SARS-CoV-2 HaploGraph: visualization of SARS-CoV-2 haplotypes spread in Japan

So Nakagawa[1,2,3,4†*], Toshiaki Katayama[5†], Lihua Jin[6†], Jiaqi Wu[2], Kirill Kryukov[1], Rise Oyachi[7], Junko S Takeuchi[8], Takatomo Fujisawa[1], Satomi Asano[1], Momoka Komatsu[9], Jun-ichi Onami[10], Takashi Abe[1,9*], Masanori Arita[1,11*]

1. Bioinformation and DDBJ Center, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
2. Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa 259-1193, Japan
3. Micro/Nano Technology Center, Tokai University, Hiratsuka, Kanagawa 259-1292, Japan
4. Institute of Medical Sciences, Tokai University, Isehara, Kanagawa 259-1193, Japan
5. Database Center for Life Science, 178-4-4 Wakashiba, Kashiwa, Chiba 277-0871, Japan
6. Genomus Co., Ltd., Sagamihara, Kanagawa 252-0226, Japan
7. Department of Applied Biochemistry, School of Engineering, Tokai University, Hiratsuka, Kanagawa 259-1292, Japan
8. Center for Clinical Sciences, National Center for Global Health and Medicine, Shinjuku, Tokyo 162-8655, Japan
9. Smart Information Systems, Faculty of Engineering, Niigata University, Niigata, Niigata 950-2181, Japan
10. Research Center for Open Science and Data Platform, National Institute of Informatics, Chiyoda, Tokyo 101-8430, Japan
11. RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa 230-0045, Japan

†These authors contributed equally.

*Corresponding authors
E-mail: so@tokai.ac.jp (So Nakagawa), takaabe@ie.niigata-u.ac.jp (Takashi Abe), and arita@nig.ac.jp (Masanori Arita)

Abbreviations: COVID-19, coronavirus disease of 2019; GISAID, global initiative on sharing all influenza data; MHLW, Ministry of Health, Labour and Welfare; NIID, National Institute of Infectious Diseases; PANGO, phylogenetic assignment of named global outbreak; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SNV, single nucleotide variant; VOC, variant of concern; WHO, World Health Organization

Abstract

Since the early phase of the coronavirus disease 2019 (COVID-19) pandemic, a number of research institutes have been sequencing and sharing high-quality severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes to trace the route of infection in Japan. To provide insight into the spread of COVID-19, we developed a web platform named SARS-CoV-2 HaploGraph to visualize the emergence timing and geographical transmission of the SARS-CoV-2 haplotypes. Using data from the GISAID EpiCoV database as of June 4, 2022, we created a haplotype naming system by determining the ancestral haplotype for each wave and showed prefectural or region-specific haplotypes in each of the four epidemic waves in Japan. The SARS-CoV-2 HaploGraph allows for interactive tracking of virus evolution and geographical prevalence of haplotypes and aids in developing effective public health control strategies during the global pandemic. The code and the data used for this study are publicly available at: https://github.com/ktym/covid19/.

Keywords: COVID-19, haplotype, genomic surveillance, SARS-CoV-2, web visualization

Introduction

Coronavirus disease 2019 (COVID-19) has become the leading cause of morbidity and mortality worldwide since the end of 2019. COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with a genome size of 29.9 kb (Khailany et al., 2020; Morens et al., 2020). SARS-CoV-2 is highly transmissible with a broad tissue tropism, causing rapid respiratory and gastrointestinal illnesses and long-term ramifications such as myocardial inflammation, which is a significant risk for the elderly (Harrison et al., 2020). Globally, as of July 25, 2022, 572,239,451 confirmed cases of COVID-19, including 6,390,401 deaths, have been reported to WHO [World Health Organization [WHO], 2022a], which makes COVID-19 one of the deadliest infectious diseases in history, as suggested by previous reports (Morens et al., 2020; Sampath et al., 2021). COVID-19 has become an enormous threat to public health and social and economic activities in many countries, including Japan (Smallwood et al., 2022; Nicola et al., 2020).

Japan has faced six COVID-19 epidemic waves caused by different lineages of SARS-CoV-2 by June 2022 (Outbreak.info, 2023). The first COVID-19 case in Japan was confirmed on January 15, 2020, from a person who had a history of staying in Wuhan, China, shortly after the COVID-19 outbreak [Ministry of Health, Labour and Welfare (MHLW), 2020]. Reports of additional cases followed, including the passengers from Wuhan via chartered flights in January and the crew members and passengers on the Diamond Princess cruise ship in February (Arima et al., 2020; Yamagishi et al., 2020). By March, the number of confirmed cases continued to increase and peaked in April (Arima et al., 2021). It was reported that the first wave was contributed by at least two introductions of distinct strains: the earlier introduction in January and February from China, and the latter in March and April mainly from Europe, of which the Spike (S) protein contains an amino acid replacement of aspartate to glycine at position 614 (D614G) (Sekizuka et al., 2020a mSphere).

The second and third epidemic waves were caused by PANGO (Phylogenetic Assignment of Named Global Outbreak, Rambaut et al., 2020) lineages B.1.1.284 and B.1.1.214 [National Institute of Infectious Diseases (NIID), 2021a], respectively. It was reported that the two Japan-specific lineages, B.1.1.284 and B.1.1.214, were derived from the B.1.1.114 lineage which was mainly found in Europe from March to April 2020 (NIID, 2021a). The second wave started with an increasing number of cases in late May 2020, then peaked during late July through August, declined in September, and plateaued through the first half of October (Arima et al., 2021). The third epidemic wave began to rise in the latter half of October 2020 and has resulted in substantial COVID-19 morbidity and mortality,

surpassing both the total and fatal case counts from the first two waves combined. All indicators continued to increase through December 2020 and peaked in January 2021 in the third wave (Arima et al., 2021; Outbreak.info, 2023).

