

1 **Title:**
2 Pandemic dynamics of COVID-19 using epidemic stage, instantaneous reproductive number
3 and pathogen genome identity (GENI) score: modeling molecular epidemiology
4

5
6 **Running Head:** Integrated molecular epidemiology and modeling
7

8 **Authors:**
9 DJ Darwin R. Bandy^{1,2} and Bart C. Weimer^{1*}
10

11 **Affiliations:**
12 ¹University of California Davis, School of Veterinary Medicine, 100K Pathogen Genome Project,
13 Davis, CA 95616, USA; ²University of the Philippines Los Baños, College of Veterinary
14 Medicine, Department of Veterinary Paraclinical Sciences, Laguna 4031, Philippines
15

16
17
18 ***corresponding author:** bcweimer@ucdavis.edu; +1-530-760-9550
19
20

21
22 **Key words:**
23 reproductive number, COVID-19, corona virus, infectious disease, genomic epidemiology
24
25

26 **Research in context**

27 Reproductive number is (R) an epidemiological parameter that defines outbreak transmission
28 dynamics. While early estimates of R exist for COVID-19, the sample size is relatively small
29 (<2000 individuals) taken during the early stages of the disease in China. The outbreak is now a
30 pandemic and a more comprehensive assessment is needed to guide public health efforts in
31 making informed decisions to control regional outbreaks. Commonly, R is computed using a
32 sliding window approach, hence assessment of impact of intervention is more difficult to
33 estimate and often underestimates the dynamic nature of R as the outbreak progresses and
34 expands to different regions of the world. Parallel to epidemiological metrics, pathogen whole
35 genome sequencing is being used to infer transmission dynamics. Viral genome analysis
36 requires expert knowledge in understanding viral genomics that can be integrated with the rapid
37 responses needed for public health to advance outbreak mitigation. This study establishes
38 integrative approaches of genome sequencing with established epidemiological outbreak
39 metrics to provide an easily understandable estimate of transmission dynamics aimed at public
40 health response using evidence-based estimates.

41 42 **Added value of this study**

43 Estimates of R are dynamic within the progression of the epidemic curve. Using the framework
44 defined in this study with dynamic estimates of R specific to each epidemic stage combined with
45 whole genome sequencing led to creation of a novel metric called GENI (pathogen genome
46 identity) that provides genomic evolution and variation in the context of the outbreak dynamics.
47 The GENI scores were directly linked and proportional to outbreak changes when using disease
48 incidence from epidemic stages (index, takeoff, exponential, and decline). By simulating short
49 and standard (2 day and 7 day, respectively) serial intervals, we calculated instantaneous R
50 followed by a global comparison that was associated with changes in GENI. This approach
51 quantified R values that are impacted by public health intervention to change the outbreak

52 trajectory and were linked to case incidence (i.e. exponential expansion or decelerating) by
53 country. Integrating viral whole genome sequences to estimate GENI we were able to infer
54 circulation time, local transmission, and index case introduction. Systematic integration of viral
55 whole genome sequences with epidemiological parameters resulted in a simplified approach in
56 assessing the status of outbreak that facilitates decisions using evidence from genomics and
57 epidemiology in combination.

58

59 **Implications of all the available evidence**

60 This study created a framework of evidence-based intervention by integrating whole genome
61 sequencing and epidemiology during the COVID-19 pandemic. Calculating instantaneous R at
62 different stages of the epicurve for different countries provided an evidence-based assessment
63 of control measures as well as the underlying genomic variation globally that changed the
64 outbreak trajectory for all countries examined. Use of the GENI score translates sequencing
65 data into a public health metric that can be directly integrated in epidemiology for outbreak
66 intervention and global preparedness systems.

67

68 **Abstract**

69 **Background:** Global spread of COVID-19 created an unprecedented infectious disease

70 crisis that progressed to a pandemic with >180,000 cases in >100 countries. Reproductive
71 number (R) is an outbreak metric estimating the transmission of a pathogen. Initial R values
72 were published based on the early outbreak in China with limited number of cases with whole
73 genome sequencing. Initial comparisons failed to show a direct relationship viral genomic
74 diversity and epidemic severity was not established for SARS-Cov-2.

75 **Methods:** Each country's COVID-19 outbreak status was classified according to epicurve
76 stage (index, takeoff, exponential, decline). Instantaneous R estimates (Wallinga and Teunis
77 method) with a short and standard serial interval examined asymptomatic spread. Whole
78 genome sequences were used to quantify the pathogen genome identity score that were used

79 to estimate transmission time and epicurve stage. Transmission time was estimated based on
80 evolutionary rate of 2 mutations/month.

81 **Findings:** The country-specific R revealed variable infection dynamics between and within
82 outbreak stages. Outside China, R estimates revealed propagating epidemics poised to move
83 into the takeoff and exponential stages. Population density and local temperatures had variable
84 relationship to the outbreaks. GENI scores differentiated countries in index stage with cryptic
85 transmission. Integration of incidence data with genome variation directly increases in cases
86 with increased genome variation.

87 **Interpretation:** R was dynamic for each country and during the outbreak stage. Integrating
88 the outbreak dynamic, dynamic R, and genome variation found a direct association between
89 cases and genome variation. Synergistically, GENI provides an evidence-based transmission
90 metric that can be determined by sequencing the virus from each case. We calculated an
91 instantaneous country-specific R at different stages of outbreaks and formulated a novel metric
92 for infection dynamics using viral genome sequences to capture gaps in untraceable
93 transmission. Integrating epidemiology with genome sequencing allows evidence-based
94 dynamic disease outbreak tracking with predictive evidence.

95
96 **Funding:** Philippine California Advanced Research Institute (Quezon City, Philippines) and the
97 Weimer laboratory.

98 99 **Introduction**

100 Outbreaks are defined by the reproductive number (R)^{1,2} a common measure of
101 transmission. Probability of further disease spread is evaluated based on the threshold value
102 with likely expansion for values >2 and decline with values of <1. R is the main component for
103 computing the needed proportion of the population to be vaccinated based on herd immunity³.
104 The expansion of COVID-19 was determined with the earliest estimate of R = 2.2 (95% CI, 1.4

105 to 3.9) using serial intervals for 424 patients in Wuhan, China⁴. Recalculation with 2033 cases
106 estimated $R = 2.2$ to 3.6 ⁵. However, estimates of R for other countries where cases were found
107 as the outbreak grew in China were not done routinely and currently a fixed estimate R is used
108 based on the refined estimate from China. However, this is falling short in predicting the spread
109 of the pandemic and expansion within individual locations, suggesting that R is not likely to be
110 constant and likely to be dynamic for each outbreak location that results in underestimates of
111 the spread rate. This limitation is hindering epidemic dynamics as previously noted due to the
112 parameter is context specific and dynamic^{1,2}. Hence, there is a need to rapidly estimate country
113 specific R values during the epidemic. This will provide global comparisons of expansion at
114 each location.

115 The Wallinga and Teunis method for R estimation requires input of outbreak incidences and
116 the serial interval (i.e. the period between the manifestation of symptoms in the primary case
117 and the onset of symptoms in secondary cases)⁶. This approach was implemented in a web
118 resource to estimate R during epidemics⁷. A key advantage of the approach is the ease of
119 production of credible intervals compared to other maximum likelihood estimation approaches.
120 Yet to be done is integration of viral genetic variation with R estimates but one study found that
121 there was no obvious relationship between R , severity of the epidemic and COVID-19 genome
122 diversity²⁰.

