

1 **Genetic determination of regional connectivity in modelling the spread of COVID-19**
2 **outbreak for improved mitigation strategies**

3 Leonidas Salichos^{1,2*}, Jonathan Warrell^{2*}, Hannah Cevasco², Alvin Chung², Mark Gerstein^{2,3,4,5}

4
5 **ABSTRACT**

6 Covid-19 has resulted in the death of more than 1,500,000 individuals. Due to the pandemic's
7 severity, thousands of genomes have been sequenced and publicly stored with extensive records,
8 an unprecedented amount of data for an outbreak in a single year. Simultaneously, prediction
9 models offered region-specific and often contradicting results, while states or countries
10 implemented mitigation strategies with little information on success, precision, or agreement
11 with neighboring regions. Even though viral transmissions have been already documented in a
12 historical and geographical context, few studies aimed to model geographic and temporal flow
13 from viral sequence information. Here, using a case study of 7 states, we model the flow of the
14 Covid-19 outbreak with respect to phylogenetic information, viral migration, inter- and intra-
15 regional connectivity, epidemiologic and demographic characteristics. By assessing regional
16 connectivity from genomic variants, we can significantly improve predictions in modeling the
17 viral spread and intensity.
18 Contrary to previous results, our study shows that the vast majority of the first outbreak can be
19 traced to very few lineages, despite the existence of multiple worldwide transmissions.
20 Moreover, our results show that while the distance from hotspots is initially important,
21 connectivity becomes increasingly significant as the virus establishes itself. Similarly, isolated
22 local strategies -such as relying on herd immunity- can negatively impact neighboring states. Our

23 work suggests that we can achieve more efficient unified mitigation strategies with selective
24 interventions.

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42 **INTRODUCTION**

43 Covid-19 related deaths have surpassed 1,500,000 worldwide and 330,000 in the United States.

44 Due to the importance of the pandemic, many resources are available for COVID-19 genome

45 research, including GenBank, GISAID, and Nextstrain¹⁻³. Due to the severity of the pandemic

46 combined with the advent of sequencing technologies, the amount of sequencing data within
47 such a short time period for a single outbreak is unprecedented. GISAID is currently the largest
48 COVID-19 database with more than 309,000 SARS-CoV2 genomes ². This is compared to 1760
49 sequences of influenza A/H3N2 collected from 2013 to 2020. These numbers are comparable or
50 surpassing the number of HIV or HCV sequences in the Los Alamos national database ^{4,5}. These
51 COVID-19 genomes represent the spread of the pandemic from China to 188 countries
52 worldwide, with more sequences added every day.

53 Recent studies have modelled the transmission, diversity and spatial phylogeography of the virus
54 mostly in a historical context ⁶⁻¹². According to studies, Covid-19 first arrived in Washington ¹¹,
55 in what is considered a cryptic infection ^{11,13}. However, known cases in persons with no relevant
56 travel history also occurred in California in late January/early February¹¹. While the first lineages
57 arrived from China in Washington and California, subsequent infectious lineages (notably in
58 New York) appear to represent importations from Europe ^{11,14}. In this context, early results also
59 suggested multiple worldwide transmissions responsible for the outbreak in the North-East of
60 United States ¹².

61 From the beginning of the pandemic, different approaches have been developed for the modeling
62 of the outbreak that use either epidemiological, demographic data ¹⁵⁻²⁰. Many -sometimes
63 contradicting- prediction models offered temporal, and locally isolated results based on local
64 outbreaks ^{16-18,21,22}, while each country implemented their strategy to combat the outbreak,
65 including controversial approaches such as “herd immunity”²³⁻²⁵. At the same time, different
66 forms of local lockdowns have been tested to successfully mitigate viral spread as non-
67 pharmaceutical interventions (NPIs) ^{22,26-28}. While previous studies offer extremely valuable
68 insights into the history of viral transmission and the effectiveness of locally implemented NPIs,

69 they tend to overlook the inland spread (migration) of the virus in order to provide a unified
70 mitigation strategy that complements local implementations.

71 In this study, we show a strong association between the temporal and geographical spread of the
72 virus. By using a case study of seven states [New York (NY), New Jersey (NJ), Connecticut
73 (CT), Massachusetts (MA), Pennsylvania (PA), Maryland (MD) and Virginia(VA)], we utilize
74 the concepts of ingrowing, incoming and outgoing viral connectivity between states and regions
75 as factors that influence the spread and viral transmission. We then use regression and random
76 walk models to show the importance of these concepts combined with epidemiological and
77 demographic factors -such as transmission rates and Urbanization Index in providing more
78 informative predictions and explaining the temporal and geographic spread of the pandemic. The
79 significance of modeling the spread of the viral wave through geographical routes and regional
80 connectivity reveals broader implications and opportunities for the consideration of more
81 efficient mitigation strategies in blocking viral migration with additional selective interventions.

82

83 **Initial Distance From Hotspots Determines Outbreak Severity**

84 Using the numbers of ‘deaths per 1 million population’ as a proxy for regional outbreak severity,
85 first we aimed to assess the association between distance from initial viral hotspots and the
86 severity of the viral outbreak. To assign a geographic location for each state, we used the
87 longitude and latitude from its respective larger city. Then, we considered the distance from New
88 York City (New York), Seattle (Washington) and New Orleans (Louisiana) as the three initial
89 hotspots of the outbreak. Introducing New York City (NYC) as a single initial hotspot, showed a
90 high negative correlation ($r=-0.37$) between the severity of the outbreak and the distance from
91 hotspot. By including Seattle (or San Francisco) as a second hotspot, the association increased

92 ($r=-0.43$). Finally, by fitting a logarithmic curve, the association increased further to $R^2=0.35$
93 (figure 1). The inclusion of New Orleans as a third hotspot did not improve our results,
94 indicating an isolated outbreak. On the contrary, by removing Louisiana as an outlier, we
95 improved the predictability of the logarithmic curve to $R^2=0.4$. These results suggest a very high
96 association between the outbreak's severity during the first wave and the distance from the two
97 initial hotspots.

98 By fitting a log curve for the case study of seven states (NY, CT, MA, NJ, PA, MD and VA), we
99 were able to associate the distance from NYC and the severity of the initial spread by explaining
100 70% of the variance. Additional factors strongly linked to the spread and severity of the epidemic
101 include Urbanization Index and maximum effective reproduction rate R_t per state during the first
102 wave, as retrieved from "<https://rt.live/us/>" (figure 1).