The fourth wave was caused by the first variant of concern (VOC), Alpha (including B.1.1.7 and Q lineages), which was first identified in the United Kingdom (UK) in late summer to early autumn 2020 (Volz et al., 2021). The fourth wave started to increase in Japan in March 2021 and reached its peak in May 2021 (NIID, 2021a; WHO, 2022a). The epidemiological studies estimating the reproductive numbers conducted in the UK showed that the Alpha variant containing N501Y replacement in the S protein has a 35% to 100% higher transmissibility than the pre-existing SARS-CoV-2 variants (Graham et al., 2021; Davies et al., 2021; Volz et al., 2021; Leung et al., 2021). The N501Y mutation enhances the Alpha variant's affinity for angiotensin-converting enzyme 2 (ACE2), the cellular receptor, that facilitates viral entry (Luan et al., 2021). An epidemiological study in Japan also indicated that the Alpha variant had an approximately 1.9–2.3-fold higher transmissibility than the pre-existing virus in the Japanese population (Tanaka et al., 2021). Alpha variants were not only more transmissible than pre-existing SARS-CoV-2 variants but also caused more severe illness and increased mortality (Challen et al., 2021; Grint et al., 2021; Davies et al., 2021).

The Delta variant (B.1.617.2 and AY lineages) that was first detected in India in October 2020 (WHO, 2022b; Mlcochova et al., 2021) began spreading rapidly throughout the country from July to August 2021 causing the 5th wave in Japan (Abe and Arita, 2021; Koyama et al., 2022). The AY.29 which is a sub-lineage of the B.1.617.2, containing C5239T and T5514C mutations, spread around the end of June and early July 2021 and became the predominant strain in Japan, reaching its peak in September 2021 (Abe and Arita, 2021; Koyama et al., 2022).The Delta variant lacks the N501Y, the mutation prominent in the Alpha variant, but carries several mutations within the S protein such as L452R, T478K, and P681R, which confer resistance to monoclonal antibody treatments (Gupta et al., 2021). The P681R replacement also was found to increase fusogenicity and pathogenicity that characterized the specific feature of the Delta variant (Saito et al., 2022). Another mutation within the N protein, R203M, increases viral mRNA delivery and expression, allowing the Delta variant to produce >50-fold more viral particles (Syed et al., 2021). Studies reported that Delta was 63%–167% more transmissible than Alpha in the USA and showed 1.4 times higher transmissibility than Alpha in Japan (Earnest et al., 2022; Ito et al., 2021).

The sixth wave in Japan was caused by the Omicron variants, mainly by BA.1 (i.e., B.1.1.529.1) (Iketani et al., 2022; Desingu et al., 2022). The first case of Omicron infection

was reported in “Japan quarantine” on November 30, 2021, from a traveler from Namibia (MHLW, 2021). Following the first case, the rapid spread of the Omicron variant caused the largest increase of COVID-19 cases in Japan and reached its peak in February 2022 (Outbreak.info, 2023). The major strain of the sixth wave shifted from BA.1 to BA.1.1 up to March 2021, which was later replaced by BA.2, a sister lineage of BA.1 (NIID, 2022). The Omicron (BA.1) variant containing over 30 amino acid replacements in the S protein, showed a higher affinity for ACE2 compared with Delta. A marked change in antigenicity increased Omicron's evasion of therapeutic monoclonal and vaccine-elicited polyclonal neutralizing antibodies after two doses (Meng et al., 2022; Okumura et al., 2022). Another study suggested that Omicron has spread more rapidly than the Delta variant in several countries and revealed that Omicron showed lower fusogenicity and attenuated pathogenicity compared to Delta and ancestral SARS-CoV-2 (Suzuki et al., 2022).

In this study, to obtain key insights into the spread of COVID-19 in Japan in more detail, we created a website visualizing the transmission of SARS-CoV-2 haplotypes named SARS-CoV-2 HaploGraph (https://ktym.github.io/covid19/). The SARS-CoV-2 HaploGraph utilized major SARS-CoV-2 lineages for each epidemic wave that were obtained by conducting phylogenetic analyses of SARS-CoV-2 genomes sampled in Japan. The reference haplotype for each lineage was identified, and their mutations were compared with those of other haplotypes to trace the transmission of COVID-19, taking into account the sampled date, location, and the number of sequences stored in the global initiative on sharing all influenza data (GISAID) EpiCoV database as of June 4, 2022 (https://www.gisaid.org; Khare et al., 2021). Next, the haplotypes of the isolates were determined by comparing their genomic variations to the reference haplotypes of the major lineages of the six epidemic waves in Japan. For each haplotype, we calculated its occurrence and observation period for each prefecture of Japan. The SARS-CoV-2 HaploGraph aids in understanding the dynamics of the SARS-CoV-2 variants spreading in Japan.

Results

Japan-related SARS-CoV-2 genomes were divided into the following three groups using the GISAID metadata: 1) “Domestic”: genomes sampled from patients infected in Japan (284,819 genomes); 2) “Quarantine”: genomes sampled from patients infected outside of Japan (10,644 genomes); and 3) “International”: genomes sampled outside Japan from patients infected in Japan (92 genomes). For the entries from the COVID-19 outbreak on the cruise ship - Diamond Princess (Sekizuka et al., 2020b), 72 entries were assigned to the “Quarantine”. From these data, we used genomes without any undetermined nucleotide base. The numbers of genomes categorized as “Domestic”, “Quarantine”, and “International” were 251,761, 9,199, and 33 (88.4%, 86.4%, and 35.9% of all genomes), respectively.