123 COVID-19 has reached global spread in all continents except Antarctica and was defined to
124 be a pandemic by the World Health Organization (WHO) in March 2020⁸⁻¹⁰. The outbreak
125 dynamics are different between countries as well as varying within individual countries. In part
126 this is due to varying and diverse healthcare systems, socio-cultural contexts, and rigorous
127 testing. Considering the lack of containment globally, except in Singapore, Hong Kong, and
128 Taiwan, we hypothesized that previously calculated R values do not provide reliable estimates
129 because they are more dynamic than is being considered and that influx of new cases and viral
130 mutation are likely sustaining expansion. While viral sequencing is occurring, it is not being

131 effectively integrated with epidemiological information because there is no existing framework
132 for that to systematically occur.

133 In spite of no clear path for deep integration of viral variation the current pandemic has
134 demonstrated the public health unity for sharing COVID-19 whole genome sequences with an
135 unprecedented openness. By quickly sharing the genome sequences it enables investigation of
136 the genome variation during the outbreak using multiple approaches and samples of the virus
137 genome space. It is approaching a viral population scale, which provides additional information
138 that cannot be gleaned with few sequences. Prior work established the value of estimating
139 transmission dynamics of rapidly evolving RNA viruses and highlights the capability to infer
140 transmission during outbreaks coupled with pathogen genomes^{11,12}. This approach was
141 validated in EBOV and MERS. Each virus variant is separated by only several mutations yet
142 produces new dynamics during the outbreak^{13,14}. Rapidly evolving pathogens undergo genome
143 sequence mutation, selection pressure, random drift and stochastic events between infected
144 individuals¹¹. Even small changes in the genome enable transmission that is determined by
145 accounting for the mutations between isolate sequences. It is recognized that the COVID-19
146 genome is changing over the outbreak but there is controversy about the impact and specifics of
147 the exact mutations. In this study, we used incidence data to derive R and compared country
148 specific COVID-19 infection dynamics with viral population genome diversity. By incorporating
149 R, epidemic curve timing, and viral genome diversity we created a systematic framework that
150 deduced how viral genome diversity can be used to describe epidemiological features of an
151 outbreak before new cases were observed. This was done by creating a genome diversity
152 metric that was directly and systematic integrated to provide context and allowed quantification
153 of the infection dynamics globally that are divergent from the early estimates with genomic
154 evidence. We call this approach pathogen genome identity (GENI) scoring system. Using GENI
155 differentiated each stage of the outbreak. It also indicated cryptic local transmission from

156 surveillance systems. This a defining advantage of using sequences as previous cryptic
157 transmission can be inferred in the genomic sequences.

158
159 **Methods**

160 Incidence data is based on daily Chinese CDC and WHO situations reports as compiled by
161 the Center for Systems Science and Engineering (CSSE) by the John Hopkins University
162 (Baltimore, MD, USA) that was accessed on March 1, 2020¹⁵. We constructed epidemic curves
163 or epicurves from the incidence data and classified country status accordingly. We defined four
164 groups that characterize increasing expansion with a decline phase.

165 The extracted time series case data served as the input for determining instantaneous
166 reproductive number on a daily basis to effectively capture dynamic changes due to new
167 detected cases and reduction of cases due to social distancing and nonpharmaceutical
168 interventions. The prior value for R was selected at 2 and prior standard deviation of 5 to allow
169 fluctuations in reporting of cases in the exponential phase. As there is limited access to
170 epidemiological data of case, parametric with uncertainty (offset gamma) distributional estimate
171 of serial interval was used. A mean of 2 and 7 days, with standard deviation of 1 was used to
172 capture short and standard serial interval assumptions using 50 samplings of serial interval
173 distribution. The Wallinga and Teunis method, as implemented by Ferguson⁷ is a likelihood-
174 based estimation procedure that captures the temporal pattern of effective reproduction
175 numbers from an observed epidemic curve. R was calculated using the web application
176 EpiEstim App (<https://shiny.dide.imperial.ac.uk/epiestim/>)⁷. The descriptive statistics were used
177 to compute mean and confidence intervals of the instantaneous reproductive number.

178 GENI score was anchored on the principle of rapid pathogen evolution between
179 transmission events. This requires defining a suitable reference sequence of the outbreak,
180 which is on the early stages the sequence nearest to the timepoint of the index case. For the
181 case of COVID-19, the reference sequence is Wuhan seafood market pneumonia virus isolate
182 Wuhan-Hu-1 NC_045512.2¹⁶. Publicly available virus sequences were retrieved from GISAID

183 (supplementary Table 1) with whole genome variant determination using Snippy v4.6.0¹⁷⁻¹⁹. The
184 average mutation per isolate was divided to the total epidemic curve days to derive a daily
185 epidemic mutation rate and scaled to a monthly rate. We calculated the average nucleotide
186 change per month to be 1.7 (95% CI 1.4-2.0), which was within boundaries of another estimate
187 with the substitution rate of 0.9×10^{-3} (95% CI $0.5-1.4 \times 10^{-3}$) substitutions per site per year²⁰.
188 We derived a transformed value of this rate before integrating it with epidemiological
189 information. The output from the variant calling step was then used to determine GENI score by
190 calculating the nucleotide difference. The basis for GENI score cutoffs to estimate transmission
191 dates are derived from accepted evolutionary inference of mutation rates of COVID-19.

192 We defined four epidemic curve (epicurves) stages to provide a clear method to define
193 increases in the outbreak. The 'index stage' is characterized by the first report (index case) or
194 limited local transmission indicated by intermittent zero incidence creating undulating epicurve.
195 Secondly, which is distinctly different from stage 1, is the 'takeoff stage' in which the troughs are
196 almost at same level of the previous peak and no longer touches zero, suggesting sustained local
197 transmission. The 'exponential stage' is characterized by the classical hockey stick like sharp
198 uptrend where the outbreak is moving quickly and large number of new cases are emerging. The
199 last stage is 'decline' and is noted when the outbreak has reached the peak and cases being
200 reported are lower than the peak, which will ultimately result in few to no new cases being
201 reported, yet viral circulation is likely still occurring.

202 **Results**

203 We determined the outbreak dynamics of pandemic COVID-19 by classifying each country's
204 status according to epicurve stages with a framework of a) index b) takeoff c) exponential
205 d) decline as a clear method that can be used to benchmark metrics that include R and viral
206 genome diversity. First, we calculated R using the instantaneous method using two serial
207 intervals (2 and 7 days; Table 1). As of March 1, 2020, this framework defined global epicurves
208 of COVID-19 outbreaks as gaining momentum globally with 52 countries were in the index

209 stage. Three countries were in the exponential stage and five countries in the takeoff stage
210 (Figure 1). China was the only country that reached the peak of the epicurve and characterized
211 to be in the decline stage - decreasing cases. At this point there was no evidence of any other
212 country near the decline stage and some countries that were poised to move into the takeoff
213 and exponential phase.

214 Instantaneous R sensitively described real-time shifts of COVID-19 incidence captured
215 within each epicurve stage (Figure 2). The decline stage in China was reflected by a decrease
216 in R estimates in the latter stages the outbreak and relative to the early estimates: 1.6 (95 % CI
217 0.4-2.9) and 1.8 (95 % CI 1.0-2.7) for 2- and 7-days serial interval, respectively. Superspreading
218 events inflated R estimates seen in exponential stage that was observed in South Korea: 2.8
219 (95% CI 0.6-5.3) and 25.6 (95 % CI 3.0-48.2) for 2- and 7-days serial interval, respectively.
220 Efficient disease control was instituted in Singapore enabling it to remain in the index stage
221 while Japan was moving to the takeoff stage characterized by increased R estimates 3.6 (95%
222 CI 0.4-7.3) 2.2 (95% CI 1.3-3.0) for 2- and 7-days serial interval, respectively. The R estimates
223 overlaps for all exemplar country outbreak stages in the two serial interval scenarios, suggesting
224 that the transmission could be as short as 2 days. These estimates were relatively lower than
225 previously reported, bringing to light possibility of transmission in the incubation period that is
226 associated with rapidly expanding outbreaks, which is currently being observed in many
227 European countries.

