Molecular Phylogenesis and Spatiotemporal Spread of SARS-CoV-2 in Southeast Asia

Mingjian Zhu[1], Jian Shen[1], Qianli Zeng[2], Joanna Weihui Tan[3], Jirapat Kleepbua[4], Ian Chew[5], Jia Xian Law[6], Sien Ping Chew[7], Anita Tangathajinda[4], Natthjija Latthitham[4] and Lanjuan Li[1]*

[1] State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2] Shanghai Institute of Biological Products, Shanghai, China, [3] Faculty of Arts and Social Sciences, National University of Singapore, Singapore, Singapore, [4] Thammasat University Hospital, Pathum Thani, Thailand, [5] Zhejiang University School of Medicine, Hangzhou, China, [6] Tuanku Ja'afar Hospital, Seremban, Malaysia, [7] Shanghai Jiao Tong University School of Medicine, Shanghai, China

Background: The ongoing coronavirus disease 2019 (COVID-19) pandemic has posed an unprecedented challenge to public health in Southeast Asia, a tropical region with limited resources. This study aimed to investigate the evolutionary dynamics and spatiotemporal patterns of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the region.

Materials and Methods: A total of 1491 complete SARS-CoV-2 genome sequences from 10 Southeast Asian countries were downloaded from the Global Initiative on Sharing Avian Influenza Data (GISAID) database on November 17, 2020. The evolutionary relationships were assessed using maximum likelihood (ML) and time-scaled Bayesian phylogenetic analyses, and the phylogenetic clustering was tested using principal component analysis (PCA). The spatial patterns of SARS-CoV-2 spread within Southeast Asia were inferred using the Bayesian stochastic search variable selection (BSSVS) model. The effective population size (Ne) trajectory was inferred using the Bayesian Skygrid model.

Results: Four major clades (including one potentially endemic) were identified based on the maximum clade credibility (MCC) tree. Similar clustering was yielded by PCA; the first three PCs explained 46.9% of the total genomic variations among the samples. The time to the most recent common ancestor (tMRCA) and the evolutionary rate of SARS-CoV-2 circulating in Southeast Asia were estimated to be November 28, 2019 (September 7, 2019 to January 4, 2020) and 1.446×10^{-3} (1.292×10^{-3} to 1.613×10^{-3}) substitutions per site per year, respectively. Singapore and Thailand were the two most probable root positions, with posterior probabilities of 0.549 and 0.413, respectively. There were high-support transmission links (Bayes factors exceeding 1,000) in Singapore, Malaysia, and Indonesia; Malaysia involved the highest number (7) of inferred transmission links within the region. A twice-accelerated viral population expansion, followed by a temporary setback, was inferred during the early stages of the pandemic in Southeast Asia.

Conclusions: With available genomic data, we illustrate the phylogeography and phylodynamics of SARS-CoV-2 circulating in Southeast Asia. Continuous genomic surveillance and enhanced strategic collaboration should be listed as priorities to curb the pandemic, especially for regional communities dominated by developing countries.

Keywords: SARS-CoV-2, phylogenetic tree, phylogeography, phylodynamics, effective population size (Ne), Bayesian inference, Southeast Asia

INTRODUCTION

Coronavirus disease 2019 (COVID-19) is an infectious disease, first reported in Wuhan, China, in late December 2019 (1, 2). COVID-19 is caused by a positive-sense, single-stranded RNA (+ssRNA) virus of zoonotic origin (3, 4). Given the close genetic similarity, it was classified as a new member of the genus Betacoronavirus and named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (5). SARS-CoV-2 has a genome of approximately 30 kb, encoding four structural proteins, namely spike (S), membrane (M), envelope (E), and nucleocapsid (N) proteins, as well as several non-structural proteins (6). The S protein on the surface contains the receptor-binding domain (RBD), which is critical for mediating cell entry and is relatively prone to mutations (7).

As a global public health threat, the COVID-19 outbreak was declared a pandemic by the WHO on March 11, 2020 (8). During the first 8 months of 2020, more than 25 million cases and 800,000 deaths were reported worldwide (9). The WHO has reported that 90% of all countries experienced disruptions to essential health services, with low- and middle-income countries (LMICs) facing the greatest difficulties (10). The contribution of Southeast Asia, which is dominated by underdeveloped economies, to the global total of COVID-19 cases is nearly 2% at the time of writing since the first infection within the region was detected in Thailand on January 13, 2020 (9, 11). Unlike in other regions, the epidemic situation in Southeast Asia is unevenly distributed and is subject to further uncertainties (11, 12).

Genomic epidemiology, particularly phylodynamic analysis, is crucial for inferring evolutionary history and revealing epidemic patterns (13–15). Bayesian inference, incorporating prior information, has been applied to the phylodynamics of viral infectious diseases, including COVID-19 (16–18). During the ongoing COVID-19 pandemic, highly informative genome sequence data were generated and publicly shared through online platforms, such as the Global Initiative on Sharing Avian Influenza Data (GISAID). Genomic data revealed a continuous evolution of SARS-CoV-2 in almost all genomic regions, potentially in a country-specific manner (17). Although several COVID-19 genomic epidemiological studies have been conducted, most have focused on North America, Western Europe, and East Asia or individual countries (19–21). A large-scale phylogeographic study reconstructed the worldwide spread of the main SARS-CoV-2 clades; however, it did not provide details on the spread of the virus within each continent (especially within Asia, the largest continent) (22). Yap et al. conducted a study on Southeast Asia; they used phylogenetic and mutational analysis to identify three central variants with possible ongoing adaptation among 142 genomes sampled before mid-April 2020 (23). Islam et al. demonstrated the emergence of European and North American mutant variants of SARS-CoV-2 in South-East Asia and constructed a transmission map using Nextstrain with 329 sequences (24). Limited by the amount of genomic data available, the existing research findings should be considered an early snapshot of the evolution and dissemination of the SARS-CoV-2 strains circulating in the region. Moreover, the geographical scope of both studies covers only a few Southeast Asian countries. Thus, an updated and complete picture of SARS-CoV-2 phylodynamics within the region is needed.

In this study, Bayesian methods were applied to bridge current knowledge gaps by inferring the time-scaled phylogeny of SARS-CoV-2 strains and the spread patterns of this pandemic in Southeast Asia, a region relatively ignored by researchers. The results of this study provide valuable insight into the current pandemic and have the potential to inform future epidemic interventions at a regional level for Southeast Asia and beyond.

MATERIALS AND METHODS

A workflow schema was created to describe the key implementation steps and logic of this study (**Supplementary Figure 1**).

Sequence Data Acquisition and Filtering

SARS-CoV-2 genome sequences were downloaded from the GISAID database (https://www.gisaid.org/) on November 17, 2020. The following inclusion and exclusion criteria were applied to select samples for analysis: (1) they should originate from Southeast Asia (member states of the Association of Southeast Asian Nations and Timor-Leste); (2) only complete viral genomes from infected individuals (more than 29 kb in length) were included; (3) high-quality sequencing and assembly (the proportion of undetermined nucleotide bases <5%) was required; (4) sequences with unknown or incomplete sampling dates were excluded; (5) sequences with Vero cell labels were excluded. After filtering was done, the final dataset comprised 1,491 genomes of SARS-CoV-2 strains from 10 Southeast Asian countries (except Laos). These samples were obtained between January 5, 2020 and August 31, 2020. The distributions of the sequence samples included for analysis are summarized in **Supplementary Table 1**. In addition, the accession numbers and the XML file are listed as **Supplementary Files 1,2**, respectively.