The haplotype analysis was conducted for the major variants of the six epidemic waves in Japan. The results for the first five waves are summarized in Table 1. The start and end dates followed the record of the Ministry of Health, Labour and Welfare (MHLW) of Japan (MHLW, 2022). The details of major variants for each epidemic wave were described in the following sections.

Table 1. Statistics for GISAID entries of the five COVID-19 epidemic waves in Japan (# indicates the number)

Epidemic wave	**1st**	**2nd**	**3rd**	**4th**	**5th**
Start date	2020-01-01	2020-06-14	2020-10-10	2021-03-01	2021-06-21
End date	2020-06-13	2020-10-09	2021-02-28	2021-06-20	2021-12-16
Major PANGO ID (VOC)	B.1.1	B.1.1.284	B.1.1.214	B.1.1.7 and Q (Alpha)	B.1.617.2 and AY (Delta)
Reference GISAID ID (# of SNVs)	-	EPI_ISL_692598 (12)	EPI_ISL_686874 (9)	Haplotype 1: EPI_ISL_1931259 (29); Haplotype 2: EPI_ISL_1430629 (29)	EPI_ISL_1927416 (38)
# of total	3,841	7,955	19,300	42,328	90,716

genomes (before filtering)	(4,259)	(8,521)	(20,540)	(44,966)	(106,668)
# of B.1.1.284 (proportion)	90	5,912 (74.3%)	2,952	64	0
# of B.1.1.214 (proportion)	5	1,949	14,667 (76.0%)	1,415	1
# of B.1.1.7/Q.* (proportion)	7	0	20	22,462 (53.1%)	8,224
# of B.1.617.2/AY.29.* (proportion)	0	0	1	192	76,979 (84.9%)

The “total genomes” shows the number of GISAID entries after filtering the genomes with any undetermined sequence. The number of genomes of the dominant PANGO IDs of the five waves is listed by their sampled date. Note that the genomes of particular PANGO IDs were detected before and after the wave period in which the particular PANGO ID was predominant. The percentage in the parenthesis indicates the proportion of the genomes of the dominant PANGO ID in the “total genomes” sampled in the corresponding wave period. Entries without month, date, or prefecture information are not included in these statistics. An asterisk (*) indicates any numbers in the sub-lineages of the given PANGO IDs.

1st wave in Japan

To obtain the major haplotype in the first wave, we examined the GISAID entries sampled from January to June 2020; the period covers from the beginning until the end of the first wave (Arima et al., 2021). Note that the wave period definition applied in this analysis is slightly different from that noted in Table 1. In total, 4,951 entries were classified into the 3 categories: “Domestic” (4,786 genomes), “Quarantine” (159 genomes), and “International” (6 genomes). We further examined the PANGO IDs of the first wave and found 69, 4, and 18 unique PANGO lineages in each of the “Domestic”, “Quarantine”, and “International” categories, respectively. After merging the three categories, the number of unique PANGO lineages became 77. The PANGO ID with the largest number of entries (3,057) was B.1.1, which accounted for 61.8% of 4,951 all Japan-related first-wave entries, followed by 7.6% (374) of B.1.1.48. We focused on the B.1.1 and its sub-lineages and investigated the haplotypes of the Japan-related B.1.1 entries of the first wave. For those

entries, the number of SNVs ranged from 4 to 17. Two GISAID entries (GISAID ID: EPI_ISL_684814 and EPI_ISL_685167) having the smallest number of SNVs in the "Domestic" B.1.1 lineage contain the same combination of 4 SNVs, of which sampling dates are April 6, 2020, and April 14, 2020, respectively. Other "Domestic" B.1.1 and the sister lineages also contained those 4 SNVs; however, the 4 SNVs were also found in 90.2% (47,138 entries) of non-Japan B.1.1 samples. Thus, no Japan-specific SNVs were found in the "Domestic" B.1.1 lineage. Therefore, it was impossible to distinguish between the B.1.1 strains of the first wave sampled in Japan and those from outside the country using their SNVs. For this reason, SARS-CoV-2 haplotypes of the first wave in Japan were not visualized in the HaploGraph.

2nd wave in Japan - B.1.1.284

The second wave was reported to be mainly caused by B.1.1.284, a sub-lineage of B.1.1 (NIID, 2021a). We extracted 9,673 genomes of B.1.1.284 and found that almost all (9,665) entries were related to Japan — 9,660 "Domestic", 2 "Quarantine", and 3 "International" entries — leaving only 8 non-Japan entries, which is consistent with the previous study [Tokyo Metropolitan Institute of Public Health (TMIPH), 2022]. We then examined the SNVs specific for B.1.1.284 lineage by comparing them with the SNVs of the ancestor lineage B.1.1. The total number of B.1.1 entries in the GISAID database is 56,003, including 3,720 "Domestic", 112 "Quarantine", 6 "International", and 52,165 non-Japan entries. We divided these 4 categories into two groups: B.1.1 Japan infected group ("Domestic" and "International") and B.1.1 non-Japan infected group ("Quarantine" and non-Japan). Five SNVs (C241T, C3037T, C14408T, A23403G, and GGG28881AAC) were shared (≥90% fraction) between the two groups, while one SNV (i.e., C313T) showed a high fraction only in Japan (90.3% and 5.5% for Japan infected and non-Japan infected B.1.1, respectively) (Table S1). The earliest B.1.1 entries containing all the six mutations, sampled in Japan and in non-Japan countries (Switzerland), are EPI_ISL_1429762 (March 8, 2020) and EPI_ISL_527916 (February 26, 2020), respectively. Although the C313T mutation itself may not have occurred in Japan, the B.1.1 variant with this mutation spread in Japan and served as the origin of the second wave because C313T was fixed in the B.1.1.284 lineage (100.0%). The B.1.1.284 lineage further obtained 6 major SNVs (≥98.7%): T4346C, C9286T, C10376T, C14708T, C28725T, and G29692T. We found that the earliest entry of the B.1.1.284 lineage is EPI_ISL_692598 (May 22, 2021), which contains all the 12 aforementioned SNVs but does not contain other SNVs. Therefore, we used the haplotype of EPI_ISL_692598 as the reference for the B.1.1.284 lineage.