228 Low detection of COVID-19 was observed in representative countries in the index stage with
229 low R values (<2) that can be attributed to effectiveness of social distancing intervention (i.e.
230 Hong Kong) or under detection for countries with limited testing (i.e. United States) (Figure 3a).
231 Sustained local transmission was occurring in five countries that were progressing into takeoff
232 stage (Japan, Germany, Spain, Kuwait and France) as measured by R values (>2) (Figure 3b).
233 The magnitude of spread was apparent with relatively higher R estimates (>10) in Italy, Iran and
234 South Korea, which demonstrated sudden surges in incidence due to prior undetected clusters

235 in part but other factors may contribute to this observation (Figure 3b). This significantly
236 increased the instantaneous R estimates versus other methods of estimation but allows a more
237 obvious depiction of the surge of cases allows differentiation of the takeoff stage from
238 exponential stage.

239 We further examined the value of computing country-specific instantaneous R by
240 comparing different temperature range (tropical versus temperate) and population density.
241 Population density of key cities (Table 2) and the higher temperature range values were used
242 for selected countries; however, no direct link was observed. Increases in the South Korean
243 outbreak was associated with a secretive religious group Shinsheonji (73% cases of COVID-19
244 in South Korea) located mainly in Daegu with a lower population density 883/km² as compared
245 to the rest of the areas with an outbreak²¹ and likely explain the outbreak expansion in the early
246 epicurve. Religious beliefs that modify health seeking behavior particularly reporting clinical
247 signs of COVID-19 combined with continued large group gathering prevented early detection of
248 the outbreak. While most countries (Table 2) have cooler temperatures (10-6°C), Singapore's
249 temperature higher indicated that local transmission occurred at higher temperatures and
250 suggests that temperature shifts will not likely change transmission. These commonly accepted
251 environmental and behavioral activities did not explain the epicurve. This led to the hypothesis
252 that the viral genomic variation underpinned changes in cases during outbreaks in each country.

253 We determined the relationship of epicurve stage with viral genetic variation using a
254 metric that merges absolute genome variation with the rate of genome change to create the
255 GENI metric that anchored population genome diversity with the rate of evolution for the SARS-
256 Cov-2. To examine how the viral genome diversity was associated with the epicurve stages we
257 first examined the index stage (Singapore) and the exponential (South Korea). Integration of
258 GENI scores successfully distinguished the index from exponential stage (Figure 4). An
259 increase in GENI scores was associated with exponential stage with a median score of 4,
260 suggesting that the viral diversity and rate of mutation played was directly proportional to case

261 increases during this stage. Singapore (index stage) effectively controlled the disease before
262 becoming exponential had a GENI median score of 2. This was found in multiple time points
263 during the outbreak were multiple mutation events were directly associated with increases in
264 cases. While China is in the decline stage the retrospective association with R, cases, and
265 GENI provided longitudinal evidence of multiple expansion in cases with mutation events in the
266 viral genome, especially early in the epicurve. The repeated viral mutations and epicurve
267 expansion were associated in each time point over 3 months, in three countries, and in three
268 different outbreak stages. This finding is useful in integrating virus genome diversity and
269 evolution into assessment of outbreak status in an outbreak between countries but also within
270 the epicurve when combined into a triad with instantaneous R estimates. The proportionality of
271 GENI scores with the epicurve stage indicates its value in determining the outbreak status and
272 the importance of generating population scale genome sequence resources.

273 A framework to merge epidemiology and population genomics was derived from this
274 study as a systematic method for molecular epidemiology (Fig. 5). It requires dynamic
275 measurements be taken for R and longitudinal efforts to determine each virus whole genome
276 sequence. Using this triad of measurements accurately and quickly provided insight to measure
277 outbreak progress but also provides an evidence-based method for interventions. This study
278 demonstrated an advancement of how to use population genomics in a viral situation where the
279 mutation rate is fast and the genome diversity of the population is extraordinarily high. GENI
280 provided a missing method that defines how to use viral genome mutation dynamics and
281 genome population diversity, which is only observable using large numbers of genomes, that
282 occurs during an outbreak.

283
284 **Discussion:**

285 Public health response is proportional to the severity and transmission dynamics of an
286 infectious disease outbreak. This requires epidemiological metrics that can be used as decision
287 criteria, and ideally, they can be used to assess impact of the intervention. In this work we

288 determined that R is much more dynamic in the COVID-19 pandemic than previously
289 appreciated by country as well as over the outbreak within each country (Fig 2-3). The
290 instantaneous R estimation with a serial interval of 2 was extremely sensitive to shifts in the
291 epicurve during the index phase (Fig 2-3). Singapore is an excellent example of effectively
292 controlling and containing the COVID-19 outbreak. They previously designated a response
293 system called Dorscon (Disease Outbreak Response System Condition)²² providing a
294 systematic approach to control so that they have not moved past the index phase. In contrast,
295 most other countries in this phase are poised to move into the takeoff phase (Fig 3). The
296 transition into the takeoff phase signified a transition from a 2-day serial interval to a 7-day serial
297 interval that was more sensitive to shifts in the epicurve.

298 While estimates of R alone is insightful in retrospect, gaps in epidemiological surveillance
299 due to several factors creates blind spots that hindered the ability to determine interventions. To
300 overcome this limitation, we merged GENI estimates based on whole genome sequence
301 variation and mutation rate with the epicurve and R and provided a predictive triad of
302 measurement that resulted in insight that accurately refined case expansion (Fig. 4). Each
303 phase of the outbreak was characterized with mutations that led to new cases in established
304 outbreaks by case definition. The merged information indicate that China found variation in the
305 viral sequence much earlier than the outbreak cases increased. Independent of the phase
306 framework merging sequence variants with the epicurve found that new cases were observed in
307 the same timeframe as new sequence variants were found. Previous studies that the
308 relationship of genomic diversity with epidemic severity (i.e. R) found no clear link²⁰. However,
309 by merging instantaneous R, the epicurve stage, and the GENI index it is clear that a link exists
310 for each country examined that resulted in a direct link between outbreak dynamics and the
311 absolute genomic mutation with the mutation rate. The GENI index provides a basis to examine
312 imported cases or locally spreading, both of which addressed this current work using
313 established metric - R and novel integration of viral whole genome sequences to define changes

314 in the sequence that are directly linked to increases in cases. This leads to an epidemiological
315 metric that is scientifically robust and at the same time can convey complex biological properties
316 to enable an efficient characterization of an outbreak in combination. Transforming complex
317 pathogen characteristics was made usable to public health and medical field using the GENI
318 score as a complete merged information set with other characteristics of the outbreak.

319 Previous outbreaks, such as Ebola, employed state of the art analysis using phylodynamics
320 that is anchored on the genetic evolution¹³. Inference such as time to most recent common
321 ancestor allowed estimation of outbreak origin, population size, and R – yet this was not
322 integrated into the outbreak dynamics and stage of advancement in the outbreak. This type of
323 analysis is possible because genomic sequences carry temporal signals and when used in
324 context with sample from different timepoints, previous divergence can be determined. The
325 GENI score includes these signals and expands their use by merging them with the outbreak
326 dynamic using the population genome variation as well as the mutation rate.

