nCoV-2019 origin time estimates

Using 20 genomes

Louis du Plessis, Oliver Pybus

Last modified: 24 Jan 2020

1 Data and Assumptions

There is a total of 25 genomes available[1]. We ignored two genomes (EPI_ISL_402120 and EPI_ISL_403928). There are two clusters with shared mutations, one from Shenzhen (with 3 genomes) and one from Zhuhai (with 2 genomes). Within each cluster all genomes are identical. We removed two genomes from the Shenzhen cluster and one from the Zhuhai cluster (reducing the clusters to single representatives, retaining only the oldest genome from each cluster).

Table 1: Dataset used (20 genomes with 20 observed mutations).

accession	strain	date	snps
EPI_ISL_404227	BetaCoV/Zhejiang/WZ-01/2020	2020-01-16	2
EPI_ISL_404228	BetaCoV/Zhejiang/WZ-02/2020	2020-01-17	0
EPI_ISL_402132	BetaCoV/Wuhan/HBCDC-HB-01/2019	2019-12-30	1
EPI_ISL_402127	BetaCoV/Wuhan/WIV02/2019	2019-12-30	2
EPI_ISL_402128	BetaCoV/Wuhan/WIV05/2019	2019-12-30	2
EPI_ISL_402129	BetaCoV/Wuhan/WIV06/2019	2019-12-30	0
EPI_ISL_402130	BetaCoV/Wuhan/WIV07/2019	2019-12-30	2
EPI_ISL_403963	BetaCoV/Nonthaburi/74/2020	2020-01-13	0
EPI_ISL_403962	BetaCoV/Nonthaburi/61/2020	2020-01-08	0
EPI_ISL_402119	BetaCoV/Wuhan/IVDC-HB-01/2019	2019-12-30	0
EPI_ISL_402121	BetaCoV/Wuhan/IVDC-HB-05/2019	2019-12-30	2
EPI_ISL_402124	BetaCoV/Wuhan/WIV04/2019	2019-12-30	0
EPI_ISL_402123	BetaCoV/Wuhan/IPBCAMS-WH-01/2019	2019-12-24	3
EPI_ISL_402125	BetaCoV/Wuhan-Hu-1/2019	2019-12-26	0
EPI_ISL_403931	BetaCoV/Wuhan/IPBCAMS-WH-02/2019	2019-12-30	0
EPI_ISL_403930	BetaCoV/Wuhan/IPBCAMS-WH-03/2019	2019-12-30	1
EPI_ISL_403929	BetaCoV/Wuhan/IPBCAMS-WH-04/2019	2019-12-30	0
EPI_ISL_403936	BetaCoV/Guangdong/20SF028/2020	2020-01-17	1
EPI_ISL_403934	BetaCoV/Guangdong/20SF014/2020	2020-01-15	1
EPI_ISL_403932	BetaCoV/Guangdong/20SF012/2020	2020-01-14	3

Assumptions

- Assume that all mutations are unique. This means we can estimate total tree length of the sample tree from the number of observed mutations and an estimate of the evolutionary rate. If we further assume the sample tree has a star topology then we can use the estimated tree length to infer the tMRCA of the sample.

[1] http://virological.org/t/phylogenetic-analysis-of-23-ncov-2019-genomes-2020-01-23/335

- Assume observed mutations are not sequencing errors.
- Assume sampling times are correct.
- Assume that the sample is representative of the outbreak as a whole.

2 Estimates of the tree length and tree height

We follow the methodology of Van Dooren (2003). Assume a homogenous Poisson process model for the evolutionary rate. The time since the start of the process is t, the evolutionary rate λ and the number of mutations obverved n. Given a rate and a number of observed mutations, the likelihood is:

$$L(t|\lambda, n) = \frac{e^{-\lambda t}}{n!}(\lambda t)^n$$

By maximising the likelihood we can obtain an ML-estimate of the time since the start of the process (in this case the total tree length) and 95% CIs using likelihood ratio tests under the χ^2_1 approximation for the likelihood ratio statistic. Note that because we only have very few observed mutations this is not a very good way of estimating CIs. Also, note that λ is measured in substitutions/year and **not** in substitutions/site/year (s/s/y). If the rate is in s/s/y then substitute $\mu\ell$ for λ where μ is the rate in s/s/y and ℓ is the length of the genome (in bp) ($\mu = \frac{\lambda}{\ell}$).

