1	Unlocking capacities of genomics for the COVID-19 response and future pandemics
2	Authors and Affiliations
3	Sergey Knyazev ¹
4	sergey.n.knyazev@gmail.com
5	ORCID: https://orcid.org/0000-0003-0385-1831
6	Twitter: SeKnyaz
7	
8	Karishma Chhugani ²
9	<u>chhugani@usc.edu</u>
10	ORCID: https://orcid.org/0000-0001-8764-9352
11	Varuni Sarwal ³
12	sarwal8@gmail.com
13	ORCID: https://orcid.org/0000-0001-7563-9835
14	
15	Ram Ayyala⁴
16	ramayyala@gmail.com
17	ORCID: <u>https://orcid.org/0000-0001-7275-271X</u>
18	
19	Harman Singh⁵
20	harmansingh.iitd@gmail.com
21	ORCID: https://orcid.org/0000-0002-3970-6276
22	
23	Smruthi Karthikeyan ⁶
24	skarthikeyan@health.ucsd.edu
25	Dhrithi Deshpande ²
26	dhrithideshpande@gmail.com

- 27 ORCID: <u>https://orcid.org/0000-0003-3794-7364</u>
- 28
- 29 Pelin Icer Baykal^{7,8}
- 30 <u>icer.pelin@gmail.com</u>
- 31
- 32 Zoia Comarova⁹
- 33 <u>zoia.comarova@gmail.com</u>
- 34
- 35 Angela Lu²
- 36 <u>alu52904@usc.edu</u>
- 37
- 38 Yuri Porozov^{10,11}
- 39 <u>yuri.porozov@gmail.com</u>
- 40 ORCID: <u>https://orcid.org/0000-0002-5450-3135</u>
- 41
- 42 Tetyana I. Vasylyeva¹²
- 43 tvasylyeva@health.ucsd.edu
- 44
- 45 Joel O. Wertheim¹²
- 46 jwertheim@health.ucsd.edu
- 47
- 48 Braden T. Tierney¹³
- 49 <u>btt4001@med.cornell.edu</u>
- 50
- 51 Charles Y. Chiu^{14,15,16}
- 52 ORCID: https://orcid.org/0000-0003-2915-2094

- 53 <u>charles.chiu@ucsf.edu</u>
 54
 55 Ren Sun^{17,18}
 56 <u>rensun@hku.hk</u>
- 57
- 58 Aiping Wu^{19,20}
- 59 wap@ism.cams.cn
- 60
- 61 Malak S. Abedalthagafi^{21,22}
- 62 malthagafi@kacst.edu.sa
- 63 ORCID: <u>https://orcid.org/0000-0003-1786-3366</u>
- 64 Twitter: @malakabed
- 65
- 66 Victoria M. Pak^{23,24}
- 67 vpak@emory.edu
- 68
- 69 Shivashankar H. Nagaraj^{25,26}
- 70 <u>shivashank@gmail.com</u>
- 71
- 72 Adam L. Smith⁹
- 73 <u>smithada@usc.edu</u>
- 74
- 75 Pavel Skums²⁷
- 76 <u>pskums@gsu.edu</u>
- 77 ORCID: <u>https://orcid.org/0000-0003-4007-5624</u>
- 78
- 79 Bogdan Pasaniuc^{1,28,29,30,31}

80	pasaniuc@ucla.edu
81	
82	Andrey Komissarov ³²
83	andrey.komissarov@influenza.spb.ru
84	ORCID: https://orcid.org/0000-0003-1733-1255
85	
86	Christopher E. Mason ^{33,34,35,36}
87	chm2042@med.cornell.edu
88	ORCID: <u>https://orcid.org/0000-0002-1850-1642</u>
89	
90	Eric Bortz ³⁷
91	<u>ebortz@alaska.edu</u>
92	
93	Philippe Lemey ³⁸
94	philippe.lemey@kuleuven.be
95	ORCID: https://orcid.org/0000-0003-2826-5353
95 96	ORCID: <u>https://orcid.org/0000-0003-2826-5353</u>
95 96 97	ORCID: <u>https://orcid.org/0000-0003-2826-5353</u> Fyodor Kondrashov ³⁹
95 96 97 98	ORCID: <u>https://orcid.org/0000-0003-2826-5353</u> Fyodor Kondrashov ³⁹ <u>fyodor.kondrashov@ist.ac.at</u>
95 96 97 98 99	ORCID: <u>https://orcid.org/0000-0003-2826-5353</u> Fyodor Kondrashov ³⁹ <u>fyodor.kondrashov@ist.ac.at</u>
95 96 97 98 99 100	ORCID: <u>https://orcid.org/0000-0003-2826-5353</u> Fyodor Kondrashov ³⁹ <u>fyodor.kondrashov@ist.ac.at</u> Niko Beerenwinkel ^{7,8}
95 96 97 98 99 100 101	ORCID: <u>https://orcid.org/0000-0003-2826-5353</u> Fyodor Kondrashov ³⁹ <u>fyodor.kondrashov@ist.ac.at</u> Niko Beerenwinkel ^{7,8} <u>niko.beerenwinkel@bsse.ethz.ch</u>
95 96 97 98 99 100 101 102	ORCID: https://orcid.org/0000-0003-2826-5353 Fyodor Kondrashov ³⁹ fyodor.kondrashov@ist.ac.at Niko Beerenwinkel ^{7,8} <u>niko.beerenwinkel@bsse.ethz.ch</u> ORCID: https://orcid.org/0000-0002-0573-6119
95 96 97 98 99 100 101 102 103	ORCID: https://orcid.org/0000-0003-2826-5353 Fyodor Kondrashov ³⁹ fyodor.kondrashov@ist.ac.at Niko Beerenwinkel ^{7,8} niko.beerenwinkel@bsse.ethz.ch ORCID: https://orcid.org/0000-0002-0573-6119 https://twitter.com/cbg_ethz
95 96 97 98 99 100 101 102 103 104	ORCID: https://orcid.org/0000-0003-2826-5353 Fyodor Kondrashov ³⁹ fyodor.kondrashov@ist.ac.at Niko Beerenwinkel ^{7,8} niko.beerenwinkel@bsse.ethz.ch ORCID: https://orcid.org/0000-0002-0573-6119 https://twitter.com/cbg_ethz

- 106 <u>ttylam@hku.hk</u>
- 107 ORCID: <u>https://orcid.org/0000-0002-9769-1527</u>
- 108
- 109 Nicholas C. Wu^{43,44,45,46}
- 110 <u>nicwu@illinois.edu</u>
- 111 ORCID: <u>http://orcid.org/0000-0002-9078-6697</u>
- 112
- 113 Alex Zelikovsky⁴⁷
- 114 <u>alexz@gsu.edu</u>
- 115 ORCID: <u>https://orcid.org/0000-0003-4424-4691</u>
- 116
- 117 Rob Knight^{6,48,49,50}
- 118 robknight@eng.ucsd.edu
- 119
- 120 Keith A. Crandall⁵¹
- 121 kcrandall@gwu.edu
- 122
- 123 Serghei Mangul⁵²
- 124 serghei.mangul@gmail.com
- 125 ORCID: <u>https://orcid.org/0000-0003-4770-3443</u>
- 126 Twitter: smangul1
- 127
- 128
- 129
- 130 Affiliations
- 131 1. Department of Pathology and Laboratory Medicine, David Geffen School of Medicine,
- 132 University of California, Los Angeles, Los Angeles, CA 90095, USA

