Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2) using the whole genomic data

Preprint · February 2020

DOI: 10.12074/202002.00033

CITATIONS
9

READS
184,861

4 authors, including:

Wen-Bin Yu
Xishuangbanna Tropical Botanical Garden
130 PUBLICATIONS **2,133** CITATIONS

SEE PROFILE

Guang-da Tang
South China Agricultural University
43 PUBLICATIONS **499** CITATIONS

SEE PROFILE

R. T. Corlett
Xishuangbanna Tropical Botanical Garden
331 PUBLICATIONS **17,333** CITATIONS

SEE PROFILE

ZOOLOGICAL RESEARCH

Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data

Wen-Bin Yu[1,2,*], Guang-Da Tang[3,4], Li Zhang[5], Richard T. Corlett[1,2]

[1] *Center for Integrative Conservation*, *Xishuangbanna Tropical Botanical Garden*, *Chinese Academy of Sciences*, *Mengla*, *Yunnan* 666303, *China*

[2] *Center of Conservation Biology*, *Core Botanical Gardens*, *Chinese Academy of Sciences*, *Mengla*, *Yunnan* 666303, *China*

[3] *Henry Fok College of Biology and Agriculture*, *Shaoguan University*, *Shaoguan*, *Guangdong* 512005, *China*

[4] *College of Forestry and Landscape Architecture*, *South China Agricultural University*, *Guangzhou*, *Guangdong* 510642, *China*

[5] *Chinese Institute for Brain Research*, *Beijing* 102206, *China*

ABSTRACT

The outbreak of COVID-19 started in mid-December 2019 in Wuhan, China. Up to 29 February 2020, SARS-CoV-2 (HCoV-19 / 2019-nCoV) had infected more than 85 000 people in the world. In this study, we used 93 complete genomes of SARS-CoV-2 from the GISAID EpiFlu^{TM} database to investigate the evolution and human-to-human transmissions of SARS-CoV-2 in the first two months of the outbreak. We constructed haplotypes of the SARS-CoV-2 genomes, performed phylogenomic analyses and estimated the potential population size changes of the virus. The date of population expansion was calculated based on the expansion parameter tau (τ) using the formula $t=\tau/2u$. A total of 120 substitution sites with 119 codons, including 79 non-synonymous and 40 synonymous substitutions, were found in eight coding-regions in the SARS-CoV-2 genomes. Forty non-synonymous substitutions are potentially associated with virus adaptation. No combinations were detected. The 58 haplotypes (31 found in samples from China and 31 from outside China) were identified in 93 viral genomes under study and could be classified into five groups. By applying the reported bat coronavirus genome (bat-RaTG13-CoV) as the outgroup, we found that haplotypes H13 and H38 might be considered as ancestral haplotypes, and later H1 was derived from the intermediate haplotype H3. The population size of the SARS-CoV-2 was estimated to have undergone a recent expansion on 06 January 2020, and an early expansion on 08 December 2019. Furthermore, phyloepidemiologic approaches have recovered specific directions of human-to-human transmissions and the potential sources for international infected cases.

Keywords: COVID-19; HCoV-19; SARS-CoV-2; Novel pneumonia outbreak; Human-to-human transmission; Phyloepidemiology

Received: 01 March 2020; Accepted: 27 April 2020; Online: 30 April 2020

Foundation items: This study was supported by grants from Ten Thousand Talents Program of Yunnan for Top-notch Young Talents, and the open research project of "Cross-Cooperative Team" of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences

*Corresponding author, E-mail: yuwenbin@xtbg.ac.cn

DOI: 10.24272/j.issn.2095-8137.2020.022

INTRODUCTION

Betacoronaviruses are characterized by enveloped, positive-sense, single-stranded RNA, and hosted in animals, particularly mammals (Cui et al., 2019). Before December 2019, four species/strains of Betacoronavirus, HKU1, MERS-CoV, OC43, and SARS-CoV, had been reported to cause severe human diseases (Cui et al., 2019). The fifth species/strain, a novel betacoronavirus SARS-CoV-2 / HCoV-19 / 2019-nCoV (Gorbalenya et al., 2020; Jiang et al., 2020) causing human pneumonia (i.e., COVID-19), was first reported in Wuhan, Hubei, Central China (Wu et al., 2020a; Zhou et al., 2020; Zhu et al., 2020). Up to 29 February 2020, SARS-CoV-2 had infected more than 85 000 people in all provinces/regions of China, and another 59 countries/regions across Africa, Asia, Europe, North America, Oceania, and South America (Wikipedia, 2020). Because SARS-CoV-2 can transmit from human to human (Li et al., 2020), the massive exodus of people before the Chinese Spring Festival boosted the infection frequencies, as predicted (Wu et al., 2020b). Daily confirmed infection cases were more than 2 000 between 30 January and 16 February, 2020, and the highest was more than 15 100 (Wikipedia, 2020), almost twice the total number for SARS-CoV (Chan-Yeung & Xu, 2003).

As a member of subgenus *Sarbecovirus*, SARS-CoV-2 has been suggested to be of bat origin (Lu et al., 2020b; Zhou et al., 2020), and may have been transmitted to humans through non-bat intermediate mammals (e.g., pangolins (Cyranoski, 2020; Lam et al., 2020; Wong et al., 2020; Xiao et al., 2020; Zhang et al., 2020b)). Medical information for the first 41 infected patients in Wuhan showed that 27 patients were linked to the Huanan Seafood Wholesale Market (abbreviated as Huanan market in the text below) (Huang et al., 2020; Li et al., 2020), which sold living wild mammals. This suggests a high possibility that SARS-CoV-2 originated in the market, then the infected people transmitted it to other people outside of the market. However, this conclusion has been challenged because the first identified infected person and 12 others had no link to the Huanan Market. Some researchers have therefore argued that the Huanan Market was not the original and/or only source of SARS-CoV-2 transmission to humans (Cohen, 2020). The market was closed on 01 January 2020, making it very difficult to identify the intermediate animal vectors of SARS-CoV-2. In the absence of information on potential intermediary reservoirs, the origin and transmission pattern of SARS-CoV-2 are still unresolved (Wong et al., 2020).

Since the outbreak of COVID-19 was first identified in Wuhan in mid-December 2019, the first infected individuals identified in other provinces and regions of China, and other countries, during January 2020, have been assumed to have been infected in Wuhan or through contact with people from Wuhan (Chan et al., 2020; Holshue et al., 2020; Phan et al., 2020; Rothe et al., 2020). In this study, we used 93 genomes of SARS-CoV-2 from the GISAID EpiFlu^{TM} database (Shu & McCauley, 2017) (access date 12 February 2020) to decode the evolution and transmissions of SARS-CoV-2 in the first two months of its spread. Our aims were to: (1) characterize genomic variations of SARS-CoV-2; (2) infer the evolutionary relationships of the worldwide samples; and (3) deduce the transmission history of SARS-CoV-2 within Wuhan and out of Wuhan to the world.