327 This inherit information is not limited to viruses. Another recent example in a bacterial setting
328 was the cholerae outbreak in Haiti wherein the phylogenetic analysis resolved the origin of the
329 pathogen²³. However, for this analysis to succeed, a substantial database of genome
330 sequences is needed, collected across time and geographic location to enable placement in a
331 phylogenetic context. As outbreaks as bound to happen in the future, investment in cataloguing
332 the genomic space of pathogens is as ever important^{24,25}. It is critical to obtain COVID-19
333 sequences from humans as well as other animals that have zoonotic potential, as was
334 demonstrated previously with zoonotic *Campylobacter* species^{26,27}. Creating sequence
335 repositories of pathogens is critical and underway for various pathogens²⁵ as well as COVID-
336 19¹⁸.

337 Prior work forewarned the practice of being overly dependent on early estimates of R
338 alone²⁸. By having the most accurate possible information for a dynamic metric and taking into
339 account the complex dynamics that factor in the calculation of R along with merging this the

340 genomics of the pathogen is a robust and insightful method to assess outbreak dynamics, as
341 demonstrated in this study. Openness and data sharing of incidence reports and sequences at
342 unprecedented scale is being done in this pandemic and it is paying rewards²⁹. Leveraging on
343 these resources opens unexpected collaboration and avenues for applying relevant
344 bioinformatic and disease modelling skills across the scientific community to solve global public
345 health problems. Examples that hindered this were observed in several countries that led to
346 cryptic spread of the disease in countries. Additionally, lacking the epidemiological infrastructure
347 and genome sequencing capabilities limit this approach that is not acceptable for modern public
348 health. However, without the appropriate technical skills in the performing complicated
349 phylogenetic inference, utility of such innovation will be limited. Establishing a protocol for
350 merging epidemiology and genomics was defined in this work (Fig. 5) and can be instituted
351 globally.

352 **Conclusion**

353 This study integrated population genomics into epidemiological methods to provide a framework
354 for molecular epidemiology. Specifically, this study demonstrated using epicurves,
355 instantaneous R estimates, and GENI specific case increases in COVID-19 are directly
356 associated with viral mutation. It was demonstrated that the pandemic is poised to become
357 larger and that mutation will be associated with the increase in cases. Exemplar outbreaks, such
358 as Singapore, found increases in cases with viral mutations that were effectively controlled.
359 However, other outbreaks had expanding R estimates during the outbreak, as well as numerous
360 viral mutation events. Use of epicurve stages, instantaneous R estimates, and GENI provided a
361 robust and accurate framework to monitor outbreak progression to different stages with direct
362 association between cases and increases in each metric.
363

364 **Acknowledgement**

365

366 We gratefully acknowledge the authors, originating and submitting laboratories of the
367 sequences from GISAID's COVID-19 Genome Database. We also thank the global community
368 for rapid information sharing that enabled integration of these data.

369
370 **References**

- 371 1 Heesterbeek, J. A. P. & Dietz, K. The concept of R_0 in epidemic theory. *Statistica*
372 *Neerlandica* **50**, doi:doi:10.1111/j.1467-9574.1996.tb01482.x.
- 373 2 Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T. & Jacobsen, K. H. Complexity of
374 the Basic Reproduction Number (R_0). *Emerg Infect Dis* **25**, 1-4,
375 doi:10.3201/eid2501.171901 (2019).
- 376 3 Fine, P., Eames, K. & Heymann, D. L. "Herd immunity": a rough guide. *Clin Infect Dis*
377 **52**, 911-916, doi:10.1093/cid/cir007 (2011).
- 378 4 Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-
379 Infected Pneumonia. *N Engl J Med*, doi:10.1056/NEJMoa2001316 (2020).
- 380 5 Zhao, S. *et al.* The basic reproduction number of novel coronavirus (2019-nCoV)
381 estimation based on exponential growth in the early outbreak in China from 2019 to
382 2020: A reply to Dhungana. *Int J Infect Dis*, doi:10.1016/j.ijid.2020.02.025 (2020).
- 383 6 Wallinga, J. & Teunis, P. Different Epidemic Curves for Severe Acute Respiratory
384 Syndrome Reveal Similar Impacts of Control Measures. *American Journal of*
385 *Epidemiology* **Volume 160**, Pages 509–516, doi: <https://doi.org/10.1093/aje/kwh255>.
- 386 7 Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to
387 estimate time-varying reproduction numbers during epidemics. *Am J Epidemiol* **178**,
388 1505-1512, doi:10.1093/aje/kwt133 (2013).
- 389 8 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus:
390 implications for virus origins and receptor binding. *The Lancet* **395**, 565-574,
391 doi:10.1016/s0140-6736(20)30251-8 (2020).
- 392 9 Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl*
393 *J Med* **382**, 727-733, doi:10.1056/NEJMoa2001017 (2020).
- 394 10 Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic
395 and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a
396 modelling study. *The Lancet* **395**, 689-697, doi:10.1016/s0140-6736(20)30260-9 (2020).
- 397 11 Campbell, F., Strang, C., Ferguson, N., Cori, A. & Jombart, T. When are pathogen
398 genome sequences informative of transmission events? *PLoS Pathog* **14**, e1006885,
399 doi:10.1371/journal.ppat.1006885 (2018).
- 400 12 Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission
401 from whole-genome sequence data. *Mol Biol Evol* **31**, 1869-1879,
402 doi:10.1093/molbev/msu121 (2014).
- 403 13 Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission
404 during the 2014 outbreak. *Science*, doi:10.1126/science.1259657 (2014).
- 405 14 Cotten, M. *et al.* Transmission and evolution of the Middle East respiratory syndrome
406 coronavirus in Saudi Arabia: a descriptive genomic study. *The Lancet* **382**, 1993-2002,
407 doi:10.1016/s0140-6736(13)61887-5 (2013).
- 408 15 Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19
409 in real time. *The Lancet Infectious Diseases*, doi:10.1016/s1473-3099(20)30120-1
410 (2020).
- 411 16 Coordinators, N. R. Database resources of the National Center for Biotechnology
412 Information. *Nucleic Acids Res* **46**, D8-D13, doi:10.1093/nar/gkx1095 (2018).