To convert from the tree length to a tMRCA (tree height), by assuming that the sample tree has a star topology, we have to take into account **all** genomes we have observed, **including** genomes without any observed mutations. Since not all genomes were sampled on the same date, the tree has a minimum length (m), accounting for the sum of all branch lengths from the collection date of the oldest to the most recent sample, which for this dataset is **217 days**. If t_0 is the collection date of the oldest sample and k is the number of samples collected, then the tMRCA (tree height) is given by:

$$t_{MRCA}(t, k, m, t_0) = \begin{cases} t_0 - \frac{(t-m)}{k} & t \geq m \\ t_0 & t < m \end{cases}$$

where m and t are measured in years. Note that it is very possible that the estimated tree length from the Poisson process model is less than m (here 217 days), in which case the treeheight is arbitrarily set to the oldest sampling time. The methodology is explained in the example shown in Figure 1.

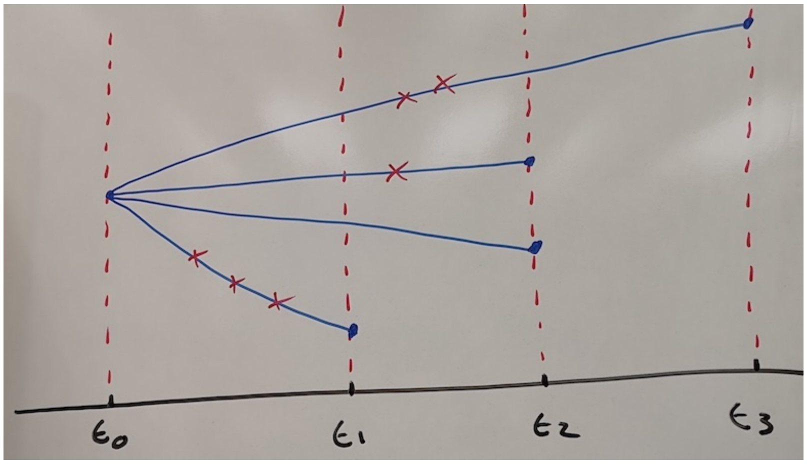

Figure 1: Example scenario for converting tree length estimates of a star topology to a tree height estimate when genomes have different collection dates.

In the example in Figure 1 there are $k = 4$ genomes, sampled at t_1, t_2 and t_3, with a total of $n = 5$ mutations between them (one genome has no mutations). We are trying to estimate the tree length (t) and the tMRCA (t_0). The tree length $t = (t_1 - t_0) + 2(t_2 - t_0) + (t_3 - t_0)$, which can also be written as $t = 4(t_1 - t_0) + 2(t_2 - t_1) + (t_3 - t_1)$. The minimum length of the tree occurs when $t_0 = t_1$, thus $m = 2(t_2 - t_1) + (t_3 - t_1)$.

