Origin time and epidemic dynamics of the 2019 novel coronavirus

Chi Zhang[1,2,*] and Mei Wang

January 25, 2020

[1]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China

[2]Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Beijing 100044, China

∗Corresponding author: E-mail: zhangchi@ivpp.ac.cn

Abstract

The 2019 novel coronavirus (2019-nCoV) have emerged from Wuhan, China. Studying the epidemic dynamics is crucial for further surveillance and control of the outbreak. We employed a Bayesian framework to infer the time-calibrated phylogeny and the epidemic dynamics represented by the effective reproductive number (R_e) changing over time from the genomic sequences available from GISAID. The origin time is estimated to be December 17, 2019 (95% CI: December 5, 2019 – December 23, 2019). R_e changes drastically over time, with highest value of 2.5 (0.5, 4.9) and lowest value of 0.3 (0.006, 1.47) (median and 95% highest posterior density interval). This study provides an early insight of the 2019-nCoV epidemic.

Introduction

An outbreak of a novel coronavirus (2019-nCoV) was reported in Wuhan, a city in central China (WHO; https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china). Coronaviruses cause diseases range from common cold to severe pneumonia. Two fatal coronavirus epidemics over the last two decades were severe acute respiratory syndrome (SARS) in 2003 and Middle East respiratory syndrome (MERS) in 2012 (WHO; https://www.who.int/health-topics/coronavirus). Human to human transmission has been confirmed for this new type of coronavirus (Wang et al. 2020) and more than 1,000 cases have been reported as of January 25, 2020 (https://bnonews.com/index.php/2020/01/the-latest-coronavirus-cases).

Studying the virus epidemic dynamics is crucial for further surveillance and control of the outbreak. Phylogeny of the viruses is a proxy of the transmission chain. Early studies of 2019-nCoV have focused on their molecular features and phylogenetic relationship with the close relatives (Zhou et al. 2020; Chan et al. 2020). The phylogenetic analyses typically ignored the sampling times and measursed branch lengths by expected number of substitutioins per site. They are also lacking a stochastic process to model the epidemic dynamics over time. However, both the timing and dynamics are critical to understand the early outbreak of 2019-nCoV. Furthermore, various sources of information and uncertainties are hard to be integrated in the analyses without employing a Bayesian approach.

In this study, we used the birth-death skyline serial (BDSS) model (Stadler et al. 2013) to infer the phylogeny, divergence times and epidemic dynamics of 2019-nCoV. This approach takes the genomic sequences and sampling times of the viruses as input, and co-estimtes the phylogeny and key epidemic parameters in a Bayesian framework. To our knowledge, this is the first study to perform such estimation on 2019-nCoV.

Results and Discussion

The phylogeny in Figure 1 shows the divergence times and relationships of the 24 BetaCoV viruses. The samples from Wuhan form a paraphyletic group, while the rest of the samples form a monophyletic clade. These two clades are not divergent from each other as their sequences are quite similar, which indicates that the outbreak is still in an early stage. The patients from Guangdong and Zhejiang had traveled from Wuhan and they had been infected before travelling (Chan et al. 2020). Note that this phylogeny is a maximum clade credibility (MCC) tree summarized from the posterior samples, which represents a best estimate of the topology. The probabilities in most clades are lower than 0.5 and would form polytomies if summarized as a 50% majority-rule consensus tree as shown in GISAID (https://www.gisaid.org).

The time of origin is estimated with median 31.9 days and 95% highest posterior density (HPD) interval from 26.3 to 43.8 days before the date of the latest sample (January 18, 2020), that is, December 17, 2019 (December 5, 2019 to December 23, 2019) (Table 1). This is on average 14 days before the cluster was initially reported (WHO, https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china), showing again an early phase of the epidemic.

The effective reproductive number (R_e) changing over time represents the epidemic dynamics of 2019-nCoV (Figure 2). R_e is defined as the number of expected secondary infections caused by an infected individual during the epidemic (Stadler et al. 2013). Thus, $R_e >$ 1.0 means that the number of cases are increasing and the epidemic is growing, whereas $R_e <$ 1.0 means that the epidemic is declining and will die out. Interestingly, the epidemic has an early boost before December 31 with R_e at around 2.0, and the number increases to over 2.5 from January 13 to 15, 2020. This is in agreement with some other studies reporting R_e between 3.3 and 5.5 (Read et al. 2020; Zhao et al. 2020). In comparison, the estimated R_e was 2.7 to 3.6 for SARS during the precontrol phase in HongKong (Riley et al. 2003). R_e is not consistenly larger than 1.0 as we also observe decreased values between January 1 to 10 and in the last three days. However, this is likely an artifact of lacking samples in these time intervals. Keep in mind that we used only 24 samples in our analysis, which is less than 2% of the reported number of infected patients. With more viurses sequenced, we would expect more reliable estimates which would provide better insights into the epidemic of 2019-nCoV.

