Clock and TMRCA based on 27 genomes

Novel 2019 coronavirus

nCoV-2019 Genomic Epidemiology

Kristian_Andersen #1 January 31, 2020, 7:29am

Estimates of the clock and TMRCA for 2019-nCoV based on 27 genomes

January 25, 2020
Kristian Andersen, Scripps Research
kristian@andersen-lab.com

Following up on the analyses provided by Andrew Rambaut this is a brief report estimating the evolutionary rate and timing of the epidemic (date of the most recent ancestor (MRCA)) based on 27 publicly shared n2019-nCoV genome sequences. Compared to earlier analyses where several parameters had to be fixed, there is now enough information content in the sequences to obtain reasonable estimates of the clock and TMRCA without fixing parameters. This work is for information purposes only and is not intended for publication. All the data used here is provided by the laboratories listed below through NCBI Genbank or GISAID.

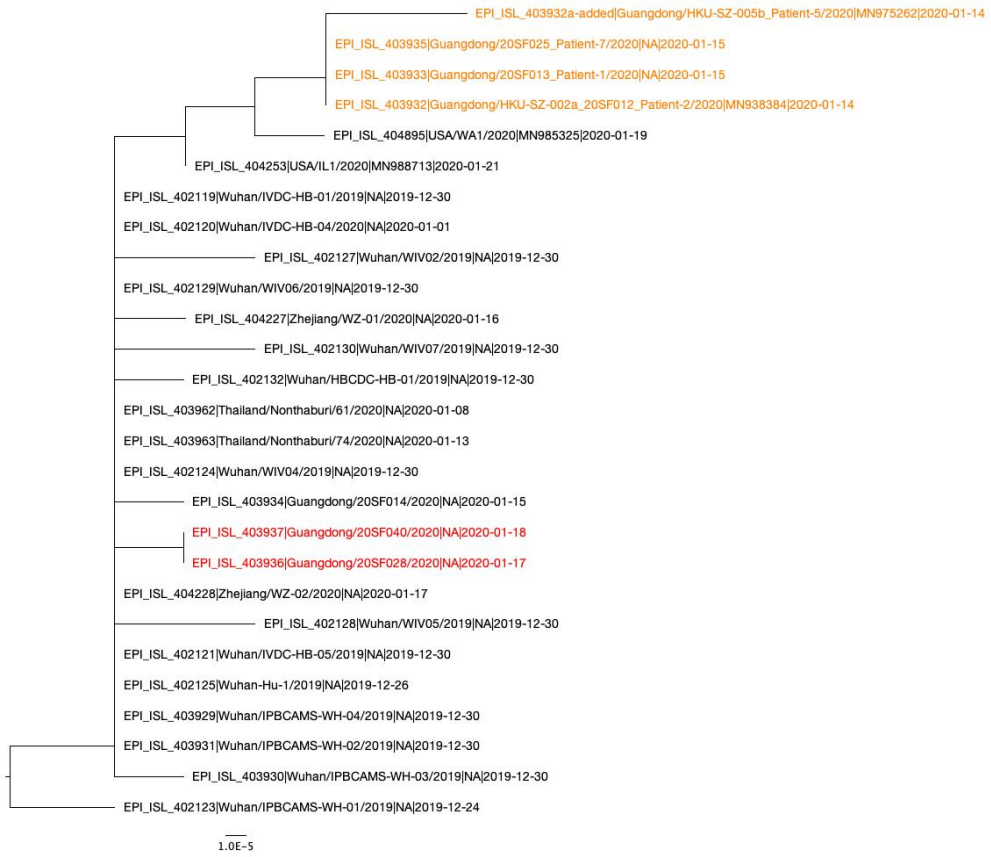
Data

As of January 25, 2020, 28 full-length nCoV-2019 genomes and 1 partial genome are available on the **GISAID** platform. The partial genome (EPI_ISL_402126) and one with too many sequencing errors (EPI_ISL_403928) were eliminated from these analyses. The final dataset contained 27 full-length nCoV-2019 genomes with 41 SNPs in total, 9 of them masked because of likely sequencing errors (leaving 32 SNPs in the dataset). Acknowledgements of the genome sequences used in this analysis are in the table at the end of this document.

Phylogenetic Tree

A phylogenetic tree was created using PhyML and in agreement with previous analyses, still shows limited genetic variation in the sampled viruses, which is consistent with a recent common ancestor. Two distinct clusters can also be seen from the tree, consistent with **reported human clusters of cases** (shown in orange and red).

