Unlocking capacities of genomics for the COVID-19 response and future pandemics

During the COVID-19 pandemic, genomics and bioinformatics have emerged as essential public health tools. The genomic data acquired using these methods have supported the global health response, facilitated the development of testing methods and allowed the timely tracking of novel SARS-CoV-2 variants. Yet the virtually unlimited potential for rapid generation and analysis of genomic data is also coupled with unique technical, scientific and organizational challenges. Here, we discuss the application of genomic and computational methods for efficient data-driven COVID-19 response, the advantages of the democratization of viral sequencing around the world and the challenges associated with viral genome data collection and processing.

Sergey Knyazev, Karishma Chhugani, Varuni Sarwal, Ram Ayyala, Harman Singh, Smruthi Karthikeyan, Dhrithi Deshpande, Pelin Icer Baykal, Zoia Comarova, Angela Lu, Yuri Porozov, Tetyana I. Vasylyeva, Joel O. Wertheim, Braden T. Tierney, Charles Y. Chiu, Ren Sun, Aiping Wu, Malak S. Abedalthagafi, Victoria M. Pak, Shivashankar H. Nagaraj, Adam L. Smith, Pavel Skums, Bogdan Pasaniuc, Andrey Komissarov, Christopher E. Mason, Eric Bortz, Philippe Lemey, Fyodor Kondrashov, Niko Beerenwinkel, Tommy Tsan-Yuk Lam, Nicholas C. Wu, Alex Zelikovsky, Rob Knight, Keith A. Crandall and Serghei Mangul

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a highly contagious pathogen that caused the COVID-19 pandemic, which reached an unprecedented scale of infection not seen since the influenza pandemic of 1918–1919. Within a month of its first reported case in Wuhan, China, in December 2019, the virus had spread to many regions within China as well as in several neighboring countries, including Thailand, Korea and Japan. As international flights continued to operate, SARS-CoV-2 rapidly spread to Europe and North America[1].

During this time, it became clear that the genomic toolkits are essential for public health decision-making, including testing for COVID-19, monitoring for emergence of new virus variants with altered biological or immunological properties, identification of at-risk individuals and informing of epidemiological models that describe outbreaks in communities[2]. This has allowed the observation of SARS-CoV-2 genome evolution in almost real time and the rapid tracking of SARS-CoV-2 genetic lineages and variants of interest and concern (VOIs, VOCs), which in turn have facilitated the development of clinical tests for SARS-CoV-2 and the prediction of vaccine efficacy against viral variants[3,4]. However, to reach the full potential of genomic data for future public health surveillance and outbreak response, we believe it is necessary to expand and coordinate best practices in genomics and bioinformatics that have now been field tested during the COVID-19 response[5]. Herein, we discuss the genomic techniques and corresponding bioinformatics algorithms that are addressing many of the pressing public health issues associated with COVID-19.

Genomics-based methods enabled early warnings of COVID-19 pandemic

As a local team of health professionals was investigating a small local outbreak of pneumonia consisting of the first 59 suspected cases from Wuhan in December 2019, they quickly discovered that they were dealing with a novel virus of unknown origin[6]. This rapid discovery was made possible by modern robust and accurate genomic and bioinformatic tools that, although now used routinely, did not exist a couple of decades ago. By 30 January 2020, when the World Health Organization (WHO) declared a Public Health Emergency of International Concern (PHEIC), 339 SARS-CoV-2 genomes had already been sequenced and characterized[1].

To investigate the newly emerging outbreak, scientists in China performed whole-genome sequencing of specimens, followed by de novo assembly and end-mapping to annotate the complete 29,903-nucleotide-long SARS-CoV-2 genome. Bioinformatics analysis revealed that the genome organization of SARS-CoV-2 was consistent with a single-stranded, positive-sense RNA virus from the genus Betacoronavirus[7]. Additionally, sequence alignment tools including BLAST[8] were used to search for related species of the newly discovered virus in the NCBI GenBank database, revealing alarming similarities to SARS-CoV (SARS-CoV-1), as well as a much higher similarity with Betacoronavirus from bats, suggestive of a zoonotic origin for the virus. Some SARS-CoV-2 genome fragments, in addition, have highest similarity to the corresponding fragments from pangolins, which suggests that recombination events between strains may have occurred during the virus' evolution. Subsequent analyses that included additional sarbecovirus genomes from bats and pangolins further scrutinized the evolution and recombination history of these viruses, finding that the lineage that gave rise to SARS-CoV-2 had been probably circulating unnoticed in bats for decades[9,10].

Genomics-based methods shaped the effective COVID-19 response. Once the SARS-CoV-2 genome was sequenced, the authors immediately publicly deposited the genome in GenBank[7,11]. This timely open-access release of the virus genome sequence was a laudable decision that allowed informed scientific analyses and pandemic preparation to begin immediately.

As the pandemic progressed, the increased availability of modern sequencing technologies prompted the collection of SARS-CoV-2 viral genomic data on an

unprecedented scale. Within a month, on average about 1,300 genomes were being submitted per day. Within six months of start of the pandemic (by May 2020), GISAID had 110,000 full-length SARS-CoV-2 genome sequences. By December 2021, two years into the pandemic, 67,000 genomes per day were being deposited into public viral genome data repositories such as GISAID, COG-UK and GenBank, which currently contain over 6 million SARS-CoV-2 genomes[12–14] (Fig. 1a and Supplementary Table 1). The unprecedented volume of data collection for SARS-CoV-2 is evident by contrast with HIV genomic data collection: for HIV, which has consistently held the attention of public health officials and the general public since the 1980s, fewer than 16,000 full-length genome sequences have been collected by the biggest public HIV sequence database, at the Los Alamos National Laboratory in the United States, over the past 40 years[15] (Fig. 1a).