MATERIALS AND METHODS

To decode the evolutionary history of SARS-CoV-2, we retrieved 96 complete genomes from GISAID (Supplementary Table S1, access by 12 February 2020) (Shu & McCauley, 2017). The genome EPI_ISL_402131 (bat-RaTG13-CoV, hereafter) from GISAID was also included as the outgroup, because it is the closest sister betacoronavirus available to SARS-CoV-2 (Zhou et al., 2020). The 97 genome sequences were aligned using MAFFT (Katoh & Standley, 2013), then the alignment was manually checked using Geneious (Biomatters, New Zealand). In the alignment, we found that EPI_ISL_404253 contains six ambiguous sites at variable positions and EPI_ISL_407079 and EPI_ISL_408978 have 175 “N” and 1476 “N” bases, respectively, so these three genomes were excluded in this study. In addition, four genomes (EPI_ISL_407071, EPI_ISL_407894, EPI_ISL_407896, and EPI_ISL_409067) have their own private ambiguous sites, which were conservatively replaced by the common nucleotide at that position in the alignment; Notably, EPI_ISL_406592 (H15) and EPI_ISL_406595 (H17) had excessive amounts of private variable sites, which were possibly affected by sequencing errors. In the alignment, the 5' untranslated region (UTR) and 3' UTR regions contain missing and ambiguous sites, so these regions were excluded in the following analyses.

The alignment was then imported into DnaSP (Rozas et al., 2017) for haplotype analyses. Population size changes were estimated based on a constant population size hypothesis using DnaSP, in combination with neutrality tests (Tajima’s D and Fu’s F_s). We also used Arlequin (Excoffier & Lischer, 2010) to test the sudden population expansion hypothesis and to calculate the expansion parameter tau (τ), since the sudden population expansion was not rejected. We used the formula $t=\tau/2u$ (Rogers & Harpending, 1992) to estimate the time since expansion (in days). In the formula, u is the cumulative substitution rate per year for the genome sequence, so we used the formula $u=\mu k$ to calculate it, where μ is the substitution rate per site per year, and k is the genome sequence length (29 358 bp for the coding sequence (CDS) matrix). The substitution rate was set as 0.92×10^{-3} (95% CI, 0.33×10^{-3}–1.46×10^{-3}) substitution/site/year based on the most recent estimation for SARS-CoV-2 (Rambaut, 2020). To adjust the time, we used a mean value of the expansion time calculated from the three substitution rates, i.e., 0.33×10^{-3}, 0.92×10^{-3}, and 1.46×10^{-3} substitution/site/year. In addition, the expansion date was estimated based on the sampling date from hospitalized patients. The estimated date should be later than the “real” date of massive human-to-human transmission events.

Phylogenetic networks of the haplotype coding region matrix and 120 substitution sites of SARS-CoV-2 (Supplementary Datasets) were inferred using SplitsTree (Huson & Bryant, 2006). A median-joining network of haplotypes was generated by the NETWORK program (Bandelt et al., 1999, 2020) with the reported bat coronavirus (bat-RaTG13-CoV, (Zhou et al., 2020)) as the outgroup. Transversions were arbitrarily weighted three times as high as transitions. Hypervariable sites (if number of mutations ≥5) were weighted as 1, and the other sites were weighted as 10. Genetically, SARS-CoV-2 and bat-RaTG13-CoV, as well as SARS-CoV, have been proposed as the same species (Gorbalenya et al., 2020), and genome sequence identity between bat-RaTG13-CoV and SARS-CoV-2 was 96.2%. We carefully used three datasets (i.e., four core substitution sites, 120 substitution sites, and 1 235 substitution sites, Supplementary Datasets) to evaluate the relationship between bat-RaTG13-CoV and four associated/central haplotypes of SARS-CoV-2 (H1, H3, H13, and H38). Phylogenomic analyses of haplotypes were performed using IQ-TREE (Minh et al., 2020). We conducted likelihood mapping and SH-like approximate likelihood ratio tests to assess the phylogenetic information and branch supports, respectively.

RESULTS AND DISCUSSION

Genomic variations of SARS-CoV-2

Genome size of SARS-CoV-2 varied from 29 782 bp to 29 903 bp. The aligned matrix was 29 910 bp in length, including 140 variable sites. The CDS regions contained 120 substitution sites (Supplementary Figure S1), which were classified as 58 haplotypes (Supplementary Table S2). Nucleotide diversity (P_i) was $0.15\times10^{-3}\pm0.02\times10^{-3}$ (standard deviation, *SD*, hereafter). Haplotype diversity (H_d) was 0.953±0.016 (*SD*) and variance of H_d was 0.26×10^{-3}.

There were 120 substitution sites found in eight coding sequence (CDS) regions of SARS-CoV-2 (Figure 1, Supplementary Table S2), including 79 transitions (65.83%) and 41 transversions (34.17%). A chi-squared test showed that the distribution of substitution sites across CDS regions in the genome was even (χ^2=1.958, d_f=9, P=0.99). Substitution sites occurring at the 1st to 3rd frame positions were 27 (25.55%), 44 (40.0%), and 49 (44.55%), respectively. The 120 substitution sites were associated with 119 codons, including 79 non-synonymous (65.83%) and 40 synonymous (33.61%) substitutions. Forty non-synonymous substitutions (50.63%) changed the biochemical properties of the amino acid (AA), and are therefore potentially associated with virus adaptation. The current samplings showed that the H1 haplotype has been found in 19 patients, but most haplotypes were just sequenced once, suggesting that the haplotype H1 was rapidly circulated at an early stage of human-to-human transmissions (Figure 2, Supplementary Table S1).

In comparisons with published genomes of SARS-CoV (Luk et al., 2019) and MERS-CoV (Cotten et al., 2013), genomic variations of SARS-CoV-2 are still low, without evident

