ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

A comprehensive study of the phylodynamics of SARS-CoV-2 in Europe

Master Thesis

Cecilia Valenzuela Agüí

`ceciliav@student.ethz.ch`

Department of Biosystems Science and Engineering
Computational Evolution
ETH Zürich

Supervisors:
Prof. Dr. Tanja Stadler, Dr. Timothy Vaughan, Sarah Nadeau

March 5, 2021

Acknowledgements

I thank Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam non-umy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Abstract

The abstract should be short, stating what you did and what the most important result is. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Contents

CHAPTER 1

Introduction

We have many sources of information, but all of them are imperfect in some aspect. Bayesian phylodynamics incorporates different beliefs and data to get a detailed picture of the epidemic. Genetic sequences as an objective source of information about virus evolution and transmision together with viral phyloge- netics, transpotation data and geographic data to understand the dynamics and case counts to guide the magnitutde of the numbers.

We lack absolute numbers of the pandemic time resoluted without reporting issues dependent. We lack understading of the dynamics, from where to where, how often, how long.

We can use travel data as a proxy, or we can incorporate it in a model. Lemey et al. Difference we get absolute numbers, not only the rates. And we use BD model instead of coalescent, better to describe beginning of a epidemic. While due to the complexite and time consuming model we dow scaled to europe and only thhe initial phase with the first countries to experiment relevant sars cov 2 epidemics.

the first reported cases were in China in early December [],

In a phylodynamic analysis, we aim to integrate genetic and epidemiological information to understand pathogen evolution and transmission dynamics. We can learn about the epidemic spread and the interactions between hosts from the imprint that these events leave in virus phylogenies. This is possible for RNA viruses, and in specific for SARS-CoV-2, because virus genetic evolution and epidemic processes are in the same time scale [?] [?]. More precisely, the field of phylodynamics studies how phylogenetic trees are being generated and infers the population parameters behind that process.

Better in the
Introduction?

Material and Methods

SARS-CoV 2 genome data set

We create a dataset of 360 genetic sequences from December 2019 to March 8 2020, obtained from publicly available data on GISAID [**?**] (accessed on November 2020). We follow the Nextstrain workflow for the curation of the dataset [1]. Sequences with incomplete collection date, less than 27.000 bases in length or with more than 3.000 unknown bases are omitted. Also, sequences from known clusters of transmission or from the same patient are excluded. The resulting dataset of x sequences is aligned with MAFFT. The begining and the end of the alignment are masked respectively by 100 and 50 sequences as well as sites 13402, 24389 and 24390, identified by Nextstrain as prone to sequencing errors.

> look number of seqs

To focus on the early dynamics in Europe we select sequences from China, origin of the epidemic; France and Germany, the European countries with the earliest cases; and Italy and Spain, the European countries with the biggest outbreaks in March. To take into account the dynamics in other regions of Europe we include a group of 50 sequences from other European countries. We limit our sample of Chinese genomes to sequences until January 23, the starting date of the lockdown in Hubei.

Due to the large (and unprecendented) number of available genetic sequences for SARS-CoV-2, we need to subsample the alignment. Each sequence is subsampled with a probability equal to the probability of having a reported case in that country the day of sample collection, inversely weighted by the probability of having a sequence in GISAID that day in the country. With this subsampling protocol, we aim to get a constant sampling proportion across the full period for each country.

This dataset of genetic sequences is the main source of information in our phylodynamic analysis. The goal is to infer the phylogenetic tree, i.e. the evolutionary tree-shaped relationship among the sequences, together with the epidemiological transmission parameters that gave rise to it. These parameters are defined within a population dynamic model and will inform us about the epidemic that the viral genetic sequences come from.

The multitype birth death model

To study the early dynamics of SARS-CoV-2 in the European countries, we use a simplified version of the multitype birth-death model described in [2], following the analysis in [3]. Birth death models are compartmental population models with high flexibility that describe the process of epidemiological transmission. The stochastic formulation of these models are used in phylodynamic analyses [4]. In the multitype version, we consider a structured population in types or subpopulations with characteristic within-subpopulation dynamics and migrations between them. In our case, the subpopulations are the different locations of the samples.

