The early cryptic transmission and evolution of SARS-CoV-2 in human hosts

Libing Shen[1#*], Funan He[2#], Zhao Zhang[3#],

Author Affiliations:

[1]Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, 200031, P.R. China

[2]School of Life Sciences, Fudan University, Shanghai, 200433, China

[3]Department of Biochemistry and Molecular Biology, McGovern Medical School at The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

# These authors are contributed equally to this work.

* Corresponding author: Libing Shen; Email: shenlibing@ion.ac.cn

Summary

Background Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is first identified in Wuhan City at the end of December 2019 and responsible for the ongoing coronavirus disease 2019 (COVID-19) pandemic. So far this pandemic has claimed more than one million lives all over the world and its origin is still unknown.

Methods In this study, we developed a method to search the least mutated strain using SARS-CoV-2 whole genome sequences. By parsimony principle, the least mutated strain should be the phylogenetic root for all SARS-CoV-2s. We further investigated the SARS-CoV-2's adaptive evolutionary process in human hosts using the least mutated strain as the phylogenetic root and analyzed its strain diversity in different countries/regions.

Findings According to their coding region identity, we classified 4571 SARS-CoV-2 genome sequences into 2449 viral strains collected from human hosts between December 2019 and July 2020. We found that the SARS-CoV-2 (NC_045512) strain first identified in Wuhan is not the least mutated strain. There are 41 SARS-CoV-2 strains harboring fewer global point mutations than the NC_045512 strain in our dataset. The least mutated strain can be found in eight countries across four continents due to SARS-CoV-2's low mutability. Eight positive selection sites are identified in five SARS-CoV-2's genes and four of them were present in the early stage of SARS-CoV-2's human-to-human transmission. The NC_045512 strain has two positive selection sites, one in RNA-dependent RNA polymerase (L314P) and the other in spike protein (G614D). The statistical analysis of the SARS-CoV-2's strain diversity in different countries/regions shows that the Indian subcontinent has the highest strain diversity. Furthermore, based on the SARS-CoV-2's mutation rate, we estimate that the earliest SARS-CoV-2 transmission in human hosts could be traced back to July or August of 2019.

Interpretation Our result shows that Wuhan is not the place where human-to-human

SARS-CoV-2 transmission first happened. Before it spread to Wuhan, SARS-CoV-2 has already experienced adaptive evolution during its human-to-human transmission. The positive selection sites could contribute to the different clinical features of different SARS-CoV-2 strains. Both the least mutated strain's geographic information and the strain diversity suggest that the Indian subcontinent might be the place where the earliest human-to-human SARS-CoV-2 transmission occurred, which was three or four months prior to the Wuhan outbreak. Our study helps to elucidate the early cryptic transmission and evolution of SARS-CoV-2 in human hosts and provide the new thinking for the global management of the COVID-19 pandemic.

Introduction

The coronavirus disease 2019 (COVID-19) is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is clustered together with the SARS-CoVs in bat and regarded as a SARS-like virus [1]. It can cause a variety of symptoms in human host including fever, cough, fatigue, loss of sense of smell, and pneumonia [2]. Although it has a lower mortality rate than SARS-CoV, SARS-CoV-2 exhibits a high contagious transmission pattern with low detectability, making this virus more threating than any coronavirus before. Now COVID-19 is an ongoing global pandemic and has infected more than 30 million people and claimed over one million lives across 200 countries and regions.

The first known COVID-19 epidemic broke out in December, 2019, in Wuhan City, China. The news reported that the early SARS-CoV-2's cases were tightly linked to a sea food market in Wuhan, but its origin is still veiled in secrecy. Even the evolutionary relationship within SARS-CoV-2s is very obscure. The bat RaTG13 coronavirus is usually used as the outgroup to root the SARS-CoV-2's phylogenetic tree in some practices [3]. However, the bat RaTG13 coronavirus was discovered in 2013 and the great sequence divergence between the bat virus and SARS-CoV-2s created a problem called long-branch attraction [4]. That is the fast evolving SARS-CoV-2 branches which would be placed close to the bat RaTG13 coronavirus in phylogenetic analysis as if they were the basal group to the other SARS-CoV-2s.

The first complete SARS-CoV-2 genome was collected in Wuhan and sequenced in Shanghai, China [5]. Many researchers used it as the phylogenetic root for analyzing SARS-CoV-2's evolution. However, one question is whether it is truly the first SARS-CoV-2 strain responsible for infecting human host. If it was not, then the results in some studies could be devious. In this work, we developed a method to search the least mutated SARS-CoV-2 strain by globally comparing the point mutations between two viral sequences in the whole dataset. Our method does not need any outgroup and thus circumvents the long-branch attraction trap. By

parsimony principle, the SARS-CoV-2 strain with the least number of mutations should be the phylogenetic root to the other SARS-CoVs. Our result shows that the SARS-CoV-2 sequence first identified in Wuhan is not the least mutated strain, which means that Wuhan cannot be the first place where human-to-human SARS-CoV-2 transmission happened. Our study further reveals some interesting details in the early stage of SARS-CoV-2's evolution in human host and indicates that the Indian subcontinent might be the place of earliest human-to-human SARS-CoV-2 transmission.

Data and Method

SARS-CoV-2 genome data

We retrieved the SARS-CoV-2 genome data from Coronavirus genomes of NCBI Datasets (https://www.ncbi.nlm.nih.gov/datasets/coronavirus/genomes/). We only downloaded the sequence of human host only. The NCBI dataset has both SARS-CoV-2's gene and genome sequences. We only keep the sequence longer than 29000 nucleotides and without any ambiguous site (such as N) for this study. After removing the short and low-quality sequences, 4678 SARS-CoV-2 genome sequences were kept for the next step analysis.