Overall, this study provides an early insight of the 2019-nCoV epidemic by inferring key epidemiological parameters from the virus sequences. Such estimates would help public health officials to coordinate effectively to control the outbreak.

Methods

The original data downloaded from GISAID (https://www.gisaid.org) consists of 26 sequences of 2019-nCoV (as of January 24, 2020). We excluded two outliers, one is the virus from Kanagawa, Japan (EPI_ISL_402126) which contained only a small segment of 369bp, another is from Wuhan, China (EPI_ISL_403928) which contains suspiciously many mutations and has genetic distance about ten times longer than the rest of the sequences (GISAID; https://www.gisaid.org). Sequence alignment was done using MUSCLE {Edgar:2004bo}, resulting in a total length of 29904bp for the whole genome. The sampling times of the viruses ranged from December 24, 2019 (EPI_ISL_402123) to January 18, 2020 (EPI_ISL_403937) and they were used as fixed ages (in unit of years) in subsequent analysis.

We used the BDSS model (Stadler et al. 2013) implemented in the BDSKY package for BEAST 2 (Bouckaert et al. 2019) to infer the phylogeny, divergence times and epidemic dynamics of 2019-nCoV. The model has an important epidemiological parameter, the effective reproductive number R_e, defined as the number of expected secondary infections caused by an infected individual during the epidemic. The model allows R_e to change over time, making it feasible to estimate its dynamics (Stadler et al. 2013). The other two parameters are the becoming noninfectious rate δ and sampling proportion p, which were assumed constant over time. For the priors, R_e was assigned a lognormal(0, 1.25) distribution with median 1.0 and 95% quantiles between 0.13 and 7.82. δ was given a lognormal (2, 1.25) distribution with median 7.39 and mean 16.1, expecting the infectious period of an individual ($1/\delta$) to be about a month. The sampling proportion of infected individuals p of was a beta(1, 9) distribution with mean 0.1. The BDSS process starts from the origin time t_0, which was assigned a lognormal(–2, 1.5) distribution with median 0.135 (years before the lasest sampling time). Time from the origin to the lastest sample was divided into ten equally spaced intervals where R_e was estimated individually in each interval.

We used the lognormal independent relaxed clock (Drummond et al. 2006; Rannala and Yang 2007) to model evolutionary rate variation along the branches. The mean clock rate r was assigned a gamma(2, 0.0005) prior with mean of 0.001 substitutions per site per year and the standard deviation σ was a gamma (0.54, 0.38) prior with mean 0.2 and median 0.1 by default.

The substitution model used was HKY+Γ_4 (Hasegawa et al. 1985; Yang 1994) in which the transition-transversion rate ratio κ was set a lognormal(1, 1.25) prior and the gamma shape parameter α was an exponential(1) prior.

The analysis was performed in the BEAST 2 platform (Bouckaert et al. 2019). We ran 150 million MCMC iterations and sampled every 5000 iterations. The first 30% samples were discarded as burn-in. Convergence was diagnosed in Tracer (Rambaut et al. 2018) to confirm that independent runs gave consensus results and all parameters had effective sample size (ESS) larger than 100. The remaining 70% samples were used to build the maximum clade credibility (MCC) tree and to summarize the parameter estimates. The common ancestor heights were used to annotate the clade ages in the MCC tree.

Acknowledgments

CZ is supported by the 100 Young Talents Program of Chinese Academy of Sciences.

References

Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., De Maio N., Matschiner M., Mendes F.K., Müller N.F., Ogilvie H.A., Plessis du L., Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M.A., Wu C.-H., Xie D., Zhang C., Stadler T., Drummond A.J. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15:e1006650.