The process starts with one infected host in one of the subpopulations, e.g in subpopulation i, who can infect another individual at rate λ_i, become uninfectious at rate μ_i by death or recovery, migrate to another deme j at rate m_{ij} or be sequenced at rate ψ_i to become part of the phylogenectic tree. This process depicts the full transmission dynamics and specifically, the generation of the phylogenetic tree that we observed from our sequence data.

Under this model, we are able to compute the likelihood of the multitype birth-death parameters for a given tree. This likelihood is derived in [2] by considering the probability of an individual evolved as observed in the tree. This derivation is analogous to the work in [5], which is based on ideas from [6].

We parameterize our model in terms of the effective reproductive number $R_i = \frac{\lambda_i}{\mu_i + \psi_i}$, a key value in epidemic control and understanding, the rate of becoming uninfectious $\delta_i = \mu_i + \psi_i$, the probability of and individual to be sequenced $s_i = \frac{\psi_i}{\mu_i + \psi_i}$ and the migration rates between locations m_{ij}.

Epidemic trajectories and structured trees

We will refer to the full sequence of transmissions, recoveries/deaths, migrations and sampling events as an epidemic trajectory. One set of sequences, and therefore one phylogenetic tree, is the product of one epidemic trajectory. Moreover, the tree represents a fraction of the events in the epidemic trajectory, those that involve the sampled individuals. In epidemiology, we aim to learn about the true epidemic trajectory since we usually have incomplete information caused by unreported cases or unknown transmission chains.

From the sequences metadata, we know the location of the tree tips. However, if we know the epidemic trajectory we also know the location of the lineage at any point in time in the tree. We will refer to the phylogenies with ancentral locations mapped onto the tree as structured trees [7]. In Fig 2.2 we show the epidemic trajectories corresponding to two different subpopulations and the structured tree of a set of samples. The change of ancestral location, represented by a change in color from blue to red in the tree, is caused by a migration event from one subpopulation to the other, depicted in the epidemic trajectories.
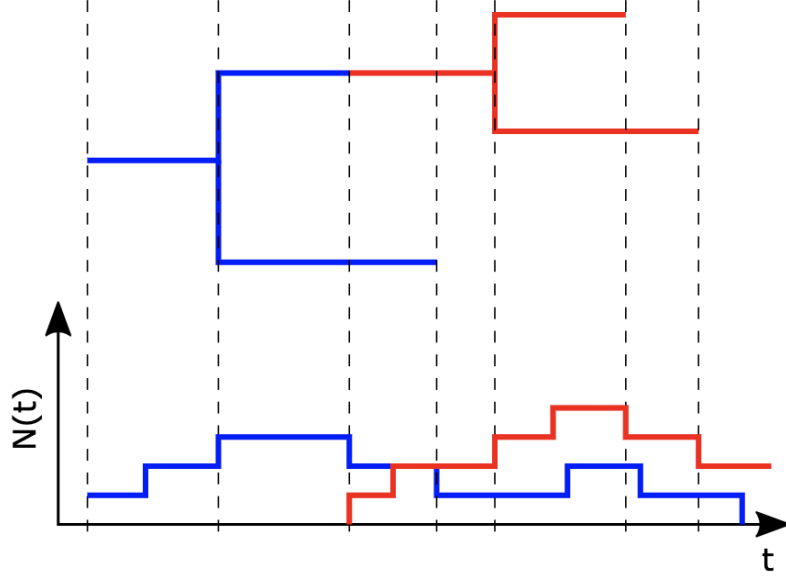
Figure 2.1: Example of structured tree (up) and epidemic trajectories (bottom) for two types. The horizontal axis represents time, and for the epidemic trajectories the vertical axis is the population size N(t). Figure from Tim Vaughan.

Use figure from EpinInf paper?

Bayesian Inference with BDMM-Prime

We want to infer the epidemic trajectories under the multitype birth-death model fit to a set of samples collected throughout SARS-CoV-2 early epidemics in Europe. In order to do this, we have to estimate the joint posterior distribution of the structured tree $\mathcal{T}_c$, the epidemic trajectory $\mathcal{E}$, the substitution model parameters θ and the multi-type birth death parameters η. This posterior probability can be expressed with Bayes' formula as:

$$P(\mathcal{T}_c, \mathcal{E}, \theta, \eta | A) = \frac{P(A|\mathcal{T}_c, \mathcal{E}, \theta, \eta)P(\mathcal{T}_c, \mathcal{E}, \theta, \eta)}{P(A)} \tag{2.1}$$

where A is the sequence alignment.