Coding region extraction and alignment

We used the complete coding region of the SARS-CoV-2 first identified in Wuhan (NCBI accession ID: NC_045512) as the reference template to extract the coding region for the filtered SARS-CoV-2 genome sequences. Each SARS-CoV-2 genome sequence was aligned to the coding region of NC_045512. We used MUSCLE 3.8.31 to perform the alignment with default parameters [6]. The sequence with incomplete opening reading frame, deletion or insertion according to NC_045512 was excluded from this study. Then the coding regions of all SARS-CoVs were merged into a single file using an in-house Perl script. The final alignment contains 4571 SARS-CoV-2 sequences with 29409 nucleotide sites (Supplemental data 1).

Global point mutation calculation

In this study, we calculated the number of global point mutations in a SARS-CoV-2 genome sequence by a pair-wise comparison fashion. Each sequence was compared with the other sequences in the final SARS-CoV-2 alignment and the number of point mutations from each comparison was summed up. By doing so, one can obtain the number of global point mutations for each sequence. According to parsimony principle, the basal sequence should have the least number of global point mutations (Figure S1).

Mutation rate, transition-transversion ratio and phylogenetic analyses

Mutation rate was calculated as per nucleotide per year using our final alignment result. Transition-transversion ratio was computed with MEGA X [7]. The phylogenetic tree of 4571 SARS-CoV-2 genome sequences was constructed with FastTree 2.1.7 using the GTR model and gamma distribution [8]. The neighbor-joining tree for mutation profile analyses were also constructed with MEGA X using gamma distribution of 4 rates [7].

Positive selection analysis and statistical analysis

We extracted the coding region for each SARS-CoV-2 gene in the final alignment using the NC_045512 SARS-CoV-2's gene as the reference template. The HyPhy program in MEGA X and the CODEML program in PAML 4.7 package were used to detect the positive selection site in the codon alignment of each SARS-CoV-2 protein set [9, 10]. For CODEML, we chose the site-specific model for selection signal detection. The site-specific model was performed by comparing the model M2a (positive selection) vs. the null model M1a (nearly neutral). The R package (version 3.4.4) was used to perform Chi-squared test and *P*-value smaller than 0.05 was viewed as statistically significant in this study.

Result

Number of strains and mutation rates

The length of the SARS-CoV-2 genome first identified in Wuhan is 29903 nucleotides [11]. Its 5'UTR has 265 nucleotides and 3'UTR has 229 nucleotides. Thus, its gene coding region has 29409 nucleotides and almost accounts for 98.3% of total genome. We observed 3182 mutation sites in the final alignment and each SARS-CoV-2 averagely has less than one mutation site (Supplemental data 2). The number of global point mutations for each SARS-CoV-2 can be found in Supplemental data 3. According to the coding region sequence identity, 4571 SARS-CoV-2 genome sequences can be classified into 2449 viral strains (Supplemental data 4). The number of SARS-CoV-2 sequences varies from strain to strain. The largest strain contains 318 SARS-CoV-2 sequences while the smallest has only one sequence. Notably, the identical sequences within one strain were often collected from different dates and different countries or regions. For example, the NC_045512 strain can be found in five countries and its collection date ranges from 2019/12/30 to 2020/04/04. This result proposes that SARS-CoV-2 has a very low mutability and is capable of long time and range transmission.

Interestingly, our result shows that the NC_045512 strain is not the least mutated strain. There are 41 strains having fewer global point mutations than it (Supplemental data 3 and 4). The least mutated strain has 15552 global point mutations whereas the NC_045512 one has 18421 global point mutations. The least mutated strain with 41 sequences are collected from Bangladesh, USA, Greece, Australia, India, Italy, Czech Republic, Russia and Serbia (Supplemental data 4). Their collection date ranges from 2020/03/01 to 2020/06/12. Using its number of global point mutations, we calculated the mutation rate for SARS-CoV-2 as follows. Firstly, we assume that there was no mutation in the least mutated strain and 15552 total point mutations are shared by the rest 2448 SARS-CoV-2 strains. Thus, there are average 6.35 point mutations in each strain. For our final alignment with 29409 nucleotide sites, the expected mutation rate was 2.16×10^{-4} per nucleotide for SARS-CoV-2's coding region. Considering that the first SARS-CoV-2 patient was hospitalized on 12 December 2019 in Wuhan City and the final collection date in our dataset is 2020/07/12 [5], 15552 total point mutations

were accumulated in seven months and we estimate that the mutation rate is 3.88×10^{-4} per nucleotide per year for SARS-CoV-2's coding region, which is much lower than those of SARS-CoV and MERS-CoV (Table 1). The mutation rates and the transition-transversion bias for SARS-CoV and MERS-CoV were collected or calculated from literatures [12-15]. Moreover, the transition-transversion bias for SARS-CoV-2 is higher than those of SARS-CoV and MERS-CoV. Both its mutation rate and transition-transversion bias show that SARS-CoV-2 has a very stable genome.

Phylogenetic analyses and mutation profile analyses

Based on parsimony principle, the strain with the least number of mutations should be the basal clade, i.e. the root, in a phylogenetic tree. We first reconstructed the phylogenetic tree for 4571 SARS-CoV-2s and marked out the 42 less mutated strains named 0001 to 0042 on it. Figure 1 shows that more than half of less mutated strains are very close to the least mutated strain (0001) while the NC_045512 strain (0042) is distant from the other less mutated strains. Although the tree is unrooted, the distribution of less mutated strains proposes that the strain 0001 is more likely to be the root of the tree than the strain 0042.

Next, we tested the possibility of being the root of the tree for the strain 0001 and the strain 0042. First, we reconstructed the phylogenetic trees with only 42 less mutated strains. Second, we used the strain 0001 and the strain 0042 as the template to annotate the number of mutations for the rest 41 strains (Supplemental data 5 and 6). Third, after combined the phylogenetic tree with two mutation profiles, we rooted the tree with either the strain 0001 or the strain 0042.

