Detection of SARS-CoV-2 variants requires urgent global coordination

Running Title: Genomic capacity and SARS-CoV-2 detection

Carlos M. Duarte*, Tahira Jamil, Takashi Gojobori and Intikhab Alam

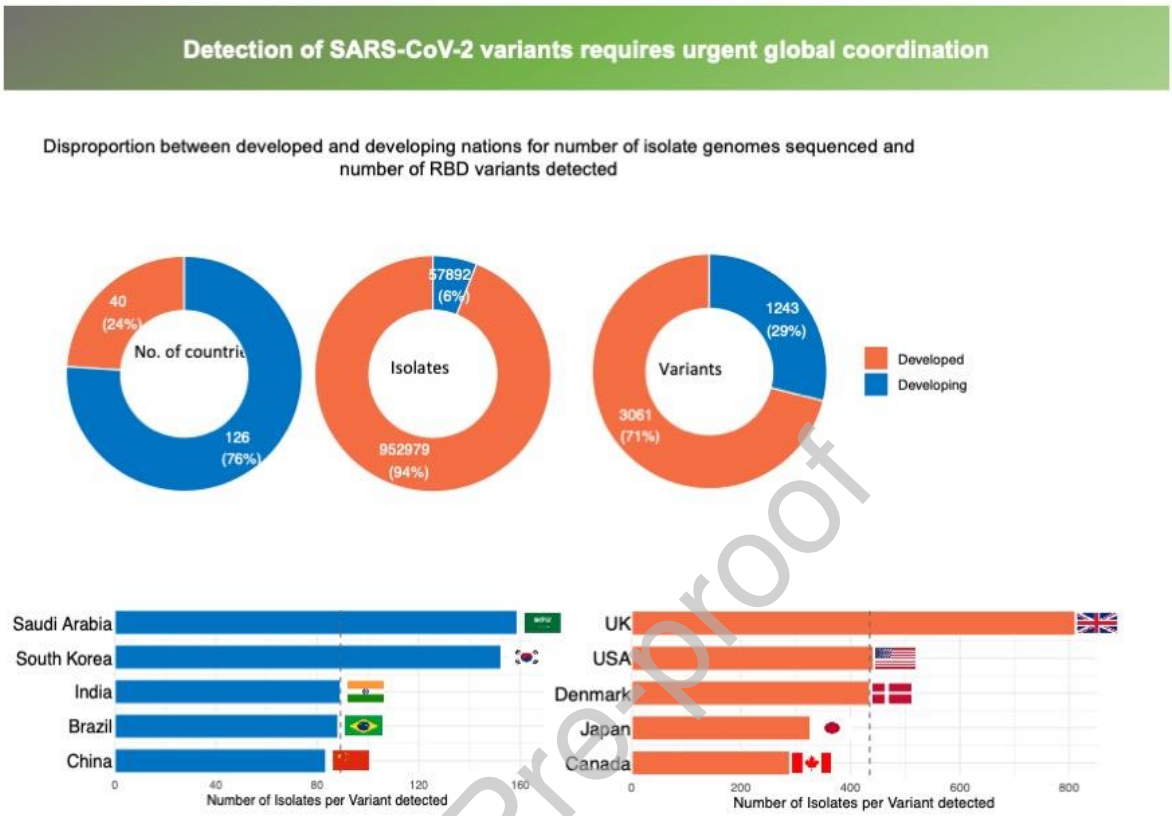
Computational Bioscience Research Centre (CBRC), King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

*Corresponding author, carlos.duarte@kaust.edu.sa, +966 542510757

Highlights

- Eight nations contributed 79% of all SARS-CoV-2 isolates sequenced.
- Two-thirds of SARS-CoV-2 variants found were reported by 5 nations.
- The number of variants detected increases at the square root of sequencing effort.
- International cooperation for SARS-CoV-2 sequencing is urgently needed.
- Effective monitoring and prevention requires a global pathogen sequencing mechanism.

Graphical Abstract



Abstract

Objectives: We assessed the effort deployed by different nations and territories to sequence SARS-CoV-2 isolates, thus allowing the detection of variants, known and novel, of concern.

Design: We analyzed the sources of over one million full genome sequences of SARS-CoV-2 virus available in the COVID-19 virus Mutation Tracker (CovMT) to determine the number of variants in the RBD region of the genome determining infectivity detected in the various nations and territories.