Maximum-Likelihood (ML) Phylogenetic and Temporal Signal Analyses

Sequencing data were aligned using MAFFT v7.467 (Osaka University, Suita, Osaka, Japan) (25). An ML phylogenetic tree was constructed using the IQ-TREE v.2.0.6 software (Free Software Foundation, Boston, Massachusetts, USA) (26). The best-fit nucleotide substitution model (general time reversible with four gamma categories, GTR + G4) was selected according to the Bayesian information criterion (BIC) using the jModelTest v2.1.7 tool (Free Software Foundation, Boston, Massachusetts, USA) (27). A regression of root-to-tip genetic distance against sampling time was performed to evaluate the temporal signal in the dataset using the TempEst v1.5.3 tool (28). The estimated genetic distance was computed based on the maximum composite likelihood metric.

Time-Scaled Phylogenetic Analysis and Principal Component Analysis (PCA)

Time-scaled phylogenies for whole genomes were analyzed through Bayesian inference combined with Markov chain Monte Carlo (MCMC) methods using the BEAST v1.10.4 tool (29, 30). Bayesian phylogenetic analyses were performed based on the uncorrelated lognormal (UCLN) relaxed clock and three independent MCMC procedures. MCMC chains were run for 120 million steps, sampling every 300 steps from the posterior distribution. The convergence of all parameters was assessed by calculating effective sample sizes, setting their threshold at 200 using the Tracer v1.7.1 tool (30). After discarding the first 50 million steps as burn-in, 700,000 trees were combined using the LogCombiner v1.10.4 tool. The final tree was summarized as maximum clade credibility (MCC) tree with the TreeAnnotator v1.10.4 tool (both tools were used through the BEAST package). Node ages were annotated using median values. Geographic locations of the internal states on the tree were not observed and were treated as hidden states, which are associated with uncertainties. Branches were color-coded based on the locations of sampling for tips.

To assess the reliability of the major clades generated from the time-scaled phylogenetic tree, we performed PCA. Genome-wide variation data were extracted by the global pairwise sequence alignment of 1491 Southeast Asian SARS-CoV-2 genomes against the NC_045512 reference genome (downloaded from NCBI http://www.ncbi.nlm.nih.gov/) using EMBOSS Stretcher v.6.6.0 (European Bioinformatics Institute, Hinxton, Cambridgeshire, England) (31). PCA was performed using the PLINK v1.9 tool (32, 33), and the first three principal components (PCs) were plotted using the SPSS statistical software v26.0 (IBM Corp., Armonk, New York, USA).

Molecular Phylogeography and Demographic History Analyses

The two-way asymmetric Bayesian stochastic search variable selection (BSSVS) model was used to reconstruct the molecular phylogeography of SARS-CoV-2 within Southeast Asia. The Bayes factor (BF) was used to measure the likelihood of an epidemiological association between two sequence source countries (16). A BF value exceeding 10 (roughly similar to $P < 0.01$) indicates that the epidemiological association is strongly supported by the examined data, whereas a value less than 3 (roughly similar to $P > 0.05$) suggests that the epidemiological association is not statistically significant (16, 34, 35). Regarding demographic history, changes in effective population size (Ne) over time were inferred based on the coalescent theory using the Bayesian Skygrid approach. Both analyses were conducted using the BEAST v1.10.4 software.

To visually display the spatiotemporal spread of the pandemic in Southeast Asia during the study, a dynamic webpage was created using the Apache ECharts v4.1.0 tool (Apache Software Foundation, Forest Hill, Maryland, USA) (36).

RESULTS

ML Phylogeny and Temporal Structure

The ML phylogenetic tree based on complete genomes showed the molecular evolution of SARS-CoV-2 isolates from Southeast Asia (**Figure 1A**). Using root-to-tip regression analysis, a positive correlation between genetic divergence and sampling time was found; the retrieved R^2 value of 0.339 (F = 764.817; $P < 0.01$) suggested the presence of a temporal signal in the dataset (**Figure 1B**).

Time-Scaled Phylogenetic Reconstruction of SARS-CoV-2 in Southeast Asia

Phylogenetic relationships among 1,491 SARS-CoV-2 genomes were reconstructed using Bayesian methods and visualized in the form of an MCC tree (**Figure 2A**). The estimated root position and age were Singapore/Thailand (posterior probabilities 0.549 and 0.413, respectively) and November 28, 2019 [95% highest posterior density (HPD) interval: September 7, 2019 to January 4, 2020], respectively. Posterior probabilities for all Southeast Asian countries, as root positions, are presented in **Supplementary Table 2**. Furthermore, the evolutionary rate of SARS-CoV-2 isolated from Southeast Asia was estimated to be 1.446×10^{-3} (95% HPD: 1.292×10^{-3} to 1.613×10^{-3}) substitutions per site per year.

Four distinct phylogenetic clades (labeled I–IV) were identified, which were consistent with the topological structure of the ML phylogenetic tree. Clades I and II were relatively early in terms of divergence time, containing fewer sequences, with 160 and 119 sequences for Clades I and II, respectively. Clade III had 387 sequences and presented high genetic diversity with two separate subclades formed in mid-February 2020 and had the widest distribution across Southeast Asia. Of the sequences, 55.3% clustered in Clade IV (number of sequences: 825), most of which originated from Singapore, followed by Malaysia. All sequences from this region were relatively concentrated in one or two clades, except those originating from Singapore.

