Xingguang Li    ORCID iD: 0000-0002-3470-2196

Potential of large ‘first generation’ human-to-human transmission of 2019-nCoV

Xingguang Li1,,##, Junjie ZAI2,,*, Xiaomei Wang1,,*, and Yi Li1,,#

1. Hubei Engineering Research Center of Viral Vector, Wuhan University of Bioengineering, Wuhan 430415, China.

2. Immunology Innovation Team, School of Medicine, Ningbo University, Ningbo 315211, China.

Correspondence to:

Dr. Xingguang Li, Hubei Engineering Research Center of Viral Vector, Wuhan University of Bioengineering, Wuhan, 430415, China. Tel: +86-027-89648139, E-mail: xingguanglee@hotmail.com.

Prof. Yi Li, Hubei Engineering Research Center of Viral Vector, Wuhan University of Bioengineering, Wuhan, 430415, China. Tel: +86-027-89648361, E-mail: liyi@whsw.edu.cn.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jmv.25693.

This article is protected by copyright. All rights reserved.
Abstract

To investigate the genetic diversity, time origin, and evolutionary history of the 2019-nCoV outbreak in China and Thailand, a total of 12 genome sequences of the virus with known sampling date (24 December 2019 and 13 January 2020) and geographic location (primarily Wuhan city, Hubei Province, China, but also Bangkok, Thailand) were analyzed. Phylogenetic and likelihood-mapping analyses of these genome sequences were performed. Based on our results, the star-like signal and topology of 2019-nCoV may be indicative of potentially large ‘first generation’ human-to-human virus transmission. We estimated that 2019-nCoV likely originated in Wuhan on 9 November 2019 (95% credible interval: 25 September 2019 and 19 December 2019), and that Wuhan is the major hub for the spread of the 2019-nCoV outbreak in China and elsewhere. Our results could be useful for designing effective prevention strategies for 2019-nCoV in China and beyond.

Keywords

2019-nCoV; TMRCA; outbreak; Wuhan; Bangkok
Introduction

The current outbreak of the novel coronavirus 2019-nCoV, which was first reported in the Chinese city of Wuhan on 31 December 2019, can cause serious pneumonia and is now known to spread from person to person. From 24 January 2020, hundreds of millions of people will travel to their hometowns or overseas for the Chinese New Year holidays, which is China’s most significant annual holiday. Authorities in China and around the world have mounted an enormous operation to track and screen travelers from Wuhan in central China, where the first cases were documented. As of 26 January 2020, there have been 2744 cases of 2019-nCoV confirmed in mainland China, including 80 deaths, 461 serious, and 51 discharged, as well as eight in Hong Kong, five in Macao, and four in Taiwan. Thirty-three exported cases to other countries have been confirmed, including seven in Thailand, three in Japan, three in South Korea, three in the United States, two in Vietnam, four in Singapore, three in Malaysia, one in Nepal, three in France, and four in Australia. Of note, the 2019-nCoV disease has not been found in humans previously; therefore, determining how 2019-nCoV spreads is the most urgent question surrounding the outbreak. It is also imperative that we discover whether 2019-nCoV has the potential to cause an outbreak similar to the 2002–2003 outbreak caused by severe acute respiratory syndrome coronavirus (SARS-CoV)\(^1\text{-}^3\), which emerged in southern China and eventually killed 774 people in 37 countries. 2019-nCoV and SARS-CoV are
members of the large family of coronaviruses, which also includes the viruses
responsible for the Middle East respiratory syndrome (MERS)\textsuperscript{4,5}. Field studies have
revealed the original source of SARS-CoV and MERS-CoV to be the bat\textsuperscript{6-9}, with
masked palm civets (a mammal native to Asia and Africa)\textsuperscript{10-12} and camels\textsuperscript{13,14},
respectively, serving as intermediate hosts between bats and humans. To date,
2019-nCoV is most closely associated with SARS and related viruses that circulate in
bats\textsuperscript{15}. The hypothesis that 2019-nCoV jumped from an animal at the Wuhan Huanan
Seafood Wholesale Market, which sold processed meats and live consumable animals,
is strongly supported by previous study suggesting that this virus came from snakes\textsuperscript{15}.
Therefore, the 2019-nCoV outbreak implies that human consumption of wild animals
should be limited to prevent zoonotic disease infection.

Professor Zhong Nanshan, a SARS intervention specialist, is the respiratory
researcher leading the Chinese government’s expert panel on the 2019-nCoV
outbreak. After a visit to Wuhan city on 20 January 2020, Professor Zhong confirmed
that 2019-nCoV is spreading between people, and further confirmed that 14 medical
workers had been infected by one person, raising concerns that certain people may be
‘super-spreaders’ of the virus. Determining the genome sequences of 2019-nCoV
strains can offer key information about viral origin and dissemination. Thus far,
researchers have played a crucial role in the rapid sequencing, publishing, and sharing
of genome sequences obtained from 2019-nCoV strains found in infected people from China and Thailand.