Chan J.F.-W., Yuan S., Kok K.-H., To K.K.-W., Chu H., Yang J., Xing F., Liu J., Yip C.C.-Y., Poon R.W.-S., Tsoi H.-W., Lo S.K.-F., Chan K.-H., Poon V.K.-M., Chan W.-M., Ip J.D., Cai J.-P., Cheng V.C.-C., Chen H., Hui C.K.-M., Yuen K.-Y. 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67:901–904.

Rannala B., Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Syst. Biol. 56:453–466.

Read J.M., Bridgen J.R., Cummings D.A., Ho A., Jewell C.P. 2020. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. medRxiv.

Riley S., Fraser C., Donnelly C.A., Ghani A.C., Abu-Raddad L.J., Hedley A.J., Leung G.M., Ho L.-M., Lam T.-H., Thach T.Q., Chau P., Chan K.-P., Lo S.-V., Leung P.-Y., Tsang T., Ho W., Lee K.-H., Lau E.M.C., Ferguson N.M., Anderson R.M. 2003. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science. 300:1961–1966.

Stadler T., Kühnert D., Bonhoeffer S., Drummond A.J. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc. Natl. Acad. Sci. USA. 110:228–233.

Wang C., Horby P.W., Hayden F.G., Gao G.F. 2020. A novel coronavirus outbreak of global health concern. The Lancet.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Zhao S., Ran J., MUSA S.S., Yang G., Lou Y., Gao D., Yang L., He D. 2020. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. bioRxiv.

Zhou P., Yang X.-L., Wang X.-G., Hu B., Zhang L., Zhang W., Si H.-R., Zhu Y., Li B., Huang C.-L., Chen H.-D., Chen J., Luo Y., Guo H., Jiang R.-D., Liu M.-Q., Chen Y., Shen X.-R., Wang X., Zheng X.-S., Zhao K., Chen Q.-J., Deng F., Liu L.-L., Yan B., Zhan F.-X., Wang Y.-Y., Xiao G., Shi Z.-L. 2020. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. bioRxiv.

Table 1. Posterior estimates of key model parameters

	median and 95% HPD interval
t_0	0.0873 (0.072, 0.120)
R_{e1}	1.32 (0.013, 5.92)
R_{e2}	1.72 (0.015, 5.58)
R_{e3}	1.07 (0.015, 4.50)
R_{e4}	1.79 (0.015, 3.51)
R_{e5}	0.28 (0.006, 2.61)
R_{e6}	0.31 (0.006, 1.47)
R_{e7}	0.59 (0.014, 1.93)
R_{e8}	0.59 (0.012, 2.10)
R_{e9}	2.54 (0.52, 4.87)
R_{e10}	0.84 (0.014, 2.73)
δ	183.25 (15.45, 451.97)
p	0.0099 (8.8E–7, 0.18)
r	0.0015 (0.00067, 0.0030)
σ	0.79 (2.3E–7, 1.40)
κ	7.38 (2.74, 16.05)
α	0.73 (0.0015, 3.00)

Note: time unit is years.