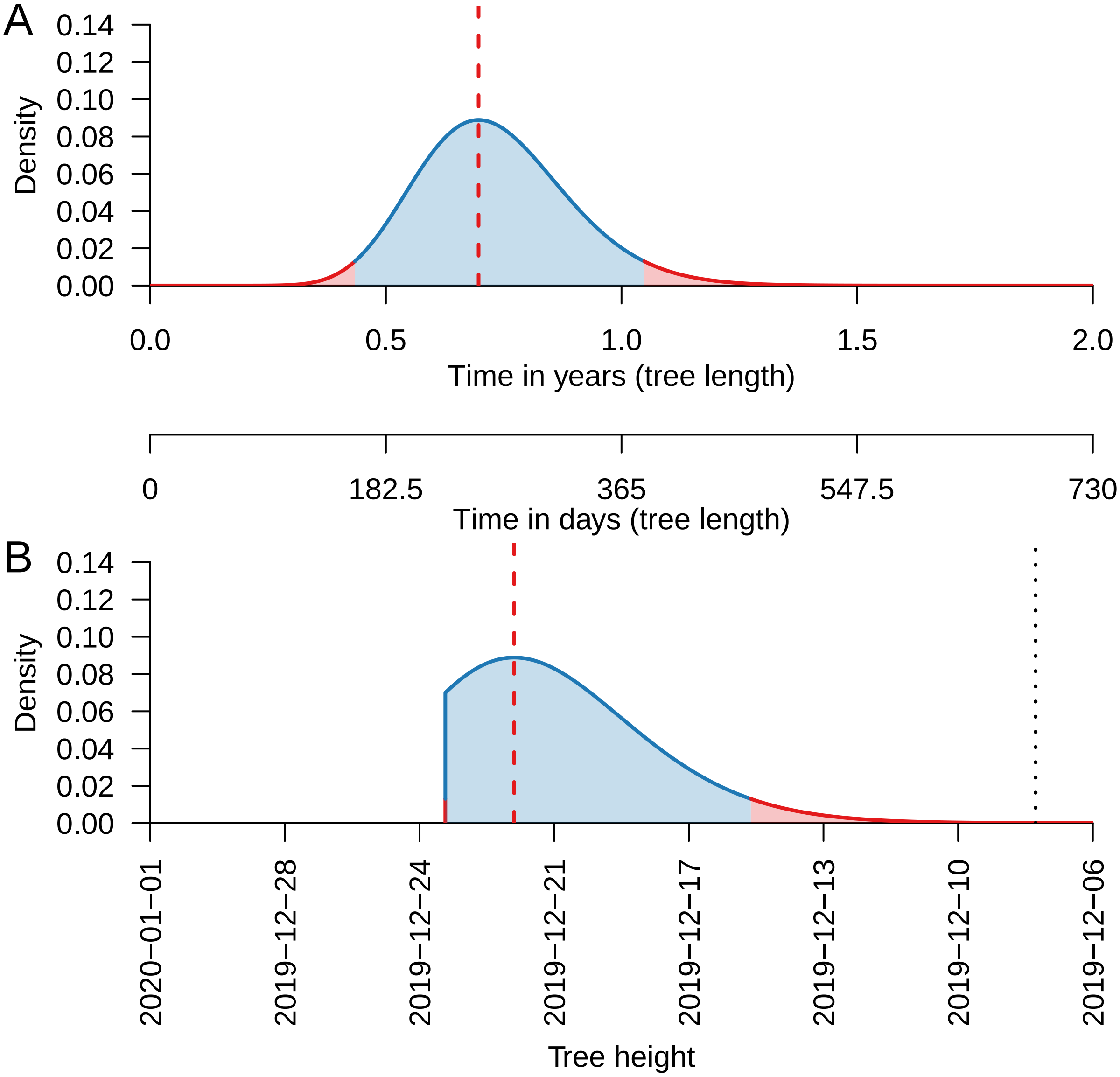

Figure 2: Likelihood curve for tree length (A) and tree height (B) calibrated by the point estimate from Dudas et al. (2018). ML estimate is the dashed line, regions outside the 95% CI are shaded in red. The dotted black line indicates the earliest reported date of symptom onset.

Assuming $\mu = 0.96 \times 10^{-3}$ s/s/y (the rate reported for MERS-CoV by Dudas et al. (2018)) and a genome length of $\ell = 29903$ bp, the length of the genome on Genbank[2], we get Figure 2. The ML-estimate of the total tree length is 0.6966915 years or 254.29 days, with CI [158.5, 382.58] days. Using the method illustrated in Figure 1 we obtain an ML estimate of the sample tMRCA of 2019-12-22 with CI [2019-12-24, 2019-12-16] (note upper limit is cut off by the oldest sample).