Results: The number of variants detected increased as the square root of sequencing effort of sequencing effort by nations. Eight nations contributed 79% of all SARS-CoV-2 isolates sequenced, with 2/3's of all unique variants, adding to 1118 RBD variants, reported by 5 nations.

The median number of isolates sequenced required to detect, on average, one novel RBD variant is 24.05, a threshold only achieved by 70 nations.

Conclusions: Many developing nations have not contributed any sequences due to lack of capacity, with a risk of dangerous virus variants in these undersampled regions spreading globally before being detected. A collaborative program to sequence SARS-CoV-2 isolates, and other pathogens of concern, is needed to monitor, track and control the pandemic.

Keywords

SARS-CoV-2, variants, mutations, sequencing, capacity, detection

The rapid diversification of SARS-CoV-2 variants, with the total number of variants (i.e., unique sequences across the entire genome) reaching 539,933 as of Apr 11, 2021 (COVID-19 virus Mutation Tracker (CovMT), https://www.cbrc.kaust.edu.sa/covmt, Alam et al. 2021) based on genomic data from world leading SARS-CoV-2 repository gisaid.org) is raising concern as some of these variants, particularly the E484K mutation, may overcome immune defenses produced by previously infected or vaccinated people (Tada et al. 2021).

Novel SARS-CoV-2 variants are initially diluted in the population of infected persons, thereby leading to low detection power, allowing the more infective variants to reach high abundances and possibly percolate across national borders before being detected. Detecting new SARS-CoV-2 variants requires whole genome sequencing, with the total number of good quality reported genome sequences of isolates, with >90% base coverage, reaching 1,010,872 by Apr 11, 2021.

Detecting variants is particularly important in the RBD region, as this is believed to determine SARS-CoV-2 infectivity (Greaney et al. 2021).

Most of the reported RBD variants were first detected in the UK, USA, Denmark, Germany, Maxico, Switzerland and South Africa, accounting for 64% of the total of 1118 RBD variants reported to date, including the so-called UK (N501Y, B.1.1.7), S. African (K417N+E484K+N501Y, B.1.351) and Brazilian (E484K+N501Y, P.1 & E484K, P.2) variants of concern (Sabino et al. 2021). One of the most concerning RBD mutations is E484K, also reported to be acquired by the UK variant, involved in evasion of antibodies, rendering existing vaccines less effective (Greaney et al. 2021, Sabino et al. 2021). A total of 87 nations, including most developing nations and territories, particularly in Africa and Island states, have not yet reported any RBD variant (Fig. 1a) because no isolate sampled in those nations and territories has been sequenced. Indeed, sequencing effort is highly skewed (Furuse 2021), with the top eight nations contributing to this effort (UK, USA, Denmark, Japan, Australia, Canada, Switzerland, and the Netherlands, in order of contribution), with more than 10,000 reported isolate genomes sequenced each, having reported 82% of all genome sequences globally (Fig. 1b).

The number of RBD variants detected in isolates sequenced in any one nation increases as the square root of sequencing effort, with > 10,000 sequenced isolates required to considerably reduce the rate of discovery with further sequencing effort (Fig. 1c). The median number of isolates sequenced required to detect, on average, one novel RBD variant is 24.05, a threshold only achieved by 70 nations (Fig. 1c). Hence, many nations with large numbers of infected cases fall, particularly developing nations but also some developed nations (Fig. 1c), have reached very

short of the sequencing effort required to detect new RBD variants that may be present in the population.

SARS-CoV-2 sequence data derive mostly from clinical diagnostic samples, particularly from infected individuals with high viral loads, that provide enough RNA for the sequencing of nearly complete genomes (Chiara et al. 2021). Handling the samples for RNA extraction requires biosafety level (BSL) 2 laboratories, and extracted RNA should stored at -80 ^{o}C to avoid degradation until sequenced, using various next generation sequencing strategies, mostly using Illumina sequencing, with specific library protocols available (Chiara et al. 2021). The sequences retrieved need then be assembled to construct a high-quality full viral genome. Chiara et al. (2021) provide a useful overview of the different procedures involved. The resulting viral genomes should then be made available by depositing it in the GISAID EpiCov portal (Shu and McCauley 2017), the most widely used repository of SARS-CoV-2 genomic data. However, while relatively standard, these technologies and the infrastructure and skill sets required may not be available everywhere, particularly in developing nations.