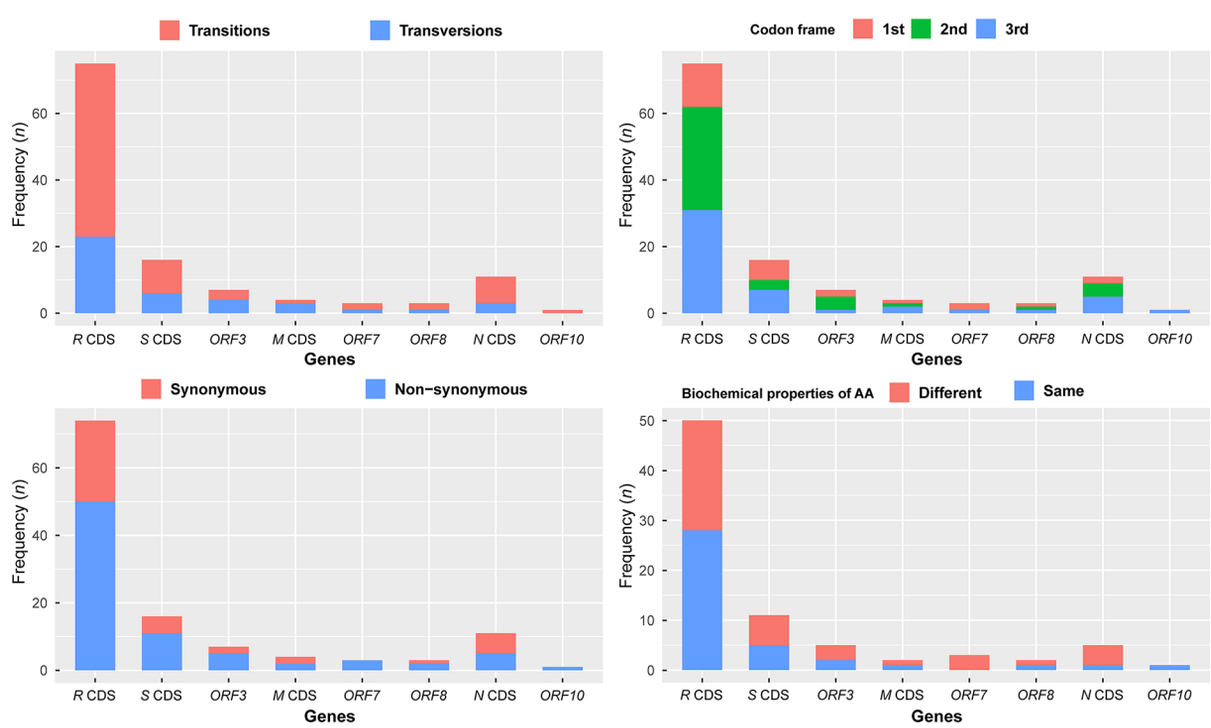


Figure 1 Summary information for 120 substitution sites crossing eight coding sequence regions in the aligned SARS-CoV-2 genomic sequences

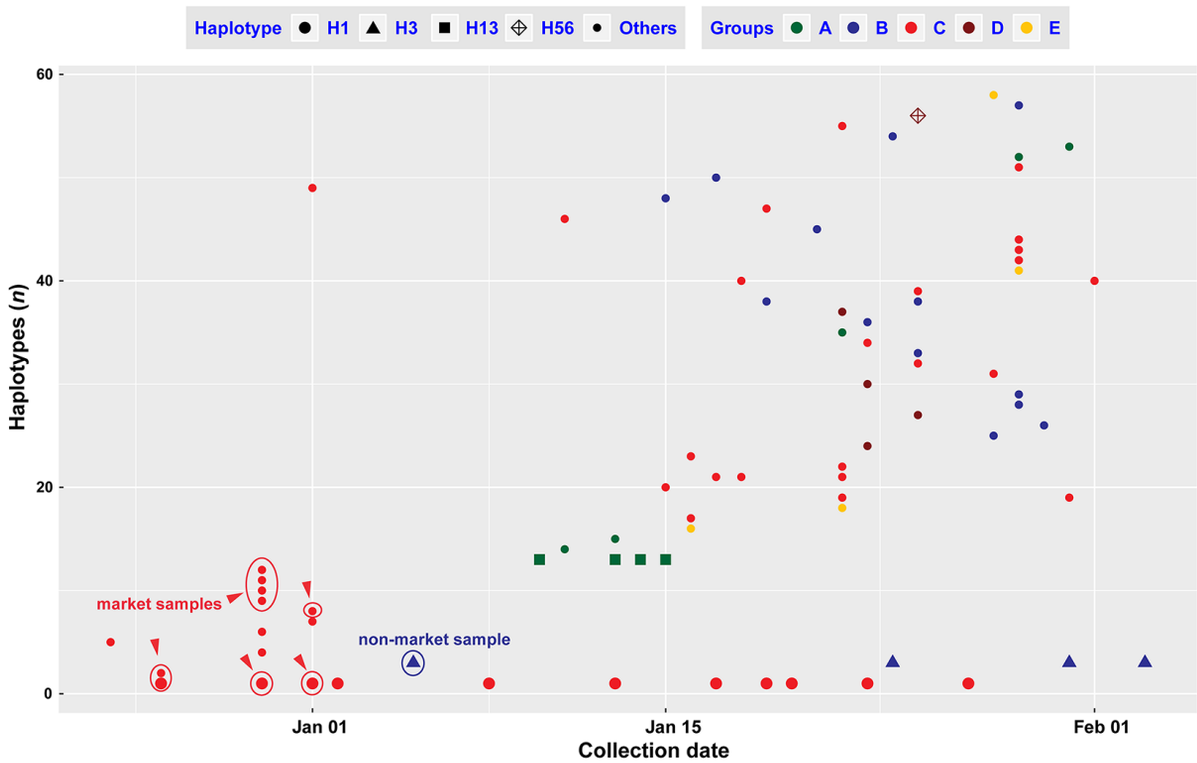


Figure 2 Genomic haplotypes of SARS-CoV-2 changes between the collection dates of samples

The confirmed samples from the Huanan Seafood Wholesale Market are indicated using red circles, and a confirmed sample with no link to the market is indicated using a blue circle.

recombination sites/regions (R_m=2, P=1.0) at this time. According to the collection dates of the sequenced samples, haplotypes H1 and H3 were found in two samples at intervals of more than 30 days, and multiple samples over 20 days (Figure 2, Supplementary Table S1). Although the incubation period can be over 24 days, there was only one case of this out of 1 099 observations (Guan et al., 2020). Estimation of the substitution rates using 90 genomes of SARS-CoV-2 (Rambaut, 2020) showed that the rate for SARS-CoV-2 was close to or lower than the rates for MERS-CoV (Cotten et al., 2014; Dudas et al., 2018) and SARS-CoV (Zhao et al., 2004). Due to the mild symptoms and low mortality (Yang et al., 2020; Zhang et al., 2020a), the immune systems of the infected humans may provide a suitable environment for propagation of SARS-CoV-2 (Andersen et al., 2020). SARS-CoV-2 is highly infectious (Yang et al., 2020) and is able to infect humans not only through the mucous membranes of the nose and mouth, but also use the mucous membranes in the eyes (Lu et al., 2020a), which may boost regional circulation and large-scale spread. Some large mutations may have occurred in Wuhan or other regions, but the strict quarantine policy over China since 23 January 2020 may have reduced the circulation and spreading of some mutants.

Of the 93 genomes of SARS-CoV-2, 39 (41.93%) were from infected patients in 11 countries outside China and encoded 31 haplotypes (H_d=0.987±0.009 (*SD*), P_i=0.16×10^{-3}±0.01×10^{-3}), with 27 nationally/regionally private haplotypes. The 54 genomes (58.07%) from China also encoded 31 haplotypes (H_d=0.906±0.001 (*SD*), P_i=0.14×10^{-3}±0.03×10^{-3}). A proportion *Z*-test showed significant differences in haplotype diversity of samples between China and other countries (χ^2=4.024, d_f=1, P<0.05). The high haplotype diversity found in samples from other countries may be because the sampling dates were mostly after 22 January 2020, while those in China were before this date (Supplementary Table S1 and Figure S2). In addition, the low level of radiation exposure on long-distance international flights (Bottollier-Depois et al., 2000) may have accelerated mutation rates of SARS-CoV-2 (Shibai et al., 2017).