It has been reported that B.1.1.284 and B.1.1.214 — the main lineages of the second and the third wave, respectively — were derived from a European strain B.1.1.114 (NIID, 2021a). We found that the early B.1.1.114 strain contains 7 SNVs and shares 6 of them with B.1.1.284 and B.1.1.214, including the C313T. However, the remaining SNV (G26849T) was not found in either B.1.1.284 or B.1.1.214. For this reason, we conclude that B.1.1.114 is not the ancestral lineage of B.1.1.284 and B.1.1.214, but their sister lineage.

The total number of haplotypes of the “Domestic” B.1.1.284 lineage (9,251 genomes) was 3,474. The reference haplotype (Haplotype 1) was the third largest one containing 220 entries, with the earliest entry EPI_ISL_692598 (May 22, 2020) and the latest entry EPI_ISL_902206 (September 16, 2020), both of which were sampled in Saitama Prefecture, indicating that the identical SARS-CoV-2 haplotype continuously circulated around 4 months (118 days) (Table 2 and Figure 1).

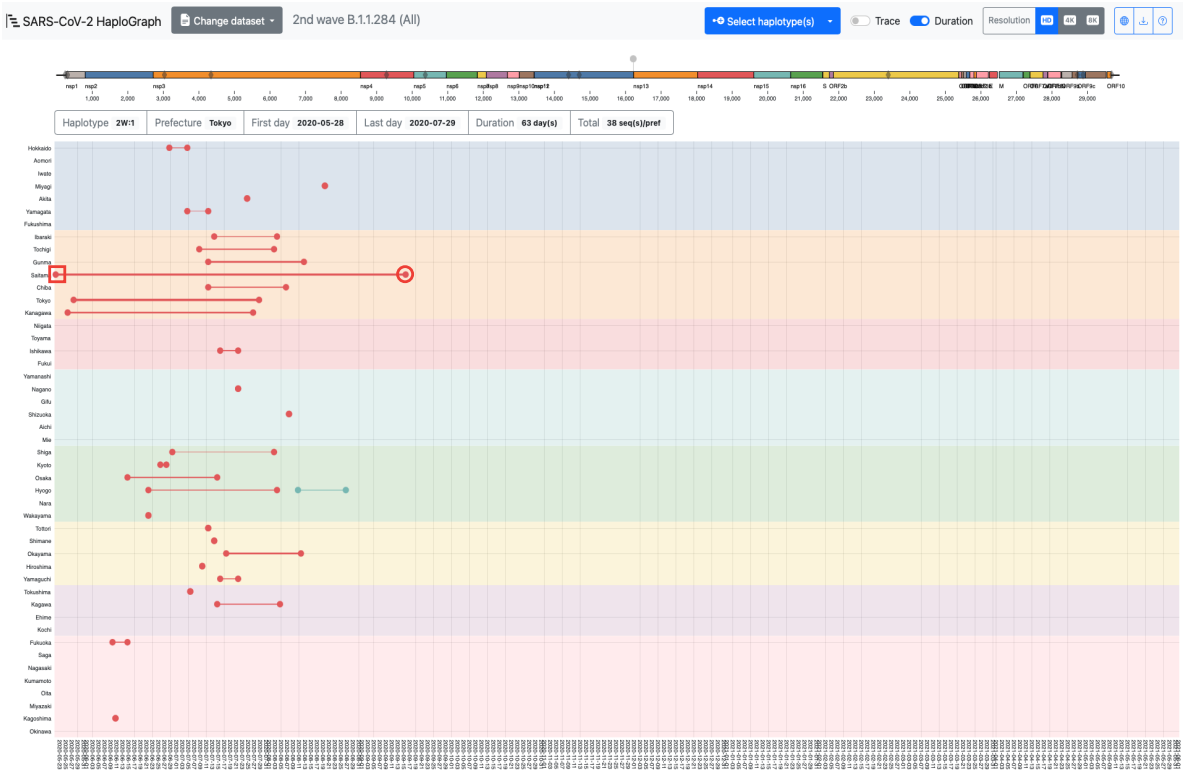

Figure 1. SARS-CoV-2 HaploGraph website visualizes the haplotype “1” and haplotype “1.1” of the second epidemic wave in Japan. On the top menu bar, from the “Change dataset”, the dataset for an epidemic wave can be selected. Two datasets are available for each wave: “All” and “50%”. “All” contains all haplotypes, and “50%” includes all haplotypes up to the smallest frequency haplotype whose sum exceeds 50% when all frequencies are ordered from highest to lowest. Then, the “Select haplotype(s)”, one or more haplotype(s) can be selected: “1” and “1.1” of the second wave shown in this figure. The X-