If the tree is rooted with the strain 0001, the mutation profile shows a total of 56 mutations (Figure 3a). If the tree is rooted with the strain 0042, the mutation profile shows a total of 164 mutations (Figure 3b). It is noteworthy that two reverse mutations occurred on the strain 0007 and 0008, respectively, if the tree is rooted with the strain 0042. Compared with the strain 0042, the strain 0004 has five mutations at

the position 794 (C to T), 2772 (C to T), 14143 (C to T), 23138 (A to G), and 25298 (G to T); the strain 0041 has six mutations at the position 794 (C to T), 2772 (C to T), 14143 (C to T), 23138 (A to G), 25298 (G to T) and 27879 (T to C); the strain 0007 has four mutations at the position 794 (C to T), 14143 (C to T), 23138 (A to G), and 25298 (G to T); the strain 0008 has four mutations at the position 794 (C to T), 2772 (C to T), 14143 (C to T), and 25298 (G to T).

However, based the topology and branch length of the phylogenetic tree in Figure 2b, the strain 0007 and 0008 should have six mutations compared with the strain 0042. Thus, two hypothetical reverse mutations should have happened in the strain 0007 and 0008, respectively, if one believed that the strain 0042 was the ancestor to 41 SARS-CoV-2 strains in Figure 2b. For a coding region of 29409 nucleotides, the probability that reverse mutation happened at one site would be $1/(29409\times29409\times3) = 3.85\times10^{-10}$. The probability that four reverse mutations happened at the same time would be 3.85×10^{-10} to the power of four, which equals 2.2×10^{-38}. Table 1 has already shown that SARS-CoV-2 is a virus with very low mutation rate. Thus, it is nearly impossible that the strain 0042 could be the root of SARS-CoV-2's phylogenetic tree. In another word, the SARS-CoV-2 genome sequence first identified in Wuhan cannot be the ancestor of the other SARS-CoV-2s, which exclude the possibility that human-to-human SARS-CoV-2 transmission first happened in Wuhan City, China.

Positive selection analysis for SARS-CoV-2 genes

Viral transmission is usually accompanied by positive selection event which facilitates the virus jumping between different hosts or help it adapt to a new host population. We explored the possible positive selection events in 2449 SARS-CoV-2 strains. Eight positive selection sites were detected in *Orf1ab1*, *Orflab2*, spike (*S*), *Orf3a* and nucleocapsid (*N*) genes (Table 2). Three of them were located in *Orf1ab1* gene. Orf1ab1 and Orflab2 genes totally encodes 16 predicted non-structural proteins (nsps) [5]. Two out of three positive selection sites in Orf1ab1 were in nonstructural protein 2 (nsp2) and one of them was in nonstructural protein 6 (nsp6). Nsp2 interacts

with the host protein complex involved in mitochondrial biogenesis and intracellular signaling [16], whereas nsp6 can induce double-membrane vesicles and restrict autophagosome expansion [17, 18]. The positive selection site for *Orflab2* is in nonstructural protein 12 (nsp12) which is a RNA dependent RNA polymerase (RdRp) [19]. *S* gene, a very important protein for SARS-CoV-2's transmission, encodes glycoprotein spike which binds to human ACE2 receptor [20]. *Orf3a* gene encodes a transmembrane protein which can induce apoptosis in cells [21]. *N* gene binds to the viral RNA genome and processes the virus particle assembly and release [22].

Interestingly, four positive selection sites have already presented in 42 less mutated strains. They are the epitome of the early stage of SARS-CoV-2's transmission in human hosts. The phylogenetic analysis shows that the positive selection events in *Orf1ab1* and *Orf3a* are mainly present in the strains geographically associated with USA whereas the positive selection event in *Orflab2* is present in the European strains (Figure 3). The G614D mutation in *S* protein is recurrent and independently emerges in the strain 0042 (Wuhan City, China) and the strain 0008 (California, USA). The CODEML result shows that 614D exhibits a weak positive selection signal even in 42 less mutated strains (lnL_0 = -5068.97 , lnL_1 = -5065.4, *P*-value = 0.168), although the *P*-value is not significant with this very limited dataset.

Statistical analysis on the distribution of the numbers of SARS-CoV-2 sequences and strains in different countries and regions

The phylogenetic analyses combined with mutation profiles argue that the SARS-CoV-2 strain first identified in Wuhan City is highly impossible to be the ancestor of all SARS-CoV-2 in our study. By parsimony principle, the least mutated strain should be the ancestor of all SARS-CoV-2s. The problem is that it can be found in eight countries including Australia, Bangladesh, Greece, USA, Russia (Serbia), Italy, India, and Czech Republic. Theoretically, anyone of them could be the birthplace of the least mutated strain. Due to its low mutability, SARS-CoV-2s is capable of preserving its genome integrity after a long time and range of transmission.

Since phylogenetic analysis is unable to provide the clue for SARS-CoV-2's origin, one common sense is that the birthplace of a virus usually has a high diversity of viral strains. In this study, we have total 4571 SARS-CoV-2 genome sequences from 2449 strains. Each country or region has a percentage of SARS-CoV-2's sequences and percentage of SARS-CoV-2's strains. For example, USA has 3297 SARS-CoV-2 sequences from 1631 strain. The percentage of SARS-CoV-2's sequences from USA is 71.13% (3297/4571) and the percentage of SARS-CoV-2's strains from USA is 66.6% (1631/2449) in our dataset. Using Chi-square test, we can examine whether these two percentages are significantly different or not in 17 countries/regions (Table 3).

