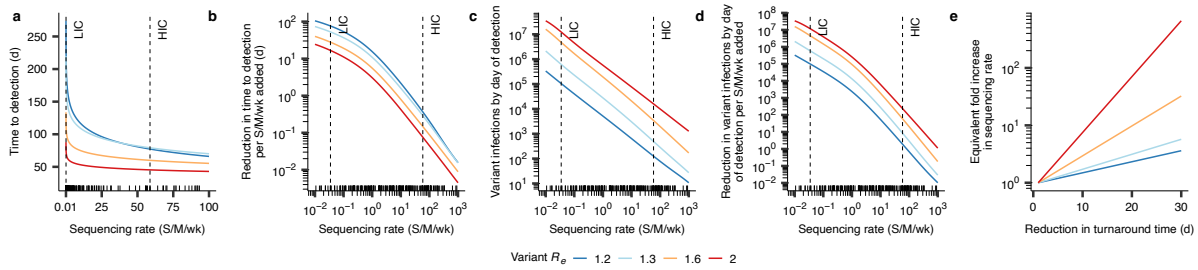


Fig. 2. The dependence of time to variant detection on sequencing rate and turnaround time for a single country. In all panels, lines are coloured by values of variant R_e, with a distinct scenario of variant emergence for each value of variant R_e; sequencing turnaround time was assumed to be 14 days. **a)** Relationship between sequencing rate and the number of days until the variant will have been detected with 95% confidence. The small black tick marks on the *x*-axes in this plot and in **b-d** show country-specific SARS-CoV-2 sequencing rates for 2022. Vertical dotted lines correspond to the median SARS-CoV-2 sequencing rates for high-income (HIC) and low-income (LIC) countries in 2022. **b)** Relationship between sequencing rate and the reduction in time to variant detection that results from increasing the existing sequencing rate (*x*-axis) by 1 S/M/wk. **c)** Relationship between sequencing rate and the number of variant infections by the day the variant will have been detected with 95% confidence. **d)** Relationship between sequencing rate and the reduction in the number of variant infections by the day of detection that results from increasing the existing sequencing rate (*x*-axis) by 1 S/M/wk. **e)** Relationship between a reduction in turnaround time (in days) and the fold increase in sequencing rate that would be required to effect the same reduction in time to detection if turnaround time was kept constant.

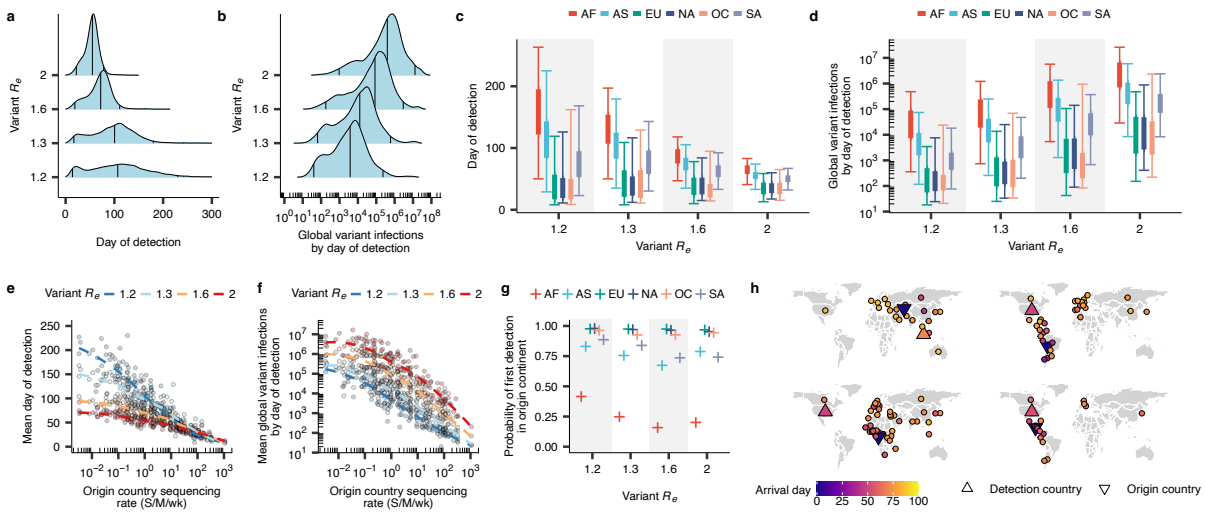


Fig. 3. The global time to variant detection based on the SARS-CoV-2 genomic sequencing landscape in 2022. a) The distribution of days to variant detection for different values of variant R_e, each with a distinct scenario of variant emergence *(n* = 10,000 for each variant R_e). Vertical lines correspond to the median and 95% CI. **b)** The distribution of the number of global variant infections by the day of variant detection. **c)** The time to variant detection by variant origin continent. Thin and thick lines correspond to 95% and 50% CIs, respectively. **d)** The number of global variant infections by the day of detection by variant origin continent, analogous to **c**. **e)** The relationship between a country's sequencing rate and the mean time to first global detection of a variant emerging in that country. **f)** The relationship between a country's sequencing rate and the mean number of global variant infections by the day of detection of a variant emerging in that country. **g)** The probability that the variant is first detected in its origin continent, by origin continent. **h)** Four example simulations of dynamics of variant spread and detection. Each point represents a country that has seen at least one variant infection by the day the variant is detected, coloured by the day of the first infection. Triangles and inverted triangles depict the country where the variant is first detected and first emerged, respectively. Simulations are for variant R_e of 1.6.