SARS-CoV-2 sequencing data collected all over the world and rapidly shared in online databases ultimately aided public health officials and governments in making better-informed decisions[16]. However, to fully explore the potential of such databases, a few issues still need to be resolved. Despite the unprecedented pace overall, inevitable delays caused by shortage of sequencing capacity, and in some regions political interference, led to problems in the logistical chain in these regions, including in sample collection, transporting and shipping samples[17]. Depending on the country and the strength of its public health infrastructure, the median time lag from collection to submission can differ greatly, ranging from one day to one year. Several factors influence the rate and scale of viral genomic sequencing across the globe. Countries with minimal sequencing capacity are likely to encounter outbreaks of higher severity, leading to blind spots of genomic surveillance that can facilitate the spread of new variants to other countries[17]. On average, high-income countries shared about 100 times more sequences per capita than low-income countries (Fig. 1b and Supplementary Fig. 2). However, some African countries with a low GDP per capita were able to sequence a comparable number of viral genomes to middle- and high-income countries[18]. This preparedness can be attributed to previous global initiatives to support African countries in mitigating outbreaks of other viruses that have enhanced sequencing capacities in the region. Africa provides a remarkable example of the necessity of international cooperation and of approaches that could be implemented in other parts of the world to improve pandemic response globally (Fig. 1c). In general, however, the number of shared coronavirus genomes per capita is correlated with the country's GDP per capita (Fig. 1d).

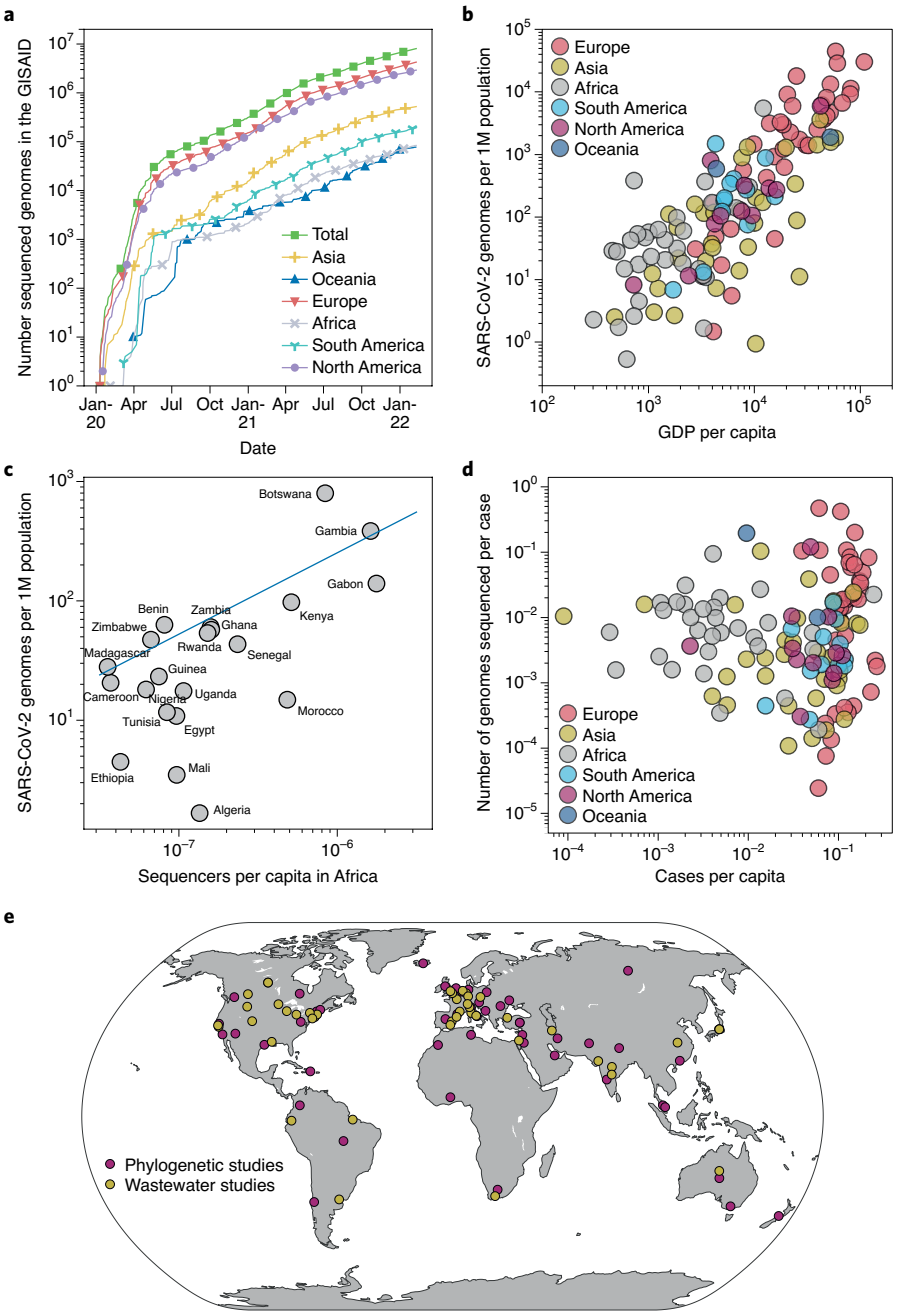

Fig. 1 | Available SARS-CoV-2 genomic sequencing data and its usage for outbreak investigation. a, The number of SARS-CoV-2 genomes sequenced in different regions according to Global Initiative On Sharing All Influenza Data (GISAID) between January 2020 and December 2021. **b**, The number of available SARS-CoV-2 sequences in GISAID per 1 million (1M) individuals vs. the number of cases per capita for each country or region up to March 2021. **c**, The number of available SARS-CoV-2 sequences in GISAID per 1M individuals vs. the number of sequencers per capita for each country in Africa up to March 2021. Blue line is the correlation of all data points on the plot. **d**, The number of available SARS-CoV-2 sequences in GISAID per number of reported COVID-19 cases vs. the number of reported COVID-19 cases per capita for each country or region from December 2019 up to December 2021. **e**, Global outbreak investigations by phylogenetic analysis (red) and wastewater studies (yellow); dots are placed in the geographical centers of each country or region.

Moving forward, several important data-sharing issues need to be addressed to facilitate open and rapid sharing of viral genome data. For example, it is important that scientists depositing sequencing data be able to trust that their rights will be respected by data users and that their authorship rights will not be violated[19]. The GISAID data access mechanism proved its ability to address these concerns and overcome obstacles to the international sharing of virus data, making GISAID the largest repository of influenza and SARS-CoV-2 genomic data[16,20].

