Phylodynamic Analyses based on 93 genomes

Insights on the clock rate, TMRCA, time of epidemic origin, R_0, R_e, and overall number of cases

Jeremie Scire[1,2], Timothy G. Vaughan[1,2], and Tanja Stadler[1,2,*]

[1]*Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland*
[2]*Swiss Institute of Bioinformatics (SIB), Switzerland*
[*]*Correspondence to be addressed to: tanja.stadler@bsse.ethz.ch*

February 25, 2020

In this report, we outline phylodynamic analyses based on full-length COVID-19 genomes downloaded from GISAID and NCBI. Acknowledgements for and details of the genome sequences are given in Table 6 of the post by A. Rambaut.

The "full alignment" of 93 sequences was obtained following the exact same procedure as described in that post. We considered a second alignment, coined the "cluster-free alignment", where only one representative from each known epidemiologically-linked transmission cluster was included (obtained as in post), resulting in a 75-sequence alignment. The third considered alignment is the "outside-China alignment" where we eliminated all sequences sampled within China from the cluster-free alignment, resulting in 38 sequences [1]. The data is characterized in Table 1.

We use these three alignments in BEAST2 [1] in order to quantify:

- **Clock rate** (i.e. an **evolutionary dynamics** aspect)
- **TMRCA and time of origin of the outbreak** (TMRCA = time of most recent common ancestor of the samples; this is a proxy for the time of origin of the outbreak, i.e. an important **past state** of the epidemic)
- R_0 **and** R_e (i.e. an **epidemiological dynamics** aspect)
- **overall number of cases** (i.e. an important **current state** of the epidemic)

[1]We note that the cluster-free alignment has three less outside-China sequences compared to the full alignment and the outside-China alignment as it was compiled prior to updating the other two alignments.

Alignment	# seq.	# seq. before Wuhan quarantine	time of first seq.	time of last seq.	total # of cases upon last seq.
cluster-free	75	48	27	Feb. 3, 2020	17238
outside-China	38	10	28	Feb. 8, 2020	34598

Table 1: Characterisation of the sequence alignments considered in the epidemiological analyses. The total number of cases at the time of the Wuhan quarantine onset on Jan. 23, 2020 was 571 within China (including the cases of Jan. 23; 10 cases were confirmed outside of China). For January 23, we have 6 sequences in the cluster-free alignment and 4 sequences in the outside-China alignment; for each alignment, we assign half of them to the sequences before the onset of the quarantine and the other half to the sequences after the onset of the quarantine. The case data was taken from the summary of the WHO situation reports.

Evolutionary analysis: Quantifying clock rate and TMRCA

The product of clock rate and TMRCA can generally be inferred from genetic sequencing data. However, the clock rate and the TMRCA are non-identifiable in the absence of measurable evolution; we also call this the absence of clock signal. Several posts explored the amount of clock signal in the COVID-19 sequencing data (A. Rambaut, S. Duchene, L. du Plessis, O. Pybus & L. du Plessis). We explored the sensitivity of the clock rate estimates on the choice of tree prior, similar to what was done for the early sequences in the 2013-2016 Ebola outbreak [3].

Method

We chose a number of coalescent and birth-death based tree priors and inferred the phylogenetic trees together with the evolutionary and epidemiological parameters. The choice of models and prior distributions for the model parameters is shown in Table 2.

In summary, we have three different coalescent frameworks and three different birth-death frameworks. For the birth-death framework, we additionally have three different options for the priors on the removal rate δ. In total, we thus perform twelve analyses both on the full alignment and on the cluster-free alignment. The rational for considering the cluster-free alignment is that epidemiologically-linked transmission clusters have been mentioned to potentially bias clock rate estimates, see discussion below this post.

For the structured models, we assigned each tip one of three demes at random. For birth-death models, we further analytically integrate over the tip deme assignment as explained in [6]. Such deme assignment is envisioned to take into account population structure implicitly.

