Global disparities in SARS-CoV-2 genomic surveillance

Anderson F. Brito[†, *, 1, 2]; Elizaveta Semenova[†, 1, 3]; Gytis Dudas[†, 1, 4]; Gabriel W. Hassler[5]; Chaney C. Kalinich[1, 6]; Moritz U.G. Kraemer[7]; Joses Ho[8, 9]; Houriiyah Tegally[10]; George Githinji[11, 12]; Charles N. Agoti[11]; Lucy E. Matkin[7]; Charles Whittaker[13, 14]; Danish Covid-19 Genome Consortium; COVID-19 Impact Project; Network for Genomic Surveillance in South Africa (NGS-SA); GISAID core curation team; Benjamin P Howden[15]; Vitali Sintchenko[16, 17]; Neta S. Zuckerman[18]; Orna Mor[18]; Heather M Blankenship[19]; Tulio de Oliveira[10, 20, 21, 22]; Raymond T. P. Lin[23]; Marilda Mendonça Siqueira[24]; Paola Cristina Resende[24]; Ana Tereza R. Vasconcelos[25]; Fernando R. Spilki[26]; Renato Santana Aguiar[27, 28]; Ivailo Alexiev[29]; Ivan N. Ivanov[29]; Ivva Philipova[29]; Christine V. F. Carrington[30]; Nikita S. D. Sahadeo[30]; Céline Gurry[8]; Sebastian Maurer-Stroh[8, 9, 23]; Dhamari Naidoo[31]; Karin J von Eije[32, 33]; Mark D. Perkins[33]; Maria van Kerkhove[33]; Sarah C. Hill[34]; Ester C. Sabino[35]; Oliver G. Pybus[7, 34]; Christopher Dye[7]; Samir Bhatt[13, 14, 36]; Seth Flaxman[37]; Marc A. Suchard[5, 38, 39]; Nathan D. Grubaugh[‡, 1, 40]; Guy Baele[‡, 41]; Nuno R. Faria[‡, *, 7, 13, 16, 35]

1. Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA
2. Instituto Todos pela Saúde, São Paulo, São Paulo, Brazil
3. Department of Mathematics, Imperial College London, London, UK
4. Gothenburg Global Biodiversity Centre, Gothenburg, Sweden
5. Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, USA
6. Yale School of Medicine, Yale University, New Haven, Connecticut, USA
7. Department of Zoology, University of Oxford, Oxford, United Kingdom
8. GISAID Global Data Science Initiative, Munich, Germany
9. Bioinformatics Institute & ID Labs, Agency for Science Technology and Research, Singapore, Singapore
10. KwaZulu–Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine and Medical Sciences, University of KwaZulu–Natal, Durban, South Africa
11. KEMRI-Wellcome Trust Research Programme, Kenya
12. Department of Biochemistry and Biotechnology, Pwani University, Kenya
13. MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, United Kingdom
14. The Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), School of Public Health, Imperial College London, London, United Kingdom
15. Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology and Immunology, The University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia
16. Sydney Institute for Infectious Diseases, The University of Sydney, Sydney, New South Wales, Australia
17. Institute of Clinical Pathology and Medical Research, NSW Health Pathology, Westmead, New South Wales 2145, Australia
18. Central Virology Laboratory, Israel Ministry of Health, Sheba Medical Center, Israel
19. Michigan Department of Health and Human Services, Bureau of Laboratories, Lansing, Michigan, USA

20. Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa
21. Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa
22. Department of Global Health, University of Washington, Seattle, Washington, USA
23. National Centre for Infectious Diseases, Singapore
24. Laboratory of Respiratory Viruses and Measles, FIOCRUZ, Rio de Janeiro, Brazil
25. Laboratório de Bioinformática, Laboratório Nacional de Computação Científica, Petrópolis, Brazil
26. Feevale University, Institute of Health Sciences, Novo Hamburgo, Rio Grande do Sul, Brazil
27. Laboratório de Biologia Integrativa, Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
28. Instituto D'Or de Pesquisa e Ensino (IDOR), Rio de Janeiro, Brazil
29. National Center of Infectious and Parasitic Diseases, Sofia, Bulgaria
30. Department of Preclinical Sciences, Faculty of Medical Sciences, The University of the West Indies, St. Augustine, Trinidad and Tobago
31. Health Emergencies Programme, World Health Organization Regional Office for South-East Asia, New Delhi, India
32. Department of Medical Microbiology and Infection Prevention, Division of Clinical Virology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
33. Emerging Diseases and Zoonoses Unit, Health Emergencies Programme, World Health Organization, Geneva, Switzerland
34. Royal Veterinary College, Hawkshead, United Kingdom
35. Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil
36. Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark
37. Department of Computer Science, University of Oxford, Oxford, United Kingdom
38. Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, California, USA
39. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, USA
40. Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA
41. Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium

* Co-corresponding authors (AFB: andersonfbrito@gmail.com and NRF: nfaria@ic.ac.uk)
† Co-first authorship
‡ Co-senior authorship

Abstract: Genomic sequencing provides critical information to track the evolution and spread of SARS-CoV-2, optimize molecular tests, treatments and vaccines, and guide public health responses. To investigate the spatiotemporal heterogeneity in the global SARS-CoV-2 genomic surveillance, we estimated the impact of sequencing intensity and turnaround times (TAT) on variant detection in 167 countries. Most countries submit genomes >21 days after sample collection, and 77% of low and middle income countries sequenced <0.5% of their cases. We found that sequencing at least 0.5% of the cases, with a TAT <21 days, could be a benchmark for SARS-CoV-2 genomic surveillance efforts. Socioeconomic inequalities substantially impact our ability to quickly detect SARS-CoV-2 variants, and undermine the global pandemic preparedness.

One-Sentence Summary: Socioeconomic inequalities impacted the SARS-CoV-2 genomic surveillance, and undermined the global pandemic preparedness.