103

104 **Major 7-state Outbreak is Related to Few European Lineages**

105 To create a dataset of world reference sequences, on June 25th, we sampled 50 sequences
106 spanning the 5 Covid-19 lineages as determined by Nextstrain (Figure 2i). Lineages 19A and
107 19B represent the earliest detected infections closely associated with the Wuhan epidemic. Using
108 the GISAID database, we downloaded all available sequences for our 7 states case study until the
109 6th of August 2020. These sequences represent the first viral wave in the USA. Then, we inferred
110 a phylogenetic tree for each individual state using Bayesian inference with date constraints under
111 a Yule process population model. For all states, the major outbreak clusters with a specific
112 European lineage (reference sequence: HF1465 FRA). For New York, Connecticut and
113 Massachusetts this lineage clearly constitutes the dominant outbreak. For New Jersey,

114 Pennsylvania, Maryland and Virginia, a secondary outbreak -which also circulates in NY, CT
115 and MA- appears significant (figure 2, S1-7).

116

117 **Assessing Viral Connectivity Between States**

118 By selecting a set of (whenever possible) 50 reference sequences per state (in addition to the set
119 of world reference sequences), we built a phylogenetic tree that includes sequences from all 7
120 states. We then inferred a connectivity map between the different states by parsing the tree's
121 bipartitions (figure 3i). For this, we examined all possible connected pairs of sequences that
122 cluster together, while moving hierarchically from smaller to larger bipartitions without double-
123 counting. To establish directionality between pairs, we used sampling dates. For example, the
124 pair (NY-PV09151_USA_NY_2020-03-22, CT-UW-6574_USA_CT_2020-04-03) would be
125 counted as NY -> CT, which denotes one incoming transitional connectivity from NY to CT.
126 Similarly, the pair (NY-PV08434_USA_NY_2020-03-18, NY-NYUMC659_USA_NY_2020-
127 03-18) would be counted as ingrown connectivity NY -> NY. Overall, the NY outbreak showed
128 the highest connectivity, while VA and MA showed the lowest. Interestingly, even though CT
129 showed high connectivity comparable to NJ, the decreased number of outgoing versus incoming
130 connections explains the low connectivity shown by MA. This is also supported by the
131 outbreak's high transitional connectivity from NY to CT (NY ->CT) rendering CT as a potential
132 bottleneck (figure 3).

133

134 **Urbanism and Transitional Connectivity Increase Outbreak's Severity**

135 Previously, by considering the geographic distance from the initial hotspots NY and WA, we
136 found a strong association between the distance and the severity of the outbreak for the first

137 month of the epidemic. Additional factors associated with the spread included urbanism, the
138 maximum R_t per state, as well as the neighboring states' maximum R_t . Here, we test the
139 importance of various features in predicting the per-state death rate across the first wave of the
140 pandemic (March to August 2020). We include in our analysis features including the estimated
141 incoming, outgoing and ingrowing transmission rates between states, and a transmission-based
142 normalized distance of each state from New York representing the viral flow (described below).
143 The full feature set includes: maximum Reproduction rate per state (R_t) usually in April,
144 Urbanization Index (U), Geographic or transition-based distance from New York (D or trD), and
145 incoming, outgoing and ingrowing transmission rates. To determine the importance of regional
146 transitional connectivity in addition to these factors in explaining and predicting the outbreak
147 intensity during the whole first wave, we built 4 regression models with increasing complexity
148 that combine phylogenetic information with epidemiological data from 10 dates (April 29th, May
149 1st 8th and 15th, June 11th 18th 25th, July 2nd, 29th, August 23rd).

150 In our simplest model (figure 4i) we examined the role of urbanism, distance (D) from NYC and
151 virus' maximum reproduction rate R_t . U and D showed high and increasing significance
152 throughout the whole first wave, while the use of maximum R_t -as obtained by
153 '<https://rt.live/us/>'- showed maximum significance at the beginning of the outbreak but
154 eventually decreased. This is possibly because max R_t only represents the virus' reproductive
155 rate during the first stage of the outbreak (i.e in March-April), before the lockdowns. In our
156 second model (figure 4ii), we substituted D with Transitional Distance (trD), a weighted version
157 of distance as a proxy for viral flow which considers transitional connectivity between states
158 (e.g. NY -> CT, NY -> NY, CT -> NY, NY -> NJ, .. etc) using random walks between the states
159 (see methods). By replacing D with trD we were able to significantly increase our model's

160 predictability throughout the first wave ($p=0.0003$, figure S8). In our third model (figure 4iii), we
161 returned to using the geographic distance D , but this time we also included each states' total
162 incoming, outgoing and ingrowing rates. Finally, in our 4th model (figure 4iv), we again replaced
163 D with trD , while also including states' incoming and outgoing rates. While our 4th model also
164 integrates transitional connectivity in trD , this information is also used in calculating each state's
165 incoming, outgoing and ingrowing connectivity. Therefore, as expected, factors trD , incoming
166 and outgoing rates often behave in a complementary manner. However, model 4 is still
167 significantly more informative than model 3 ($p=0.0273$). Moreover, model 4 indicates that the
168 initial importance of trD during the beginning of the outbreak, is gradually replaced by the state's
169 connectivity rate, as the outbreak spreads away from the initial hotspots.

170

171 **A Case Study for Selective Mitigation Strategies Based on Regional Connectivity**

172 Using our second regression model with normalized transitional distance trD , we predicted the
173 total number of deaths by removing one by one each geographic connection between every
174 geographically linked state pair according to figure 3iv. Our results suggested that by enforcing a
175 blockade between New York and Pennsylvania, as well as between Maryland and Pennsylvania
176 would result in saving around 450 and 200 deaths per million individuals respectively, after the
177 lockdowns. This is a particularly interesting result, since our model seems to take into account
178 the drop in deaths in specific states after the imposed lockdowns (based on epidemiological data
179 from NJ, NY, MA and CT) and respond to the temporal flow of the pandemic resulting in later
180 death peaks in states like Virginia (figure S9). This becomes more evident in figure 4vi, where
181 we depict the temporal effect of each blockade in reducing the number of total deaths per million
182 individuals.

183

184 **New York's Second Wave**

185 To understand the origin of New York's second wave, we inferred a new phylogenetic tree, this
186 time by including all sequences from GISAID after August 2020 and until November 24th
187 (figure 5). In addition to these new sequences, we also included our set of 50 world reference
188 sequences and 50 from NY as mentioned previously. Our results indicate that about half of the
189 second NY outbreak, has been re-introduced from Europe, possibly from Great Britain. This
190 appears to be a completely new lineage, previously unseen in New York.