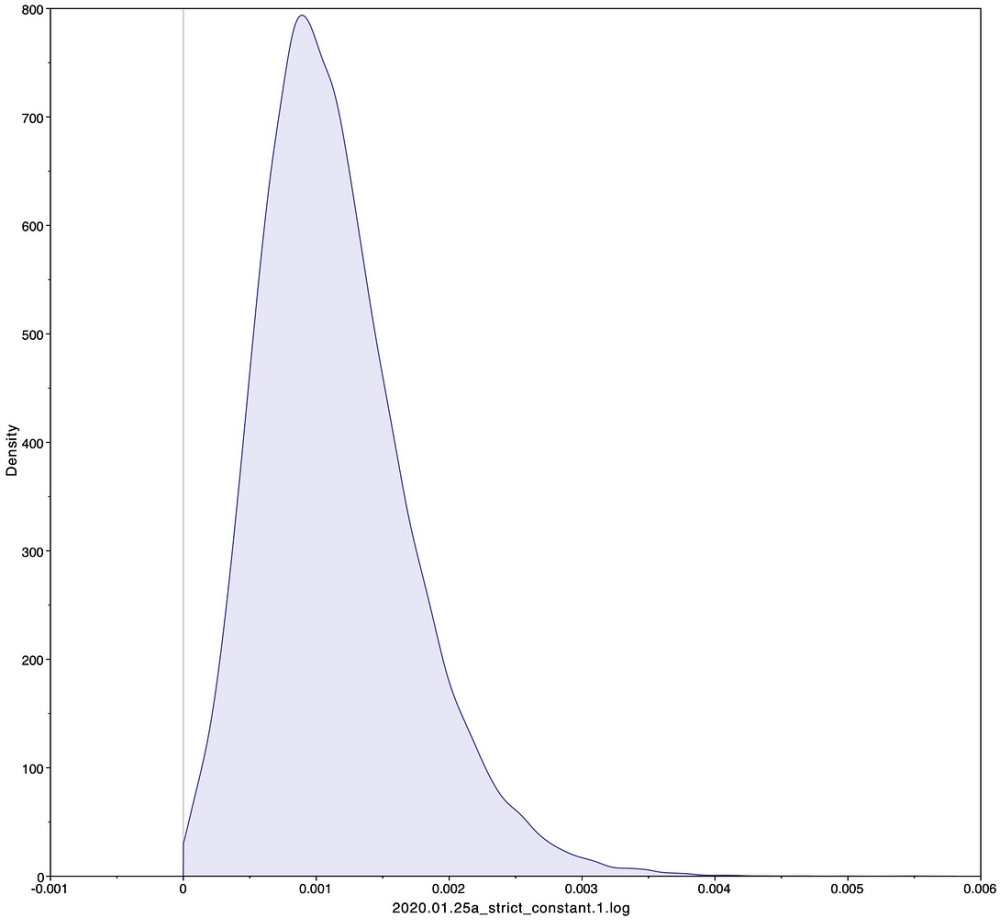
We are starting to see more structure in the tree and overall the genetic data is highly suggestive of a single-point introduction into the human population followed by sustained human-to-human transmission. This introduction was likely via either a single infected animal or a small cluster of recently infected animals directly into either a single human individual or a small cluster of human individuals. All subsequent cases are the result of human-to-human transmission with no further evidence of zoonotic transmissions.



Evolutionary rate

To estimate the substitution rate of nCoV-2019, I used **BEAST** with a simple model consisting of HKYγ, strict clock with a CTMC rate prior, and a constant tree prior. The median estimate for the substitution rate is very similar to other RNA viruses, including SARS-CoV, Ebola virus, Zika virus, and others at ~ 1E-3 subs/site/year. The range is still wide, but should improve as more sequence data is produced.