The importance of genomic surveillance

Twenty months into the COVID-19 pandemic, many countries continue to face large epidemics of SARS-CoV-2 infections (*1*), mostly driven by the emergence and spread of novel viral variants (*2*). Genomic surveillance has been critical to the study of SARS-CoV-2 evolution and spread, to the design and optimization of diagnostic tools and vaccines, and to the early identification and assessment of viral lineages with altered epidemiological characteristics, including variants of concern (VOCs) such as Alpha/B.1.1.7; Beta/B.1.351; Gamma/P.1; and Delta/B.1.617.2. These lineages pose increased global public health risks due to their greater transmissibility and potential immune escape from neutralizing antibodies induced by natural infections and/or vaccines (*3*, *4*). Variants of interest (VOIs) also require continued monitoring for changes in transmissibility, disease severity, or antigenicity (*5*). To help guide public health responses to evolving variants, it is essential to track the diversity of SARS-CoV-2 lineages circulating worldwide in near real-time (*3*, *6*, *7*). An unprecedented number of SARS-CoV-2 viral genomes have now been released in publicly accessible databases, with >4 million consensus genome sequences shared via the EpiCoV database at the GISAID data science initiative (*8*) and >1.2 million high-throughput sequencing datasets and >1.7 million consensus sequences in National Center for Biotechnology Information as of October 1st, 2021. Until then, and as a comparison, 324,992 influenza genome sequences have been shared in the GISAID database. Despite improvements in models for equitable sharing of pathogen genomic data (*9*), there are striking differences in the intensity of genomic surveillance within and among countries worldwide. Here we examine global publicly-accessible SARS-CoV-2 genomic surveillance data from the first 15 months of the COVID-19 pandemic to identify key aspects associated with sequencing intensity and timely variant detection, and investigate the consequences of surveillance disparities.

Global disparities in SARS-CoV-2 genomic surveillance

To investigate spatial and temporal heterogeneity in SARS-CoV-2 genome sequencing intensity, we explored the percentage of COVID-19 cases sequenced each week per country from February 2020 to March 2021 (**Fig. 1A; Table S1**). It has been proposed that at least 5% of SARS-CoV-2 positive samples should be sequenced to detect viral lineages at a prevalence of 0.1 to 1.0% (*10*). Only 16 countries (or 9.6%) worldwide sequenced 5% or more of their total confirmed cases, while 100 out of 167 countries had <0.5% of confirmed cases sequenced (**Fig. 1B; Fig. S1**). A total of 72 countries had <25% of their genomes sequenced locally, and relied mostly on sequencing capacity in other countries to get their cases sequenced (**Fig. S2; Table S2**). Among high-income countries (HICs) and low- and middle-income countries (LMICs), while the number of reported cases was relatively similar until late March 2021 (65.3 and 61.2 million, respectively), HICs shared on average 16.5–fold more sequences per reported case (1.81% and 0.11% for HIC and LMICs, respectively) (**Table S3**). A moderate negative correlation between weekly sequencing percentages and reported COVID-19 incidence was observed (cases/100K pop., $r^2 = -0.52$; *p*-value < 0.001), suggesting that countries with low incidence (**Fig. 1C; Fig. S3**) were able to sequence higher proportions of cases. Exceptionally, some countries, such as Denmark and the UK, faced high weekly COVID-19 incidence in late 2020 but were still able to maintain sequencing intensity >10% in most weeks (32% and 8% of total confirmed cases, respectively) (**Fig. 1A-B**; **Fig. S3**).

Most countries in Africa and Asia, despite reporting low COVID-19 incidence, did not reach genomic surveillance levels similar to the Gambia (8.6%), Japan (7.3%), Hong Kong (12.3%), New Zealand (3.8%) and Australia (5.9%), which also experienced low COVID-19 incidences (**Fig. 1B-C**; **Fig. S3**). Likewise, sequencing of >0.5% of cases has not been achieved in most Latin American countries, particularly during periods of high incidence (**Fig. S3**). This finding is robust to under-ascertainment of reported cases due to more limited availability of diagnostic tests. Our

study also revealed an absence of SARS-CoV-2 genomes in public databases from >20 LMICs; for some countries, the only available information on the diversity of circulating lineages has been obtained from travel-related infections sequenced abroad (**Fig. S2**). Overall, most countries did not achieve high or moderate percentages (0.1% to 1%) of sequenced cases each week of the pandemic (**Fig. 1**; **Fig. S3**).

We also described turnaround time (TAT; defined as the time in days between sample collection and genome submission) of SARS-CoV-2 genome sequencing across 19 geographic regions (**Fig. 1D**; see also (*11*)). On average, virus sequences were deposited in public databases 48 days after sample collection, but in 2021, following the detection of the Alpha VOC, efforts were made in nearly all geographic regions to decrease TAT, and provide faster responses (**Fig. 1D**; see **Fig. S4** for weekly changes in TAT across regions). Rapid generation and sharing of pathogen sequence data from regularly-collected samples is essential to maximize public health impact of genomic data (*12*, *13*). The VOCs Alpha and Gamma, for example, reached up to 50% frequency within 2 to 3 months of their emergence in the U.K. and Manaus, respectively (*14*, *15*). Thus, quick TAT is essential for the timely recognition and assessment of transmissibility potential of VOCs.

Sampling strategies for rapid variant detection

We investigated the impact of genome sequencing intensity and TAT on the detection of SARS-CoV-2 lineages. Similar to what has been observed in the UK (*14*), the number of globally observed lineages correlates with the number SARS-CoV-2 genomes available per country (Pearson's $r = 0.96$, p-value<0.0001) and the overall proportion of sequenced cases in each country (Pearson's $r = 0.48$, p-value<0.0001) (**Fig. S5**). This implies that limited genome sequencing intensity may affect the identification and response to new viral lineages with altered epidemiological and antigenic characteristics. To investigate strategies for rapid variant detection, we simulated the impact of the percentage of sequenced cases and TAT on the reliable detection of previously-identified SARS-CoV-2 lineages using metadata from Denmark, which has one of the most comprehensive SARS-CoV-2 genome surveillance systems (see **Materials and Methods**). Because several countries have opportunistically selected samples for sequencing based on testing characteristics, e.g. spike gene target failures of a commonly-used PCR assay, or additional, often unspecified characteristics, such as imported cases or severe disease, we focused on analysing data collected prior to November 2020 (**Fig. S6**).