Bioinformatics methods cab accurately track SARS-CoV-2 genomic evolution. As SARS-CoV-2 spread through the world population over the first year of the pandemic, it gradually evolved into several viral lineages[21–24]. Statistical analysis of collected SARS-CoV-2 genomes showed that SARS-CoV-2 has a mutation rate of at least tenfold lower than that of seasonal influenza[25]. This lower mutation rate initially gave hope for efficient control of the pandemic through vaccination because the slower a virus mutates, the less chances it has to adapt to vaccines. However, given the large number of COVID-19 cases (>277 million and climbing, according to the WHO) and possibly because of SARS-CoV-2 recombination events, new variants continue to evolve, which are currently being classified as variants under investigation (VUIs), of interest (VOIs) and of concern (VOCs) according to their epidemiological, biological and/or immunological properties. Indeed, some variants acquired numerous mutations in a rapid fashion (variants Alpha and Omicron) and/or showed evidence of immune escape (variant Omicron). Notably, it was observed that immunodeficient individuals who experience unusually long periods of SARS-CoV-2 infection can provide a plausible environment for faster SARS-CoV-2 evolution because their immune systems allows viral immune escape[26].

Before the COVID-19 pandemic, the public health community had had experience tracking and responding to genome evolution of viruses such as the influenza viruses that cause season flu. The Global Influenza Surveillance and Response System (GISRS) was established by the WHO for timely collection and genetic and antigenic characterization of these viruses[27]. Sharing of virus sequence data in the GISAID database along with the Nextstrain[28] online phylogenetic tool are used for biannual selection of influenza A and B vaccine seed strains and to help understand viral genomic evolution and antigenic drift. GISAID and Nextstrain were both promptly adopted for collecting and analyzing SARS-CoV-2 genomic data, becoming the largest global system for tracking SARS-CoV-2 evolution and monitoring new variants.

The widespread application of sequencing technologies became possible because of extensive efforts by the scientific community to benchmark and standardize sequencing protocols and open-source bioinformatics workflows for accurate consensus genome assembly[29]. However, the use of proprietary next-generation sequencing solutions and software has been more commonplace in well-resourced national and state/province-level public health labs. The accessibility of tiled primer sequences (such as ARCTIC or midnight primer sets) and lower costs of Illumina and Oxford Nanopore sequencing, along with open-access bioinformatics workflows, supported sequencing in dozens of regional public health labs and academic institutions across the world. By 24 December 2021, 80.49% of available SARS-CoV-2 genomic data at GISAID had been generated by Illumina sequencers, 12.46% by Oxford Nanopore, 3.85% by Pacbio, 1.59% by IonTorrent, 1.29% by BGI, 0.31% by Sanger and 0.02% by Qiagen (Supplementary Fig. 1a). NCBI GenBank contains 91.04% genomic data sequenced by Illumina, 8.1% by Oxford Nanopore, 0.47% by IonTorrent, < 0.01% by PacBio and 0.38% unspecified (Supplementary Fig. 1b).

This democratization of viral sequencing methods has helped build pathogen sequencing capacity in low- to middle-income countries and has fostered insights into the genomic epidemiology of SARS-CoV-2, including the emergence and spread of variants, for example in Colombia (VOI Mu), Ukraine (VOC Delta), the Philippines (VOC Alpha), the UK (VOC Alpha, as it moved to the United States) and South Africa, where immune-evasive VOC Omicron was identified by genome sequencing[30–33].

Bioinformatics methods enable tracking COVID-19 geographical spread in real time. As viruses evolve, tracking the appearance of new mutations and the locations where they were introduced can reveal geographical transmission routes. These routes help distinguish imported cases from those due to community transmission, aiding the identification of high-risk transmission routes that can be subject to enhanced public health control[34]. Comparative genomic analyses to study COVID-19 outbreak transmission dynamics have mostly been conducted using classic maximum-likelihood (ML) phylogenetic methods[35]. Unfortunately, ML methods are not scalable enough to handle the large volumes of SARS-CoV-2 genomic data available. For ML, therefore, it is often necessary to reduce sample size and consider only a fraction of the data in order to conduct the analysis, which can potentially compromise the accuracy of the results. Alternatively, more scalable approximate maximum-parsimony methods (MP) can be used for phylogeny reconstruction from dense SARS-CoV-2 data[36]. Indeed, it has been shown theoretically that with dense enough sampling, MP produces an ML tree under certain ML models[37–39]. Another approach has been to use network-based methods, which are significantly faster but theoretically less accurate than phylogeny-based methods[40–42].

The public availability of diverse SARS-CoV-2 genome sequences from around the world has facilitated the efficient and accurate tracking of local and global SARS-CoV-2 transmission routes[43–45] (Supplementary Fig. 3). Phylogenetics methods (Supplementary Table 2) revealed that SARS-CoV-2 was introduced into Europe from China and into the United States from China and Europe[34,46–48] and have also been used to track domestic transmission chains and differentiate them from international ones. In the United States, for example, studies showed that SARS-CoV-2 was likely introduced into Connecticut via a domestic transmission route, and the most successful viral introductions in Arizona were also likely via domestic travel[34,49]. The New York City area experienced multiple introductions of SARS-CoV-2, primarily from Europe[50]. Similarly, phylogenetic analysis suggested that SARS-CoV-2 was likely introduced into France from several countries, including China, Italy, the United Arab Emirates, Egypt and Madagascar[51] (Fig. 1e and Supplementary Table 2).

Differences in sampling across geographical locations and over time represent a considerable challenge to the accurate reconstruction of spatial transmission patterns. However, additional data, such as travel information and epidemiological estimates, may help mitigate difficulties due to non-uniform sampling across geographical locations and time and may contribute to a more complete picture of viral spread. This has been illustrated by a study of SARS-CoV-2 importation and establishment in the UK[52]. Large-scale genomic data resulted in estimates of the number and timing of introductions events, but combining these data with epidemiological and

travel data made it possible to identify the spatiotemporal origins of these introductions. Such additional data sources are also increasingly being integrated into phylodynamic inferences. For example, a study of the contribution of persistence versus new introductions to the second COVID-19 wave in Europe made use of Google mobility data to inform the phylogeographic component of the genomic reconstruction[53]. The individual travel history of sampled individuals can also be formally incorporated into such analyses[54].