axis or Y-axis indicates the sampling date or prefecture. In each prefecture (47 prefectures in total), the dots indicate the earliest or the latest sampled dates. The dots are connected with a line, the thickness of which is correlated with the number of entries sampled in that prefecture. A schematic representation of the SARS-CoV-2 genome is shown in a colored bar. On the genome bar, the gray diamonds indicate the genomic locations of the mutations contained in the reference haplotype of the selected wave. The grey pins above the genome bar indicate the schematic genomic location of the mutation(s) contained in the selected haplotype(s): “1.1” shown in the grey dots and line in this figure indicates SNVs compared to the reference haplotype of each epidemic wave. The 6 boxes under the genome bar describe: 1) “Haplotype”, the wave and haplotype name; 2) “Prefecture”, the prefecture where the haplotype was sampled; 3) “First day”, the earliest date, and 4) “Last day”, the latest date of the entries sampled in the prefecture; 5) “Duration”, the time span by day(s); and 6) “Total”, the total number entries of the haplotypes sampled in the prefecture. These particulars change when a different prefecture within its “Duration” period is touched by the cursor. When the "Trace" option is on, the earliest entries of different prefectures will be connected by a line in time order. However, it doesn't mean that those earliest entries were related to the spread of the haplotype. In this figure, the earliest or latest entry of the haplotype “1” is surrounded by an open square or an open circle of the same color as the haplotype. Note that this visualization is not by the function of the viewer. The colors for displaying the selected haplotypes are adopted randomly by the viewer.

The two largest haplotypes of the second wave were Haplotype “1.73” (479 entries) and “1.70” (298 entries), with T8076C and T22020C mutation compared to the Haplotype “1”, respectively. Haplotype “1.73” (the earliest sampled on June 20, 2020, in Hyogo; the latest sampled on August 27, 2020, in Tochigi) was mainly sampled in Fukuoka, Hyogo, and Kumamoto (64.3%), the prefectures located in the southwestern part of Japan. Haplotype “1.70” (the earliest sampled on May 22, 2020, in Tokyo; the latest sampled on October 29, 2020, in Fukuoka) was mainly sampled in the greater Tokyo area (Tokyo, Saitama, and Kanagawa; 60.8%) (Table 2 and Figure2).

Table 2. The representative haplotypes of the second wave that showed distinctive features

Haplotype	Main feature	Obtained SNVs compared to “1”	Date	Prefecture

“1” (Reference)	Third largest	-	From 2020-05-22 To 2020-09-16	Saitama Saitama
“1.73”	Largest	T8076C	From 2020-06-20 To 2020-08-27	Hyogo Tochigi
“1.7”	Second largest	T22020C	From 2020-05-22 To 2020-10-29	Tokyo Fukuoka
“1.45.4”	Prefectural preference	A8031G, C7728T	From 2020-07-22 To 2020-10-05	Mostly sampled in Okinawa
“1.70.13.20.1.9.2”	Prefectural specificity	A27633G, C18247T, C23939T, C29642T, G25489A, T22020C	From 2020-12-10 To 2021-01-04	All sampled in Iwate
“s.2”	Includes the entry with the latest date	24 SNVs	From 2021-05-31 To 2021-05-31	Chiba Chiba

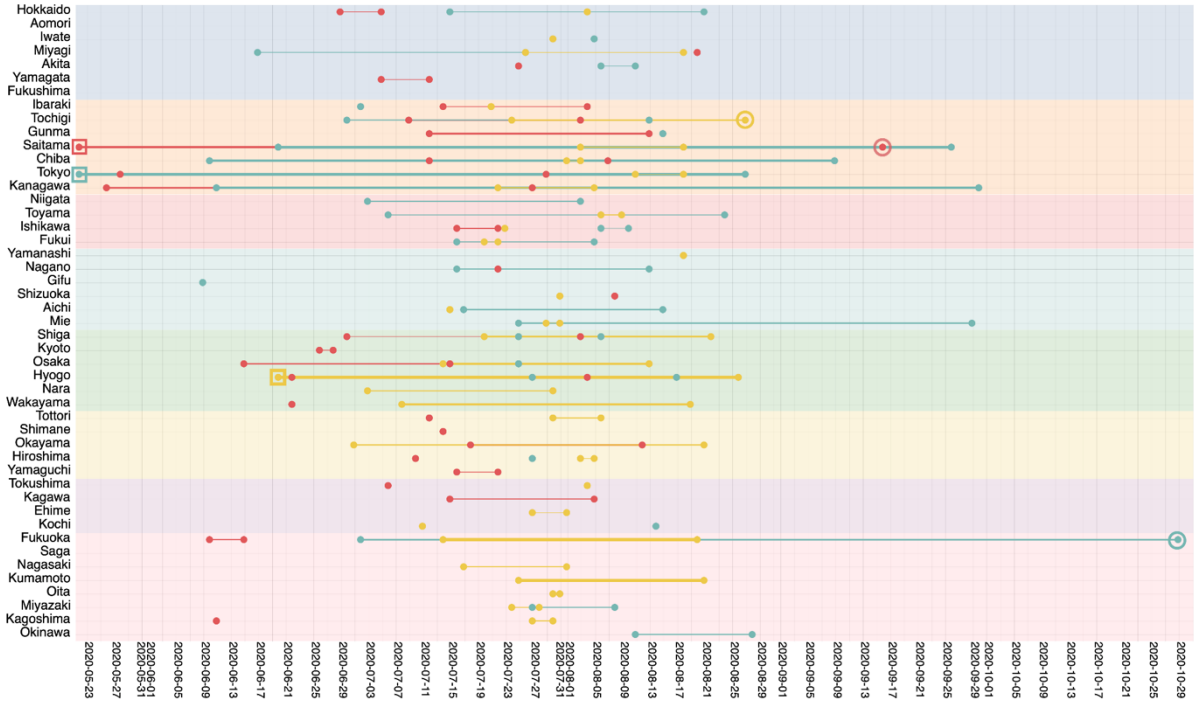

Figure 2. HaploGraph visualization for the haplotypes “1”, “1.73”, and “1.70” of the second epidemic wave. Dots and lines in red, yellow, and cyan indicate the reference haplotype “1”, the largest haplotype “1.73”, and the second-largest haplotype “1.70”, respectively. The earliest or latest entry of each haplotype is surrounded by a square or a circle of the corresponding color. For the details of this figure, please see Figure 1.