The first three PCs explained 46.9% of the total genomic variations among the samples. A three-dimensional scatter plot of PC1 (32.7%), PC2 (10.0%), and PC3 (4.2%) revealed a clustering pattern that was roughly the same as the branching of the MCC tree (**Figure 2B**).
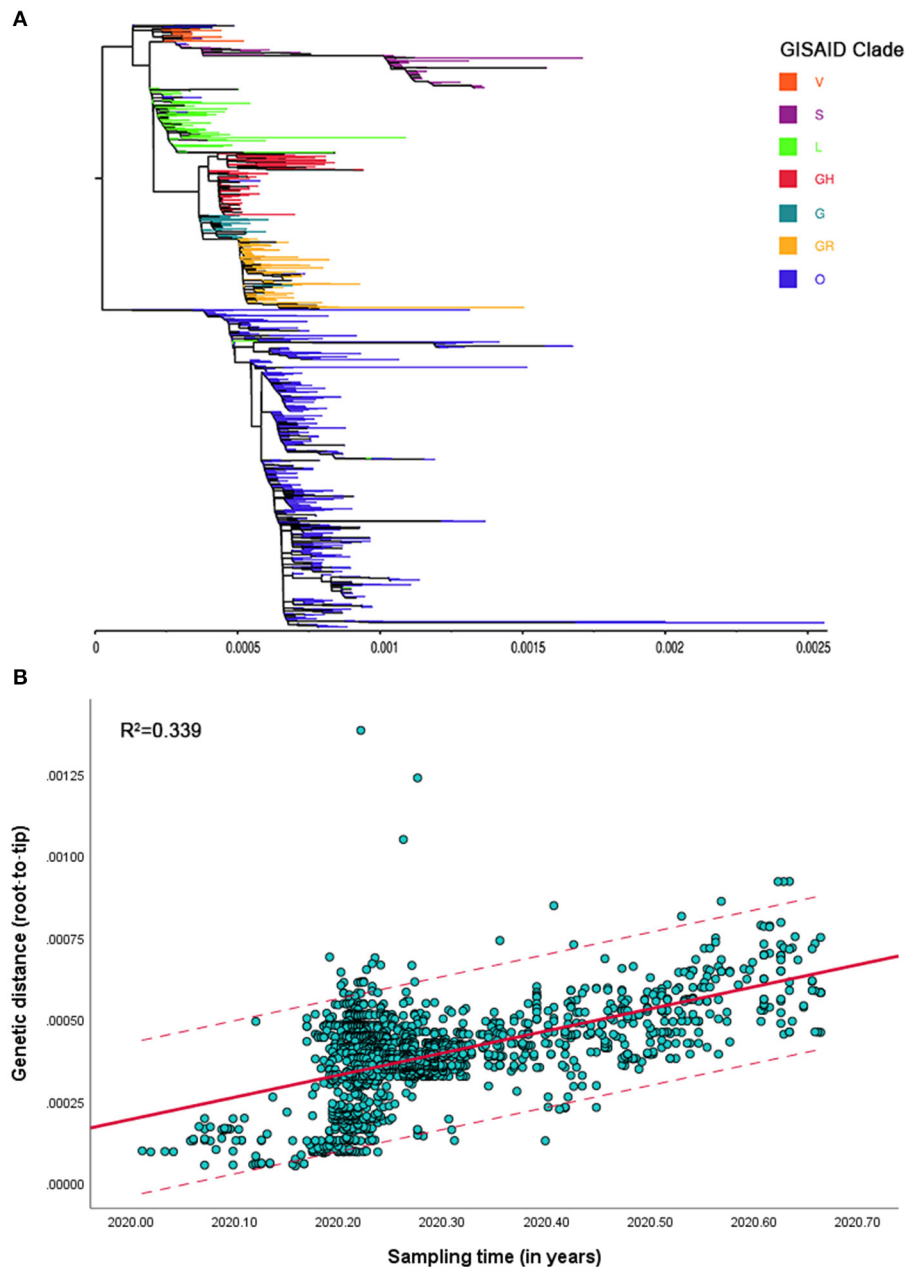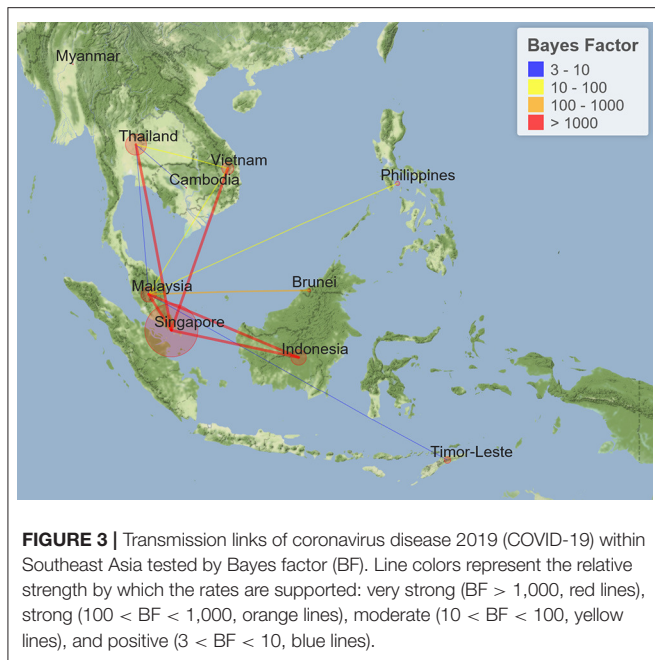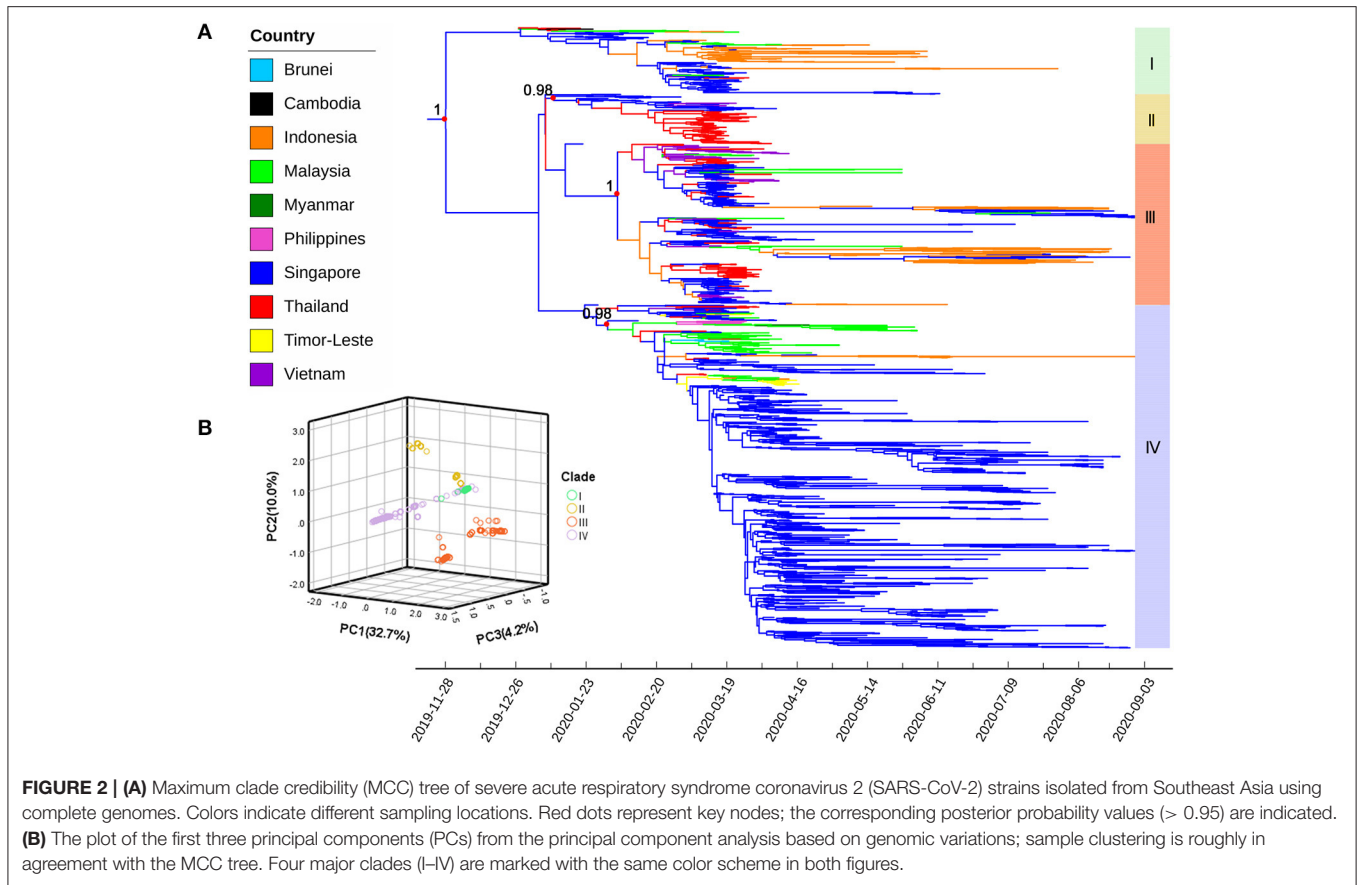
FIGURE 1 | (A) Maximum-likelihood (ML) phylogenetic tree of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) strains isolated from Southeast Asia using complete genomes. **(B)** Root-to-tip regression based on the ML phylogenetic tree; the solid red line and dotted lines represent the regression line and upper or lower limits of 95% confidence interval (CI), respectively.