- 133 2. Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, 134 University of Southern California, 1985 Zonal Avenue, Room 713. Los Angeles, CA 90089, 135 USA 3. Department of Computer Science, University of California Los Angeles, 580 Portola Plaza, 136 137 Los Angeles, CA 90095, USA 138 4. Department of Translational Biomedical Informatics, University of Southern California, 139 1450 Biggy Street, NRT 2509, Los Angeles, CA 90033 5. Department of Electrical Engineering, Indian Institute of Technology, Hauz Khas, New 140 141 Delhi, 110016, India 6. Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA 142 143 7. Department of Biosystems Science and Engineering, ETH Zurich, Basel, 4058, 144 Switzerland 145 8. SIB Swiss Institute of Bioinformatics, Basel, 4058, Switzerland 146 9. Astani Department of Civil and Environmental Engineering, University of Southern 147 California, 3620 South Vermont Avenue, Los Angeles, CA 90089 148 10. World-Class Research Center "Digital biodesign and personalized healthcare", I.M. 149 Sechenov First Moscow State Medical University, Moscow, Russia 150 11. Department of Computational Biology, Sirius University of Science and Technology, Sochi, Russia 151 152 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 153 154 14. Department of Laboratory Medicine, University of California San Francisco, San 155 Francisco, CA, 94143, USA 156 15. Department of Medicine, Division of Infectious Diseases, University of California San 157 Francisco, San Francisco, CA, 94143, USA 16. UCSF-Abbott Viral Diagnostics and Discovery Center, University of California San 158 159 Francisco, San Francisco, CA, 94143, USA 160 17. Department of Molecular and Medical Pharmacology, University of California, Los 161 Angeles, Los Angeles, CA 90095, USA 162 18. School of Biomedical Sciences, LKS Faculty of Medicine, The University of Hong Kong, 163 Hong Kong, PR China 164 19. Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union 165 Medical College, Beijing, 100005, China 166 20. Suzhou Institute of Systems Medicine, Suzhou, 215123, China 167 21. Genomics Research Department, Saudi Human Genome Project, King Fahad Medical 168 City and King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia 22. King Salman Center for Disability Research, Riyadh 12512, Saudi Arabia 169 170 23. Emory University, School of Nursing, Atlanta, GA 171 24. Emory University, Rollins School of Public Health, Department of Epidemiology, Atlanta, 172 GA 173 25. Centre for Genomics and Personalised Health, Queensland University of Technology, 174 Brisbane, QLD 4059, Australia
- 175 26. Translational Research Institute, Brisbane, Australia

176 27. Department of Computer Science, College of Art and Science, Georgia State University, 177 25 Park Place, Atlanta, GA 30303, USA 28. Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los 178 179 Angeles, CA 90095, USA 180 29. Department of Human Genetics, David Geffen School of Medicine, University of California, 181 Los Angeles, Los Angeles, CA 90095, USA 182 30. Department of Computational Medicine, David Geffen School of Medicine, University of 183 California, Los Angeles, Los Angeles, CA 90095, USA 184 31. Institute of Precision Health, University of California, Los Angeles, Los Angeles, CA 185 90095, USA 186 32. Smorodintsev Research Institute of Influenza, Saint Petersburg, Russia 187 33. Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA 188 34. The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational 189 Biomedicine, Weill Cornell Medicine, New York, NY, USA 190 35. The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, 191 NY, USA 192 36. The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, 193 USA 194 37. Department of Biological Sciences, University of Alaska Anchorage, 3211 Providence Dr, 195 Anchorage AK 99508 38. Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven 196 197 - University of Leuven, Leuven, Belgium 198 39. Institute of Science and Technology Austria, 1 Am Campus, Klosterneuburg, 3400, Austria 199 40. State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The 200 University of Hong Kong, Hong Kong SAR, P. R. China 201 41. Laboratory of Data Discovery for Health Limited, 19W Hong Kong Science & Technology 202 Parks, Hong Kong SAR, P. R. China 203 42. Centre for Immunology & Infection Limited, 17W Hong Kong Science & Technology Parks, 204 Hong Kong SAR, P. R. China 205 43. Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 206 61801. USA 207 44. Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 208 Urbana, IL 61801, USA 209 45. Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-210 Champaign, Urbana, IL 61801, USA 211 46. Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL 212 61801, USA 213 47. Department of Computer Science, College of Art and Science, Georgia State University, 214 1 Park Place, Floor 6, Atlanta, GA 30303, USA 215 48. Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA 216 49. Department of Computer Science & Engineering, University of California, San Diego, La 217 Jolla, CA, USA 218 50. Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA

- 51. Computational Biology Institute and Department of Biostatistics & Bioinformatics, Milken
 Institute School of Public Health, George Washington University, Washington, DC 20052
 52. Department of Clinical Pharmacy, School of Pharmacy, University of Southern California,
- 222 1540 Alcazar Street, Los Angeles, CA 90033, USA
- 223
- 224
- 225 Equal Contributions statement:

These authors contributed equally: Sergey Knyazev and Karishma Chhugani (joint first coauthors). These authors contributed equally: Varuni Sarwal, Ram Ayyala, and Harman Singh (joint second co-authors).

- 229 Corresponding Author
- 230 Correspondence to Serghei Mangul (serghei.mangul@gmail.com)
- 231
- 232
- 233

234 Standfirst

235 During the COVID-19 pandemic, genomics and bioinformatics have emerged as essential 236 public health tools. The genomic data acquired using these methods have supported the 237 global health response, facilitated development of testing methods, and allowed timely 238 tracking of novel SARS-CoV-2 variants. Yet the virtually unlimited potential for rapid 239 generation and analysis of genomic data is also coupled with unique technical, scientific, 240 and organizational challenges. Here, we discuss the application of genomic and 241 computational methods for the efficient data driven COVID-19 response, advantages of 242 democratization of viral sequencing around the world, and challenges associated with viral 243 genome data collection and processing.

244 Introduction

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 the country 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 America¹.

251 During this time, it became clear that the genomic toolkits are essential for public health decision 252 making, including testing for COVID-19, monitoring for emergence of new virus variants with 253 altered biological or immunological properties, identification of at-risk individuals and informing 254 epidemiological models that describe outbreaks in communities². This has allowed for the 255 observation of SARS-CoV-2 genome evolution in almost real time, rapid tracking of SARS-CoV-256 2 genetic lineages, and variants of interest and concern (VOIs, VOCs) which in turn have 257 facilitated the development of SARS-CoV-2 clinical tests and prediction of vaccines efficacy 258 against viral variants^{3,4}. However, to reach the full potential of genomic data for future public health 259 surveillance and outbreak response, we believe it is necessary to expand and coordinate best 260 practices in genomics and bioinformatics that have now been field tested during the COVID-19 261 response⁵. Herein, we discuss the genomic techniques and corresponding bioinformatics 262 algorithms that are addressing many of the pressing public health issues associated with COVID-263 19.

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

When 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 origin⁶. This rapid discovery was made possible by modern robust and accurate genomic and bioinformatic tools, which while now used routinely, did not exist a couple of decades ago. On January 30, 2020, when WHO declared a Public Health Emergency of International Concern (PHEIC) 339 SARS-CoV-2 genomes had already been sequenced and characterized¹.