We assumed a recommended scenario of random sampling, whereby samples for virus genomic sequencing are selected independently of sample metadata such as age, sex, or clinical symptoms (*15*). When calculating the probability of detecting at least one genome of a rare lineage (0–5% prevalence) under different sequencing intensities, we found that sequencing at least 300 genomes per week is required to detect, with a 95% probability, a lineage that is circulating in a population at a weekly prevalence of 1%. For a weekly prevalence of 5%, this number decreases to 75 genomes per week (**Fig. 2A**). These figures are independent of outbreak and population size of a given location, and can only tell if a lineage is present, not how prevalent it is, and furthermore assumes representative sampling. On average, genome surveillance programmes in high income countries should be able to detect circulating virus lineages at 5% prevalence with maximum probability under the assumption of random sampling (**Fig. 2B; Table 1**). However, under a scenario of random sampling, low income countries that sequence an average of 9 genomes per week may miss a SARS-CoV-2 lineage circulating at up to 26% prevalence. This will present a substantial limitation to the lines of inquiry available to such countries from genome sequencing data (**Table 1)**. Within the range 0.05–5% sequences per case considered here, increasing sampling intensity and at a lesser extent reducing TAT strongly improves the rapid detection of viral lineages (**Fig. 2B**).

Next, we simulated 25 scenarios with 100 replicates, in which we varied sampling frequency (from 0.05% to 5%) and TAT (from 35 to 7 days) to compute the probabilities of detecting at least one genome of a given lineage before the lineage reaches a cumulative size of 100 cases (**Fig. 2B**), using as "ground truth" a dataset from a well characterized setting (see **Materials and Methods** and **Fig. S6**). The simulated scenario shows that when sequencing percentages of 5% per week and turnaround times of 7 days are achieved in a given setting, there is a 48% probability of detecting a viral lineage before it reaches 100 cases randomly selected from the population. When the proportion of sequenced cases per week decreases by 100-fold, to 0.05%, the probability of the timely detection of a viral lineage before it reaches 100 cases decreases to 4.8% for turnaround times of 7 days, and further declines to 2.6% when turnaround time is 35 days (**Fig. 2B**).

For an optimistic scenario of 0.5% of sequenced cases (achieved by 69% of HICs and 23% of LMICs) and a turnaround time of 21 days (achieved by 14% of the HICs and 3% of the LMICs) (**Table S4**), we found a 20% probability of detecting a lineage before it reaches 100 cases. Throughout the pandemic, many countries reported weekly incidences as high as 100 cases per 100,000 inhabitants (**Fig. 1C, Fig. S3**). For example, in such a scenario of high incidence, for Manaus (2.2 million inhabitants, Amazonas state, Brazil), the 0.5% sequencing threshold would correspond to 11 randomly selected genomes per week. With a 21-day turnaround time, this would allow the detection of a given lineage with a 20% probability (**Fig. 2B**). For São Paulo (12.4 million inhabitants), this number increases to 62 genomes per week. For Brazil (212.6 million inhabitants), this would correspond to 1,063 weekly genomes selected from a random population of samples, in the above mentioned scenario of high incidence. Although the 0.5% ratio of sequenced cases per week in near real-time is a reasonable benchmark for SARS-CoV-2 genomic surveillance in over two thirds of high income country settings (**Table S4**), this often comes as a result of close coordination between diagnostic centers and well-funded, decentralized infrastructures to integrate sequencing data and sample-associated metadata (see e.g. (*16*)).

Factors associated with genomic surveillance capacity

While many HICs were able to rely on previously established networks and laboratory infrastructure to perform molecular testing and sequencing (*17*, *18*), many LMICs – including Brazil, South Africa, and India where three VOCs are believed to have emerged (*19–22*) - have faced additional challenges to rapid expansion of genomic surveillance (*18*, *23*, *24*). Pathogen genomics complements but often competes for limited resources with other aspects of pandemic response, for instance, surveillance and testing capacity, medical supplies, laboratory reagents, public health and social measures, vaccine development, and supplies (*25*). To investigate how socioeconomic factors can impact SARS-CoV-2 genomic surveillance response around the

world, we explored the correlation between the percentage of sequenced COVID-19 cases in each country, and 20 country-level socioeconomic and health quality covariates (**Table S5**). We found that the percentage of sequenced cases are significantly associated with expenditure on research and development (R&D) per capita (r^2=0.47, p-value<0.0001), GDP per capita (0.37, p-value<0.0001), socio-demographic index (0.31, p-value<0.001), and established influenza virus genomic surveillance capacity prior to the COVID-19 pandemic (0.30, p-value<0.001) (**Fig. 3; Table S6**).

Before January 2020, only 67% (113 out of 167) of the countries that uploaded SARS-CoV-2 genomes to public databases had shared influenza virus genome sequences. When we compared breakdown by income class, we observed that the majority of UMCs (77%) and HICs (78%) sequencing SARS-CoV-2 had already reported influenza virus sequences in public databases up to 2019. For LICs and LMCs countries, this number drops to 39% and 54%, respectively, suggesting that many LICs and LMCs initiated genome sequencing programmes during the COVID-19 pandemic. While disparities in investment in national health, research, and development continue to impact the ability of countries to scale up genomic surveillance intensity (*6*, *18*, *26*), the uptake in genomic surveillance by many LMICs and the association of sequencing efforts with established genomic surveillance capacity provide an encouraging picture for future pandemic preparedness programmes.

When we explored correlations with mean turnaround time (**Table S7**), we found that universal health coverage (r^2=-0.45, p-value<0.0001), healthcare access and quality index (-0.44, p-value<0.0001), socio-demographic index (-0.42, p-value<0.0001), and health expenditure per capita (-0.4, p-value<0.0001) are significantly correlated with mean turnaround times (**Fig. S7, Table S7**). Our results quantify only correlations between socioeconomic covariates, sequencing intensity, and turnaround time, and cannot be interpreted as causal. Future studies should focus on additional variables, such as training laboratory and bioinformatic personnel, costs associated with imported consumables, and shipment delays that may be exacerbated by border closures and travel restrictions (*6*, *23*, *24*, *26*, *27*). Other factors associated with delays in reporting VOCs include social and political stigma and perceived negative impact on travel when reporting potential VOCs, and concerns of having findings scooped and published by other researchers (*28*). Longer turnaround times are also expected in countries where virus genomics activities are focused on retrospective genomic studies to investigate SARS-CoV-2 reinfections (*29*), vaccine breakthrough infections (*30*), and past epidemic dynamics (*31*, *32*).

