The coronavirus disease of 2019, COVID-19, has been causing devastating economic losses and mental health stress to societies around the world. An outbreak in the central region of China proceeded to spread to the rest of mainland China and its neighboring regions. Three locales of Japan, South Korea, and Taiwan, mainland China’s neighbors to the east, each reported their index cases of COVID-19 around mid-January 2020. The etiological agent of the severe pneumonic disease is a human-infecting coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 genome sequences that are collected at a high rate are available at the Global Initiative on Sharing All Influenza Data (GISAID) to help the fight against COVID-19. Here, we detect phylogenetic and epidemiological signals in viral genomes sampled from COVID-19 patients in the three locales of East Asia using phylodynamic models implemented in a Bayesian statistical method of BEAST2 software. We estimate and compare the dates of epidemic origin and the basic reproductive number for South Korea was greater than one.

**Keywords:** Bayesian inference, COVID-19, genomic epidemiology, Markov chain Monte Carlo, SARS-CoV-2

In December 2019, patients with unusual pneumonia were hospitalized near a seafood wholesale market in Wuhan, a city of 11 million people, in China (Zhu et al., 2020). Most of these patients suffered from symptoms of respiratory illness including cough, shortness of breath, fever, chill, muscle pain, loss of taste or smell, and fatigue (Zhou et al., 2020). People with underlying medical conditions were reported to be more vulnerable than others to the newly emerging pneumonic disease (Prem et al., 2020).

On December 31, 2019, the existence of these cases was acknowledged by Chinese health authorities including the Chinese Center of Disease Control and Prevention (China CDC), and some of the patients were claimed to be epidemiologically associated with the seafood wholesale market, which was closed on January 1, 2020 (Li et al., 2020). On February 11, 2020, the World Health Organization (WHO) named the contagious disease COVID-19, an acronym for “Coronavirus Disease of 2019.” The infectious disease has been, since then, causing devastating economic losses (Ayiittey et al., 2020) and mental health stress (Bao et al., 2020) to societies around the world.
The initial outbreak in Wuhan proceeded to spread to the rest of mainland China during the first two months of 2020. It was quickly propagated from mainland China to its neighbors in East and South-East Asia. In the midst of the peak of the epidemic in Asia, confirmed cases and deaths due to the infectious disease began to rise in Europe and America.

By mid-May 2020, many countries were considering lifting their strict social distancing measures because of economic concerns. Some health officials warned the public of a potential second wave of COVID-19 in the coming fall of 2020. As of June 29, 2020, more than ten million people had tested positive for the coronavirus disease and nearly five hundred thousand people had died from COVID-19 (World Health Organization, 2020).

In early January, the China CDC announced a new coronavirus as the etiological agent of the severe pneumonic disease (Li et al., 2020). In mid-January 2020, the coronavirus responsible for COVID-19 was provisionally named 2019-nCoV or 2019 novel coronavirus (Zhu et al., 2020). Later, in February 2020, the International Committee on Taxonomy of Viruses (ICTV) renamed the virus severe acute respiratory syndrome coronavirus 2 or SARS-CoV-2 because of its phylogenetic proximity to SARS-CoV that had caused outbreaks in China and a few neighboring countries nearly two decades earlier (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020).

Japan, South Korea, and Taiwan, mainland China’s neighbors to the east, each reported their index cases of COVID-19 around mid-January 2020. However, the COVID-19 situations in these three places varied in the number of confirmed, recovered, and fatal cases (May 9, 2020, accessed https://covid19.who.int). For example, confirmed cases of COVID-19 in Japan and Taiwan increased very slowly until mid-March 2020. The number of confirmed cases in South Korea until February 17, 2020 was only 30 and was similar to that for Japan after accounting for the population sizes of the two locales. However, after mid-February 2020, the number of confirmed cases in South Korea surged due to the so-called the crisis of Shincheonji Church. For a while, South Korea ranked second in the number of confirmed cases of COVID-19 among all of the countries of the world (Choe, 2020).