272 To investigate the newly emerging outbreak, scientists in China performed whole-genome 273 sequencing on specimens, followed by de-novo assembly and end mapping to annotate the 274 complete 29,903 nucleotides long SARS-CoV-2 genome. Bioinformatics analysis revealed that 275 the genome organization of SARS-CoV-2 was consistent with single-stranded, positive-sense 276 RNA from the genus Betacoronavirus⁷. Additionally, sequence alignment tools including BLAST⁸ 277 were used to search for related species of the newly discovered virus in the NCBI GenBank 278 database, revealing alarming similarities to SARS-CoV (SARS-CoV-1), and a much higher 279 similarity with Betacoronavirus from bats, proposing zoonotic origin of the virus. Some SARS-280 CoV-2 genome fragments, in addition, have the highest similarity to the corresponding fragments 281 from pangolins, which suggests that there were possible recombination events. Subsequent 282 analyses including on additional sarbecovirus genomes from bats and pangolins further 283 scrutinized the evolution and recombination history, and found that the lineage giving rise to 284 SARS-CoV-2 had been probably circulating unnoticed in bats for decades^{9,10}.

285 Genomics-based methods shaped the effective COVID-19 response

286 Once the SARS-CoV-2 genome was sequenced, the authors immediately publicly deposited the 287 genome to GenBank^{7,11}. Timely open access release of the virus genome sequence was a 288 laudable decision that allowed informed scientific analyses and pandemic preparation to begin 289 immediately.

290 As the pandemic progressed, increased availability of modern sequencing technologies prompted 291 the collection of SARS-CoV-2 viral genomic data at an unprecedented scale. Within a month, on 292 average about 1,300 genomes were being submitted per day. Within six months of the pandemic 293 (May 2020), GISAID had 110,000 SARS-CoV-2 full-length genome sequences. By December 294 2021, two years into the pandemic, 67,000 genomes per day were being deposited into public 295 viral genome data repositories like GISAID, COG-UK, and GenBank, which currently contain over 6 million SARS-CoV-2 genomes^{12–14} (Figure 1a, Table S1). The unprecedented volume of data 296 297 collection for SARS-CoV-2 is seen when contrasted with HIV genomic data collection. HIV that 298 consistently captivated the attention of public health officials and the general public since 1980's, 299 has fewer than 16,000 full-length genome sequences collected by the biggest public HIV 300 database at Los Alamos sequence National Laboratory over the past 40 years¹⁵ (Figure 1a).

301 Sequencing data collected all over the world and rapidly shared on online databases ultimately 302 aided public health officials and governments in making better-informed decisions¹⁶. However, to 303 fully explore the potential of such databases, there are a few issues which need to be solved. 304 Despite the unprecedented pace overall, inevitable delays caused by shortage of sequencing 305 capacity and political interference in some regions led to problems in the logistical chain in these 306 regions, including in sample collection, transporting, and shipping samples¹⁷. Depending on the 307 country and the strength of their public health infrastructure, the median collection to submission 308 time lag differs, ranging from one day to one year. Several factors impact the rate and scale of 309 viral genomic sequencing across the globe. Countries with minimal sequencing are likely to 310 encounter outbreaks of higher severity, leading to blind spots of genomic surveillance that can 311 facilitate the spread of new variants to other countries¹⁷. On average high-income countries 312 shared about 100 times more sequences per capita than low-income countries (Figures 1b and

S2). However, some African countries with a low GDP per capita were able to sequence a comparable number of viral genomes of middle- and high-income countries¹⁸. This preparedness can be attributed to previous global initiatives to support African countries in mitigating outbreaks of other viruses that has enhanced the sequencing capacity of the region. Africa provides a remarkable example of the necessity of international cooperation that could be implemented in other parts of the globe to improve pandemic response (Figure 1c). The number of shared coronavirus genomes per capita is correlated with the country's GDP per capita (Figure 1d).

Moving forward, several important data sharing issues need to be addressed to facilitate open and rapid viral genome data sharing. Scientists depositing sequencing data should trust that their rights will be respected by data users and that their authorship rights will not be violated¹⁹. For instance, GISAID data access mechanism proved its ability to overcome these obstacles to the international sharing of virus data, making GISAID the largest repository of influenza and SARS-CoV-2 genomic data^{16,20}.

326 Bioinformatics methods are capable of accurately tracking SARS-CoV-2 genomic 327 evolution

328 As SARS-CoV-2 has spread through the world population over the first year of the pandemic, it gradually evolved into several viral lineages²¹⁻²⁴. Based on statistical analysis of collected SARS-329 330 CoV-2 genomes, it was shown that SARS-CoV-2 has a mutation rate of at least 10-fold lower than seasonal influenza²⁵. The lower mutation rate initially gave hope for efficient control of the 331 332 pandemic through vaccination because the slower the virus mutates, the less chances it has to 333 adapt to vaccines. However, given the large number of COVID-19 cases (>277 million and 334 climbing, according to WHO) and possibly because of SARS-CoV-2 recombination events, new 335 variants continue to evolve, which are being classified as variant under investigation (VUI), of 336 interest (VOI), and of concern (VOC) according to their epidemiological, biological and/or 337 immunological properties. Indeed, some variants acquired numerous mutations in a rapid fashion 338 (variants Alpha and Omicron) and showed evidence of immune escape (Omicron). Notably, it was 339 observed that immunodeficient individuals with unusually long periods of SARS-CoV-2 infection 340 can create a plausible environment for faster SARS-CoV-2 evolution because their immune 341 system allows for viral immune escape²⁶.

342 Prior to the COVID-19 pandemic, the public health community has had experience with tracking 343 and responding to genome evolution for viruses such as the seasonal flu causing influenza 344 viruses. The Global Influenza Surveillance and Response System (GISRS) was established by 345 WHO for timely collection, genetic and antigenic characterization of these viruses²⁷. Sharing of 346 virus sequence data in the GISAID database along with Nextstrain²⁸ online phylogenetic tool was 347 utilized for biannual influenza A and B vaccine seed strain selection and understanding viral 348 genomic evolution and antigenic drift. GISAID and Nextstrain were both promptly adopted for 349 collecting and analyzing SARS-CoV-2 genomic data, becoming the largest global system for 350 tracking SARS-CoV-2 evolution and monitoring of the 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 open353 source bioinformatics workflows for accurate consensus genome assembly²⁹. However, the use 354 of proprietary next-generation sequencing solutions and software has been more commonplace 355 in well-resourced national and state/province level public health labs. The accessibility of tiled 356 primer sequences (e.g., ARCTIC or midnight primer sets), lower costs of Illumina and Oxford 357 Nanopore sequencing along with open access bioinformatics workflows supported sequencing in 358 dozens of regional public health labs and academic institutions across the world. By December 359 24th, 2021, 80.49% of available SARS-CoV-2 genomic data at GISAID was generated by Illumina sequencers, 12.46% by Oxford Nanopore, and 3.85% by Pacbio, 1.59% by IonTorrent, 1.29% by 360 361 BGI, 0.31% by Sanger and 0.02% by QIAGEN (Figure S1a). NCBI GenBank has 91.04% genomic 362 data sequenced by Illumina, 8.1% by Oxford Nanopore, 0.47% by IonTorrent, < 0.01% by PacBio, 363 and 0.38% unspecified (Figure S1b).