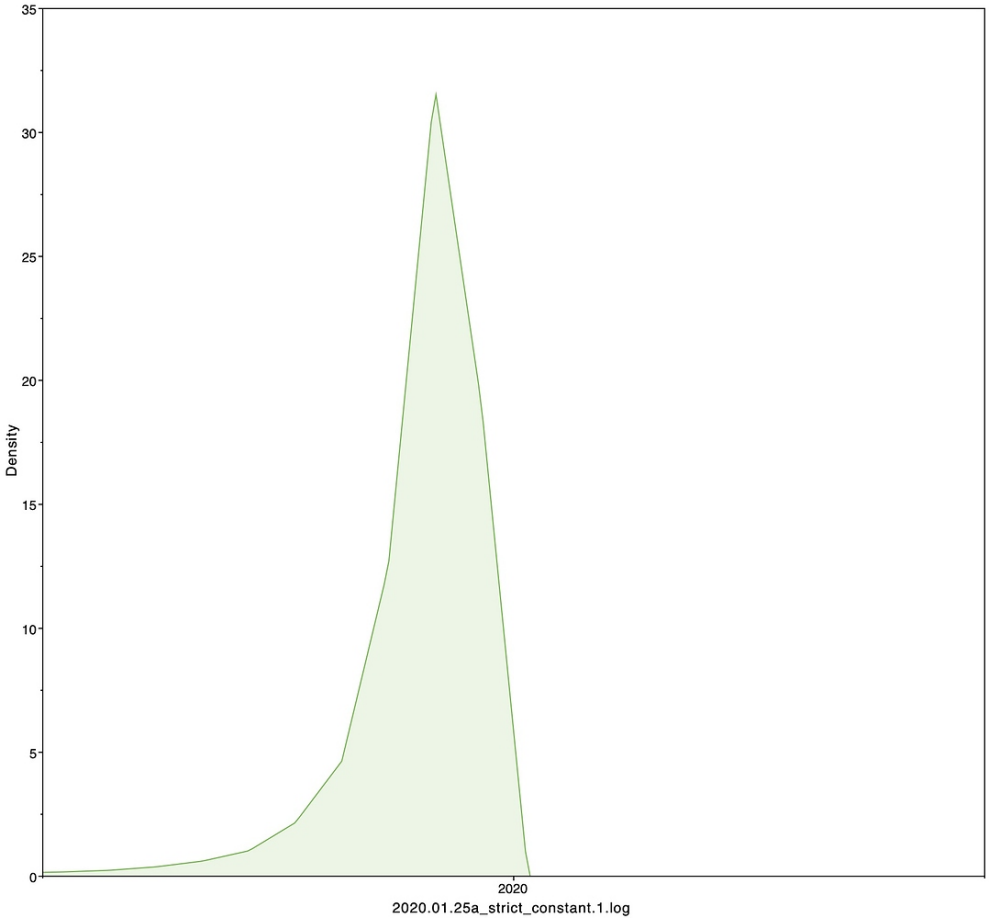
Median	95% HPD
1.067E-3	4.031E-6 - 5.53E-3



Date of the MRCA

I next estimated the date of the MRCA of the sampled nCoV-2019 genomes, corresponding to the point of the ancestral virus of all the sampled cases was in the same host - in other words, the initial spillover event leading to the outbreak. The first case symptoms was recently reported to be **December 1, 2019**, although WHO has previously reported this as **December 8, 2019**. The estimate from BEAST is in agreement with these dates, giving a median date of December 2, 2019. This date is also consistent with **prior phylogenetic analyses** using fixed rates of the evolutionary rate of 2019-nCoV.

Median	95% HPD
02 Dec 2019	01 Oct 2019 - 22 Dec 2019



Caveats

Earlier versions of our alignments had significant issues with sequencing errors (I estimate up to 50%). I believe that this issue is minimized in this dataset, with only 9/41 SNPs looking suspicious (and therefore masked in these analyses). That said, there is still limited variation in the sampled genomes and even small artefacts and sequencing errors could greatly influence the estimates.

The clock and TMRCA estimates have large intervals and the median values should be interpreted with caution. The ranges are more appropriate for interpretation of the dates, as opposed to any one of point media values mentioned above. They are all likely to change - possibly significantly - as more patients are sampled and genomes produced.

Acknowledgements and Genome Availability

Strain	Authors	Source	Lab
EPI_ISL_402119	Wenjie Tan, et al.	GISAID	National Institute for Viral Disease Control and Prevention, China CDC
EPI_ISL_402120	Wenjie Tan, et al.	GISAID	National Institute for Viral Disease Control and Prevention, China CDC
EPI_ISL_402121	Wenjie Tan, et al.	GISAID	National Institute for Viral Disease Control and Prevention, China CDC
EPI_ISL_402123	Lili Ren, et al.	GISAID	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College