SARS-CoV-2 genome sequences are being collected at a high rate. As of May 9, 2020, only four months after the announcement of the first SARS-CoV-2 genome, more than 17 thousand SARS-CoV-2 genomes have been sequenced. Most of them are available at the Global Initiative on Sharing All Influenza Data (Shu and McCauley, 2017; GISAID, https://www.gisaid.org). The GISAID database has been instrumental in the fight against the coronavirus pandemic. Applications of the SARS-CoV-2 genome sequence data span a range of research topics, including the classification of SARS-CoV-2 (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020), the comparative genomic analysis of SARS-CoV and SARS-CoV-2 (Xu et al., 2020), the origin of SARS-CoV-2 (Andersen et al., 2020), phylogenetic network analyses (Forster et al., 2020), the proposal (Ji et al., 2020) or rejection (Paraskevis et al., 2020) of recombination as an evolutionary force shaping the SARS-CoV-2 genome, and phylodynamic analyses (Hadfield et al., 2018; Lai et al., 2020).

Phylodynamics, the intersection of epidemiology and phylogenetics, uses genetic sequences of infectious agents such as viruses or sometimes bacteria (Kühnert et al., 2018) to make epidemiological inferences (Grenfell et al., 2004). Recent statistical advances (Suchard et al., 2018) and high-throughput genome sequencing of viral infectious agents (Gire et al., 2014) are helping to make phylodynamic inferences an important supplement to traditional epidemiological techniques. The resulting improvement could facilitate more informed public health decisions regarding contagious diseases.

One of the epidemiologically important parameters in phylodynamics is the basic reproductive number, which is the expected number of infections to which each newly infected individual will give rise in a fully susceptible population. The epidemic is expected to die out if the basic reproductive number is less than 1 while it is likely to spread if the number is greater than 1. Knowledge of the basic reproduction number informs public health decisions such as what proportion of the population needs to be vaccinated.

Here, we analyze viral genomes isolated from COVID-19 patients in Japan, South Korea, and Taiwan. We compare inferences from these three locales of East Asia. We focus on estimating and comparing the dates of epidemic origin and the basic reproductive numbers. We also explore the sensitivity of our Bayesian inferences to the number and sampled distribution of viral genomes.
Materials and Methods

Data preparation regarding genome alignment is described in SARS-CoV-2 genome data. Subsampling of genome sequences for phylodynamic analysis is given in the subsection Data preparation for phylodynamic analyses. Bayesian phylodynamic analyses are detailed in the subsection Phylodynamic analyses.

SARS-CoV-2 genome data

A dataset of 32,198 SARS-CoV-2 sequences was downloaded on June 20, 2020 along with information of geographical sampling location and collection dates from the GISAID database (https://www.gisaid.org). A reference SARS-CoV-2 genome sequence (NCBI accession: MN908947.3) was added to the sequence dataset in order to utilize genome annotation to later refine the alignment between the genomes. Three steps of sequence filtering of the dataset were conducted. First, sequences with no precise sample collection dates were removed. Second, sequences with ambiguous nucleotides were excluded for the purpose of quality control. Third, sequences that were shorter than 29 kb were filtered out because the reference SARS-CoV-2 genome was substantially longer.

The remaining sequences were aligned using the MAFFT software (Katoh and Standley, 2013). Two refinement steps with the alignment were followed. First, the 5'- and 3'-untranslated regions according to the reference genome were trimmed from the alignment because the 5'- and 3'-ends of the genome alignment substantially varied in length. Second, alignment columns representing insertions relative to the reference sequence were removed because such columns were deemed particularly likely to be sequencing artefacts.

The resulting sequence alignment includes 8,510 time-stamped viral isolates with alignment length of 29,409 bp and will be referred to as the ALL8510 data set. The aligned ALL8510 data was utilized to infer a phylogenetic tree using IQ-TREE (Minh et al., 2020).

Data preparation for phylodynamic analyses

Three locales in East Asia were considered for a phylodynamic analysis: Japan, South Korea, and Taiwan. Among the ALL8510 genomes, the oldest dates of sample collection for both Japan and Taiwan were January 23, 2020. The corresponding oldest date for South Korea was January 25, 2020. The index cases for Japan, South Korea, and Taiwan were reported January 16, 20, and 21, 2020, respectively. The most recent dates of sample collection for Japan and Taiwan were May 7 and April 27, 2020, respectively. However, the most recent date for a South Korean isolate was February 27, 2020. To infer the rate of evolution and the epidemic origin dates, all of the available data were used for the three locales: 96 sequences for Japan, 30 for South Korea, and 96 for Taiwan.