Bayesian Phylogeographical Inference of SARS-CoV-2 in Southeast Asia

Probabilities of epidemiological associations among different sampling locations were calculated using the two-way asymmetric BSSVS; results are shown in **Figure 3**. During the study period, Singapore was estimated to have significant epidemiological associations with Thailand, Indonesia, Vietnam, and Malaysia, with BF values exceeding 1,000. In addition to neighboring Singapore, Malaysia was significantly associated with Indonesia, Brunei, the Philippines, Vietnam, Timor-Leste, and Thailand, with BF values ranging from 1,732 to 5. These associations make Malaysia the country with the highest number of transmission links in Southeast Asia. In mainland Southeast Asia (except Peninsular Malaysia), there were two estimated transmission links in the early stage: Thailand–Vietnam and Thailand–Cambodia, with BF values of 15 and 9, respectively.

FIGURE 2 | (A) Maximum clade credibility (MCC) tree of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) strains isolated from Southeast Asia using complete genomes. Colors indicate different sampling locations. Red dots represent key nodes; the corresponding posterior probability values (> 0.95) are indicated. **(B)** The plot of the first three principal components (PCs) from the principal component analysis based on genomic variations; sample clustering is roughly in agreement with the MCC tree. Four major clades (I–IV) are marked with the same color scheme in both figures.

FIGURE 3 | Transmission links of coronavirus disease 2019 (COVID-19) within Southeast Asia tested by Bayes factor (BF). Line colors represent the relative strength by which the rates are supported: very strong (BF > 1,000, red lines), strong (100 < BF < 1,000, orange lines), moderate (10 < BF < 100, yellow lines), and positive (3 < BF < 10, blue lines).

TABLE 1 | Two-way Bayes factor tests for SARS-CoV-2 transmission links in Southeast Asia.

Transmission link	Bayes factor*
Brunei-Malaysia	405
Cambodia-Thailand	9
Indonesia-Malaysia	1732
Indonesia-Singapore	652305
Malaysia-Philippines	14
Malaysia-Singapore	1088
Malaysia-Thailand	5
Malaysia-Timor-Leste	7
Malaysia-Vietnam	29
Singapore-Thailand	INFINITE
Singapore-Vietnam	203843
Thailand-Vietnam	15

*Only significant non-zero transmission links (Bayes factor > 3) are listed.

were visualized on a dynamic map (http://gos.mingleadgene.com/gosweb/virusdb/dataChartBk.html).

Population Dynamics of SARS-CoV-2 in Southeast Asia

The inferred effective population size trajectory of SARS-CoV-2 circulating in Southeast Asia is shown in the Bayesian

Significant transmission links with corresponding BF values are listed in **Table 1**. Additionally, time-varying transmission links
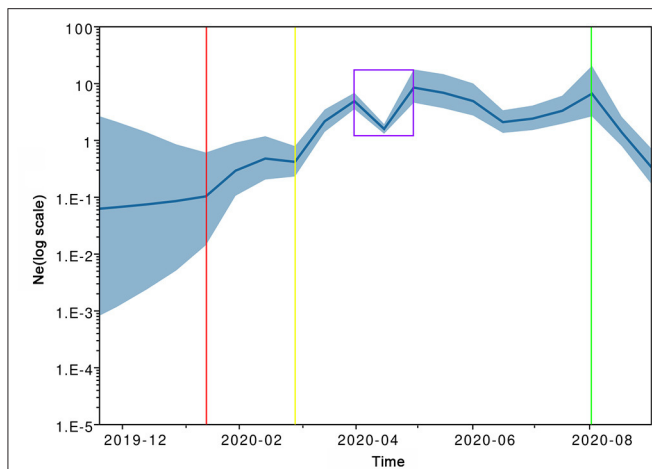
FIGURE 4 | Historical changes in effective population size (Ne) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) circulating in Southeast Asia by September 2020. The blue line represents posterior point estimates; the blue band represents 95% Bayesian credible interval (BCI). The red, yellow, and green vertical lines mark the watersheds corresponding to two accelerated expansions and one significant contraction; the purple box marks a temporary setback followed by a rapid rebound.

Skygrid plot (**Figure 4**). From the tMRCA to late March 2020, the effective population of the virus generally grew in a continuous manner and with two accelerations (around mid-January 2020 and late February 2020, respectively). In April 2020, the population size underwent a sudden fluctuation and hit a lower point in the middle of that month. Over the following 3 months, the viral population seemed to remain relatively steady. A pronounced population contraction took place in August 2020.

DISCUSSION

To the best of our knowledge, this is the first study to investigate the molecular phylogeny and spatiotemporal spread of SARS-CoV-2 in Southeast Asia using the Bayesian framework. Phylogenetic relationships among 1,491 SARS-CoV-2 isolates from 10 Southeast Asian countries were determined by analyzing their complete genomic sequences. Furthermore, the regional transmission dynamics and effective population size changes during the first 8 months of 2020 were illustrated.

The previous finding that the phylodynamic threshold of SARS-CoV-2 was attained by February 2, 2020, provided a reasonable basis for us to infer its evolutionary history through genomic analyses [37]. Although positive temporal signals had been detected in viral genome sequence data collected in the early stages, they were generally minor, with their corresponding R^2 (in root-to-tip regressions) values being smaller than 0.2 [37, 38]. The improved temporal structure of our dataset may be the consequence of sufficient accumulation of variations over a longer time span rather than an increase in the number of genome samples. Exact sampling dates were used, based on this feature, to calibrate the molecular clock.

