COVID-19 CG: Tracking SARS-CoV-2 mutations by locations and dates of interest

Albert Tian Chen[1], Kevin Altschuler, Shing Hei Zhan[2]*, Yujia Alina Chan[1]*, Benjamin E. Deverman[1]*

[1]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, United States of America.

[2]Department of Zoology & Biodiversity Research Centre, the University of British Columbia, Vancouver BC, Canada.

*Correspondence to: bdeverma@broadinstitute.org, alinac@broadinstitute.org, zhan@zoology.ubc.ca

Lead Contact

Benjamin E. Deverman
Stanley Center for Psychiatric Research
Broad Institute of MIT and Harvard
75 Ames St
Cambridge, MA 02142
Phone (626) 755-7523
bdeverma@broadinstitute.org

Abstract

COVID-19 CG is an open resource for tracking SARS-CoV-2 single-nucleotide variations (SNVs) and lineages while filtering by location, date, gene, and mutation of interest. COVID-19 CG provides significant time, labor, and cost-saving utility to diverse projects on SARS-CoV-2 transmission, evolution, emergence, immune interactions, diagnostics, therapeutics, vaccines, and intervention tracking. Here, we describe case studies in which users can interrogate (1) SNVs in the SARS-CoV-2 Spike receptor binding domain (RBD) across different geographic regions to inform the design and testing of therapeutics, (2) SNVs that may impact the sensitivity of commonly used diagnostic primers, and (3) the recent emergence of a dominant lineage harboring an S477N RBD mutation in Australia. To accelerate COVID-19 research and public health efforts, COVID-19 CG will be continually upgraded with new features for users to quickly and reliably pinpoint mutations as the virus evolves throughout the pandemic and in response to therapeutic and public health interventions.

Introduction

Since the beginning of the pandemic, SARS-CoV-2 genomic data has been accumulating at an unprecedented rate (90,000+ virus genomes as of September, 2020 on the GISAID database) (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017). Numerous countries have mobilized to sequence thousands of SARS-CoV-2 genomes upon the occurrence of local outbreaks, collectively and consistently contributing more than 10,000 genomes per month (**Figure S1A, B**). It is important to note that, despite the slow accumulation of potentially functional (nonsynonymous) mutations, there has been a steady increase in the number of variants with more than 6 nonsynonymous mutations compared to the WIV04 reference, an early isolate of SARS-CoV-2 that was collected in Wuhan in December, 2019 (**Figure S1C**). To evaluate the outcomes of anti-COVID-19 measures and detect keystone events of virus evolution, it is important to track changes in SARS-CoV-2 mutation and population dynamics in a location and date-specific manner. Indeed, several countries and the National Institutes of Health (NIH) have recognized how critical it is to collect SARS-CoV-2 genomic data to support contact tracing efforts and to inform public health decisions – these are paramount to the re-opening of countries and inter-regional travel (Collins 2020; Rockett et al. 2020; Oude Munnink, et al. 2020; Gudbjartsson et al. 2020; Pybus et al. 2020). Yet, the quantity and complexity of SARS-CoV-2 genomic data (and metadata) make it challenging and costly for the majority of scientists to stay abreast of SARS-CoV-2 mutations in a way that is meaningful to their specific research goals. Currently, each group or organization has to independently expend labor, computing costs, and, most importantly, time to curate and analyze the genomic data from GISAID before they can generate specific hypotheses about SARS-CoV-2 lineages and mutations in their population(s) of interest.

58 **Results**

59 To address this challenge, we built COVID-19 CoV Genetics (COVID-19 CG, covidcg.org), a
60 performant, interactive, and fully-scalable web application that tracks SARS-CoV-2 single-
61 nucleotide variants (SNVs) and lineages without sub-sampling. COVID-19 CG is a free, open
62 access interface that allows users to adapt analyses according to their dates and locations of
63 interest (**Figure 1A,B**; data processing workflow in **Figure S2**). Users can also select and
64 compare trends in SARS-CoV-2 lineage or SNV frequency across multiple locations (**Figure**
65 **1C**) as we will demonstrate using case studies. COVID-19 CG provides functionalities that, to
66 the best of our knowledge, cannot be found in other existing public browsers, and was designed
67 to empower these specific user groups:

68

69 ***Vaccine and therapeutics developers*** can inform the design and testing of their vaccine,
70 antibody, or small molecule by using COVID-19 CG to rapidly identify all of the variants in their
71 targeted SARS-CoV-2 protein or antigen, alongside the frequency of each variant in their
72 geographic location(s) of interest. Scientists can use COVID-19 CG to generate hypotheses and
73 experimentally determine whether the variants present in the location of vaccine/therapeutic
74 implementation may impact their product-specific interaction interface or antigen.

75

76 **Case study of SNVs in the receptor binding domain (RBD) of the SARS-CoV-2 Spike:**
77 Analyzing SNVs by geography and time is critical as the frequency of each SNV may vary
78 significantly across different regions over time. For instance, an S477N mutation in the RBD has
79 become dominant in Oceania (57.5% of Oceanian genotypes, all time) although it constitutes
80 only 4.36% of SARS-CoV-2 genotypes globally and has not been detected in Africa, Asia, or
81 South America (**Figure 2A**). SNV frequency in a given region can also shift over time, e.g., an
82 RBD N439K mutation not found in Ireland prior to July is now present in 79.5% of the genomes
83 collected mid-July through August (**Figure 2B**). Another rare RBD S477N mutation, which was

84 found in only 1.05% of the Australian SARS-CoV-2 sequences before June, now constitutes
85 more than 90% of the sequenced June through August isolates (**Figure 2C**). This geographical
86 and temporal variation is important to incorporate into the design and testing of therapeutic
87 antibodies (such as those under development as therapeutics by Regeneron that specifically
88 target the SARS-CoV-2 Spike RBD), as well as mRNA or recombinant protein-based vaccines.
89 This will help to assure developers of the efficacy of their therapeutics and vaccines against the
90 SARS-CoV-2 variants that are present in the intended location of implementation.
91
92 In addition, COVID-19 CG can be harnessed to track changes in SARS-CoV-2 evolution post-
93 implementation of therapeutics and vaccines. It will be crucial to watch for rare escape variants
94 that could resist drug- or immune-based interventions to eventually become the dominant
95 SARS-CoV-2 variant in the community. This need was particularly emphasized by a Regeneron
96 study that demonstrated that single amino acid variants could evolve rapidly in the SARS-CoV-2
97 Spike to ablate binding to antibodies that had been previously selected for their ability to
98 neutralize all known RBD variants; these amino acid variations were found either inside or
99 outside of the targeted RBD region, and some are already present at low frequency among
100 human isolates globally (Baum et al., 2020). The authors, Baum et al., suggested that these
101 rare escape variants could be selected under the pressure of single antibody treatment, and,
102 therefore, advocated for the application of cocktails of antibodies that bind to different epitopes
103 to minimize SARS-CoV-2 mutational escape. A recent study by Greaney et al. generated high-
104 resolution 'escape maps' delineating RBD mutations that could potentially result in virus escape
105 from neutralization by ten different human antibodies (Greaney et al., 2020). Based on lessons
106 learnt from the rise of multidrug resistant bacteria and cancer cells, it will be of the utmost
107 importance to continue tracking SARS-CoV-2 evolution even when multiple vaccines and
108 therapeutics are implemented in a given human population.
109