Strain	Authors	Source	Lab
EPI_ISL_402124	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402125	Zhang, et al.	GISAID	National Institute for Communicable Disease Control and Prevention (ICDC) Chinese Center for Disease Control and Prevention (China CDC)
EPI_ISL_402127	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402128	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402129	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402130	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402132	Bin Fang, et al.	GISAID	Hubei Provincial Center for Disease Control and Prevention
EPI_ISL_403929	Lili Ren, et al.	GISAID	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
EPI_ISL_403930	Lili Ren, et al.	GISAID	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
EPI_ISL_403931	Lili Ren, et al.	GISAID	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
EPI_ISL_403932	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403933	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403934	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403935	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403936	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403937	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403962	Pilailuk, et al.	GISAID	Department of Medical Sciences, Ministry of Public Health, Thailand
EPI_ISL_403963	Pilailuk, et al.	GISAID	Department of Medical Sciences, Ministry of Public Health, Thailand
EPI_ISL_404227	Yin Chen, et al.	GISAID	Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention
EPI_ISL_404228	Yanjun Zhang, et al.	GISAID	Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention
EPI_ISL_404253	Ying Tao, et al.	GISAID	IL Department of Public Health Chicago Laboratory
EPI_ISL_404895	Queen, et al.	GISAID	Division of Viral Diseases, Centers for Disease Control and Prevention

Strain	Authors	Source	Lab
MN975262	Chan et al.	Genbank	State Key Laboratory of Emerging Infectious Diseases

How close are we to the 'phylodynamic threshold'? A simulation study

OliverPybus #2 January 26, 2020, 9:39pm

Thanks Kristian. Would it be possible to share the XML and MCC tree files here?

I'd like to add a note of caution on these early results. BEAST can sometimes give very plausible sounding rate estimates even when there is no temporal signal in the data, due to the strong influence of the tree and rate priors. One way to check this is to randomise the dates on the tips multiple times. If you get the same rates as your current estimate, then your current estimate may be dominated by the prior. If you impose a strong alternate prior on the TMRCA, does the data meaningfully pull the posterior away from that prior? And TempEst can visualise temporal signal.

Of course, the other TMRCA estimates here on **virological.org** rely on an external estimate of the evolutionary rate from other viruses, which has its own problems. So it's worth seeking consensus across multiple methods - which I think, so far, we have.

Kristian_Andersen #3 January 27, 2020, 12:32am

Hey Oli,

Yup, totally agree caution is needed here and I think that is mentioned above and also reflected by the HPDs. The signal is fairly robust, but of course dependent on the older samples as expected. Tip randomization leads to different results, although I have not explores this in-depth - I'm currently running additional analyses.

As for the temporal signal, the slope of the RTT is very similar to the estimate above at 8E-4 s/s/y and the intercept is November 27, 2019 - so again, very close to the estimate above. This is using the oldest sample as the root.

These are by no means perfect analyses and more sampling is for sure needed, but I think we're starting to get reasonable estimates as long as we consider the HPDs and interpret the values with those in mind.

I unfortunately can't share the raw files here as that'd violate the GISAID policy, but I'll share them with you in private, assuming you have GISAID access.

OliverPybus #4 January 27, 2020, 9:40am

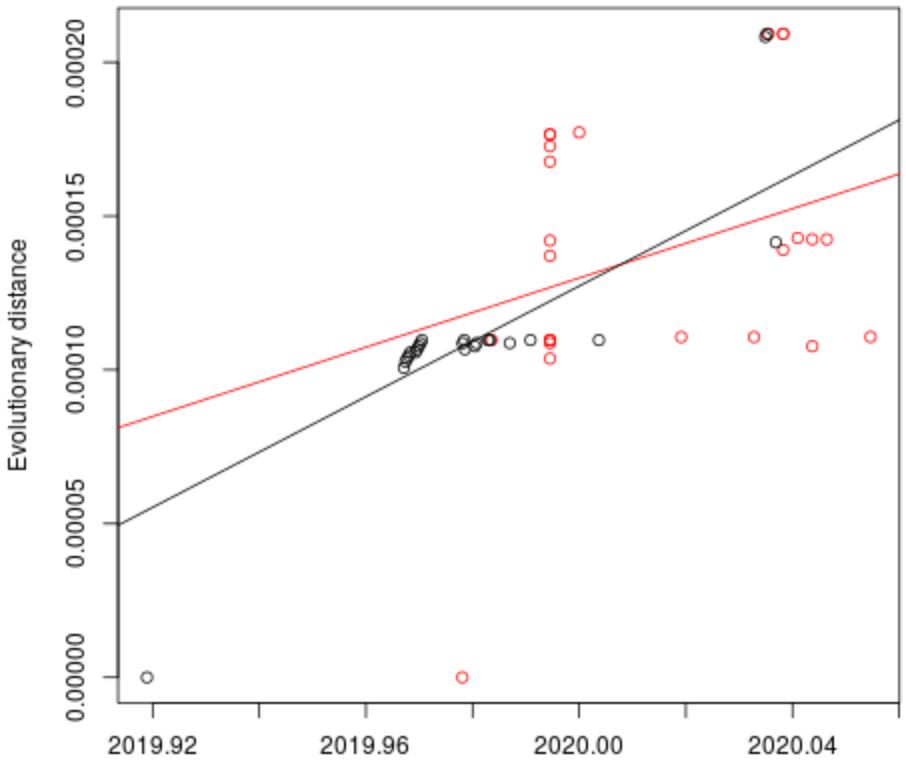
It's important to note that the HPDs are also influenced by the priors. So interpreting the results with the HPDs in mind does not ameliorate the problem I mentioned. If we are to present BEAST TMRCAs at this stage I think it's