Additionally, phylogenetics can be used to monitor the effectiveness of global travel restrictions and lockdowns. For example, it was shown that the risk of domestic transmission of SARS-CoV-2 in Connecticut already exceeded that of international introduction at the time federal travel restrictions were imposed, highlighting the critical need for local surveillance[34]. Similarly, in Brazil, three clades of European origin were established before the initiation of travel bans and lockdowns[55]. In the UK, lineages introduced before national lockdown were shown to be larger and more dispersed, and lineage importation and regional lineage diversity declined after lockdown[52]. Phylogenetics showed that several international introductions of SARS-CoV-2 likely occurred in Morocco as a result of violations of imposed lockdowns involving sea trade[56]. In Australia, lockdown effectiveness was validated using SARS-CoV-2 genomic data coupled with agent-based modeling, a computation tool to simulate the interactions of autonomous agents such as individuals[57]. Phylogenetic modeling of over 11,000 SARS-CoV-2 genomes collected in Switzerland throughout 2020 enabled estimation of the effects of different public health measures, including lockdown, border closure and test–trace–isolate efforts[58]. Similarly, comparative phylodynamics analysis of SARS-CoV-2 transmission dynamics in the neighboring Eastern European countries of Belarus and Ukraine, which followed highly different COVID-19 containment policies, allowed an assessment of the effectiveness of public health intervention measures in this region, and highlighted the roles of regional political and social factors in virus spread[59].

Genomics methods enable wastewater-based monitoring of SARS-CoV-2 epidemiology. The presence of trace viral genomic material in wastewater has been successfully exploited to track antibiotic use[60] and tobacco consumption[61] and for the monitoring of several respiratory and enteric viruses, including poliovirus[62]. Although COVID-19 is primarily associated with respiratory symptoms, SARS-CoV-2 is regularly shed in the feces of infected individuals[63]. As of December 2021, wastewater-based surveillance to track SARS-CoV-2 viral infection dynamics[64] had been implemented in many countries around the world (Fig. 1e).

Wastewater-based epidemiology has been shown to provide more balanced estimates of viral prevalence rates in a population than clinical testing alone due to inherent limitations in testing resources and/or testing uptake rates, especially in underserved communities. Combining clinical diagnostics with wastewater-based surveillance can provide a more comprehensive community-level profile of both symptomatic and asymptomatic cases, enabling identification of hospital capacity needs[65–72]. Another important advantage of wastewater monitoring is the ability to detect early-stage outbreaks before they become widespread[62,73–76]. Although tracking of SARS-CoV-2 viral RNA via quantitative PCR (qPCR)-based methods can reveal temporal changes of virus prevalence in a given population, it cannot provide underlying epidemiological information to identify transmission or genomic details of emerging variants. Tracking viral genomic sequences from wastewater significantly improves community prevalence estimates and also provides detection of emerging variants. Tracking SARS-CoV-2 viral genomic sequences from wastewater using a targeted tiled amplicon-based sequencing approach would significantly ameliorate community prevalence estimates and also detect emerging variants[77].

Wastewater genomic epidemiology can also act as a surrogate for elucidating strain geospatial distributions, helping identify outbreak clusters and track prevailing and newly emerging variants, and covering even areas with insufficient clinical testing rates. However, the highly variable nature of wastewater, low viral loads, fragmented RNA and the presence of multiple genotypes in a single sample makes it challenging to obtain good-quality genome sequences and discern lineages with a high degree of accuracy[78].

The commonly used tools used for discerning viral lineages in clinical samples, such as pangolin[3] and UShER[79], cannot deconvolute the multiple lineages that are commonly observed in a single wastewater sample and at best detect the most dominant one. As existing lineage-calling methods require a single consensus sequence to perform assignment, they are ill-equipped to capture the diversity present in mixed viral samples. Hence, tools to robustly identify the multiple lineages and their relative proportions present in wastewater are critical in understanding and interpreting the underlying sequence data obtained from these samples. For example, a depth-weighted demixing algorithm, Freyja[80], employs a 'barcode' library of lineage-defining mutations to represent each viral variant and can be used to recover relative abundances of different lineages within samples. This approach enabled the early detection of emerging VOCs in wastewater up to 14 days before their first clinical detection and also identified multiple instances of cryptic transmission not observed via clinical genomic surveillance[81]. Similar algorithms for mutation calling, haplotype reconstruction and population characterization in viral specimens can also be used to deconvolute the mixture of variants present in a wastewater sample[82,83]. By searching for signature mutations co-occurring on the same amplicon, variant B.1.1.7 was detected in wastewater eight days before the first patient sample tested positive for the variant[84]. Similarly, RNA transcript quantification methods, such as Kallisto, can be used to estimate the relative abundance of SARS-CoV-2 variants in wastewater[85]. Both digital PCR-based and sequencing-based estimates of variant abundance in wastewater have been used to derive the fitness advantage of a recently introduced variant, an important epidemiological parameter for assessing the expected transmissibility and spread of such a variant[84,86].

Alternatively, viral genomes in wastewater can be sequenced via next-generation sequencing approaches after enriching for a wider array of RNA viruses present in a sample through a hybrid probe-capture approach. This approach allows characterization of the prevalent SARS-CoV-2 genomic variants in a defined local region and the dynamics of other pathogenic viruses present in the sample[87–89]. Shotgun metagenomic and metatranscriptomic sequencing (i.e., community-based sequencing approaches) can provide a comprehensive snapshot of the viral community ecology and thereby aid the tracking of viruses of clinical significance in a community.

As SARS-CoV-2 transitions to become an endemic pathogen, wastewater genomic sequencing offers a scalable, less expensive, long-term passive surveillance tool to track emerging variants in the population. A global metagenomics approach has been suggested to detect, collect and store samples in preparation for future pandemics[90,91]. Resources such as GISAID, GenomeTrakr[92,93]

Table 1 | Online services with SARS-CoV-2 genome resources and analytics

Resource	Description	Link
GISAID	Platform for assembled genome sharing and analysis	https://www.gisaid.org/
NCBI GenBank	Sequence Read Archive (SRA)	https://www.ncbi.nlm.nih.gov/sars-cov-2/
COG-UK	United Kingdom sequences database	https://www.cogconsortium.uk/
PANGO	Lineage analytics	https://cov-lineages.org/
Nextstrain	Phylogenetic analysis	https://nextstrain.org/
WBEC	Wastewater analytics	https://www.covid19wbec.org/
COVID-3D	Structural changes of lineages	http://biosig.unimelb.edu.au/covid3d/
Outbreak.info	Variants reports	https://outbreak.info/
CoVizu	Global and local variant distribution analytical tool	https://filogeneti.ca/covizu/
CoVsurver	GISAID quality check and annotation tool identifying phenotypically or epidemiologically interesting candidate amino acid (aa) changes for further research	https://corona.bii.a-star.edu.sg/, https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/
KSA-KAUST	COVID-19 virus mutation tracker	https://www.cbrc.kaust.edu.sa/covmt/
COVID Genes	Shotgun RNA-seq viral data and host responses	https://covidgenes.weill.cornell.edu/

and the US National Wastewater Surveillance System (CDC-NWSS)[94] could facilitate the above efforts.