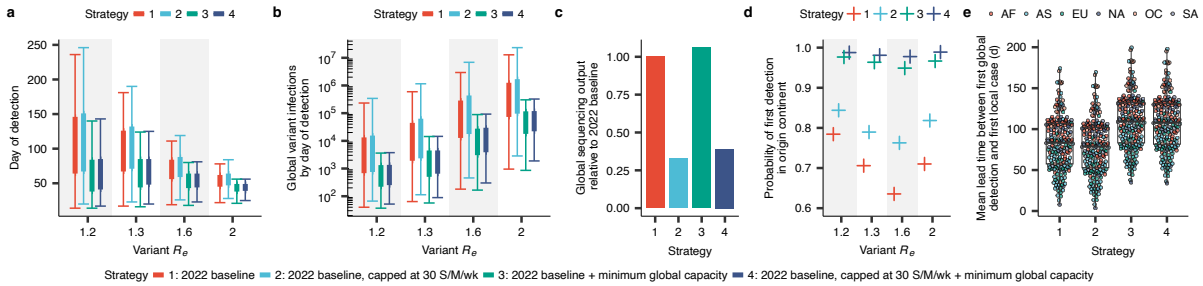


Fig. 4. The time to detection under varying global distributions of global respiratory virus genomic sequencing infrastructure. a) Comparison of time to variant detection for different global strategies for the global distribution of genomic surveillance infrastructure. Each value of variant R_e corresponds to a distinct scenario of variant emergence *(n* = 10,000 for each). Thin and thick lines correspond to 95% and 50% CIs, respectively. **b)** The cumulative number of global variant infections by the day of variant detection by strategy, analogous to **a**. **c)** Total global sequencing output relative to the 2022 baseline by strategy. **d)** The probability that the variant is first detected in its origin continent, by strategy. **e)** Comparison of the mean time between the first detection of the variant globally, and the first local within-country infection, by strategy, for individual countries, averaged across values of variant R_e. Each point corresponds to a country, coloured by continent. Boxplots show the median, first and third quartiles, and minimum and maximum values.

Online Methods

Sequence metadata analysis

We downloaded metadata corresponding to all SARS-CoV-2 genomes in the GISAID[13] database with collection date between January 1st 2022 and January 1st 2023 and submission date up to July 1st 2023 (n = 6,894,449). For each country with at least one sequence in the dataset, we computed the weekly sequencing rate by dividing the number of viruses sampled in that country by 52 and the country's population size in millions, yielding a sequencing rate in units of sequences per million people per week (S/M/wk). Population sizes for July 1st 2022 were extracted from the United Nations World Population Prospects 2022 (https://population.un.org/wpp/Download/Standard/MostUsed/). For each sequence, we computed the turnaround time from the number of days between the sample collection and submission day in GISAID. We extracted countries' *per capita* gross domestic product (GDP) for 2022, or the most recent year before 2022 if data for 2022 was unavailable, from the World Bank (https://data.worldbank.org/indicator/NY.GDP.PCAP.CD (last updated 2023/10/26)). We extracted income classifications for each country for fiscal year 2024 from the World Bank (https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups). For the analysis of seasonal influenza sequencing output, we downloaded metadata for all seasonal influenza haemagglutinin sequences from humans with sampling date between January 1st 2018 and January 1st 2019 (n = 28,992) as the last full year for which sequencing would have been minimally affected by the COVID-19 pandemic.

Surveillance simulations

In all analyses, we assumed that a variant virus emerges in the context of circulating wildtype virus. In our simulations, both variant and wildtype epidemiological dynamics are described by a susceptible-infected-recovered (SIR) compartmental model with infectious period $1/\gamma$ equal to 5 days for both viruses, with no interactions between genotypes. We simulated variant epidemics under a range of values of variant R_e at time of introduction (variant R_e = 1.2, 1.3, 1.6, and 2). In the main text, we assumed a different scenario of variant emergence for each value of variant R_e, characterized by a wildtype (wt) R_e at time of variant introduction and a wildtype prevalence at time of variant introduction (variant R_e = 1.2: wt R_e

= 1, wt prevalence = 0.1%; variant R_e = 1.3: wt R_e = 1.05, wt prevalence = 0.2%; variant R_e = 1.6: wt R_e = 1.1, wt prevalence = 0.5%; variant R_e = 2: wt R_e = 1, wt prevalence = 2%). These scenarios were chosen such that circulation dynamics of wildtype and variant were comparable (e.g. the emergence of a highly transmissible variant in the background of high wildtype prevalence). In the Extended Data, we show the same analyses for all combinations of variant R_e and scenario of variant emergence (e.g. a variant with R_e = 2 with wildtype dynamics corresponding to the scenario for variant R_e = 1.2 (wt R_e = 1, wt prevalence = 0.1%)). Epidemic dynamics for each scenario in the main text are shown in Extended Data Fig. 2. We note that, for any combination of initial variant proportion and variant proportion logistic growth rate, including those not explicitly discussed in this study, the mathematical model derived below can be used to compute the expected time to variant detection, for any sequencing rate.

For the single-country analyses presented in Fig. 2, we assumed a population of 100 million and turnaround time of two weeks. We deterministically simulated variant and wildtype epidemics, starting with one variant-infected individual, and computed the variant proportion *f(t)* through time. For each sequencing rate and given *f(t)*, we computed the expected day of detection with 95% confidence as the day on which the probability that zero wildtype sequences would have been binomially sampled up to and including that day declined below 0.05. On each day, the total number of samples to sequence was assumed to be a Poisson-valued random variable with rate given by the sequencing rate. For each sequencing rate, the day of detection was computed as the median across 100 replicates. To compute the equivalent fold increase in sequencing rate for each reduction in turnaround time, we computed the slope of a linear model that relates the logarithm of the sequencing rate to the simulated day of detection for $1 < n < 100$ S/M/wk.

Mathematical model

To derive a mathematical model for the relationship between sequencing rate and time to variant detection, we based our analyses on the premise that, starting from a single introduction, the proportion of all new infections of a particular virus type that is attributable to a variant virus with a transmission advantage at time *t* follows a logistic growth function $f(t) = \frac{1}{1+\frac{1-f_0}{f_0}e^{-st}}$. Here f_0 is the initial variant proportion relative to all circulating virus and *s*

is the variant proportion's logistic growth rate. Given that we are sampling to sequence n samples per unit time, the binomial probability that the variant is detected at or before time step τ is

$$P(t \leq \tau) = 1 - \prod_{t=0}^{\tau} \left[1 - \frac{1}{1+\frac{1-f_0}{f_0}e^{-st}}\right]^n \quad \text{[Eq. 1]}$$

If we define $q = 1 - P(t \leq \tau)$ (i.e. the probability that the variant will not be detected before or during time step τ), we can rewrite Eq. 1 as:

$$\tau = \frac{\ln[(q^{-s/n}-1)/f_0]+1}{s} \quad \text{[Eq. 2]}$$

Hence, for given s, f_0, and n, and q, Eq. 2. computes the day on which the variant will have been detected with confidence level 1-q. We validated Eq. 2 by comparing the simulated time to detection as shown in Fig. 2 to time to detection predicted using Eq. 2. Details are given in the Supplementary Information.