110 ***Diagnostics developers*** can evaluate their probe, primer, or point-of-care diagnostic according
111 to user-defined regional and temporal SARS-CoV-2 genomic variation. More than 665
112 established primers/probes are built into COVID-19 CG, and new diagnostics will be continually
113 incorporated into the browser. Users can also input custom coordinates or sequences to
114 evaluate their own target sequences and design new diagnostics.
115
116 **Case study of SNVs that could impact the sensitivity of diagnostic primers:** A recent
117 preprint alerted us to the finding that a common G29140T SNV, found in 22.3% of the study's
118 samples from Madera County, California, was adversely affecting SARS-CoV-2 detection by the
119 NIID_2019-nCoV_N_F2 diagnostic primer used at their sequencing center; the single SNV
120 caused a ~30-fold drop in the quantity of amplicon produced by the NIID_2019-nCov_N_F2/R2
121 primer pair (Vanaerschot et al., 2020). We used COVID-19 CG to detect other SNVs that could
122 impact the use of this primer pair, discovering that there are SARS-CoV-2 variants in several
123 countries with a different C29144T mutation at the very 3' end of the same NIID_2019-
124 nCoV_N_F2 primer (**Figure 3A**). As the authors of the preprint, Vanaerschot et al., noted, SNVs
125 could impact assay accuracy if diagnostic primers and probes are also being used to quantify
126 viral loads in patients. We found that at least ten other primer pairs could potentially be at risk in
127 different geographical regions due to SNVs that appear proximal to the 3' ends of primers
128 (**Figure 3B-K**): China-CDC-N-F and R; NIH, Thailand, WH-NIC N-F; US CDC 2019-nCoV-N1-R;
129 US CDC 2019-nCoV-N2-F; ARTIC-V3_nCoV-2019_11_RIGHT; ARTIC-V3_nCoV-
130 2019_13_LEFT; ARTIC-V3_nCoV-2019_34_LEFT; ARTIC-V3_nCoV-2019_39_LEFT (note that
131 the ARTIC primers are used for nanopore sequencing) (Tyson et al., 2020); WHO
132 N_Sarbarco_R1; and Institut Pasteur, Paris 12759Rv. We advocate that labs and clinics use
133 COVID-19 CG (https://covidcg.org) to check their most commonly used primers and probes
134 against the SARS-CoV-2 sequences that are prevalent in their geographic regions.
135

136 ***Researchers and public health professionals*** can use COVID-19 CG to gain insights as to
137 how the virus is evolving in a given population over time (e.g., in which genes are mutations
138 occurring, and do these lead to structural or phenotypic changes?). For example, users can
139 track D614G distributions across any region of interest over time. **Figure 4** shows a variety of
140 different D614G population dynamics in different areas. Nonetheless, we strongly caution
141 against inferring (i) chains or directionality of transmission and (ii) changes in the transmissibility
142 of any SARS-CoV-2 SNV based on population dynamics alone. Inconsistent sampling, sampling
143 biases, differences in founder host population traits (even median patient age), superspreading
144 events, regionally and temporally differential travel restrictions, and numerous other factors
145 instead of virus biological differences can influence the global distribution of SNVs (Grubaugh et
146 al., 2020).
147
148 **Case study of Australia's new dominant SARS-CoV-2 variant:** We discovered that the
149 SARS-CoV-2 Spike S477N mutation has become more prevalent in Australia (**Figure 5A**).
150 Globally, the S477N mutation was first detected in a single sample of lineage B.1.1.25 that was
151 collected on January 26, 2020 in Victoria, Australia; this is now the dominant SARS-CoV-2
152 variant in the region (**Figure 5B, C**). In particular, the set of SNVs that co-occur with the S477N
153 mutation in Australia (all time, as well as prior to May, 2020 before the most recent outbreak)
154 are different from the set of co-occurring SNVs in the United Kingdom (**Figure 5C**) —
155 suggesting that the S477N mutation occurred separately in the Australian and the UK lineages.
156 However, COVID-19 CG only reflects data contributed to GISAID. Variants of interest could be
157 present in other countries, but not yet known to the public because the sequencing centers in
158 those countries have not collected or deposited their data in GISAID. Furthermore, in instances
159 where only a singular, sporadic variant is detected (no sustained transmission), there is also the
160 possibility of sequencing error resulting in incorrect lineage assignment. Due to these caveats,
161 the genetic data must be used in combination with other types of data, such as from contact

162 tracing efforts, before it is possible to draw conclusions about the international transmission of
163 SARS-CoV-2 variants. In the case of the S477N variant that is now dominating in Australia, the
164 sequencing data alone indicate that the local transmission of this variant since January, 2020 in
165 Australia cannot be ruled out.
166
167 **Discussion**
168 COVID-19 CG (https://covidcg.org) was designed to be modular in order to continually integrate
169 datasets from other COVID-19 initiatives. We anticipate building modules for users to **(1)** map
170 emerging mutations onto structural interaction interfaces of interest (e.g., between virus protein
171 and therapeutic antibodies or host proteins) using existing and future structures on the Protein
172 Data Bank (PDB), **(2)** visualize mutations in isolates of interest in the context of different virus
173 protein phenotypes or mutational constraints of antibody epitopes according to emerging
174 genotype-to-phenotype maps (Greaney et al., 2020; Starr et al., 2020), **(3)** compare SARS-CoV-
175 2 mutations in different host species (e.g., humans versus minks) (Oude Munnink et al., 2020b),
176 **(4)** rapidly determine when and where each lineage or SNV has been detected around the
177 world, and **(5)** overlay important policy events or travel restrictions over time on the lineage or
178 SNV tracker to help guide user date range selection. In addition, as more detailed metadata is
179 generated by COVID-19 studies and initiatives, we will update the application to enable filtering
180 according to patient traits such as gender, age, ethnicity, and medical condition (e.g.,
181 symptoms, hospitalization).
182
183 COVID-19 CG (https://covidcg.org) was built to help scientists and professionals worldwide, with
184 varying levels of bioinformatics expertise, in their real-time analysis of SARS-CoV-2 genetic
185 data. We hope that COVID-19 CG will also motivate decision makers to sustain or accelerate
186 their sequencing of virus isolates in their geographical area for the purposes of informing
187 vaccine, therapeutics, and policy development. Collecting virus genomic data is particularly

188 relevant to regions that are experiencing increases in COVID-19 cases. If only sparse genomic
189 data are sampled, we risk the late detection of SARS-CoV-2 variants that exhibit enhanced
190 virulence or resistance against therapeutics or vaccination programs in these pandemic
191 hotspots. Furthermore, the widespread implementation of vaccines and antibody therapies
192 could stimulate the emergence and selection of new escape variants (Baum et al., 2020). To
193 compound these risks, SARS-CoV-2 transmission from humans to minks (and back into
194 humans) has already been detected at farms across the Netherlands, Denmark, Spain, and the
195 United States (Oude Munnink et al., 2020b). This process of species crossing, if left unchecked,
196 can result in the emergence of diverse SARS-CoV-2 variants.
197
198 Coordinated sequencing and contact tracing efforts (e.g., in the UK, Singapore, the
199 Netherlands, Italy, California, and Australia) emphasize the urgency of establishing open access
200 platforms to evaluate trends in virus introduction into each country or region in real time. Local
201 policymakers, public health researchers, and scientists can use **global** SARS-CoV-2 genetic
202 data, in complementation with contact tracing data, to better understand which lineages were
203 imported into their region (from which potential international locations), whether these were
204 introduced multiple times, and if particular lineages are dying out or persisting. Labs in
205 numerous countries are already making incredible efforts to sequence the SARS-CoV-2 variants
206 circulating in their local populations (**Figure 6**). When each country actively contributes to the
207 database of SARS-CoV-2 genomes, this protects against sampling biases that can impact the
208 ability to perform phylogenetic analysis and interpret global SARS-CoV-2 data. Towards this
209 goal that affects all of humanity, we advocate for the increased sequencing of SARS-CoV-2
210 isolates from patients (and infected animals) around the world, and for these data to be shared
211 in as timely a manner as possible.