Population size expansion of SARS-CoV-2

We used a variety of parameters to estimate the population dynamics of SARS-CoV-2. Constant population size of SARS-CoV-2 was rejected (Ramos-Onsins and Rozas's R^2=0.025, P<0.001; Raggedness r=0.011, P<0.05) using DnaSP (Rozas et al., 2017) (also see Supplementary Figure S3), while both Fu's test (F_s=–67.681.964, P<0.001) and Tajima's D test (D=–2.701, P<0.001) indicated that the population size of SARS-CoV-2 was rapidly increasing. Mismatch distribution analysis using Arlequin (Excoffier & Lischer, 2010) strongly supported that the population of SARS-CoV-2 underwent sudden expansion (τ=2.887, Sum of Squared deviation,

SSD=0.541×10^{-3}, *P*=0.88, Harpending's Raggedness index, *R*=0.010, *P*=0.88). The calculated expansion was 28.72 days (95% Confident Interval: 12.29–54.36 days) ago. Of the 93 genomes, the latest one was sampled on 03 February 2020, so the estimated expansion date was on 06 January 2020 (95% CI: 11 December 2019–22 January 2020), which may be related to the New Year holiday. Before 06 January 2020, 129 patients were identified as SARS-CoV-2 infected through field investigations (Li et al., 2020). Of 22 genomes (17.05% of 129 patients) sequenced before 06 January 2020 in Wuhan, China, 13 haplotypes (22.41% of 58 haplotypes) were recovered, which were H1 and its derived descendant haplotypes, and H3 (Figures 2 and 3A). Coincidentally, the China CDC (Chinese Center for Disease Control and Prevention) started to activate a Level-2 emergency response on 06 January 2020 (Li et al., 2020). The China CDC's emergency response greatly reduced public activities and travel, and might have reduced the local circulation and large-scale spread in the following weeks of January.

Furthermore, mismatch distribution analysis of the 22 genomes before 06 January 2020 also showed a sudden population expansion of SARS-CoV-2 at an earlier stage of transmission (τ=2.818, *SSD*=0.010, *P*=0.41, *R*=0.046, *P*=0.57, Tajima's *D*=–2.241, *P*<0.001; Fu's F_s=–7.834, *P*<0.001). This earlier population expansion time was estimated at 28.38 days (95% CI: 12.00–54.36 days) before 05 January 2020, which was the latest sampling date of the 22 genomes. This earlier expansion date was thus estimated to have occurred on 08 December 2019 (95% CI: 13 November 2019–26 December 2019), when there was only one infected patient officially reported (Huang et al., 2020; Li et al., 2020). This suggests that SARS-CoV-2 might have already circulated widely among humans in Wuhan before December 2019, probably beginning in mid to late November (Rambaut, 2020).

Evolutionary relationships of SARS-CoV-2 haplotypes

Phylogenetic networks showed that the 58 haplotypes were clustered into two main clades (Figure 4). Clade I included 19 haplotypes and Clade II included 39 haplotypes. The outgroup bat-RaTG13-CoV was connected to Clade I, supposed to be an ancestral clade for Clade II. The long branches of H15 and H17 correspond to an excessive amount of mutations, which are possibly affected by sequencing errors, but this is still to be determined. Three different datasets were used to infer evolutionary networks, which consistently supported H13 and H38 as the potentially ancestral haplotypes, i.e., the outgroup bat-RaTG13-CoV could connect to both H13 and H38, or H38 alone, or through a medium vector mv1 (an intermediate host or the first infected humans) connected to both H13 and H38 by single mutations at positions 18067 (S, synonymous substitution) and/or 29102 (S), referring to the numbering of the alignment length 29 910 bp (Figure 5). Five main groups can be recognized in the network using the dataset of 120 substitution sites (Figure 3A). The H1, H3, and H13 were three core haplotypes, so that Groups A–C were recognized using them as the central (i.e., ancestral super-spreader) haplotypes. Groups D and E were recognized based on two new super-spreader haplotypes, H56 and a medium vector mv2, which was a hypothesized (often ancestral) haplotype not sampled in the current samples. These two groups can be also treated as subgroups of Group C. Moreover, the SH-like approximate likelihood ratio test further enhanced the

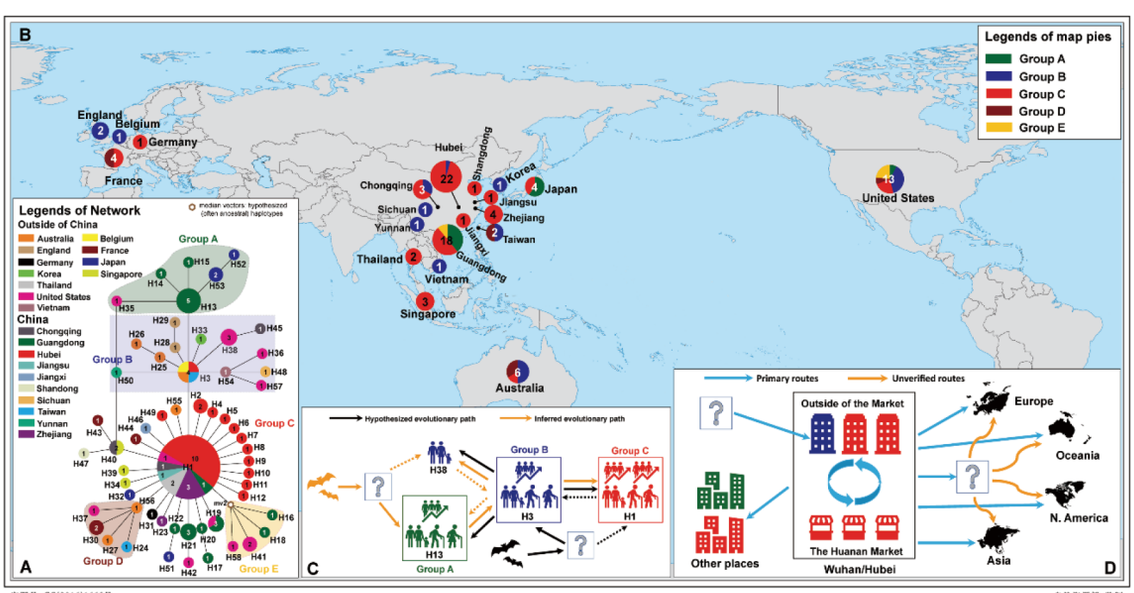


Figure 3 Evolutionary relationship and geographical distribution of 58 haplotypes of SARS-CoV-2 (A, B)

Proposed evolutionary paths (C) of haplotypes and possible transmission and spreading routes (D) are also inferred based on evolutionary analyses and epidemiologic research. Sample sizes of haplotypes and regions are annotated in the circles.