Next-generation whole-genome sequencing of SARS-CoV-2 virus isolates is essential to trace the spread and transmission chains of outbreaks, as well as for monitoring its evolution and diversification (Chiara el al. 2021). Hence, the COVID-19 pandemic has led to an unprecedented effort of full-genome sequencing in an effort to detect new variants and deploy defense strategies, such as mobility limitations and quarantines, as exemplified efforts across nations to limit the spread of the variants referred as UK (B.1.1.7), S. African (501Y.V2) and Brazil (P1) variants, some of which are more infective and/or may scape immune defenses (Kuzmina et al. 2021, Li et al. 2021).

Delivering the full power of whole-genome sequence to detect, assess and manage risks associated with the evolution of new SARS-CoV-2 variants requires a global effort, particularly focused on the areas experiencing high number of infections, as these are the areas where new variants are more likely to be originated. The high inequality in the capacity for advanced genomic sequencing is a manifestation of the growing gap in R&D capacity in health science between developing and developed nations. In a pandemic situation, the risk of failing to detect potentially-dangerous variants until they spread and become prominent is shared by all nations, so massive national efforts to sequence SARS-CoV-2 variants isolated from the national population may not be effective if dangerous variants are not detected where they originate, allowing their spread. Hence, a mechanism for nations with a demonstrated high capacity, such as the 10 top nations in sequencing effort (Fig. 1), to assist with sequencing of samples collected in developing nations lacking the capacity is absolutely required. Recently WHO provided a guideline to improve sequencing efforts (https://apps.who.int/iris/rest/bitstreams/1326052/retrieve) and suggested that samples for sequencing be sent to an established international sequencing laboratory in a third country where sequencing capacity maybe lacking (https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/).

Creating a global, coherent, and collaborative program for pathogen sequencing is not just required to respond to the COVID-19 pandemic, but a permanent mechanism is required to maintain an effective monitoring and prevention system in the future. In a world with unprecedented connectivity, global collaboration and generosity in sharing genomic sequencing capacity and data is not just an act of generosity; it is also an act of self-interest for every nation.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Source

This research was funded by King Abdullah University of Science and Technology through funding provided to CMD and TG.

Ethical Approval

This research did not require ethical approval.

References

Alam I, Radovanovic A, Incitti R, Kamau AA, Alarwai M, Azahar EI, Gojobori T, et al. CovMT: an interactive SARS-CoV-2 mutation tracker, with a focus on critical variants. The Lancet Infectious Diseases 2021; DOI:https://doi.org/10.1016/S1473-3099(21)00078-5

Chiara M, D'Erchia CM, Gissi AM, Manzari C, Parisi C, Resta A, Zambelli N, Picardi F, Pavesi E, Horner G, Pesole G. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Briefings in Bioinformatics 2021; 22: 616-630.

Furuse Y. Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. Int J Infect Dis 2021;103: 305-307.

Greaney AJ, Starr TN, Gilchuk P, et al. Complete mapping of mutations to the SARSCoV-2 spike receptor-binding domain that escape antibody recognition. Cell host & microbe 2021; 29: 44-57. e9.

Kuzmina A, Khalaila Y, Voloshin O, Keren-Naus A, Bohehm L, Raviv Y, Shemer-Avni Y, Rosenberg E, Taube R. SARS CoV-2 spike variants exhibit differential infectivity and

neutralization resistance to convalescent or post-vaccination sera. Cell Host & Microbiome 2021; 29: 522-528.e2

Li Q, Nie J, Wu J, Zhang L, Ding R, Wang H, Zhang Y, Li T, Liu S, Zhang M, Zhao C. SARS-CoV-2 501Y. V2 variants lack higher infectivity but do have immune escape. Cell 2021; https://doi.org/10.1016/j.cell.2021.02.042

Sabino EC, Buss LF, Carvalho MP, et al. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. The Lancet 2021; 397: 452-455.

Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data—from vision to reality. Eurosurveillance 2017;22:30494.

Tada T, Dcosta BM, Samanovic-Golden M, et al. Neutralization of viruses with European, South African, and United States SARS-CoV-2 variant spike proteins by convalescent sera and BNT162b2 mRNA vaccine-elicited antibodies. bioRxiv 2021.