vital to carefully compare the posterior estimate with the joint prior.

---

erik.volz #5 January 27, 2020, 11:07am

I have done a root-to-tip regression using the latest build from nextstrain (ML by iqtree, rooted in treedater). The results are not very compelling:




```
Root-to-tip mean rate: 0.00056301772302758
Root-to-tip p value: 0.139650296902652
Root-to-tip R squared (variance explained): 0.0923786229950628
```

treedater is not able to get a stable estimate of the rate with these data. If I fix the rate at 8e-4 I estimate the following dates:

```
"2019-12-02 09:54:53 UTC"
                     2.5%                      97.5%
"2019-11-02 22:07:58 UTC" "2019-12-16 09:36:07 UTC"
```

I suspect that getting robust & precise estimates of the TMRCA will require more samples, ideally including the animal reservoir, but having a better estimate of the rate would help a lot too.

Phylodynamic analysis of nCoV-2019 genomes – 29-Jan-2020

Kristian_Andersen #6 January 28, 2020, 4:43am

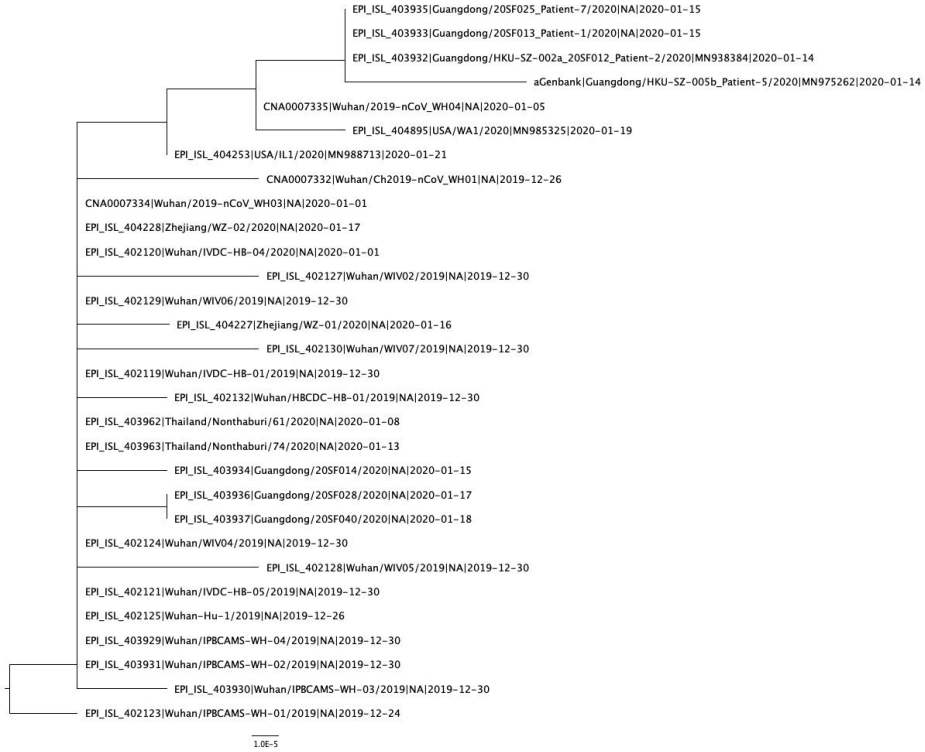
Three more high-quality full-length genomes were added to the **China National GeneBank**. I updated the analyses to include these new genomes and at the suggestion of Oliver Pybus, also provide an additional analysis randomizing tip dates to show that the estimates are informed more by the data than choice of priors. The main conclusions remain the same, although we now obtain significantly better mixing of the MCMC and narrower 95% HPDs. Despite this, please note that this is still work in progress and precaution should be taken interpreting the values and all estimates should be considered based on their intervals and not point values.