Outlook. The unprecedented volume of available SARS-CoV-2 genomic data coupled with available bioinformatics tools accelerated the prompt and effective characterization of SARS-CoV-2 genomes and provided tools enabling epidemiologists and public health officials to more effectively respond to the COVID-19 pandemic. Numerous independent efforts across the globe used bioinformatics methods, thereby demonstrating the utility of genomics-based approaches and creating a solid foundation for the response to COVID-19 and future pandemics. This was achieved by the standardization of methodology, protocol and data sharing, and applications of SARS-CoV-2 genomic data in epidemiological investigations.

Genome-based surveillance has been shown to be beneficial in addressing COVID-19. However, the unprecedented volume of sequencing data, currently six million complete SARS-CoV-2 genome sequences in databases, pose a challenge to the current systems of data storage, processing and bioinformatics analysis[16,19,95]. Owing to various technological burdens, such systems were still in the early stages of development when SARS-CoV-19 emerged in December 2019. COVID-19 has led to the mobilization of financial, scientific and developmental resources in record time, with numerous global surveillance systems providing resources for outbreak response using SARS-CoV-2 genome analysis (Table 1). A notable example is the timely deployment of GISAID and Nextstrain to address the COVID-19 response. This technology has played a leading role in centralizing efforts to collect and analyze SARS-CoV-2 genomic data.

Emerging VOCs, VOIs and VUIs are likely to continue shaping the course of the COVID-19 pandemic. Global genomics-based surveillance for new variants, in our view, will continue to play a leading role, with information on all SARS-CoV-2 lineages being collected and made available online for the rapid evaluation of their impact on transmission, virulence and vaccine escape[96,97]. We believe that targeted genomic surveillance of SARS-CoV-2 in immunocompromised patients can provide useful insights into the mechanisms of appearance of newly emerging VOCs. This can be done by applying bioinformatics tools for intra-host population analysis similar to those that are already available for other RNA viruses, such as HCV and HIV[82,98–101].

Efficient early detection and tracking of potentially dangerous variants requires real-time data from all countries[102]. The European Commission, for example, recommended achieving a capacity to sequence at least 5% of positive test results, which can be a good global standard. Yet, many underdeveloped countries face insurmountable logistic, technological and financial barriers to operating sequencing centers to accommodate this scale of testing, suggesting that developed countries should share responsibility for global surveilance[103]. Following the example of many African countries, countries in other regions that are currently lacking in viral genomic sequencing capability could establish additional sequencing centers. In regions where that is not practical, a logistically efficient system to obtain samples and deliver them to sequencing centers in other countries might be an appealing alternative.

In our view, there are three potential benefits of a standard genome epidemiological sequencing system. The immediate benefit is that this improve the timeliness and accuracy with which emerging VOIs and VOCs can be tracked. A longer-term goal is an improved ability to learn about the evolutionary pressures driving the emergence of novel, potentially dangerous variants. Presently, VOCs are declared based on their increased transmissibility or virulence, or decreased effectiveness of public health and social measures, available diagnostics, vaccines and therapeutics. Learning more about the evolutionary dynamics of emergent strains may lead to predictions of VOIs based on genomic sequence alone, further improving response times. Finally, a truly global system of pathogen genome sequencing and analysis is likely to improve our ability to combat future pandemics.

Global coordination of genomic data surveys will also allow wider application of wastewater-based or environmental-based virus surveillance[104]. Currently, wastewater-based monitoring lacks the granularity of clinical diagnostic testing and cannot discern a particular area of an outbreak when the wastewater treatment plant serves a large population. Sampling at a higher spatial resolution within the sewer system, or even at a building-level scale, could potentially provide early indications of viral outbreaks and help monitor their progression[105]. ❐

Sergey Knyazev[1,51], Karishma Chhugani[2,51], Varuni Sarwal[3], Ram Ayyala[4], Harman Singh[5], Smruthi Karthikeyan[6], Dhrithi Deshpande[2], Pelin Icer Baykal[7,8], Zoia Comarova[9], Angela Lu[2], Yuri Porozov[10,11], Tetyana I. Vasylyeva[12], Joel O. Wertheim[12], Braden T. Tierney[13], Charles Y. Chiu[14,15,16], Ren Sun[17,18], Aiping Wu[19,20], Malak S. Abedalthagafi[21,22], Victoria M. Pak[23,24], Shivashankar H. Nagaraj[25,26], Adam L. Smith[9], Pavel Skums[27],

Bogdan Pasaniuc[1,28,29,30,31], Andrey Komissarov[32], Christopher E. Mason[13,33,34,35], Eric Bortz[36], Philippe Lemey[37], Fyodor Kondrashov[38], Niko Beerenwinkel[7,8], Tommy Tsan-Yuk Lam[39,40,41], Nicholas C. Wu[42,43,44,45], Alex Zelikovsky[27], Rob Knight[6,46,47,48], Keith A. Crandall[49] and Serghei Mangul[50] ✉