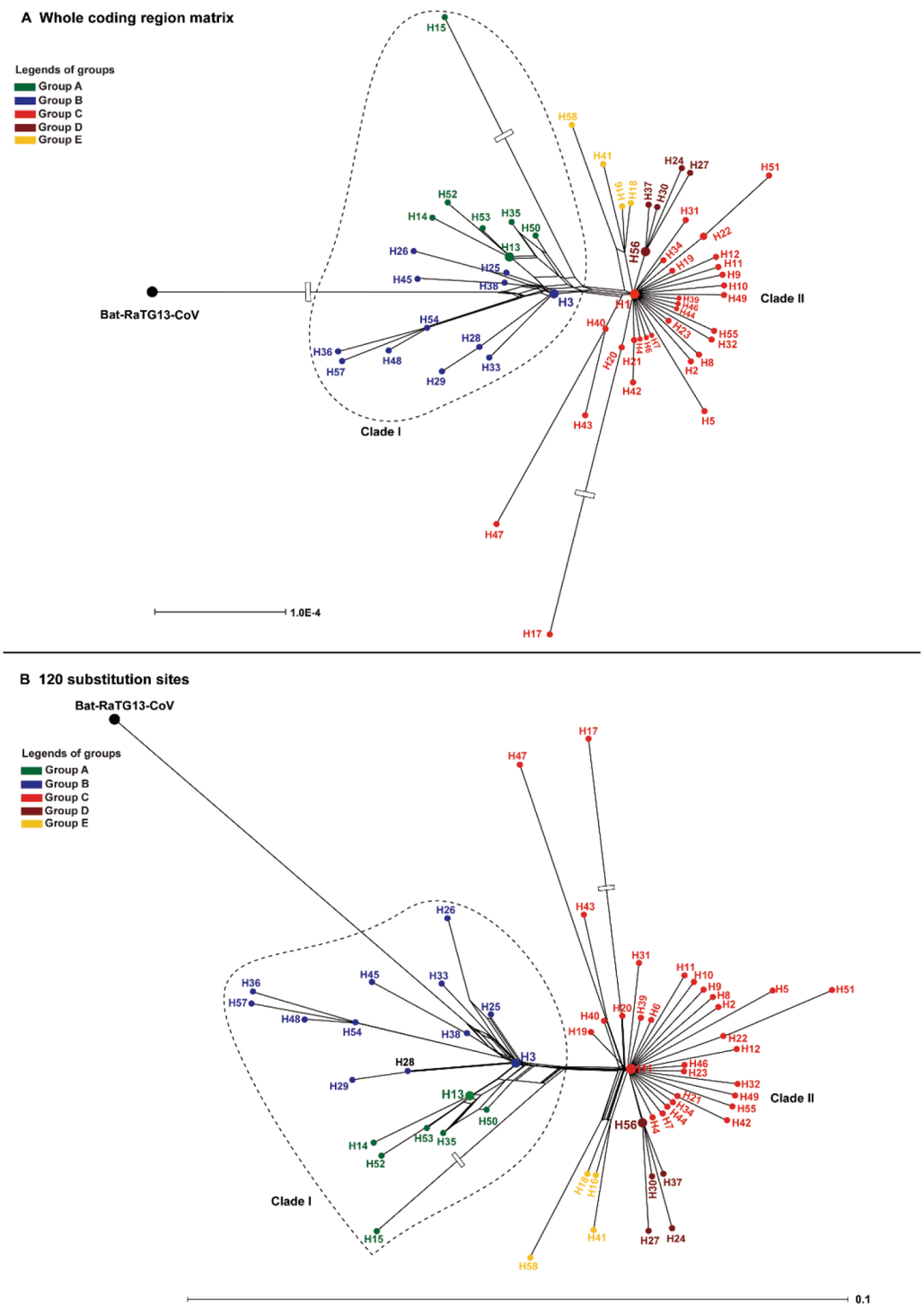
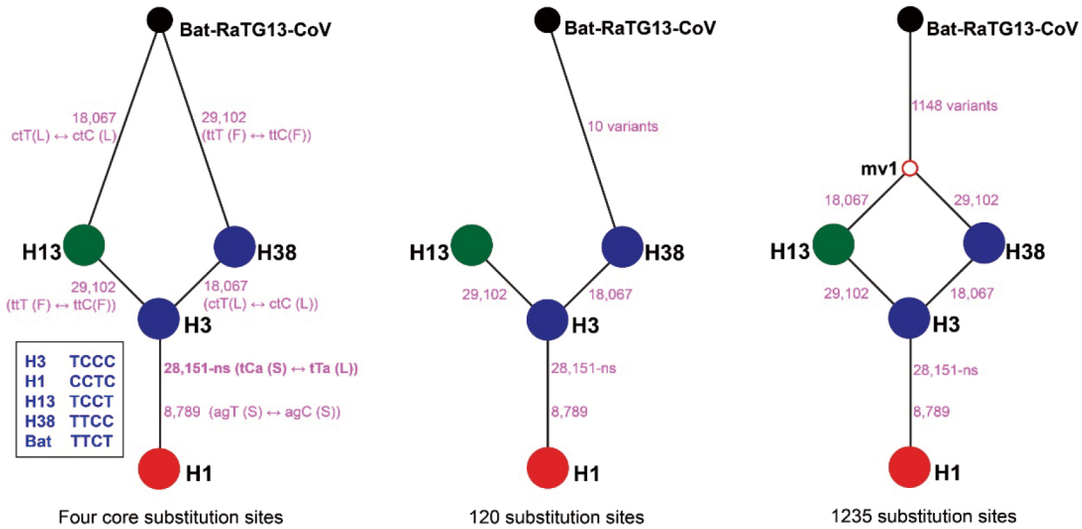


Figure 4 Phylogenetic networks of 58 haplotypes of SARS-CoV-2 with the outgroup bat-RaTG13-CoV using the whole coding region matrix (A) and 120 substitution sites (B), referring to all variable sites of coding regions in SARS-CoV-2 haplotypes

The bottom scale bars represent the number of substitutions per site. The open box on branches of bat-RaTG13-CoV (A), H15 (A, B), and H17 (A, B) indicating a long branch was clipped.

phylogenetic relationship retrieved from 58 haplotypes, that either H13 or H38 (with H45) appeared in the basal lineages (Supplementary Figure S4), although it was difficult to distinguish which one, H13 or H38, originated earlier



The marked mutation sites refer to synonymous variants, unless otherwise with a suffix "ns" to indicate nonsynonymous variant.

Figure 5 The inferred relationships between the outgroup bat-RaTG13-CoV and four associated/central node haplotypes (H1, H3, H13, and H38) of SARS-CoV-2 using three datasets

The dataset of four core substitution sites is the four variable sites shared among bat-RaTG13-CoV and the four central node haplotypes. The dataset of 120 substitution sites refers to all variable sites of coding regions in SARS-CoV-2 haplotypes. The dataset of 1235 substitution sites refers to all variable sites of coding regions among bat-RaTG13-CoV and SARS-CoV-2 haplotypes.

(Supplementary Figure S5).

In the network, four satellite haplotypes and H35 connected to H13 (Group A), and nine satellite haplotypes and H38+H45 and H50 connected to H3 (Group B). The connections between the H3 and H1 are two mutations at positions 8789 (S) and 28151 (NS, non-synonymous substitution) (Figure 5), the latter mutation changed both residues and the biochemical properties of the AA. This biochemical change may be associated with the infectivity of SARS-CoV-2. The H1 haplotype, the most abundant, included 19 samples, while 26 satellite haplotypes and H40+(H43 and H47) haplotypes are directly derived from H1 (Group C). Moreover, five haplotypes of Group D and four haplotypes of Group E were also derived from H1.