191

192 **METHODS**

193 **Data Availability**

194 All data are available in public databases. SARS-CoV2 genomes were retrieved from the
195 GISAID database ². Epidemiological data concerning the daily and total deaths per million
196 individuals have been retrieved from Worldometer 'worldometers.info/coronavirus/'. Maximum
197 reproduction rates have been retrieved from The Covid Tracking Project
198 "<https://covidtracking.com/>" and '<https://rt.live/us/>' ²⁹. For the first wave of Covid-19 outbreak in
199 United States we collected a total of 3,133 sequences for the states of New York (NY),
200 Connecticut (CT), Massachusetts (MA), New Jersey (NJ), Pennsylvania (PA), Maryland (MD)
201 and Virginia (VA) that were sampled between 01/29/2020 and 07/05/2020. More specifically, we
202 collected 1505, 353, 418, 45, 112, 178, 522 sequences from each state, respectively. For the
203 second wave of Covid-19 in New York, we collected a total of 112 sequences sampled between
204 08/01/2020 to 10/18/2020.

205 World reference sequences: For the use of reference sequences representing the global pandemic,
206 we randomly selected 50 sequences spanning Nextstrain lineages 19A, 19B, 20A, 20B and 20C
207 (see figure 2i).

208 State reference sequences: We randomly selected up to 50 reference sequences from each state,
209 prioritizing selection of one sequence per bipartition with higher than 50% posterior probability.
210 Excluding world reference sequences, 50 sequences were selected from NY, CT, MA and VA
211 while 43, 37, 22 were selected from NJ, PA and MD, respectively.

212

213 **Phylogenetic Analysis**

214 By retrieving the genomic sequences from GISAID (Supplementary table), we used MAFFT
215 (45) to build multiple sequence alignments for every state based on nucleotide sequence data.
216 Then, using BEAST (31, 32), we performed Bayesian phylogenetic analysis with time
217 constraints based on sampling dates, under a GTR evolutionary model. To determine the
218 appropriate growth models and population size, we tested various growth models including a i)
219 Yule process, ii) exponential growth, iii) logistic growth iv) Bayesian Skyline v) Birth–Death
220 skyline. The BEAST suite also includes multiple software tools that aid in selecting appropriate
221 models and parameters (BEAUti) to infer a phylogenetic tree using Bayesian inference,
222 coalescent theory and speciation with respect to the time of sequence collection. We evaluated
223 the efficacy of these models using Tracer v1.7.1(46). The best model (Yule process) for this data
224 was selected based on the estimated sample size, posterior probabilities, and reports on algorithm
225 convergence.

226

227 **Estimating Transitional Connectivity**

228 Using custom scripts, we were able to parse the inferred phylogenetic trees into groups of
229 sequences based on the tree bipartitions. Then, by further parsing the groups in ascending order
230 based on group size (from groups of 2 to $X=10$), we determined all possible pairs and state
231 connectivity based on dates. For example, pair of sequences {*NY-PV09151_USA_NY_2020-03-*
232 *22* and *CT-UW-6574_USA_CT_2020-04-03*} would depict an outgoing connectivity between
233 New York (NY) and Connecticut (CT) denoted as NY>CT +1 (see figure 3i). In the manuscript
234 we show results for a strict/conservative approach where pair inconsistencies are dropped, and
235 sequences cannot be considered as incoming twice.

236

237 **Maximum Reproduction Rate R_t**

238 To calculate the maximum reproduction rate R_t , we used the maximum R_t value for each state
239 from '<https://rt.live/us/>' during the first wave of the pandemic (until August 2020). R_t represents
240 the effective reproduction rate of the virus calculated for each locale. It allows to estimate how
241 many secondary infections are likely to occur from a single infection in a specific area.

States	Maximum Reproduction rate R_t	Neighboring max R_t (Average)
New York	5.3	3.3
New Jersey	4	3.32
Connecticut	3.1	3.53
Massachusetts	2.8	2.88
Maryland	2.9	2.66
Pennsylvania	3.2	3.25
Virginia	2.4	2.45

Delaware	2
North Carolina	2.5
West Virginia	1.7
Tennessee	2.7

242

243

244 **Urbanization Index**

245 For the Urbanization Index, as an indication of how “urban” a state is, we used the definition and
246 data from 538 (<https://fivethirtyeight.com/>). FiveThirtyEight’s urbanization index is calculated
247 as the natural logarithm of the average number of people living within a five-mile radius of a
248 given resident.

State	Urbanism
New York	12.56
New Jersey	12.24
Connecticut	11.41
Massachusetts	11.84
Maryland	11.71
Pennsylvania	11.15
Virginia	10.91

249

250

251 **Regression Analysis Models**

252 We perform multiple linear regression analyses in order to assess the importance of each factor
253 on the prediction of the per-state death rate. We use data from 7 states (NY, CT, MA, PA, NJ,
254 VA, MD), over a series of 10 timepoints from April 29 to July 23. We regress the per-state death
255 rate (the cumulative ratio of deaths to cases from the earliest date) on either three variables
256 (Transmission rate (R0), Urbanism, Distance from NYC) or six variables (Transmission rate
257 (R0), Urbanism, Distance from NYC, ingoing/outgoing/ingrowing rates per-state). Prior to the
258 analysis, we Z-score all variables (enforcing zero mean and unit covariance). For distance from
259 NYC, we use either the geographic distance between the state's capital and NYC, or the
260 transition distance as defined below. For each model, we calculate the log-likelihood by fitting a
261 variance parameter to the predicted outputs and using a Gaussian noise model. Hence, we set
262 $\sigma_t^2 = (1/N)\sum_{i=1:N}(y_{it} - \beta_t x_{it})^2$, where N is the number of states, β_t and x_{it} are the vectors of
263 coefficients and features associated with state i at time t respectively, and y_{it} is the associated
264 death rate. We calculate the log-likelihood at time t as $L_t = \sum_i \log(\text{Gauss}(y_{it} - \beta_t x_{it}; \mathbf{0}, \sigma_t))$,
265 where *Gauss* is the probability density function of a normal distribution. We then compare the
266 log-likelihood differences of pairs of models over time using the Pearson Correlation Coefficient
267 (differences versus temporal ordering).

268

269 **Random Walk Model**

270 We define the transmission distance of a state from NYC as the expected first arrival time at that
271 state of a Markov random walk starting at NYC, using the transition probabilities between states
272 inferred from the phylogenetic analysis. Hence, we set $d_{ij} = E(\min(\{t | s_t = j\}) | s_0 = i)$ for the
273 directed transmission-distance between states i and j (which is not a metric), where $s_t = i$
274 indicates that the state at time t in a sampled random walk is i , and $E(\cdot)$ denotes expectation. To

275 estimate these distances, we run 1000 such random walks for 1000 time-steps and use the
276 empirical mean time of first arrival at each state across samples. As above, we Z-score the
277 resulting distances for each state.

