Dear recommender and reviewers,

First of all we apologize for our slow response, which, as you can guess, originates from our work on the COVID-19 epidemic.

We have performed several additional analyses and are now very confident about the robustness of the results we present.

More specifically, here is a list of the major changes we made or the additional analyses we performed:
- we tested for potential sampling effects by subsampling our dataset,
- we tried to improve the date of the tMRCA by adding external sequences to the phylogeny,
- we performed an analysis using the BDMM model in BEAST2,
- we now use a symmetric prior for the difference between the two host types,
- we derive doubling time differences more carefully,
- we investigate 100 trees from the BEAST posterior distribution,
- we improved the parametric bootstrap to better visualise the effect of using the posterior distribution or a uniform distribution based on the posterior distribution quantiles,
- we updated the figures for improved clarity.

Below you will find more detailed responses to your queries.

We thank you again for your time and interest!

Response to David Rasmussen

Dear authors,

Both of the original reviewers have now reviewed your revised manuscript. Unfortunately, both still raise substantial concerns that their original concerns were not adequately addressed or deserve more attention.

I strongly agree with the opinion of the reviewers, especially since their criticisms concern the main findings of the paper regarding transmission dynamics in the different risk groups. I would especially urge the authors to carefully address the concerns of the reviewers with respect to:

1) Sampling differences between the "new" and "classic" risk groups and how this impacts the estimated epidemic growth rates in each group.

To evaluate how our ABC framework is sensitive to sampling biases we ran ABC analyses on phylogenies containing only half of the ‘new’ hosts sequences. Despite this

under-sampling of the new hosts, results remained qualitatively similar to those estimated with the full phylogeny.

2) The comments from Reviewer #2 about the priors on the ratio of the transmission rates between risk groups. I agree with the reviewer here that that the priors should give symmetric or equal prior probability to either risk group having a higher transmission rate.

As detailed below, we now use a prior of the assortativity parameter (\nu) where the probability to have \nu < 1 is equal to the probability to have \nu > 1.

Responses to Chris Wymant

Apologies to the authors for my delayed response.

This previous substantive concern has not been addressed: "The main conclusion of the paper is that the epidemic in new hosts is growing faster than that in classical hosts, however the confidence with which this conclusion can be made is not stated. The doubling time in classical hosts since 1997 is estimated to be 0.58 - 10.13 years, and the doubling time in new hosts 0 - 3.51 years. The relevant quantity for the conclusion is the posterior for the ratio of these two parameters. The authors do present a ratio comparing the two hosts with regards to the reproduction number, and find that the confidence intervals do not exclude 1; if the same is true of the doubling time, which seems plausible given the similar parameter dependencies of R_0 and doubling time, the main conclusion is not supported. The same point applies to the other host parameters inferred to be different: assortativity and recovery/removal rate. The parameters themselves need not be redefined, but the posteriors of their ratios should be examined to support claims of differences" (and the precision with which these differences have been measured). This applies both to the difference between the two host types, and also to the difference in the effective reproduction number before and after the third generation tests, where a difference is currently presented despite strongly overlapping confidence intervals in the two separate values. If the ratio of doubling times does indeed have a credibility interval excluding 1, I would suggest reporting it as 10 [x-y] times smaller for new hosts instead of an order of magnitude lower, to also communicate the precision of the estimate.

Thank you for the suggestion.

We computed the ratio of the doubling times of classical hosts after 1997 over the doubling times for new hosts to estimate the current difference and, although the median is 10, the 95% credibility interval does not exclude 1 and is [0.62;149.99], however the 75% credibility interval does exclude 1 and is [3.39;25.61].

We find that the doubling time of ‘new’ hosts is higher than the doubling time of ‘classical’ hosts before 1997 when the *R_e* of the ‘classical’ hosts is higher than the *R_e* of the ‘new’

hosts. We computed the ratio of the R_e of ‘new’ hosts over the *R_e* of ‘classical’ hosts after 1997 and, the median is 1.14 and the 95% credibility interval is [0.56;3.25].

We computed the ratio of the *R_0* of the classical hosts before and after 1997 ($R0_{1,t1}/Re_{1,t2}$). The median of the ratio is 1.26 and its 95% credibility interval is [0.88;2.06]. However, its 75% credibility interval is [1.01;1.68].