- 413 17 Seemann, T. *Rapid haploid variant calling and core genome alignment*,
414 <https://github.com/tseemann/snippy> (2020).
- 415 18 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from
416 vision to reality. *Euro Surveill* **22**, doi:10.2807/1560-7917.ES.2017.22.13.30494 (2017).
- 417 19 Bandoy, D. Large scale enterohemorrhagic E coli population genomic analysis using
418 whole genome typing reveals recombination clusters and potential drug target
419 *F1000Research* **8**, doi:10.12688/f1000research.17620.1 (2019).
- 420 20 Bedford, T. *Phylogenetic estimation of incidence and prevalence of novel coronavirus*
421 *(nCoV) infections through time*, <https://github.com/blab/ncov-phylogenetics> (2020).
- 422 21 Control, K. C. f. D. *The update of COVID-19 in ROK*,
423 <https://www.cdc.go.kr/board/board.es?mid=a30402000000&bid=0030> (2020).
- 424 22 Niehus, R., De Salazar, P. M., Taylor, A. & Lipsitch, M. Quantifying bias of COVID-19
425 prevalence and severity estimates in Wuhan, China that depend on reported cases in
426 international travelers *medrxiv*, doi:10.1101/2020.02.13.20022707 (2020).
- 427 23 Eppinger, M. *et al.* Genomic epidemiology of the Haitian cholera outbreak: a single
428 introduction followed by rapid, extensive, and continued spread characterized the onset
429 of the epidemic. *mBio* **5**, e01721, doi:10.1128/mBio.01721-14 (2014).
- 430 24 Weis, A. M. *et al.* Large-Scale Release of Campylobacter Draft Genomes: Resources for
431 Food Safety and Public Health from the 100K Pathogen Genome Project. *Genome*
432 *Announc* **5**, doi:10.1128/genomeA.00925-16 (2017).
- 433 25 Weimer, B. C. 100K Pathogen Genome Project. *Genome Announc* **5**,
434 doi:10.1128/genomeA.00594-17 (2017).
- 435 26 Taff, C. C. *et al.* Influence of Host Ecology and Behavior on Campylobacter jejuni
436 Prevalence and Environmental Contamination Risk in a Synanthropic Wild Bird Species.
437 *Applied and Environmental Microbiology* **82**, 4811-4820, doi:10.1128/aem.01456-16
438 (2016).
- 439 27 Weis, A. M. *et al.* Genomic Comparisons and Zoonotic Potential of Campylobacter
440 Between Birds, Primates, and Livestock. *Appl Environ Microbiol*, 7165-7175,
441 doi:10.1128/AEM.01746-16 (2016).
- 442 28 Wilder-Smith, A. & Freedman, D. O. Isolation, quarantine, social distancing and
443 community containment: pivotal role for old-style public health measures in the novel
444 coronavirus (2019-nCoV) outbreak. *J Travel Med*, doi:10.1093/jtm/taaa020 (2020).
- 445 29 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**,
446 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).
- 447
448

449 **Figure legends:**

450 Figure 1. Distribution of country classification based on COVID-19 epicurve status.

451

452 Figure 2. Instantaneous reproductive number estimates for different stages of the COVID-19
453 epidemic curve: a) index (Singapore) b) takeoff (Japan) c) Exponential (South Korea) d) decline
454 (China) in short (2 days) and standard (7 days) serial interval. Decelerating stage of epidemic
455 curve results to a reproductive number lower than 2 for both serial intervals, epidemic curve with
456 multiple introductions yields 2-day serial interval with higher reproductive number and
457 exponential serial interval yields higher reproductive number for the 7-day serial interval. Dot (.)
458 the surge in the epidemic curve of China corresponds to the alteration of the case definition of
459 COVID-19 by broadening confirmed cases with pneumonia confirmed with CT (computed
460 tomography) scan. South Korea's higher reproductive number is due to cryptic transmission
461 associated with a secretive cult with altered health seeking behavior.

462

463 Figure 3. Epicurve estimates with different serial intervals. Panel A represents Epicurves and
464 instantaneous R values for index stage countries using 2- and 7-day serial interval. Panel B
465 Global dynamics of COVID-19 using instantaneous estimate of reproductive number with 2-day
466 serial interval. Under preincubation period infectivity scenario, reproductive numbers globally
467 increasing (> 2). Italy's $R = 8$ is highest due to late detection of infection clusters. This higher R
468 estimate is due to a huge bump in cases combined with diagnostic gap of low-level incidence.
469 The same surge dynamics is seen in South Korea. Global dynamics of COVID-19 using
470 instantaneous estimate of reproductive number with 7-day serial interval. Italy's R value inflates
471 to 57 with the 7-day serial interval assumption and overlaps with the lower threshold of 2 day
472 serial interval R estimate. This estimation depicts a decreasing pattern for countries multiple
473 introductions like Singapore, Hong Kong.

474

475 Figure 4. Relationship of pathogen genome identity (GENI) score with the temporal signal along
476 the epidemic curve. Local transmission is captured by virus mutation as expressed in GENI
477 score values. GENI scores of SARS-COV2 isolates are relative to Wuhan reference strain
478 Wuhan-Hu-1 NC_045512.2. The red line in the China epicurve represents the time before an
479 outbreak was determined yet genome sequences were circulating. The blue shaded curves
480 indicate GENE scores directly overlaid with the outbreak curve. The dotted line represents the
481 common point in time as a reference for visualization. The GENI score and epicurve show
482 similarity except in China as the outbreak advanced to takeoff and exponential the GENI score
483 increased while in the index stage example of Singapore the outbreak was contained and the
484 GENI score remained <2.

485
486 Figure 5. Integration of genomic and classical epidemiology for outbreak investigation. The
487 foundation of epidemiology is the accurate and timely reporting of cases which enable the
488 calculation of the number. Genomic Identity (GENI) score is formulated from genomic data of
489 pathogens to differentiate imported cases versus local transmission and measure time of cryptic
490 spread. Together these two epidemic values deliver insight that can be directly used for making
491 decision criteria for public health intervention.

492

493

494 **Table 1.** Country-specific Instantaneous Reproductive Number (R) estimates for COVID-19 as
495 of March 1, 2020.

Country	Cases	Instantaneous Reproductive Number (R) with different serial intervals	
		2 days	7 days
Mainland China	79251	1.6	2.1
South Korea	3150	2.8	25.6
Italy	1128	8	57.0
Iran	593	2.8	17.1
Japan	241	3.6	2.2
Singapore	102	3.3	1.6
France	100	2.9	16.9
Hong Kong	95	2.6	1.6
Germany	79	3.1	17.2
United States	70	4.3	1.7
Kuwait	45	2.6	15.3
Spain	45	3.7	10.8
Thailand	42	3.8	1.7

496
497

498 **Table 2.** Epidemiological Parameters and instantaneous R estimates. The population density for
499 South Korea is based on Daegu where 75% of the cases are reported.

500

	Reproductive Number (R)	Temperature (°C) during outbreak	Population Density (people/km ²)	Interpretation in consideration of the epidemiological curve
Singapore	3.3	32	8136	Imported cases, limited local transmission
France	2.9	10	4300	Imported, Local transmission >1-2 month
Italy	8	10	7200	Imported cases, Local transmission >1 month
United States	4.3	9	8444	Imported cases, Local transmission >2 month
South Korea	2.8	6	883	Imported cases, Local transmission >1-2 month

501
502

503 **Table 3.** Relationship of Pathogen Genome Identity (GENI) Score derived from mutational
504 difference from the index genome (Wuhan isolate of COVID-19 or cluster isolate reference from
505 multiple outbreak regions outside of territory).

506

Equivalent Pathogen Genome Identity (GENI) score for COVID-19	Clinical Interpretation and Epidemiological Inference	Notes
0-2	No difference from index case isolate genome or reference, imported case if there is no prior report, indicative of acute transmission <1 month	Reference genome is primarily earliest isolate available.
3-4	recent local transmission (average 1-2 months) if there are no prior report of cases	Subsequent outbreak clusters can serve as sources of introduction hence near neighbor reference has to be selected to generate an accurate GENI score.
>4	sustained local transmission (greater than 2 months) if there is no prior report of cases	Subsequent outbreak clusters can serve as sources of introduction hence near neighbor reference has to be selected to generate an accurate GENI score.