We make the following independence assumptions:

$$P(A|\mathcal{T}_c, \mathcal{E}, \theta, \eta) = P(A|\mathcal{T}, \theta) \tag{2.2}$$

$$P(\mathcal{T}_c, \mathcal{E}, \theta, \eta) = P(\mathcal{T}_c|\mathcal{T}, \eta)P(\mathcal{E}|\mathcal{T}_c, \eta)P(\mathcal{T}|\eta)P(\theta)P(\eta) \tag{2.3}$$

Where $\mathcal{T}$ is the rooted time tree without ancestral locations. Thus, we can express Equation 2.1 in terms of the conditional distributions:

$$P(\mathcal{T}_c, \mathcal{E}, \theta, \eta | A) = P(\mathcal{T}_c | \mathcal{T}, \eta) P(\mathcal{E} | \mathcal{T}_c, \eta) \frac{P(A | \mathcal{T}, \theta) P(\mathcal{T} | \eta) P(\theta) P(\eta)}{P(A)} \tag{2.4}$$

$$= P(\mathcal{T}_c | \mathcal{T}, \eta) P(\mathcal{E} | \mathcal{T}_c, \eta) P(\mathcal{T}, \theta, \eta | A) \tag{2.5}$$

We use a Markov chain Monte Carlo (MCMC) Metropolis-Hastings algorihtm to approximate this posterior. Since the MCMC only uses ratios of posterior probabilities, we avoid calculating the marginal likelihood $P(A)$. This algorithm is implemented in BDMM-Prime package [8] for BEAST 2.6.3 [9]. BDMM-Prime first samples from $P(\mathcal{T}, \theta, \eta | A)$, then in a second step these samples are augmented by sampling from $P(\mathcal{T}_c | \mathcal{T}, \eta)$ and $P(\mathcal{E} | \mathcal{T}_c, \eta)$ and adding these variables to obtain the overall posterior $P(\mathcal{T}_c, \mathcal{E}, \theta, \eta | A)$. In this way, $P(\mathcal{T}_c | \mathcal{T}, \eta)$ and $P(\mathcal{E} | \mathcal{T}_c, \eta)$ are only calculated for a subset of samples instead of every MCMC step.

While it is also possible to include $\mathcal{T}_c$ and $\mathcal{E}$ directly into the Bayes' rule as implemented in EpiInf package[10], the factorization of the posterior in these three terms allow us to use the standard birth-death-sampling tree prior implementations in BDMM [2] [11] to compute $P(\mathcal{T}, \theta, \eta | A)$. This will speedup the analysis with almost no overhead compared to the standard BDMM inference and can be used in pre-existing multitype analyses withoud additional MCMC.

Stochastic mapping of ancestral locations and epidemic trajectories

To sample from $P(\mathcal{T}_c | \mathcal{T}, \eta)$, BDMM-Prime implements a stochastic mapping algorithm based on the work by [?]. A set of differential equations is numerically integrated over $\mathcal{T}$ to obtain the marginal probability of a location value at any point along each branch. From this probabilities, we can derive time-dependent rates that define the changes of the location along the tree according to a continuous-time Markov process. Then, we can simulate forward in time from the root of the tree the location trajectories down the tree edges. For a detailed derivation of the stochastic mapping of the algorithm refer to [?].

In the case of $P(\mathcal{E} | \mathcal{T}_c, \eta)$, following Bayes'rule:

$$P(\mathcal{E} | \mathcal{T}_c, \eta) = \frac{P(\mathcal{T}_c, \eta | \mathcal{E}) P(\mathcal{E})}{P(\mathcal{T}_c, \eta)} = \frac{P(\mathcal{T}_c | \mathcal{E}) P(\mathcal{E} | \eta)}{P(\mathcal{T}_c | \eta)} \tag{2.6}$$

We can easily compute the probability of a structured tree for a given epidemic trajectory $P(\mathcal{T}_c | \mathcal{E})$ as described in [10]. Each node in the structured tree must correspond to a compatible event in the epidemic trajectory for this probability to be nonzero. If we simulate an epidemic trajectory directly from $P(\mathcal{E} | \eta)$, it is very likely that the simulated events will not match with the events in the tree and $P(\mathcal{T}_c | \mathcal{E})$ will be 0. To avoid this problem, we simulate from $P^*(\mathcal{E} | \eta)$, which guarantees trajectories with non-zero probabilities by enforcing the tree events.