Two of the 17 haplotypes having more than 50 entries showed prefectural preference or specificity: the haplotype “1.45.4” (obtaining A8031G and C7728T) in Okinawa (89.84%; 115 out of 128 entries); and “1.70.13.20.1.9.2” (obtaining A27633G, C18247T, C23939T, C29642T, G25489A, and T22020C) in Iwate Prefecture (100.0%; 51 entries). Among all 3,474 haplotypes of the second wave, 243 haplotypes (7.0%) showed the largest sample size in Tokyo. The latest entry of the “Domestic” B.1.1.284 variant was EPI_ISL_3190941, sampled in Chiba on May 31, 2021. It belongs to the haplotype “s.2”, obtaining 21 SNVs compared to the haplotype type “1” (33 SNVs compared to the Wuhan-Hu-1 reference) (Table 2 and Figure S1).

3rd wave in Japan - B.1.1.214

The leading cause for the third wave of Japan was a Japan-specific lineage B.1.1.214, a sub-lineage of B.1.1 (NIID, 2021a). We used the same approach as for the second wave and obtained 18,978 entries of B.1.1.214 (18,954 “Domestic”, 9 “Quarantine”, 4 “International”, and 11 non-Japan entries). The B.1.1.214 lineage possesses 9 SNVs with high frequencies (≥99.9%): C241T, C313T, C3037T, C14408T, C18167T, G21518T, A23403G, GGG28881AAC, and G28975T. Among them, we defined 4 SNVs (C313T, C18167T, G21518T, and G28975T) as the B.1.1.214 specific fixed SNVs, because the remaining 5 SNVs (C241T, C3037T, C14408T, A23403G, and GGG28881AAC) were also observed in non-Japan B.1.1 variants. As we noted, among the 4 specific SNVs, the C313T mutation was also found in the B.1.1.284 lineage in the 2nd wave, suggesting that B.1.1.214 and B.1.1.284 shared the common ancestor that probably emerged in Japan. The earliest entry of B.1.1.214 is EPI_ISL_686874 (June 1, 2020), which contains all the 9 B.1.1.214 specific SNVs. Therefore, we used the haplotype of EPI_ISL_686874 as the reference for the B.1.1.214 lineage.

We found 7,788 haplotypes in the “Domestic” B.1.1.214 lineage. The haplotype “1” including the EPI_ISL_686874 had only 6 entries. The haplotype having the largest number of entries (416) was “1.7.21.x.3.x.5”, which contains six substitutions (C12049T, C17502T, C29353T, C6380T, C8917T, and G23587T) in addition to the haplotype “1”. Compared to the second wave, the major haplotypes of the third wave contained various substitutions, probably because the time lag between the emergence of the original haplotype “1” and the beginning of the 3rd wave was long. The reference haplotype of the third wave was sampled between June 1 and July 20, 2020, while the third wave started in the middle of October 2020 (and lasted until the end of February 2021) (Table 3 and Figure 3). The largest haplotype, “1.7.21.x.3.x.5”, was sampled evenly in the northernmost part of Japan (29.6% in Hokkaido) and the southernmost part (other than Okinawa) of Japan (38.6% in Oita and

323 Fukuoka), with the earliest entry sampled on Oct 2, 2020, in Hokkaido and the latest entry
324 sampled on February 17, 2021, in Chiba (Table 3 and Figure 3).
325
326 **Table 3. The representative haplotypes of the third wave that showed distinctive**
327 **features**

Haplotype	Main feature	Obtained SNVs compared to “1”	Date	Location
“1” (Reference)	Only 6 entries	-	From 2020-06-01 To 2020-07-20	Tochigi Kanagawa
“1.7.21.x.3.x.5”	Largest	C12049T, C17502T, C29353T, C6380T, C8917T, G23587T	From 2020-05-15 To 2021-02-17	Hokkaido Chiba
“1.7”	Second largest	C8917T	From 2020-06-22 To 2020-11-12	Okayama Hokkaido
“c.1.x.10”	Third largest	A6592G, G1738A, G28541A, T25673C, T6235C	From 2020-10-01 To 2021-01-30	Hyogo Nara
“c.1.x.6.x.43.x.4.1”	Prefectural specificity	A27712G, A3546G, A6592G, C16887T, C29272T, C6636T, G1738A, G26217T, G28541A, GTCATGTTA510	From 2021-01-16 To 2021-01-31	All sampled in Yamaguchi
“1.6.7.1.2.9.x.11.2”	Prefectural specificity	A23720G, C15240T, C6433T, C9207T, G10565A, G25489T, G29527T, G8371T	From 2021-01-22 To 2021-03-13	All sampled in Fukuoka
“s.1”	Includes the entry with the latest date	19 SNVs	From 2021-07-29 To 2021-07-29	Tokyo Tokyo