If use the percentage of SARS-CoV-2's sequences in a country to multiply the total number of strains, one will get the expected number of SARS-CoV-2's strains in a country from the sequence percentage's perspective, vice versa for the percentage of SARS-CoV-2's strains. For example, if we multiply 71.13% by 2449, we will have 1766 which is the expected number of SARS-CoV-2's strains in USA; if we multiply 66.6% by 4571, we will have 3044 which is the expected number of SARS-CoV-2's sequences in USA. Then Chi-squared test can be used to examine the observed numbers and the expected numbers and in a contingency table.

We examined the percentages of SARS-CoV-2's sequences and strains in 17 countries and regions (Table 4). USA, India, Bangladesh, and Saudi Arabia have significant *P*-values. For USA, the statistical significance proposes that its percentage of SARS-CoV-2's sequences is much higher than its percentage of SARS-CoV-2's strains in our dataset, a sign of low viral strain diversity. In India and Bangladesh, the statistical significance is too high to be overlooked. It proposes that the percentage of strains is much higher that of sequences in these two countries, which a sign of high viral strain diversity. Both India and Bangladesh are located in the Indian subcontinent. Saudi Arabia with a relative high strain diversity is also geographically

close to the Indian subcontinent, just across the Arabian Sea. With the evidence that SARS-CoV-2 has a very low mutability and the least mutated strain can be found in both India and Bangladesh, it is natural for someone to deduce that the earliest human-to-human SARS-CoV-2 transmission might occur on the Indian subcontinent.

The possible date of the earliest human-to-human SARS-CoV-2 transmission

Based on its mutation rate, a SARS-CoV-2 accumulates about 11.41 point mutations in its gene coding region per year. There are three nucleotide differences between the least mutated strain and the NC_045512 strain. It proposes that the divergence time between two strains is about three or four months (0.951 point mutation per month). Since the first SARS-CoV-2 patient was hospitalized on 12 December 2019 in Wuhan City, the SARS-CoV-2 cryptic transmission in Wuhan must have started in the middle or late November of 2019 for SARS-CoV-2 has an incubation period of 2 to 14 days [23]. Therefore, the earliest human-to-human SARS-CoV-2 transmission could be traced back to July or August, 2019.

Discussion

There are a lot of mysteries to SARS-CoV-2 so far. The first mysterious thing is that different SARS-CoV-2 strains evolve at different rates. Some of them are very stable and capable of long time and range transmission without any mutation. It seems that SARS-CoV-2 has a genome-replicase complex with extremely high fidelity which grants SARS-CoV-2 a much lower mutation rate than SARS-CoV and MERS-CoV. *Orf1ab* gene encodes 16 nonstructural proteins for coronavirus, most of which have certain replicase/transcriptase function [24]. However, which nonstructural protein is mainly responsible for the stability of SARS-CoV-2's genome remains under experimental investigation.

The slow evolutionary rate of SARS-CoV-2s means that only a few informative sites could be used for their phylogenetic analysis. Thus, the phylogenetic tree is only suitable for the probable estimation of SARS-CoV-2s' evolutionary relationship and

classification, which creates difficulty for tracing its origin and transmission. On the other hand, the slow evolutionary rate also makes multiple mutations at the same site or reverse mutation highly unlikely. For our method based on parsimony principle, it is an advantage. Our analyses show that the NC_045512 strain first identified in Wuhan City was not the ancestral sequence to all SARS-CoV-2s. Thus, it is impossible that human-to-human SARS-CoV-2 transmission first happened in Wuhan, China. SARS-CoV-2 must have spread undetectably in human population for some time before it was identified in Wuhan.

Positive selection helps a virus adapt to new hosts and enhance its transmission and spread. In order to coexist with its hosts, the general trend during viral evolution is to decrease its pathogenicity and meanwhile increase its transmissibility. One positively selected site (G614D) in spike gene has been systematically studied [25], but the work of Bette et al proposed that D614G mutation was more adaptive, while our work found that G614D was actually positively selected. Their work showed that D614G mutation could increase upper respiratory tract viral loads but not with increased disease severity [25]. For another perspective, this result means that G614D mutation has a lower detectability than D614G. Stealth is a clear advantage in SARS-CoV-2's transmission. T265I in *Orf1ab1* (nsp2) and Q57H in *Orf3a* seem to be more likely associated with pathogenesis, because *Orf1ab1* (nsp2) involves in mitochondrial biogenesis and intracellular signaling and *Orf3a* can induce apoptosis in cells [15, 16]. If both T265I and Q57H were associated with less disease severity, it would explain why United States didn't notice this virus at the first place. *Orflab2* (nsp12 RNA polymerase) has been shown to be a potential antiviral drug target [26, 27]. Hypothetically, if an European patient with COVID-19 which was mistaken by his/her doctor as a severe viral flu was treated with a certain kind of antiviral drug, L314P in *Orflab2* would be a drug resistance mutation. If it was the case, it is no wonder that SARS-CoV-2 would catch doctor's attention in Wuhan and finally lead to its identification, because it becomes difficult to treat after it was cryptically imported to Wuhan from Europe and the COVID-19 mortality rate would be higher in China

and Europe than the other countries and regions. Certainly, the true adaptive features of T265I in *Orf1ab1*, Q57H in *Orf3a*, and L314P in *Orflab2* are unknown and needs further study.

