Response to Reviewers

Reviewed by Luca Ferretti, 2020-07-22 11:49

This manuscript presents an exhaustive phylodynamic analysis of the early phase of the French COVID-19 epidemic.

The focus is on the dominant clade (B.1) since it is the only informative one for phylodynamic analyses. The lack of sequences from the initial days of the epidemic hinders a more detailed reconstruction of the early dynamics due to lack of resolution.

Nevertheless, the results are quite strong and informative. The beginning of the French epidemic is dated within a reasonable interval (mid-January to early February) consistent with epidemiological evidence for the European epidemic.

The doubling time is also consistent with the epidemiological evidence, although with large uncertainties.

Thanks!

The authors also find an increase in the doubling time along the epidemic, that would be extremely interesting in terms of a slowdown due to the non-pharmaceutical interventions implemented in France. However, I wonder if either the non-uniform sampling rate or the geographical spread of the strains could have affected this result. Intuitively, both could cause an upward bias to the inferred growth rate in the early part of the tree. In my opinion, while very suggestive, the evidence presented here is not conclusive

Ideally, we would increase the density of the sampling. Unfortunately, this was a constraint of the study.

We think it would take a completely different study to redo this analysis with more data, because one of the originality of the current work was to be performed early Apr with the available sequences.

Besides, as can be checked on GISAID, the number of sequences for France has remained extremely low. This means that such a more complete study cannot be performed yet. This is also discussed in further details in our response to Reviewer #3 below.

If we cannot significantly increase the sampling, we can decrease it. In order to try and account for this sampling issue, we have designed subsets of the phylogeny (shown in Supplementary figures). The estimates obtained for these phylogenies are in agreement with that in the main text.

Finally, the BDSKY model model perhaps best illustrate the sampling issue you mention. Indeed, as we acknowledge, it is difficult to make inferences in the early part of the phylogeny due to the sparse sampling (reproduction number R1). We then see a decreasing trend but, as for the doubling time, our resolution is limited.

=> We have rephrased the abstract and added a sentence in the Discussion to make it clear that the evidence is suggestive.

The birth-death skyline analysis used to infer the duration of the infectious period is very intriguing. The authors find an infectious period of 4-6 days from phylodynamic evidence. Note that in the model by Stadler et al 2013, infectiousness is constant in time until individuals are not infectious anymore, while COVID-19 has a bell-like profile of infectiousness centered around 4-6 days post infection. Hence, it is difficult to assess the agreement between the result of the authors and the known generation time distribution of COVID-19.

Here we think there is a slight misunderstanding. What Stadler *et al* show in their Figures 2C or 3C is the variation of the recovery rate over time (using a skyline technique with 10 temporal estimates). Here, we assume this value is constant over time and show its posterior distribution.

=> We clarified the figure caption to address this issue.

Assuming that the relevant comparison would be with the duration of the infectious period, the estimates in this manuscript would be reasonably close to the epidemiological evidence.

Yes, this is our point.

Instead, if the relevant comparison would be the one with the relation between exponential growth rate and R0 (Lotka-Volterra equation), the model used by the authors would lead to a significant underestimation of R0 in the initial phase or an overestimation of the the infectious interval, as well as an overestimation of Rt when the epidemic is decreasing. This could be one of the reason behind the suspiciously low (although very uncertain) value of the inferred Rt in the first period of the epidemic, and the fact that Rt>1 in later phases. Anyway, the values of the infectious period and Rt are in the right ballpark.

Again, we think there is a misunderstanding in the fact that what we represent is a distribution.

Overall, this is a very good and clear manuscript that provides an excellent example of the power of phylodynamics to infer quantities of epidemiological interest.

Thanks!

Reviewed by anonymous reviewer, 2020-07-05 02:47

Danesh et al. present a study of the phylodynamics of SARS-CoV-2 sequences from France early in the outbreak. Their analysis is based on 196 genomic sequences collected early in the outbreak (January 24 - March 24 2020) and they estimate several key epidemiological parameters from the sequence data. While this work is important and timely, some clarifications would strengthen the manuscript.

Thanks, we tried to clarify the main points.