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 emergence and spread of variants, for example in Colombia (VOI Mu), Ukraine (VOC Delta), the Philippines (VOC Alpha), in the U.K. (VOC Alpha) as it moved to the U.S., and in South Africa, where immune evasive VOC Omicron was identified by genome sequencing^{30–33}.

370 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 371 372 introduced can reveal geographical transmission routes. These routes help distinguish imported 373 cases from community transmission, aiding the identification of high-risk transmission routes that 374 can be subject to enhanced public health control³⁴. Comparative genomic analyses for studying 375 COVID-19 outbreak transmission dynamics have been mostly conducted using classic maximum 376 likelihood (ML) phylogenetic methods³⁵. Unfortunately, ML methods are not scalable enough to 377 handle large volumes of SARS-CoV-2 genomic data available. It is often a requirement, therefore, 378 for ML to reduce sample size and to consider only a fraction of the data in order to conduct the 379 analysis, which can potentially compromise the accuracy of the results. Alternatively, more 380 scalable approximate maximum parsimony methods (MP) can be used for phylogeny reconstruction for SARS-CoV-2 dense data³⁶. Indeed, it was shown theoretically that with dense 381 382 enough sampling, MP produces an ML tree under certain maximum likelihood models^{37–39}. 383 Another approach has been to use network-based methods, which are significantly faster but 384 theoretically less accurate than phylogeny-based methods⁴⁰⁻⁴².

385 Diverse publicly available SARS-CoV-2 genome sequences from around the world have aided 386 efficient and accurate tracking of local and global SARS-CoV-2 transmission routes^{43–45} (Figure 387 S3). Phylogenetics methods (listed in Table S2) revealed that SARS-CoV-2 was introduced into Europe from China and into the US from China and Europe^{34,46–48} and have also been used to 388 389 track domestic transmission chains and differentiate them from international ones. For example, 390 studies showed that SARS-CoV-2 was likely introduced in Connecticut via a domestic 391 transmission route while the most successful viral introductions in Arizona were likely via domestic travel^{34,49}. New York City area experienced multiple introductions of SARS-CoV-2, primarily from 392 393 Europe⁵⁰. 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 Madagascar⁵¹ (Figure 1e, Tables S2).

396 Differences in sampling across geographical locations and over time represent a considerable 397 challenge to accurately reconstruct spatial transmission patterns. However, additional data such 398 as travel information and epidemiological estimates may help mitigate non-uniform sampling 399 across geographical locations and time and contribute to a more complete picture of viral spread. 400 This has been illustrated by a study of SARS-CoV-2 importation and establishment in the UK⁵². 401 Large-scale genomic data resulted in estimates of the number and timing of introductions events, 402 but its combination with epidemiological and travel data allowed identification of the 403 spatiotemporal origins of these introductions. Such additional data sources are also being 404 increasingly integrated in phylodynamic inferences. For example, a study of the contribution of 405 persistence versus new introductions to the second COVID-19 wave in Europe made use of 406 Google mobility data to inform the phylogeographic component of the genomic reconstruction⁵³. 407 Furthermore, the individual travel history of sampled individuals can be formally incorporated in 408 such analyses⁵⁴.

409 Additionally, phylogenetics can be used to monitor the effectiveness of global travel restrictions 410 and lockdowns. For example, it was shown that the risk of domestic transmission of SARS-CoV-411 2 in Connecticut already exceeded that of international introduction at the time federal travel restrictions were imposed, highlighting the critical need for local surveillance³⁴. Similarly in Brazil, 412 413 three clades of European origin were established prior to the initiation of travel bans and 414 lockdowns⁵⁵. In the UK, lineages introduced prior to national lockdown were shown to be larger 415 and more dispersed and lineage importation and regional lineage diversity declined after lockdown⁵². Phylogenetics showed that, due to violations of imposed lockdowns with sea trade, 416 417 several SARS-CoV-2 international introductions likely occurred in Morocco⁵⁶. In Australia, 418 lockdown effectiveness was validated using SARS-CoV-2 genomic data coupled with agent-419 based modeling, a computation tool to simulate the interactions of autonomous agents such as 420 individuals⁵⁷. Phylogenetic modeling of over 11,000 SARS-CoV-2 genomes collected in 421 Switzerland throughout 2020 enabled estimating the effect of different public health measures, including lockdown, border closure, and test-trace-isolate efforts⁵⁸. Similarly, comparative 422 423 phylodynamics analysis of SARS-CoV-2 transmission dynamics in neighboring Eastern European 424 countries of Belarus and Ukraine, that followed highly different COVID-19 containment policies, 425 allowed to assess the effectiveness of public health intervention measures in this region, and 426 highlight the role of regional political and social factors in the virus spread⁵⁹.

427 Genomics methods enable wastewater-based monitoring of SARS-CoV-2 epidemiology

The presence of trace viral genomic material in wastewater has been successfully employed to track antibiotic use⁶⁰, tobacco consumption⁶¹ and the monitoring of several respiratory and enteric viruses including poliovirus⁶². Although COVID-19 is primarily associated with respiratory symptoms, SARS-CoV-2 is regularly shed in feces of infected individuals⁶³. As of December 2021, wastewater-based surveillance for tracking SARS-CoV-2 viral infection dynamics⁶⁴ has been implemented in many countries around the world (Figure 1e).

434 Wastewater-based epidemiology has been shown to provide more balanced estimates of viral 435 prevalence rates in a population than clinical testing alone due to inherent limitations in testing 436 resources and/or testing uptake rates especially in underserved communities. Combining clinical 437 diagnostics with wastewater-based surveillance can provide a more comprehensive community-438 level profile of both symptomatic and asymptomatic cases, enabling identification of hospital 439 capacity needs^{65–72}. Additionally, an important advantage of wastewater monitoring is the ability 440 to detect early-stage outbreaks before they become widespread^{62,73-76}. Although tracking of 441 SARS-CoV-2 viral RNA via gPCR-based methods can reveal temporal changes of virus 442 prevalence in a given population, it cannot provide underlying epidemiological information for 443 identifying transmission or genomic details on emerging variants. Tracking viral genomic 444 sequences from wastewater significantly ameliorates community prevalence estimates and also 445 detects emerging variants. Tracking SARS-CoV-2 viral genomic sequences from wastewater 446 using a targeted tiled amplicon-based sequencing approach would significantly ameliorate 447 community prevalence estimates and also detect emerging variants⁷⁷.

Wastewater genomic epidemiology can also act as a surrogate to elucidate strain geospatial distributions, helping identify outbreak clusters and track prevailing and newly emerging variants, 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 accuracy⁷⁸.