Metapopulation model

We used a metapopulation model that couples local SIR dynamics with global migration to simulate the global spread of a variant, given a single index country. Given a rate of movement w_{nm} from population m to n, the expected number of variant-infected (I_n) and variant-susceptible (S_n) people in population n with population size N_n, given transmission rate β and recovery rate γ, is described by

$$\partial_t I_n = \frac{\beta S_n I_n}{N_n} - \gamma I_n + \sum_{m \neq n} (w_{nm} I_m - w_{mn} I_n)$$

$$\partial_t S_n = \frac{-\beta S_n I_n}{N_n} + \sum_{m \neq n} (w_{nm} S_m - w_{mn} S_n)$$

This model is the basis of the model used by Brockmann et al.[20] to fit empirical arrival times for multiple respiratory viruses to global air transportation data. We used the estimated pairwise number of trips between all countries from the Global Transnational Mobility (GTM)[41] to inform w_{nm}. Specifically, for any two countries n and m we computed w_{nm} by dividing the number of trips from country m to n in the year 2016 by the population size of country m and by 365. Modelled arrival times using the GTM have been shown to strongly correlate with those from the global air transportation network[42]. For each value of variant R_e, we performed 10,000 independent simulations of the metapopulation model, assuming that the probability a variant virus would emerge in a particular country is proportional to the

country's relative population size (simulations initialized in Africa: $n = 1793$; Asia: $n = 5946$, Europe: $n = 934$; North America: $n = 739$; Oceania: $n = 54$; South America: $n = 534$). We integrated the model forward in time at a daily timescale using a tau-leap algorithm, which also furnishes the epidemic dynamics and global spread with stochasticity. Each simulation was initialized with an infected population of 10 individuals. We validated the model by comparing the simulated spread dynamics to simulations using an equivalent model in GLEAMviz 7.2[23,24] (www.gleamviz.org), a global metapopulation model that incorporates dynamics of air travel and mobility. Details are given in the Supplementary Information.

Genomic surveillance simulations

We performed the genomic surveillance simulations using empirical turnaround times and sampling rates for each country, using data for 2022. For each sequence in GISAID, we computed the time T between the sample's collection date and submission date. In some cases, sequence analysis might have been performed but the sequence would only later be deposited in GISAID. Hence, given the computed turnaround time T, we assumed that a sequence's adjusted turnaround time $\tilde{T}$ was equal to ϕT, for $0 < \phi < 1$. In the main text, $\phi = 1$, and we performed sensitivity analyses for $\phi = 0.25$ and $\phi = 0.5$. For each country c, the turnaround-time specific sequencing rate in unit of sequences per day $n_{x,c}$, for each value of turnaround time x in days, was equal to the country's total sequencing rate in sequences per day multiplied by the proportion of sequences from that country with $\tilde{T} = x$.

For each country, for each simulation, starting from the first day on which the number of new variant infections exceeded 10 onwards, we deterministically simulated the wildtype epidemic dynamics. For each value of variant R_e, we assumed the same scenario of variant emergence (characterized by a wildtype prevalence and wildtype R_e) as in the single-country analyses presented in Figure 3. In Extended Data Fig. 6, we show the same analyses for all combinations of variant R_e and scenario of variant emergence. Until the first day on which variant incidence exceeded 10, wildtype incidence was assumed to be equal to wildtype incidence on the first day of the simulated wildtype epidemic, to account for the stochasticity observed when the number of infections was small and the potential for stochastic variant extinction.

Using the simulated variant and wildtype incidence on each day, we computed the variant proportion through time *f(t)*. For each country *c*, on each day *t*, we used the simulated country-specific variant proportion $f_c(t)$ to simulate genomic surveillance by, for each value of turnaround time *x,* generating a sample count $\tilde{n}_{x,c} \sim \text{Poisson}(n_{x,c})$, using the estimated turnaround-time specific sequencing rates $n_{x,c}$ described above, and simulating the total number of variant samples $V_c(t) = \sum_{x=0}^{t} v_{x,c}$, with $v_{x,c} \sim \text{Binomial}(\tilde{n}_{x,c}, f_c(t\text{-}x))$. In each of 10,000 replicate simulations, and for each strategy for the global distribution of surveillance infrastructure, we computed the detection day as the first day *t* on which $V_c(t)$ was at least one in at least one country *c*. We defined the detection country as the first country for which this held.

To investigate the sensitivity of our results with respect to mobility rates, we multiplied each country's mobility rate by 3 and 1/3, representing faster and slower spread, respectively, and re-simulated the epidemic dynamics. We applied the same genomic surveillance simulations to the epidemic simulations with increased and reduce mobility rates, respectively, to assess the sensitivity of our results to the mobility dynamics underlying global variant spread.

Global surveillance strategies

We investigated four strategies for the global distribution of sequencing infrastructure:

Strategy 1: the 2022 baseline. For each country, turnaround time-specific sequencing rates were extracted from GISAID metadata.

Strategy 2: Equivalent to strategy 1, but individual countries' sequencing output capped at 30 S/M/wk. Countries that sequenced at rates exceeding 30 S/M/wk had their sequencing output capped by dividing sequencing rate uniformly across all values of turnaround time such that total output across all values of turnaround time was equal to 30 S/M/wk.

Strategy 3: the 2022 baseline + a global minimum sequencing capacity of 2 S/M/wk at 14 day turnaround time in each country. If a country already satisfied this requirement (i.e., the sum of turnaround time-specific sequencing rates with turnaround time ≤14 days was equal to or greater than 2 S/M/wk), its sequencing rates were unchanged relative to strategy 1. If a

492 country satisfied the sequencing rate across all values of turnaround time, but not within the
493 required two-week turnaround time, the deficit in S/M/wk in the sum of turnaround time-
494 specific sequencing rates with turnaround time ≤14 days was uniformly removed from the
495 sequencing rates exceeding 14 days and added to the sequencing rate corresponding to a
496 turnaround time of 14 days. Hence, in this scenario, total sequencing output remained
497 unchanged, and the minimum sequencing capacity was attained by reducing turnaround time.
498 If a country did not satisfy the minimum sequencing rate at all, all sequencing output
499 corresponding to a sequencing rate >14 days was set to a turnaround time of 14 days. The
500 remaining deficit in S/M/wk in the sum of turnaround time-specific sequencing rates with
501 turnaround time ≤14 days was added to the sequencing rate corresponding to a turnaround
502 time of 14 days.
503
504 Strategy 4: In countries that, after capping according to strategy 2, did not satisfy the
505 minimum sequencing rate of 2 S/M/wk at 14 day turnaround time, this minimum was ensured
506 analogous to Strategy 3.
507