212 **Experimental procedures**

213 **Data Pipeline**

214 Our data processing pipeline is written with the Snakemake scalable bioinformatics workflow
215 engine (Koster and Rahmann, 2012), which modularizes our workflow and enables
216 reproducibility and compatibility with cloud-computing. All code and relevant documentation are
217 hosted on an open-source, publicly available GitHub repository (https://github.com/vector-
218 engineering/COVID19-CG), providing example data for users to validate our pipeline.

219

220 **Data Acquisition**

221 SARS-CoV-2 sequences and metadata are downloaded daily from the GISAID EpiCov™
222 database (https://epicov.org/epi3/start), by navigating to the "Browse" tab and selecting
223 sequences by "Submission Date". As of 2020-09-04, only 10,000 sequences can be
224 downloaded from this selection at a time, so the selection is adjusted to include no more than
225 10,000 sequences. "Sequences", "Patient status metadata", and "Sequencing technology
226 metadata" are downloaded separately, stored in separate folders, and renamed for ingestion
227 into the data processing pipeline (see below).

228

229 **Sequence Preprocessing**

230 Based on best practices, we filter out sequences meeting any of the following criteria: (1)
231 Present on the NextStrain's exclusion list
232 (https://github.com/nextstrain/ncov/blob/master/defaults/exclude.txt), (2) Isolates from non-
233 humans (animals, environmental samples, etc), (3) genome length less than 29,700 nt, or (4)
234 >5% ambiguous base calls. Sequences which pass all preprocessing filters are carried onto the
235 next steps.

236

237 **Metadata Cleaning**

238 We clean metadata with the aim of preserving the original intent of the authors and data
239 submitters while presenting simpler and unified versions to end users. Sequencing metadata is
240 cleaned to remove obvious typos, and to unify labels with the same meaning, e.g., "MinION"
241 and "Nanopore MinION". Location metadata is cleaned with the goal of simplifying the location
242 selector in the sidebar. Locations with excessive children are collapsed to the nearest upper
243 hierarchical grouping. E.g., if a state has individual data for 200+ towns, these towns will be
244 collapsed to the county level in order to facilitate easier data browsing. Typos and clear
245 identities are also unified to prevent the display of duplicate locations in the application.
246
247 **SNV Assignments**
248 SNVs and insertions/deletions (indels) at the nucleotide and amino acid level are determined by
249 aligning each sequence to the WIV04 reference sequence (WIV04 is a high quality December,
250 2019 isolate that is 100% identical to the first publicly available SARS-CoV-2 genome reference
251 Wuhan-Hu-1/NC_045512.2, excepting the sequences at the end of the genomes) using
252 *bowtie2*. Spurious SNVs and probable sequencing errors, defined as less than 3 global
253 occurrences, are filtered out prior to downstream analysis. SNVs involving ambiguous base
254 calls (“N” in the original sequences) are ignored. Indels resulting in frameshifts are ignored, and
255 SNVs/indels occurring in non-protein-coding regions are ignored when determining SNVs/indels
256 on the AA level.
257
258 **Lineage/Clade Analysis**
259 Viral lineages, as defined by the *pangolin* tool (Rambaut et al., 2020), and clades (Tang et al.,
260 2020) are provided by GISAID. In accordance with *pangolin*, SNVs present in >90% of
261 sequences within each lineage/clade will be assigned as lineage/clade-defining SNVs.
262
263 **Application Compilation**

The web application is written in Javascript, and primarily uses the libraries React.js, MobX, and Vega. The code is compiled into javascript bundles by webpack. All sequence data is compressed and injected inline as JSON into the javascript bundle – no server is needed to serve data to end users. The compiled application files can then be hosted on any static server.

Application Deployment

COVID CG (https://covidcg.org) is hosted by Google Cloud Run. The application code is assembled into a Docker image (see Dockerfile), with a build environment (node.js) and deployment environment (NGINX).

Acknowledgments

B.D. and A.C. are supported by awards from the National Institute of Neurological Disorders and Stroke (UG3NS111689) and a Brain Initiative award funded through the National Institute of Mental Health (UG3MH120096) and from the Stanley Center for Psychiatric Research. Y.A.C. is supported by a Broad Institute SPARC award and the Stanley Center for Psychiatric Research. We gratefully acknowledge all of the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this resource is based.

Data Availability

All of the data shown in this manuscript and displayed on COVID CG (https://covidcg.org) are downloaded from the GISAID EpiCov™ database (https://www.gisaid.org). All code and relevant documentation are hosted on an open-source, publicly available GitHub repository (https://github.com/vector-engineering/COVID19-CG).

290 **Author Contributions**

291 Y.A.C., S.H.Z, and A.T.C. conceived the project and browser. B.E.D. supervised the work.
292 A.T.C. and K.A. developed the COVID CG web browser with input from all of the authors. S.H.Z.
293 advised the implementation of lineage and clade analysis. Y.A.C., B.E.D., and A.T.C. prepared
294 the figures, analyzed the data, and wrote the manuscript with input from all authors.

295

296 **Declaration of Interests**

297 Shing Hei Zhan is a Co-founder and Director of Bioinformatics at Fusion Genomics Corporation,
298 which develops molecular diagnostic assays for infectious diseases. The other authors declare
299 no competing interests.

300

301 **References**

302 Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, Giordano S, Lanza K, Negron N,
303 Ni M, Wei Y, Atwal GS, Murphy AJ, Stahl N, Yancopoulos GD & Kyratsous CA (2020) Antibody
304 cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual
305 antibodies. Science. Available at: http://dx.doi.org/10.1126/science.abd0831

306 Collins F (2020) Genome Data Help Track Community Spread of COVID-19. NIH Director's
307 Blog. Available at: https://directorsblog.nih.gov/2020/07/21/genome-data-helps-track-
308 community-spread-of-covid-19/

309 Dong E, Du H & Gardner L (2020) An interactive web-based dashboard to track COVID-19 in
310 real time. Lancet Infect. Dis. 20, 533–534.

311 van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M & Balloux F (2020) No evidence for
312 increased transmissibility from recurrent mutations in SARS-CoV-2. Genomics, 501.