The Huanan Seafood Wholesale Market boosted human-to-human transmission at an early stage

Phylogenetic networks showed that bat-RaTG13-CoV was nested with Group B in Clade I, and Clade II tends to be derived from Clade I, i.e., H1 and its descendant haplotypes were new mutants from an ancestral haplotype in Clade I. The rooted network suggested two potential evolutionary paths of available haplotypes that can be from H13 through H3 to H1 and H38, or from H38 through H3 to H1 and H13 (Figure 3C). Both scenarios suggested that H3 might be the ancestral haplotype of H1. H13 was only recovered from five Shenzhen (Guangdong Province) samples, including patient 2 of the familial cluster (Chan et al., 2020). Two derived haplotypes were also only found in Shenzhen of Guangdong Province (H14 from the grandson of patient 2), and the other three haplotypes were found in three samples from Japan and one sample from Arizona in the United States (Figure 3). According to an epidemiological study, the Shenzhen family could have been infected during their visit to Wuhan (Chan et al., 2020). This suggests that H13 might have originated from Wuhan. Genetically, haplotypes of Group A have links to only Wuhan haplotype H3 (only EPI_ISL_406801). It is possible that H13 was newly derived from H3 (Figure 3C) and did not spread in Wuhan, or that three repatriated Japanese might be infected by an unknown source of H13 in Wuhan, China or somewhere else (The Asahi Shimbun, 2020), or that no samples have been sequenced yet. H38 has three genomes from the same patient (Supplementary Table S1), who was the first identified infected patient in the United States (Holshue et al., 2020). This patient might have been infected while visiting his family in Wuhan, China, or was infected in some other place. The original source of H38 can be explained as that of H13, which can be also derived from H3 (Figure 3C), and the derived H45 was from a Chongqing patient who was reported as working in Wuhan and had no link to the Huanan Market.

The H3 haplotype has only one sample from Wuhan, which was not linked to the Huanan Market (Lu et al., 2020b), and the other samples in this group were from outside of Wuhan (Figure 3A). Noteworthily, all the samples from the Market belonged to H1 or its derived haplotypes (H2, H8-H12, see Figure 2 and Supplementary Table S1), indicating that there were circulated infections within the market in the short term. Other researchers have argued that the source of the coronavirus in the Market should be imported from elsewhere, or at least it should be not the single source of SARS-CoV-2

(Cohen, 2020). In this study, evolutionary relationships indicated that H1 and its descendant haplotypes from the Market should be derived from H3 (Figures 3, 4). H3 mutated to the H1 by two substitutions, and none of the currently available Market samples encoded H3, suggesting that H3 might have originated and spread outside of the Market before an early stage of population expansion. The non-synonymous mutation from H3 to H1 might have enhanced the infectiousness of SARS-CoV-2, and a functional characterization should be performed to confirm this speculation. It is possible that SARS-CoV-2 in the Market had been transmitted from other places (Figure 3D), or at least, that the Market did not host the original source of SARS-CoV-2 (Cohen, 2020). As the first identified infected patients had no link to the Market (Huang et al., 2020), it is possible that infected humans transmitted the H1 haplotype of SARS-CoV-2 to workers or sellers in the market, after which it rapidly circulated there due to its special surroundings. The crowded market boosted SARS-CoV-2 transmissions to buyers and spread it to the whole city in early December 2019, corresponding to the estimated population expansion time. Due to insufficient sampling from Wuhan in the currently available samples, it is not clear whether H3 never appeared in the Market, or H1 was quickly derived from H3 to adapt in the Market.

Regional and worldwide circulation and spread

Of the 54 genomes from patients in China, Chongqing (3 samples), Guangdong (18), Hubei (22), Taiwan (2), and Zhejiang (4) have more than two samples, and the other five provinces have one sample. Hubei (Wuhan) samples dated from 24 December 2019 to 05 January 2020 encoded 13 haplotypes, belonging to Groups C (H1 and 11 satellite haplotypes) and B (only H3). These relationships indicated a rapid transmission and circulation of SARS-CoV-2 in Wuhan at an early stage of human-to-human transmissions. H1 (no satellite haplotypes) and H3 are the ancestors of haplotypes outside of Wuhan/Hubei because most of early confirmed patients might have history in Wuhan or Hubei. Eighteen Guangdong samples, collected from 10–23 January 2020, encoded 15 haplotypes, belonging to Groups A, C, and E, showing that there were multiple sources imported into Guangdong. Three haplotypes (H14, H15, and H17) may have evolved locally, indicating that human-to-human transmissions happened when SARS-CoV-2 initially spread to Shenzhen in Guangdong Province (Chan et al., 2020). Two samples from Taiwan Province, China, encoded H3 and H24 in Groups B and D, respectively, and three samples from Chongqing encoded H1, H40, and H45 in Groups B and C, respectively. There were two sources imported into these two provinces. Four Zhejiang samples encoded H1 and H24 in Group C, which might be only imported from the source of the H1 haplotype.

The samples outside China encoded 31 haplotypes belonging to Groups A–E. Of these, 27 haplotypes are private by regional samplings, only two samples from Thailand were the H1 haplotype, one each from Australia and Belgium were the H3 haplotype, one from the United States was the H19 haplotype, and one from Singapore was the H40 haplotype. Twelve samples, encoding 10 haplotypes, were from patients in five countries in Asia. Six haplotypes linked to H1, and two each linked to H3 and H1, respectively, indicating the 12 patients were infected by different sources. Human-to-human transmissions may have happened from patients with H53 to H52 haplotypes in Tokyo, Japan. Five Australian samples, encoding six haplotypes in Groups B, C, and D, were from patients of three states. Patients with H3, H25 and H26, and with H55 were in Groups B and C, respectively, and human-to-human transmission might have happened from the patients with H25 to H26, who were in a same tour group in Queensland (AAP reporters, 2020). The connection between the patients with H56 and H27 is not clear. One possibility is that there was an intermediary spreader with H56, who also transmitted SARS-CoV-2 to other patients in France, the United States, and Taiwan Province of China. Eight European samples, encoding seven haplotypes, were from patients in four countries. The patients in England were reported as a household transmission from H28 to H29 (Lillie et al., 2020). The patients in France may have been infected by three different sources, i.e., H44 was linked to H1, H43 might link to H40 (in Chongqing, Singapore or somewhere else), and H30 might link to an intermediary spreader with H56. Of the 13 genomes from the United States, three were from the same patient in Washington encoding the same haplotype H38, while the other ten samples encoded eight haplotypes, covering all five groups (Figure 3A, B), so the sources of infections are complicated. There is no evidence of human-to-human transmission in the United States from these 11 cases. To clarify the exact origins of these haplotypes outside China, we need more epidemiological investigative efforts and more SARS-CoV-2 genomic data from patients at the early stage of transmissions.

Phylogenetic approaches provide insights into the epidemiology of SARS-CoV-2

Epidemiological study of SARS-CoV-2 using traditional approaches is very difficult, because it was not identified as a new coronavirus until 29 December, and some infected people with mild symptoms or without symptoms (Heymann & Shindo, 2020; Rothe et al., 2020; Wu & McGoogan, 2020) may have been overlooked in late November and early December. Evolutionary analyses suggested that the source of the H1 haplotype in the Huanan Market was imported from elsewhere, as has been suggested by other researchers (Cohen, 2020). The rooted network suggested that H13 and H38 should be ancestral haplotypes that connected to the outgroup bat-RaTG13-CoV through a hypothesized intermediate haplotype (Figure 5). The most common ancestral haplotype was missed because the currently available samples do not include the first identified infected patient and other patients from early December, and because of the relatively high mutation rate of the viral genome. If there

are any frozen samples from those patients, it would be worth doing genomic sequencing for phyloepidemiologic study to help to locate the birthplace of SARS-CoV-2. Meanwhile, we expect that the H13 and H38 haplotypes might be found in some samples from infected patients in Wuhan or in other places across the world if more samples are sequenced in future. This will be very helpful in the search for the original sources of SARS-CoV-2, because both H13 and H38 tend to be ancestral haplotypes.