[2] https://www.ncbi.nlm.nih.gov/nuccore/MN908947

We can incorporate uncertainty in the rate estimate by marginalising across the rate. Incoroporating the rate range uniformly into the Poisson model (i.e. assuming a conservative uniform prior for the rate between λ_l and λ_u) we get:

$$L(t|n) = \int_{\lambda=\lambda_l}^{\lambda_u} \frac{e^{-\lambda t}}{n!} \frac{(\lambda t)^n}{(\lambda_u - \lambda_l)} d\lambda$$

which has solution:

$$L(t|n) = \frac{1}{t(\lambda_u - \lambda_l)} \sum_{i=0}^{n} \frac{1}{i!} \left[e^{-\lambda_l t} (\lambda_l t)^i - e^{-\lambda_u t} (\lambda_u t)^i \right]$$

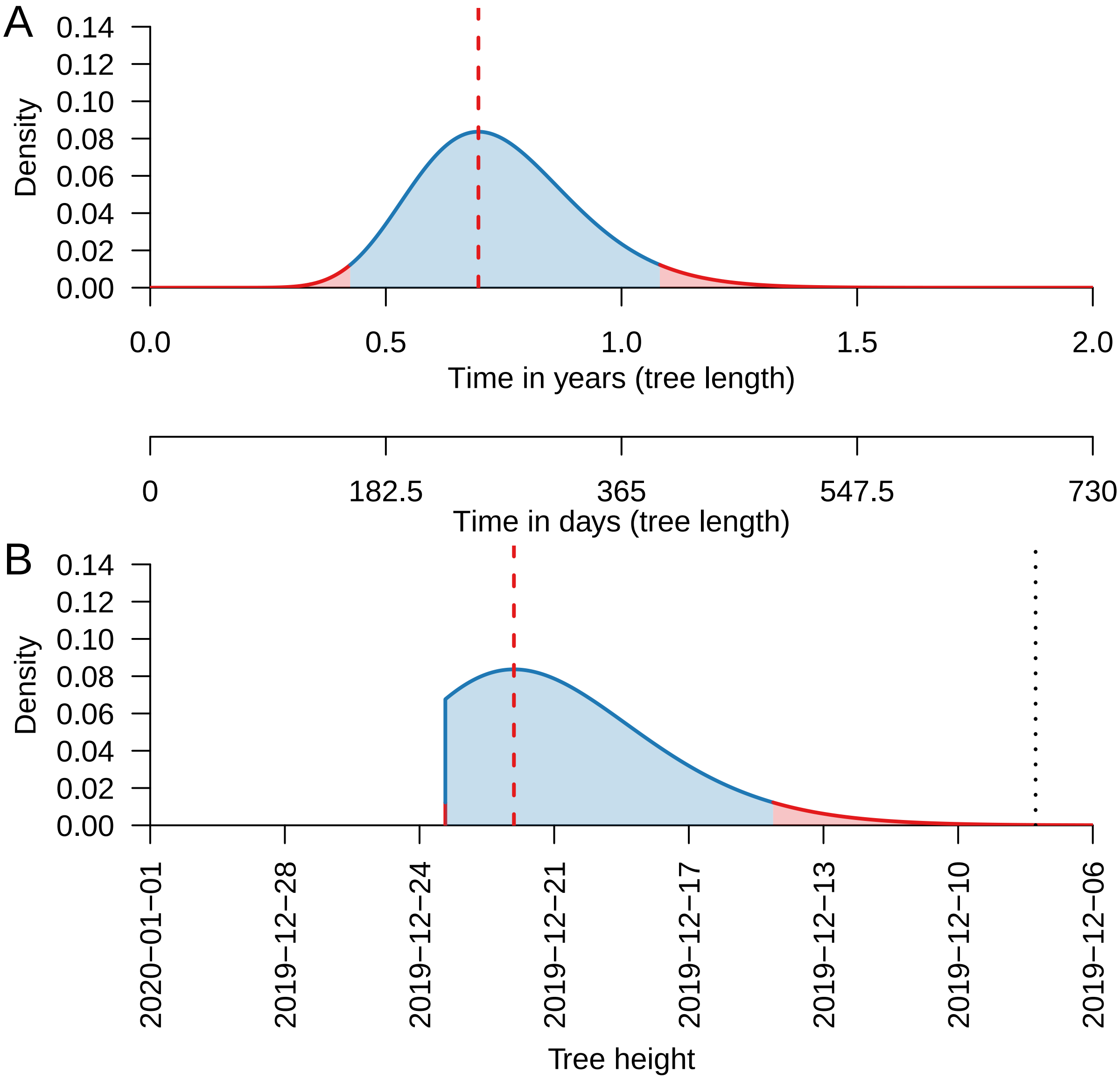

Figure 3: Likelihood curve for tree length (A) and tree height (B) calibrated by the rate range reported in Dudas et al. (2018). ML estimate is the dashed line, regions outside the 95% CI are shaded in red. The dotted black line indicates the earliest reported date of symptom onset.

Using the rate range reported by Dudas et al. (2018) we get Figure 3. The ML-estimate of the total tree length is 0.696379 years or 254.18 days, with CI [154.92, 394.68] days. Using the method illustrated in Figure 1 we obtain an ML estimate of the sample tMRCA of 2019-12-22 with CI [2019-12-24, 2019-12-15] (note upper limit is cut off by the oldest sample).