313 Elbe S & Buckland-Merrett G (2017) Data, disease and diplomacy: GISAID's innovative
314 contribution to global health. Glob Chall 1, 33–46.

315 Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, Hilton SK, Huddleston J,
316 Eguia R, Crawford KHD, Dingens AS, Nargi RS, Sutton RE, Suryadevara N, Rothlauf PW, Liu
317 Z, Whelan SPJ, Carnahan RH, Crowe JE & Bloom JD (2020) Complete mapping of mutations to
318 the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. bioRxiv,
319 2020.09.10.292078. Available at: https://www.biorxiv.org/content/10.1101/2020.09.10.292078v1

320 Grubaugh ND, Hanage WP & Rasmussen AL (2020) Making Sense of Mutation: What D614G
321 Means for the COVID-19 Pandemic Remains Unclear. Cell. Available at:
322 https://linkinghub.elsevier.com/retrieve/pii/S0092867420308175

323 Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL,
324 Saemundsdottir J, Sigurdsson A, Sulem P, Agustsdottir AB, Eiriksdottir B, Fridriksdottir R,
325 Gardarsdottir EE, Georgsson G, Gretarsdottir OS, Gudmundsson KR, Gunnarsdottir TR,
326 Gylfason A, Holm H, Jensson BO, Jonasdottir A, Jonsson F, Josefsdottir KS, Kristjansson T,
327 Magnusdottir DN, le Roux L, Sigmundsdottir G, Sveinbjornsson G, Sveinsdottir KE, Sveinsdottir
328 M, Thorarensen EA, Thorbjornsson B, Löve A, Masson G, Jonsdottir I, Möller AD, Gudnason T,
329 Kristinsson KG, Thorsteinsdottir U & Stefansson K (2020) Spread of SARS-CoV-2 in the
330 Icelandic Population. N. Engl. J. Med. 382, 2302–2315.

331 Koster J & Rahmann S (2012) Snakemake--a scalable bioinformatics workflow engine.
332 Bioinformatics 28, 2520–2522. Available at: http://dx.doi.org/10.1093/bioinformatics/bts480

333 Langmead B, Trapnell C, Pop M & Salzberg SL (2009) Ultrafast and memory-efficient alignment
334 of short DNA sequences to the human genome. Genome Biology 10, R25. Available at:
335 http://dx.doi.org/10.1186/gb-2009-10-3-r25

336 Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, Kamga SK,
337 Schapendonk C, Pronk M, Lexmond P, van der Linden A, Bestebroer T, Chestakova I,
338 Overmars RJ, van Nieuwkoop S, Molenkamp R, van der Eijk AA, GeurtsvanKessel C, Vennema
339 H, Meijer A, Rambaut A, van Dissel J, Sikkema RS, Timen A, Koopmans M & Dutch-Covid-19
340 response team (2020) Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed
341 public health decision-making in the Netherlands. Nat. Med. Available at:
342 http://dx.doi.org/10.1038/s41591-020-0997-y

343 Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R,
344 van der Spek A, Tolsma P, Rietveld A, Brouwer M, Bouwmeester-Vincken N, Harders F, Honing
345 RH der, Wegdam-Blans MCA, Bouwstra R, GeurtsvanKessel C, van der Eijk A, Velkers F, Smit
346 L, Stegeman A, van der Poel WHM & Koopmans M (2020) Jumping back and forth:
347 anthropozoonotic and zoonotic transmission of SARS-CoV-2 on mink farms. bioRxiv,
348 2020.09.01.277152. Available at: https://www.biorxiv.org/content/10.1101/2020.09.01.277152v1

349 Pybus OG & Rambaut A with du Plessis L, Zarebski AE, Kraemer MUG, Raghwani J, Gutiérrez
350 B, Hill V, McCrone J, Colquhoun R, Jackson B, O'Toole A, Ashworth J, on behalf of the COG-
351 UK consortium (2020) Preliminary analysis of SARS-CoV-2 importation & establishment of UK
352 transmission lineages. Virological. Available at: https://virological.org/t/preliminary-analysis-of-
353 sars-cov-2-importation-establishment-of-uk-transmission-lineages/507

354 Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L & Pybus OG
355 (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic
356 epidemiology. Nat Microbiol. Available at: http://dx.doi.org/10.1038/s41564-020-0770-5

357 Rockett RJ, Arnott A, Lam C, Sadsad R, Timms V, Gray K-A, Eden J-S, Chang S, Gall M,
358 Draper J, Sim EM, Bachmann NL, Carter I, Basile K, Byun R, O'Sullivan MV, Chen SC-A,
359 Maddocks S, Sorrell TC, Dwyer DE, Holmes EC, Kok J, Prokopenko M & Sintchenko V (2020)

360 Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-
361 based modeling. Nat. Med. Available at: http://dx.doi.org/10.1038/s41591-020-1000-7

362 Shu Y & McCauley J (2017) GISAID: Global initiative on sharing all influenza data - from vision
363 to reality. Euro Surveill. 22. Available at: http://dx.doi.org/10.2807/1560-
364 7917.ES.2017.22.13.30494

365 Starr TN, Greaney AJ, Hilton SK, Crawford KHD, Navarro MJ, Bowen JE, Alejandra Tortorici M,
366 Walls AC, Veesler D & Bloom JD (2020) Deep mutational scanning of SARS-CoV-2 receptor
367 binding domain reveals constraints on folding and ACE2 binding. bioRxiv, 2020.06.17.157982.
368 Available at: https://www.biorxiv.org/content/10.1101/2020.06.17.157982v1

369 Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J & Lu J
370 (2020) On the origin and continuing evolution of SARS-CoV-2. Natl. Sci. Rev. 7, 1012–1023.

371 Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, Choi JH, Lapointe H,
372 Kamelian K, Smith AD, Prystajecky N, Goodfellow I, Wilson SJ, Harrigan R, Snutch TP, Loman
373 NJ & Quick J (2020) Improvements to the ARTIC multiplex PCR method for SARS-CoV-2
374 genome sequencing using nanopore. bioRxiv, 2020.09.04.283077. Available at:
375 https://www.biorxiv.org/content/10.1101/2020.09.04.283077v1

376 Vanaerschot M, Mann SA, Webber JT, Kamm J, Bell SM, Bell J, Hong SN, Nguyen MP, Chan
377 LY, Bhatt KD, Tan M, Detweiler AM, Espinosa A, Wu W, Batson J, Dynerman D, CLIAHUB
378 Consortium, Wadford DA, Puschnik AS, Neff N, Ahyong V, Miller S, Ayscue P, Tato CM, Paul S,
379 Kistler A, DeRisi JL & Crawford ED (2020) Identification of a polymorphism in the N gene of
380 SARS-CoV-2 that adversely impacts detection by a widely-used RT-PCR assay. bioRxiv,
381 2020.08.25.265074. Available at:
382 https://www.biorxiv.org/content/10.1101/2020.08.25.265074v1.full