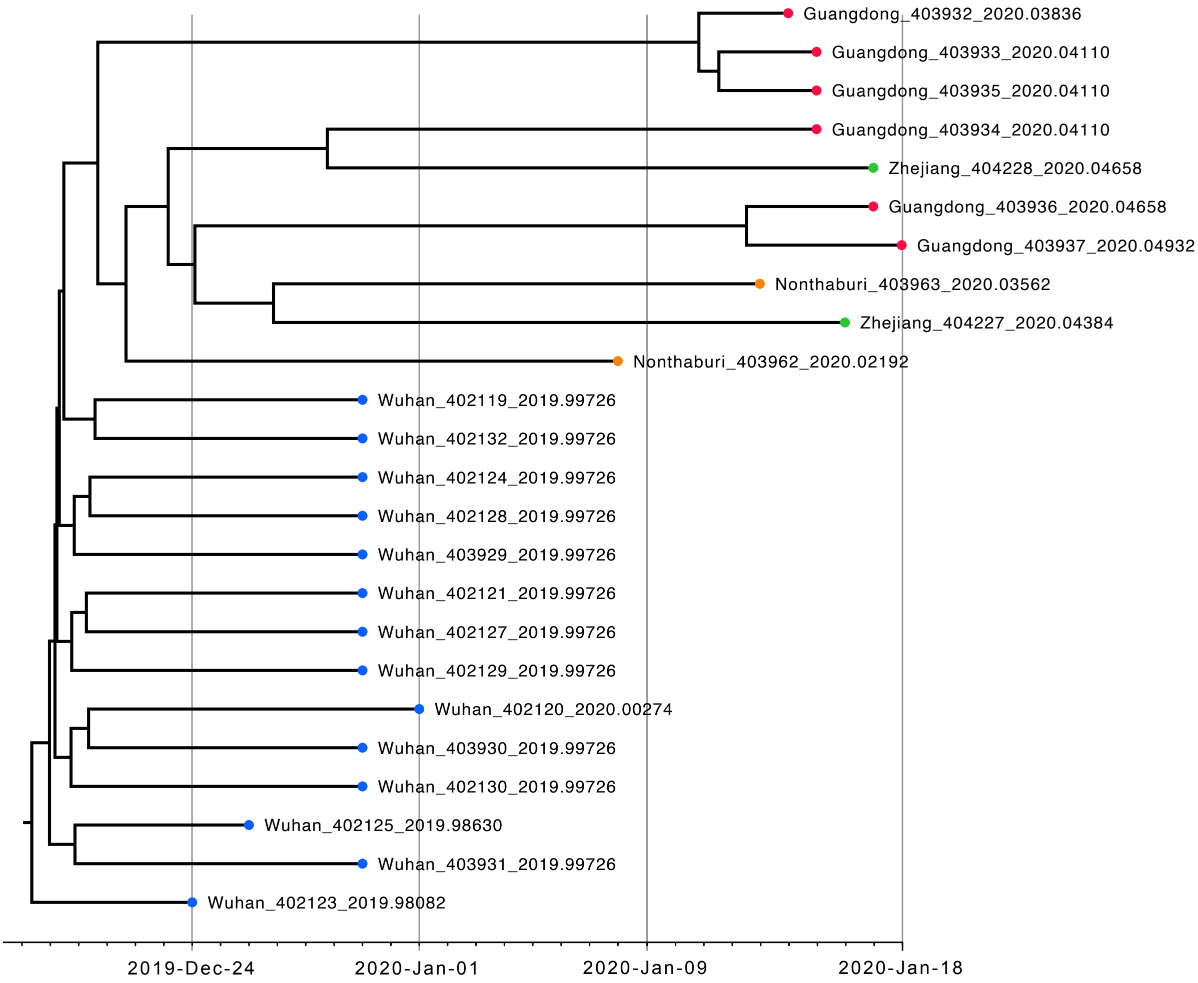

Figure 1. Maximum clade credibility (MCC) tree summarized from the MCMC sample. The common ancestor heights were used to annotate the clade ages.

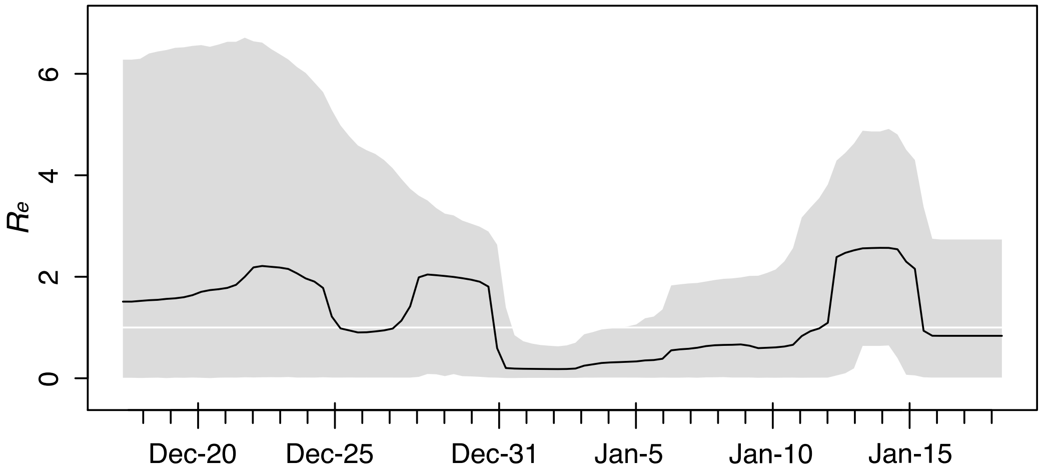

Figure 2. Skyline plot of the effective reproductive number (R_e) changing over time. The solid line indicates the median and the gray area indicates the 95% HPD interval.