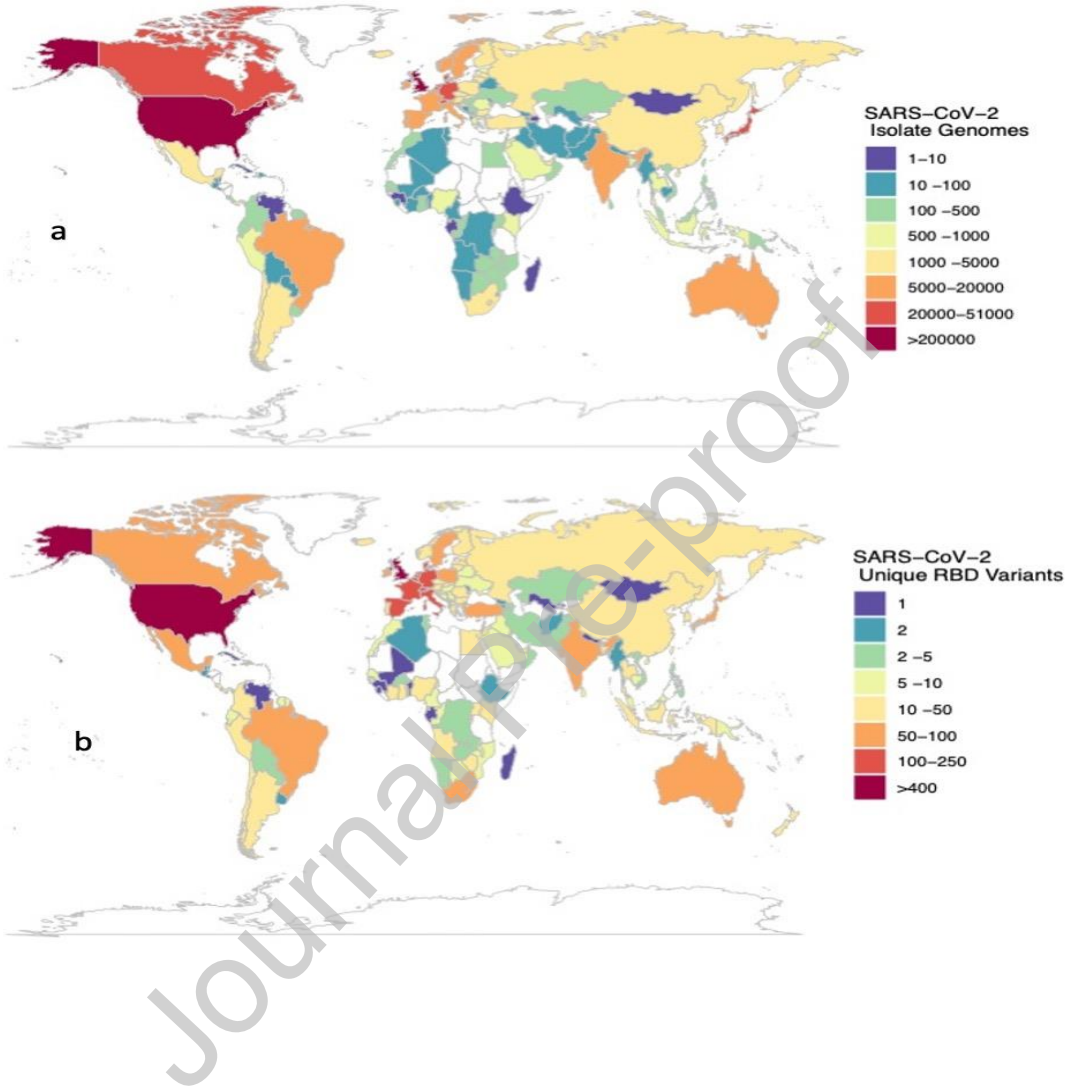


Figure 1. (a) The number of sequenced isolate genomes sequenced and (b) the number of RBD variants detected across nations as reported by March 01, 2021. Data obtained from world variants table available at COVID-19 virus Mutation Tracker (CovMT) (1), where RBD variants are operationally defined as each set of SARS-CoV-2 genomes that generate the exact same amino acid sequence in the RBD region of the spike protein.