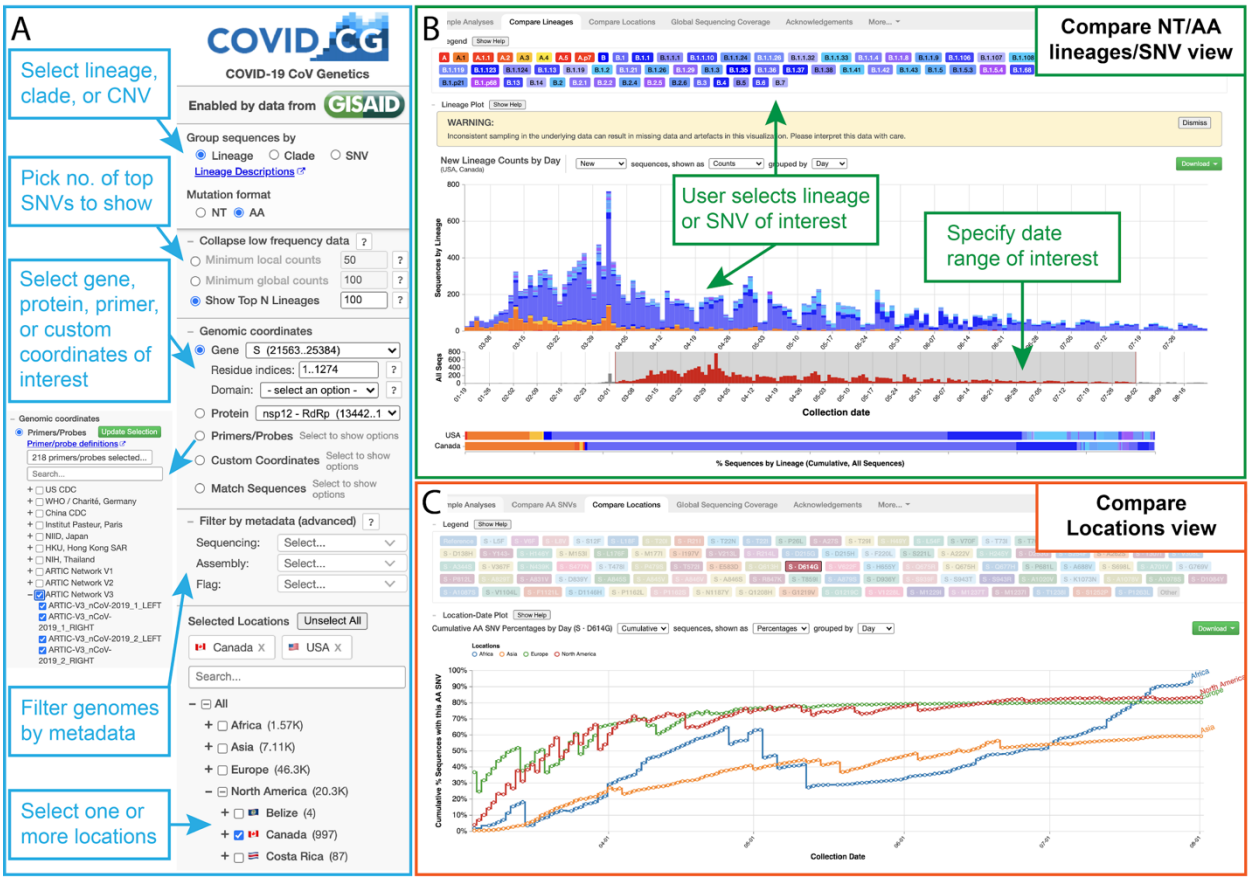


Figure 1. The COVID-19 CG (https://covidcg.org) browser interface. **(A)** Users can select SARS-CoV-2 genomes according to lineage, clade, or SNV, virus gene or protein, and location(s). Genomes can also be filtered by metadata, and specifically analyzed at genomic coordinates of interest, such as the target sites of 665 commonly used diagnostic primers and probes. **(B)** In the "Compare lineages or SNVs" tab, users can visualize SARS-CoV-2 lineages or SNVs by location, define their date range of interest, and see the corresponding SNVs at the nucleotide or amino acid level. **(C)** In the "Compare locations" tab, users can compare the frequencies of specific lineages or SNVs in multiple locations over time.