Provided we weight these trajectories accordingly we can efficiently sample from $P(\mathcal{E}|\mathcal{T}_c, \eta)$:

$$P(\mathcal{E}^{(a)}|\mathcal{T}_c, \eta) = w_a^* P^*(\mathcal{E}^{(a)}|\eta) \propto P(\mathcal{T}_c|\mathcal{E}^{(a)}) \frac{P(\mathcal{E}^{(a)}|\eta)}{P^*(\mathcal{E}^{(a)}|\eta)} P^*(\mathcal{E}^{(a)}|\eta) \qquad (2.7)$$

The epidemic trajectories are simulated with an adaptive tau-leaping algorithm [12] in BDMM-Prime. This method is based on the Gillespie algorithm [13] but the simulation time is divided in small intervals and the number of events is drawn from a poisson distribution for each interval. This algorithm is more efficient that Gillespie when we have a high number of individuals since we have to update the rates less often.

We simulate a fixed number of trajectories, called particles, for each set of population parameters η and structured tree $\mathcal{T}_c$. The trajectories are simulated until some point and are weigthed as described in Equation 2.7. We resample using importance sampling to get rid of low-weighted trajectories and then resume the simulation of all the particles from the sampled trajectory ending point. This process is repeated until the end of the simulation. Then, we sample one single trajectory from the set of final trajectories with important sampling and record each event and its timing to future analysis [?]. All these steps are implemented by BDMM-Prime trajectory logger and we use it in our analysis.

We conducted two different analyses. One with constant migration rates as in the standard multi-type birth death model and the other one with time-changing migration rates informed by travel data. We use the BDMM-Prime package [8] implemented in BEAST 2.6.3 [?]. We run 10 parallel chains for analysis, of 10^7 iterations each with different initial values. These chains are assesed for convergence using Tracer v.1.7.1 and then combined after removing a 10% burnin. We check that the effective sample size (ESS) is greater than 200 for all parameters. The stochastic mapping of structured trees and epidemic trajectories is performed in a second step using the trace and tree logs from the MCMC. We use 300, 1000, 3000 and 10000 particles in the tau-leaping simulation of epidemic trajectories, and a tolerance of 0.03 for selecting tau leap length.

Incorporating air travel data and geographic distances

In order to inform our migration rates with external information, and to avoid having to estimate a large number of migration rates we implement a generalized linear model (GLM) for the migration matrix based on [14] and [?]. This GLM model parametrizes the migration rates as a log linear function of a number of potential predictors, Equation 2.8. For each predictor x_i, the GLM parameterization includes a coefficient β_i, which quantifies the effect size of the

predictor, and a binary indicator variable δ_i, that allows the predictor to be included or excluded from the model. The parameter c represents the overal magnitude of migration. This means that if every indicator is 0, every migration rate will be equal to this parameter. In the MCMC, we estimate the effect size of each of the predictors, as well as their inclusion probability $\mathbb{E}[\delta_i]$.

$$m_{ij} = c\exp(\sum_{i=1}^{p}\delta_i\beta_i xij) \tag{2.8}$$

In a similar way to the GLM formulation in [15] we consider as predictors air travel data and geographical distance between locations. We define three time periods in our analysis: from the origin of the epidemic to January 23 (Hubei lockdown), from January 23 to end of February, from March 1 to March 8. The international travel during these months decreased drastically due to the increasing awareness about the SARS-CoV-2 pandemic as is shown in Figure ??. Thus, we would expect the migration rates to change during the time of the analysis.

The air travel data is obtained from EUROSTATS transport datasets *avia_ paexcc* and *avia_ paincc*, in particular the passengers carried departures values for December 2019, January, February and March 2020. We compute the average daily flux of passengers for each of the time periods. The geographical distance is defined as the great circle pairwise distance between the centroids of the countries, computed based on the Natural Earth project (the 1:50m resolution version) world map 2013. The predictors are log transformed, a pseudocount is added to make all values positive, and then they are standardized, following the description in [14].