278

279 **Mitigation Analysis**

280 In order to break the link between geographically adjacent states s_1 and s_2 , we set a reduction
281 factor $r = 0.1$, and update the transmission probabilities as: $P'(s_b|s_a) = r \cdot P(s_b|s_a)$, and

282 $P'(s_a|s_a) = P(s_a|s_a) + (1 - r) \cdot P(s_b|s_a)$, where $P'(s_a|s_b)$ is the updated transition

283 probability between states s_a and s_b . We make such updates for $a = 1, b = 2$ and $a = 2, b = 1$

284 simultaneously, hence breaking the link in both directions. We then recalculate the distances

285 $d_{ij}^{s_1s_2}$, i.e. the distance between states i and j , given the link between s_1 and s_2 has been broken.

286 We then use these to estimate the overall predicted reduction in the death-rate given the break as:

287 $\Delta_{s_1s_2} = \sum_{it} w_i \cdot (y'_{it} - y_{it})$, where w_i is a weighting factor proportional to the population of state

288 i (and $\sum_i w_i = 1$), and y'_{it} is the predicted death-rate for state i at time t when $d_{ij}^{s_1s_2}$ is substituted

289 for d_{ij} in the predictive model from the Regression analysis.

290

291 **DISCUSSION**

292 Previous studies have provided an important historic view of travel history^{8,9,11-14,30} and viral

293 spread of Covid-19^{6,11,12,15} using genetic variability. Others, data driven, have modeled the

294 spread of the virus and effectiveness of government interventions^{20,27,28}. So far, the only

295 acclaimed and efficient non-pharmaceutical interventions in our arsenal are forms of regional

296 lockdowns^{19,22,26,31,32}, while other strategies relying on 'herd immunity' have also been

297 suggested and disputed²³⁻²⁵.

298 Here, we used SARS-CoV2 genomes to determine regional connectivity in a case study of 7
299 states, where New York acted as an initial hotspot. By combining epidemiological demographic
300 and genetic information, we used four regression models to evaluate the importance of different
301 factors that contribute to outbreak severity throughout the first viral wave.

302 Our results can explain the discordance between regions and strategies, especially between the
303 first and second pandemic waves. For example, states within distance from hotspots are able to
304 deal with a milder initial outbreak, before the virus establishes at a later timepoint, depending on
305 transitional distance (i.e., the speed of the wave) and regional connectivity. Similarly, states with
306 lower connectivity (e.g., naturally or physically isolated regions) can be more efficient in battling
307 the viral spread, as they deal with reduced viral wave and incoming infections. This also suggests
308 that reducing incoming transmission routes (through pharmaceutical or non-pharmaceutical
309 interventions) can have a significant effect in addition to local mitigation strategies such as
310 lockdowns. This does not necessarily mean complete isolation, but rather a blockade on
311 transmission routes with high connectivity. However, our results also suggest that states deciding
312 to follow less stringent mitigation strategies are also largely responsible for their outgoing viral
313 connectivity, affecting neighboring regions, while often taking advantage of the low incoming
314 connectivity resulting from neighboring lockdowns in return.

315 By deriving genetic connectivity between regions using genomic information, we combined
316 genetic information with demographic and epidemiological data to create a model and a proxy
317 for the flow of the viral wave in order to study factors that temporally contribute to the severity
318 of local outbreaks throughout the pandemic. Then we used this model to consider the outcome of
319 selective intervention strategies using geographic blockades. Overall, our results suggest that
320 unified mitigation strategies are more efficient in tackling a pandemic, while also providing a

321 framework within which to pursue these strategies. Our framework can be implemented for both
322 pharmaceutical (e.g vaccination) or non-pharmaceutical interventions (e.g., lockdowns,
323 blockades).

324

325 **Author Contributions**

326 L.S. conceived of the project, designed, developed, performed, and analyzed experiments. J.W.
327 developed and performed the regression and random walk models. H.C. performed the
328 phylogenetic analysis. A.C. performed the regression and random walk models. L.S. drafted the
329 paper. L.S., J.W., and M.G. wrote the paper. All authors read and approved the final paper.

330

331 **Competing interests**

332 The authors declare no competing interests

333

334 **Corresponding Authors**

335 Correspondence to Leonidas Salichos and Mark Gerstein

336

337 **Author Information**

338 *Affiliations:*

339 *1. Biological and Chemical Sciences, New York Institute of Technology, Old Westbury, NY,*

340 *11568, USA.*

341 *2. Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT,*

342 *06520, USA.*

343 3. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT,
344 06520, USA.

345 4. Department of Computer Science, Yale University, New Haven, CT, 06520, USA.

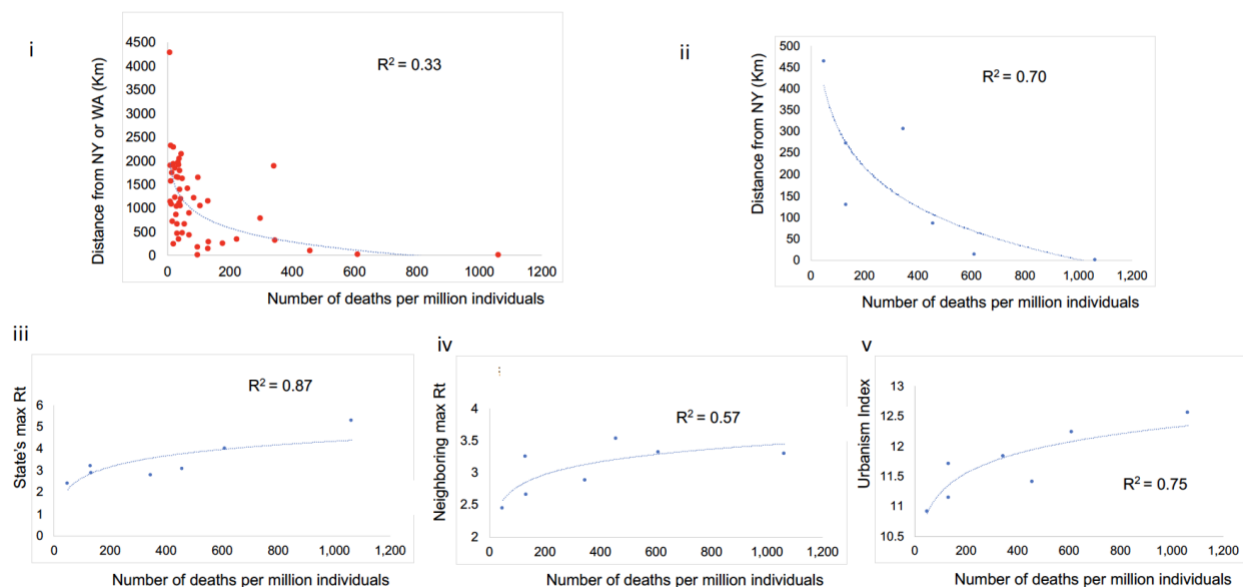
346 5. Center for Biomedical Data Science, Yale University, New Haven, CT, 06520, USA.