-In Figure 3, the estimated doubling time from the sequences from the second half of the epidemic (France 61-2 set) is lower than the doubling time for the sequences from the whole epidemic (France 122a) or the doubling time for the sequences from the first three quarters of the epidemic (France 81). This does not appear to be consistent with the interpretation that adding more recent sequences increases estimated doubling time. Was estimated doubling time lower at the end of the time period examined as well as at the beginning?

Thank you for pointing this issue that we indeed forgot to discuss. It does make sense to hypothesise that the doubling time would be similar or even higher to that estimated using all sequences (France122). Our interpretation here is that we reach the limits of phylogenetic inference in terms of genetic variation. In particular, the estimate for the date of the most recent common ancestor shifts from Jan 31 to Feb 8 even though we should be sampling the same epidemic.

=> We now explain that the sampling can vary depending on the time period and that the phylogenetic signal may be limited.

-The methods section states that 196 sequences are analyzed. However, 204 sequence ids are listed in Supplemental Table 1. The set of sequences used for the analysis should be clarified.

Apologies: we uploaded a revised table, where the sequences included in the main analysis are highlighted. As indicated in the text, we removed sequences that were too short or of too low quality.

-In a few places, more detail on the methods used would be helpful. In particular, it would be helpful to provide more detail on the steps taken to align and clean the data using the augur pipeline (what parameters were used to filter and align the sequences?) Parameters used to run RDP, SMS, PhyML should be listed (where default parameters were used this should be specified). The authors should also consider including the phylogenetic tree in Figure 1 as a supplemental file.

Thank you for th suggestion. We now specify the following default parameters and added the phylogeny in a Newick format.

RDP parameters were default and we also now cite another study that did not detect recombination events in SARS-Cov-2 genomes

PhyML parameters were imported from the results of the SMS analysis.

The sole SMS parameter was the use of the AIC selection criterion. The results from SMS are indicated in the main text.

-The motivation for the molecular clock settings chosen (above, below and equal to a previously reported value) are described only in the methods; it would be helpful to have this information in the results section when Figure 2 is discussed.

=> We decided to move the Methods section before the Results, which is less friendly for a larger audience but

definitely clarifies the reading.

It would also be helpful in the caption of Figure 2 to specify that the clock rate is in substitutions per site per year (this is also just described in the methods).

=> We added the unit as well as the code to analyse the substitution rate values in the colour legend.

Also, in Figure 2, the models should be listed in the legend in either increasing or decreasing order of clock rate (right now the slowest of the three clocks is in the middle of the legend, which is confusing).

=> We edited the caption for further clarity to indicate that the number in the legend corresponds to a substitution rate.

--In the introduction (line 14), the Liu et al. reference is not the right one for the genomic sequence of SARS-CoV-2. For the initial sequencing of the virus, cite Wu et al., 2020, A New Coronavirus Associated With Human Respiratory Disease in China and Zhou et al., 2020, A pneumonia outbreak associated with a new coronavirus of probable bat origin.

=> We now cite Wu et al and Zhou et al.

Typographical suggestions:
-title: epidemics → epidemic
-line 2: pandemics → pandemic
-line 19-20: "Early results allowed to better understand the origin of SARS-Cov-2 and identify" → "Early results allowed better understanding of the origin of SARS-CoV-2 and identification of"
-line 34: "among which the temporal reproduction number" → "including the temporal reproduction number"
-line 42: "epidemics" → epidemic"
-Fig 1 legend: "because outside the main clade" → "because they are outside the main clade"
-Figure 2 legend: "fix molecular clock" → "fixed molecular clock"
-line 79: In the following of the work → in the following work
-line 87: In appendix → in the appendix
-line 95: in smaller dataset → in the smaller dataset
-line 159: "if we use a," → delete comma
-line 234: bayesian → Bayesian (fix capitalization)
-line 250: as previous models → as in previous models
-line 264: delete comma

Thank you for the detailed suggestions! We apologize for the numerous typoes and tried to carefully correct them in the revised version.