Our results have proven that Wuhan is impossible to be the original place of human-to-human SARS-CoV-2 transmission. The questions then become what was its transmission route to Wuhan and where human-to-human transmission first happened. The phylogenetic analysis proposes that SARS-CoV-2 was cryptically transmitted to Wuhan from Europe and experienced at least one positive selection event during the process. The least mutated strain is found, but the problem is that it covers eight countries from four continents. It seems kind of farfetched to use statistical analysis to deduce that the Indian subcontinent is the place of the earliest human-to-human SARS-CoV-2 transmission. However, the geographic vicinity of Bangladesh, India, and Saudi Arabia proposes that the significantly high SARS-CoV-2's strain diversity is unlikely to be an artifact in these three countries. The least mutated strain presenting in both Bangladesh and India makes the earliest human-to-human SARS-CoV-2 transmission occurred in the Indian subcontinent a very sound scientific hypothesis. If the Indian subcontinent is the place of the earliest human-to-human SARS-CoV-2 transmission, one will naturally ask why SARS-CoV-2 wasn't identified there in the first place. The Indian subcontinent has a tropical climate and a very young population, both of which contribute to a mild symptom of COVID-19. As known for all, the hygiene condition is imperfect and the public medical system is less efficient in the subcontinent. Thus, it is conceivable that a virus with flu-like symptom could spread undetectably for several months there. If our hypothesis is correct, the prediction is that the subcontinent will exhibit the highest SARS-CoV-2's strain diversity as the virus sequencing continues.

Compared with the least mutated strain, the NC_045512 strain has three mutations. Based on SARS-CoV-2's mutation rate, we estimate that the beginning date for the earliest human-to-human transmission in the subcontinent was July or August, 2019.

From May to June 2019, the second longest recorded heat wave had rampaged in northern-central India and Pakistan, which created a serious water crisis in this region (https://en.wikipedia.org/wiki/2019_heat_wave_in_India_and_Pakistan). The water shortage made wild animals such as monkeys engage in the deadly fight over water among each other and would have surely increased the chance of human-wild animal interactions. We speculated that the zoonotic transmission of SARS-CoV-2 might be associated with this unusual heat wave as well. If it was the case, the heat wave would explain why SARS-CoV-2 is able to rapidly spread in the summer of 2020 while SARS-CoV and MERS-CoV usually slow down their spread in high temperatures. Before it was identified in Wuhan in December, 2019, SARS-CoV-2 had been spread to four continents without detection. In this regard, the COVID-19 pandemic is inevitable in 2020 and the Wuhan epidemic is only a part of it. We hope that our work could shed some light on this cryptic virus, but there are still many unanswered questions around SARS-CoV-2. If the zoonotic transmission of SARS-CoV-2 occurred in the subcontinent, where is the exact location of transmission? What is the natural host for SARS-CoV-2? The animal could be in or outside the subcontinent. It is notable that SARS-CoV-2-related coronaviruses have been found in Malayan pangolins and one of them has the highly similar receptor-binding domain to SARS-CoV-2 [28]. Malayan pangolin is widely distributed in Southeast Asia [29], where is geographically next to the Indian subcontinent. Certainly, the true origin of SARS-CoV-2 will not be revealed until its authentic natural host is found.

Reference

1. Xu X, Chen P, Wang J, Feng J, Zhou H, Li X, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Science China Life sciences. 2020.

2. Grant MC, Geoghegan L, Arbyn M, Mohammed Z, McGuinness L, Clarke EL, et al. The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries. PloS one. 2020; **15**(6): e0234765.

3. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences of the United States of America. 2020;

117(17): 9241-3.
4. Gribaldo S, Philippe H. Ancient phylogenetic relationships. Theoretical population biology. 2002; **61**(4): 391-408.
5. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020; **579**(7798): 265-9.
6. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004; **32**(5): 1792-7.
7. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular biology and evolution. 2018; **35**(6): 1547-9.
8. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PloS one. 2010; **5**(3): e9490.
9. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005; **21**(5): 676-9.
10. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution. 2007; **24**(8): 1586-91.
11. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. Author Correction: A new coronavirus associated with human respiratory disease in China. Nature. 2020; **580**(7803): E7.
12. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science. 2004; **303**(5664): 1666-9.
13. Pavlovic-Lazetic GM, Mitic NS, Tomovic AM, Pavlovic MD, Beljanski MV. SARS-CoV genome polymorphism: a bioinformatics study. Genomics, proteomics & bioinformatics. 2005; **3**(1): 18-35.
14. Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, et al. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. mBio. 2014; **5**(1).
15. Zhang Z, Shen L, Gu X. Evolutionary Dynamics of MERS-CoV: Potential Recombination, Positive Selection and Transmission. Scientific reports. 2016; **6**: 25049.
16. Cornillez-Ty CT, Liao L, Yates JR, 3rd, Kuhn P, Buchmeier MJ. Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. Journal of virology. 2009; **83**(19): 10314-8.
17. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. mBio. 2013; **4**(4).
18. Cottam EM, Whelband MC, Wileman T. Coronavirus NSP6 restricts autophagosome expansion. Autophagy. 2014; **10**(8): 1426-41.
19. te Velthuis AJ, Arnold JJ, Cameron CE, van den Worm SH, Snijder EJ. The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. Nucleic acids research. 2010; **38**(1): 203-14.
20. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020; **181**(2): 281-92 e6.
21. Ren Y, Shu T, Wu D, Mu J, Wang C, Huang M, et al. The ORF3a protein of SARS-CoV-2 induces apoptosis in cells. Cellular & molecular immunology. 2020; **17**(8): 881-3.
22. Zeng W, Liu G, Ma H, Zhao D, Yang Y, Liu M, et al. Biochemical characterization of SARS-CoV-2 nucleocapsid protein. Biochemical and biophysical research communications. 2020; **527**(3): 618-23.
23. Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung SM, et al. Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. Journal of clinical medicine. 2020; **9**(2).
24. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA

proofreading machine regulates replication fidelity and diversity. RNA biology. 2011; **8**(2): 270-9.
25. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020; **182**(4): 812-27 e19.
26. Gordon CJ, Tchesnokov EP, Feng JY, Porter DP, Gotte M. The antiviral compound remdesivir potently inhibits RNA-dependent RNA polymerase from Middle East respiratory syndrome coronavirus. The Journal of biological chemistry. 2020; **295**(15): 4773-9.
27. Ruan Z, Liu C, Guo Y, He Z, Huang X, Jia X, et al. SARS-CoV-2 and SARS-CoV: Virtual screening of potential inhibitors targeting RNA-dependent RNA polymerase activity (NSP12). Journal of medical virology. 2020.
28. Lam TT, Jia N, Zhang YW, Shum MH, Jiang JF, Zhu HC, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature. 2020; **583**(7815): 282-5.
29. Francis CM, Barrett P. A guide to the mammals of Southeast Asia. Princeton, N.J.: Princeton University Press; 2008.