In the present study, we employed state-of-the-art methods to investigate the time origin and potential rapid expansion of this virus in humans based on 12 genome sequences of 2019-nCoV with known sampling date (24 December 2019 and 13 January 2020) and geographic location (Wuhan, Hubei Province, China, and Bangkok, Thailand). Our study should provide insight into the time of origin and evolutionary history of the 2019-nCoV outbreak in China and Thailand. This research may be useful for designing effective 2019-nCoV prevention strategies in China and beyond.

Materials and methods

Collation of 2019-nCoV genome sequence dataset

As of 19 January 2020, 13 genome sequences of 2019-nCoV have been released on GISAID (http://gisaid.org/) and one strain (Virus name: BetaCoV/Wuhan-Hu-1/2019; Accession ID: EPI_ISL_402125) has been released on GenBank (https://www.ncbi.nlm.nih.gov/nuccore/MN908947) by the Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China. This strain was the first released 2019-nCoV genome sequence but has since been updated three times. In this study, we used the latest version (MN908947.3), and therefore the

This article is protected by copyright. All rights reserved.
genome sequence of strain BetaCoV/Wuhan-Hu-1/2019 from GISAID (EPI_ISL_402125) was excluded. Isolate BetaCoV/Wuhan/IVDC-HB-04/2020 (EPI_ISL_402120) shows evidence of sequencing artefacts and thus was also excluded in this study. The final dataset included 12 genome sequences of 2019-nCoV with known sampling date and city (10 from individuals in Wuhan, Hubei Province, China, with samples collected between 24 and 30 December 2019, and two from Chinese individuals in Bangkok, Thailand, who had recently traveled from Wuhan, with samples collected between 8 and 13 January 2020) (Supplementary Table S1). For this dataset, the 2019-nCoV genome sequences were aligned using MAFFT v7.22216 and then manually curated using BioEdit v7.2.517.

Phylogenetic analyses

To study the amount of evolutionary information contained in the dataset, likelihood-mapping analysis18 was performed using TREE-PUZZLE v5.319, with 30,000 randomly chosen quartets for the entire tree. For each sequence quartet, three unrooted tree topologies are possible. For a random sample of quartets, the likelihoods for the three possible topologies are reported as dots in an equilateral triangle. The distribution of points in different sections of this triangle indicates the tree-likeness of the data: the three corners represent fully resolved tree topologies, indicating the presence of a tree-like phylogenetic signal; the center represents the sets of points where all three trees are equally supported, indicating a star-like phylogenetic signal;
and the three areas on the sides indicate support for conflicting tree topologies. To infer phylogeny, we used the maximum-likelihood approach with the Hasegawa-Kishino-Yano nucleotide substitution model with gamma-distributed rate variation among sites (HKY + G) in PhyML v3.1. Support for the inferred relationships was determined by bootstrap analysis with 1 000 replicates. To reconstruct the evolutionary history of 2019-nCoV, we used Bayesian inference through a Markov chain Monte Carlo (MCMC) framework implemented in BEAST v1.8.4, with BEAGLE library v2.1.2 used to increase computational performance. As no temporal signal was present in the dataset, we employed HKY + G, as well as a constant size coalescent tree prior and strict molecular clock model assuming an evolutionary rate of $4.59 \times 10^{-4}$ substitutions per site per year to estimate the time to most recent common ancestor (TMRCA) for this dataset based on previous genetic analysis. We also placed a log-normal prior (mean = $4.59 \times 10^{-4}$ substitutions per site per year, 95% Bayesian credible interval: $1.178 \times 10^{-4} - 3 \times 10^{-3}$ substitutions per site per year) on the evolutionary rate with a strict molecular clock model based on previous studies. To ensure adequate mixing of model parameters, MCMC chains were run for 100 million steps with sampling every 10 000 steps from the posterior distribution. Convergence was evaluated by calculating the effective sample sizes of the parameters using Tracer v1.7.1. Trees were summarized as maximum clade
credibility (MCC) trees using TreeAnnotator after discarding the first 10% as burn-in, and then visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

**Ancestral reconstructions of discrete traits**

To perform ancestral reconstruction of the unobserved sampling locations (k = 2), discrete phylogeographic analyses were performed using the empirical tree distributions generated for our 2019-nCoV dataset. The location exchange process was modeled using asymmetric continuous-time Markov chains (CTMC) with an approximate CTMC conditional reference prior on the overall rate scaler and a uniform prior distribution. Bayesian analysis was run using BEAST v1.8.4 and BEAGLE library v2.1.2 for an MCMC chain of 100 million iterations, with 10 000 samples of all parameters and 10 000 trees for our dataset.