Model specifications and priors

Since we focus on the early epidemic outbreak, we expect un unimpeded spread of the virus that can be described by the exponential growth of the infected population and no significant decrease in the number of susceptibles over time [16]. Thus, we assume a constant basic reproductive number R_0 for each of the European countries before the lockdown in the Lombardy region in March 8. However, in China the first wave of the epidedmic had almost concluded in March so we can not make the same assumption. We fixed the reproductive number in China and include three different periods, changing on January 23 and February 10 based on [17] [18] [19]. Before January 23, we use the estimation in [20] [21] to fix the basic reproductive number R_0 in China to 2.9 . From January 23 to February 10 and from February 11 to March 8, we calculate the average effective reproductive number based on the estimations in [22] [23] and obtain the R_e in China to be 1.1 and 0.43 respectively.

In Table 2.1 we show the values and prior distributions used in the Bayesian inference of the multitype birth death model and substitution parameters. Several
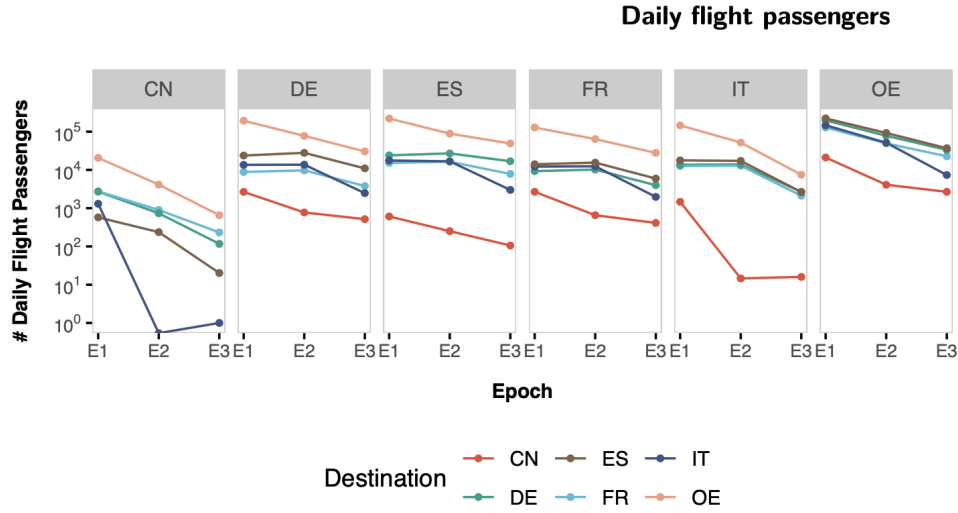
Figure 2.2:

todo

of these model specifications are identical to those from [3].

Data availability The SARS-CoV-2 genetic data can be downloaded from www.gisaid.org. Supplementary Table **??** lists accession numbers for the genetic sequences used in this study. Reported SARS-CoV-2 case counts are obtained from the country official agencies indicated in Table **??**. Flight data can be obtained from EUROSTATS https://ec.europa.eu/eurostat/.

Code availability and analyses reproducibility All the code is available at ... To ensure the reproducibility of the analysis, we have implemented the whole workflow with Snakemake [24]. The workflow is a modified version of Nextrain workflow [1] with the additional rules for BEAST 2 analysis and results processing.

Parameter	Value or Prior	Rationale
Nucleotide substitution model	HKY + Gamma	Unequal transition/transversion rates, unequal base frequencies, rate heterogeneity among sites with 4 categories
Clock rate	0.0008	Approximatelly 24 mutations per year \cite{10}
Location of origin	Hubei	Putative pandemic origin
Time of origin	Lognormal (-1, 0.2)	Median 26 October, 95% IQR 22 August to 8 December 2019
European countries reproductive number	Lognormal (0.8, 0.5)	Median 2.2, 95% IQR 0.8 to 5.9
China reproductive number	Origin - Jan 23: 2.9 Jan 23 - Feb 10: 1.1 Feb 11 - Mar 8: 0.43	Fixed based on \cite{}.
Becoming uninfectious rate	36.5 y-1	Period between infections and becoming uninfectious assumed exponentially distributed with a mean of 10 days
Sampling proportion		Upper bounds based on reported cases:
China	Uniform (0, 0.12)	
France	Uniform (0, 0.07)	
Germany	Uniform (0, 0.07)	
Italy	Uniform (0, 0.01)	
Spain	Uniform (0, 0.08)	
Other European	Uniform (0, 0.03)	
Sampling start date	December 23, 2019	Just before first sample from China
Sampling end date (only China)	January 23, 2020	Only included Chinese sequences before Lockdown in Hubei
Migration rates (no GLM analysis)	Lognormal(0,1)	Median, IQR
Migration rates (GLM analysis)	Uniform(0,50)	7.3 days as mean of exponential time to travel for every individual
Migration rates change times (GLM analysis)	December 23, 2019 and March 1, 2020	
Global scaler GLM parameter	December 23, 2019	Median, IQR
Coefficients GLM predictors	Normal (0, 2)	Median 0, IQR
Binary indicators GLM predictors	0 - 1 equiprobable	BitFlipOperator.

