COVID-19 epidemic severity is associated with timing of non-pharmaceutical interventions

Supplementary Methods

1. Overview of phylogenetics and phylodynamics pipeline

Genetic data cleaning and preparation

SARS-CoV-2 sequences were downloaded from GISAID (gisaid.org) on June 7th 2020. Sequences were removed if they were not sampled from human hosts or if sampling dates were not exact (day/month/year). We dropped 80% of sequences collected from the UK after March 15th to reduce bias due to sampling[1]. Remaining sequences were aligned using MAFFT v7[2]. After alignment, we removed sequences with >20% of nucleotide sites missing and cut sequences to the beginning of the first and the end of the last open reading frames. To eliminate badly aligned sequences and sequences problematic for time-resolved phylogenetic analyses, we performed an additional round of data cleaning. We split our data into subsets (~2,000 sequences) and to each subset added a set of 500 sequences spanning the time period from the first SARS-CoV-2 sample (24/12/2019) to the last sample date. We constructed maximum likelihood phylogenies for each subset using IQtree v1.6[3], and dropped sequences considered to be outliers. Outliers were defined as sequences a) with a mean cophenetic distance >= 3 standard deviations from the mean phylogeny cophenetic distance or b) that did not conform to the molecular clock based on a time-scaled analysis with *treedater* v0.5.0[4]. We included only unique sequences; where replicate sets existed, we removed all but the earliest sequence. All data cleaning was performed in R v3.6.1.

Inclusion/ exclusion of sites

Sites were eligible for analysis if there were at least 100 hundred sequences available from that location on GISAID on June 7th (n=78). Fourteen sites with fewer sequences were also analysed, for reasons explained below. Among the 82 sites, we then excluded sites for the following reasons. Our model requires samples to be collected at random across a population and with a range of dates that enables reconstruction of a molecular clock. We excluded locations where samples were known to have been collected as a result of contact tracing or where travellers had been preferentially sequenced (n=8) [5,6] . Unfortunately, that information was unavailable for many sites. We chose to exclude identical sequences in case they resulted from contact tracing; but this choice introduces a different kind of bias, as groups of identical sequences are a feature of early rapidly spreading epidemics[7]. Fortunately, in our simulations, exclusion of identical sequences from different individuals did not overly bias results (see below). When data were available for sites located within each other (e.g. New Orleans in Louisiana), the smaller geographic unit was preferentially selected (n=21), and some regions were excluded because they were too large geographically to fit our model assumption of random mixing (n=3). One exception to the former rule is Valencia, which was analysed as

"Communitat Valenciana" because labeling of the latter was more systematic. Wuhan and Hubei were not analysed because we could not have estimated viral origin without including non-human samples. Fourteen sites with <100 sequences were analyzed because these regions were among the first on GISAID to have at least 20 sequences available. Fifty-seven sites were included in our final analysis. Details of inclusion/ exclusion and sample sizes for each site are displayed in Supplementary Table 1.

Model-based phylodynamic inference of epidemic size and reproduction numbers.
For each geographical location under investigation, we selected up to 150 unique regional sequences from the GISAID alignment, as well as exogenous sequences representing the international reservoir. Fifty exogenous sequences encompassing the full time-range of GISAID samples were selected each time at random as background, and to these we added sequences from GISAID that were <= 2 substitutions away from the sequences in the regional dataset calculated. Pairwise genetic distances were calculated using TN93 (https://github.com/veg/tn93). For each regional dataset, we then constructed a phylogeny in IQtree. Polytomies were resolved at random 10 times, each time generating a new starting tree for the analysis in BEAST2, totalling 10 independent chains.

The phylodynamic model is designed to estimate epidemiological parameters from sequence data and is implemented in BEAST2. The model of epidemic dynamics is based on a susceptible-exposed-infectious-recovered (SEIR) model and is described in the next section.

Each of the 10 runs was set up for 20 million steps. Subsequently, log files were examined for convergence in Tracer v1.7.1, problematic runs excluded, and log files and trajectory files were combined and cleaned using the sarscov2 R package (available at http://github.com/emvolz-phylodynamics/sarscov2Rutils).