**Identifying pathways of virus spread using graph hierarchies**

To identify a subset of well-supported migration events among 2019-nCoV strains, we used a Bayesian stochastic search variable selection procedure (BSSVS) with a hierarchical prior on location indicators (0-1), which allow CTMC rates to shrink to zero with some probability, using BEAST v1.8.4. Strongly supported rates of virus movement (Bayes factor > 10) were identified using SpreaD3 v0.9.6.

This article is protected by copyright. All rights reserved.
Results

Demographic characteristics of dataset

The dataset included 12 genome sequences of 2019-nCoV strains from Wuhan, China ($n = 10$), with sampling dates between 24 and 30 December 2019, and from Bangkok, Thailand ($n = 2$), with sampling date between 8 and 13 January 2020 (Supplementary Table S1). The samples were primarily from Wuhan (83.33%), which is also the city of the original 2019-nCoV outbreak.

Likelihood-mapping and phylogenetic analyses

For this dataset, our likelihood-mapping analysis revealed a strong star-like topology signal (Fig. 1), indicating very limited genetic diversity for 2019-nCoV as of now. Furthermore, phylogenetic analysis of the 12 genome sequences of 2019-nCoV also showed a lack of genetic diversity, in accordance with the likelihood-mapping analysis (Fig. 2). The correlation between sampling date and divergence to the root of midpoint-rooted maximum-likelihood phylogeny indicated no temporal signal (Fig. 3). Therefore, we used an assumed evolutionary rate of $4.59 \times 10^{-4}$ substitutions per site per year from previous study\textsuperscript{23}. The estimated TMRCA for 2019-nCoV was 9 November 2019 (95% Bayesian credible interval: 25 September 2019 and 19 December 2019). We also used an assumed evolutionary rate range (95% Bayesian credible interval: $1.178 \times 10^{-4} - 3 \times 10^{-3}$ substitutions per site per year) for this dataset.
based on previous studies\textsuperscript{24-26}, and estimated that the TMRCA for 2019-nCoV was 3 September 2019 (95% Bayesian credible interval: 25 March 2019 and 20 December 2019).

**Dynamics analysis of ancestral discrete traits**

Our phylogeographic analysis revealed that the most probable root location of the 2019-nCoV ancestor was in Wuhan, China (posterior probability = 0.98) (Fig. 4). Our results also revealed that Wuhan acted as a diffusion epicenter compared to Bangkok. Most viral transmissions between epidemiologically linked cities were from Wuhan to Bangkok (mean estimate 1.96; 95% credible interval: 0.99–2.42; Bayes factor = 1 631).

**Discussion**

From the 12 genome sequences of 2019-nCoV, which included city of origin and date of sampling, our likelihood-mapping analysis confirmed that this novel virus has spread rapidly, indicating that the scale of transmission may be large given the virus spillover from animal reservoir to humans and high secondary human-to-human transmission. Based on confirmation from Professor Zhong, who is leading the government’s expert panel on the outbreak, 14 health-care workers appear to have been infected by one person who carried the virus. This suggests that epidemic super-spreaders may have been generated at the very beginning of the outbreak and
occurred once this novel coronavirus jumped from animals to humans. In addition, as mass travel for the Chinese New Year holiday could spread this virus farther and faster, it is critical to understand the Wuhan 2019-nCOV outbreak and predict how easily the virus can transmit between humans and whether the outbreak has the potential to persist. Based on our time-scale phylogenetic analysis of 12 genome sequences, the observed lack or limited diversity of the 2019-nCOV strains suggests that the common ancestor of this virus in humans possibly occurred on 9 November 2019.

However, we acknowledge that the estimates presented here relate only to a limited number of 2019-nCOV genome sequences and our results are estimated with some uncertainty. Therefore, our conclusions should be considered preliminary and explained with caution. As the outbreak continues, it is possible that the addition of new sequences to our analyses could change the above results significantly. The generation of additional genomic datasets from Wuhan, as well as other regions of China and beyond, and from humans and animals that have transmitted 2019-nCOV, will help clarify the extent of human-to-human transmission and identify the genetic changes that allowed this virus to jump species. Future study should include comparative analyses of between-host and within-host transmission of 2019-nCOV, and finer scale epidemiology to elucidate the transmission dynamics of different hosts through time and space. Our results show that the 2019-nCOV outbreak has been
shaped by migration and travel and therefore it is important to consider both local, national, and international strategies when designing interventions to end 2019-nCOV transmission in China and beyond.