347 *Equal contribution

348

349 FIGURES

Figure 1



350

351 **Figure 1.** Using data collected on the 29th of April we show the logarithmic association between

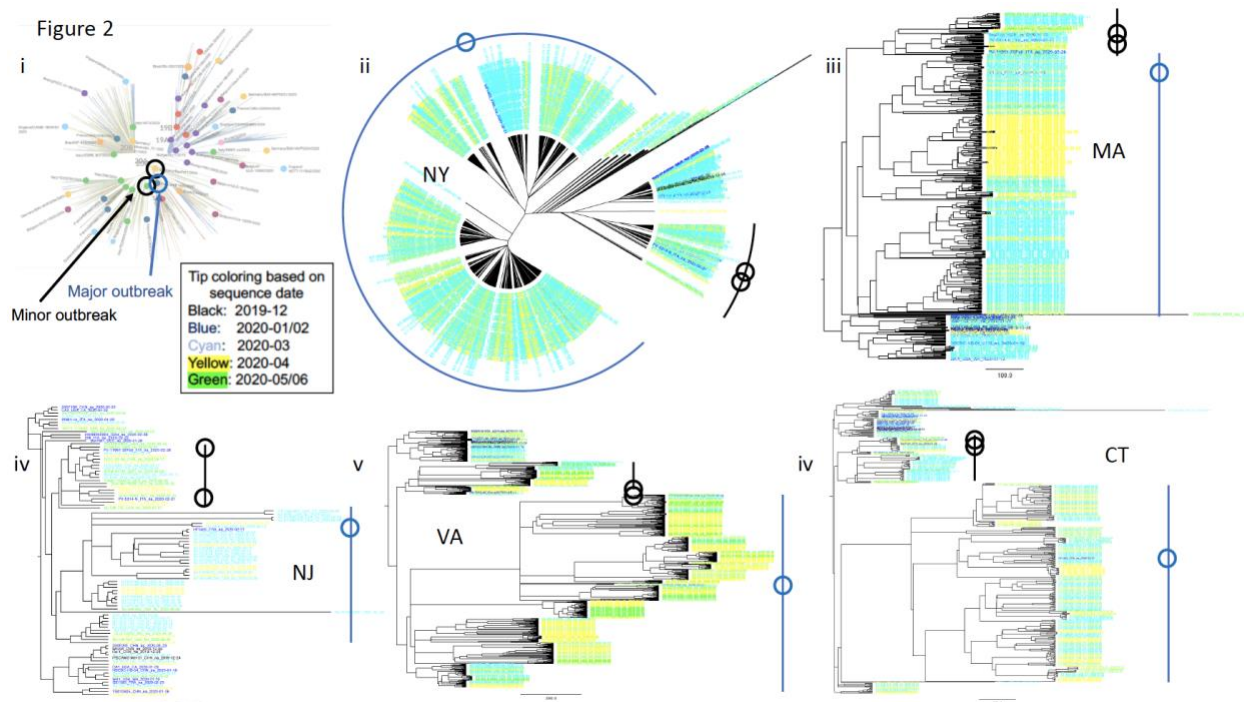
352 i) the number of deaths per million individuals for every state and the distance from hotspots

353 (New York or Washington). Using a case study of 7 states (New York, New Jersey, Connecticut,

354 Massachusetts, Pennsylvania, Virginia and Maryland), we show the logarithmic association

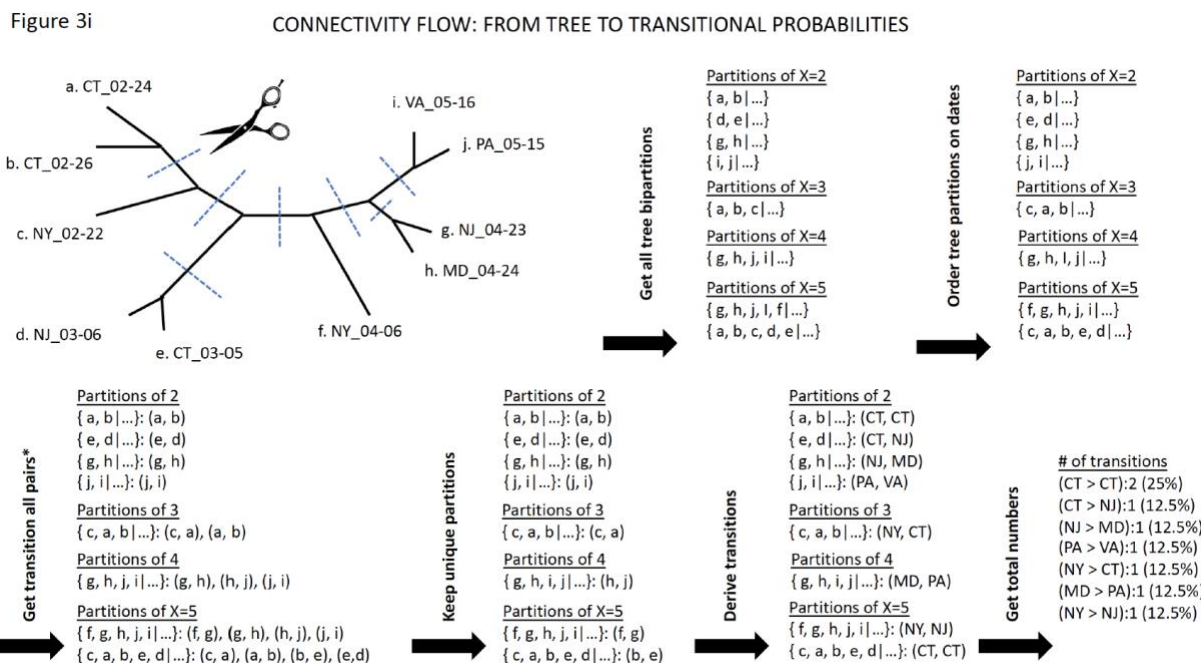
355 between the number of deaths per million individuals versus ii) the Distance from New York

356 city, iii) each state's maximum reproduction rate R_t , iv) each state's average neighboring
357 maximum R_t , and v) each state's urbanization index.