328

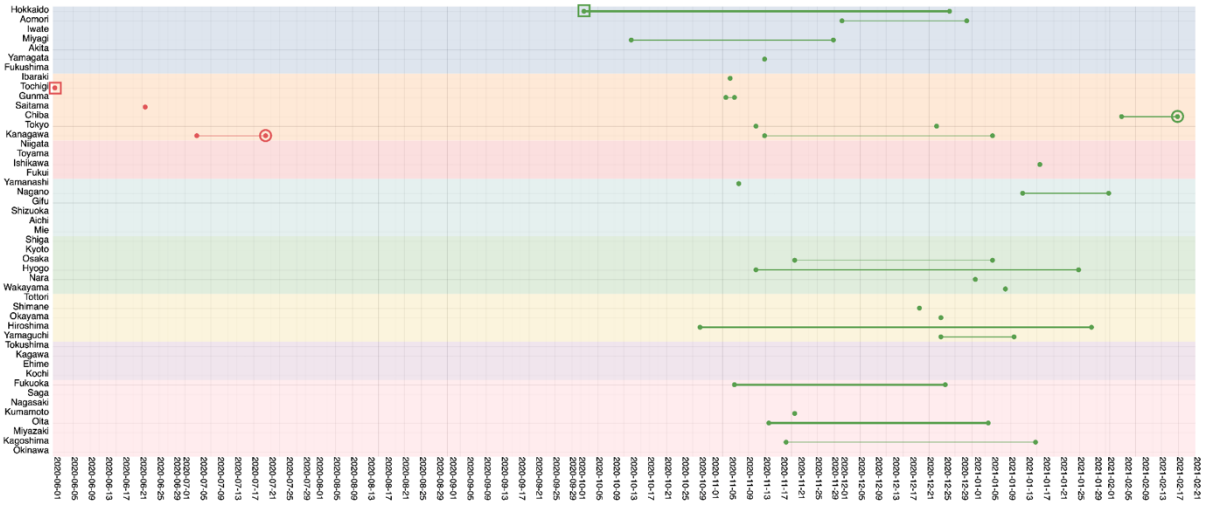

Figure 3. HaploGraph visualization for the haplotypes "1" and "1.7.21.x.3.x.5" of the third epidemic wave. Dots and lines in red and green indicate the reference haplotype "1" and the largest haplotype "1.7.21.x.3.x.5", respectively. The earliest or latest entry of haplotype is indicated by a square or a circle of the corresponding color, respectively.

The second largest haplotype of the third wave was the haplotype "1.7" (405 entries), obtaining C8917T with the earliest entry sampled on June 22, 2020, in Okayama and the latest entry sampled on November 12, 2020, in Hokkaido. The third largest haplotype was "c.1.x.10" (obtaining A6592G, G1738A, G28541A, T25673C, and T6235C; 267 entries) with the earliest entry sampled on October 1, 2020, in Hyogo and the latest entry sampled on January 30, 2021, in Nara. The second and third largest haplotypes were sampled in various prefectures as shown in Figure S2.

Additionally, there are 2 haplotypes showing prefectural specificity, such as the haplotype "c.1.x.6.x.43.x.4.1" (56 entries obtaining A27712G, A3546G, A6592G, C16887T, C29272T, C6636T, G1738A, G26217T, and G28541A substitutions and a GTCATGTTA510 deletion), which were all sampled in Yamaguchi; and the haplotype "1.6.7.1.2.9.x.11.2" (50 entries obtaining A23720G, C15240T, C6433T, C9207T, G10565A, G25489T, G29527T, and G8371T), which were all sampled in Fukuoka. Among all the 7,788 haplotypes of the third wave, 10.8% (841 haplotypes) showed the largest sample size in Tokyo. The latest entry of the "Domestic" B.1.1.214 variant was sampled in Tokyo on July 29, 2021. It belongs to the haplotype "s.1", obtaining 19 SNVs compared to the haplotype "1" (28 SNVs compared to the Wuhan-Hu-1 reference) (Table 3 and Figure 4).

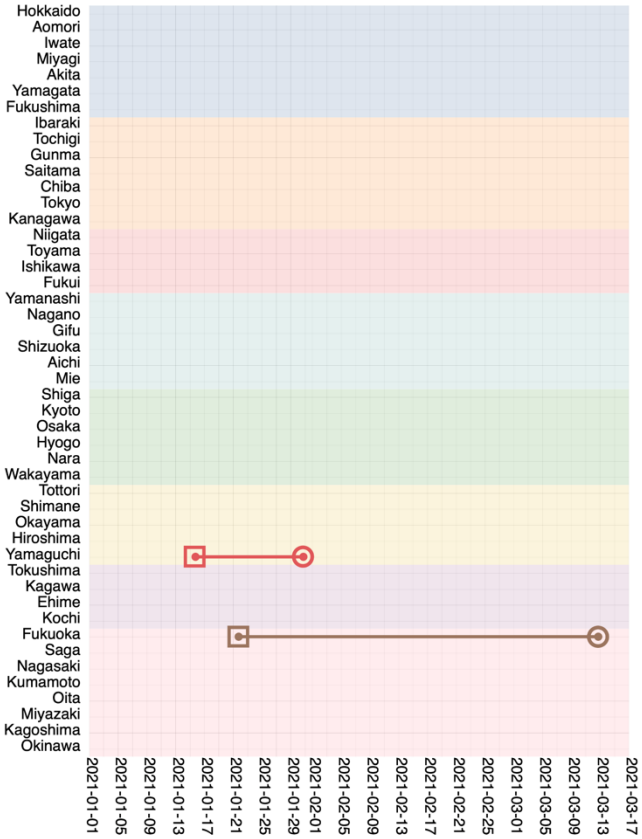

Figure 4. HaploGraph visualization for the haplotypes "c.1.x.6.x.43.x.4.1"and "1.6.7.1.2.9.x.11.2" of the third epidemic wave that showed prefectural specificity. Dots and lines in red and brown indicate the haplotypes "c.1.x.6.x.43.x.4.1" and "1.6.7.1.2.9.x.11.2", respectively. The earliest or latest entry of each haplotype is surrounded by a square or a circle of the corresponding color. The haplotype "c.1.x.6.x.43.x.4.1" was all sampled in Yamaguchi prefecture, and the "1.6.7.1.2.9.x.11.2" was all sampled in Fukuoka prefecture.