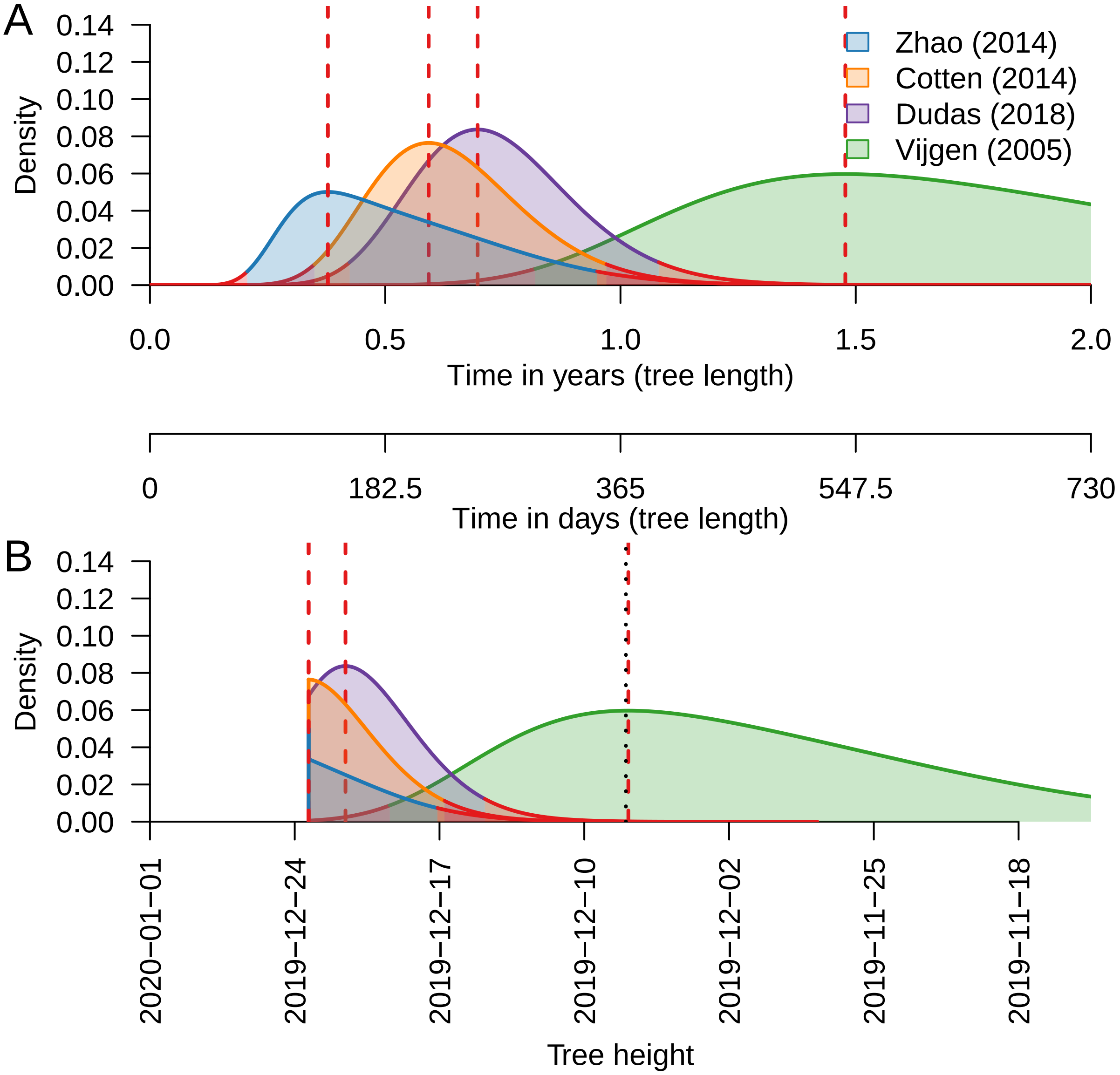

Figure 4: Likelihood curve for tree length (A) and tree height (B) calibrated by the rate ranges reported in different studies. ML estimates are the dashed lines, regions outside the 95% CI are shaded in red. The dotted black line indicates the earliest reported date of symptom onset.

Using rate ranges reported for various different coronavirus studies we get Figure 4 and the results below:

Calibration	Rate range (s/s/y)	Tree length (days)	tMRCA (date)
Zhao et al. (2004)	8×10^{-4}, 0.00238	138.04 [75.48, 346.89]	2019-12-24 [2019-12-18, 2019-12-24]
Cotten et al. (2014)	8.8×10^{-4}, 0.00137	216.24 [127.67, 354.05]	2019-12-24 [2019-12-17, 2019-12-24]
Dudas et al. (2018)	8.3×10^{-4}, 0.00109	254.18 [154.92, 394.68]	2019-12-22 [2019-12-15, 2019-12-24]
Vijgen et al. (2005)	2.7×10^{-4}, 6×10^{-4}	539.45 [298.84, 1076.48]	2019-12-08 [2019-11-11, 2019-12-20]

3 Conclusions

Using the rate estimates from Zhao et al. (2004), Cotten et al. (2014) or Dudas et al. (2018) to calibrate the model results in a tMRCA estimate for the **sample tree** that is more recent than the earliest reported date of symptom onset (8 December[3]). Rate estimates from Vijgen et al. (2005) result in an ML-estimate for the tMRCA that coincides with the earliest reported date of symptom onset, however that should not be taken as evidence that this rate estimate is more accurate. The rate reported in Vijgen et al. (2005) is considerably slower than the rates reported in the the other studies, resulting in an earlier tMRCA with a wider distribution. On the other hand the rate estimates reported in Zhao et al. (2004) have a very long tail, possibly biasing the model to faster rates and an unrealistically recent tMRCA, resulting in most of the tMRCA distribution being more recent than 24 December (Figure 4B).

The fact that our tMRCA is more recent than the date of first reported symptoms could be because (i) the true rate of evolution is slower than the rates reported in Dudas et al. (2018) and Cotten et al. (2014) or (ii) because the number of mutations in the genomes have been underestimated. The former is unlikely because observed molecular clock rates are expected to be higher when estimated over very short timescales, as is the case here. Secondly, the probability of true variants not being present in the observed genomes is unlikely. A third explanation is that there are unsampled cases. This could be either unsampled circulating diversity or earlier lineages that went extinct (Figure 5). This hypothesis implies the existence of a virus lineage in humans briefly before the tMRCA of the sample used here, thereby resolving the short discrepancy between the date of first reported symptoms and the estimated tMRCA of the sample tree.