Conclusions

Strengthening pathogen genomic surveillance efforts worldwide, but particularly in LMICs, should be a global priority to improve pandemic preparedness. Our findings demonstrate that global SARS-CoV-2 genomic surveillance efforts are currently highly unbalanced, and contingent upon socioeconomic factors and pre-pandemic laboratory and surveillance capacity. Our results suggest that sequencing 0.5% of total confirmed cases, with a TAT below 21 days, could provide a benchmark for genomic surveillance studies targeting SARS-CoV-2 and future emerging viruses. Ongoing surveys to understand barriers to virus genome sequencing and sampling selection strategies will provide valuable information for future surveillance programmes. Implementation of metagenomic approaches for virus discovery followed by virus-genome specific sequencing approaches could help overcome existing limitations of molecular and syndromic surveillance strategies (*33*). Adoption of standardized protocols for representative genomic surveillance strategies (*15*, *34*), rapid integration of sequence and sample-associated metadata, and collaboration between academia, public health laboratories and other stakeholders will be essential to maximize cost-effectiveness and public health impact of genomic surveillance. While a random sampling strategy may provide accurate information into SARS-CoV-2 variant emergence and frequency estimation, we note that genome sampling strategies should be considered pathogen- and question-specific (*15*). For example, non-random selection of samples stratified by disease severity may be required to identify genes or mutations associated with clinical outcomes.

Our findings call for strengthening equitable strategies that increase confidence in data sharing for improving global genomic surveillance (*28*). There are several global efforts underway to improve genomic sequencing capacities around the world, including the AFRO-Africa Centre for Disease Control, the Pan American Health Organization COVIGEN Network, South East Asian SARS-CoV-2 Genomics Consortium, and the ACT-A WHO Global Risk Monitoring Framework. These global efforts must be made to improve in-country genomic surveillance capacity and guarantee sustainable research funding for low and middle income countries. Improved pathogen surveillance at the human, animal and human-animal interfaces is also urgently needed (*35*). Retaining existing and expanding local capacity efforts acquired during the SARS-CoV-2 pandemic will be critical to contain and respond to the next “Disease X" (*35*).

References and Notes

1. WHO. *WHO COVID-19 Explorer* (2021), (available at https://worldhealthorg.shinyapps.io/covid/).

2. A. S. Lauring, E. B. Hodcroft, Genetic Variants of SARS-CoV-2—What Do They Mean? *JAMA*. **325**, 529–531 (2021).

3. The Lancet, Genomic sequencing in pandemics. *Lancet*. **397**, 445 (2021).

4. CDC, Cases, Data, and Surveillance (2021), (available at https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html).

5. WHO, Tracking SARS-CoV-2 variants. *WHO* (2021), (available at https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/).

6. S. C. Inzaule, S. K. Tessema, Y. Kebede, A. E. Ogwell Ouma, J. N. Nkengasong, Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *Lancet Infect. Dis.* (2021), doi:10.1016/S1473-3099(20)30939-7.

7. N. D. Grubaugh, E. B. Hodcroft, J. R. Fauver, A. L. Phelan, M. Cevik, Public health actions to control new SARS-CoV-2 variants. *Cell*. **184**, 1127–1132 (2021).

8. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* **22** (2017), doi:10.2807/1560-7917.es.2017.22.13.30494.

9. A. Maxmen, One million coronavirus sequences: popular genome site hits mega milestone. *Nature*. **593**, 21 (2021).

10. D. Vavrek, L. Speroni, K. J. Curnow, M. Oberholzer, V. Moeder, P. G. Febbo, Genomic surveillance at scale is required to detect newly emerging strains at an early timepoint, , doi:10.1101/2021.01.12.21249613.

11. K. Kalia, G. Saberwal, G. Sharma, The lag in SARS-CoV-2 genome submissions to GISAID. *Nat. Biotechnol.* (2021), doi:10.1038/s41587-021-01040-0.

12. T. R. Frieden, C. T. Lee, A. F. Bochner, M. Buissonnière, A. McClelland, 7-1-7: an organising principle, target, and accountability metric to make the world safer from pandemics. *Lancet*. **398**, 638–640 (2021).

13. WHO, Policy statement on data sharing by WHO in the context of public health emergencies. *Wkly. Epidemiol. Rec.* **91**, 237–240 (2016).

14. L. du Plessis, J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, N. R. Faria, B. Jackson, N. J. Loman, Á. O'Toole, S. M. Nicholls, K. V. Parag, E. Scher, T. I. Vasylyeva, E. M. Volz, A. Watts, I. I. Bogoch, K. Khan, COVID-19 Genomics UK (COG-UK) Consortium, D. M. Aanensen, M. U. G. Kraemer, A. Rambaut, O. G. Pybus, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*. **371**, 708–712 (2021).

15. WHO, Guidance for surveillance of SARS-CoV-2 variants: Interim guidance, 9 August 2021. *WHO* (2021), (available at https://www.who.int/publications/i/item/WHO_2019-nCoV_surveillance_variants).

16. S. M. Nicholls, R. Poplawski, M. J. Bull, A. Underwood, M. Chapman, K. Abu-Dahab, B. Taylor, R. M. Colquhoun, W. P. M. Rowe, B. Jackson, V. Hill, Á. O'Toole, S. Rey, J. Southgate, R. Amato, R. Livett, S. Gonçalves, E. M. Harrison, S. J. Peacock, D. M. Aanensen, A. Rambaut, T. R. Connor, N. J. Loman, COVID-19 Genomics UK (COG-UK) Consortium, CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Genome Biol.* **22**, 196 (2021).

17. D. Cyranoski, Alarming COVID variants show vital role of genomic surveillance. *Nature*. **589**, 337–338 (2021).

18. M. Wadman, Blind spots thwart global coronavirus tracking. *Science*. **372**, 773–774 (2021).

19. N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. da S. Candido, S. Mishra, M. A. E. Crispim, F. C. S. Sales, I. Hawryluk, J. T. McCrone, R. J. G. Hulswit, L. A. M. Franco, M. S. Ramundo, J. G. de Jesus, P. S. Andrade, T. M. Coletti, G. M. Ferreira, C. A. M. Silva, E. R. Manuli, R. H. M. Pereira, P. S. Peixoto, M. U. G. Kraemer, N. Gaburo Jr, C. da C. Camilo, H. Hoeltgebaum, W. M. Souza, E. C. Rocha, L. M. de Souza, M. C. de Pinho, L. J.