454 The commonly used tools used for discerning viral lineages in clinical samples such as pangolin³ 455 and UShER⁷⁹ cannot deconvolute the multiple lineages that are commonly observed in a single 456 wastewater sample and at best detect the most dominant one. As existing lineage calling methods 457 require a single consensus sequence to perform assignment, they are ill-equipped to capture the 458 diversity present in mixed viral samples. Hence, tools to robustly identify the multiple lineages and 459 their relative proportions present in wastewater are critical in understanding and interpreting the 460 underlying sequence data obtained from these samples. For example, a depth-weighted demixing 461 algorithm Freyja⁸⁰, uses a "barcode" library of lineage defining mutations to represent each viral 462 variant and can be used to recover relative abundance in the sample. This approach enabled the 463 early detection of emerging VOCs in wastewater up to 14 days in advance of first clinical detection 464 and also identified multiple instances of cryptic transmission not observed via clinical genomic 465 surveillance⁸¹. Similar algorithms for mutation calling, haplotype reconstruction, and population 466 characterization in viral specimens, can also be used to deconvolute the mixture of variants 467 present in a wastewater sample^{82,83}. By searching for signature mutations co-occurring on the 468 same amplicon, variant B.1.1.7 in wastewater was detected eight days before the first patient 469 sample was tested positive for the variant⁸⁴. Similarly, RNA transcript quantification methods, 470 such as Kallisto, can be used to estimate the relative abundance of SARS-CoV-2 variants in 471 wastewater⁸⁵. Both digital PCR-based and sequencing-based estimates of variant abundance in 472 wastewater have been used to derive the fitness advantage of a recently introduced variant, an 473 important epidemiological parameter to assess the expected transmissibility and spread of the 474 variant^{86,87}.

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 dynamics of other pathogenic viruses present in the sample^{88–90}. Shotgun metagenomic and metatranscriptomic sequencing (i.e. communitybased sequencing approaches) can provide a comprehensive snapshot of the viral community ecology and thereby aid in 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 for tracking emerging variants in the population. A global metagenomics approach has been suggested to detect, collect, and store samples in preparation for future pandemics^{91,92}. Resources such as GISAID, GenomeTrakr^{93,94} and CDC-NWSS⁹⁵ (National Wastewater Surveillance System)⁹⁵ could facilitate the above efforts.

488 Outlook

489 The unprecedented volume of available SARS-CoV-2 genomic data coupled with available 490 bioinformatics tools accelerated the prompt and effective characterization of SARS-CoV-2 491 genomes and provided tools to epidemiologists and public health officials to more effectively 492 respond to the COVID-19 pandemic. Numerous independent efforts across the globe utilized 493 bioinformatics methods that demonstrated the utility of genomics-based approaches and created 494 a solid foundation for the response to COVID-19 and future pandemics. This was achieved by the 495 standardization of methodology, protocol and data sharing, and applications of using SARS-CoV-496 2 genomic data in epidemiological investigations.

497 Genome-based surveillance has been shown to be beneficial in addressing COVID-19. However, 498 the unprecedented volume of sequencing data, currently six million complete SARS-CoV-2 499 genome sequences in databases, challenged the current systems of data storage, processing, 500 and bioinformatics analysis^{16,19,96}. Due to various technological burdens, such systems were still 501 in the early stages of development in December of 2019. COVID-19 has led to the mobilization of 502 financial, scientific, and developmental resources in record time, with numerous global 503 surveillance systems that provided resources for outbreak response using SARS-CoV-2 genome 504 analysis (Table 1). A notable example is the timely deployment of GISAID and Nextstrain for 505 addressing the COVID-19 response. This technology has taken a lead in centralizing efforts to 506 collect and analyse SARS-CoV-2 genomic data.

507 Emerging VOCs, VOIs, and VUIs are likely to continue shaping the course of the COVID-19 508 pandemic. Global genomics-based surveillance for new variants, in our view, will continue to play 509 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^{97,98}. 510 511 Targeted genomic surveillance of SARS-CoV-2 in immunocompromised patients, in our view, can 512 provide useful insights into the mechanisms of appearance of newly emerging VOCs. This can 513 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^{82,99–102}. 514

515 Efficient early detection and tracking of potentially dangerous variants requires real-time data from 516 all countries¹⁰³. The European Commission, for example, recommended gaining a capacity of 517 sequencing of at least 5% of positive test results, which can be a good global standard. Yet, many 518 underdeveloped countries in the world face insurmountable logistic, technological, and financial 519 barriers to operating sequencing centers to accommodate this scale, suggesting that developed 520 countries share responsibility for global surveilance¹⁰⁴. Following the example of many African 521 countries, additional sequencing centers in countries without viral genomic sequencing could be 522 established. In regions where that is not practical, a logistically efficient system of obtaining and 523 delivering samples to sequencing centers in other countries might be an appealing alternative.

524 In our view, there are three potential benefits of a standard genome epidemiological sequencing 525 system. The immediate benefit will allow improved timeliness and accuracy of tracking emerging 526 VOI and VOC. A longer-term goal is an improved ability to learn about evolutionary pressures 527 driving the emergence of novel, potentially dangerous variants. Presently, VOC are declared 528 based on their increased transmissibility or virulence, or decreased effectiveness of public health 529 and social measures, available diagnostics, vaccines and therapeutics. Learning more about 530 evolutionary dynamics of emergent strains may lead to predictions of VOI based on genomic 531 sequence alone, further improving response times. Finally, a truly global system of pathogen 532 genome sequencing and analysis is likely to improve our ability to combat future pandemics.

533 Global coordination of genomic data surveys will also allow for wider application of wastewater-534 based or environmental-based virus surveillance¹⁰⁵. Currently, wastewater-based monitoring 535 lacks the granularity of clinical diagnostic testing and cannot discern a particular area of an 536 outbreak when the wastewater treatment plant serves a large population. Sampling at a higher 537 spatial resolution within the sewer system or even at a building-level scale could potentially 538 provide early indications of viral outbreaks and help monitor their progression¹⁰⁶.

539 Acknowledgements

540 We thank William M. Switzer and Ellsworth M. Campbell from the Division of HIV/AIDS 541 Prevention, Centers for Disease Control and Prevention, Atlanta, 30333 GA, USA for useful 542 discussions and suggestions. We thank Jason Ladner from the Pathogen and Microbiome 543 Institute, Northern Arizona University, Flagstaff, AZ for providing suggestions and feedback. We 544 also thank numerous anonymous reviewers who helped improve our manuscript by their valuable 545 comments on the manuscript.

546 Author Contributions

S.M. conceived of the presented idea and supervised the project. S.M., S. Kna, K.C., led the
project. S.M., S. Kna, K.C., S. Kar, D.D., P.I.B., Z.C., A.L., Y.P., T.I.V., J.O.W, B.T.T. C.C., R.S.,
A.W., M.S.A., V.M.P, S.H.N, A.L.S., P.S., A.Z., R.K., and K. Cran 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.