To compare epidemiological parameters from the three locales, we focused on viral data sampled in the same time period. Because the most recently isolated South Korea virus was isolation on February 27, 2020 whereas the most recent isolates from both Japan and Taiwan were later than that, our phylodynamic analyses of each of the three locals considered only sequences isolation on February 27, 2020 or before. Subject to this constraint, the SARS-CoV-2 genome alignment contained 72 sequences from Japan, 30 from South Korea, and 14 from Taiwan.

Comparison of parameters such as rate of evolution, the epidemic origin dates, and the basic reproductive number among the locales needs cautious interpretations because the numbers of sequences from the three locales were different. Parameter uncertainties will tend to be larger for locales with smaller sample sizes. Phylodynamic models employed in the current study are relatively parameter-rich. This means the number of sequences per locale could be too small for the data to provide much information. This is especially the case for Taiwan where only 14 viral isolates are available that meet our criteria. For Bayesian analyses, prior and posterior distributions are expected to be quite similar when the data have little information. In our phylodynamic analyses, we carefully contrast prior and posterior distributions. We also form a fourth data set with 116 sequences by pooling the individual data sets from the three locales. We will refer to this data set as the JKT116 data set. We contrast and compare the inferences from the JKT116 data to the results from the three individual locales.

To evaluate the sensitivity of the analysis to phylodynamic sampling, we used the JKT116 data as the basis. Random sampling from the ALL8510 alignment generated 100 data sets, each of which matched the same number of sequences and sample collection date range as the pooled data set. We refer to
these 100 data sets as the SUBSAMPLE data sets. Because the SUBSAMPLE data sets would be collected under a sampling scheme that differed from that of the JKT116 data set, the epidemiological parameters corresponding to the SUBSAMPLE data sets would be distributed differently from that of the JKT116 data. If the phylodynamic analysis was sensitive enough to detect the epidemiological signal in the JKT116 data, the analysis could capture different signals in the SUBSAMPLE data sets.

Phylodynamic analyses

The prepared data sets were analyzed with the Bayesian statistical framework implemented in version 2.6.2 of the BEAST2 software (Bouckaert et al., 2019). This software was employed to estimate the rate of viral evolution, the dates of epidemic origin, and the basic reproductive number (Stadler et al., 2013; Kühnert et al., 2014). The BEAST2 inference procedure included three phylodynamic models described in BEAST2 as tree priors: the exponential growth model as “Coalescent Exponential Population” prior (Stadler et al., 2012), the constant birth-death model as “Birth Death Skyline Serial” prior (Boskova et al., 2014), and the birth-death SIR model as “Phylodynamic: Birth Death SIR (serial)” prior (Kühnert et al., 2014). Here, the three models are denoted CE, BD, and BDSIR. The SUBSAMPLE data sets for evaluating the sensitivity of phylodynamic analyses were analyzed only with the BD model. This is because the CE model does not directly incorporate the basic reproductive number while the BDSIR model was computationally too intensive to get a satisfactory estimate of the posterior distribution for some data sets. Model parameters and the prior distributions of those parameters are summarized in Table 1. Each analysis used the HKY (Hasegawa-Kishino-Yano) nucleotide substitution model (Hasegawa et al., 1985) with empirically estimated nucleotide frequencies and with the strict molecular clock.

With the exception of the SUBSAMPLE data sets, three independent Markov chain Monte Carlo (MCMC) chains were run for each combination of model and data set. To ensure convergence of the chains, Tracer version 1.7.1 (Rambaut et al., 2018) was employed. MCMC chains for models CE and BD were run for 10 million generations with sampling of every 1,000 generations while chains for model BDSIR were run for 100 million generations with sampling of every 10,000 generations. Each chain thereby resulted in 10,000 MCMC samples and the initial 10% of these samples was treated as a burn-in period and excluded from posterior distribution approximations. Because MCMC runs for JKT116 data set with model BD needed longer chains, they used the same chain length for model BDSIR. Combining the three independent chains, the effective sample size was greater than 200 for all of the parameters.