[1]*Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA.* [2]*Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, CA, USA.* [3]*Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA.* [4]*Department of Translational Biomedical Informatics, University of Southern California, Los Angeles, CA, USA.* [5]*Department of Electrical Engineering, Indian Institute of Technology, Hauz Khas, New Delhi, India.* [6]*Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA.* [7]*Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.* [8]*SIB Swiss Institute of Bioinformatics, Basel, Switzerland.* [9]*Astani Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA, USA.* [10]*World-Class Research Center "Digital biodesign and personalized healthcare", I.M. Sechenov First Moscow State Medical University, Moscow, Russia.* [11]*Department of Computational Biology, Sirius University of Science and Technology, Sochi, Russia.* [12]*Department of Medicine, University of California, San Diego, La Jolla, CA, USA.* [13]*Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA.* [14]*Department of Laboratory Medicine, University of California, San Francisco, San Francisco, CA, USA.* [15]*Department of Medicine, Division of Infectious Diseases, University of California, San Francisco, San Francisco, CA, USA.* [16]*UCSF-Abbott Viral Diagnostics and Discovery Center, University of California, San Francisco, San Francisco, CA, USA.* [17]*Department of Molecular and Medical Pharmacology, University of California, Los Angeles, Los Angeles, CA, USA.* [18]*School of Biomedical Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, P.R. China.* [19]*Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China.* [20]*Suzhou Institute of Systems Medicine, Suzhou, China.* [21]*Genomics Research Department, Saudi Human Genome Project, King Fahad Medical City and King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia.* [22]*King Salman Center for Disability Research, Riyadh, Saudi Arabia.* [23]*Emory University, School of Nursing, Atlanta, GA, CA, USA.* [24]*Emory University, Rollins School of Public Health, Department of Epidemiology, Atlanta, GA, CA, USA.* [25]*Centre for Genomics and Personalised Health, Queensland University of Technology, Brisbane, Queensland, Australia.* [26]*Translational Research Institute, Brisbane, Queensland, Australia.* [27]*Department of Computer Science, College of Art and Science, Georgia State University, Atlanta, GA, USA.* [28]*Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA.* [29]*Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA.* [30]*Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA.* [31]*Institute of Precision Health, University of California, Los Angeles, Los Angeles, CA, USA.* [32]*Smorodintsev Research Institute of Influenza, Saint Petersburg, Russia.* [33]*The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA.* [34]*The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA.* [35]*The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA.* [36]*Department of Biological Sciences, University of Alaska Anchorage, Anchorage, AK, CA, USA.* [37]*Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven–University of Leuven, Leuven, Belgium.* [38]*Institute of Science and Technology Austria, Klosterneuburg, Austria.* [39]*State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, P.R. China.* [40]*Laboratory of Data Discovery for Health Limited, Hong Kong SAR, P.R. China.* [41]*Centre for Immunology & Infection Limited, Hong Kong SAR, P.R. China.* [42]*Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA.* [43]*Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA.* [44]*Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA.* [45]*Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL, USA.* [46]*Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA.* [47]*Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA, USA.* [48]*Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA.* [49]*Computational Biology Institute and Department of Biostatistics & Bioinformatics, Milken Institute School of Public Health, George Washington University, Washington, DC, USA.* [50]*Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, USA.* [51]*These authors contributed equally: Sergey Knyazev, Karishma Chhugani.*
✉*e-mail: serghei.mangul@gmail.com*

Published online: 8 April 2022
https://doi.org/10.1038/s41592-022-01444-z

References

1. Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. *Lancet* **395**, 470–473 (2020).
2. Grubaugh, N. D. et al. *Nat. Microbiol.* **4**, 10–19 (2019).
3. Rambaut, A. et al. *Nat. Microbiol.* **5**, 1403–1407 (2020).
4. Karim, S. S. A. & Karim, Q. A. *Lancet* **398**, 2126–2128 (2021).
5. Rockefeller Foundation. The Rockefeller Foundation releases new action plan to accelerate development of a national system for gathering and sharing information on SARS-CoV-2 genomic variants and other pathogens (2021); https://www.rockefellerfoundation.org/news/the-rockefeller-foundation-releases-new-action-plan-to-accelerate-development-of-a-national-system-for-gathering-and-sharing-information-on-sars-cov-2-genomic-variants-and-other-pathogens/
6. Huang, C. et al. *Lancet* **395**, 497–506 (2020).
7. Wu, F. et al. *Nature* **579**, 265–269 (2020).
8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. Mol. Biol.* **215**, 403–410 (1990).
9. Boni, M. F. et al. *Nat. Microbiol.* **5**, 1408–1417 (2020).
10. Wang, H., Pipes, L. & Nielsen, R. *Virus. Evol.* **7**, veaa098 (2021).
11. Dolgin, E. *Nature* **597**, 318–324 (2021).
12. Shu, Y. & McCauley, J. *Euro Surveill.* **22**, 30494 (2017).
13. The COVID-19 Genomics UK (COG-UK). *Lancet Microbe* **3**, e99–e100 (2020).
14. Fernandes, J. D. et al. *Nat. Genet.* **52**, 991–998 (2020).
15. Kuiken, C., Korber, B. & Shafer, R. W. *AIDS Rev.* **5**, 52–61 (2003).
16. Maxmen, A. *Nature* **593**, 21 (2021).
17. Kalia, K., Saberwal, G. & Sharma, G. *Nat. Biotechnol.* **39**, 1058–1060 (2021).
18. Inzaule, S. C., Tessema, S. K., Kebede, Y., Ogwell Ouma, A. E. & Nkengasong, J. N. *Lancet Infect. Dis.* **9**, e281–e289 (2021).
19. Van Noorden, R. *Nature* **590**, 195–196 (2021).
20. Elbe, S. & Buckland-Merrett, G. *Glob. Chall.* **1**, 33–46 (2017).
21. Geoghegan, J. L. & Holmes, E. C. *Nat. Rev. Genet.* **19**, 756–769 (2018).
22. van Dorp, L. et al. *Infect. Genet. Evol.* **83**, 104351 (2020).
23. Zhang, Y.-Z. & Holmes, E. C. *Cell* **181**, 223–227 (2020).
24. Korber, B. et al. *Cell* **182**, 812–827.e19 (2020).
25. Tao, K. et al. *Nat. Rev. Genet.* **22**, 757–773 (2021).
26. Kemp, S. A. et al. *Nature* **592**, 277–282 (2021).
27. Hay, A. J. & McCauley, J. W. *Influenza Other Respir. Viruses* **12**, 551–557 (2018).
28. Hadfield, J. et al. *Bioinformatics* **34**, 4121–4123 (2018).
29. Bull, R. A. et al. *Nat. Commun.* **11**, 6272 (2020).
30. Laiton-Donato, K. et al. Preprint at bioRxiv https://doi.org/10.1101/2020.06.26.20135715 (2020).
31. Yakovleva, A. et al. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-1044446/v1 (2021).
32. Tablizo, F. A. et al. *Microbiol. Resour. Announc.* **10**, e00219–2 (2021).
33. Viana, R. et al. *Nature* https://doi.org/10.1038/s41586-022-04411-y (2022).
34. Fauver, J. R. et al. *Cell* **181**, 990–996.e5 (2020).
35. Morel, B. et al. *Mol. Biol. Evol.* **38**, 1777–1791 (2021).
36. Novikov, D. et al. *J. Comput. Biol.* **28**, 1130–1141 (2021).
37. Steel, M. & Penny, D. *Appl. Math. Lett.* **17**, 785–790 (2004).
38. Woolley, S. M., Posada, D. & Crandall, K. A. *PLoS One* **3**, e1913 (2008).
39. Wertheim, J. O., Steel, M. & Sanderson, M. *J. Syst. Biol.* **71**, 426–438 (2021).

40. Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J. & Wertheim, J. O. *Mol. Biol. Evol.* **35**, 1812–1819 (2018).
41. Knyazev, S. et al. in *Data Bioinformatics Research and Applications: 17th International Symposium, ISBRA 2021, Shenzhen, China* (eds. Wei, Y., Li, M., Skums, P. & Cai, Z.) 165–175 (Springer, 2021).
42. Campbell, E. M. et al. *PLoS Comput. Biol.* **17**, e1009300 (2021).
43. Blair, C. & Ané, C. *Syst. Biol.* **69**, 593–601 (2020).
44. Martin, M. A., VanInsberghe, D. & Koelle, K. *Science* **371**, 466–467 (2021).
45. Hodcroft, E. B. et al. *Nature* **595**, 707–712 (2021).
46. McNamara, R. P. et al. *Cell Rep.* **33**, 108352 (2020).
47. Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S. & Stadler, T. *Proc. Natl Acad. Sci. USA* **118**, e2012008118 (2021).
48. Worobey, M. et al. *Science* **370**, 564–570 (2020).
49. Ladner, J. T. et al. *MBio* **11**, e02107–20 (2020).
50. Gonzalez-Reiche, A. S. et al. *Science* **369**, 297–301 (2020).
51. Gámbaro, F. et al. *Euro Surveill.* **25**, 2001200 (2020).
52. du Plessis, L. et al. *Science* **371**, 708–712 (2021).
53. Lemey, P. et al. *Nature* **595**, 713–717 (2021).
54. Lemey, P. et al. *Nat. Commun.* **11**, 5110 (2020).
55. Candido, D. D. S. et al. *J. Travel Med.* **27**, taaa042 (2020).
56. Badaoui, B., Sadki, K., Talbi, C., Salah, D. & Tazi, L. *Biosaf. Health* **3**, 124–127 (2021).
57. Rockett, R. J. et al. *Nat. Med.* **26**, 1398–1404 (2020).
58. Nadeau, S. A. et al. Preprint at medRxiv https://doi.org/10.1101/2021.11.11.21266107 (2021).
59. Nemira, A. et al. *Commun. Med.* **1**, 31 (2021).
60. Fahrenfeld, N. & Bisceglia, K. J. *Environ. Sci. Water Res. Technol.* **2**, 788–799 (2016).
61. Castiglioni, S., Senta, I., Borsotti, A., Davoli, E. & Zuccato, E. *Tob. Control* **24**, 38–42 (2015).
62. Sims, N. & Kasprzyk-Hordern, B. *Environ. Int.* **139**, 105689 (2020).
63. Chen, Y. et al. *J. Med. Virol.* **92**, 833–840 (2020).
64. COVID-19 wastewater epidemiology SARS-CoV-2. https://www.covid19wbec.org (accessed 12 November 2021).
65. Weidhaas, J. et al. *Sci. Total Environ.* **775**, 145790 (2020).
66. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. *Environ. Sci. Technol. Lett.* **7**, 511–516 (2020).
67. Ahmed, W. et al. *Sci. Total Environ.* **728**, 138764 (2020).
68. Gonzalez, R. et al. *Water Res.* **186**, 116296 (2020).
69. Peccia, J. et al. *Nat. Biotechnol.* **38**, 1164–1167 (2020).
70. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. *Environ. Sci. Technol. Lett.* **7**, 511–516 (2020).
71. Wu, F. et al. *Sci. Total Environ.* **805**, 150121 (2020).
72. Karthikeyan, S. et al. *mSystems* https://doi.org/10.1128/mSystems.00045-21 (2022).
73. Farkas, K., Hillary, L. S., Malham, S. K., McDonald, J. E. & Jones, D. L. *Current Opin. Environ. Sci. Health* **17**, 14–20 (2020).
74. Larsen, D. A. & Wigginton, K. R. *Nat. Biotechnol.* **38**, 1151–1153 (2020).
75. Schmidt, C. *Nat. Biotechnol.* **38**, 917–920 (2020).
76. Rothman, J. A. et al. *Appl. Environ. Microbiol.* **87**, e0144821 (2021).
77. DNA Pipelines R&D et al. *Protocols.io* https://www.protocols.io/view/covid-19-artic-v3-illumina-library-construction-an-bgxjjxkn/metadata (2020).
78. Sharkey, M. E. et al. *Sci. Total Environ.* **798**, 149177 (2021).
79. Turakhia, Y. et al. *Nat. Genet.* **53**, 809–816 (2021).
80. Andersen Lab. Freyja: depth-weighted de-mixing. Code at *GitHub* https://github.com/andersen-lab/Freyja (accessed 28 September 2021).
81. Karthikeyan, S. et al. Preprint at bioRxiv https://doi.org/10.1101/2021.12.21.21268143 (2021).
82. Knyazev, S., Hughes, L., Skums, P. & Zelikovsky, A. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbaa101 (2020).
83. Posada-Céspedes, S. et al. *Bioinformatics* **37**, 1673–1680 (2021).
84. Jahn, K. et al. Preprint at *medRxiv* https://doi.org/10.1101/2021.01.08.21249379 (2021).
85. Baaijens, J. A. et al. Preprint at *medRxiv* https://doi.org/10.1101/2021.08.31.21262938 (2021).
86. Caduff, L. et al. Preprint at *medRxiv* https://doi.org/10.1101/2021.08.22.21262024 (2021).
87. Crits-Christoph, A. et al. *mBio* https://doi.org/10.1128/mBio.02703-20 (2020).
88. Izquierdo-Lara, R. W. et al. *Emerg. Infect. Dis.* **27**, 1405–1415 (2020).
89. Nagy-Szakal, D. et al. *Microbiol. Spectr.* **9**, e0019721 (2021).
90. Carbo, E. C. et al. *J. Clin. Virol.* **131**, 104594 (2020).
91. Bedford, J. et al. *Nature* **575**, 130–136 (2019).
92. Center for Food Safety & Applied Nutrition. Wastewater surveillance for SARS-CoV-2 variants. (US Food and Drug Administration, 2021); https://www.fda.gov/food/whole-genome-sequencing-wgs-program/wastewater-surveillance-sars-cov-2-variants
93. BioProject. https://www.ncbi.nlm.nih.gov/bioproject/757291 (2021).
94. Centers for Disease Control and Prevention. National Wastewater Surveillance System (NWSS) (CDC, 2021); https://www.cdc.gov/healthywater/surveillance/wastewater-surveillance/wastewater-surveillance.html
95. Hodcroft, E. B. et al. *Nature* **591**, 30–33 (2021).
96. Rambaut, A. et al. *Nat. Microbiol.* **5**, 1403–1407 (2020).
97. Maxmen, A. *Nature* https://doi.org/10.1038/d41586-021-00490-5 (2021).
98. Knyazev, S. et al. *Nucleic Acids Res.* **49**, e102 (2021).
99. Sapoval, N. et al. *Genome Res.* **31**, 635–644 (2021).
100. Lythgoe, K. A. et al. *Science* **372**, eabg0821 (2021).
101. Butler, D. et al. *Nat. Commun.* **12**, 1660 (2021).
102. Kissler, S. M. et al. *N. Engl. J. Med.* **385**, 2489–2491 (2021).
103. MacKay, M. J. et al. *Nat. Biotechnol.* **38**, 1021–1024 (2020).
104. Danko, D. et al. *Cell* **184**, 3376–3393.e17 (2021).
105. Bogler, A. et al. *Nat. Sustain.* **3**, 981–990 (2020).