T. Araujo, F. S. V. Malta, A. B. de Lima, J. do P. Silva, D. A. G. Zauli, A. C. de S. Ferreira, R. P. Schnekenberg, D. J. Laydon, P. G. T. Walker, H. M. Schlüter, A. L. P. Dos Santos, M. S. Vidal, V. S. Del Caro, R. M. F. Filho, H. M. Dos Santos, R. S. Aguiar, J. L. Proença-Modena, B. Nelson, J. A. Hay, M. Monod, X. Miscouridou, H. Coupland, R. Sonabend, M. Vollmer, A. Gandy, C. A. Prete Jr, V. H. Nascimento, M. A. Suchard, T. A. Bowden, S. L. K. Pond, C.-H. Wu, O. Ratmann, N. M. Ferguson, C. Dye, N. J. Loman, P. Lemey, A. Rambaut, N. A. Fraiji, M. do P. S. S. Carvalho, O. G. Pybus, S. Flaxman, S. Bhatt, E. C. Sabino, Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*. **372**, 815–821 (2021).

20. J. Raghwani, L. du Plessis, J. T. McCrone, S. C. Hill, K. V. Parag, J. Thézé, D. Kumar, A. Puvar, R. Pandit, O. G. Pybus, G. Fournié, M. Joshi, C. Joshi, Genomic epidemiology of early SARS-CoV-2 transmission dynamics in Gujarat, India. *bioRxiv* (2021), , doi:10.1101/2021.08.31.21262680.

21. H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M. Allam, A. Ismail, T. Mohale, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, F. Petruccione, A. Sigal, D. Hardie, G. Marais, N.-Y. Hsiao, S. Korsman, M.-A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O. Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C. K. Wibmer, B. T. Sewell, J. Lourenço, L. C. J. Alcantara, S. L. Kosakovsky Pond, S. Weaver, D. Martin, R. J. Lessells, J. N. Bhiman, C. Williamson, T. de Oliveira, Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*. **592**, 438–443 (2021).

22. F. G. Naveca, V. Nascimento, V. C. de Souza, A. de L. Corado, F. Nascimento, G. Silva, Á. Costa, D. Duarte, K. Pessoa, M. Mejía, M. J. Brandão, M. Jesus, L. Gonçalves, C. F. da Costa, V. Sampaio, D. Barros, M. Silva, T. Mattos, G. Pontes, L. Abdalla, J. H. Santos, I. Arantes, F. Z. Dezordi, M. M. Siqueira, G. L. Wallau, P. C. Resende, E. Delatorre, T. Gräf, G. Bello, COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. *Nat. Med.* **27**, 1230–1238 (2021).

23. M. A. Benítez, C. Velasco, A. R. Sequeira, J. Henríquez, F. M. Menezes, F. Paolucci, Responses to COVID-19 in five Latin American countries. *Health Policy Technol*. **9**, 525–559 (2020).

24. Y. A. Adebisi, G. I. Oke, P. S. Ademola, I. G. Chinemelum, I. O. Ogunkola, D. E. Lucero-Prisno Iii, SARS-CoV-2 diagnostic testing in Africa: needs and challenges. *Pan Afr. Med. J.* **35**, 4 (2020).

25. S. J. Salyer, J. Maeda, S. Sembuche, Y. Kebede, A. Tshangela, M. Moussif, C. Ihekweazu, N. Mayet, E. Abate, A. O. Ouma, J. Nkengasong, The first and second waves of the COVID-19 pandemic in Africa: a cross-sectional study. *Lancet*. **397**, 1265–1275 (2021).

26. S. J. Becker, J. Taylor, J. M. Sharfstein, Identifying and tracking SARS-CoV-2 variants - A challenge and an opportunity. *N. Engl. J. Med.* **385**, 389–391 (2021).

27. A. Maxmen, Why US coronavirus tracking can't keep up with concerning variants. *Nature*. **592**, 336–337 (2021).

28. A. Maxmen, Why some researchers oppose unrestricted sharing of coronavirus data. *Nature* (2021), doi:10.1038/d41586-021-01194-6.

29. J. Wang, C. Kaperak, T. Sato, A. Sakuraba, COVID-19 reinfection: a rapid systematic review of case reports and case series. *J. Investig. Med.* **69**, 1253–1255 (2021).

30. T. Kustin, N. Harel, U. Finkel, S. Perchik, S. Harari, M. Tahor, I. Caspi, R. Levy, M. Leshchinsky, S. Ken Dror, G. Bergerzon, H. Gadban, F. Gadban, E. Eliassian, O. Shimron, L. Saleh, H. Ben-Zvi, E. Keren Taraday, D. Amichay, A. Ben-Dor, D. Sagas, M. Strauss, Y. Shemer Avni, A. Huppert, E. Kepten, R. D. Balicer, D. Netzer, S. Ben-Shachar, A. Stern, Evidence for increased breakthrough rates of SARS-CoV-2 variants of concern in BNT162b2-mRNA-vaccinated individuals. *Nat. Med.* (2021), doi:10.1038/s41591-021-01413-7.

31. C. Alteri, V. Cento, A. Piralla, V. Costabile, M. Tallarita, L. Colagrossi, S. Renica, F. Giardina, F. Novazzi, S. Gaiarsa, E. Matarazzo, M. Antonello, C. Vismara, R. Fumagalli, O. M. Epis, M. Puoti, C. F. Perno, F. Baldanti, Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. *Nat. Commun.* **12**, 434 (2021).

32. B. B. Oude Munnink, D. F. Nieuwenhuijse, M. Stein, Á. O'Toole, M. Haverkate, M. Mollers, S. K. Kamga, C. Schapendonk, M. Pronk, P. Lexmond, A. van der Linden, T. Bestebroer, I. Chestakova, R. J. Overmars, S. van Nieuwkoop, R. Molenkamp, A. A. van der Eijk, C. GeurtsvanKessel, H. Vennema, A. Meijer, A. Rambaut, J. van Dissel, R. S. Sikkema, A. Timen, M. Koopmans, Dutch-Covid-19 response team, Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat. Med.* **26**, 1405–1410 (2020).

33. J. L. Gardy, N. J. Loman, Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).

34. WHO, Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health (2021), (available at https://apps.who.int/iris/bitstream/handle/10665/338480/9789240018440-eng.pdf).

35. M. D. Van Kerkhove, M. J. Ryan, T. A. Ghebreyesus, Preparing for "Disease X." *Science* (2021), doi:10.1126/science.abm7796.

36. GISAID. *GISAID* (2021), (available at https://www.gisaid.org/).

37. UN, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019, Online Edition. Rev. 1. *United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019, Online Edition. Rev. 1.* (2019).

38. World Bank, World bank country and lending groups – world bank data help desk. *The World Bank* (2021), (available at https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups).