507

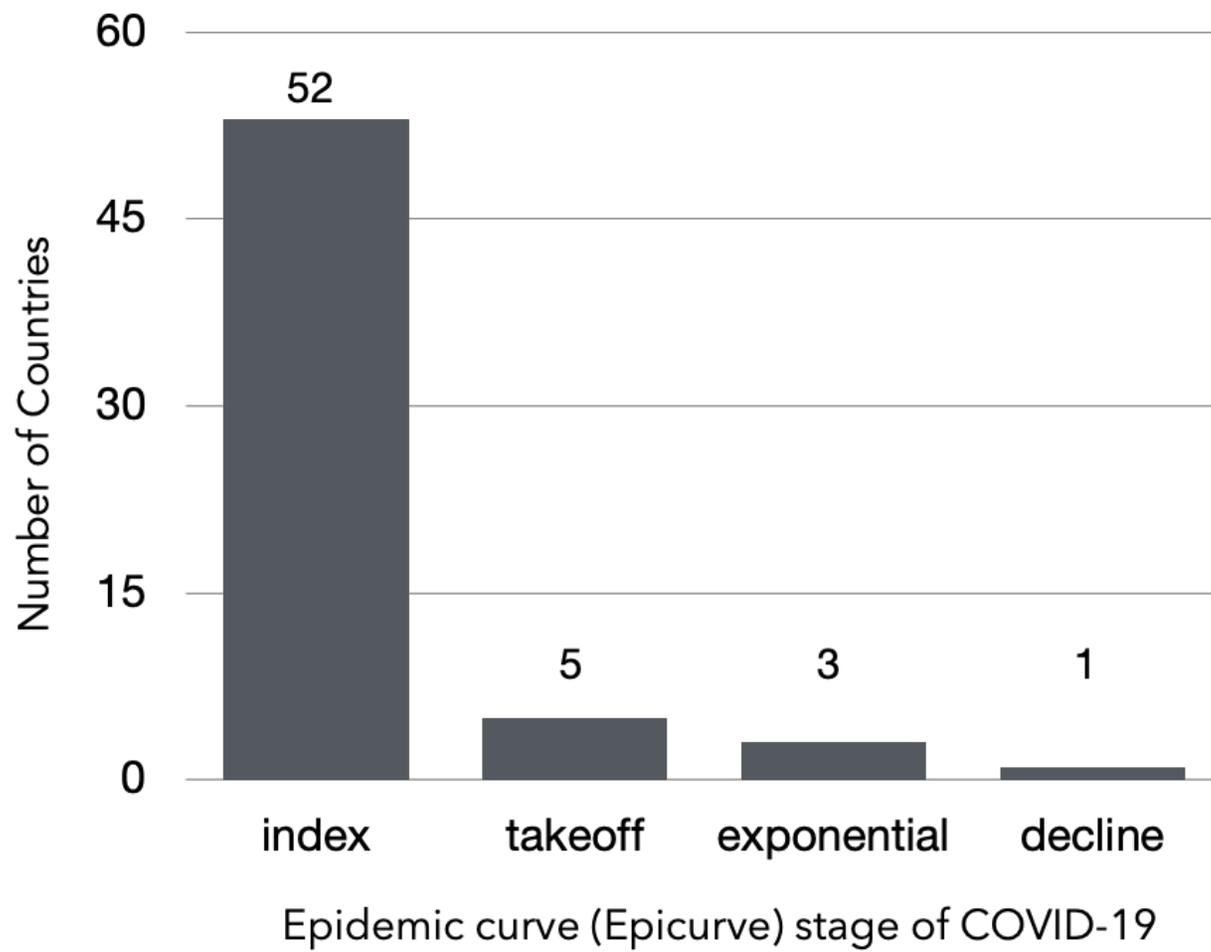


Fig 1

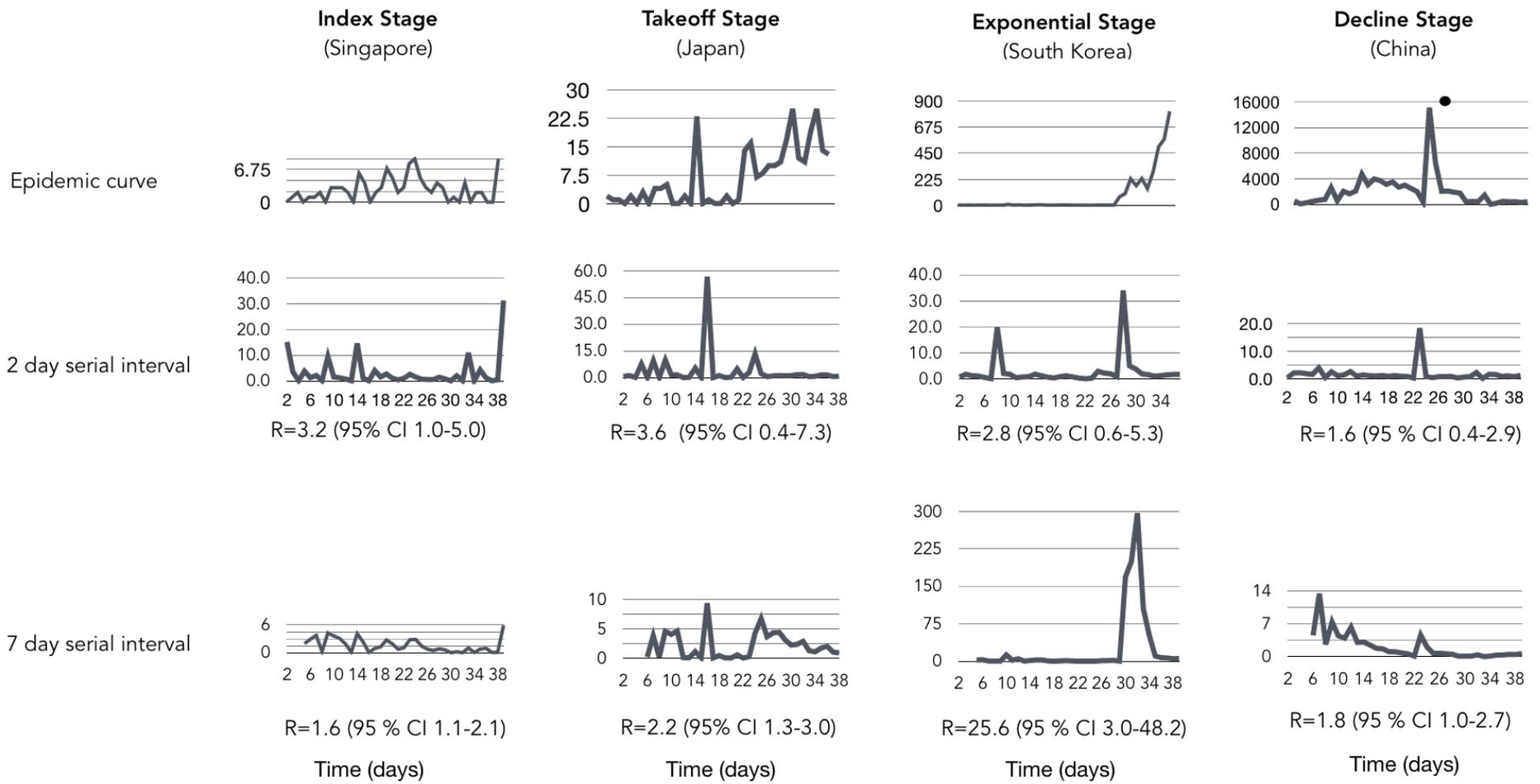


Figure 2

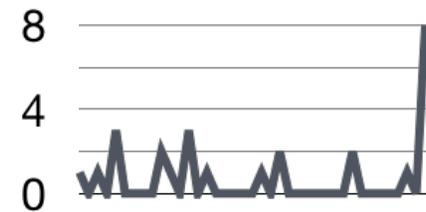
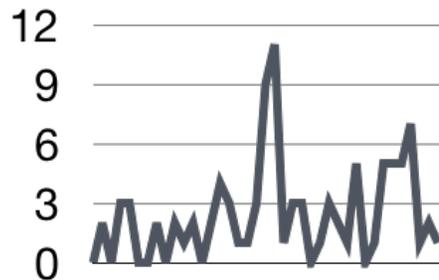
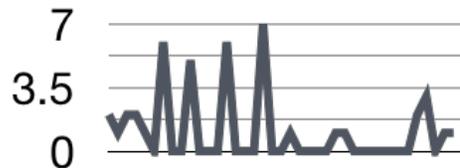
Index Stage