Estimating the time of regional viral introductions
For each region, we estimated the timing of viral introductions through time-resolved phylogenetic analysis and parsimony reconstruction. We included all sequences available for that region as well as all close exogenous sequences (<= 2 substitutions away) in a maximum likelihood (ML) tree, built using IQtree. Within the ML tree, we resolved polytomies at random, and estimated rooted time-scaled phylogenies using *treedater*, repeating the procedure 100 times. The mean clock rate of evolution was constrained between 0.00075 and 0.0015 substitutions per site per year. Branch lengths were smoothed by enforcing a minimum number of substitutions per site on each branch and by sampling from the distribution estimated by *treedater*. Finally, we reconstructed the ancestral state of nodes and dated and counted importation events, repeating the procedure 25 times. We calculated the weighted mean from the distribution of viral introduction times and call it the central epidemic seed time (CEST). All functions are available and documented within the sarscov2 R package (https://github.com/emvolz-phylodynamics/sarscov2Rutils). For each site we then calculated the time between CEST and the maximum NPI and looked for relationships between this delay and the severity of the epidemic at each site. Ou results are reported within the main body of the text. As a sensitivity analysis, as well as the CEST, we also calculated the 5th and 25th percentiles of the distribution of viral introduction times and recalculated the delay to maximum NPI for each definition. Results stayed broadly consistent with those from our analysis using CEST, but effects were stronger for CEST than for other definitions of seeding time (data not shown).

Non-parametric phylodynamic inference
We applied a *skygrowth* model[8,9] (version 0.3.1) to estimate viral effective population size through time and growth rates of effective population size which under appropriate conditions can be used as a proxy statistic for epidemic prevalence[10]. Growth rates were estimated using *skygrowth* 0.3.1[8] using Markov chain Monte Carlo (MCMC) and 500 thousand iterations for each time tree and using an Exponential(10^{-4}) prior for the smoothing parameter. This method was applied to trees generated as described above for estimating time of viral introductions. The final results were produced by averaging across 100 time trees estimated for each region. Code to reproduce this analysis is contained in the sarscov2 R package (*skygrowth1* function, https://github.com/emvolz-phylodynamics/sarscov2Rutils).

2. SEIJR phylodynamic model reconstruction of simulated epidemics

Model for Infectious disease dynamics: Susceptible-Exposed-Infected(IJ)-Recovered

Terminology: The phylodynamic model is designed to estimate epidemiological parameters using a combination of sequence data from a *region* (e.g. a city, county or other small territory) and *exogenous* sequences from a much larger international reservoir.

Essential metadata: Location (region or exogenous) and date of sampling.

Mathematical model: The mathematical model is based on previous development of SEIR-type models for Ebola virus[11] and implemented in a structured coalescent framework in the PhyDyn package[12]. A related model was applied in the early stage of the SARS-CoV-2 epidemic to estimate global case numbers[13] and has also been applied in studies of local Chinese[14] and Israeli[5] SARS-CoV-2 sequence data. The phylodynamic model is designed to account for

- Nonlinear epidemic dynamics in the region,
- A realistic distribution of generation times with incubation and infectious periods,
- Migration of lineages between region and exogenous demes; and
- Variance in transmission rates which has a large influence on epidemic size estimates.

The model of epidemic dynamics within a region is based on a susceptible-exposed-infectious-recovered (SEIR) model. We elaborate this model with an additional compartment J which has a higher transmission rate (τ-fold higher) than the I compartment. Upon leaving the incubation period individuals progress to the J compartment with probability p_h, or otherwise to I.

The model is implemented as a system of ordinary differential equations:

$$\dot{S}(t) = -\left(\beta I(t) + \beta\tau J(t)\right)\frac{S(t)}{S(t)+E(t)+I(t)+J(t)+R(t)}$$
$$\dot{E}(t) = \left(\beta I(t) + \beta\tau J(t)\right)\frac{S(t)}{S(t)+E(t)+I(t)+J(t)+R(t)} - \gamma_0 E(t)$$
$$\dot{I}(t) = \gamma_0(1-p_h)E(t) - \gamma_1 I(t)$$
$$\dot{J}(t) = \gamma_0 p_h E(t) - \gamma_1 J(t)$$
$$\dot{R}(t) = \gamma_1(I(t)+J(t))$$

Parameters: β is the per-capita transmission rate. τ is the ratio of transmission rate in the high to low risk categories. γ_0 is the rate of progression from incubating individuals to infectious individuals (note that this does not describe which individuals are symptomatic). γ_1 is the rate of recovery once infectious.