References

1. Ladner, J. T. & Sahl, J. W. Towards a post-pandemic future for global pathogen genome sequencing. *PLOS Biol.* **21**, e3002225 (2023).
2. Hill, V. *et al.* Toward a global virus genomic surveillance network. *Cell Host Microbe* **31**, 861–873 (2023).
3. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
4. Cherian, S. *et al.* SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganisms* vol. 9 at https://doi.org/10.3390/microorganisms9071542 (2021).
5. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science (80-.).* **372**, eabg3055 (2021).
6. Chen, Z. *et al.* Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat. Genet.* **54**, 499–507 (2022).
7. Tegally, H. *et al.* The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance. *Science (80-.).* **378**, eabq5358 (2023).
8. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. *Nat. Commun.* **13**, 7003 (2022).
9. Knyazev, S. *et al.* Unlocking capacities of genomics for the COVID-19 response and future pandemics. *Nat. Methods* **19**, 374–380 (2022).
10. Hill, V., Ruis, C., Bajaj, S., Pybus, O. G. & Kraemer, M. U. G. Progress and challenges in virus genomic epidemiology. *Trends Parasitol.* **37**, 1038–1049 (2021).
11. Lin, C. *et al.* Towards equitable access to public health pathogen genomics in the Western Pacific. *Lancet Reg. Heal. – West. Pacific* **18**, (2022).
12. Wohl, S., Lee, E. C., DiPrete, B. L. & Lessler, J. Sample size calculations for pathogen variant surveillance in the presence of biological and systematic biases. *Cell Reports Med.* **4**, (2023).
13. Méder, Z. Z. & Somogyi, R. Optimal capacity sharing for global genomic surveillance. *Epidemics* **43**, 100690 (2023).
14. Boyle, L. *et al.* Selective sweeps in SARS-CoV-2 variant competition. *Proc. Natl. Acad. Sci.* **119**, e2213879119 (2022).
15. Wilkinson, E. *et al.* A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science (80-.).* **374**, 423–431 (2021).

16. Jones, K. E. *et al.* Global trends in emerging infectious diseases. *Nature* **451**, 990–993 (2008).
17. Fraser, C. *et al.* Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings. *Science (80-.).* **324**, 1557–1561 (2009).
18. Davis, J. T. *et al.* Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. *Nature* **600**, 127–132 (2021).
19. Hodcroft, E. B. *et al.* Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
20. Brockmann, D. & Helbing, D. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. *Science (80-.).* **342**, 1337–1342 (2013).
21. Tegally, H. *et al.* Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern. *Cell* **186**, 3277-3290.e16 (2023).
22. Bedford, T., Cobey, S., Beerli, P. & Pascual, M. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLOS Pathog.* **6**, e1000918- (2010).
23. Balcan, D. *et al.* Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Med.* **7**, 45 (2009).
24. Van den Broeck, W. *et al.* The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infect. Dis.* **11**, 37 (2011).
25. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* vol. 22 at https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 (2017).
26. Kalia, K., Saberwal, G. & Sharma, G. The lag in SARS-CoV-2 genome submissions to GISAID. *Nat. Biotechnol.* **39**, 1058–1060 (2021).
27. Sahadeo, N. S. D. *et al.* Implementation of genomic surveillance of SARS-CoV-2 in the Caribbean: Lessons learned for sustainability in resource-limited settings. *PLOS Glob. Public Heal.* **3**, e0001455 (2023).
28. World Health Organization (WHO). *Updated working definitions and primary actions for SARS-CoV-2 variants, 4 October 2023*. (2023).
29. Oude Munnink, B. B. *et al.* The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat. Med.* **27**, 1518–1524 (2021).

30. Genomic sequencing in pandemics. *Lancet* **397**, 445 (2021).
31. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **22**, 757–773 (2021).
32. Madhi, S. A. *et al.* Population Immunity and Covid-19 Severity with Omicron Variant in South Africa. *N. Engl. J. Med.* **386**, 1314–1326 (2022).
33. Hill, V. *et al.* The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol.* **8**, veac080 (2022).
34. Pandemics move faster than funders. *Lancet* **402**, 367 (2023).
35. Han, A. X. *et al.* SARS-CoV-2 diagnostic testing rates determine the sensitivity of genomic surveillance programs. *Nat. Genet.* **55**, 26–33 (2023).
36. Salyer, S. J. *et al.* The first and second waves of the COVID-19 pandemic in Africa: a cross-sectional study. *Lancet* **397**, 1265–1275 (2021).
37. Ellingford, J. M. *et al.* Genomic and healthcare dynamics of nosocomial SARS-CoV-2 transmission. *Elife* **10**, e65453 (2021).
38. Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* **6**, veaa061 (2020).
39. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).
40. Inzaule, S. C., Tessema, S. K., Kebede, Y., Ogwell Ouma, A. E. & Nkengasong, J. N. Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *Lancet Infect. Dis.* **21**, e281–e289 (2021).
41. Deutschmann, E., Recchi, E. & Vespe, M. Assessing Transnational Human Mobility on a Global Scale BT - Migration Research in a Digitized World: Using Innovative Technology to Tackle Methodological Challenges. in (eds. Pötzschke, S. & Rinken, S.) 169–192 (Springer International Publishing, 2022). doi:10.1007/978-3-031-01319-5_9.
42. Klamser, P. P. *et al.* Inferring country-specific import risk of diseases from the world air transportation network. *arXiv e-prints* arXiv:2304.12087 at https://doi.org/10.48550/arXiv.2304.12087 (2023).

Data availability

Data on global population sizes are available from the United Nations World Population Prospects 2022 (https://population.un.org/wpp/Download/Standard/MostUsed/). Data on country GDP (https://data.worldbank.org/indicator/NY.GDP.PCAP.CD) and income

classification (https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups) is available from the World Bank. The Global Transnational Mobility Dataset is available from the Global Mobilities Project (https://migrationpolicycentre.eu/globalmobilities/dataset/). Metadata on global SARS-CoV-2 and seasonal influenza virus sequencing rates were extracted from GISAID (www.gisaid.org).

Code availability

Custom code and data used to generate the results in this study is publicly available at https://github.com/AMC-LAEB/genomic_surveillance_equity. Raw global epidemic simulation output is available at https://zenodo.org/records/10051237.