Both Singapore and Thailand presented the highest posterior probabilities as root positions, with a slight gap and a sum exceeding 0.95. The first confirmed cases of COVID-19 in Southeast Asia were found in Thailand (January 13, 2020) and Singapore (January 23, 2020) [39]. More importantly, these two countries were among the first within the region where local transmission of the virus emerged (on January 31, 2020 and February 3, 2020, respectively). There was no direct correlation between relevant infection clusters [40, 41]. Thus, we speculate that the outbreak in this region probably originated in both places almost simultaneously, constituting a "dual origin" pattern. In the Western Hemisphere, Europe is another possible example of a within-region COVID-19 outbreak of dual origin. Evidence based on genomic epidemiology indicated that Italy and Germany might have been introduced to SARS-CoV-2 early and independently and seeded the epidemic in Europe [20, 42, 43]. Bayesian time-scaled phylogenetic analysis with the tip-dating method revealed that rootage or the time to the most recent common ancestor (tMRCA) converged to late November 2019, which is consistent with reported estimates based on analyses of sequences collected worldwide [37, 38]. However, tMRCA estimates varied regionally, as follows: mid-November 2019 for East Asia [18], early December 2019 for Africa [44], late December 2019 for Oceania and North America [20, 45], late January 2020 for Europe [20], and mid-February 2020 for South America [46]. Such regional differences reflect the non-synchronicity of viral introductions globally. In comparison, our estimate for Southeast Asia seems to be the earliest (except for East Asia), despite its relatively broad 95% HPD interval. Notably, SARS-CoV-2 related coronaviruses (SC2r-CoVs), which share high nucleotide identity with SARS-CoV-2, have been detected in *Rhinolophus* bats from Southeast Asia. Thus, Southeast Asia could be an important candidate for the origin of SARS-CoV-2 [47, 48]. The median evolutionary rate for SARS-CoV-2 isolates from Southeast Asia during the study was 1.446×10^{-3} substitutions per site per year, which is slightly higher than the results ($\sim 1.1 \times 10^{-3}$) reported in most studies [37, 38, 45, 49, 50]. It is unclear whether this subtle difference is related to lineage composition or climatic conditions. However, it was significantly higher than that reported for early African strains, with a median of 4.133×10^{-4} substitutions per site per year [44]. This difference seems to be largely attributable to the estimation bias caused by considering only early viral sequences with a small sample size in the African study.

The evolutionary branches in the phylogenetic tree are a manifestation of genetic diversity. However, determining distinct clades or subgroups is a major challenge in analyzing genetic datasets [51]. In this study, the MCC tree showed that SARS-CoV-2 strains circulating in Southeast Asia had evolved into at least four major clades, a finding supported by PCA results and sample points labeled with the major clades of the MCC tree clustered in the corresponding PCA plot. Notably, the first PC (PC1) from PCA explained nearly one-third of total genomic variation among samples, indicating limited genetic diversity in SARS-CoV-2 isolates from Southeast Asia. This finding was consistent with a global-level study and could be interpreted as the effect of pure genetic drift [52]. Most of the analyzed samples

were from Singapore and were incorporated into all clades. Singapore, an international financial center and free trade port, has a well-developed transportation network. Many studies have highlighted the role of travel (especially international air travel) in promoting the spatial transmission of SARS-CoV-2 (53, 54), which can be used to explain the diverse lineage composition of virus strains in Singapore. Notably, we found that the largest clade, namely Clade IV, appeared to have a continuous and relatively independent evolution centered in Singapore, despite its late divergence time. According to GISAID, up to 96.0% of sequences within Clade IV belong to GISAID-Clade O (O meaning "Others"), whereas 3.6% belong to GISAID-Clade V (55). Therefore, this clade likely represents a unique group of variants evolved from GISAID-Clade V and warrants further analysis. Most strains from Thailand were clustered in Clade II, which was evolutionarily older and roughly corresponded to GISAID-Clade S under the framework of GISAID. In the early days of the epidemic, Chinese scientists reported that GISAID-Clade S strains resulted from relatively conservative evolution and presented low infection rates (56). This is a potential explanation for why COVID-19 in Thailand was more effectively contained than in other countries (11, 40). A high concentration in one subclade of Clade III was observed for the strains from Vietnam, and all belonged to GISAID-Clade GR. GISAID-Clade GR, with the genetic marker S-D614G, was dominant in the WHO European Region (57), and the second wave of COVID-19 in Vietnam was triggered by a series of cases imported from Europe during March 2020, when approximately 85% of Vietnamese strains included in this study were sampled (58). The results of our study add circumstantial evidence to such a transmission link between Vietnam and Europe. Indonesia, the most populous country in Southeast Asia, is the world's largest island country and comprises more than 17,000 islands. Phylogenetic analysis revealed that its lineage composition was also very diverse, with the coexistence of both old and new strains, even if they were relatively concentrated in only two evolutionary clades (Clade I and a subclade of Clade III). This could be the result of spatially and temporally separate introduction events related to fragmented topography (59).

Quantitative and visualizing transmission links were further inferred by Bayesian phylogeographical analysis based on molecular associations among SARS-CoV-2 isolates from different countries in Southeast Asia. In maritime Southeast Asia, the high-support transmission links with BF values exceeding 1,000 in Singapore, Malaysia, and Indonesia have constituted a "transmission triangle" since March 2020. This epidemic pattern could be supported by the surveillance results of the Ministry of Health (MOH) Malaysia. As of August 27, 2020, the first and third countries of origin of imported COVID-19 cases in Malaysia were Indonesia and Singapore, respectively (60). Interestingly, a triangular cross-border regional development plan (the Indonesia–Malaysia–Singapore Growth Triangle, IMS-GT) has been in effect for more than 20 years, strengthening the integration of capital, land, and labor between these countries. Such a pattern of cross-border transmission is more likely to be attributable to certain negative externalities of the development of IMS-GT than simply geographical proximity

(61). Sri Petaling tabligh, a large religious gathering, took place in Malaysia at the end of February 2020, which subsequently triggered the widespread of COVID-19 in Southeast Asia (62). It was reported that the 4-month-long infection cluster eventually led to 2,550 Malaysian cases from 7 states and 825 non-Malaysian cases from 28 countries (63). At the molecular level, results from our phylogeographical analysis suggested that Malaysia possessed transmission links with seven other countries in the region, which is consistent with the colocalization of related isolates in the MCC tree. By combining macro- and micro-epidemiological evidence, Malaysia was inferred to be a potential regional epicenter during the early stages of the COVID-19 outbreak. In mainland Southeast Asia, two early transmission links in Thailand, Vietnam, and Cambodia were probably indirect, considering that most of their early cases were imported from China (40, 58, 64). Both Bayesian phylogenetic analysis and BF analysis revealed that the Thai strains were closely related to the Singaporean strains. Singapore is one of the largest destination countries for labor export from Thailand (65). According to the Royal Thai Ministry of Public Health (MOPH), some Thai nationals working in Singapore returned during the pandemic and subsequently tested positive for SARS-CoV-2 (66). Hereby, we speculated that the mobility of migrant workers could contribute in part to the Singapore–Thailand transmission link (67).