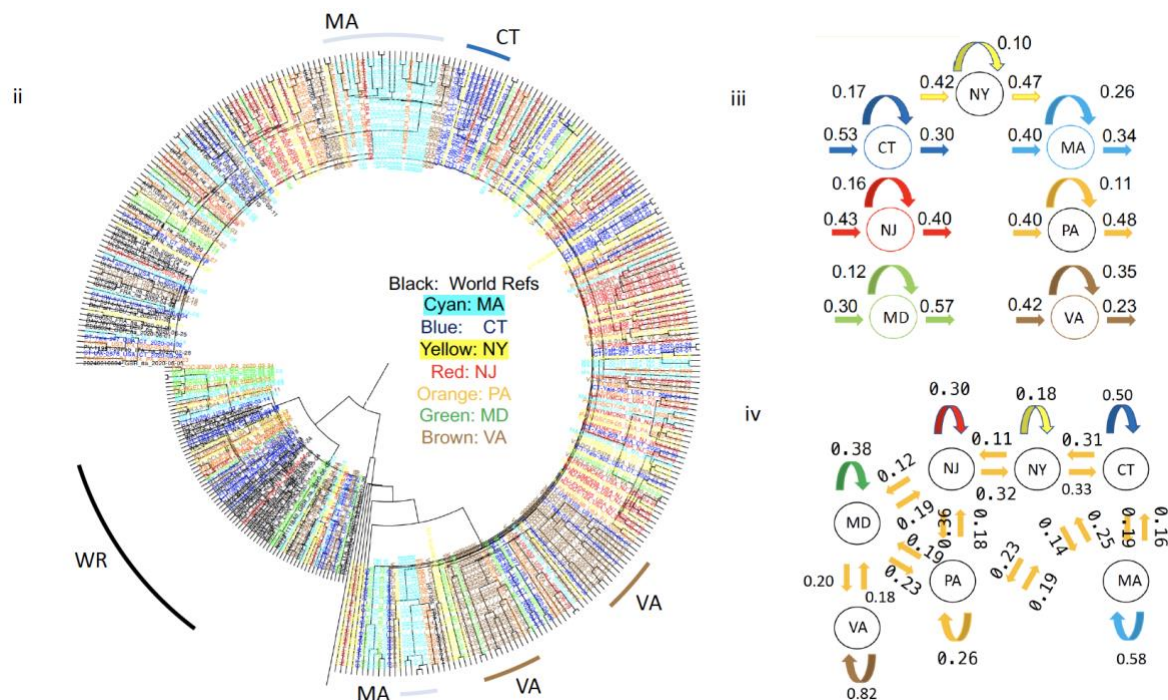


358
359 **Figure 2.** According to a Nextstrain adaptation, there were five main initial lineages of the
360 pandemic (19A, 19B, 20A, 20B, 20C), which can be used to suggest the original routes of the
361 transmission in the United States. In 2i) we show the topology of world reference sequences as
362 collected spanning the Nextstrain tree on June 25th. From these randomly collected sequences,
363 sample HF1465 FRA is the only sequence that consistently clusters with each state's major
364 outbreak (blue line). Two other reference sequences (from Italy and Germany) cluster -again
365 consistently- with each state's minor outbreak (black dotted line), suggesting that most of the
366 outbreak derives from these specific lineages. In (2ii) we show the unrooted tree of the New
367 York outbreak, which we consider as the outbreak epicenter. In (2iii-vi) we show the
368 phylogenetic tree analysis for Massachusetts, New Jersey, Virginia and Connecticut as rooted by
369 the older lineage that contains sequences from Wuhan dating in 2019.

370



371



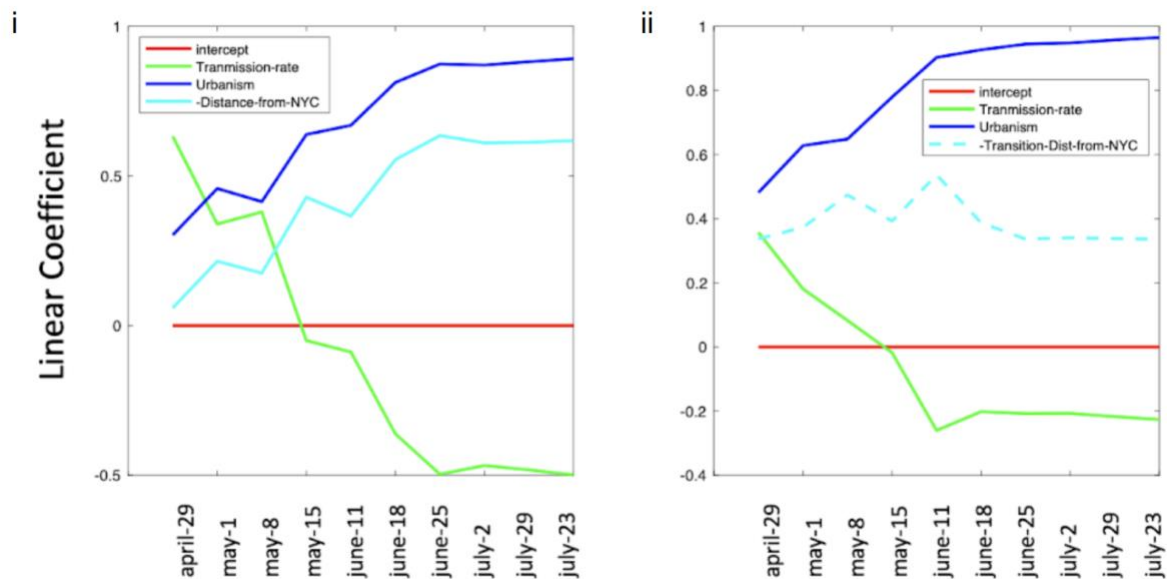
372

373 **Figure 3.** In i) we use a tree adaptation example to explain the workflow that we implemented