Finally, we computed the ratio of \gamma_2/\gamma_1 and the 95% credibility interval does exclude 1.

We added these details in the manuscript.

I am uncertain, but think my concern about differences in sampling has not been addressed. My concern was that higher sampling in the new group than in the classical group might lead to the data observed, and to the resulting conclusion, even if there were no difference in transmission dynamics in the two groups. I think the authors' reply can be summarised as saying that if the new group was under-sampled, spread there would be even faster than currently estimated. But the same is also true in the classical group, and if it were true there to a greater extent - i.e. lower sampling there - then the true difference in transmission dynamics between the groups would be less than that estimated here. In the absence of any information of the true underlying population sizes, the most defensible assumption on sampling might be that equal proportions of both populations end up diagnosed, and not that equal proportions of both populations end up included in the present study, given that the classical group seems to have been deliberately downsampled and by a known factor.

To further investigate potential sampling issues, we pruned the target phylogeny to get different phylogenies with half of the ‘new’ hosts tips. We then ran our ABC framework on each tree and found that the results estimations were similar to those from the full phylogeny (although with wider confidence intervals).

These results suggest that our ABC inference is not strongly affected by sampling biases.

Unless I have misunderstood, the authors' choice of the prior Unif(1, 10) for the parameter nu means it was impossible to come to any conclusion other than a greater transmission rate for new hosts. The prior for nu should have equal weight above and below 1 if we are to learn how the data inform the ratio of transmission rates; here it has zero weight below 1. I think that treating the two hosts equally a priori (i.e. imposing a host-type interchangeability symmetry) implies that the prior for log(nu) should be symmetrical about 0. Thank you for the suggestion. We now run the same ABC analyses with a prior of \nu in Unif(0,10) to include both hypotheses. Re-articulating my previous concern, this was that 100% of the prior probability mass was for nu greater than 1, when it should be 50% if the prior is to be uninformative of differences between the host types. In modifying the prior from Unif(1,10) to Unif(0,10), it is now the case that 90% of the prior probability mass is greater than 1; this is still far from 50%. More specifically, for parameter priors to be agnostic about the hosts, they should be unchanged by relabelling the two host types new <-> classical. For

parameters describing a multiplicative difference, the prior p should be such that it is equally likely that the one group's parameter is x times larger as it that the other group's parameter is x larger. In other words p should satisfy int*{1/a}^{1/b} p(x) dx = int*{b}^{a} p(x) dx for any a and b. Equivalently, probability mass should be symmetric about zero on a logarithmic scale for x, i.e. F(log(x)) dlog(x) = F(log(x))/x dx for any symmetric function F. The simplest example is F=1, i.e. a prior of 1/x, truncated above a point L and below a point 1/L to give a proper prior. Given the previously plotted shape of the posterior for nu, I am confident that the authors' qualitative result will be unchanged with this more appropriate prior. However the paper's conclusion would be considerably more persuasive if it arises from an uninformative prior than from a prior geared 90% towards that conclusion. Also judging by the posteriors in Figure 2, the aforementioned truncation point L for the parameter nu looks like it ought to be greater than 10; similarly, priors allowing for higher values of gamma*2 and R0*1^{t2} seem justified given the data (posteriors squashing to the prior boundary).

Thank you for the suggestion!

We now use a symmetric prior distribution between 0 to 10 for the \nu parameter. The values of the first and the third quartiles are 0.5 and 5.5, respectively, and the median is 1.0.

Minor suggestions:

For a number of points where I requested clarification, this was provided in the response without indicating whether it was also added to the manuscript. For the authors to check and consider, to taste.

With the shift in nomenclature from basic reproductive number to effective reproductive number, it would be appropriate to drop the subscript 0 in R_0.

Yes thank you. We made the change in the manuscript.