We also model an exponentially growing (rate ρ) reservoir $Y(t)$ for imported lineages into the *region*.

Migration is modeled as a bidirectional process which only depends on the size of variables in the region compartment and thus migration does not influence epidemic dynamics; it will only influence the inferred probability that a lineage resides within the region.
For a compartment X (E,I, or J), η is the per lineage rate of migration out of the region and the total rate of migration in and out of the region is ηX.

Table S A: Parameters and priors of the SEIJR model.

Parameter	**Symbol**	**Prior**
Initial infected	E_0	Exponential(1)
Initial susceptible	S_0	Exponential[1]
Migration rate	η	Exponential(10)[2]
Reproduction number	R_0	Lognorm(0.88, sd log=0.5)
Clock rate	ω	Uniform(.0005,.005)[3]
Transition/transversion	κ	Lognorm(1, sd log=1.25)

1. Prior mean for susceptible population was calibrated to individual locations based on population size.
2. Units: Migrations per lineage per year. Maximum value = 10.
3. Units: Substitutions / site / year

During phylodynamic model fitting β and ρ are estimated. Additionally, we estimate initial sizes of Y, E, and S. Other parameters are fixed based on prior information. We fix $1/\gamma_0 = 5$ days and $1/\gamma_1 = 3$ days. Parameters controlling overdispersion in transmission rates (p_h and τ) are estimated with strong priors which yields a dispersion of the reproduction number that matches a negative binomial distribution with $k = 0.22$ if $R_0 = 2$, similar to values estimated for the 2003 SARS epidemic[15].

3. Simulation of epidemics under the SEIJR model

In order to evaluate the ability of our SEIJR model to reconstruct phylodynamic history and estimate epidemic parameters, we simulated epidemics with known parameters. Thirty combinations of parameters were sampled from the uniform distributions shown in Table S B using latin hypercube sampling as implemented in the lhs R package[16]. Other parameters were fixed (Table S B).

From the 30 combinations of parameter values, we calculated the cumulative number of infections for each using phydynR v.0.2.0[17]. We then selected five sets of parameters that displayed a diversity of outcomes (different numbers of cumulative infections, as well as showing recent decreases). These parameters are shown in Table S C.

We next simulated phylogenetic trees based on the structured coalescent using the function *sim.co.tree* in phydynR. Tips in the phylogenies trees belonged to two different compartments: regional (I, n=100) and exogenous (*Y*, n=50). The seeding time of the epidemic as a whole was set to the beginning of December (2019.92), the seeding time for the regional epidemic was sampled from a uniform distribution (Table S B). Sampling began in January 2020 (2020.0).

Sequences were simulated from the phylogenies using seq-gen[18], as implemented in the phyclust R package[19]. We used an Hasegawa-Kishino-Yano (HKY) DNA substitution model[20] with a transition/transversion rate of 5.5, a clock rate of 0.001 nucleotide substitutions/ year, and relative base frequencies for the frequency of nucleotides A, C, G and T of 0.3, 0.2, 0.2 and 0.3, respectively. Sequences generated were 29,500 bases. We then used a customized function in R to deduplicate the DNA sequence alignment. All scripts to reproduce our simulations are available on github (github.com/thednainus/sarscov2simulations).

Analysis of simulated data using BEAST2

Five test datasets for analysis were generated using simulation (for parameters of each, see Table S C), and each was analysed in totality and in deduplicated form. Each dataset was processed in BEAST2 using the SEIJR model and parameter priors as described in Table S A to generate the effective reproduction number R_0, the reproduction number through time, R_t, and an estimated number of infections . True values for daily and cumulative infections fell within 95% highest posterior density (HPD) estimates for 3 out of 5 simulations (Supplementary Figure 5, simulations are labelled 1 to 5). For simulations 3 and 4 estimates of the number infected were below the true value. R_0 true values fell within 95% estimated HPD for R_0 in 4 of 5 simulations. Longitudinal estimates of R_t were within the 95% HPD in 3 out of simulations. In 4 out of 5 simulations R_t was set to decrease and that drop was captured in all 4 of the reconstructions within +/- 1 week of the true drop. De-duplication of the data based on

sequence identity did not bias results but tended to increase HPD. Timings of decrease in R_t were unaffected by deduplication.