Taken together, our results emphasize the importance of using likelihood-mapping and phylogenetic analyses to provide insights into the origin, evolutionary history, and spread of 2019-nCOV over a short time scale. These efforts, combined with epidemiological investigations, are needed to track changes in the 2019-nCOV epidemic. Understanding these epidemic dynamics in real time is increasingly important for public health in terms of guiding prevention efforts.

References


6. Lau, S. K. et al. Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related Rhinolophus bat coronavirus in China reveal bats...


This article is protected by copyright. All rights reserved.
Acknowledgements

This work was supported by a grant from National Natural Science Foundation of China (No. 31470268) to Prof. Yi Li. This study was sponsored by the K.C. Wong Magna Fund in Ningbo University. We acknowledge the authors and originating and submitting laboratories of the sequences as follows:

EPI_ISL_402119, EPI_ISL_402120, EPI_ISL_402121:

Originating and submitting lab - National Institute for Viral Disease Control and Prevention, China CDC.

Authors - Wenjie Tan, Xiang Zhao, Wenling Wang, Xuejun Ma, Yongzhong Jiang, Roujian Lu, Ji Wang, Weimin Zhou, Peihua Niu, Peipei Liu, Faxian Zhan, Weifeng Shi, Baoying Huang, Jun Liu, Li Zhao, Yao Meng, Xiaozhou He, Fei Ye, Na Zhu, Yang Li, Jing Chen, Wenbo Xu, George F. Gao, Guizhen Wu.

EPI_ISL_402123:

Originating and submitting lab - Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College.

Authors - Lili Ren, Jianwei Wang, Qi Jin, Zichun Xiang, Yongjun Li, Zhiqiang Wu, Chao Wu, Yiwei Liu.

This article is protected by copyright. All rights reserved.
EPI_ISL_402124, EPI_ISL_402127, EPI_ISL_402128, EPI_ISL_402129, EPI_ISL_402130:

*Originating lab* - Wuhan Jinyintan Hospital.

*Submitting lab* - Wuhan Institute of Virology, Chinese Academy of Sciences.

*Authors* - Peng Zhou, Xing-Lou Yang, Ding-Yu Zhang, Lei Zhang, Yan Zhu, Hao-Rui Si, Zheng-Li Shi.

MN908947:

*Originating lab* – Unknown.

*Submitting lab* - Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China.


EPI_ISL_402132:

*Originating lab* - Wuhan Jinyintan Hospital.

*Submitting lab* - Hubei Provincial Center for Disease Control and Prevention.

This article is protected by copyright. All rights reserved.
Authors - Bin Fang, Xiang Li, Xiao Yu, Linlin Liu, Bo Yang, Faxian Zhan, Guojun Ye, Xixiang Huo, Junqiang Xu, Bo Yu, Kun Cai, Jing Li, Yongzhong Jiang.

EPI_ISL_403962, EPI_ISL_403963:

Originating lab - Bamrasnaradura Hospital.

Submitting lab - Department of Medical Sciences, Ministry of Public Health, Thailand, Thai Red Cross Emerging Infectious Diseases - Health Science Centre, and Department of Disease Control, Ministry of Public Health, Thailand.

Authors - Pilailuk, Okada; Siripaporn, Phuygun; Thanutsapa, Thanadachakul; Supaporn, Wacharapluesadee; Sittiporn, Parnmen; Warawan, Wongboot; Sunthareeya, Waicharoen; Rome, Buathong; Malinee, Chittaganpitch; Nanthawan, Mekha.

Author contributions

X.L. conceived and designed the study, analyzed the data, and drafted the manuscript.

X.L., J.Z., X.W., and Y.L. interpreted the data and provided critical comments. All authors reviewed and approved the final manuscript.

Competing financial interests

The authors declare no competing interests.

This article is protected by copyright. All rights reserved.
Figures

Figure 1. Likelihood-mapping analyses of 12 genome sequences of 2019-nCOV.
Likelihoods of three tree topologies for each possible quartet (or for a random sample of quartets) are denoted by a data point in an equilateral triangle. Distribution of points in the seven areas of the triangle reflects tree-likeness of the data. Specifically, three corners represent fully resolved tree topologies; center represents an unresolved (star) phylogeny; and sides represent support for conflicting tree topologies.
Figure 2. **Maximum-likelihood phylogeny of 12 genome sequences of 2019-nCOV.**

Points are color-coded by city of origin. Tree is midpoint rooted.

Figure 3. **Regression of root-to-tip genetic distance against year of sampling for 2019-nCOV.** Points are color-coded by city of origin. Gray indicates linear regression line.
Figure 4. Maximum-clade-credibility tree estimated from 12 genome sequences of 2019-nCOV. Number near node represents most probable geographic location of descendent branches. Nodes are color-coded by city of origin.