39. IHME, Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019 (GBD 2019) Covariates 1980-2019. *Institute for Health Metrics and Evaluation (IHME)* (2020), (available at ttp://ghdx.healthdata.org/record/global-burden-disease-study-2019-gbd-2019-covariates-1980-2019).

40. IHME, Institute for Health Metrics and Evaluation (IHME). Gross Domestic Product (GDP) Estimates by Country 1950-2015. *Institute for Health Metrics and Evaluation (IHME)* (2012), (available at http://ghdx.healthdata.org/record/ihme-data/gross-domestic-product-gdp-estimates-country-1950-2015.

41. UNESCO, UNESCO Institute for Statistics. *UNESCO* (2021), (available at http://data.uis.unesco.org/Index.aspx?DataSetCode=SCN_DS&lang=en).

42. A. E. Magurran, *Measuring Biological Diversity* (John Wiley & Sons, 2013).

43. T. Y. Michaelsen, M. Bennedbaek, L. E. Christiansen, M. S. F. Jorgensen, C. H. Moller, E. A. Sorensen, S. Knutsson, J. Brandt, T. B. N. Jensen, C. Chiche-Lapierre, Others, Introduction and transmission of SARS-CoV-2 B. 1.1. 7 in Denmark. *medRxiv* (2021) (available at https://www.medrxiv.org/content/10.1101/2021.06.04.21258333v1.abstract).

Acknowledgments: We gratefully acknowledge the authors from the Originating laboratories responsible for obtaining the specimens, as well as the Submitting laboratories where the genomic data were generated and shared via GISAID, on which this research is based. An acknowledgement table can be found in Table S8 and at gisaid.org with set accession EPI_SET_20211008ez. We thank James Nokes, Isabella Lynette Ochola, and Sylvie Briand for their valuable comments. GD acknowledges Joshua Batson, whose work shared on Twitter

(@thebasepoint) inspired the creation of Figure 2A. BHP and VS acknowledge the contribution of SARS-CoV-2 genomes by members of the Communicable Diseases Genomics Network of Australia.

Funding: ES and SF acknowledges the EPSRC (EP/V002910/1). GB acknowledges support from the Internal Fondsen KU Leuven/Internal Funds KU Leuven (Grant No. C14/18/094) and the Research Foundation - Flanders ("Fonds voor Wetenschappelijk Onderzoek - Vlaanderen," G0E1420N, G098321N). GWH acknowledges support from NIH F31 AI154824. MAS acknowledges support from grants NIH R01 AI153044 and NIH U19 AI135995. MUGK acknowledges funding from the Oxford Martin School, EUH2020 project MOOD, Branco Weiss Fellowship and grants from The Rockefeller Foundation and Google.org. NDG acknowledges support from Fast Grant from Emergent Ventures at the Mercatus Center at George Mason University and CDC Contract # 75D30120C09570. OGP acknowledges support from the Oxford Martin School. NRF acknowledges support by a Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z). NRF and ECS acknowledge support by a Medical Research Council-São Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0) (http://caddecentre.org/) and by Bill & Melinda Gates Foundation (INV-034540 and INV-034652). Rede Corona-ômica BR MCTI/FINEP is affiliated to RedeVírus/MCTI (awards FINEP = 01.20.0029.000462/20, CNPq = 404096/2020-4). CCK acknowledges support from the US Public Health Service Ruth L. Kirschstein National Research Service Award (5T35HL007649-35). RSA acknowledges funding from CNPq: 312688/2017-2 and 439119/2018-9; MEC/CAPES: 14/2020 - 23072.211119/2020-10; FINEP: 0494/20 01.20.0026.00 and UFMG-NB3 1139/20 and FAPERJ: 202.922/2018.

Author Contributions: Conception: AFB, NDG, NRF; Data acquisition: Danish Covid-19 Genome Consortium, COVID-19 Impact Project, Swiss SARS-CoV-2 Sequencing Consortium, Bulgarian SARS-CoV-2 sequencing group, Network for Genomic Surveillance in South Africa

(NGS-SA), GISAID core curation team, GG, CNA, RTPL, MMS, PCR, CVFC, NSDS, SMS; Analysis: AFB, ES, GD, GWH, CCK, JH, SMS, SB, SF, MAS, GB; Interpretation: AFB, ES, GD, GWH, MUGK, JH, HT, GG, CNA, TdO, RTPL, SMS, SCH, OGP, CD, SB, SF, NDG, GB, NRF; Drafting: AFB, ES, GD, CCK, GB, NRF; Revising: AFB, ES, GD, GWH, MUGK, JH, HT, GG, CNA, LEM, CW, BPH, VS, NSZ, OM, HMB, TdO, RTPL, MMS, PCR, ATRV, FRS, RSA, IA, INI, IP, CVFC, NSDS, CG, SMS, DN, MP, MvK, SCH, ECS, OGP, CD, MAS, NDG, GB, NRF; Funding: NDG, NRF.

Conflicts of Interests: NDG is an infectious diseases consultant for Tempus Labs and the National Basketball Association. MAS receives grants and contracts from the National Institutes of Health, the US Food & Drug Administration, the US Department of Veterans Affairs and Janssen Research & Development. OGP has undertaken work for AstraZeneca on SARS-CoV-2 classification and genetic lineage nomenclature.

Data and materials availability: All data and scripts used to generate figures and tables are available at https://github.com/andersonbrito/paper_2021_metasurveillance