Table S B. List of parameters that were sampled from a uniform distribution or fixed in our simulations. *U* denotes Uniform distribution. All rate parameters have units of 1/year.

Parameters	Values
Transmission rate β	U(15, 25)
Initial number of susceptible individuals	U(1,0000, 1,000,000)
Importation rate η	U(1, 10)
Start time for sampling in the region	U(2020.10, 2020.15)
Initial number of exposed individuals	U(1, 30)
Transmission risk ratio τ	Fixed at 74
Proportion high-risk p_h	Fixed at 0.2
Exogenous growth rate	Fixed at 25
Rate of disease progression γ_0	Fixed at 73
Rate of recovery γ_1	Fixed at 121.667

Table S C. List of randomly generated parameters values used to simulate phylogenetic trees.

Simulation number	β	R_0	**Initial susceptible**	η	**Regional epidemic start**	**Initial E**
1	23.93111	3.07	272824.10	5.236837	2020.137	19.661442
2	17.38683	2.23	840389.11	9.113708	2020.111	2.882301
3	21.13614	2.71	596454.58	1.931835	2020.103	25.039326
4	24.69453	3.16	309257.92	1.727232	2020.106	29.417566
5	15.28835	1.96	47565.01	7.214645	2020.140	22.193182

References

1 Hall MD, Woolhouse MEJ, Rambaut A. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: A simulation study. *Virus Evol* 2016; **2**: vew003.

2 Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002; **30**: 3059–66.

3 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; **32**: 268–74.

4 Volz EM, Frost SDW. Scalable relaxed clock phylogenetic dating. *Virus Evol* 2017; **3**. DOI:10.1093/ve/vex025.

5 Miller D, Martin MA, Harel N, *et al.* Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Genetic and Genomic Medicine. 2020; published online May 22. DOI:10.1101/2020.05.21.20104521.

6 Seemann T, Lane C, Sherry N, *et al.* Tracking the COVID-19 pandemic in Australia using genomics. *medRxiv* 2020. https://www.medrxiv.org/content/10.1101/2020.05.12.20099929v1.abstract.

7 Lu J, du Plessis L, Liu Z, *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* 2020; **181**: 997–1003.e9.

8 Volz EM, Didelot X. Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Syst Biol* 2018; published online Feb 7. DOI:10.1093/sysbio/syy007.

9 Fountain-Jones NM, Appaw RC, Carver S, Didelot X, Volz E, Charleston M. Emerging phylogenetic structure of the SARS-CoV-2 pandemic. 2020; : 2020.05.19.103846.

10 Frost SDW, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. *Philos Trans R Soc Lond B Biol Sci* 2010; **365**: 1879–90.

11 Volz E, Pond S. Phylodynamic analysis of ebola virus in the 2014 sierra leone epidemic. *PLoS Curr* 2014; **6**. DOI:10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e.

12 Volz EM, Siveroni I. Bayesian phylodynamic inference with complex models. *PLoS Comput Biol* 2018; **14**: e1006546.

13 Volz E, Baguelin M, Bhatia S, Boonyasiri A, Cori A. et al. Report 5: phylogenetic analysis of SARS-CoV-2. https://doi.org/10.25561/77169.

14 Volz E, Fu H, Wang H, *et al.* Genomic epidemiology of a densely sampled COVID19

outbreak in China. *medRxiv* 2020.

15 Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. *Nature* 2005; **438**: 355–9.

16 Carnell R. lhs: Latin hypercube samples. *R package version 0 10, URL http://CRAN R-project org/package= lhs* 2012.

17 phydynR. Github https://github.com/emvolz-phylodynamics/phydynR (accessed Sept 6, 2020).

18 Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* 1997; **13**: 235–8.

19 Chen W-C. Phylogenetic Clustering with R package phyclust, 2010. *URL http://thirteen-01 stat iastate edu/snoweye/phyclust*; **54**.

20 Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 1985; **22**: 160–74.