Acknowledgements

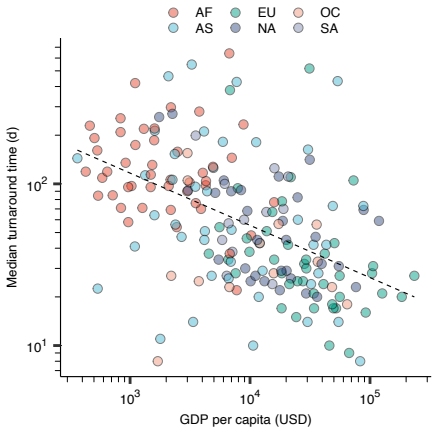
This work was supported by the European Research Council (grant number 818353). We gratefully acknowledge the originating and submitting laboratories that generated the sequence data in GISAID that this study relies on.

Author contributions

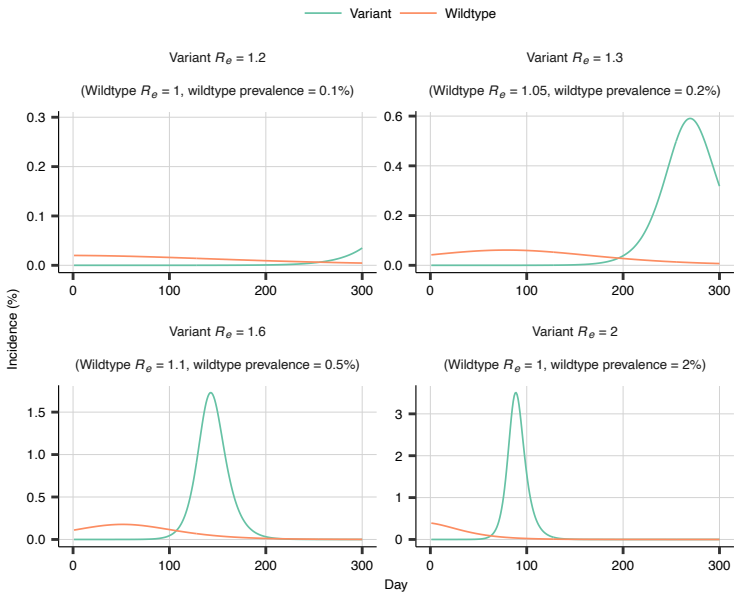
S.P.d.J., B.E.N., A.X.H., and C.A.R. designed the research; S.P.d.J. and A.X.H. performed the data analysis and modelling work; S.P.d.J., A.X.H., and C.A.R. wrote the first draft of the paper. All authors contributed to the critical revision of the paper.

Competing interests

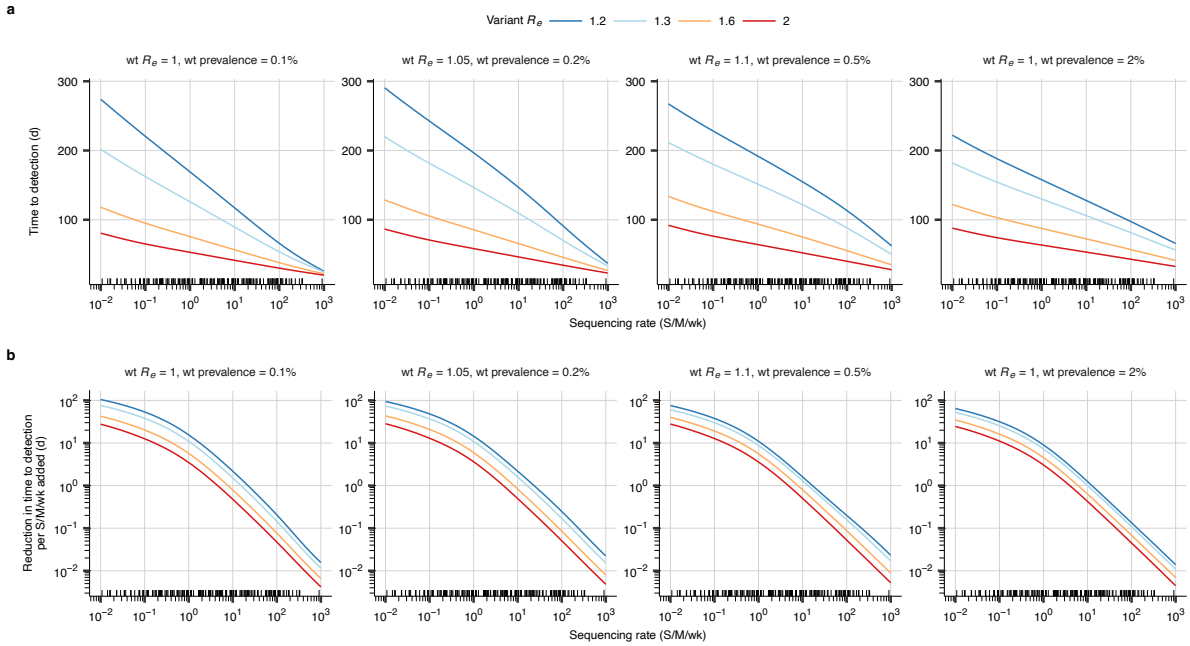
The authors declare no competing interests related to this work.