Table 2.1: Caption. Source of case data.

TODO. Vertical alignment

Discussion

Bibliography

[1] "Github - nextstrain/ncov: Nextstrain build for novel coronavirus sars-cov-2." [Online]. Available: https://github.com/nextstrain/ncov

[2] D. Kuhnert, T. Stadler, T. G. Vaughan, and A. J. Drummond, "Phylodynamics with migration: A computational framework to quantify population structure from genomic data," *Mol Biol Evol*, vol. 33, pp. 2102–2116, 2016, main reference for BDMM model

. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27189573https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4948704/pdf/msw064.pdf

[3] S. A. Nadeau, T. G. Vaughan, J. Scire, J. S. Huisman, and T. Stadler, "The origin and early spread of sars-cov-2 in europe." [Online]. Available: https://www.gisaid.org/;

[4] T. Stadler, R. Kouyos, V. VonWy, S. Yerly, J. Böni, P. Bürgisser, T. Klimkait, B. Joos, P. Rieder, D. Xie, H. F. Günthard, A. J. Drummond, and S. Bonhoeffer, "Estimating the basic reproductive number from viral sequence data," *Molecular Biology and Evolution*, vol. 29, pp. 347–357, 1 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21890480/

[5] T. Stadler, D. Kühnert, S. Bonhoeffer, and A. J. Drummond, "Birth-death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 228–233, 1 2013. [Online]. Available: https://www.pnas.org/content/110/1/228https://www.pnas.org/content/110/1/228.abstract

[6] W. P. Maddison, P. E. Midford, and S. P. Otto, "Estimating a binary character's effect on speciation and extinction," *Systematic Biology*, vol. 56, pp. 701–710, 10 2007. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/17849325/

[7] T. G. Vaughan, D. Kühnert, A. Popinga, D. Welch, and A. J. Drummond, "Efficient bayesian inference under the structured coalescent," *Bioinformatics*, vol. 30, pp. 2272–2279, 8 2014. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu201

[8] "Github - tgvaughan/bdmm-prime: Modified implementation of the bdmm multi-type birth-death model for beast 2." [Online]. Available: https://github.com/tgvaughan/BDMM-Prime

[9] R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Four-
     ment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. D. Maio,
     M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. D. Plessis,
     A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C. H.
     Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond, "Beast 2.5: An
     advanced software platform for bayesian evolutionary analysis," *PLoS Com-
     putational Biology*, vol. 15, 2019.

[10] T. G. Vaughan, G. E. Leventhal, D. A. Rasmussen, A. J. Drummond,
     D. Welch, and T. Stadler, "Estimating epidemic incidence and prevalence
     from genomic data," *Mol Biol Evol*, vol. 36, pp. 1804–1816, 2019,
     vaughan, Timothy G Leventhal, Gabriel E Rasmussen, David A
     Drummond, Alexei J Welch, David Stadler, Tanja eng Evaluation Study
     Research Support, Non-U.S. Gov't Mol Biol Evol. 2019 Aug 1;36(8):1804-
     1816. doi: 10.1093/molbev/msz106. Particle filtering paper. Nicely
     written. Compartmental models explained. Re-read to understad method.
     [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31058982https:
     //www.ncbi.nlm.nih.gov/pmc/articles/PMC6681632/pdf/msz106.pdf

[11] J. Scire, J. Barido-Sottani, D. Kühnert, T. G. Vaughan, and T. Stadler, "Im-
     proved multi-type birth-death phylodynamic inference in beast 2," *bioRxiv*,
     2020.