The model used here assumes a star topology for the tree and assumes that the 9 genomes without any observed mutationes represent the true ancestral state (i.e. they are at the root of the genetic distance tree). For a given tree length, a star topology is the tree with the smallest possible height, thus our tree height estimates should be seen as a lower bound, and any departure from a star topology is likely to result in an older tMRCA estimate (e.g. if all 20 genomes share some mutations from the index case then the tMRCA will be more recent).

A similar approach followed by Richard Neher, but not accounting for uncertainty in the rate estimate, is available on Nextstrain[4]. Using a rate of 1×10^{-3} s/s/y it results in tMRCA estimates similar to our estimates calibrated by the rate estimates in Dudas et al. (2018) and Cotten et al. (2014), and using a rate of 5×10^{-4} it is similar to ours when calibrated by the rate estimates in Vijgen et al. (2005).

[3] https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/

[4] https://nextstrain.org/ncov

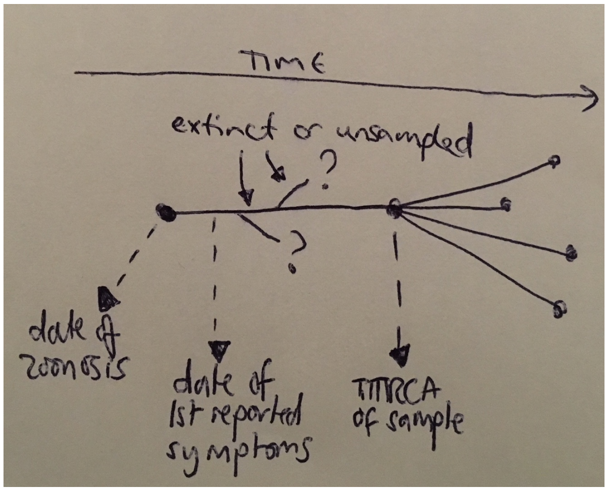

Figure 5: Possible scenarios for the outbreak.

References

Cotten, Matthew, Simon J. Watson, Alimuddin I. Zumla, Hatem Q. Makhdoom, Anne L. Palser, Swee Hoe Ong, Abdullah A. Al Rabeeah, et al. 2014. "Spread, Circulation, and Evolution of the Middle East Respiratory Syndrome Coronavirus." *mBio* 5 (1): e01062–13. doi:10.1128/mBio.01062-13.

Dudas, Gytis, Luiz Max Carvalho, Andrew Rambaut, and Trevor Bedford. 2018. "MERS-CoV Spillover at the Camel-Human Interface." *eLife* 7 (January): e31257. doi:10.7554/eLife.31257.

Van Dooren, S. 2003. "The Low Evolutionary Rate of Human T-Cell Lymphotropic Virus Type-1 Confirmed by Analysis of Vertical Transmission Chains." *Molecular Biology and Evolution* 21 (3): 603–11. doi:10.1093/molbev/msh053.

Vijgen, Leen, Els Keyaerts, Elien Moës, Inge Thoelen, Elke Wollants, Philippe Lemey, Anne-Mieke Vandamme, and Marc Van Ranst. 2005. "Complete Genomic Sequence of Human Coronavirus Oc43: Molecular Clock Analysis Suggests a Relatively Recent Zoonotic Coronavirus Transmission Event." *Journal of Virology* 79 (3): 1595–1604. doi:10.1128/JVI.79.3.1595-1604.2005.

Zhao, Zhongming, Haipeng Li, Xiaozhuang Wu, Yixi Zhong, Keqin Zhang, Ya-Ping Zhang, Eric Boerwinkle, and Yun-Xin Fu. 2004. "Moderate Mutation Rate in the SARS Coronavirus Genome and Its Implications." *BMC Evolutionary Biology* 4 (1): 21. doi:10.1186/1471-2148-4-21.