For the 100 SUBSAMPLE data sets that were analyzed with the BD model, two (rather than three) independent MCMC chains were run. These two chains were run for 100 million generations with sampling of every 10,000 generations, leaving a sample of size 10,000 from which 10% was removed as a burn-in period. When estimating the basic reproductive number with the models of BD and BDSIR, we also approximated the prior distribution of the parameter of the basic reproductive number to compare the posterior and prior estimates. We refer to

---

Table 1. Prior distributions for phylodynamic models

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model*</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>CE/BD/BDSIR</td>
<td>lognormal (1, 1.25)</td>
</tr>
<tr>
<td>Clock rate</td>
<td>CE/BD/BDSIR</td>
<td>uniform (0, infinity)</td>
</tr>
<tr>
<td>Grow rate</td>
<td>CE</td>
<td>Laplace (mu = 0.001, scale = 30.7)</td>
</tr>
<tr>
<td>Origin of epidemic</td>
<td>BDSIR</td>
<td>uniform (0,1000)</td>
</tr>
<tr>
<td>Origin of epidemic</td>
<td>BD</td>
<td>uniform (0, infinity)</td>
</tr>
<tr>
<td>Rate of become noninfectious</td>
<td>BD/BDSIR</td>
<td>uniform (0,1000)</td>
</tr>
<tr>
<td>Sampling proportion</td>
<td>BD/BDSIR</td>
<td>beta (1,1)</td>
</tr>
<tr>
<td>Basic reproductive number</td>
<td>BD</td>
<td>lognormal (0, 1)</td>
</tr>
<tr>
<td>Basic reproductive number</td>
<td>BDSIR</td>
<td>lognormal (0, 1.25)</td>
</tr>
<tr>
<td>Susceptible population</td>
<td>BDSIR</td>
<td>lognormal (8.2)</td>
</tr>
</tbody>
</table>

*phyldynamic models. CE for the exponential growth model, BD for the constant birth-death model, and BDSIR for the birth-death SIR model.
approximating the “prior” distribution here in the phylodynamic models, but that “prior” means the distribution of parameters conditional upon the sampling times with no sequence data (Boskova et al., 2018). Therefore, the “prior” distribution approximated here in the BD and BDSIR models is different from the prior distribution in usual Bayesian inference.

**Results**

**SARS-CoV-2 genome phylogeny**

A SARS-CoV-2 phylogeny was built from the ALL8510 genome alignment, clustering the virus isolates into four groups (Fig. 1A and Supplementary data Fig. S1A). Virus isolates that were sampled from Japan and Taiwan belonged to the four groups.
groups whereas those that were sampled from South Korea belonged to Group I. Several virus isolates that were sampled from Japan belonged to Group III and a few to Group II and IV. Virus isolates that were sampled from Taiwan were scattered all around the tree and belonged to all the four groups. Virus isolates that were sampled from the three locales on or before February 27, 2020 were clustered in Group I (Fig. 1B and Supplementary data Fig. S1B). Comparison of the two subfigures of Fig. 1 suggests multiple introductions of genetically variable strains of SARS-CoV-2 viruses into Taiwan (Fig. 1).

The JKT116 data set was used to build an unrooted phylogenetic tree by employing IQ-TREE (Fig. 2). Most of the virus isolates from Japan were clustered in Group C, appearing to be diverged from a single introduction into the locale. A few sequences that were from Japan were evolutionarily more related to those from South Korea in Group A. A few others from Japan commingled with others from South Korea and Taiwan in Group B. Virus isolates from Taiwan belonged to all
the groups and were more distantly related to one another than those from Japan or South Korea. Virus isolates from South Korea mostly were concentrated in Group A in the tree. A few virus isolates from South Korea belonged to Groups B and C in the tree. Overall, virus isolates from South Korea belonged to Group A, those from Japan to Group C, and those from Taiwan were scattered in the tree.

Overall, virus isolates from the different locales were not monophyletic; for example, some sequences from Taiwan were more closely related to those from either South Korea or Japan. These sequences from different locales that were more closely related to one another than those from the same locale would be likely to share a common ancestor in one of the three locales or in mainland China. Those sequences that were from a locale and more closely related to those in other locales might have been introduced by travelers from the other locales rather...
than be arisen by transmission within the same locale.