For each analysis, we run five parallel chains for 1×10^8 to 2.5×10^8 steps, depending on the tree prior configuration. We check the convergence of each individual chain (ESS > 200) and combine them to obtain the final posterior distribution. Structured coalescent runs mixed more slowly, in this case, we set the threshold for convergence

Model	Parameter	Prior distribution
Strict clock	clock rate	lognormal(-7, 1.75)
HKY	kappa Gamma shape	lognormal(1.0,1.25) Exponential(0.5)
Coalescent	N_e	lognormal(1,2); lognormal(1,1.5) (with structure)
exponential growing population	growth rate r	laplace(0,20)
with structure[4]	migration rate	lognormal(1,1.5)
Birth-death [7]	R_e removal rate δ sampling proportion s time of origin t_{or}	lognormal(0.8,0.5) lognormal(3, 0.8); fixed to 365/10 per year; fixed to 365/20 per year beta(alpha=1, beta=50) lognormal(-1.5, 0.4)
skyline[7]	change time for R_e, s	fixed to Jan. 23 (start of Wuhan quarantine)
with structure[2, 5]	migration rate	lognormal(1,1.5)

Table 2: Models and prior distributions used for the BEAST2 analyses. If no reference is provided for a model, this model is part of the core of BEAST2 [1]. We highlight that $1/\delta$ is the time of an individual being infected, i.e. the sum of the exposed period and the infectious period. We both estimate δ, or assume $1/\delta$ to be fixed to 10 or 20 days. Sampling is assumed to coincide with removal from the infectious pool. Under the birth-death skyline model, the parameters R_e and s are allowed to change.

to ESS > 100 for individual chains.

Results

The resulting estimates for clock rate, TMRCA, tree length (i.e. sum of branch lengths), and tree divergence (= clock rate x tree length) are shown in Fig. 1. As expected, the tree divergence is estimated the same across tree prior choices. The clock rate estimate and the tree height (resp. tree length) estimate are inversely correlated. The clock rate estimates for the full alignment are higher compared to the cluster-free alignment. Since the priors do not take into account the epidemiologically-linked transmission cluster, we focus now on the cluster-free alignment results.

The clock rate estimates for the unstructured coalescent are around 10^{-3} (median) while all birth-death estimates and the structured coalescent estimates are lower, with the smallest median estimates being a bit under 5×10^{-4}. The unstructured coalescent estimates agree with previous estimates (all based on the unstructured coalescent; Table 3). Tree height is estimated between 2.5-3 months (medians) which translates into the median TMRCA to be between mid-November and early December 2019.

Overall, the clock rate and tree height estimates are sensitive to the prior, with the clock rate medians varying by about a factor of 2 and the tree height medians by about 0.5 months. Additional sequences should make these estimates more robust towards the tree prior choice.

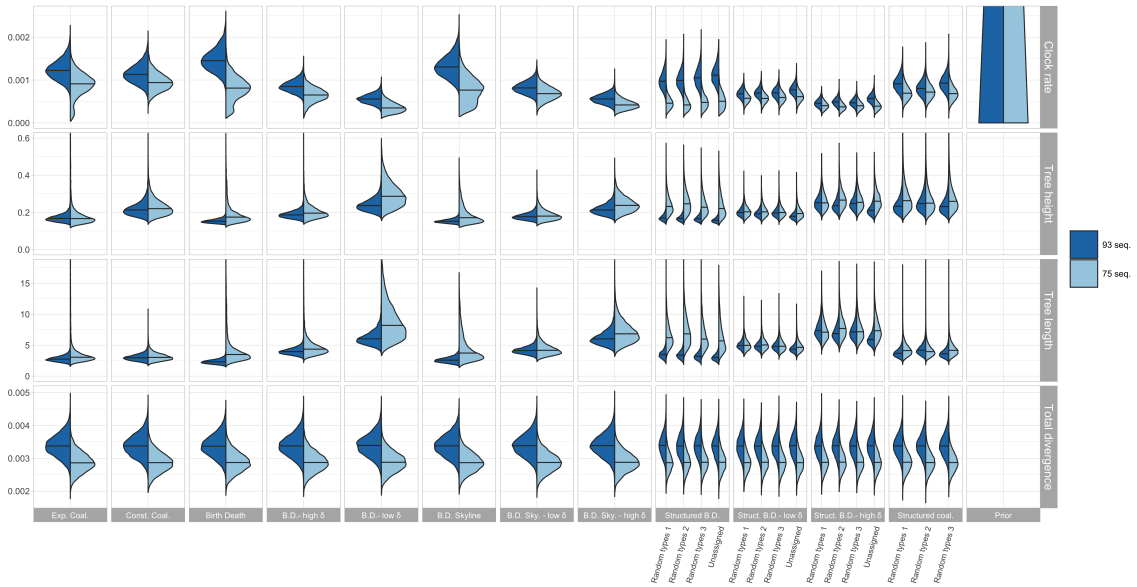

Figure 1: Posterior distribution of clock rate, TMRCA, tree length (i.e. sum of branch lengths), and tree divergence (= clock rate $\times$ tree length) for the 12 different tree prior choices.