553

554 Funding

555 Serghei Mangul: SM was partially supported by National Science Foundation grants 2041984. 556 Tommy Lam: TL is supported by NSFC Excellent Young Scientists Fund (Hong Kong and Macau) 557 (31922087), RGC Collaborative Research Fund (C7144-20GF), RGC Research Impact Fund 558 (R7021-20), the Innovation and Technology Commission's InnoHK funding (D24H), and Health 559 and Medical Research Fund (COVID190223). Pavel Skums: PS was supported by the NIH grant 560 1R01EB025022 and NSF grant 2047828. Malak Abedalthagafi: MA acknowledges King Abdulaziz 561 City for Science and Technology and the Saudi Human Genome Project for technical and financial 562 support (https://shgp.kacst.edu.sa) Nicholas Wu: NW was supported by the US National Institutes 563 of Health (R00 AI139445, DP2 AT011966, R01 AI167910). Adam Smith: AS acknowledge funding 564 from NSF grant no. 2029025. Alex Zelikovsky: AZ has been partially supported by NIH Grants 565 1R01EB025022-01 and NIH Grant 1R21CA241044-01A1. Sergey Knyazev: SK has been partly 566 supported by Molecular Basis of Disease at Georgia State University, and NIH awards R01 567 HG009120, R01 MH115676, R01 AI153827, U01 HG011715. Aiping Wu: AW has been supported 568 by the CAMS Innovation Fund for Medical Sciences (2021-I2M-1-061). Rob Knight: RK was 569 supported by NSF project 2038509, RAPID: Improving QIIME 2 and UniFrac for Viruses to 570 Respond to COVID-19. CDC project 30055281 with Scripps led by Kristian Andersen, Genomic 571 sequencing of SARS-CoV-2 to investigate local and cross-border emergence and spread. Joel 572 O. Wertheim: JOW was supported by NIH-NIAID R01 AI135992, receives funding from CDC 573 unrelated to this work. Tetyana I. Vasylyeva: TIV is supported by the Branco Weiss Fellowship. 574 Yuri Porozov: YP was supported by the Ministry of Science and Higher Education of the Russian 575 Federation within the framework of state support for the creation and development of World-Class Research Centers "Digital biodesign and personalized healthcare" No075-15-2020-926. Eric 576 577 Bortz: E.B. was supported by a US NIGMS IDeA Alaska INBRE (P20GM103395) and NIAID 578 CEIRR (75N93019R00028). C.E.M. thanks Testing for America (501c3), OpenCovidScreen 579 Foundation, Igor Tulchinsky and the WorldQuant Foundation, Bill Ackman and Olivia Flatto and 580 the Pershing Square Foundation, Ken Griffin and Citadel, the US National Institutes of Health 581 (R01AI125416, R01AI151059, R21AI129851, U01DA053941), and the Alfred P. Sloan 582 Foundation (G-2015-13964). Charles Y. Chiu: CYC is supported by US CDC Epidemiology and 583 Laboratory Capacity (ELC) for Infectious Diseases Grant 6NU50CK000539 to the California 584 Department of Public Health, the Innovative Genomics Institute (IGI) at UC Berkeley and UC San 585 Francisco, National Institutes of Health grant R33AI12945. US Centers for Disease Control and 586 Prevention contract 75D30121C10991. Andrey Komissarov: AK was partly supported by RFBR 587 grant 20-515-80017. Philippe Lemey: PL acknowledges support from the European Research 588 Council under the European Union's Horizon 2020 research and innovation programme (grant 589 agreement no. ~725422 - ReservoirDOCS), the Wellcome Trust through project 206298/Z/17/Z 590 (Artic Network) and from NIH grants R01 AI153044 and U19 AI135995. Keith Crandall: KC 591 acknowledges support from the US NSF award EEID-IOS-2109688. Fyodor Kondrashov: FK's 592 work was supported by the ERC Consolidator grant to FAK (771209-CharFL).

593 Competing interests

594 The authors have no competing interests.

595 Figure Legend

596 Figure 1. Available SARS-CoV-2 genomic sequencing data and its usage for outbreak 597 investigation (a) The number of SARS-CoV-2 genomes sequenced according to Global Initiative 598 On Sharing All Influenza Data (GISAID) between January 2020 and December 2021. (b) The 599 number of available SARS-CoV-2 sequences in GISAID per 1 million (1M) individuals for each 600 country or region vs. the number of cases per capita up to March 2021. (c) The number of 601 available SARS-CoV-2 sequences in GISAID per 1 million (1M) individuals for each country in 602 Africa vs. the number of sequencers per capita up to March 2021. Blue line is the correlation of 603 all data points on the plot. (d) The number of available SARS-CoV-2 sequences in GISAID per 604 number of reported COVID-19 cases for each country or region vs. the number of reported 605 COVID-19 cases per capita from December 2019 up to December 2021. (e) Global outbreak 606 investigations by phylogenetic analysis (red) and wastewater studies (yellow), dots were placed 607 in the geographical centers of each country or region.

608

609

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.go v/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.o rg/
COVID-3D	Structural changes of lineages	http://biosig.unimelb.edu.a u/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	https://corona.bii.a-

	annotation tool identifying phenotypically or epidemiologically interesting candidate amino acid (aa) changes for further research	star.edu.sg/ https://www.gisaid.org/epifl u-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.cor nell.edu/

610

611 References

- Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health
 concern. The Lancet vol. 395 470–473 (2020).
- 614 2. Grubaugh, N. D. et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol 4, 10–
 615 19 (2019).
- 616 3. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic
 617 epidemiology. Nat Microbiol 5, 1403–1407 (2020).
- Karim, S. S. A. & Karim, Q. A. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19
 pandemic. Lancet 398, 2126–2128 (2021).
- 5. 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. https://www.rockefellerfoundation.org/news/the-rockefeller-foundation-releasesnew-action-plan-to-accelerate-development-of-a-national-system-for-gathering-and-sharinginformation-on-sars-cov-2-genomic-variants-and-other-pathogens/.
- 625 6. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, 626 China. Lancet 395, 497–506 (2020).
- Wu, F. et al. A new coronavirus associated with human respiratory disease in China. Nature 579,
 265–269 (2020).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search
 tool. J. Mol. Biol. 215, 403–410 (1990).
- Boni, M. F. et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for
 the COVID-19 pandemic. Nat Microbiol (2020) doi:10.1038/s41564-020-0771-4.
- Wang, H., Pipes, L. & Nielsen, R. Synonymous mutations and the molecular evolution of SARSCoV-2 origins. Virus Evol 7, veaa098 (2021).
- 11. Dolgin, E. The tangled history of mRNA vaccines. Nature 597, 318–324 (2021).
- 636 12. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data--from vision to
 637 reality. Eurosurveillance 22, 30494 (2017).
- An integrated national scale SARS-CoV-2 genomic surveillance network. The Lancet Microbe
 (2020) doi:10.1016/S2666-5247(20)30054-9.
- 640 14. Fernandes, J. D. et al. The UCSC SARS-CoV-2 Genome Browser. Nat. Genet. 52, 991–998