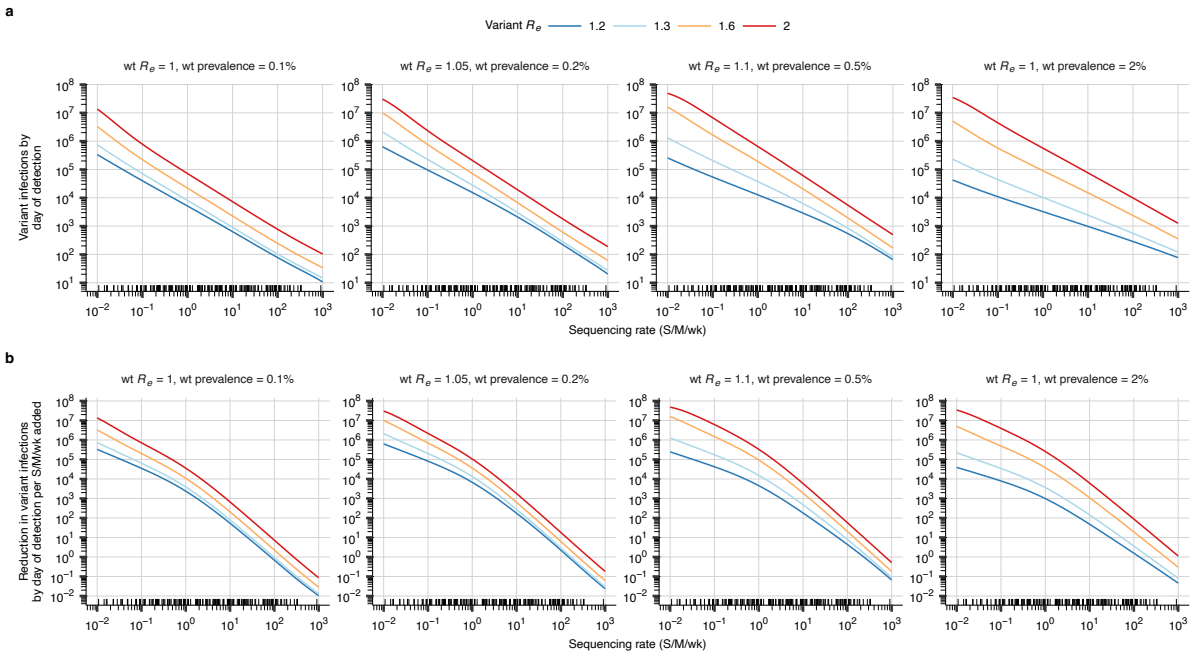
Extended Data Fig. 1. Relationship between median turnaround time and per capita GDP. Each point corresponds to a country *(n* =188), coloured by continent.



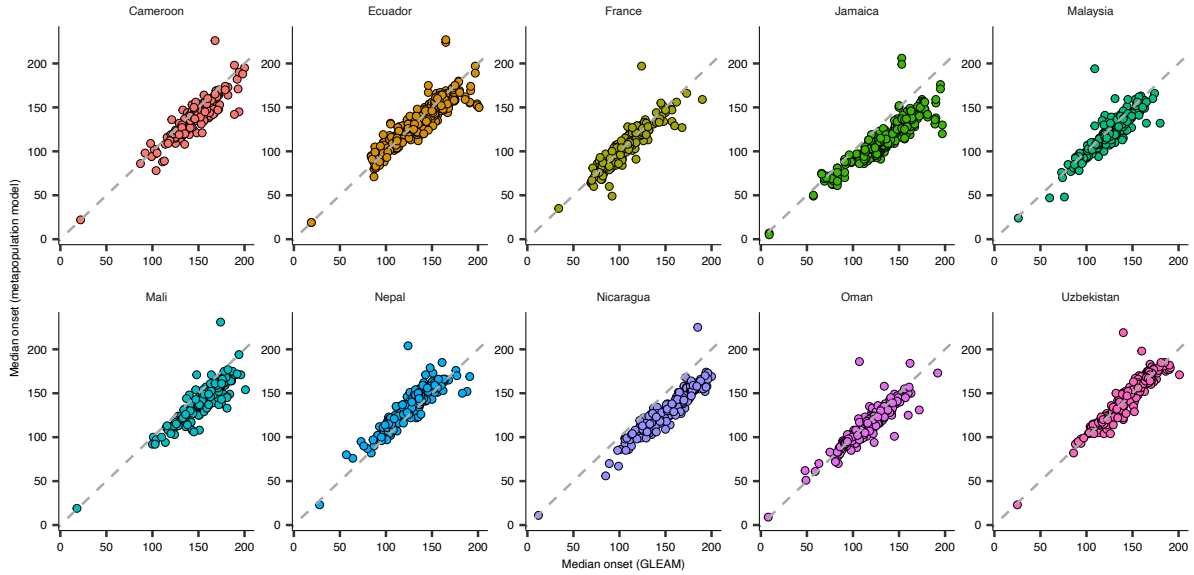
Extended Data Fig. 2. Different scenarios of variant emergence. For each of the values of variant R_e, the corresponding panel shows the epidemiological dynamics of variant and wildtype for that scenario of variant emergence, starting from the day of variant introduction. For each value of variant R_e, the scenario of variant emergence is characterised by a different value of wildtype R_e and wildtype prevalence at introduction.



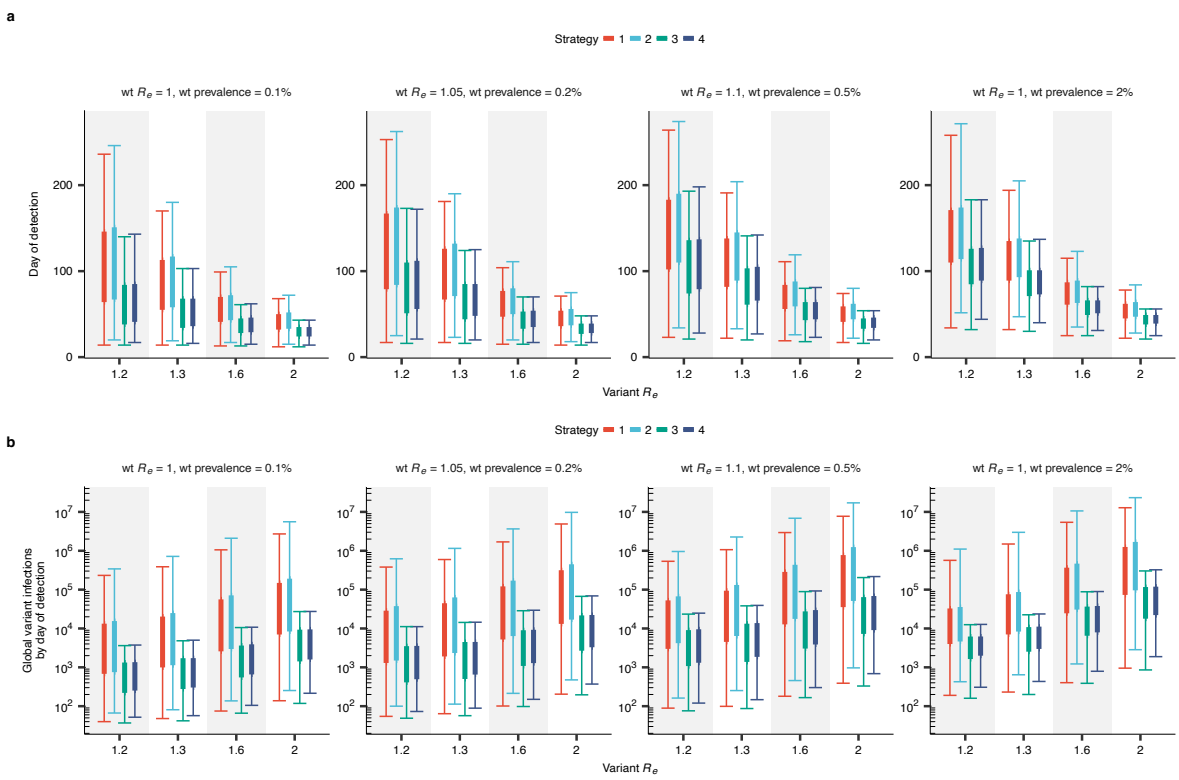
Extended Data Fig. 3. The dependence of time to variant detection on sequencing rate for varying scenario of variant emergence. a) Relationship between sequencing rate and the number of days until the variant will have been detected with 95% confidence. The small black tick marks on the *x*-axes in this plot and in **b** show country-specific SARS-CoV-2 sequencing rates for 2022. Each panel corresponds to a different scenario of variant emergence, characterized by a wildtype (wt) R_e and wildtype prevalence at introduction. In each panel, lines are colored by value of variant R_e. **b)** Relationship between sequencing rate and the reduction in time to variant detection that results from increasing the existing sequencing rate (*x*-axis) by 1 S/M/wk.