The most recent common ancestor of a locale or the genome sequence collected at the earliest date for a locale could be sampled from a patient who had a history of travel to mainland China. For example, the virus isolate of GISAID accession of EPI_ISL_406031 (labeled with a boldface font in Fig. 2) is the earliest isolate from Taiwan in the data set. This isolate was sampled from a 59-year-old Taiwanese male who worked in Wuhan until January 20, 2020 and returned to Taiwan on January 21, 2020 (Chen et al., 2020). Accession EPI_ISL_410531 (labelled with a boldface font in Fig. 2) is the earliest date in the data from Japan. It also happens to represent the most recent common ancestor of Japan in the inferred tree. This Japanese isolate belongs to lineage B and GISAID clade L, the earliest type of SARS-CoV-2. The origin of this isolate might have been in mainland China (see GISAID analysis update accessed July 6, 2020).

Rates and times

Regardless of the phylodynamic models used, the rate of molecular evolution was higher for Japan than Taiwan and the rate for Taiwan was higher than for South Korea (Table 2). The rate from the JKT116 data was similar to that estimated from Japan, presumably because the majority of viral isolates in the JKT116 data set were from Japan. The rates estimated here were larger than those estimated from a previous study (Andersen et al., 2020), but they were of the same order of magnitude.

The phylodynamic analyses yield information about when an epidemic originated in a sampled population as well as about the time of most recent common ancestry of sampled viral isolates. Out of necessity, the time of most recent common ancestry cannot have preceded the origin time. In contrast, the index case in a locale could be later than the origin time and the time of a most recent common ancestor if the earliest circulating viruses in the locale were not identified. Estimation error could explain an index case that was identified prior to the inferred origin time.

The first index case in Japan was reported on January 16, 2020. The estimate of the first infection date for Japan was January 21, 2020. The estimate of origin of the epidemic was January 18, 2020 for both the BD and BDSIR models (Table 3).

The first index case in Taiwan was reported on January 21, 2020. The estimates of the two kinds of date for Taiwan were January 10, 2020 for the first infection date and January 3, 2020 for origin of the epidemic from the BD model. The date estimates for Taiwan from the BDSIR model were similar to

| Table 2. Estimates of evolutionary rate in substitutions per site per year. Each estimate is a posterior median and the values within the brackets are for the interval of 95% high posterior density |
|-----------------|--------------------|--------------------|--------------------|
| Locale          | CE                 | BD                 | BDSIR               |
| Japan           | 1.678e-03 [1.118e-03, 2.285e-03] | 2.376e-03 [1.752e-03, 3.139e-03] | 2.410e-03 [1.784e-03, 3.081e-03] |
| South Korea     | 0.559e-03 [0.131e-03, 1.108e-03] | 1.152e-03 [0.631e-03, 1.763e-03] | 1.094e-03 [0.589e-03, 1.685e-03] |
| Taiwan          | 1.116e-03 [0.782e-03, 1.475e-03] | 1.707e-03 [1.368e-03, 2.077e-03] | 1.495e-03 [1.176e-03, 1.852e-03] |
| JKT116          | 1.637e-03 [1.152e-03, 2.178e-03] | 2.478e-03 [1.905e-03, 3.138e-03] | 2.345e-03 [1.780e-03, 2.957e-03] |

| Table 3. Estimates of dates at the root of the tree and origin of the epidemic. Each estimate is a posterior median and the values within the brackets are for the interval of 95% high posterior density |
|-----------------|--------------------|--------------------|
| Model           | Locale             | Root of the tree   | Origin of the epidemic |
|                 | Taiwan             | 2020-01-10 [2019-12-31, 2020-01-17] | 2020-01-03 [2019-12-12, 2020-01-16] |
those of the BD model. The 95% HPD of the date estimates did not include the date of the first index case in Taiwan. Because SARS-CoV-2 genome sequences from Taiwan were more diverse than those from Japan (Fig. 1A) and the rate of evolution were estimated to be smaller (Table 2), the dates for Taiwan were estimated to be earlier than those for Japan.

The first index case in South Korea was reported on January 20, 2020. The estimate of the first infection date for South Korea was January 3, 2020 and the estimate of origin of the epidemic was December 27, 2019 for the BD model (Table 3). These date estimates are earlier than those for the other locales and they are accompanied with a smaller estimate of the rate of evolution for South Korea (Table 2). Overall, the dates of the first infection and the epidemic origin with the pooled data set of JKT116 were estimated to be mid-January 2020 for the BD and BDSIR models (Table 3).