The effective population size represents the number of diverse genomes capable of producing progeny virions and indirectly reflects the number of infected individuals in the context of the epidemic (68, 69). The Bayesian Skygrid model is an improved coalescent-based model with higher accuracy in recovering population size trajectories than Bayesian Skyline and Skyride models (70). Using the Bayesian Skygrid model, we found that the SARS-CoV-2 circulating in Southeast Asia experienced a complex demographic history. In line with the recent origin of the virus, the initial effective population size of Southeast Asian SARS-CoV-2 was low (71). After a potential "adaptation period" of cross-species transmission (72), the virus experienced its first accelerated population expansion in mid-January 2020, when the first infection in Southeast Asia was confirmed (39). The second accelerated expansion of the viral population occurred in March 2020, matching the exponential growth pattern of reported cases during the same period (66). Epidemiological investigations confirmed that the early representative infection clusters in this region involved multiple mass gatherings (MGs) or superspreading events (SSEs), such as the SAFRA Jurong cluster in Singapore (involving a private dinner function held on February 15, 2020), Sri Petaling Tabligh cluster in Malaysia (involving an Islamic missionary convention held from February 27 to March 3, 2020) and Lumpinee Boxing Stadium cluster in Thailand (involving a Thai boxing match held on March 6, 2020) (8, 9, 40). These MG/SSE-related clusters were likely to be important external drivers of viral exponential expansion (73–75). It took Southeast Asian countries an average of 17 days to declare a state of emergency or lockdown after 50 cases were confirmed (76). Our reconstruction also showed that the population expansion of Southeast Asian SARS-CoV-2 suffered a temporary

setback in April 2020, and the corresponding 95% Bayesian credible interval was narrow. Accordingly, this phenomenon may reflect the efficacy of public health interventions (e.g., Movement Control Order of Malaysia effective from March 18, 2020, and the nationwide isolation order of Vietnam effective from April 1, 2020) (58, 63). However, the rapid rebound in effective population size not only highlighted the impact of shifting transmission dynamics and hidden reservoirs (68) but also underscored the importance of integrated and sustainable response strategies (53). The final population contraction, which was unexpected, possibly arose from bias due to the small sample size (18, 44).

This study has two basic limitations. First, the relatively small and unevenly distributed sample complicated the inference for the spatiotemporal patterns of viral spread in the region, as observed in all molecular epidemiological studies using real-world data (18). For example, only seven sequences were available from the Philippines, one of the worst-affected Southeast Asian nations, meaning more potential transmission links involving the Philippines remain unresolved in our model (9). We only selected the first 8 months of the epidemic to minimize sampling biases, as the research period based on the tradeoff between sampling coverage and sampling density (77, 78). Except for a few excluded sequences due to data quality control, we included nearly all the Southeast Asian SARS-CoV-2 sequences available then. Second, the virus importations from other regions may pose a potential threat to the validity of the research findings. As the COVID-19 pandemic continues to spread worldwide, it is unlikely that the SARS-CoV-2 strains circulating in any region are entirely indigenous. Unfortunately, we failed to adjust for this confounding factor directly because travel history was only available for four sequences in the dataset. However, the impact of the interregional importations of the virus seems to be relatively limited. Empirical evidence shows that compared with the long-distance spread across regions, SARS-CoV-2 is much easier to spread between neighboring countries over short distances (22, 79, 80). For Southeast Asia, international travel bans were announced at a very early stage of the pandemic (76, 81), but movement restrictions within the region, the focus of this research, were relatively loose (81). Indeed, genomic epidemiology is an important complement to traditional epidemiological investigations rather than a substitute (13). Therefore, we combined the results with macro-epidemiological statistics to a large extent and adopted a cautious approach throughout the inference process. Further investigations into the relationship between the molecular evolution of SARS-CoV-2 and its spread patterns are warranted using a more representative sample and in a broader context.

CONCLUSIONS

In conclusion, this study provides insights into the genomic epidemiology of SARS-CoV-2 in Southeast Asia, a tropical region with limited resources. The most recent common ancestor of Southeast Asian strains was estimated to have emerged in Singapore and/or Thailand since late November 2019. Our phylogenetic analysis revealed an average evolutionary rate of 1.446×10^{-3} substitutions per site per year and at least four major evolutionary branches (one of which was potentially endemic). Moreover, the spatial spread of SARS-CoV-2 within this region was characterized by a potential epicenter (Malaysia) and an inferred "transmission triangle" (Indonesia–Malaysia–Singapore). The population dynamics during the first 8 months presented a fluctuating process with early expansion. These findings have significant implications for public health interventions. Continuous-genomic surveillance and enhanced strategic collaboration should be listed as priorities in combatting the ongoing pandemic, especially for regional communities dominated by developing countries.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.685315/full#supplementary-material

Supplementary Figure 1 | Workflow schema describing key implementation steps and the logic. Steps connected by solid lines constitute the main technical route of the study, whereas those connected by dotted lines represent auxiliary or verification steps.

Supplementary Table 1 | Distributions of the SARS-CoV-2 sequence samples included in this study.

Supplementary Table 2 | Root probability and outbreak date for each country in Southeast Asia.

REFERENCES

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med.* (2020) 382:727–33. doi: 10.1056/NEJMoa2001017

2. Chan JF, Yuan S, Kok KH, To KK, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet.* (2020) 395:514–23. doi: 10.1016/S0140-6736(20)30154-9

3. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature.* (2020) 579:270–73. doi: 10.1038/s41586-020-2012-7

4. 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) 583:282–5. doi: 10.1038/s41586-020-2169-0

5. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol.* (2020) 5:536–44. doi: 10.1038/s41564-020-0695-z

6. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet.* (2020) 395:565–74. doi: 10.1016/S0140-6736(20)30251-8

7. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, et al. Cell entry mechanisms of SARS-CoV-2. *Proc Natl Acad Sci U S A.* (2020) 117:11727–34. doi: 10.1073/pnas.2003138117

8. World Health Organization. *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020.* (2020). Available online at: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-\$-\$11-march-2020 (accessed February 10, 2021)

9. World Health Organization. *Coronavirus disease (COVID-19) Weekly Epidemiological Update.* (2020). Available online at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200831-weekly-epi-update-3.pdf?sfvrsn=d7032a2a_4 (accessed February 10, 2021)

10. World Health Organization. Pulse Survey on Continuity of Essential Health Services During the Covid-19 Pandemic: Interim Report, 27 August 2020. (2020). Available online at: https://www.who.int/publications/i/item/WHO-2019-nCoV-EHS_continuity-survey-2020.1 (accessed February 10, 2021)

11. Association of Southeast Asian Nations. *ASEAN Rapid Assessment: The Impact of COVID-19 on Livelihoods across ASEAN.* (2020). Available online at: https://asean.org/?static_post=asean-rapid-assessment-impact-covid-19-livelihoods-across-asean (accessed February 10, 2021)

12. Zhu M, Jirapat K, Shen J, Chew SP, Tan JW, Latthitham N, et al. Early epidemiological features and trends of the COVID-19 outbreak in southeast asia: a population-level observational study. *Res Square [Preprint].* (2020). doi: 10.21203/rs.3.rs-34707/v2