Table 1. Mutation rates and transition vs. transversion (Ts/Tv) biases across three human infected coronaviruses.

Species	Mutation rate (mutation per nt per year)	Ts/Tv bias
SARS-CoV	3.01×10^{-3}	2.61
MERS-CoV	1.12×10^{-3}	1.87
SARS-CoV-2	3.88×10^{-4}	3.52

Table 2. Positive selection sites in five SARS-CoV-2's genes.

Gene	Mutation	*P*-value	Positively selected site	Reference position
Orf1ab1	C>T	8.8×10^{-5}	T265I*	1059
Orf1ab1	A>T	0.043	I300F*	1163
Orf1ab1	G>T	7.9×10^{-6}	L3606F*	11083
Orf1ab2	T>C	8.9×10^{-5}	L314P*	14408
S	C>T	0.012	L5F*	21575
S	G>A	4.5×10^{-5}	G614D*	23403
Orf3a	G>T	0.035	Q57H*	25563
N	C>T	0.0035	A194G*	28854

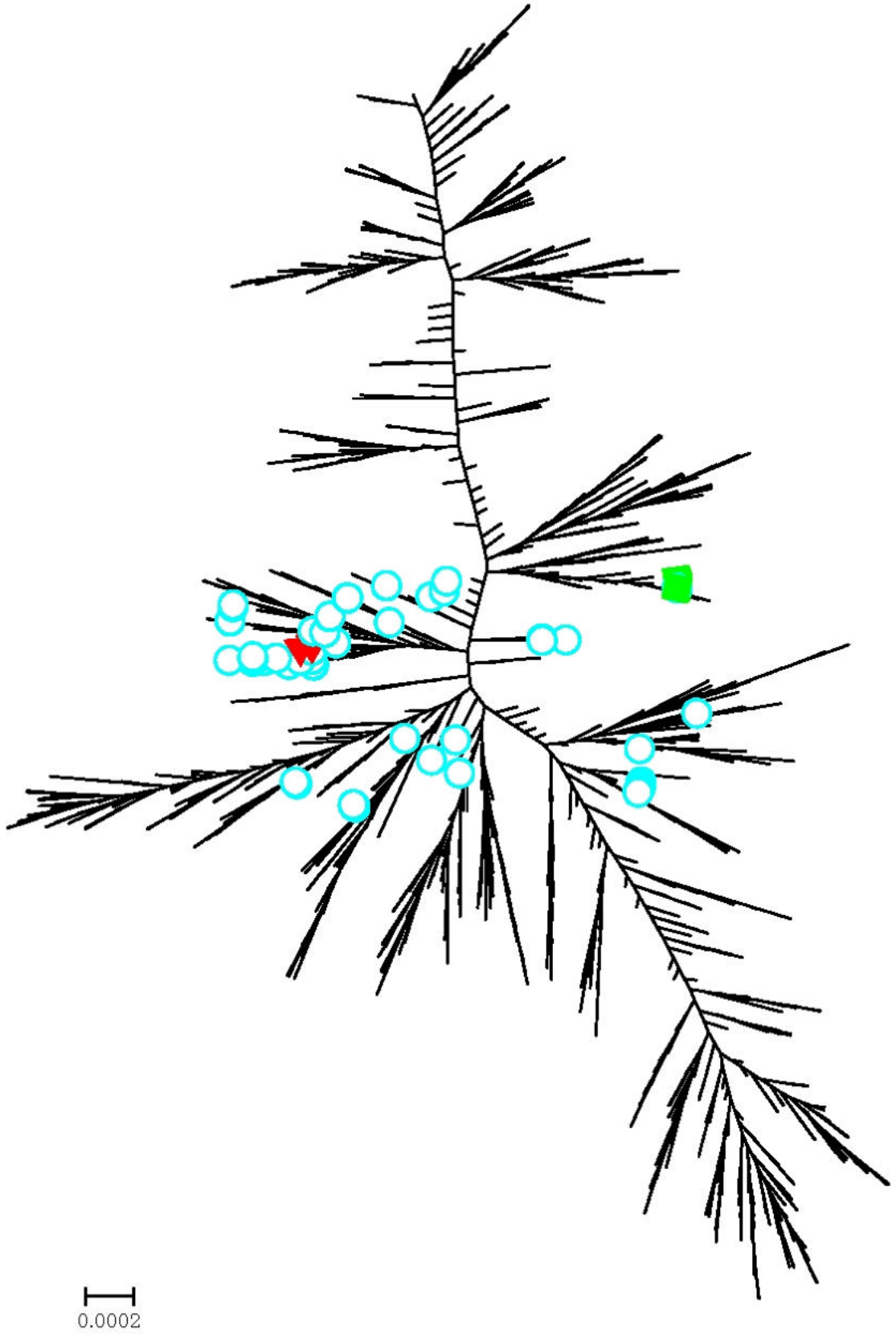
Table 3. Numbers of SARS-CoV-2 sequences and strains in 17 countries/regions.