- 641 (2020).
- 15. Kuiken, C., Korber, B. & Shafer, R. W. HIV sequence databases. AIDS Rev. 5, 52–61 (2003).
- 643 16. Maxmen, A. One million coronavirus sequences: popular genome site hits mega milestone.
 644 Nature (2021) doi:10.1038/d41586-021-01069-w.
- Kalia, K., Saberwal, G. & Sharma, G. The lag in SARS-CoV-2 genome submissions to GISAID.
 Nat. Biotechnol. 39, 1058–1060 (2021).
- 18. Inzaule, S. C., Tessema, S. K., Kebede, Y., Ogwell Ouma, A. E. & Nkengasong, J. N. Genomicinformed pathogen surveillance in Africa: opportunities and challenges. Lancet Infect. Dis. (2021)
 doi:10.1016/S1473-3099(20)30939-7.
- 450 19. Van Noorden, R. Scientists call for fully open sharing of coronavirus genome data. Nature 590,
 451 195–196 (2021).
- Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution
 to global health. Glob Chall 1, 33–46 (2017).
- 654 21. Geoghegan, J. L. & Holmes, E. C. The phylogenomics of evolving virus virulence. Nat. Rev.
 655 Genet. 19, 756–769 (2018).
- 456 22. van Dorp, L. et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2.
 457 Infect. Genet. Evol. 104351 (2020).
- 23. Zhang, Y.-Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARSCoV-2. Cell 181, 223–227 (2020).
- 660 24. Korber, B. et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases
 661 Infectivity of the COVID-19 Virus. Cell (2020) doi:10.1016/j.cell.2020.06.043.
- 5. Tao, K. et al. The biological and clinical significance of emerging SARS-CoV-2 variants. Nat. Rev.
 663 Genet. 22, 757–773 (2021).
- Kemp, S. A. et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature (2021)
 doi:10.1038/s41586-021-03291-y.
- 666 27. Hay, A. J. & McCauley, J. W. The WHO global influenza surveillance and response system
 667 (GISRS)-A future perspective. Influenza Other Respi. Viruses 12, 551–557 (2018).
- 4123 (2018).
 4123 (2018).
- 670 29. Bull, R. A. et al. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome
 671 analysis. Nat. Commun. 11, 6272 (2020).
- 672 30. Laiton-Donato, K. et al. Genomic epidemiology of SARS-CoV-2 in Colombia. bioRxiv (2020)
 673 doi:10.1101/2020.06.26.20135715.
- 31. Yakovleva, A. et al. Tracking SARS-COV-2 Variants Using Nanopore Sequencing in Ukraine in
 Summer 2021. Res Sq (2021) doi:10.21203/rs.3.rs-1044446/v1.
- Tablizo, F. A. et al. Detection and Genome Sequencing of SARS-CoV-2 Variants Belonging to
 the B.1.1.7 Lineage in the Philippines. Microbiol Resour Announc 10, (2021).
- 678 33. Viana, R. et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa.
 679 bioRxiv (2021) doi:10.1101/2021.12.19.21268028.
- 680 34. Fauver, J. R. et al. Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the
 681 United States. Cell (2020) doi:10.1016/j.cell.2020.04.021.
- 35. Morel, B. et al. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. Mol. Biol. Evol. 38, 1777–
 1791 (2021).
- 684 36. Novikov, D. et al. Scalable Reconstruction of SARS-CoV-2 Phylogeny with Recurrent Mutations.
 685 J. Comput. Biol. 28, 1130–1141 (2021).
- 686 37. Steel, M. & Penny, D. Two further links between MP and ML under the poisson model. Appl.

687 Math. Lett. 17, 785–790 (2004).

- 38. Woolley, S. M., Posada, D. & Crandall, K. A. A comparison of phylogenetic network methods
 using computer simulation. PLoS One 3, e1913 (2008).
- 690 39. Wertheim, J. O., Steel, M. & Sanderson, M. J. Accuracy in near-perfect virus phylogenies. Syst.
 691 Biol. (2021) doi:10.1093/sysbio/syab069.
- 40. Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J. & Wertheim, J. O. HIV-TRACE
 (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and
 Other Rapidly Evolving Pathogens. Mol. Biol. Evol. 35, 1812–1819 (2018).
- 41. Knyazev, S. et al. A Novel Network Representation of SARS-CoV-2 Sequencing Data. in
 Bioinformatics Research and Applications 165–175 (Springer International Publishing, 2021).
- 697 42. Campbell, E. M. et al. MicrobeTrace: Retooling molecular epidemiology for rapid public health
 698 response. PLoS Comput. Biol. 17, e1009300 (2021).
- 43. Blair, C. & Ané, C. Phylogenetic Trees and Networks Can Serve as Powerful and Complementary
 Approaches for Analysis of Genomic Data. Syst. Biol. 69, 593–601 (2020).
- 44. Martin, M. A., VanInsberghe, D. & Koelle, K. Insights from SARS-CoV-2 sequences. Science 371,
 466–467 (2021).
- 45. Hodcroft, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020.
 Nature 595, 707–712 (2021).
- McNamara, R. P. et al. High-Density Amplicon Sequencing Identifies Community Spread and
 Ongoing Evolution of SARS-CoV-2 in the Southern United States. Cell Rep. 33, 108352 (2020).
- 707 47. Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S. & Stadler, T. The origin and early spread
 708 of SARS-CoV-2 in Europe. Proc. Natl. Acad. Sci. U. S. A. 118, (2021).
- 48. Worobey, M. et al. The emergence of SARS-CoV-2 in Europe and North America. Science (2020)
 doi:10.1126/science.abc8169.
- 49. Ladner, J. T. et al. An Early Pandemic Analysis of SARS-CoV-2 Population Structure and
 Dynamics in Arizona. MBio 11, (2020).
- 50. Gonzalez-Reiche, A. S. et al. Introductions and early spread of SARS-CoV-2 in the New York
 City area. Science (2020) doi:10.1126/science.abc1917.
- 51. Gámbaro, F. et al. Introductions and early spread of SARS-CoV-2 in France, 24 January to 23
 March 2020. Euro Surveill. 25, (2020).
- 52. du Plessis, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK.
 Science 371, 708–712 (2021).
- 53. Lemey, P. et al. Untangling introductions and persistence in COVID-19 resurgence in Europe.
 Nature 595, 713–717 (2021).
- 54. Lemey, P. et al. Accommodating individual travel history and unsampled diversity in Bayesian
 phylogeographic inference of SARS-CoV-2. Nat. Commun. 11, 5110 (2020).
- 55. Candido, D. D. S. et al. Routes for COVID-19 importation in Brazil. J. Travel Med. 27, (2020).
- 56. Badaoui, B., Sadki, K., Talbi, C., Salah, D. & Tazi, L. Genetic diversity and genomic epidemiology
 of SARS-CoV-2 in Morocco. Biosaf Health 3, 124–127 (2021).
- 726 57. Rockett, R. J. et al. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome
 727 sequencing and agent-based modeling. Nat. Med. 26, 1398–1404 (2020).
- 58. Nadeau, S. A. et al. Swiss public health measures associated with reduced SARS-CoV-2
 transmission using genome data. bioRxiv (2021) doi:10.1101/2021.11.11.21266107.
- 730 59. Nemira, A. et al. SARS-CoV-2 transmission dynamics in Belarus in 2020 revealed by genomic
 731 and incidence data analysis. Communications Medicine 1, 1–9 (2021).
- 60. Fahrenfeld, N. & Bisceglia, K. J. Emerging investigators series: sewer surveillance for monitoring