13. Tang P, Croxen MA, Hasan MR, Hsiao WW, Hoang LM. Infection control in the new age of genomic epidemiology. *Am J Infect Control.* (2017) 45:170–9. doi: 10.1016/j.ajic.2016.05.015

14. Martin MA, VanInsberghe D, Koelle K. Insights from SARS-CoV-2 sequences. *Science.* (2021) 371:466–7. doi: 10.1126/science.abf3995

15. Volz EM, Koelle K, Bedford T. Viral phylodynamics. *PLoS Comput Biol.* (2013) 9:e1002947. doi: 10.1371/journal.pcbi.1002947

16. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput Biol.* (2009) 5:e1000520. doi: 10.1371/journal.pcbi.1000520

17. Li J, Li Z, Cui X, Wu C. Bayesian phylodynamic inference on the temporal evolution and global transmission of SARS-CoV-2. *J Infect.* (2020) 81:318–56. doi: 10.1016/j.jinf.2020.04.016

18. Nie Q, Li X, Chen W, Liu D, Chen Y, Li H, et al. Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Res.* (2020) 287:198098. doi: 10.1016/j.virusres.2020.198098

19. Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H, et al. Genomic epidemiology of SARS-CoV-2 in Guangdong province, China. *Cell.* (2020) 181:997–1003.e9. doi: 10.1016/j.cell.2020.04.023

20. Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, et al. The emergence of SARS-CoV-2 in Europe and North America. *Science.* (2020) 370:564–70. doi: 10.1126/science.abc8169

21. du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science.* (2021) 371:708–12. doi: 10.1126/science.abf2946

22. Gómez-Carballa A, Bello X, Pardo-Seco J, Martinón-Torres F, Salas A. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. *Genome Res.* (2020) 30:1434–48. doi: 10.1101/gr.266221.120

23. Yap PSX, Tan TS, Chan YF, Tee KK, Kamarulzaman A, Teh CSJ. An overview of the genetic variations of the SARS-CoV-2 genomes isolated in southeast asian countries. *J Microbiol Biotechnol.* (2020) 30:962–6. doi: 10.4014/jmb.2006.06009

24. Islam OK, Al-Emran HM, Hasan MS, Anwar A, Jahid MIK, Hossain MA. Emergence of European and North American mutant variants of SARS-CoV-2 in South-East Asia. *Transbound Emerg Dis.* (2021) 68:824–32. doi: 10.1111/tbed.13748

25. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* (2013) 30:772–80. doi: 10.1093/molbev/mst010

26. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. Corrigendum to: IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol.* (2020) 37:2461. doi: 10.1093/molbev/msaa131

27. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods.* (2012) 9:772. doi: 10.1038/nmeth.2109

28. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* (2016) 2:vew007. doi: 10.1093/ve/vew007

29. Alekseyenko A, Bedford T, Bielejec F, Bloomquist E, Carvalho L, Cybis G, et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* (2018) 4:vey016. doi: 10.1093/ve/vey016

30. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst Biol.* (2018) 67:901–4. doi: 10.1093/sysbio/syy032

31. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* (2019) 47:W636–41. doi: 10.1093/nar/gkz268

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* (2007) 81:559–75. doi: 10.1086/519795

33. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* (2015) 4:7. doi: 10.1186/s13742-015-0047-8

34. Su YCF, Bahl J, Joseph U, Butt KM, Peck HA, Koay ESC, et al. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nat Commun.* (2015) 6:7952. doi: 10.1038/ncomms8952

35. Keysers C, Gazzola V, Wagenmakers EJ. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nat Neurosci.* (2020) 23:788–99. doi: 10.1038/s41593-020-0660-4

36. Li D, Mei H, Shen Y, Su S, Zhang W, Wang J, et al. ECharts: A declarative framework for rapid construction of web-based visualization. *Visual Informatics.* (2018) 2:136–46. doi: 10.1016/j.visinf.2018.04.011

37. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* (2020) 6:veaa061. doi: 10.1093/ve/veaa061

38. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J Med Virol.* (2020) 92:602–11. doi: 10.1002/jmv.25731

39. Association of Southeast Asian Nations. *Risk Assessment for International Dissemination of COVID-19 to the ASEAN Region.* 2021. Available online at: https://aseanphe.org/wp-content/uploads/2021/02/Risk-Assessment-Report-1-February-2021.pdf (accessed May 10, 2021)

40. Kitphati R, Pratedrat P, Puenpa J, Chansaenroj J, Yorsaeng R, Sudhinaraset N, et al. Infection control measures of COVID-19 in Thailand: one year experience. *Southeast Asian J Trop Med Public Health.* (2021) 52:300–13. Available online at: https://journal.seameotropmednetwork.org/index.php/jtropmed/article/view/430

41. Pung R, Chiew CJ, Young BE, Chin S, Chen MI, Clapham HE, et al. Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *Lancet.* (2020) 395:1039–46. doi: 10.1016/S0140-6736(20)30528-6

42. Nadeau SA, Vaughan TG, Scire J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. *Proc Natl Acad Sci USA.* (2021) 118. doi: 10.1073/pnas.2012008118

43. Zhukova A, Blassel L, Lemoine F, Morel M, Voznica J, Gascuel O. Origin, evolution and global spread of SARS-CoV-2. *C R Biol.* (2020) 1:57–75. doi: 10.5802/crbiol.29

44. Motayo BO, Oluwasemowo OO, Olusola BA, Akinduti PA, Arege OT, Obafemi YD, et al. Evolution and genetic diversity of SARS-CoV-2 in Africa using whole genome sequences. *Int J Infect Dis.* (2021) 103:282–7. doi: 10.1016/j.ijid.2020.11.190

45. Seemann T, Lane CR, Sherry NL, Duchene S, Gonçalves da Silva A, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. *Nat Commun.* (2020) 11:4376. doi: 10.1038/s41467-020-18314-x

46. Xavier J, Giovanetti M, Adelino T, Fonseca V, Barbosa da Costa AV, Ribeiro AA, et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. *Emerg Microbes Infect.* (2020) 9:1824–34. doi: 10.1101/2020.05.05.20091611

47. Latinne A, Hu B, Olival KJ, Zhu G, Zhang L, Li H, et al. Origin and cross-species transmission of bat coronaviruses in China. *Nat Commun.* (2020) 11:4235. doi: 10.1038/s41467-020-17687-3

48. Wacharapluesadee S, Tan CW, Maneeorn P, Duengkae P, Zhu F, Joyjinda Y, et al. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat Commun.* (2021) 12:972. doi: 10.1038/s41467-021-21768-2

49. Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ.* (2020) 98:495–504. doi: 10.2471/BLT.20.253591

50. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science.* (2020) 369:1255–60. doi: 10.1126/science.abd2161.72

51. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* (2006) 2:e190. doi: 10.1371/journal.pgen.0020190

52. Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc Natl Acad Sci USA.* (2020) 117:23652–62. doi: 10.1073/pnas.2008281117

53. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science.* (2020) 368:395–400. doi: 10.1126/science.aba9757