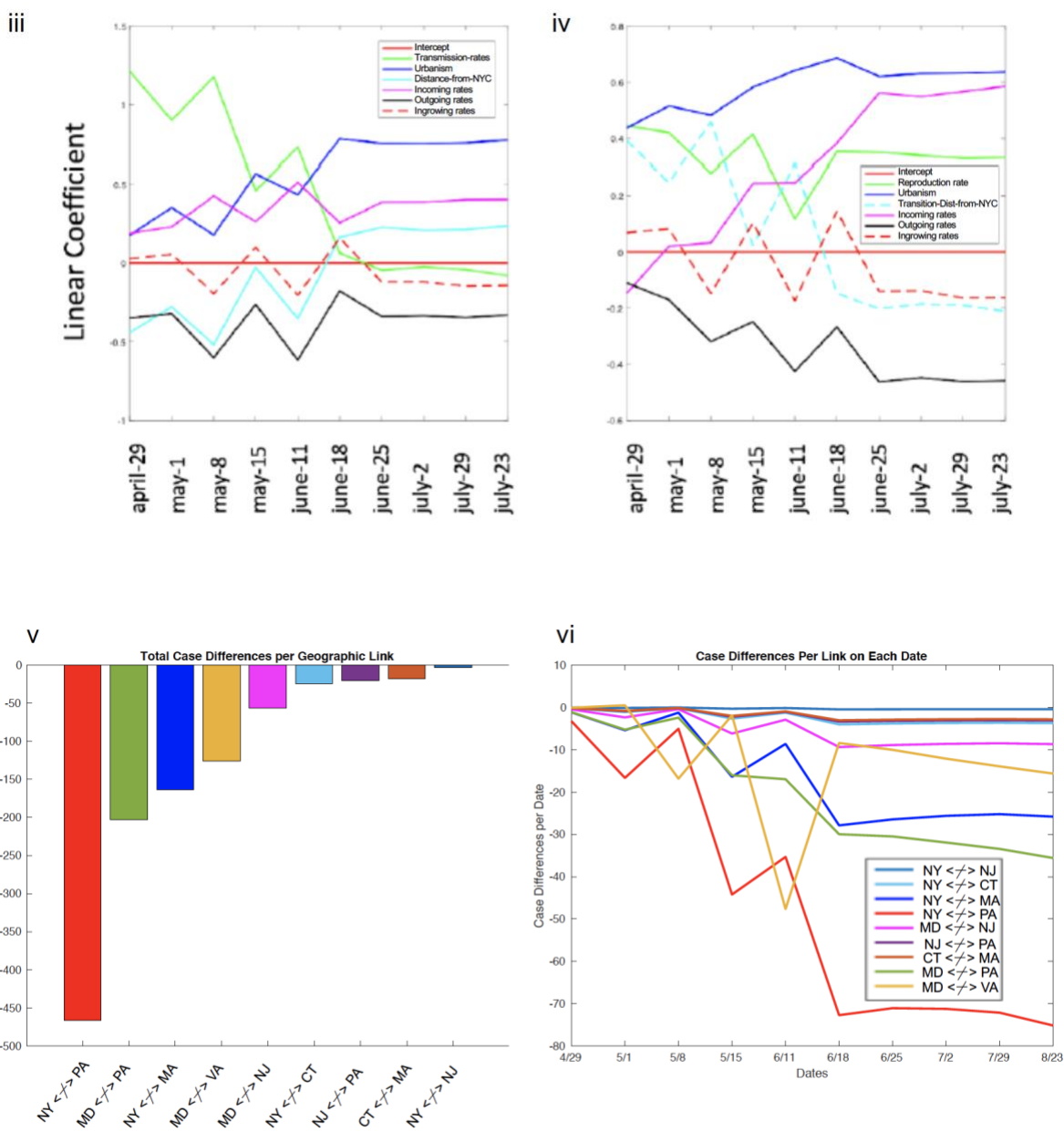
374 in order to assign directed connectivity, incoming, ingrowing and outgoing connections between

375 each state. In ii) using world reference sequences and selected reference sequences from 7 states,
376 we inferred a phylogenetic tree with time constraints for each state. Each sequence's tip color
377 corresponds to the state it was collected. Using pairing and dating information described in (i),
378 we derived iii) incoming, outgoing and ingrowing connectivity for each state and iii) transitional
379 connectivity between all states. For convenience, we only show neighboring and geographical
380 connectivity.

Figure 4



381



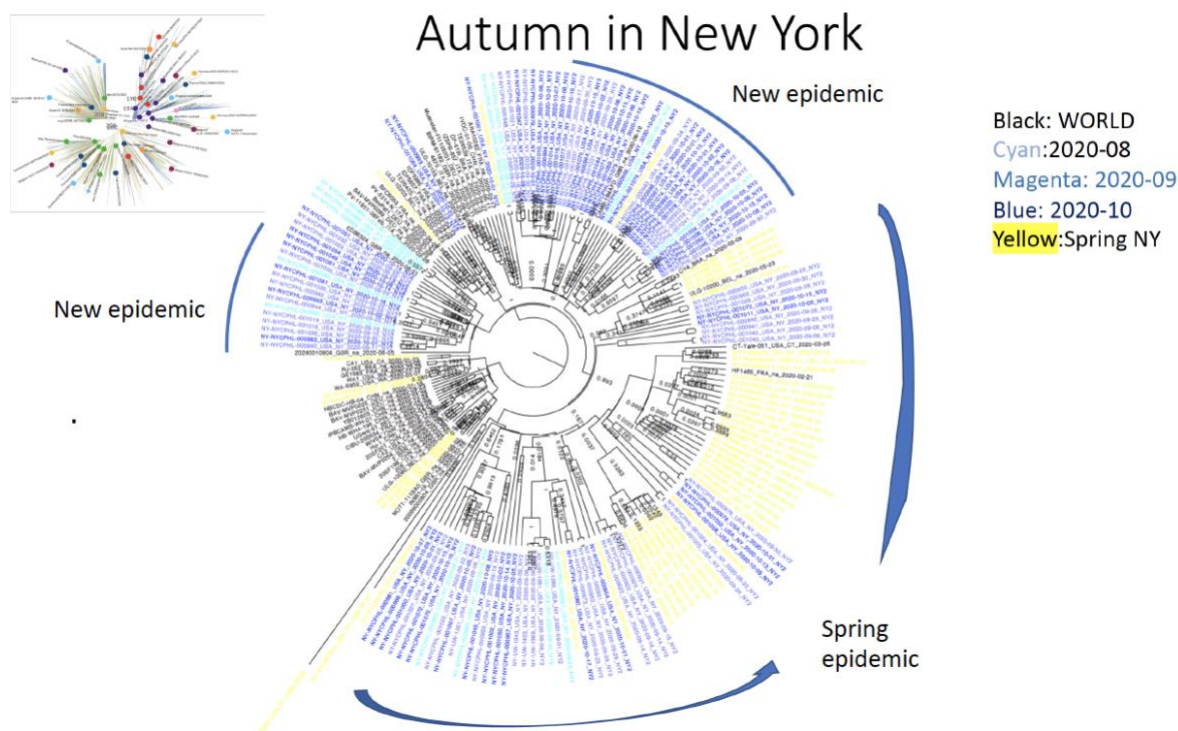
382

383

384 **Figure 4.** Predictive models with connectivity-based features. (i-ii) Models 1-2 (three factors),
 385 (iii-iv) Models 3-4 (six factors). Likelihood significance was found for models (1) vs (2) and (3)
 386 versus (4). (Model 1 vs. 2 / 3 vs. 4; $p=0.0003, 0.0273$ resp., 2-sided t-test for Pearson's r). We
 387 then estimated the sum of total deaths that would be saved if we remove any geographic link

388 between two states. In v) we show the total number of deaths per million individuals per case,
389 while in vi) we show the temporal distribution of these deaths, showing when specific links
390 become important. The link between NY and PA becomes important around May, while the link
391 between MD and PA a month later.