median	hpd low	hpd high	
0.80×10^{-3}	0.14×10^{-3}	1.31×10^{-3}	A. Rambaut (24/02)
0.92×10^{-3}	0.33×10^{-3}	1.46×10^{-3}	A. Rambaut (12/02)
1.23×10^{-3}	5.63×10^{-4}	1.98×10^{-3}	S. Duchene (03/02)
1.05×10^{-3}	3.29×10^{-4}	2.03×10^{-3}	K. Andersen (27/01)

Table 3: Previous clock rate estimates for COVID-19 based on the unstructured coalesent tree prior.

Epidemiological analysis: R_0, R_e, the time of origin of the epidemic, and the overall number of cases

Next we set out to quantify the epidemiological dynamics. We should be close to the phylodynamic threshold such that these analyses reveal informative insights into epidemic spread, see a post by E. Volz. T. Bedford reported prevalence and incidence estimates based on the genomic data. Here, we estimate R_0, R_e, the time of origin of the epidemic, the total number of cases, and the sampling proportion using birth-death models.

Method

We apply the birth-death skyline framework to the cluster-free alignment and the outside-China alignment, with both R_e and s being allowed to change through time. The full alignment is not considered here as the sampling assumption of the birth-death model is strongly violated in presence of the epidemiologically-linked transmission clusters.

In addition to the cluster-free alignment, we analysed the outside-China alignment as sampling intensity within China and outside of China is very different. Further, we are not confident to assume uniformly-at-random sampling (specified by the parameter s) within China. We argue that the outside-China sequences can be viewed as a very sparse and random sample from the pool of Chinese sequences. In particular, a case is sampled and sequenced if an infected individual travels abroad and they (or a closely linked epidemiological case) is subsequently diagnosed and sequenced. The sampling proportion s is thus a combination of travel abroad, diagnosis, and sequencing. As travel abroad changed drastically upon the start of the Wuhan quarantine, we allow s to change at that time point.

Here we assume a prior for the sampling proportion s different from what is shown in Table 2: Before the start of the Wuhan quarantine on January 23, 2020, we assume a uniform prior between 0 and 0.1 for the cluster-free alignment and a uniform prior between 0 and 0.015 for the outside-China alignment. For both alignments, we used a uniform prior between 0 and 2×10^{-3} after the onset of quarantine of Wuhan. The upper bounds for these uniform priors were determined based on the number of sequences and the number of confirmed cases (Table 1).

We argue that the duration of an infection (parameter δ) did not drastically change throughout this epidemic outbreak and in particular fixed it to two plausible values, 365/10 and 365/20 per year, as above. As the clock rate estimate is uncertain (see above), we fixed the clock rate here to two different plausible values, namely 10^{-3} and 5×10^{-4}.

The other priors were set as in Table 2. Chain length and convergence assessment was done as for the evolutionary analyses.