Data

As of January 27, 2020, 30 full-length nCoV-2019 high quality genomes are available. The final dataset contains 36 SNPs after removing 9 SNPs that are likely due to sequencing errors. Acknowledgements of the genome sequences used in this analysis are in the table at the end of this document.

Phylogenetic Tree

A phylogenetic tree was created using PhyML and in agreement with previous analyses still shows limited genetic variation in the sampled viruses, which is consistent with a recent common ancestor.

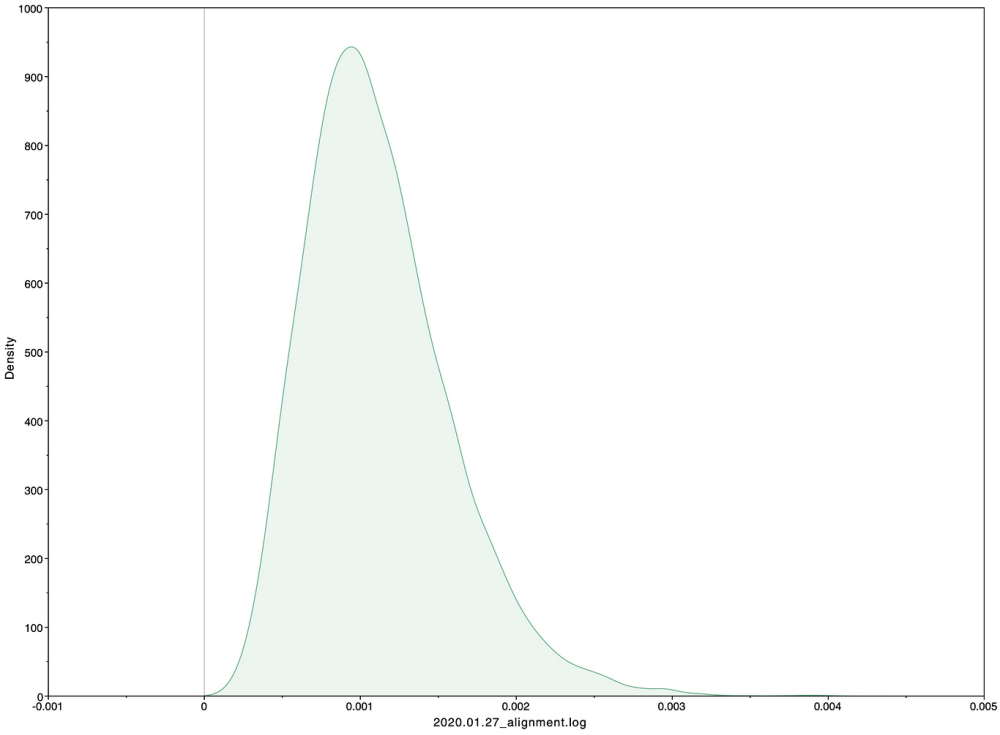


The genetic data is still highly suggestive of a single-point introduction into the human population followed by sustained human-to-human transmission with no further evidence of zoonotic transmissions.

Evolutionary rate

To estimate the substitution rate of nCoV-2019, I used **BEAST** with a simple model consisting of HKYγ, strict clock with a CTMC rate prior, and a constant tree prior. The median estimate for the substitution rate is very similar to other RNA viruses, including SARS-CoV, Ebola virus, Zika virus, and others at ~ 1E-3 subs/site/year. Compared to previous analyses, we're now starting to see better estimates of the rate.