Reviewed by anonymous reviewer, 2020-07-16 09:21

Danesh et al. perform a phylodynamic analysis of the French SARS-CoV-2 epidemic, and notably estimate the reproduction number and the duration of infection from the phylogeny. They analyze the sensitivity of their results to the sequence sampling, and observe an effect of the lockdown on the reproduction number. The values they infer for the parameters of the epidemic agree with those of contact-tracing analyses.

I found Danesh et al.'s manuscript interesting, in particular that the phylodynamic estimates overlap with contact tracing estimates, but have a few comments and suggestions to make.

Thanks!

First, I found an interpretation of the phylogeny puzzling: it seems a polytomy is interpreted by the authors as evidence for multiple introductions, but I don't understand why it would be so.

Actually we had two hypotheses: either a rapid "super-spreading event" or the introduction of sequences that had already diverged. But we agree that there is little evidence for either and that it makes more sense to point out the lack of phylogenetic signal.

=> We now only mention the multiple introductions with respect to the non-B1 clade int he phylogeny and point out the lack of phylogenetic signal associated with the polytomy.

Second, the sampling is uneven across French regions, with some regions entirely missing. I think the authors should address this problem, for instance by discussing its origin and its potential consequences on the estimated phylogeny and parameters.

As indicated in our response to Reviewer #1, we would of course have preferred to have a more thorough sampling.

In terms of the regions that are missing, there is a clear correlation with the magnitude of the epidemic wave. For instance, Occitanie was much less impacted than the Paris area or the East of France.

=> We now describe in more details in the Results and in the Discussion that some regions are under-represented and why this is consistent with the French epidemic.

Third, I think the manuscript can be made clearer by reorganizing some parts, and explaining some terms (see below for specific examples). Fourth, I found myself missing some technical information, for instance on the clock models that were used, on convergence diagnostics, and I would have liked to see a more systematic comparison between the prior and the posterior distributions (see below for specific examples).

My opinion is that those points should be addressed before any recommendation.

We tried to address these more specific comments.

More specific comments:

p3 l55: I think it would be useful if the authors could address the lack of sequences coming from region provence Alpes Côte d'Azur. It is also a missing point in Gambaro et al., but those authors circumvented this issue by arguing that they were focusing on the epidemic in the north of France. That's not what the authors here are aiming to do, but can they still talk about the epidemic in all of France if entire regions are missing?

We would have liked to access such sequences as well... We do have sequences from Valence (in Auvergne Rhône-Alpes region), which is located between Lyon and Marseilles.

As indicated above, redoing this analysis would be a completely different study since the originality of the current one is to have been performed in a time of crisis. More, as indicated above, the number of sequences available for France on GISAID currently does not allow a more thorough analysis. Indeed, when we performed the analysis, we had approximately 200 sequences spanning from Feb 27 to Mar 24 (i.e. less than a month). More than 3 months later, the total number of sequences from France on GISAID is only 409. Moreover, more than 80 of the new sequences cannot be used since the sampling date is not provided (neither is the region but this is less important). We also need to remove 2 sequences from cats, 4 with incomplete coverage and 11 from cell cultures…

=> We now further specify that the main added value of this analysis is dynamical side rather than the phylogeography, which will require a denser sampling.

p4 l70: "Another interpretation, could be independent introductions in France (up to 6 events)." : I don't understand. Independent introductions would not necessarily create a polytomy, as is shown by the sequences in black. And I'm not sure where the number 6 comes from.

As indicated to reviewer #2, we modified this interpretation to rather point out the lack of phylogenetic signal.

Fig. 1 : there seems to be a contrast between the relative number of sequences coming from Île de France in the data set and the relative number of infections or hospitalizations in Île de France. Based on the latter, I would expect many more red sequences in the phylogeny. Has there been an under-sequencing of sequences in Île de France compared to other regions ? Finally, the scattered distribution of the sequences from Île de France indicates that they may be the source of many clusters in other French regions, if the support values in the phylogeny are high enough.

In general, the sampling for sequences is extremely low since there are now 400 available and the total number of cases is estimated to be above 4 million.

Again, our goal here is not to perform a phylogeographic analysis of the epidemics but rather to analyse its speed of spread.