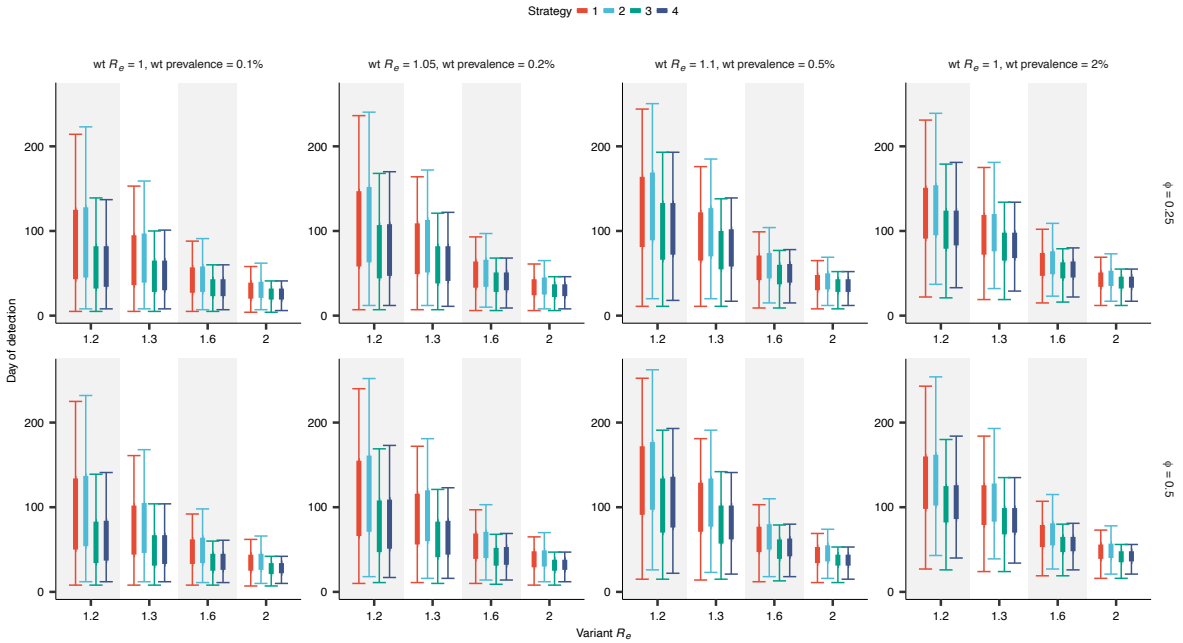
Extended Data Fig. 4. The dependence of the number of variant infections by the day of variant detection on sequencing rate for varying scenario of variant emergence. **a)** Relationship between sequencing rate and the number of variant infections by the day the variant will have been detected with 95% confidence. The small black tick marks on the *x*-axes in this plot and in **b** show country-specific SARS-CoV-2 sequencing rates for 2022. Each panel corresponds to a different scenario of variant emergence, characterized by a wildtype (wt) R_e and wildtype prevalence at introduction. In each panel, lines are colored by value of variant R_e. **b)** Relationship between sequencing rate and the reduction in the number of variant infections by the day of detection that results from increasing the existing sequencing rate (*x*-axis) by 1 S/M/wk.



Extended Data Fig. 5. Validation of the metapopulation model against GLEAM. For ten geographically representative countries, global variant spread was simulated, initialised in the country's capital city, in GLEAM. For each of the ten index countries, all global countries' epidemic onset timings as simulated using GLEAM were compared against the countries' epidemic onset timings as simulated using the epidemic model used in this study. For both models, timings were computed as the median across 10 independent simulations. Simulations are for a variant R_e of 1.6.