Figure 5



392
393 **Figure 5.** By inferring a phylogenetic tree using sequences from New York that were collected
394 after the 16th of August, together with previous world reference sequences and reference
395 sequences from New York during the first wave, we show that about half of the 2nd wave's
396 outbreak in New York constitutes a previously unseen outbreak, clustering with reference
397 sequences from Great Britain.

398

399

400

401

402

403 **BIBLIOGRAPHY**

404 1. Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res.* (2020). doi:10.1093/nar/gkz956

405 2. GISAID. GISAID Initiative. *Adv. Virus Res.* (2020).

406 3. Hadfield, J. *et al.* NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics*

407 (2018). doi:10.1093/bioinformatics/bty407

408 4. Leitner, T. *et al.* HIV Sequence Compendium 2008 Los Alamos HIV Sequence Database.

409 *HIV Seq. Compend.* (2008).

410 5. Kuiken, C., Hraber, P., Thurmond, J. & Yusim, K. The hepatitis C sequence database in

411 Los Alamos. *Nucleic Acids Res.* (2008). doi:10.1093/nar/gkm962

412 6. Candido, D. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*

413 (80-.). (2020). doi:10.1126/SCIENCE.ABD2161

414 7. Isabel, S. *et al.* Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein

415 mutation now documented worldwide. *Sci. Rep.* (2020). doi:10.1038/s41598-020-70827-z

416 8. Lemey, P. *et al.* Accommodating individual travel history and unsampled diversity in

417 Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* (2020).

418 doi:10.1038/s41467-020-18877-9

419 9. Seemann, T. *et al.* Tracking the COVID-19 pandemic in Australia using genomics. *Nat.*

420 *Commun.* (2020). doi:10.1038/s41467-020-18314-x

421 10. Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into

422 Northern California. *Science* (80-.). (2020). doi:10.1126/science.abb9263

- 423 11. Jorden, M. A. *et al.* Evidence for Limited Early Spread of COVID-19 Within the United
424 States, January–February 2020. *MMWR. Morb. Mortal. Wkly. Rep.* (2020).
425 doi:10.15585/mmwr.mm6922e1
- 426 12. Fauver, J. R. *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in
427 the United States. *Cell* (2020). doi:10.1016/j.cell.2020.04.021
- 428 13. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington State. *Science* (80-
429 .). (2020). doi:10.1101/2020.04.02.20051417
- 430 14. Worobey, M. *et al.* The emergence of SARS-CoV-2 in Europe and North America.
431 *Science* (80-.). (2020). doi:10.1126/SCIENCE.ABC8169
- 432 15. Xu, B. *et al.* Epidemiological data from the COVID-19 outbreak, real-time case
433 information. *Sci. Data* (2020). doi:10.1038/s41597-020-0448-0
- 434 16. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic
435 review and critical appraisal. *BMJ* (2020). doi:10.1136/bmj.m1328
- 436 17. Weinberger, D. M. *et al.* Estimation of Excess Deaths Associated with the COVID-19
437 Pandemic in the United States, March to May 2020. *JAMA Intern. Med.* (2020).
438 doi:10.1001/jamainternmed.2020.3391
- 439 18. Ioannidis, J. P. A., Axfors, C. & Contopoulos-Ioannidis, D. G. Population-level COVID-
440 19 mortality risk for non-elderly individuals overall and for non-elderly individuals
441 without underlying diseases in pandemic epicenters. *Environ. Res.* (2020).
442 doi:10.1016/j.envres.2020.109890
- 443 19. Eubank, S. *et al.* Commentary on Ferguson, et al., “Impact of Non-pharmaceutical

- 444 Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand". *Bull.*
445 *Math. Biol.* (2020). doi:10.1007/s11538-020-00726-x
- 446 20. Cacciapaglia, G., Cot, C. & Sannino, F. Second wave COVID-19 pandemics in Europe: a
447 temporal playbook. *Sci. Rep.* (2020). doi:10.1038/s41598-020-72611-5
- 448 21. Reiner, R. C. *et al.* Modeling COVID-19 scenarios for the United States. *Nat. Med.*
449 (2020). doi:10.1038/s41591-020-1132-9
- 450 22. Ferguson, N. *et al.* Report 9 - Impact of non-pharmaceutical interventions (NPIs) to
451 reduce COVID-19 mortality and healthcare demand | Faculty of Medicine | Imperial
452 College London. *Imp. Coll. COVID Response Team* (2020).
- 453 23. Jung, F., Krieger, V., Hufert, F. T. & Küpper, J. H. Herd immunity or suppression strategy
454 to combat COVID-19. *Clin. Hemorheol. Microcirc.* (2020). doi:10.3233/CH-209006
- 455 24. Orłowski, E. J. W. & Goldsmith, D. J. A. Four months into the COVID-19 pandemic,
456 Sweden's prized herd immunity is nowhere in sight. *J. R. Soc. Med.* (2020).
457 doi:10.1177/0141076820945282
- 458 25. Aschwanden, C. The false promise of herd immunity for COVID-19. *Nature* (2020).
459 doi:10.1038/d41586-020-02948-4
- 460 26. Farsalinos, K. *et al.* Improved strategies to counter the COVID-19 pandemic: Lockdowns
461 vs. primary and community healthcare. *Toxicol. Reports* (2021).
462 doi:10.1016/j.toxrep.2020.12.001
- 463 27. Fang, Y., Nie, Y. & Penny, M. Transmission dynamics of the COVID-19 outbreak and
464 effectiveness of government interventions: A data-driven analysis. *J. Med. Virol.* (2020).

465 doi:10.1002/jmv.25750

466 28. Brauner, J. M. *et al.* Inferring the effectiveness of government interventions against
467 COVID-19. *Science* (80-.). (2020). doi:10.1126/science.abd9338

468 29. Kevin Systrom, T. V. and M. K. Rt.live. (2020).

469 30. Mbuviha, R. & Marwala, T. Bayesian inference of COVID-19 spreading rates in South
470 Africa. *PLoS One* (2020). doi:10.1371/journal.pone.0237126

471 31. Panovska-Griffiths, J. *et al.* Determining the optimal strategy for reopening schools, the
472 impact of test and trace interventions, and the risk of occurrence of a second COVID-19
473 epidemic wave in the UK: a modelling study. *Lancet Child Adolesc. Heal.* (2020).
474 doi:10.1016/S2352-4642(20)30250-9

475 32. Stefana, A., Youngstrom, E. A., Hopwood, C. J. & Dakanalis, A. The COVID-
476 19 pandemic brings a second wave of social isolation and disrupted services. *European*
477 *Archives of Psychiatry and Clinical Neuroscience* (2020). doi:10.1007/s00406-020-
478 01137-8

479

480