146 onwards: I don't understand the hypothesis being tested here. I think it is that previously the epidemic consisted only of people who were diagnosed a long time after infection, but more recently there is a separate epidemic of people who were diagnosed shortly after infection. This only makes sense to me if diagnosis time is merely a proxy for another characteristic that is causal for difference in dynamics. Here, we wanted to validate the fact that the differences in the structure and the labels in the phylogeny are not due to the infection phase (acute vs. chronic), but rather to the epidemiological profile (classical hosts vs new hosts). This is because new/MSM hosts are all detected during acute infections. I think I now understand this: in the alternative model, nu is the difference in infectiousness for the same person in acute and chronic phases, not for two different people who tend to be diagnosed in different stages,

correct? Including the ODE for this model in the supplementary material would clarify this.

We apologize for the lack of clarity. In our alternative model, we kept two types of host and there are two infection phases (acute and chronic). The parameter \nu is still the difference in infectiousness between the two types of hosts.

We included the diagram and the ODE in the manuscript appendix.

80: "posterior to 1997" -> "after 1997"

This is now modified.

108: "To better apprehend" -> "To better understand" (or similar)

This is now modified.

113-116: consistent time units throughout would aid comparison of these numbers.

This is now clarified in the manuscript.

188-191: I suggest moving this sentence to the introduction. Specifying precisely how the two host groups were defined is helpful for following the rest of the paper (c.f. our earlier confusion).

We moved the sentence “The epidemiology of HCV infection in the cohort has been extensively described from 2000 to 2016 [40–42]. The incidence of acute HCV infection has been estimated among HIV-infected MSM between 2012 and 2016, among HIV-negative MSM enrolled in PrEP between in 2016-2017 [13] and among HIV-infected and HIV-negative MSMs from 2014 to 2017 [14].” to the Introduction.

We also now better explain that the two host categories were assigned by clinical epidemiologists.

261 (just before): I think the numerator should be (clarified to) "number of descendant leaves labelled Y"

This is now clarified in the manuscript.

Figure 2: I suggest spreading these plots over two rows, as the horizontal squashing makes them hard to read. Also using the same x axis scale for the two reproductive numbers would aid comparison.

Thank you for the suggestion, this is now modified.

Best wishes, Chris Wymant

Responses to Louis DuPlessis

The authors have improved the manuscript and addressed some of the comments from the previous round of reviews, but there are still a few issues with the manuscript.

Major comments

1. How representative are the TMRCAs of the 10 replicate trees of the posterior distribution estimated for the TMRCA is BEAST2? This should be shown. It is of necessity a very small number of replicates (10 samples are usually not enough to sufficiently represent a distribution), and if they are all very close to each other, then this doesn't address the underlying issue - that the target tree's TMRCA distribution has an HPD interval of ~35 years, which is reduced to a single median estimate in the regression-ABC analysis.

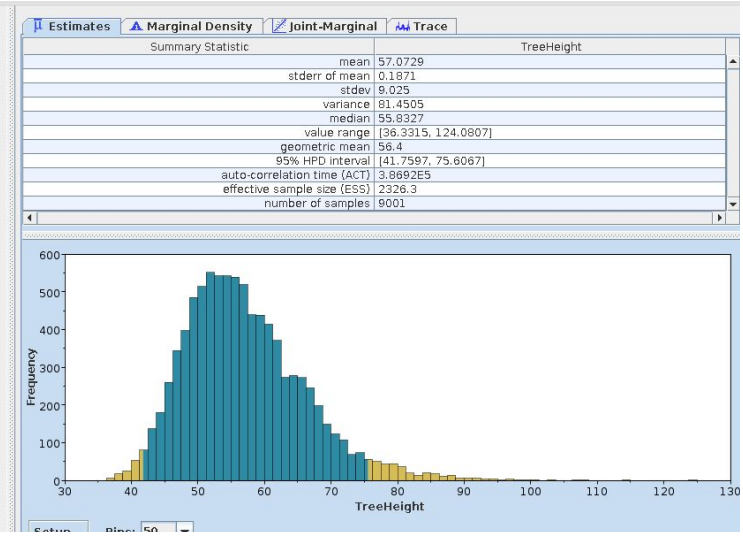
Thank you for the suggestion. Ten was indeed low, and we now use 100 replicate trees. We added a supplementary figure to show the resulting distributions for the tMRCA (Figure S6) and another supplementary figure to compare the median of these distributions to the posterior distribution of the studied tree (Figure S7).