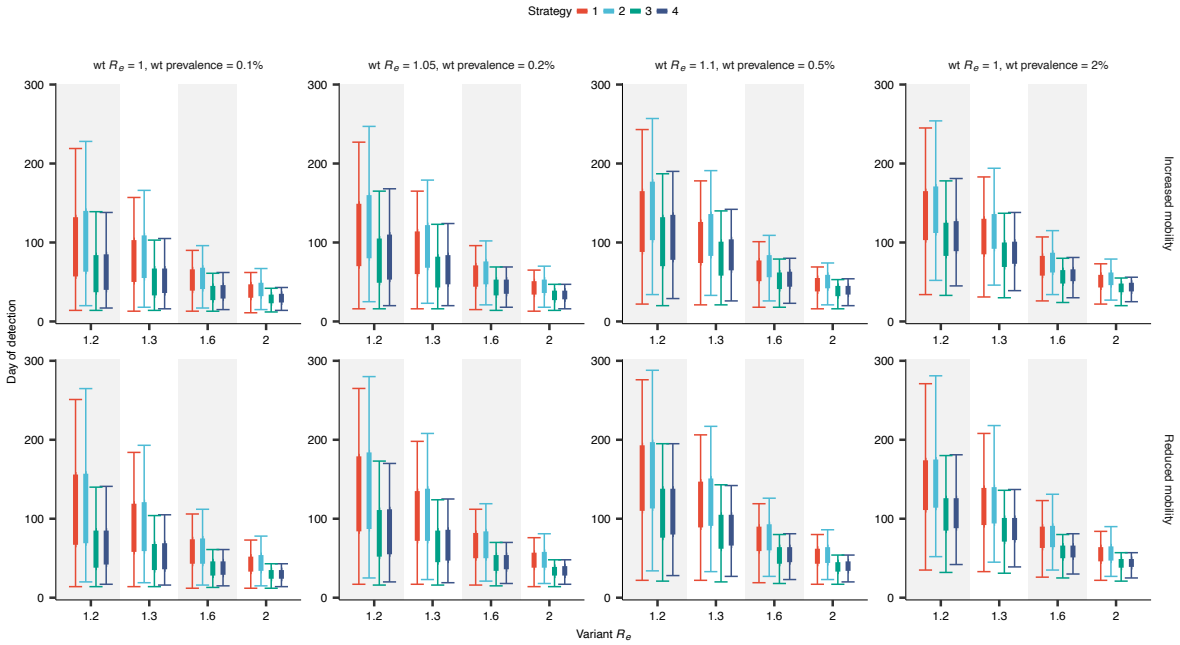


Extended Data Fig. 6. Time to global variant detection by strategy and scenario of variant emergence. a) Time to global variant detection by strategy for the global distribution of respiratory virus surveillance infrastructure, by variant R_e, for varying scenario of variant emergence (characterised by wildtype (wt) R_e and wildtype prevalence). Thin and thick lines correspond to 95% and 50% CIs, respectively. **b)** Number of global variant infections by the day of first detection by strategy for the global distribution of respiratory virus surveillance infrastructure, by variant R_e, for varying scenario of variant emergence (characterised by wildtype (wt) R_e and wildtype prevalence). Thin and thick lines correspond to 95% and 50% CIs, respectively.



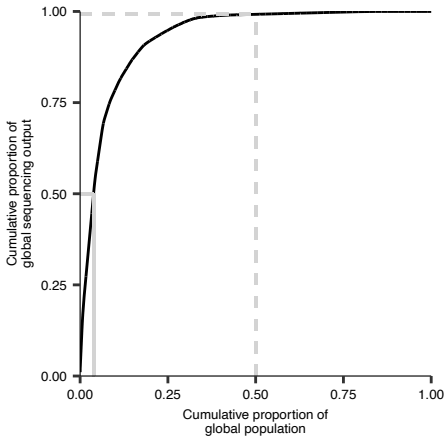
Extended Data Fig. 7. Sensitivity analysis for delay in time to GISAID submission. Time to global variant detection by strategy for the global distribution of respiratory virus surveillance infrastructure, by variant R_e, for varying scenario of variant emergence (characterised by wildtype (wt) R_e and wildtype prevalence). Thin and thick lines correspond to 95% and 50% CIs, respectively. Given a sequence in GISAID's computed turnaround time T, a sequence's adjusted turnaround time $\tilde{T}$ was equal to ϕT. These adjusted turnaround times were used to inform country-specific sequencing infrastructure in the global genomic surveillance simulations.

717



718

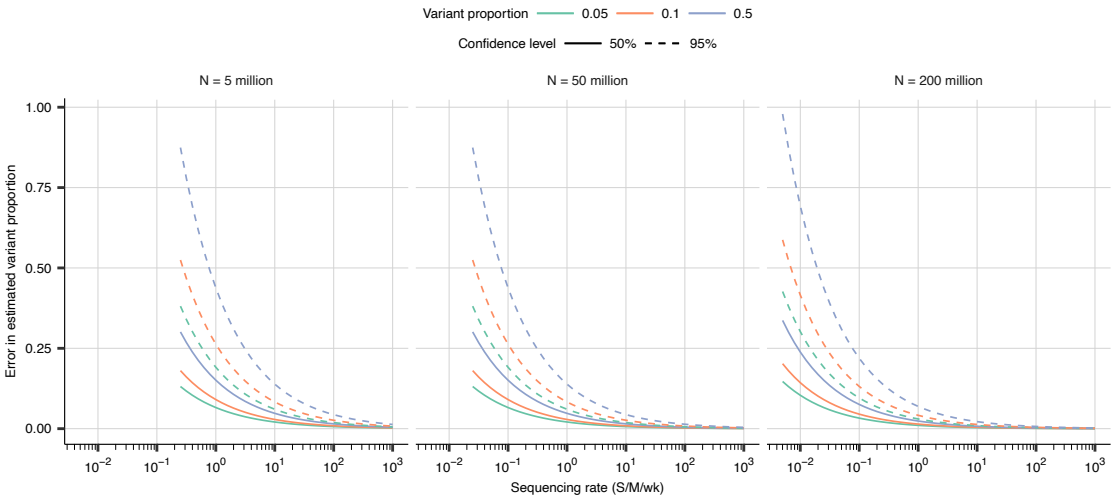
719 **Extended Data Fig. 8. Sensitivity analysis for mobility rate.** Time to global variant
720 detection by strategy for the global distribution of respiratory virus surveillance
721 infrastructure, by variant R_e, for varying scenario of variant emergence (characterised by
722 wildtype (wt) R_e and wildtype prevalence). Thin and thick lines correspond to 95% and 50%
723 CIs, respectively. Each row corresponds to a modified global mobility rate (top: baseline
724 mobility rate multiplied by 3; bottom: baseline mobility rate divided by 3).



725

726 **Extended Data Fig. 9. The global distribution of seasonal influenza sequencing output in**
727 **2018.** The cumulative proportion of the global population that accounts for a cumulative
728 proportion of global sequence output. Solid grey lines show the smallest proportion of the
729 population that accounts for 50% of sequencing output. Dashed grey lines show the smallest
730 proportion of sequencing output that is accounted for by 50% of the global population. Data
731 is for seasonal influenza sequences in GISAID collected from humans in 2018.

732
733



Extended Data Fig. 10. The relationship between sequencing rate and error in estimated variant proportion. For each sequencing rate given on the *x*-axis, for varying true variant proportion and population size, the *y*-axis shows the maximum error in the estimated weekly proportion of total infections attributable to the variant. This maximum error is presented for varying confidence (i.e. the *y*-axis represents the error that the error in the estimated variant proportion relative to the true variant proportion will be smaller than *n*% of the time, for *n* given by the confidence level).