Thailand

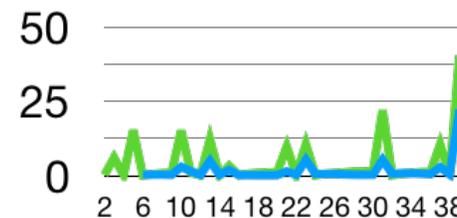
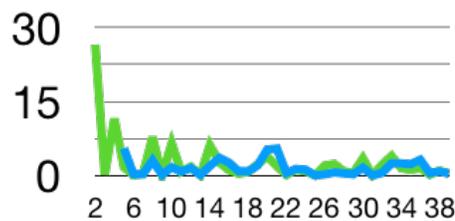
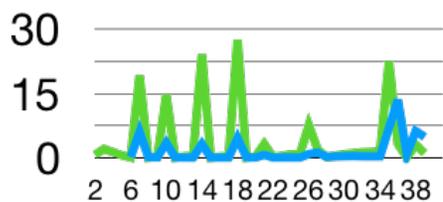
Hong Kong

United States

Epidemic curve



Reproductive Number (R)



Time (days)

Time (days)

Time (days)

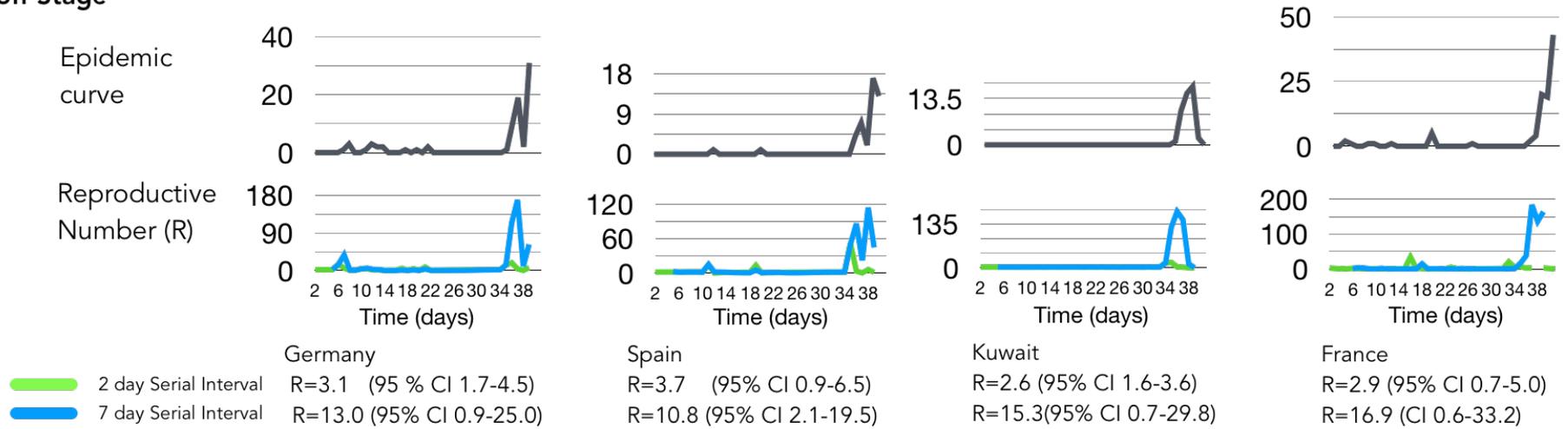
█ 2 day Serial Interval R=3.8 (95% CI 1.4-6.2)
█ 7 day Serial Interval R=1.6 (95% CI 0.6-2.7)

R=2.6 (95% CI 1.1-4.2)
 R=1.6 (95% CI 1.1-2.1)

R=4.3 (95% CI 1.7-7.0)
 R=1.7 (95% CI 0.3-3.1)