=> We now specify the sampling rate, which is likely to be below 1 in 10000.

Fig. 2: the legend of the distributions could be improved by indicating the unit notably.

We made the change also suggested by Reviewer #1.

Fig. S4: the legend needs to be improved as in Fig. S5.

We edited the captions.

Fig S3: it is not clear how the priors were specified, in particular their parameters.

The time to the most recent common ancestor is a result of the phylogenetic reconstruction. There is no prior as such for this estimate. We fit the molecular clock rate so in the end the priors are really that associated with the molecular evolution model.

p5 l90: "This was also true for the BDSKY model, where the prior shape for the recovery rate had little impact (Figure S3)": I would clarify and add something like "... but the positional parameters of the prior had an important impact."

We did not vary the position parameter of the prior for a given prior shape. Indeed, in Figure S3 what varies, besides the prior shape (uniform or lognormal) is the molecular clock. Either we assume a fixed clock value, or we estimate the value of strict molecular clock.

=> We edited the text and clarified the caption for Figure S3.

p6 l115: the data sets (France81, France61-1) are introduced here even though they have been discussed previously when commenting on Table 1.

We apologize for the lack of clarity: we now present the Methods before the results, which addresses this issue.

p6: "Since the first dataset includes more recent" : do the authors count the full dataset France 122a when writing this sentence, or do they just talk about the 3 other ones (the subsets)?

We do include France122a and reformulated the paragraph to clarify the meaning (France61-2 is introduced separately to avoid the confusion).

l125: "Adding more recent sequence data indeed leads to an increase in epidemic doubling time. Initially, with the first 61 sequences (which run from Feb 21 to Mar 12)": I am wondering if the doubling time is the only parameter that changes in this experiment. Indeed, altering the sequence sample may change other parameter estimates, which may be correlated with the doubling time. Besides, the sampling effort was probably not the same between before March 12 and after March 24. Are such differences in sampling effort accounted for in the model?

This is indeed a valid point, which actually echoes a concern from reviewer #2 (see above). The sampling is indeed different, which can be seen with the date for the MRCA.

=> We reformulated the text to highlight the fact that sampling may also change.

l130: I found this paragraph about molecular clocks confusing because it was unclear to me what clock models had been used.

Introducing the Methods before the Results helps address this confusion. We also specified that we study the effect of the substitution rate.

Fig. S6: "convergence is limited" : what does that mean? That convergence diagnostics were charateristic of a lack of convergence of the MCMC chains?

We apologize for the unclear formulation. The conclusion was drawn from the shape of the posterior distribution and its wide 95% credibility interval. We reformulated and also now show the prior distribution on the plot.

l150-155: I think it would be useful to define what the authors mean by duration of contagiousness vs distribution of infectious periods. I assume the latter is the time between successive infections in a transmission chain, but I'd like to be sure... Also, it is not clear to me why for these estimates the authors no longer consider the influence of the priors on the rate of evolution or of sampling time.

The changing terminology was indeed misleading. We have clarified what is estimated in the BDSKY model (the recovery rate and the sampling rate), and why the infection duration is estimated by taking the inverse of the sum of these rates.

The link between the infection duration estimated in the BDSKY model by Stadler et al. and the time between two infections, also known as generation time, is less straight forward. In theory, the two should be identical because the BDSKY model reflects the unfolding of the epidemics. For instance, if non-pharmaceutical interventions prevent any transmission after 4 days of infection, this will translate into a shorter "infection duration" in the BDSKY model. This is also now further explained in the text, where we refer to the "effective infection duration" to insist on the fact that this measure is different from the biological infection duration.

Concerning the influence of the priors, in Figure S8 we do vary the molecular clock value, while varying the shape of the recovery rate prior.

Fig. S3, S6, S8, S9: it would help in all figures showing the impact of the prior on the posterior distributions to also display the prior distributions.

For S3, as indicated above, we cannot really provide a prior distribution but for S6, S8, S9, and S10 we added the prior distributions on the figure.

l180: "allows to infer phylogenies" : allows one to

Done!

l215: "As acknowledge in the introduction" : acknowledged

Done!