Supplementary Materials

Materials and Methods
Table S1–S8
Fig. S1–S7

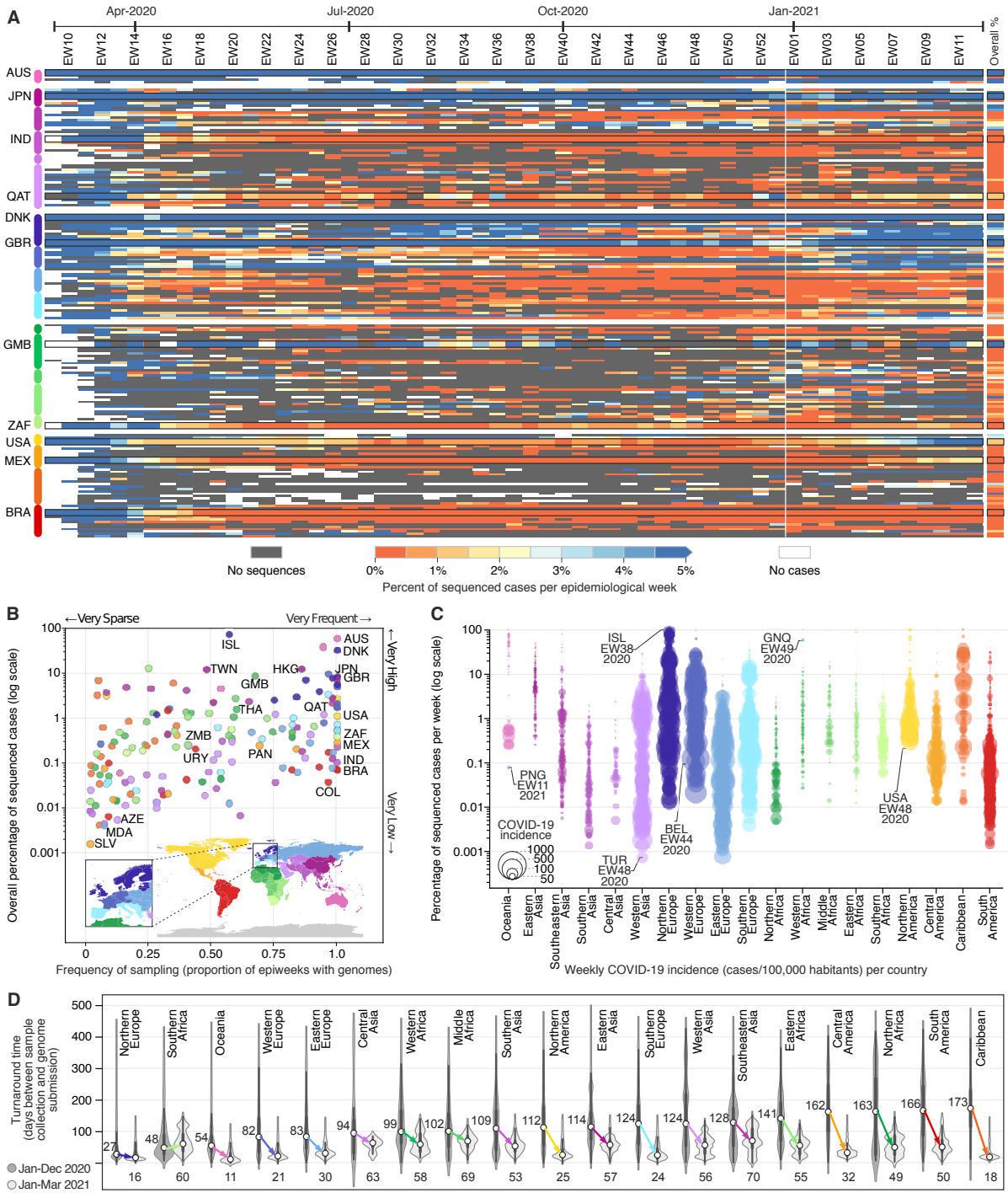

Fig. 1. Disparities in SARS-CoV-2 global genomic surveillance. **(A)** Percentage of reported cases that were sequenced per country, per epidemiological week (EW), between February 23rd, 2020 and March 27th, 2021 (based on metadata submitted to GISAID up to May 30th, 2021). Updated numbers on sequence submissions and proportion of sequenced cases are available on the GISAID Submissions Dashboard at gisaid.org. **(B)** Frequency and overall percentage of sequenced cases per country. This plot summarizes the data shown in (A), where the x-axis shows the percentage of EWs with sequenced cases, and the y-axis displays the overall percentage of cases shown in the rightmost column of panel (A). **(C)** Percentage of cases sequenced per EW per country, per geographic region (classified according to the UNSD geoscheme). Each circle represents an EW with at least one sequenced case, and their diameters highlight the incidence (cases per 100,000 habitants), e.g. "ISL-EW38-2020" shows data from week 38 in 2020, in Iceland. **(D)** Distribution of turnaround times of genomes collected in different geographic regions, in 2020 and 2021. Countries are highlighted in panels of this figure using the ISO 3166-1 nomenclature.

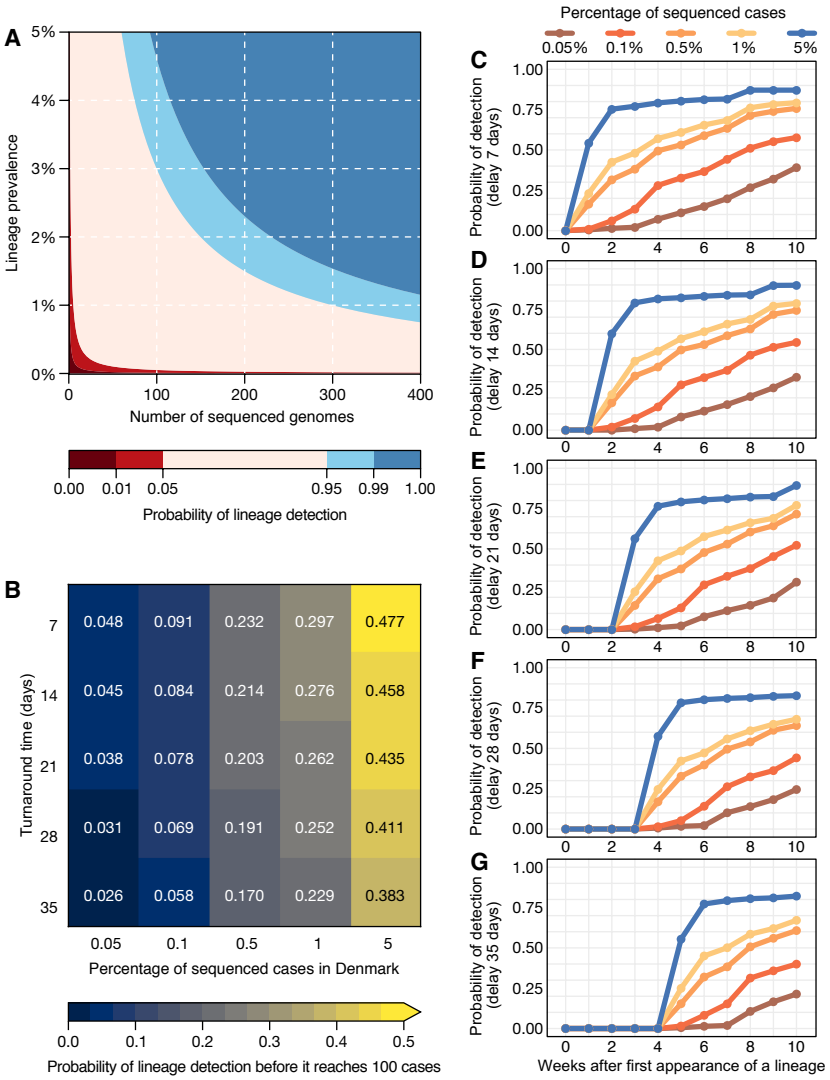

Figure 2. **Detection of SARS-CoV-2 lineages under different genomic surveillance scenarios.** (A) The probability of detecting at least one genome of a rare lineage under different sequencing regimes. (B) Relative importance of decreasing genome sequencing turnaround time (TAT) versus increasing sequencing percentage, measured as probability that a lineage found in simulated datasets was detected before it had reached 100 cases (described in **Fig. S6**). (C-G) Probability of lineage detection considering TATs of 7, 14, 21, 28 and 35 days.