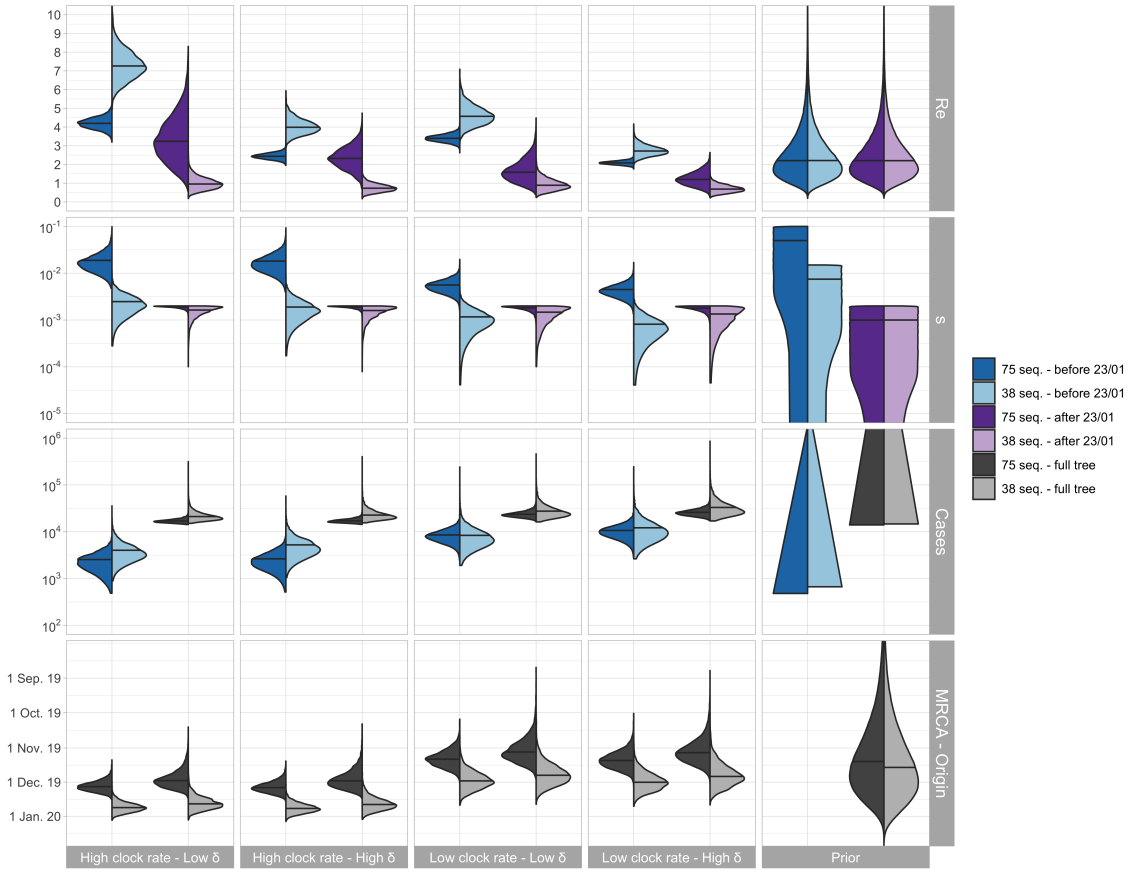

Figure 2: Posterior distribution of R_e before Wuhan quarantine (which can be interpreted as the R_0), R_e after Wuhan quarantine, sampling proportion s, total number of cases (=#sequences / s), origin t_{or}, and TMRCA for the cluster-free and the outside-China alignment. In the bottom row, in each panel, we show on the left the estimated TMRCA and on the right the estimated time of origin of the epidemic.

Results

The parameter estimates are shown in Fig. 2. First, as expected, we observe a decline of R_e upon the start of the Wuhan quarantine. For the outside-China sequences, the median drops below one. The median R_0 in China is estimated to be between 2.5-7 based on the outside-China sequenced; R_0 is estimated to be lower based on the cluster-free sequences.

We estimate the total number of cases (= # sequences / sampling proportion) to be around the number of confirmed cases; we do not trust this estimate though, see Caveats below.

The TMRCA is a bit younger than the time of origin of the outbreak, the estimated medians differ by up to a week.

Caveats

Caveats regarding the evolutionary analysis have been discussed in previous posts (A. Rambaut, S. Duchene, L. du Plessis); the same apply here.

We see two main caveats in the epidemiological analyses. First, the assumption of a constant sampling rate through time after the start of the Wuhan quarantine is violated. In particular, during the time spanned by the three most recent sequences, more than 30'000 confirmed cases accumulated. Second, we assume a random mixing within China, while this assumption is in particularly violated upon the quarantine onset (see also "Notes about demographic models" in this post). With additional future sequencing data, we can extend our models to take these characteristics into account. Until then, we are not very confident in the accuracy of our epidemiological parameter estimates.

References

[1] Remco Bouckaert, Timothy G Vaughan, Joelle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

[2] Denise Kühnert, Tanja Stadler, Timothy G Vaughan, and Alexei J Drummond. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular biology and evolution*, 33(8):2102–2116, 2016.

[3] Simon Möller, Louis du Plessis, and Tanja Stadler. Impact of the tree prior on estimating clock rates during epidemic outbreaks. *Proceedings of the National Academy of Sciences*, 115(16):4200–4205, 2018.

[4] Nicola F Müller, David Rasmussen, and Tanja Stadler. Mascot: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, 34(22):3843–3848, 2018.

[5] Jérémie Scire, Joëlle Barido-Sottani, Denise Kühnert, Timothy G Vaughan, and Tanja Stadler. Improved multi-type birth-death phylodynamic inference in beast 2. *bioRxiv*, 2020.

[6] Tanja Stadler and Sebastian Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120198, 2013.

[7] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei 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*, 110(1):228–233, 2013.