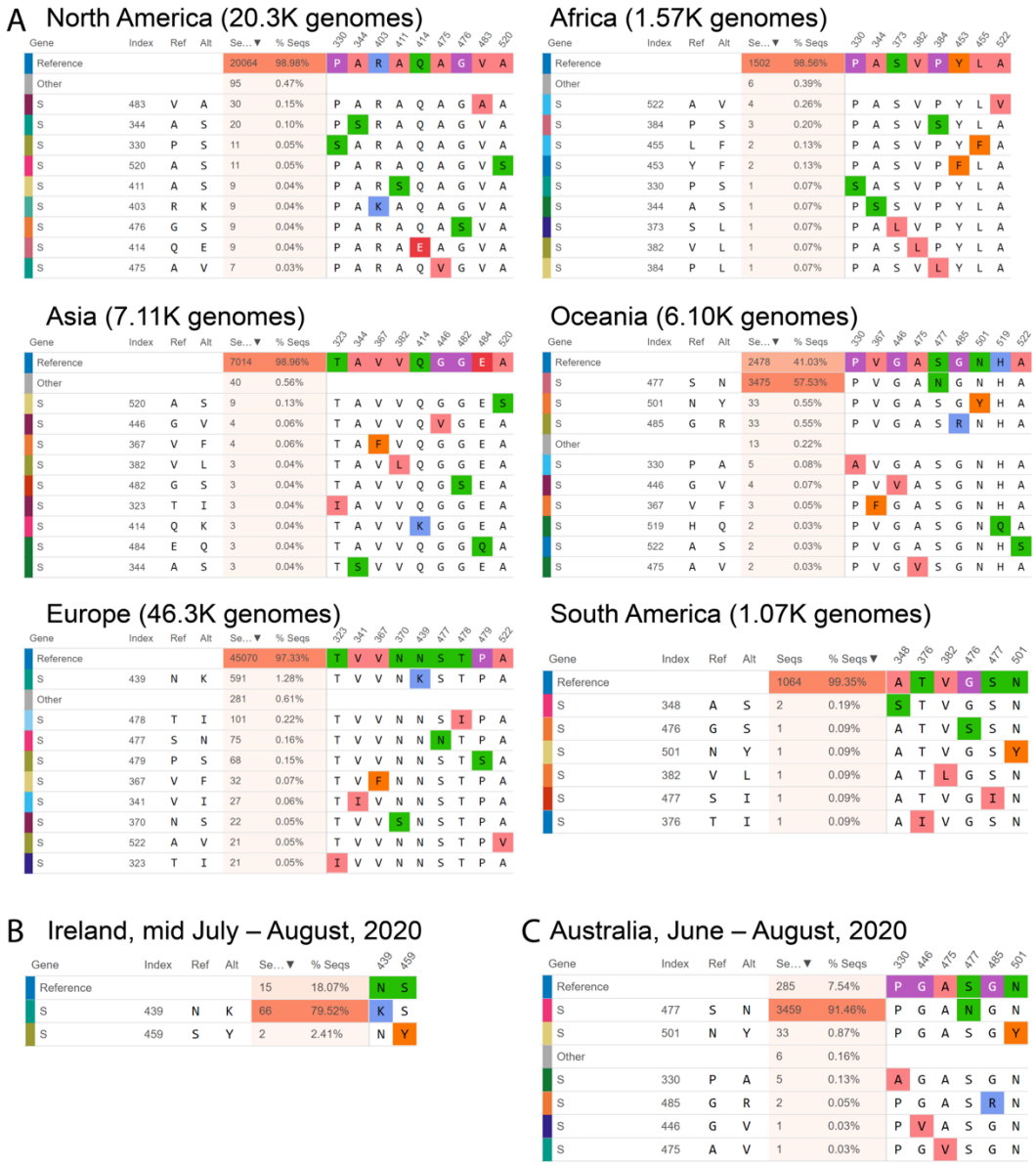


Figure 2. Mutational frequencies in the SARS-CoV-2 Spike receptor binding domain (RBD) across geographical location and time. Screen captures from the Compare AA SNVs tab are shown. **(A)** The top 10 RBD SNVs alongside the number of high quality sequences available on GISAID are shown for each region. **(B)** The top RBD SNVs for Ireland between mid July and August, 2020 are shown. The S439N mutant had not been previously detected in Ireland. **(C)** The top RBD SNVs for Australia between June and August, 2020 are shown. The S477N mutant constituted only 1.05% of the Australian SARS-CoV-2 genomes on GISAID prior to June.

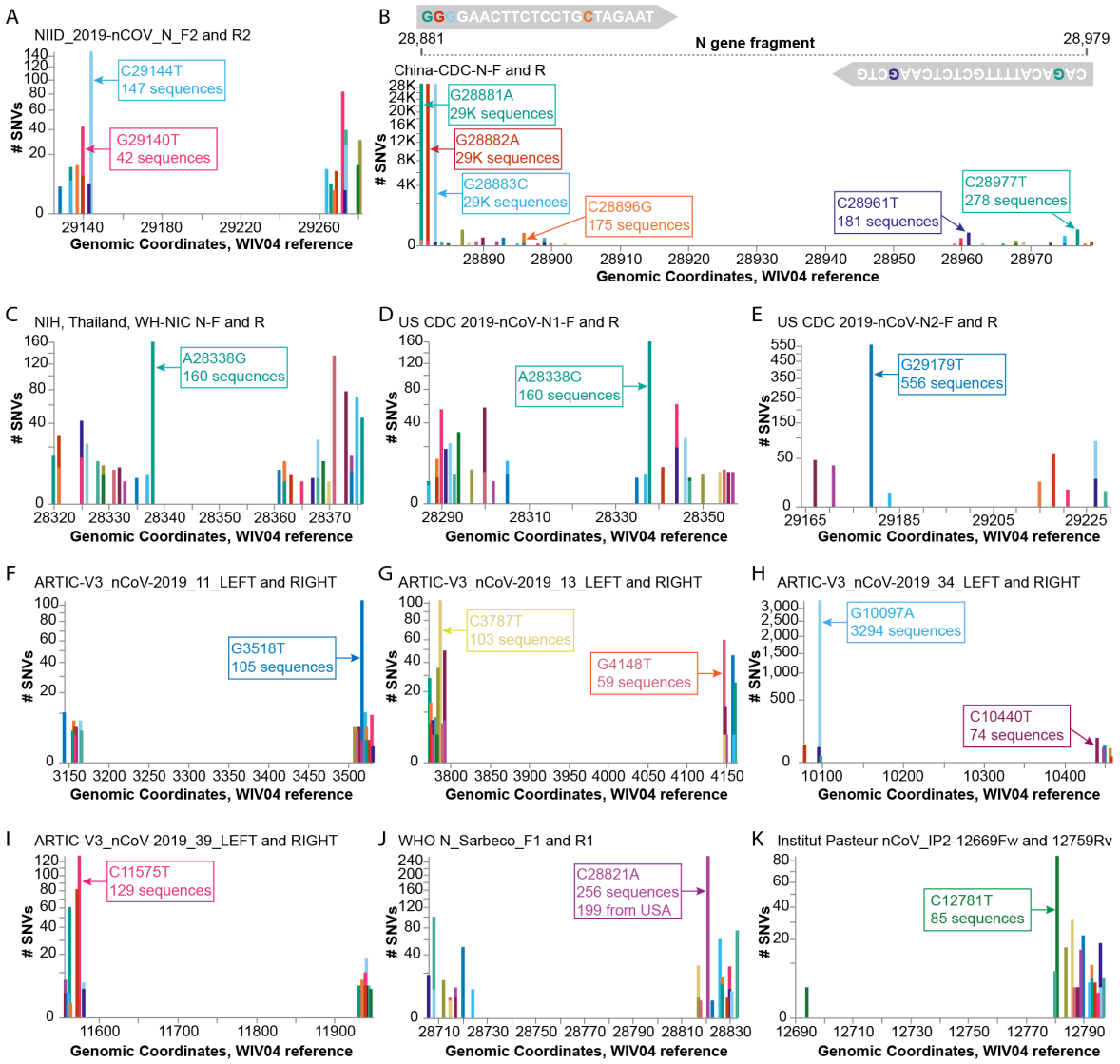


Figure 3. Investigating diagnostic-targeted regions of the SARS-CoV-2 genome for SNVs that could impact primer/probe sensitivity. Images were downloaded from the Compare NT SNVs tab. Labels for specific mutations were added. Primer pairs that contain at least one primer with potentially impactful SNVs near the 3' end are shown. None of the 11 primer pairs shown here were designed with degenerate bases. (**A)** The G29140T has been demonstrated to impact the NIID_2019-nCOV_N_F2 primer sensitivity. **(B-K)** Primer pairs affected by SNVs with a global frequency of more than 80 instances are shown. **(B)** As an example, majors SNVs are colored accordingly in the China-CDC-N-F and R (forward and reverse) primers.