**Epidemiological parameters**

Because our data sets were small, the prior and posterior distributions of the basic reproductive number parameter were compared (Table 4). A difference between the prior and posterior suggests that the data are informative. The BDSIR model is parameter-rich and therefore might require a large data set to accurately estimate the parameters in the model. Approximation of the prior distribution of the basic reproductive number with the BDSIR model with the data sets for Taiwan resulted in estimates that were much larger than 1. However, comparisons of the posterior and prior estimates with the BDSIR model in the other locales were somewhat consistent with those with the BD model.

The BD model has a smaller number of parameters than the BDSIR model. Analysis of the data set sampled from Taiwan before February 27, 2020 with the BD model resulted in very similar prior and posterior approximations (Table 4). This suggests the Taiwan data are not informative for the BD model. In contrast, the posterior approximation for South Korea was shifted relative to the prior so that the posterior median was greater than the prior median (Table 4). The South Korean data therefore seem to be informative with regard to the BD model. Although the degree of difference between the posterior and prior distributions was smaller, analysis of the data set from Japan with the BD model showed a similar pattern to that of South Korea.

We intended to compare the basic reproductive number parameters in the three locales. The data sets sampled from Taiwan and Japan before February 27, 2020 are relatively uninformative about the parameters of the BD model (Table 4). If the true basic reproductive number parameter is 1, the posterior estimate with enough data would be equal to the prior estimate. Similarly, if the data are not informative, the posterior and the prior would be the same. Because we could not distinguish the case of true parameter of the basic reproductive number being 1 and the case of the non-informativeness of the data, one needs to be cautious in interpreting the result. The difference between the prior and posterior median estimates from the South Korean data set was the largest. However, we could not confidently state that the basic reproductive number of South Korea was greater than those of Japan or Taiwan because the data sets from Japan and Taiwan are relatively uninformative about the parameter. Because we could not increase the sample size for each locale, we combined all the data sets from the three locales to increase the sample size for estimating the basic reproductive number.

<table>
<thead>
<tr>
<th>Model</th>
<th>Locale</th>
<th>Posterior</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD</td>
<td>Japan</td>
<td>1.053 [0.796, 1.375]</td>
<td>1.011 [0.769, 1.281]</td>
</tr>
<tr>
<td></td>
<td>South Korea</td>
<td>1.455 [0.880, 2.626]</td>
<td>1.159 [0.733, 1.835]</td>
</tr>
<tr>
<td></td>
<td>Taiwan</td>
<td>1.100 [0.569, 1.911]</td>
<td>1.124 [0.539, 2.021]</td>
</tr>
<tr>
<td></td>
<td>JKT116</td>
<td>1.072 [0.889, 1.340]</td>
<td>1.028 [0.825, 1.237]</td>
</tr>
<tr>
<td>BDSIR</td>
<td>Japan</td>
<td>1.320 [1.013, 1.898]</td>
<td>1.176 [0.988, 1.424]</td>
</tr>
<tr>
<td></td>
<td>South Korea</td>
<td>1.832 [1.069, 3.422]</td>
<td>1.480 [0.814, 7.175]</td>
</tr>
<tr>
<td></td>
<td>Taiwan</td>
<td>1.596 [0.917, 3.248]</td>
<td>2.324 [0.719, 18.067]</td>
</tr>
<tr>
<td></td>
<td>JKT116</td>
<td>1.284 [1.066, 1.651]</td>
<td>1.138 [1.008, 1.309]</td>
</tr>
</tbody>
</table>
number in East Asia. In other words, to address the issue of small data size in the estimation of the basic reproductive number, the pooled JKT116 data set was analyzed with the BD model. The posterior median of the reproductive number using the JKT116 data with the BD model was greater than the prior median by 0.044 (Table 4). The difference of 0.044 for the JKT116 data was compared with the distribution of differences between posterior and prior medians that were estimated using the 100 SUBSAMPLE data sets (Fig. 3). All 100 differences between the posterior and prior median estimates of the parameter was obtained for each subsampled data set and the distribution of the differences is drawn with the vertical dashed line at 0.044 that shows the corresponding difference from the JKT116 data.