54. Bielecki M, Patel D, Hinkelbein J, Komorowski M, Kester J, Ebrahim S, et al. Air travel and COVID-19 prevention in the pandemic and peri-pandemic period: a narrative review. *Travel Med Infect Dis.* (2021) 39:101915. doi: 10.1016/j.tmaid.2020.101915

55. Global Initiative on Sharing Avian Influenza Data. Clade And Lineage Nomenclature Aids in Genomic Epidemiology Studies of Active Hcov-19 Viruses. (2021). Available online at: https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/ (accessed May 10, 2021)

56. Tang XL, Wu CC, Li X, Song YH, Yao XM, Wu XK, et al. On the origin and continuing evolution of SARS-CoV-2. *Nat Sci Rev.* (2020) 7:1012–23. doi: 10.1093/nsr/nwaa036

57. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro Surveill.* (2020) 25:2001410. doi: 10.2807/1560-7917.ES.2020.25.32.2001410

58. Thai PQ, Rabaa MA, Luong DH, Tan DQ, Quang TD, Quach HL, et al. The first 100 days of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) control in Vietnam. *Clin Infect Dis.* (2021) 72:e334–42. doi: 10.1101/2020.05.12.20099242

59. Crump JA, Murdoch DR, Baker MG. Emerging infectious diseases in an island ecosystem: the New Zealand perspective. *Emerg Infect Dis.* (2001) 7:767–72. doi: 10.3201/eid0705.017501

60. Ministry of Health Malaysia. *Kenyataan Akhbar KPK 28 August 2020 – Situasi Semasa Jangkitan Penyakit Coronavirus 2019 (COVID-19) di Malaysia.* Putrajaya: Ministry of Health Malaysia. (2020). Available online at: https://kpkesihatan.com/2020/08/28/kenyataan-akhbar-kpk-28-ogos-2020-situasi-semasa-jangkitan-penyakit-coronavirus-2019-covid-19-di-malaysia/ (accessed May 10, 2021)

61. Sparke M, Sidaway JD, Bunnell T, Grundy-Warr C. Triangulating the borderless world: geographies of power in the Indonesia-Malaysia-Singapore growth Triangle. *Trans Inst Br Geogr.* (2004) 29:485–98. doi: 10.1111/j.0020-2754.2004.00143.x

62. Che Mat NF, Edinur HA, Abdul Razab MKA, Safuan S. A single mass gathering resulted in massive transmission of COVID-19 infections in Malaysia with further international spread. *J Travel Med.* (2020) 27. doi: 10.1093/jtm/taaa059

63. Hashim JH, Adman MA, Hashim Z, Mohd Radi MF, Kwan SC. COVID-19 epidemic in Malaysia: epidemic progression, challenges, and response. *Front Public Health.* (2021) 9:560592. doi: 10.3389/fpubh.2021.560592

64. WHO Cambodia Representative Office. *The First 100 Days of the Covid-19 Response: Past Investments in Health Security System Pay Off, and Learning Lessons for the Future.* (2020). Available online at: https://www.who.int/cambodia/news/feature-stories/detail/the-first-100-days-of-the-covid-19-response-past-investments-in-health-security-system-pay-off-and-learning-lessons-for-the-future (accessed February 10, 2021)

65. International Labour Organization. *Migration in ASEAN in figures: The International Labour Migration Statistics (ILMS) Database in ASEAN.* (2015). Available online at: https://apmigration.ilo.org/resources/migration-in-asean-in-figures/ (accessed May 10, 2021)

66. Ministry of Public Health Thailand. *Press Statement: Updates on the Coronavirus Diseases 2019 (COVID-19) in Thailand.* (2020). Available online at: https://ddc.moph.go.th/viralpneumonia/file/news/news_red200_150763.pdf (accessed May 10, 2021)

67. Koh D. Migrant workers and COVID-19. *Occup Environ Med.* (2020) 77:634–6. doi: 10.1136/oemed-2020-106626

68. Giovanetti M, Cella E, Benedetti F, Rife Magalis B, Fonseca V, Fabris S, et al. SARS-CoV-2 shifting transmission dynamics and hidden reservoirs potentially limit efficacy of public health interventions in Italy. *Commun Biol.* (2021) 4:489. doi: 10.1038/s42003-021-02025-0

69. Frost SD, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. *Philos Trans R Soc Lond B Biol Sci.* (2010) 365:1879–90. doi: 10.1098/rstb.2010.0060

70. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol.* (2013) 30:713–24. doi: 10.1093/molbev/mss265

71. Chaw SM, Tai JH, Chen SL, Hsieh CH, Chang SY, Yeh SH, et al. The origin and underlying driving forces of the SARS-CoV-2 outbreak. *J Biomed Sci.* (2020) 27:73. doi: 10.1186/s12929-020-00665-8

72. Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, et al. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev.* (2008) 72:457–70. doi: 10.1128/MMBR.00004-08

73. McCloskey B, Zumla A, Ippolito G, Blumberg L, Arbon P, Cicero A, et al. Mass gathering events and reducing further global spread of COVID-19: a political and public health dilemma. *Lancet.* (2020) 395:1096–9. doi: 10.1016/S0140-6736(20)30681-4

74. Frieden TR, Lee CT. Identifying and interrupting superspreading events-implications for control of severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis.* (2020) 26:1059–66. doi: 10.3201/eid2606.200495

75. Kumar S, Jha S, Rai SK. Significance of super spreader events in COVID-19. *Indian J Public Health.* (2020) 64:S139–41. doi: 10.4103/ijph.IJPH_495_20

76. UN. *Policy Brief: The Impact of COVID-19 on South-East Asia.* 2020. Available online at: https://www.un.org/sites/un2.un.org/files/sg_policy_brief_covid_south-east_asia_30_july_2020.pdf (accessed May 10, 2021)

77. Adam DC, Scotch M, MacIntyre CR. Phylodynamics of influenza A/H1N1pdm09 in India reveals circulation patterns and increased selection for clade 6b residues and other high mortality mutants. *Viruses.* (2019) 11:791. doi: 10.3390/v11090791

78. Murray M. Sampling bias in the molecular epidemiology of tuberculosis. *Emerg Infect Dis.* (2002) 8:363–9. doi: 10.3201/eid0804.000444

79. Lee EC, Wada NI, Grabowski MK, Gurley ES, Lessler J. The engines of SARS-CoV-2 spread. *Science.* (2020) 370:406–7. doi: 10.1126/science.abd8755

80. Paniz-Mondolfi A, Muñoz M, Florez C, Gomez S, Rico A, Pardo L, et al. SARS-CoV-2 spread across the Colombian-Venezuelan border. *Infect Genet Evol.* (2020) 86:104616. doi: 10.1016/j.meegid.2020.104616

81. Minh Van H, Phuong Thi Ngoc N, Thao Thi Phuong T, Long Quynh K, An Thi Minh D, Huy Van N, et al. The COVID-19 pandemic in the ASEAN: a preliminary report on the spread, burden and medical capacities. *Asian Pac J Trop Med.* (2020) 13:247–51. Available online at: https://collections.unu.edu/view/UNU:7729