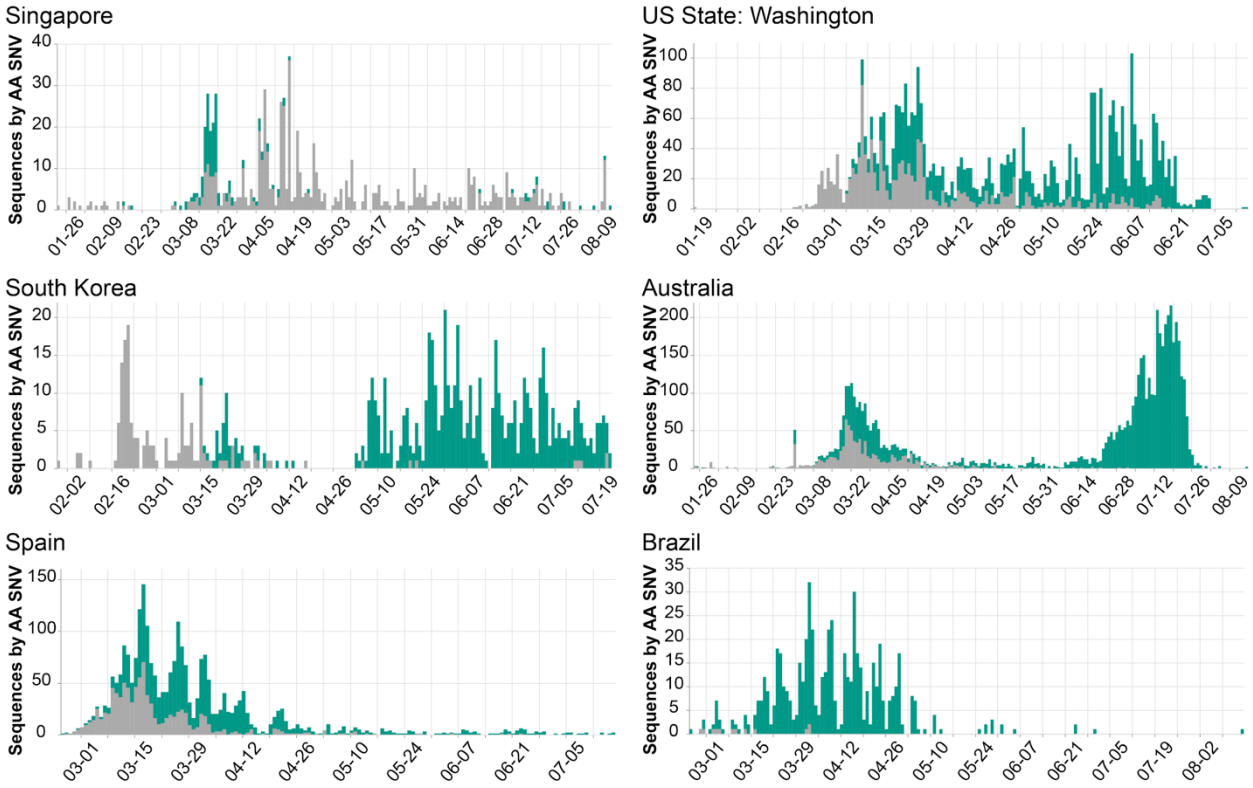


Figure 4. Population dynamics of Spike D614G in different regions. Images were downloaded from the Compare Lineages tab of covidcg.org: The Spike D614 variants are shown in grey, and the G614 variants are shown in green. Plots displaying different population dynamics were deliberately selected. Time is shown on the horizontal axis and the number of sequences is shown on the vertical axis; these differ per country depending on when and how many samples were collected and whether the sequences were deposited onto GISAID by August 31, 2020.

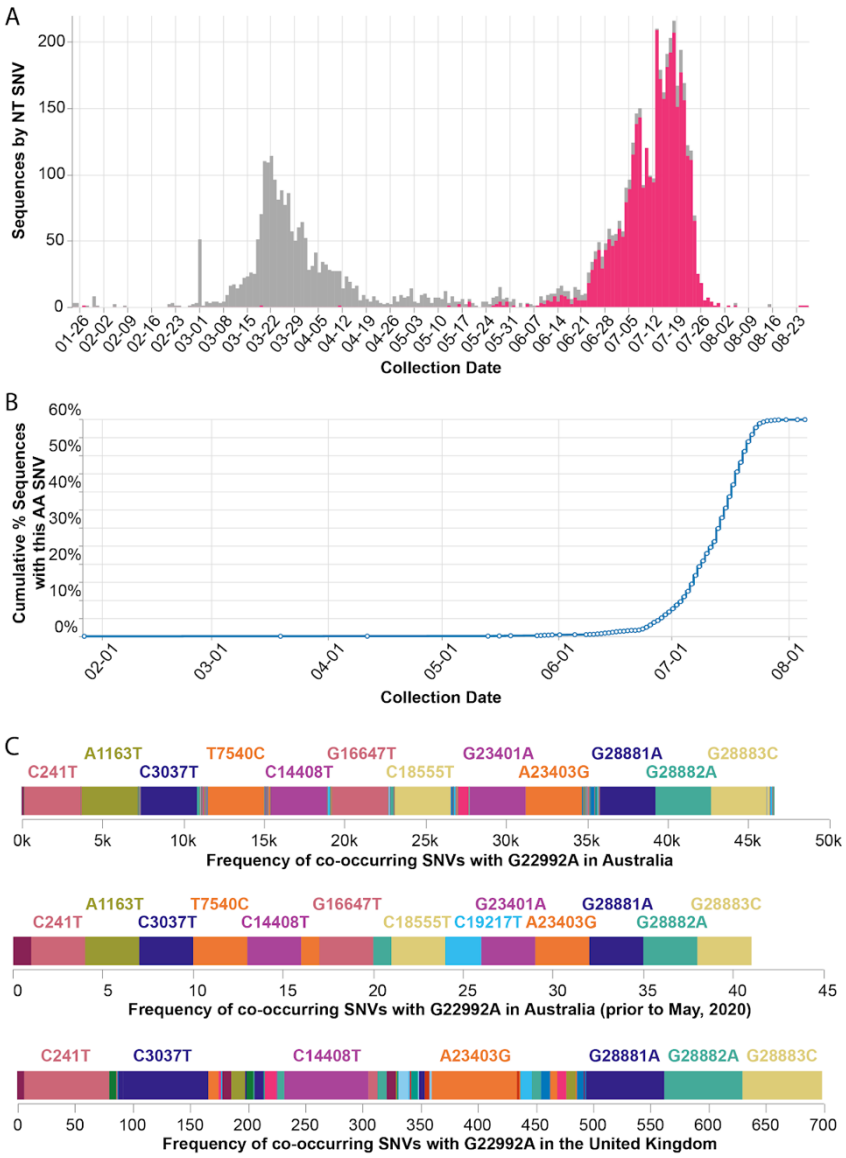


Figure 5. Frequency of the Spike S477N mutation in Australia over time. (A) Image downloaded from the Compare NT SNVs tab of covidcg.org: SARS-CoV-2 variants bearing the Spike S477N mutation (also known as the G22992A SNV; depicted in pink), the majority of which lie in the B.1.1.25 lineage, have become the most prevalent form of SARS-CoV-2 in Australia. **(B)** Image downloaded from the Compare Locations tab of covidcg.org: the cumulative percent of sequences carrying the S477N mutation in Australia. **(C)** Images downloaded from the Compare NT SNVs tab of covidcg.org: Co-occurring SNVs of G22992A (Spike S477N) in Australia, all time versus prior to May, 2020, versus in the United Kingdom.