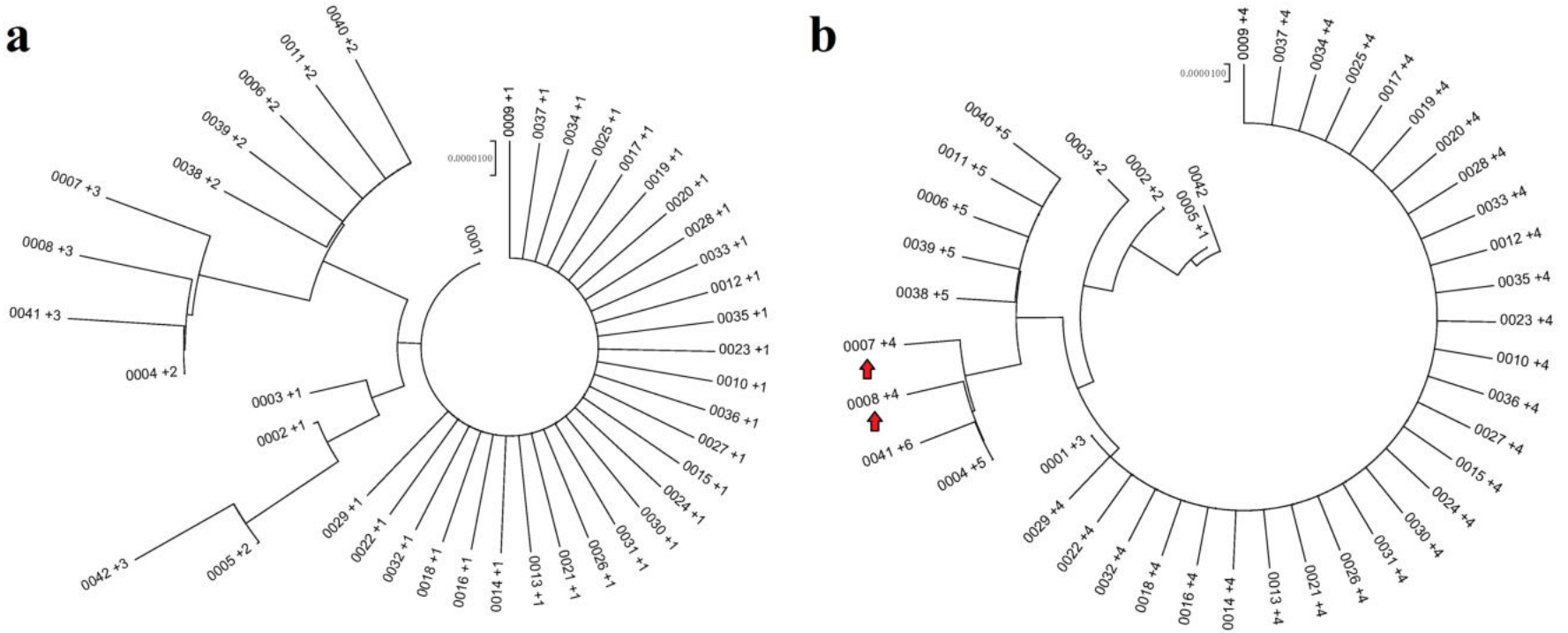
Country/Region	Number of sequences	Number of strains	Percentage of sequences	Percentage of strains
USA	3297	1631	71.13%	66.6%
India	319	247	6.98%	10.09%
Australia	263	155	5.75%	6.33%
Bangladesh	158	138	3.46%	5.63%
China	96	52	2.1%	2.12%
France	81	50	1.77%	2.04%
Greece	77	43	1.68%	1.76%
Germany	39	15	0.85%	0.61%
Taiwan, China	29	23	0.63%	0.94%
Saudi Arabia	21	19	0.46%	0.78%
Czech Republic	20	13	0.44%	0.53%
Hong Kong, China	19	12	0.42%	0.49%
Thailand	18	14	0.39%	0.49%
Spain	18	12	0.39%	0.57%
Egypt	14	8	0.31%	0.33%
Italy	9	8	0.2%	0.32%
Morocco	8	6	0.18%	0.24%