The HPD interval of the tMRCA distribution is [1941.846;1975.516] and the HPD interval of the 100 randomly picked Beast trees is [1937.534;1971.837].

2. In the previous review I suggested that including a background set of global HCV genotype 1a sequences sampled over a larger time period will help with reducing the TMRCA uncertainty (adding more sequences over a longer time period means that internal node height estimates are more accurate). The target tree can then be pruned from this bigger tree. Alternatively, I would have to be convinced that the 10 replicates are representative of the posterior TMRCA distribution of the BEAST2 analysis (point 1).

We followed your suggestion and added 106 sequences sampled from 1994 and 2018, from the study by Vanhommerig et al.(entitled Limited overlap between phylogenetic HIV and HCV clusters illustrates the dynamic sexual network structure of Dutch HIV-infected MSM, published in *AIDS* in 2017). We then re-ran Beast2, using a fixed clock rate.

Unfortunately, we found that adding these sequences did not reduce the confidence interval for the tMRCA. See the figure below.



This result is not shown in the text.

3. Line 140: I agree that if new hosts are undersampled then the analysis is conservative. However, it seems more likely to me that new hosts are oversampled. If the sampling proportion is much higher for new hosts, and high assortativity holds, then we would expect sequences from new hosts to coalesce faster, even if both populations are growing at the same rate (since the population size of new hosts would be smaller). I think the phylogeny is consistent with both the interpretation in the text (equal sampling rates and different growth rates) and with very different sampling rates and equal growth rates.

To investigate the sensitivity of our ABC framework towards the potential sampling biases, we pruned the target phylogeny to get different phylogenies with 50% of the ‘new’ hosts tips. We then ran our ABC framework on each phylogeny and found that the results estimations were similar to those from the full phylogeny (although with wider confidence intervals). These results suggest that our ABC inference is not strongly affected by sampling biases.

Minor comments

1. The statement about BEAST2 being used to root a PhyML tree is still in the legend to Figure 1.

Thank you for the suggestion, we removed the statement.

2. tf in the methods should be t3

Thanks, we made the change.

3. Even with a fixed strict clock the target tree TMRCA is still very uncertain with an HPD of ~35 years (very much the same as when co-estimating both parameters). Could the authors please double-check that? Since the TMRCA and the clock-rate usually have a very strong negative correlation to each other, fixing one should also restrict the other. This could mean that the sequences are not very informative about

the branch lengths (in substitutions/site) or topology. Perhaps the authors could check if very few nodes in the MCC tree have a high posterior probability (or alternatively look at bootstrap scores in the ML tree).

We did actually fix the clock rate. This was decided because a TempEst analysis showed that the temporal signal was indeed poor.

We further checked the Beast output, and there are few nodes in the MCC tree that do have a high posterior probability. More precisely, around 20% of the nodes have a posterior probability higher than 0.7. This is not shown in the text.

4. In the coalescent simulation, are branching times drawn from exactly those transmission events in the simulated epidemiological trajectories (step one in the simulation procedure) or are they simply drawn from the probability of two lineages to coalesce given the population sizes from the simulated trajectories?

The branching times all correspond to actual transmission events in the simulated trajectory. However, the reverse is not true since not all transmission events will correspond to a branching. Indeed, we use a hypergeometric distribution to compute the probability for 2 lineages to coalesce given the size of the compartments.

We cite the preprint of our simulator where details on the algorithms are presented in Supplementary informations.

5. As explained by the editor, a multi-type birth death model is sufficient to model the scenario. However, the revised manuscript still contains a paragraph stating that birth death models are not applicable. Two types of state-change events can be modelled in BEAST2 with BDMM, the default is indeed mutation based, but in the other states change on a transmission event (exactly at branching times).

Thanks to help from Julija Pevcerska from Dr Tanja Stadler's team at ETH Zürich in Basel, we were able to set-up an XML file to run a BDMM model with our data.

Unfortunately, even after the MCMC was run for $5 \cdot 10^8$ iterations, the resulting ESS values for all the parameters remained lower than 200. Therefore, we were unable to conclude anything from this analysis (besides the limitation of the likelihood-based approach for this dataset).