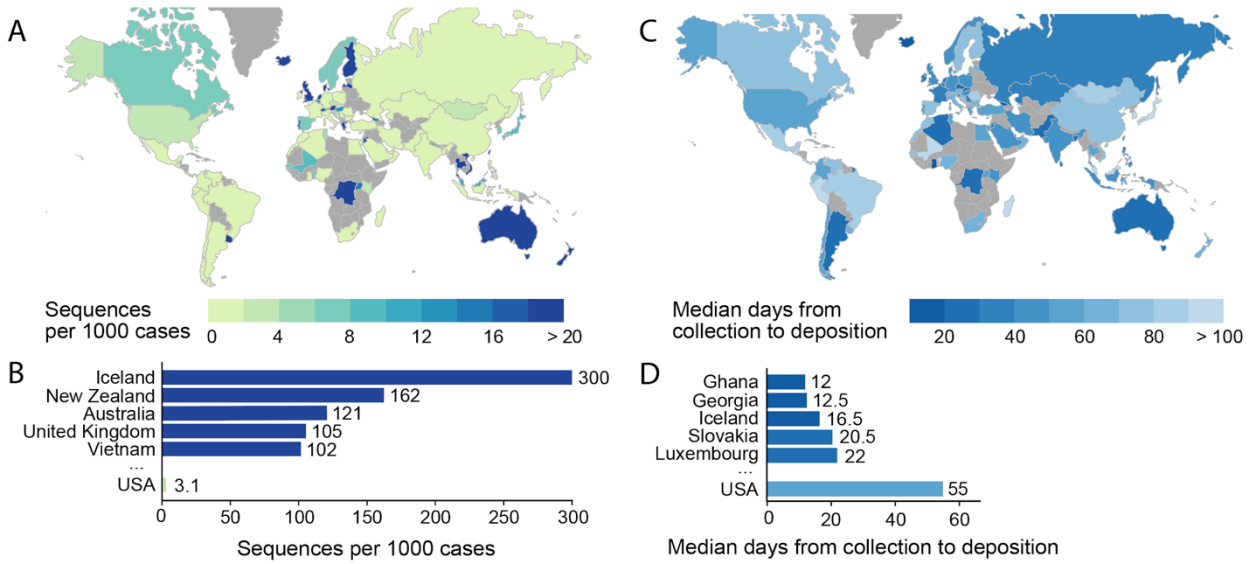
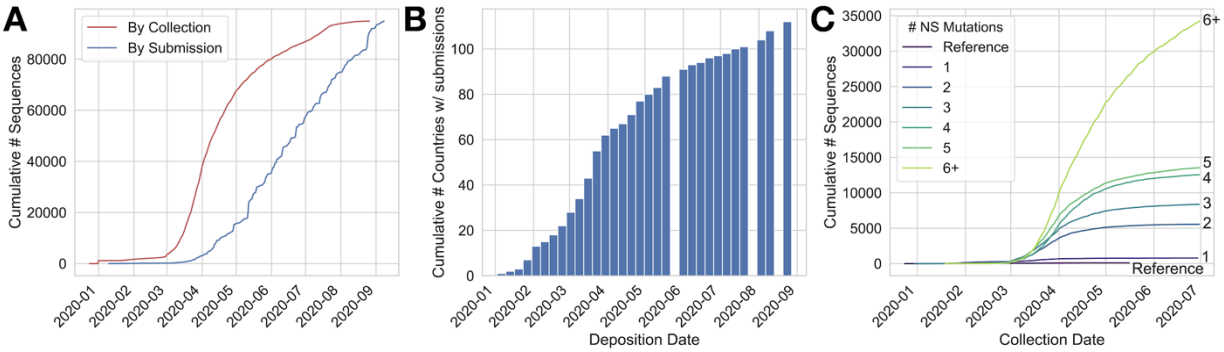


Figure 6. Global statistics of SARS-CoV-2 genomes contribution to GISAID. Interactive and more comprehensive versions of the figure panels are displayed on the Global Sequencing Coverage tab of covidcg.org. **(A)** A world map of countries labeled by the number of SARS-CoV-2 sequences contributed per 1000 cases. **(B)** A bar graph showing the sequences per 1000 cases for the top five countries and the USA. Countries with less than 500 cases were excluded from this plot. **(C)** A world map of countries labeled by median days between sample collection and sequence deposition. **(D)** A bar graph showing the median days from collection to deposition for the top five countries and the USA. These interactive displays are generated using sequencing data from the GISAID EpiCov™ database (nextmeta file) and case data from the JHU CSSE COVID-19 Data (Dong et al., 2020). Only samples that were collected between March and May, 2020 were included to avoid biases from samples that have been collected in the previous three months but not yet deposited onto GISAID.

435 **Supplemental Figures**



436 **Figure S1. The number of global SARS-CoV-2 genome sequences and mutations is**

437 **accumulating.** Data shown as of September 9, 2020. **(A)** Sequence deposition in GISAID

438 continues at a steady pace, albeit there is a lag between collection (red line) and submission

439 date (blue line). The rate of sequence submission is steady at >10,000 genomes per month. **(B)**

440 More than 100 countries have deposited SARS-CoV-2 genomes in GISAID. **(C)** The number of

441 SARS-CoV-2 variants with more than six nonsynonymous (NS) mutations continues to increase.

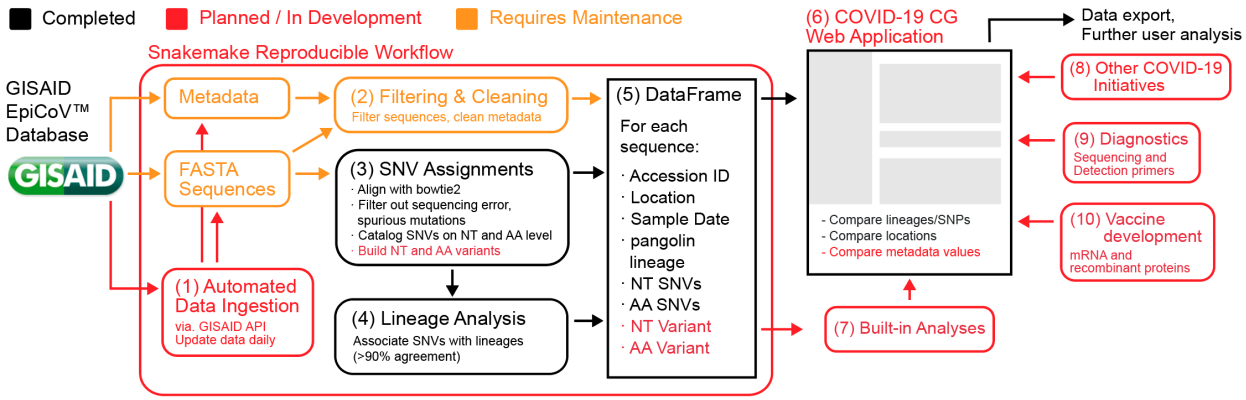


Figure S2. COVID-19 CG computational workflow. **(1)** Starting from the GISAID database, sequences are continuously updated, manually for now, but ultimately via automated data ingestion. **(2)** Based on best practices, we filter out sequences on NextStrain's exclusion list, non-human isolates, <29,700 nt, or with >5% ambiguous base calls (van Dorp et al., 2020). **(3)** SNVs at the nucleotide and amino acid level are determined by aligning (via *bowtie2*) each sequence to the WIV04 reference, a high quality December, 2019 isolate recommended by GISAID; NextStrain uses the 100% identical Wuhan-Hu-1 (Langmead et al., 2009). Importantly, spurious SNVs and probable sequencing errors are filtered out prior to downstream analysis. **(4)** Viral lineages, defined by the *pangolin* tool, are provided by GISAID. In accordance with *pangolin*, SNVs present in >90% of sequences within each lineage are assigned as lineage-defining SNVs. **(5)** The curated data and metadata, SNVs, and lineage-assigned SNVs are associated with their respective sequence identifier and compiled into a compact data set. **(6)** These data are uploaded onto the COVID-19 CG web application. **(7)** New analyses will be built into the COVID-19 CG application throughout the course of the pandemic. **(8-10)** Features and modules that integrate knowledge from other COVID-19 initiatives are continuously incorporated into COVID-19 CG.