Figure 3a

Takeoff Stage



Exponential Stage

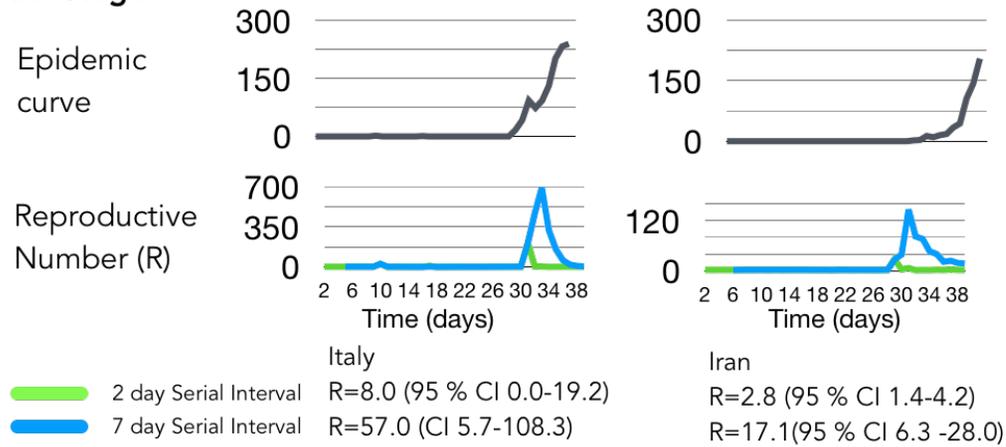


Figure 3b

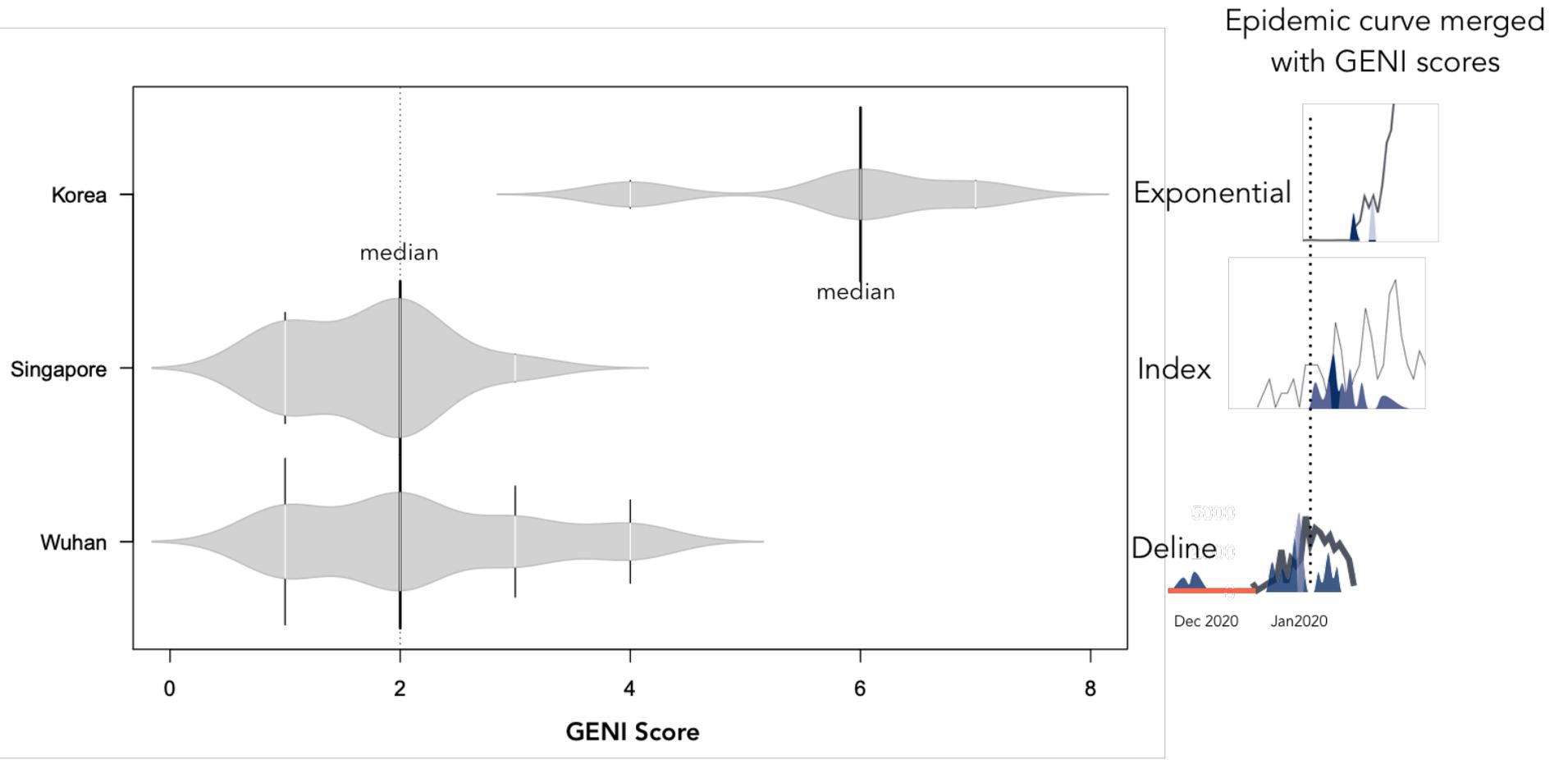
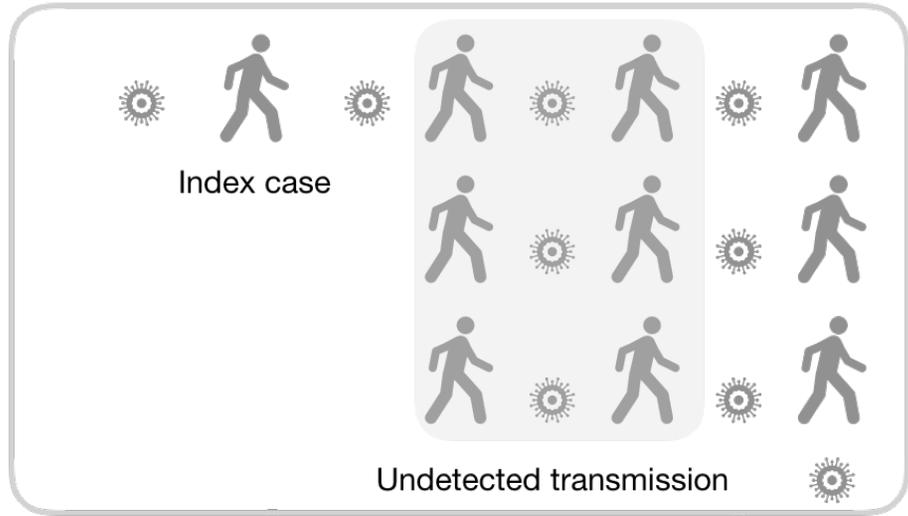


Figure 4

Infectious Disease Outbreak



Epidemiological Outbreak Investigation

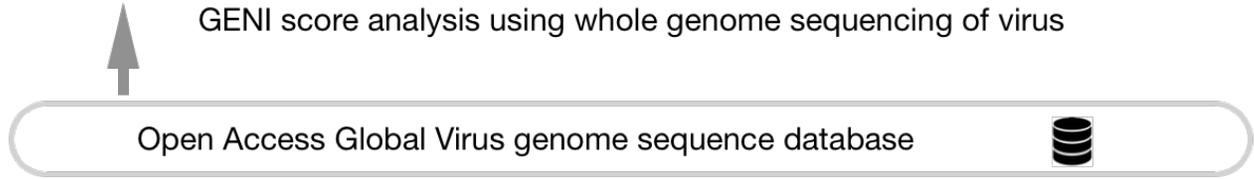
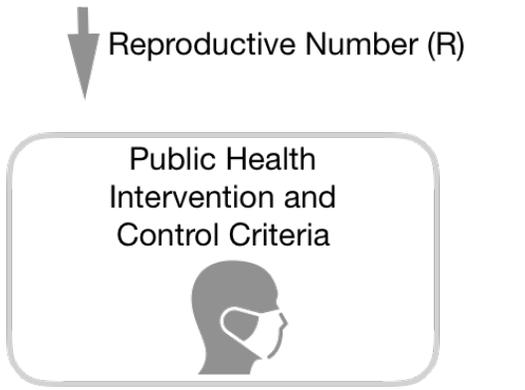
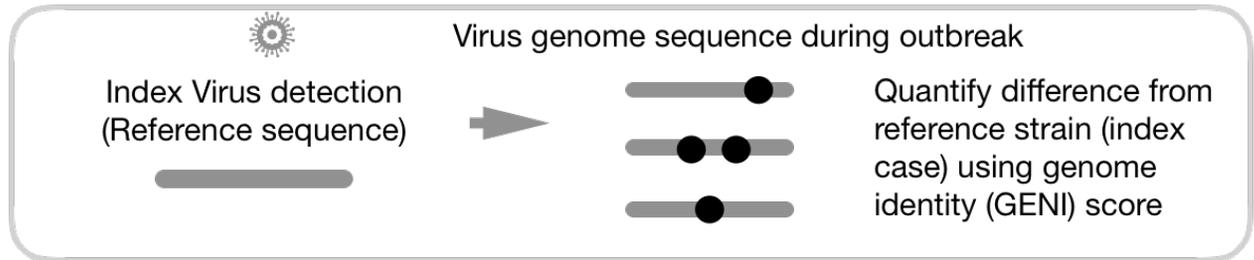
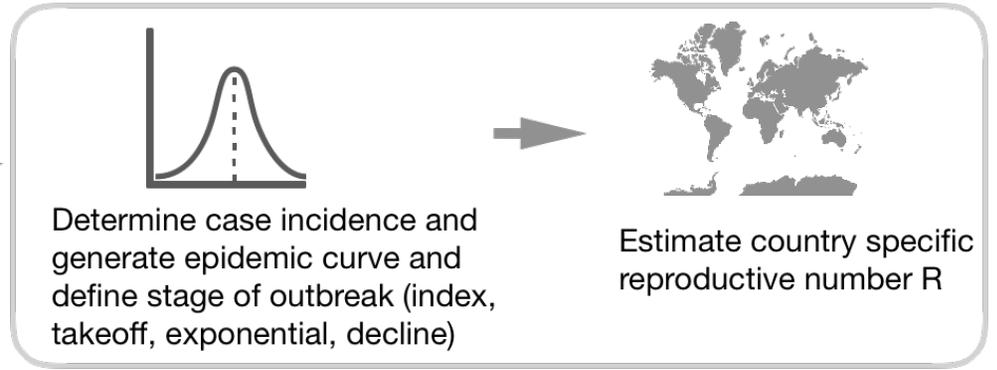


Figure 5