Acknowledgements

Our paper is dedicated to all freedom-loving people around the world, and to the people of Ukraine who fight for our freedom. We thank William M. Switzer and Ellsworth M. Campbell from the Division of HIV/AIDS Prevention, Centers for Disease Control and Prevention (CDC), Atlanta, GA, USA, for discussions and suggestions. We thank Jason Ladner from the Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, for providing suggestions and feedback. S.M. was partially supported by National Science Foundation grants 2041984. T.L. is supported by the NSFC Excellent Young Scientists Fund (Hong Kong and Macau; 31922087), Research Grants Council (RGC) Collaborative Research Fund (C7144-20GF), RGC Research Impact Fund (R7021-20), Innovation and Technology Commission's InnoHK funding (D24H) and Health and Medical Research Fund (COVID190223). P.S. was supported by US National Institutes of Health (NIH) grant 1R01EB025022 and National Science Foundation (NSF) grant 2047828. M.A. acknowledges King Abdulaziz City for Science and Technology and the Saudi Human Genome Project for technical and financial support (https://shgp.kacst.edu.sa) N.W. was supported by US NIH grants R00 AI139445, DP2 AT011966 and R01 AI167910. A.S. acknowledge funding from NSF grant no. 2029025. A.Z. has been partially supported by NIH grants 1R01EB025022-01 and 1R21CA241044-01A1. S. Knyazev has been partly supported by Molecular Basis of Disease at Georgia State University and NIH awards R01 HG009120, R01 MH115676, R01 AI153827 and U01 HG011715. A.W. has been supported by the CAMS Innovation Fund for Medical Sciences (2021-I2M-1-061). R.K. was supported by NSF project 2038509, RAPID: Improving QIIME 2 and UniFrac for Viruses to Respond to COVID-19, CDC project 30055281 with Scripps led by Kristian Andersen, Genomic sequencing of SARS-CoV-2 to investigate local and cross-border emergence and spread. J.O.W. was supported by NIH–National Institute of Allergy and Infectious Diseases (NIAID) R01 AI135992 and receives funding from the CDC unrelated to this work. T.I.V. is supported by the Branco Weiss Fellowship. Y.P. was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of state support for the creation and development of World-Class Research Centers "Digital biodesign and personalized healthcare" N○075-15-2020-926. E.B. was supported by a US National Institute of General Medical Sciences IDeA Alaska INBRE (P20GM103395) and NIAID CEIRR (75N93019R00028). C.E.M. thanks Testing for America (501c3), OpenCovidScreen Foundation, Igor Tulchinsky and the WorldQuant Foundation, Bill Ackman and Olivia Flatto and the Pershing Square Foundation, Ken Griffin and Citadel, the US National Institutes of Health (R01AI125416, R01AI151059, R21AI129851, U01DA053941), and the Alfred P. Sloan Foundation (G-2015-13964). C.Y.C. is supported by US CDC Epidemiology and Laboratory Capacity (ELC) for Infectious Diseases grant 6NU50CK000539 to the California Department of Public Health, the Innovative Genomics Institute (IGI) at the University of California, Berkeley, and University of California, San Francisco, NIH grant R33AI12945 and US CDC contract 75D30121C10991. A.K. was partly supported by RFBR grant 20-515-80017. P.L. acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. ~725422 - ReservoirDOCS), the Wellcome Trust through project 206298/Z/17/Z (Artic Network) and NIH grants R01 AI153044 and U19 AI135995. K.C. acknowledges support from the US NSF award EEID-IOS-2109688. F.K.'s work was supported by an ERC Consolidator grant to F.K. (771209–CharFL).

Author contributions

S.M. conceived of the idea presented and supervised the project. S.M., S. Knyazev and K.C. led the project. S.M., S. Knyazev, K.C., S. Karthikeyan, D.D., P.I.B., Z.C., A.L., Y.P., T.I.V., J.O.W., B.T.T., C.Y.C., R.S., A.W., M.S.A., V.M.P., S.H.N, A.L.S., P.S., A.K., B.P., C.E.M., E.B., F.K., N.C.W., N.B., P.L., S.K.N., T.T.-Y.L., A.Z., R.K. and K.A.C. contributed to the writing of the manuscript. V.S. produced figures in the main text. V.S. and H.S. produced supplementary figures. H.S. and R.A. created supplementary tables. All authors discussed the text and commented on the manuscript. All authors read and approved the final manuscript. These authors contributed equally: Sergey Knyazev and Karishma Chhugani (joint first co-authors). These authors contributed equally: Varuni Sarwal, Ram Ayyala and Harman Singh (joint second co-authors).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01444-z.