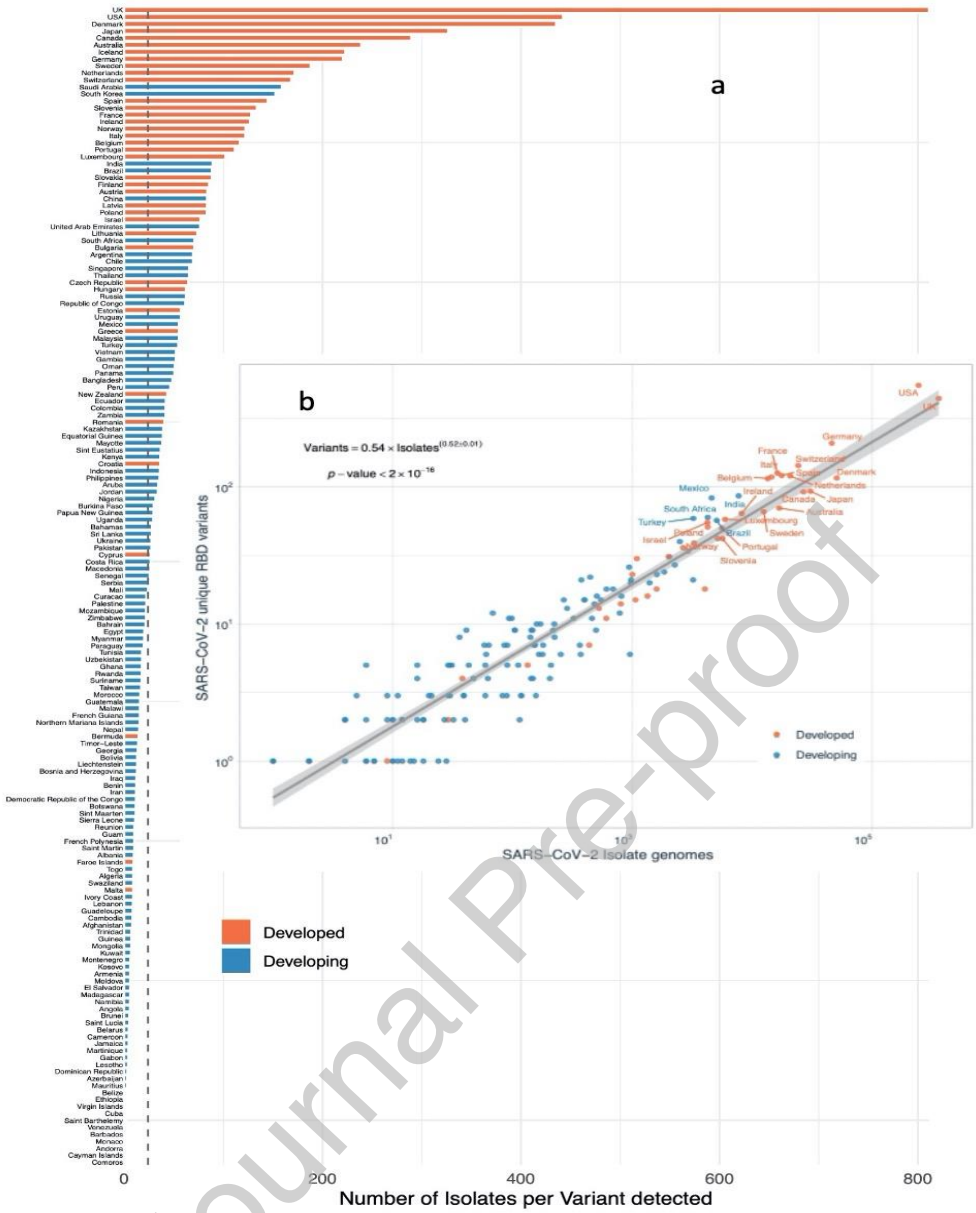


Figure 2. (a) The number of sequenced isolate genomes sequenced per RBD variants detected in developed and developing nations. The vertical dotted line shows the median number of isolates sequenced to detect the new RBD variant (24.05). (b) The relationship between the number of RBD variants detected and the number of sequenced isolate genomes sequenced in developed and developing nations. The solid line shows the fitted power law (equation in the insert), and the corresponding regression equation.