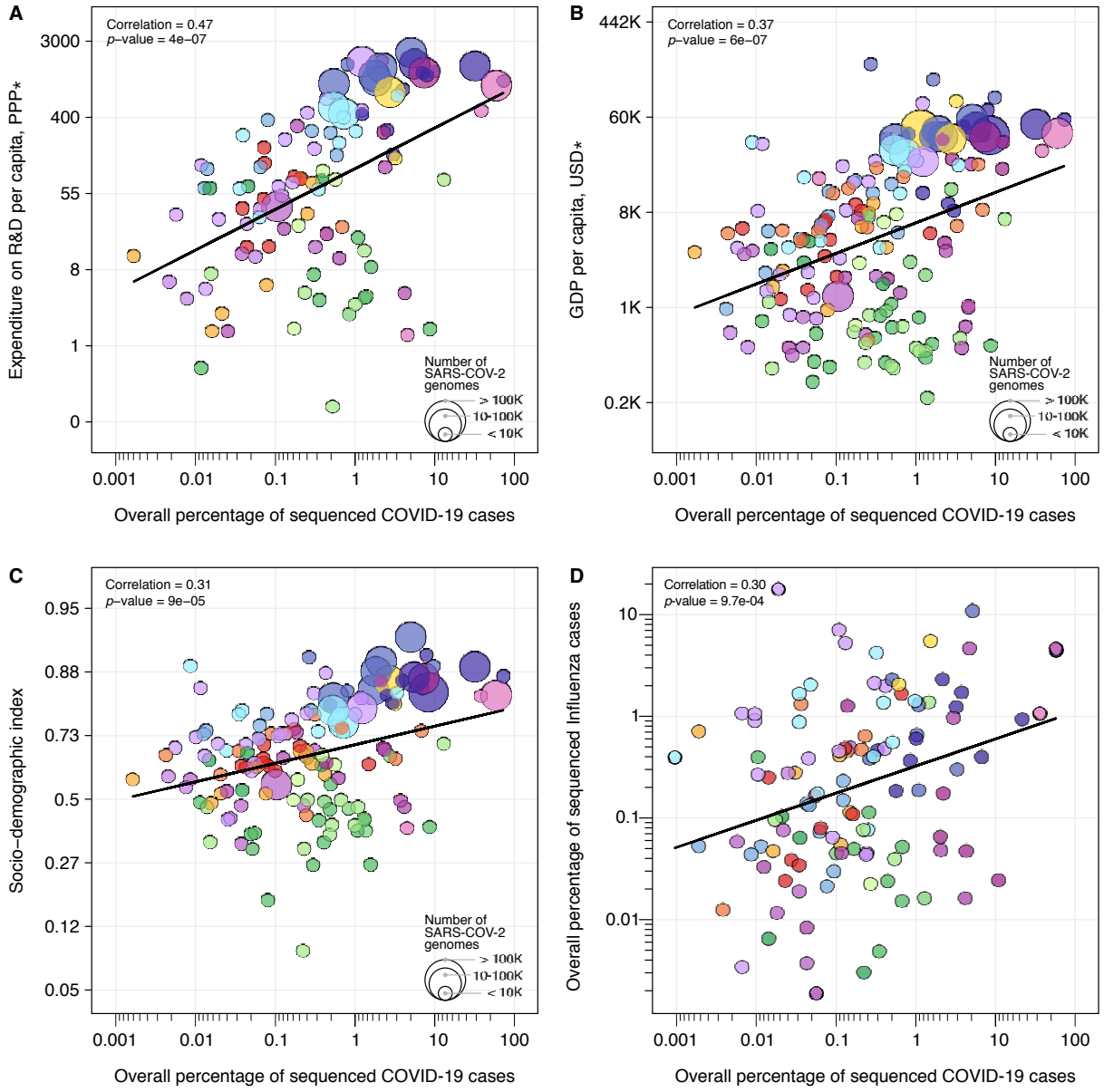

Figure 3. Case sequencing percentages and socioeconomic covariates. Covariates that show the highest correlation with the overall percentage of COVID-19 sequenced cases (during the period shown in Fig. 1A). (A) Expenditure on R&D per capita; (B) GDP per capita; (C) Socio-demographic index; (D) Overall percentage of influenza virus sequenced cases in 2019 (HA segment). For correlations between covariates and turnaround time, see **Fig. S7.** The colour scheme is the same as in Figure 1 and 2. Solid line shows the linear fit. *PPP = purchasing power parity, USD = US dollar 2005.

Table 1. Empirical country sequencing capacities at different income levels and lines of inquiry enabled at each level. Countries at each income level have markedly different sequencing capacities, allowing for different degrees of epidemic resolution and lines of inquiry. Characteristics of each income class are shown in **Table S4**.

Income class	Median weekly genomes (when sequencing at all)	Mean weekly genomes (when sequencing at all)	Probability of detecting a lineage at 5% prevalence under mean weekly sequencing regime	Maximum probable prevalence of an undetected lineage under mean weekly sequencing regime	Lines of inquiry available
Low income countries (LICs)	4	8.64	0.351	0.262	Presence/absence of prevalent lineages
Lower middle income countries (LMCs)	5	25.97	0.727	0.095	+ Quantification of lineage prevalence with some error; identification of preliminary patterns of geographic spread
Upper middle income countries (UMCs)	7	33.16	0.810	0.073	
High income countries (HICs)	38	524.80	1.000	0.005	+ Investigations of lineage dynamics, and transmissibility; high precision lineage tracking (molecular evolution and geographic spread)