The evolutionary network of haplotypes can be used to recover the directions of human-to-human transmissions at the local scale and spread at the larger scale. The central haplotype can be considered as the super-spreader haplotype, and the tip haplotype is the most recent descendant, similar to the definition and use of mtDNA haplogroups in tracing human demographic history (Yao et al., 2002). The transmission direction can be identified using the connection information of tips and branches. For example, the confirmed patients from the Huanan Market shared the common ancestral haplotype H1, indicating they might be infected from a common source, who may have been a super-spreader in the market. This approach has recovered potentially specific directions of human-to-human transmission in the Shenzhen family (H13 → H14), the Queensland tour group (H25 → H26), the England family (H28 → H29), and the Japanese (H53 → H52). It is possible that some infections could link to Wuhan or Hubei directly or indirectly, because the patients claimed connections to Wuhan or Hubei, but for some of them it is not clear exactly where they were infected. We suspect that there were super-spreaders mediating the spread of SARS-CoV-2 at the early stage of transmissions.

Our findings showed that SARS-CoV-2 has not had legitimate recombination. Thus, the haplotype-based phyloepidemiologic analyses provide a powerful way to understand the evolution of SARS-CoV-2 at the very early stage of transmission when reverse mutations and illegitimate recombination are rare. In our analysis, recombination is rejected but the outgroup bat-RaTG13-CoV is relatively highly diverged from SARS-CoV-2 haplotypes, which may affect the phyloepidemiologic analyses. Based on the estimated mutation rate of current SARS-CoV-2 viruses, the reverse mutations should be 6×10^{-3} ($0.92\times10^{-3}\times0.92\times10^{-3}$ per site per year×29 358 sites×2/12 year), with a neglectable influence on our result. But our observations leave one important question: why are ancestral haplotypes, like H13 and H38, less frequent than H1? It is highly possible that H1 acquired adaptive mutations, such as NS of site 28151, from H3 or H13 (and/or H38), evolved in an independent circulation after they jumped into intermediate hosts or directly transmitted to humans, which should be investigated in future studies if more early genome datasets are available. The exact original sources of H13 and H38 will stay as unsolved mysteries if the early stage samples were not preserved.

An early version of our manuscript was posted at ChinaXiv (DOI: 10.12074/202002.00033) on 19 February 2020. Since then, there have been many news stories stemming from our manuscript with a biased interpretation of the results. This is beyond our expectation. During the review of this manuscript, there were some reports of analyses of SARS-CoV-2 genomic variations based on a larger sample size (e.g., Forster et al., 2020; Tang et al., 2020), which showed a similar phylogenetic pattern as we present here. We expect more data-mining of the increasing number of SARS-CoV-2 genomes will provide updated insights into the origin and transmission of this virus.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

W.B.Y. conceived the research, analyzed the data, interpreted the results, and wrote the draft manuscript; W.B.Y. and G.D.T. collected data. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We are grateful to scientists and researchers for depositing whole genomic sequences of Novel Pneumonia Coronavirus (SARS-CoV-2 / HCoV-19 / 2019-nCoV) at the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu™; to GISAID database for allowing us to access the sequences for non-commercial scientific research; and to two reviewers for their valuable comments and suggestions. This study was supported by grants from Ten Thousand Talents Program of Yunnan for Top‑notch Young Talents, and the open research project of "Cross-Cooperative Team" of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences.

REFERENCES

AAP Reporters. 2020 (2020-01-30). Coronavirus outbreak: second case confirmed in Queensland. https://7news.com.au/lifestyle/health-wellbeing/qld-coronavirus-case-remains-in-isolation-c-671500.

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nature Medicine*, **26**(4): 450−452.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**(1): 37−48.

Bandelt HJ, Forster P, Röhl A. 2020 (2020-02-01). Free phylogenetic network software. https://www.fluxus-engineering.com/sharenet.htm.

Bottollier-Depois JF, Chau Q, Bouisset P, Kerlau G, Plawinski L, Lebaron-Jacobs L. 2000. Assessing exposure to cosmic radiation during long-haul flights. *Radiation Research*, **153**(5): 526−532.

Chan JFW, Yuan SF, Kok KH, To KKW, Chu H, Yang J, Xing FF, Liu JL, Yip CCY, Poon RWS, Tsoi HW, Lo SKF, Chan KH, Poon VKM, Chan WM, Ip JD, Cai JP, Cheng VCC, Chen H, Hui CKM, Yuen KY. 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*,