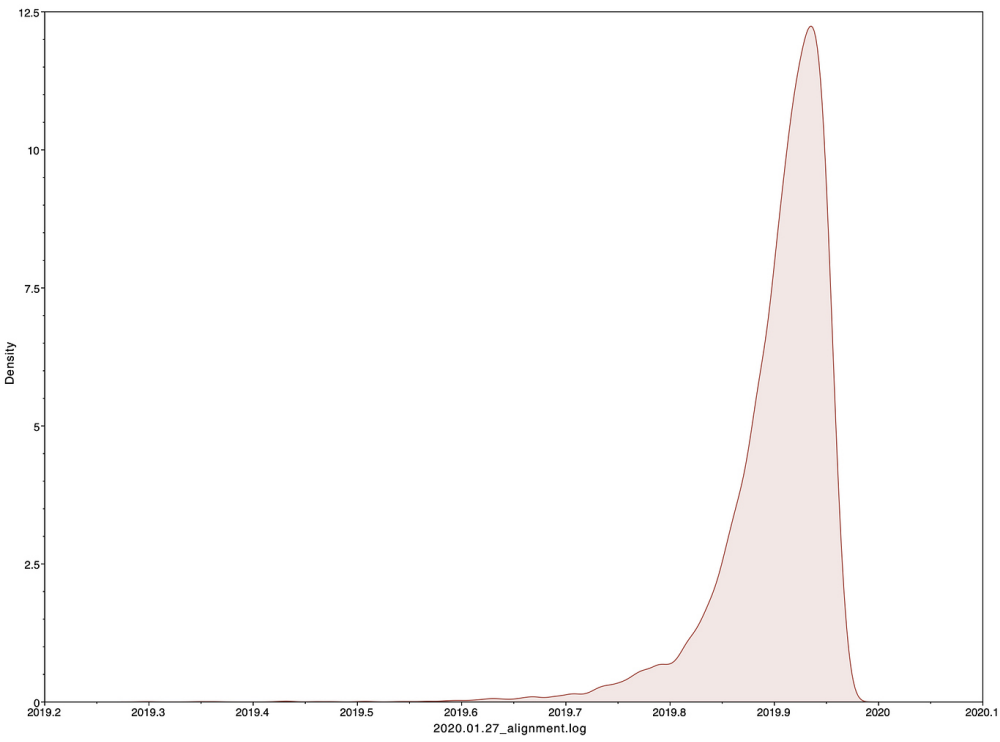
Median	95% HPD
1.05E-3	3.29E-4 - 2.03E-3



Date of the MRCA

I next estimated the date of the MRCA of the sampled nCoV-2019 genomes and the results were in agreement with previous estimates.

Median	95% HPD
01 Dec 2019	20 Oct 2019 - 20 Dec 2019



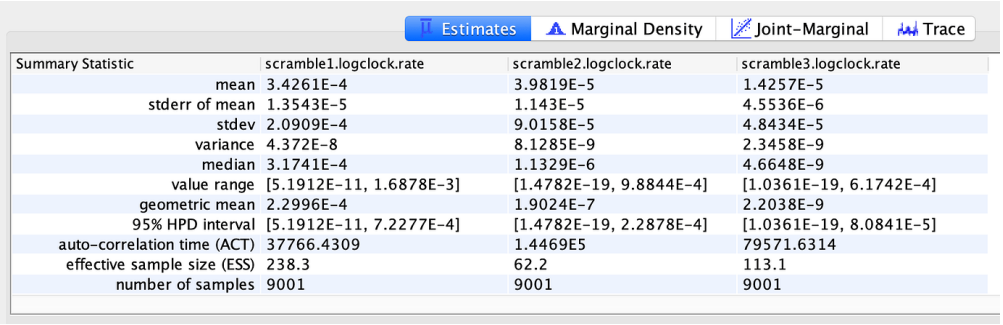
Randomization analysis

To test whether the estimates could have been strongly influenced by the priors, three independent alignments shuffling tip dates were created and analyzed with BEAST using the same model as described above. In all cases did these analyses fail to capture meaningful rate and date estimates, suggesting that the estimates above are

primarily informed by the data and not choice of priors. Despite this, all caveats described above still hold true and the dataset is still limited by size and sampling may be biased. This means that the addition of more sequencing data could likely change these estimates.

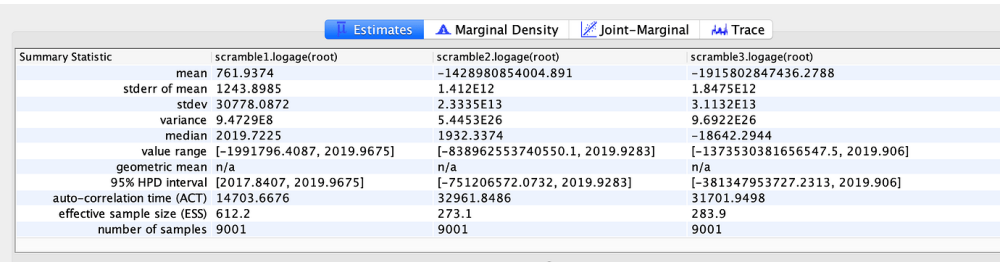
Note: ideally these analyses should be done 100+ times and include 'leave-one-out' analyses. I will try to do those at a later date.