[12] D. T. Gillespie, "Approximate accelerated stochastic simulation of
     chemically reacting systems," 2000. [Online]. Available: http://ojps.aip.
     org/jcpo/jcpcr.jsp

[13] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions,"
     1977. [Online]. Available: https://pubs.acs.org/sharingguidelines

[14] P. Lemey, A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele,
     C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, and M. A.
     Suchard, "Unifying viral genetics and human transportation data to predict
     the global transmission dynamics of human influenza h3n2," *PLoS Pathog*,
     vol. 10, p. e1003932, 2014, lemey, Philippe Rambaut, Andrew Bedford,
     Trevor Faria, Nuno Bielejec, Filip Baele, Guy Russell, Colin A Smith,
     Derek J Pybus, Oliver G Brockmann, Dirk Suchard, Marc A eng R01
     AI107034/AI/NIAID NIH HHS/ R01 HG006139/HG/NHGRI NIH HHS/
     095831/Wellcome Trust/United Kingdom 092807/Wellcome Trust/United
     Kingdom HHSN266200700010C/PHS HHS/ DP1 OD000490-01/OD/NIH
     HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S.
     Gov't Research Support, U.S. Gov't, Non-P.H.S. PLoS Pathog. 2014 Feb
     20;10(2):e1003932. doi: 10.1371/journal.ppat.1003932. eCollection 2014 Feb.
     [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24586153https:
     //escholarship.org/content/qt1wz9c9fs/qt1wz9c9fs.pdf?t=qaeerr

[15] P. Lemey, S. L. Hong, V. Hill, G. Baele, C. Poletto, V. Colizza, Áine O'Toole, J. T. McCrone, K. G. Andersen, M. Worobey, M. I. Nelson, A. Rambaut, and M. A. Suchard, "Accommodating individual travel history and unsampled diversity in bayesian phylogeographic inference of sars-cov-2," *Nature Communications*, vol. 11, p. 5110, 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-18877-9https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7547076/pdf/41467_2020_Article_18877.pdf

[16] V. Boskova, S. Bonhoeffer, and T. Stadler, "Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models," *PLOS Computational Biology*, vol. 10, p. e1003913, 2014. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003913

[17] A. Pan, L. Liu, C. Wang, H. Guo, X. Hao, Q. Wang, J. Huang, N. He, H. Yu, X. Lin, S. Wei, and T. Wu, "Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, china," *JAMA*, vol. 323, p. 1915, 5 2020. [Online]. Available: https://jamanetwork.com/journals/jama/fullarticle/2764658

[18] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study," *The Lancet*, vol. 395, pp. 689–697, 2 2020. [Online]. Available: https://www.oag.com/

[19] Y. Xiao, B. Tang, J. Wu, R. A. Cheke, and S. Tang, "Linking key intervention timing to rapid decline of the covid-19 effective reproductive number to quantify lessons from mainland china," *International Journal of Infectious Diseases*, vol. 97, pp. 296–298, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1201971220304495

[20] S. W. Park, B. M. Bolker, D. Champredon, D. J. D. Earn, M. Li, J. S. Weitz, B. T. Grenfell, and J. Dushoff, "Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: framework and applications to the novel coronavirus (sars-cov-2) outbreak," 2020. [Online]. Available: https://doi.org/10.1101/2020.01.30.20019877

[21] M. A. Billah, M. M. Miah, and M. N. Khan, "Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence," *PLOS ONE*, vol. 15, p. e0242128, 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0242128https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7657547/pdf/pone.0242128.pdf

[22] "Covid-19 re." [Online]. Available: https://ibz-shiny.ethz.ch/covid-19-re-international/

[23] J. S. Huisman, J. Scire, D. C. Angst, R. A. Neher, S. Bonhoeffer, and T. Stadler, "Estimation and worldwide monitoring of the effective

reproductive number of sars-cov-2," *medRxiv*, p. 2020.11.26.20239368, 11 2020. [Online]. Available: https://doi.org/10.1101/2020.11.26.20239368

[24] J. Köster and S. Rahmann, "Snakemake-a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, pp. 2520–2522, 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22908215/

Data

Priors

Supplementary Figures