395(10223): 514−523.
Chan-Yeung M, Xu RH. 2003. SARS: epidemiology. *Respirology*, **8**(S1): S9−S14.
Cohen J. 2020. Wuhan seafood market may not be source of novel virus spreading globally. *Science*, doi: 10.1126/science.abb0611.
Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, Al-Tawfiq JA, Alhakeem RF, Madani H, AlRabiah FA, Hajjar SA, Al-Nassir WN, Albarrak A, Flemban H, Balkhy HH, Alsubaie S, Palser AL, Gall A, Bashford-Rogers R, Rambaut A, Zumla AI, Memish ZA. 2013. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *The Lancet*, **382**(9909): 1993−2002.
Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, Al Rabeeah AA, Alhakeem RF, Assiri A, Al-Tawfiq JA, Albarrak A, Barry M, Shibl A, Alrabiah FA, Hajjar S, Balkhy HH, Flemban H, Rambaut A, Kellam P, Memish ZA. 2014. Spread, circulation, and evolution of the middle east respiratory syndrome Coronavirus. *mBio*, **5**(1): e01062−13.
Cui J, Li F, Shi ZL. 2019. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, **17**(3): 181−192.
Cyranoski D. 2020. Did pangolins spread the China coronavirus to people?. *Nature*, doi: 10.1038/d41586-020-00364-2.
Dudas G, Carvalho LM, Rambaut A, Bedford T. 2018. MERS-CoV spillover at the camel-human interface. *eLife*, **7**: e31257.
Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**(3): 564−567.
Forster P, Forster L, Renfrew C, Forster M. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, doi: 10.1073/pnas.2004999117.
Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM, Neuman BW, Penzar D, Perlman S, Poon LLM, Samborskiy DV, Sidorov IA, Sola I, Ziebuhr J, Coronaviridae Study Group of the International Committee on Taxonomy of V. 2020. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, **5**(4): 536−544.
Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC, Du B, Li LJ, Zeng G, Yuen KY, Chen RC, Tang CL, Wang T, Chen PY, Xiang J, Li SY, Wang JL, Liang ZJ, Peng YX, Wei L, Liu Y, Hu YH, Peng P, Wang JM, Liu JY, Chen Z, Li G, Zheng ZJ, Qiu SQ, Luo J, Ye CJ, Zhu SY, Zhong NS. 2020. Clinical characteristics of coronavirus disease 2019 in China. *The New England Journal of Medicine*, doi: 10.1056/NEJMoa2002032.
Heymann DL, Shindo N. 2020. COVID-19: what is next for public health?. *The Lancet*, **395**(10224): 542−545.
Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson K, Wilkerson S, Tural A, Diaz G, Cohn A, Fox L, Patel A, Gerber SI, Kim L, Tong SX, Lu XY, Lindstrom S, Pallansch MA, Weldon WC, Biggs HM, Uyeki TM, Pillai SK. 2020. First case of 2019 novel coronavirus in the United States. *The New England Journal of Medicine*, **382**(10): 929−936.
Huang CL, Wang YM, Li XW, Ren LL, Zhao JP, Hu Y, Zhang L, Fan GH, Xu JY, Gu XY, Cheng ZS, Yu T, Xia JA, Wei Y, Wu WJ, Xie XL, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie JG, Wang GF, Jiang RM, Gao ZC, Jin Q, Wang JW, Cao B. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, **395**(10223): 497−506.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**(2): 254−267.
Jiang S, Shi Z, Shu Y, Song J, Gao GF, Tan W, Guo D. 2020. A distinct name is needed for the new coronavirus. *The Lancet*, **395**(10228): 949.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**(4): 772−780.
Lam TTY, Shum MHH, Zhu HC, Tong YG, Ni XB, Liao YS, Wei W, Cheung WYM, Li WJ, Li LF, Leung GM, Holmes EC, Hu YL, Guan Y. 2020. Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv*, doi: 10.1101/2020.02.13.945485.
Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *The New England Journal of Medicine*, **382**(13): 1199−1207.
Lillie PJ, Samson A, Li A, Adams K, Capstick R, Barlow GD, Easom N, Hamilton E, Moss PJ, Evans A, Ivan M, Phe Incident Team, Taha Y, Duncan CJA, Schmid ML, the Airborne Hcid Network. 2020. Novel coronavirus disease (Covid-19): The first two patients in the UK with person to person transmission. *Journal of Infection*, **80**(5): 578−606.
Lu CW, Liu XF, Jia ZF. 2020a. 2019-nCoV transmission through the ocular surface must not be ignored. *The Lancet*, **395**(10224): e39.
Lu RJ, Zhao X, Li J, Niu PH, Yang B, Wu HL, Wang WL, Song H, Huang BY, Zhu N, Bi YH, Ma XJ, Zhan FX, Wang L, Hu T, Zhou H, Hu ZH, Zhou WM, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan JY, Xie ZH, Ma JM, Liu WJ, Wang DY, Xu WB, Holmes EC, Gao GF, Wu GZ, Chen WJ, Shi WF, Tan WJ. 2020b. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, **395**(10224): 565−574.
Luk HKH, Li X, Fung J, Lau SKP, Woo PCY. 2019. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infection, Genetics and Evolution*, **71**: 21−30.
Minh BQ, Schmidt H, Chernomor O, Schrempf D, Woodhams M, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, doi: 10.1093/molbev/msaa015.
Phan LT, Nguyen TV, Luong QC, Nguyen TV, Nguyen HT, Le HQ, Nguyen TT, Cao TM, Pham QD. 2020. Importation and human-to-human transmission of a novel coronavirus in Vietnam. *The New England Journal of Medicine*, **382**(9): 872−874.
Rambaut A. 2020 (2020-02-12). Phylodynamic analysis|129 genomes |24 Feb 2020. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356.
Rogers AR, Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9**(3): 552−569.
Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, Zimmer T, Thiel V, Janke C, Guggemos W, Seilmaier M, Drosten C, Vollmar P, Zwirglmaier K, Zange S, Wölfel R, Hoelscher M. 2020.

Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *The New England Journal of Medicine*, **382**(10): 970−971.

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, **34**(12): 3299−3302.

Shibai A, Takahashi Y, Ishizawa Y, Motooka D, Nakamura S, Ying BW, Tsuru S. 2017. Mutation accumulation under UV radiation in *Escherichia coli*. *Scientific Reports*, **7**: 14531.

Shu YL, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance*, **22**(13): 30494.

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, doi: 10.1093/nsr/nwaa1036.

The Asahi Shimbun. 2020 (2020-02-02). Japan tightens immigration as 3 more infected by coronavirus. http://www.asahi.com/ajw/articles/AJ202002020013.html.

Wikipedia. 2020 (2020-02-29). 2019-20 coronavirus outbreak. https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_outbreak.

Wong G, Bi YH, Wang QH, Chen XW, Zhang ZG, Yao YG. 2020. Zoonotic origins of human coronavirus 2019 (HCoV-19 / SARS-CoV-2): why is this work important?. *Zoological Research*, **41**(3): 213−219.

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020a. A new coronavirus associated with human respiratory disease in China. *Nature*, **579**(7798): 265−269.

Wu JT, Leung K, Leung GM. 2020b. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*, **395**(10225): 689−697.

Wu ZY, McGoogan JM. 2020. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *JAMA*, doi: 10.1001/jama.2020.2648.

Xiao KP, Zhai JQ, Feng YY, Zhou N, Zhang X, Zou JJ, Li N, Guo YQ, Li XB, Shen XJ, Zhang ZP, Shu FF, Huang WY, Li Y, Zhang ZD, Chen RA, Wu YJ, Peng SM, Huang M, Xie WJ, Cai QH, Hou FH, Liu YH, Chen W, Xiao LH, Shen YY. 2020. Isolation and characterization of 2019-nCoV-like coronavirus from malayan pangolins. *bioRxiv*, doi: 10.1101/2020.02.17.951335.

Yang Y, Lu QB, Liu MJ, Wang YX, Zhang AR, Jalali N, Dean N, Longini I, Halloran ME, Xu B, Zhang XA, Wang LP, Liu W, Fang LQ. 2020. Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China. *medRxiv*, doi: 10.1101/2020.02.10.20021675.

Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP. 2002. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *American Journal of Human Genetics*, **70**(3): 635−651.

Zhang RQ, Liu H, Li FY, Zhang B, Liu QL, Li XW, Luo LM. 2020a. Transmission and epidemiological characteristics of Novel Coronavirus (2019-nCoV) Pneumonia (NCP): preliminary evidence obtained in comparison with 2003-SARS. *medRxiv*, doi: 10.1101/2020.01.30.20019836.

Zhang T, Wu QF, Zhang ZG. 2020b. Pangolin homology associated with 2019-nCoV. *Current Biology*, **30**: 1346−1351.

Zhao ZM, Li HP, Wu XZ, Zhong YX, Zhang KQ, Zhang YP, Boerwinkle E, Fu YX. 2004. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evolutionary Biology*, **4**: 21.

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**(7798): 270−273.

Zhu N, Zhang DY, Wang WL, Li XW, Yang B, Song JD, Zhao X, Huang BY, Shi WF, Lu RJ, Niu PH, Zhan FX, Ma XJ, Wang DY, Xu WB, Wu GZ, Gao GF, Tan WJ. 2020. A novel coronavirus from patients with pneumonia in China, 2019. *The New England Journal of Medicine*, **382**(8): 727−733.