- antibiotic use and prevalence of antibiotic resistance: urban sewer epidemiology. Environmental
 Science: Water Research & Technology vol. 2 788–799 (2016).
- 61. Castiglioni, S., Senta, I., Borsotti, A., Davoli, E. & Zuccato, E. A novel approach for monitoring
 tobacco use in local communities by wastewater analysis. Tobacco Control vol. 24 38–42 (2015).
- 62. Sims, N. & Kasprzyk-Hordern, B. Future perspectives of wastewater-based epidemiology:
 Monitoring infectious disease spread and resistance to the community level. Environ. Int. 139, 105689 (2020).
- 63. Chen, Y. et al. The presence of SARS-CoV-2 RNA in the feces of COVID-19 patients. J. Med.
 Virol. 92, 833–840 (2020).
- 74264. COVID-19wastewaterepidemiologySARS-CoV-2.covid19wbec.org743https://www.covid19wbec.org.
- 65. Weidhaas, J. et al. Correlation of SARS-CoV-2 RNA in wastewater with COVID-19 diseaseburden in sewersheds. (2020).
- 66. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARSCoronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early
 Stage of the Epidemic in The Netherlands. Environ. Sci. Technol. Lett. 7, 511–516 (2020).
- Ahmed, W. et al. First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia:
 A proof of concept for the wastewater surveillance of COVID-19 in the community. Sci. Total
 Environ. 728, 138764 (2020).
- 68. Gonzalez, R. et al. COVID-19 surveillance in Southeastern Virginia using wastewater-based
 epidemiology. Water Res. 186, 116296 (2020).
- Peccia, J. et al. Measurement of SARS-CoV-2 RNA in wastewater tracks community infection
 dynamics. Nat. Biotechnol. (2020) doi:10.1038/s41587-020-0684-z.
- 756 70. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS757 Coronavirus-2 in sewage. doi:10.1101/2020.03.29.20045880.
- 758 71. Wu, F. et al. SARS-CoV-2 titers in wastewater foreshadow dynamics and clinical presentation of
 759 new COVID-19 cases. doi:10.1101/2020.06.15.20117747.
- 760 72. Karthikeyan, S. et al. High throughput wastewater SARS-CoV-2 detection enables forecasting of
 761 community infection dynamics in San Diego county. doi:10.1101/2020.11.16.20232900.
- 762 73. Farkas, K., Hillary, L. S., Malham, S. K., McDonald, J. E. & Jones, D. L. Wastewater and public
 763 health: the potential of wastewater surveillance for monitoring COVID-19. Current Opinion in
 764 Environmental Science & Health vol. 17 14–20 (2020).
- 765 74. Larsen, D. A. & Wigginton, K. R. Tracking COVID-19 with wastewater. Nat. Biotechnol. (2020)
 766 doi:10.1038/s41587-020-0690-1.
- 767 75. Schmidt, C. Watcher in the wastewater. Nat. Biotechnol. 38, 917–920 (2020).
- 768 76. Rothman, J. A. et al. RNA Viromics of Southern California Wastewater and Detection of SARS 769 CoV-2 Single-Nucleotide Variants. Appl. Environ. Microbiol. 87, e0144821 (2021).
- 770 77. R&d, D. P. et al. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol
 771 protocol metadata. protocols.io https://www.protocols.io/view/covid-19-artic-v3-illumina-library 772 construction-an-bgxjjxkn/metadata (2020).
- 773 78. Sharkey, M. E. et al. Lessons learned from SARS-CoV-2 measurements in wastewater. Sci. Total
 774 Environ. 798, 149177 (2021).
- 775 79. Turakhia, Y. et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time
 776 phylogenetics for the SARS-CoV-2 pandemic. Nat. Genet. 53, 809–816 (2021).
- 777 80. Freyja: Depth-weighted De-Mixing. (Github).
- 81. Karthikeyan, S. et al. Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant

- 779 transmission. bioRxiv (2021) doi:10.1101/2021.12.21.21268143.
- 82. Knyazev, S., Hughes, L., Skums, P. & Zelikovsky, A. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. Briefings in Bioinformatics (2020) doi:10.1093/bib/bbaa101.
- 83. Posada-Céspedes, S. et al. V-pipe: a computational pipeline for assessing viral genetic diversity
 from high-throughput data. Bioinformatics (2021) doi:10.1093/bioinformatics/btab015.
- 785 84. Jahn, K. et al. Detection of SARS-CoV-2 variants in Switzerland by genomic analysis of
 786 wastewater samples. medRxiv (2021).
- 85. Baaijens, J. A. et al. Variant abundance estimation for SARS-CoV-2 in wastewater using RNASeq quantification. medRxiv (2021) doi:10.1101/2021.08.31.21262938.
- 789 86. Caduff, L. et al. Inferring transmission fitness advantage of SARS-CoV-2 variants of concern in
 790 wastewater using digital PCR. bioRxiv (2021) doi:10.1101/2021.08.22.21262024.
- 791 87. Jahn, K. et al. Detection and surveillance of SARS-CoV-2 genomic variants in wastewater.
 792 bioRxiv (2021) doi:10.1101/2021.01.08.21249379.
- 793 88. Crits-Christoph, A. et al. Genome sequencing of sewage detects regionally prevalent SARS-CoV794 2 variants. doi:10.1101/2020.09.13.20193805.
- 89. Izquierdo Lara, R. W. et al. Monitoring SARS-CoV-2 circulation and diversity through community
 wastewater sequencing. Public and Global Health (2020) doi:10.1101/2020.09.21.20198838.
- 90. Nagy-Szakal, D. et al. Targeted Hybridization Capture of SARS-CoV-2 and Metagenomics
 Enables Genetic Variant Discovery and Nasal Microbiome Insights. Microbiol Spectr 9, e0019721
 (2021).
- 800 91. Carbo, E. C. et al. Coronavirus discovery by metagenomic sequencing: a tool for pandemic
 801 preparedness. J. Clin. Virol. 131, 104594 (2020).
- 802 92. Bedford, J. et al. A new twenty-first century science for effective epidemic response. Nature 575,
 803 130–136 (2019).
- 804 93. Center for Food Safety & Applied Nutrition. Wastewater surveillance for SARS-CoV-2 variants.
 805 U.S. Food and Drug Administration https://www.fda.gov/food/whole-genome-sequencing-wgs 806 program/wastewater-surveillance-sars-cov-2-variants (2021).
- 807 94. BioProject. https://www.ncbi.nlm.nih.gov/bioproject/757291.
- 808 95. CDC. National Wastewater Surveillance System (NWSS). Centers for Disease Control and
 809 Prevention https://www.cdc.gov/healthywater/surveillance/wastewater-surveillance/wastewater 810 surveillance.html (2021).
- 811 96. Hodcroft, E. B. et al. Want to track pandemic variants faster? Fix the bioinformatics bottleneck.
 812 Nature 591, 30–33 (2021).
- 813 97. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic
 814 epidemiology. Nat Microbiol 2020. Preprint] July 15, (2020).
- 815 98. Maxmen, A. Massive Google-funded COVID database will track variants and immunity. Nature
 816 (2021) doi:10.1038/d41586-021-00490-5.
- 817 99. Knyazev, S. et al. Accurate assembly of minority viral haplotypes from next-generation
 818 sequencing through efficient noise reduction. Nucleic Acids Res. (2021)
 819 doi:10.1093/nar/gkab576.
- 820 100. Sapoval, N. et al. SARS-CoV-2 genomic diversity and the implications for qRT-PCR diagnostics
 821 and transmission. Genome Res. 31, 635–644 (2021).
- 822 101. Lythgoe, K. A. et al. SARS-CoV-2 within-host diversity and transmission. Science 372, (2021).
- 823 102. Butler, D. et al. Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2
- 824 infection reveals unique host responses, viral diversification, and drug interactions. Nat. Commun.

825 12, 1660 (2021).

- 826 103. Kissler, S. M. et al. Viral Dynamics of SARS-CoV-2 Variants in Vaccinated and Unvaccinated
 827 Persons. N. Engl. J. Med. 385, 2489–2491 (2021).
- 104. MacKay, M. J. et al. The COVID-19 XPRIZE and the need for scalable, fast, and widespread
 testing. Nat. Biotechnol. 38, 1021–1024 (2020).
- 830 105. Danko, D. et al. A global metagenomic map of urban microbiomes and antimicrobial resistance.
 831 Cell 184, 3376–3393.e17 (2021).
- 832 106. Bogler, A. et al. Rethinking wastewater risks and monitoring in light of the COVID-19 pandemic.
- 833 Nature Sustainability (2020) doi:10.1038/s41893-020-00605-2.

834