Discussion

Traditionally, phylogenetics has mainly been a rather static study of reconstruction of the past history of species. It is becoming a dynamic study with application in other fields. Although the COVID-19 situation around the world is dire, phylodynamics is an exciting phylogenetic subfield with application to epidemiology. The three locales studied here (Japan, South Korea, and Taiwan) all neighbor mainland China and have relatively well-developed economies. Accordingly, one might expect the locales to be similarly prepared for a contagious disease such as COVID-19. However, the outcomes of their preparedness were quite different; the number of confirmed cases in either Japan or South Korea were much larger than that of Taiwan as of July 13, 2020 (https://covid19.who.int). This study of viral genetic data comes to the plausible conclusion that the dates of epidemic origin in these three East Asian locales were all about mid-January 2020. The genome data for Japan and Taiwan were not informative with regard to the phylodynamic models although the basic reproductive number parameters for the two locales were expected to be less than or equal to one because of the slow early increase in Japan and Taiwan before February 27, 2020. We infer that the basic reproductive number was a little greater in South Korea. This is in agreement with the surge of the COVID-19 cases after mid-February 2020 in South Korea.

The index cases of COVID-19 for the three locales in East Asia were announced around mid-January 2020. The Taiwan Centers for Disease Control (CDC) started inspection measures for inbound flights from Wuhan on New Year’s Eve 2019, which was around the estimated date of the epidemic origin in Taiwan. In contrast, no particular preventive measures were reported in Japan or South Korea in early to mid-January 2020. Subsequently, all three locales began to inspect inbound travelers in airports on a regular basis and to quarantine those with symptoms of COVID-19. In retrospect, preventive measures such as airport screening might have been desirable to implement earlier in Japan and South Korea.

The cumulative number of COVID-19 tests conducted in South Korea until February 23, 2020 was 26,179 (Korea CDC, 2020; So, 2020). The total tests in South Korea was over 250,000 by mid-March and over 600,000 as of May 4, 2020 (So, 2020). In contrast, the cumulative number of tests in Japan was less than 15,000 by mid-March and less than 200,000 as of May 4, 2020 (Statista Research Department, 2020). Accounting for the fact that the population size of Japan is about twice the...
size of South Korea, the per capita number of tests for South Korea was substantially greater than that for Japan. The number of cumulative COVID-19 tests in Taiwan was a little over 60,000 as of May 4, 2020, which was the smallest of all of the cumulative test numbers for the three locales (Taiwan CDC, 2020). Although we expected that the number of tests might be related with the rate of becoming noninfectious, the rate estimate of becoming noninfectious in South Korea was not greater than the estimate for Japan (Table 5). This might be partly because the two data sets of Japan and South Korea were not of the same in the size and partly because the rate of becoming noninfectious could be affected by other factors that were not considered in the model, including deaths or behavioral changes of an infected person.

Although the COVID-19 crisis has not ended, some countries have eased social distancing restrictions and health officials warn of a potential reemergence of COVID-19 in the coming fall 2020 (Abutaleb et al., 2020). Monitoring COVID-19 infections will be important for designing and implementing preventive measures. Phylodynamic inferences could aid these public health decisions, but the impact of phylodynamic inferences will be partly contingent on having a well-designed SARS-CoV-2 genome sampling strategy. As seen in this study, phylodynamic inferences can be limited by small sample sizes and by unclear sampling. The current analysis was also limited in that it restricted analysis to genome sequences that were available at GISAID. A carefully designed sampling scheme could prove beneficial when attempting to have phylodynamic analyses inform public health policies regarding COVID-19.

<table>
<thead>
<tr>
<th>Locale</th>
<th>Posterior</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan</td>
<td>245 [98, 644]</td>
<td>403 [200, 824]</td>
</tr>
<tr>
<td>South Korea</td>
<td>30 [5, 127]</td>
<td>108 [11, 360]</td>
</tr>
</tbody>
</table>

Table 5. Posterior and prior estimates of rate of becoming noninfectious per year. Each estimate is a median of either the posterior or the prior distribution and the values within the brackets are for the interval of 95% high posterior or prior density.

Acknowledgments

I thank researchers who shared SARS-CoV-2 genome sequences in the GISAID database (https://www.gisaid.org/), which was instrumental in the current study. A full table of acknowledgments for GISAID contributors can be found in Supplementary data Table 1. I thank Jeffrey Thorne for discussion. I thank the reviewers for valuable comments and suggestions that helped improve the manuscript. This work was supported by the Sungshin Women’s University Research Grant of H20200112.

References