Rate (three randomized alignments)



Summary Statistic	scramble1.logclock.rate	scramble2.logclock.rate	scramble3.logclock.rate
mean	3.4261E-4	3.9819E-5	1.4257E-5
stderr of mean	1.3543E-5	1.143E-5	4.5536E-6
stdev	2.0909E-4	9.0158E-5	4.8434E-5
variance	4.372E-8	8.1285E-9	2.3458E-9
median	3.1741E-4	1.1329E-6	4.6648E-9
value range	[5.1912E-11, 1.6878E-3]	[1.4782E-19, 9.8844E-4]	[1.0361E-19, 6.1742E-4]
geometric mean	2.2996E-4	1.9024E-7	2.2038E-9
95% HPD interval	[5.1912E-11, 7.2277E-4]	[1.4782E-19, 2.2878E-4]	[1.0361E-19, 8.0841E-5]
auto-correlation time (ACT)	37766.4309	1.4469E5	79571.6314
effective sample size (ESS)	238.3	62.2	113.1
number of samples	9001	9001	9001

TMRCA (three randomized alignments)



Summary Statistic	scramble1.logage(root)	scramble2.logage(root)	scramble3.logage(root)
mean	761.9374	-1428980854004.891	-1915802847436.2788
stderr of mean	1243.8985	1.412E12	1.8475E12
stdev	30778.0872	2.3335E13	3.1132E13
variance	9.4729E8	5.4453E26	9.6922E26
median	2019.7225	1932.3374	-18642.2944
value range	[-1991796.4087, 2019.9675]	[-838962553740550.1, 2019.9283]	[-1373530381656547.5, 2019.906]
geometric mean	n/a	n/a	n/a
95% HPD interval	[2017.8407, 2019.9675]	[-751206572.0732, 2019.9283]	[-381347953727.2313, 2019.906]
auto-correlation time (ACT)	14703.6676	32961.8486	31701.9498
effective sample size (ESS)	612.2	273.1	283.9
number of samples	9001	9001	9001

Log files can be downloaded **here**. Please contact me directly for XMLs and alignments.

Acknowledgements and Genome Availability

Strain	Authors	Source	Lab
EPI_ISL_402119	Wenjie Tan, et al.	GISAID	National Institute for Viral Disease Control and Prevention, China CDC
EPI_ISL_402120	Wenjie Tan, et al.	GISAID	National Institute for Viral Disease Control and Prevention, China CDC
EPI_ISL_402121	Wenjie Tan, et al.	GISAID	National Institute for Viral Disease Control and Prevention, China CDC
EPI_ISL_402123	Lili Ren, et al.	GISAID	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
EPI_ISL_402124	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402125	Zhang, et al.	GISAID	National Institute for Communicable Disease Control and Prevention (ICDC) Chinese Center for Disease Control and Prevention (China CDC)
EPI_ISL_402127	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences

Strain	Authors	Source	Lab
EPI_ISL_402128	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402129	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402130	Peng Zhou, et al.	GISAID	Wuhan Institute of Virology, Chinese Academy of Sciences
EPI_ISL_402132	Bin Fang, et al.	GISAID	Hubei Provincial Center for Disease Control and Prevention
EPI_ISL_403929	Lili Ren, et al.	GISAID	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
EPI_ISL_403930	Lili Ren, et al.	GISAID	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
EPI_ISL_403931	Lili Ren, et al.	GISAID	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
EPI_ISL_403932	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403933	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403934	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403935	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403936	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403937	Min Kang, et al.	GISAID	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention
EPI_ISL_403962	Pilailuk, et al.	GISAID	Department of Medical Sciences, Ministry of Public Health, Thailand
EPI_ISL_403963	Pilailuk, et al.	GISAID	Department of Medical Sciences, Ministry of Public Health, Thailand
EPI_ISL_404227	Yin Chen, et al.	GISAID	Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention
EPI_ISL_404228	Yanjun Zhang, et al.	GISAID	Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention
EPI_ISL_404253	Ying Tao, et al.	GISAID	IL Department of Public Health Chicago Laboratory
EPI_ISL_404895	Queen, et al.	GISAID	Division of Viral Diseases, Centers for Disease Control and Prevention
MN975262	Chan et al.	Genbank	State Key Laboratory of Emerging Infectious Diseases
CNA0007332	Chen et al.	China National GeneBank	BGI

Strain	Authors	Source	Lab
CNA0007334	Chen et al.	China National GeneBank	BGI
CNA0007335	Chen et al.	China National GeneBank	BGI

1 Like