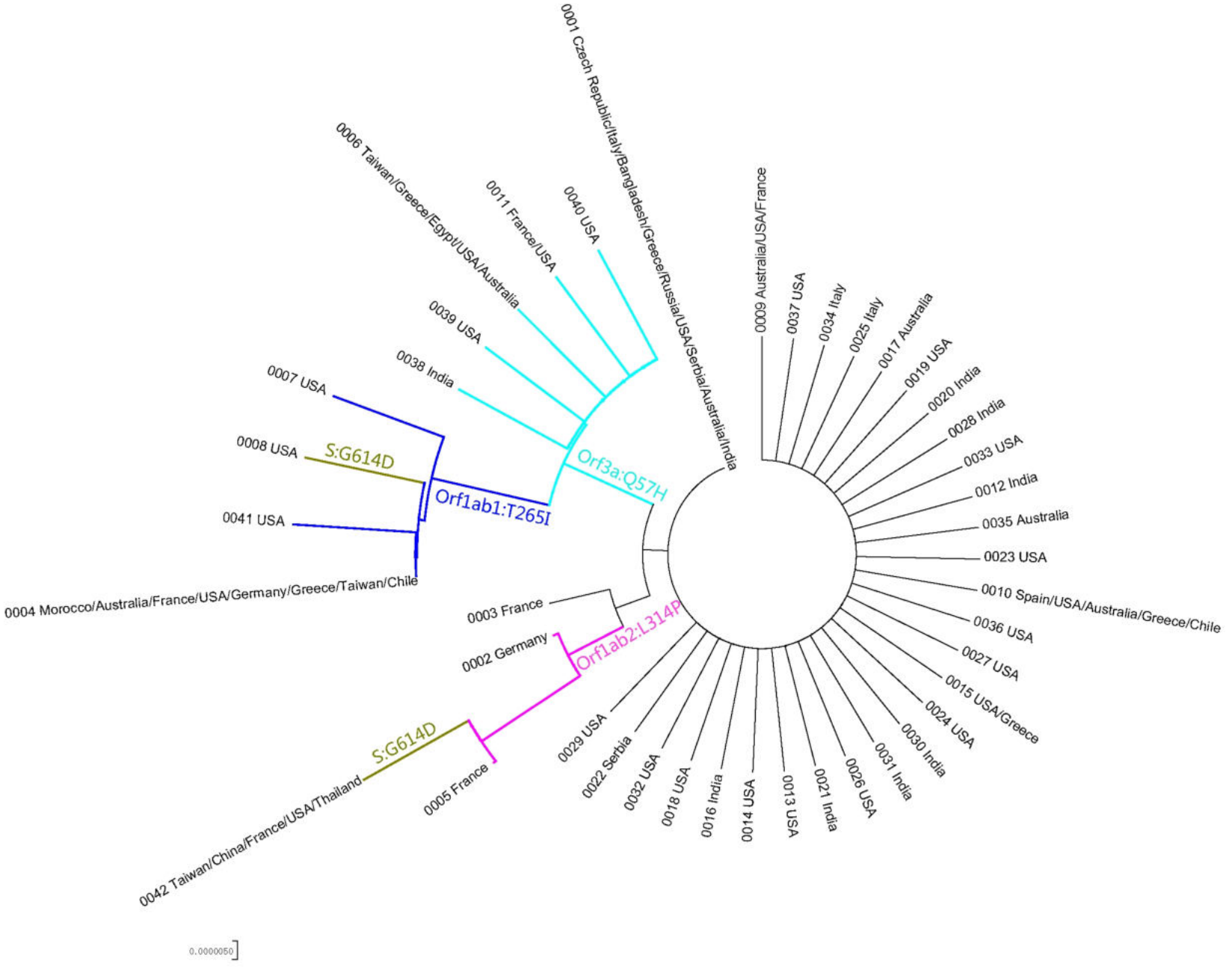
Table 4. Chi-squared tests for the observed numbers of sequences and strains and the expected numbers of sequences and strains in 17 countries/regions.

Country/Region	Category	Observed number	Expected number	*P*-value
USA	sequence	3297	3044	0.000196**
	strain	1631	1766	
India	sequence	319	461	2.66×10^{-9}**
	strain	247	171	
Australia	sequence	263	298	0.1456
	strain	155	141	
Bangladesh	sequence	158	258	1.25×10^{-8}**
	strain	138	85	
China	sequence	96	97	1
	strain	52	51	
France	sequence	81	93	0.3199
	strain	50	43	
Greece	sequence	77	80	0.8553
	strain	43	41	
Germany	sequence	39	28	0.1627
	strain	15	21	
Taiwan, China	sequence	29	43	0.09194
	strain	23	16	
Saudi Arabia	sequence	21	35	0.03917*
	strain	19	11	
Czech Republic	sequence	20	22	0.665
	strain	13	10	
Hong Kong, China	sequence	19	25	0.7214
	strain	12	11	
Thailand	sequence	18	22	0.2621
	strain	14	10	
Spain	sequence	18	22	0.6498
	strain	12	10	
Egypt	sequence	14	15	1
	strain	8	8	
Italy	sequence	9	15	0.2913
	strain	8	5	
Morocco	sequence	8	11	0.5991
	strain	6	4	

* Indicates P-value smaller than 0.05. ** Indicates P-value smaller than 0.005.


0.0002


a
0.0000100
0009 +1
0037 +1
0034 +1
0025 +1
0017 +1
0019 +1
0020 +1
0028 +1
0033 +1
0012 +1
0035 +1
0023 +1
0010 +1
0036 +1
0027 +1
0015 +1
0024 +1
0030 +1
0031 +1
0026 +1
0021 +1
0013 +1
0014 +1
0016 +1
0018 +1
0032 +1
0022 +1
0029 +1
0001
0003 +1
0002 +1
0042 +3
0005 +2
0041 +3
0004 +2
0008 +3
0007 +3
0038 +2
0039 +2
0006 +2
0011 +2
0040 +2
b
0.0000100
0009 +4
0037 +4
0034 +4
0025 +4
0017 +4
0019 +4
0020 +4
0028 +4
0033 +4
0012 +4
0035 +4
0023 +4
0010 +4
0036 +4
0027 +4
0015 +4
0024 +4
0030 +4
0031 +4
0026 +4
0021 +4
0013 +4
0014 +4
0016 +4
0018 +4
0032 +4
0022 +4
0029 +4
0001 +3
0004 +5
0041 +6
0008 +4
0007 +4
0038 +5
0039 +5
0006 +5
0011 +5
0040 +5
0003 +2
0002 +2
0042
0005 +1


0001 Czech Republic/Italy/Bangladesh/Greece/Russia/USA/Serbia/Australia/India
0006 Taiwan/Greece/Egypt/USA/Australia
0011 France/USA
0040 USA
0039 USA
0038 India
0007 USA
0008 USA
S:G614D
Orf1ab1:T265I
0041 USA
0004 Morocco/Australia/France/USA/Germany/Greece/Taiwan/Chile
Orf3a:Q57H
0009 Australia/USA/France
0037 USA
0034 Italy
0025 Italy
0017 Australia
0019 USA
0020 India
0028 India
0033 USA
0012 India
0035 Australia
0023 USA
0010 Spain/USA/Australia/Greece/Chile
0036 USA
0027 USA
0015 USA/Greece
0024 USA
0030 India
0031 India
0026 USA
0021 India
0013 USA
0014 USA
0016 India
0018 USA
0032 USA
0022 Serbia
0029 USA
0003 France
0002 Germany
Orf1ab2:L314P
S:G614D
0042 Taiwan/China/France/USA/Thailand
0005 France
0.0000050