4th wave in Japan - Alpha

The major variant in the fourth epidemic wave of Japan was reported to be the Alpha (B.1.1.7) (Hirotsu et al., 2021b); however, there were no reports on Japan-specific Alpha variants to our best knowledge. To trace the major Japan-specific Alpha variants, we compared the 51,863 Japan-related Alpha (51,859 "Domestic" and 4 "International" entries) with 1,117,535 non-Japan-related Alpha (1,117,247 non-Japan and 288 "Quarantine" entries) and revealed 4 nucleotide substitutions which were almost exclusively observed in Japan Alpha variants: G17019T (55.5%), C26464T (25.1%), T23659C (24.8%), and C11173T (11.2%). The 2nd and the 3rd SNVs (C26464T and T23659C) co-occurred with the 1st SNV (G17019T) in 28,761 genomes, while the C11173T in 5,783 genomes was independent. Thus, for the fourth wave, we identified two reference haplotypes for the two major independent lineages (34,544 genomes) which accounted for 66.61% of all Japan

Alpha genomes (51,863). Among the total 11,788 haplotypes of the Japan Alpha variants, the largest one was the reference haplotype “1” containing G17019T (1,648 entries) and the third largest one was the reference haplotype “2” containing C11173T (722 entries). The earliest and latest entries of haplotype “1” were sampled on March 8, 2021, in Hyogo and on July 1, 2021, in Hyogo; and the earliest and latest entries of haplotype “2” were sampled on February 27, 2021, in Hokkaido and on July 9, 2021, in Fukuoka (Table 4 and Figure 5).

Table 4. The representative haplotypes of the fourth wave that showed distinctive features

Haplotype	Main feature	Obtained SNVs compared to “1” or “2”	Date	Location
“1” (Reference)	Largest	-	From 2021-03-08 To 2021-07-01	Hyogo Hyogo
“2” (Reference)	Third largest	-	From 2021-02-27 To 2021-07-09	Hokkaido Fukuoka
“1.265.2”	Second largest	C26464T, T23659C	From 2021-03-13 To 2021-07-20	Osaka Kanagawa/ Ibaraki
“2.112.5”	Fourth largest	A5871G, C2710T	From 2021-06-29 To 2021-09-10	Tokyo Osaka
“1.265.2.145.3”	Prefectural preference	C15783T, C18501T, C26464T, G17019T, T23659C	From 2021-04-26 To 2021-08-19	Mostly sampled in Hokkaido
“1.265.2.454”	Prefectural preference	C26464T, G17019T, G25177T, T23659C	From 2020-04-24 To 2021-05-26	Mostly sampled in Nagasaki
“1.265.2.483.1.4.x.1”	Prefectural preference	C11518T, C2048A, C20719T, C26464T, G17019T, G28079T, T23659C, T27428C	From 2021-04-27 To 2021-05-27	Mostly sampled in Fukushima
“1.193.x.1.x.13.x.23.x.1.2”	Includes the entry with the latest date	12 SNVs	From 2021-09-22 To 2021-09-22	Tokyo Tokyo

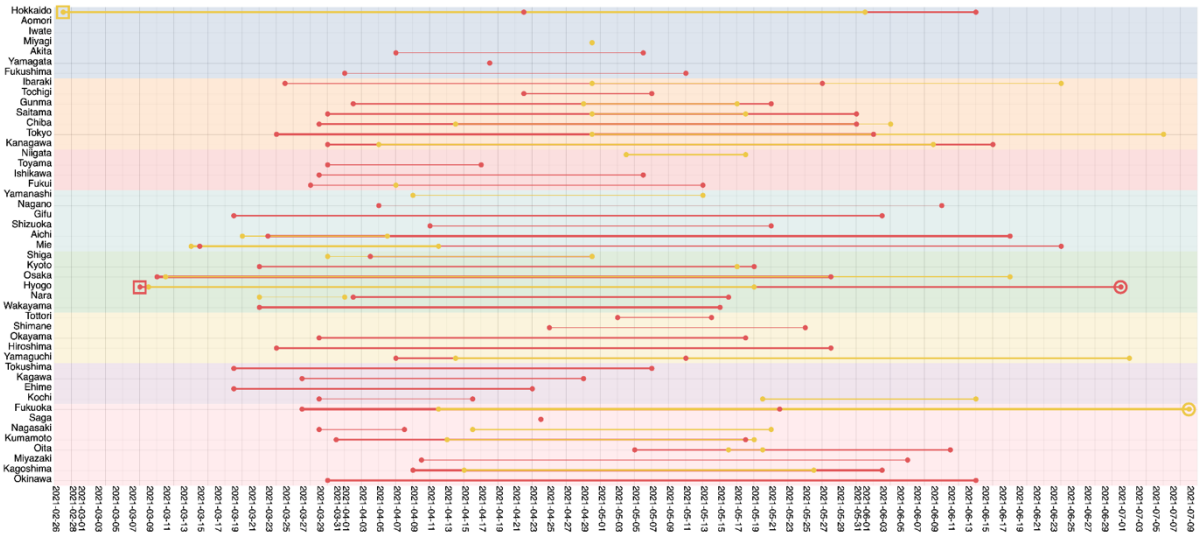

Figure 5. HaploGraph visualization for the haplotypes "1" and "2" of the fourth epidemic wave. Dots and lines in red and yellow indicate the largest haplotype "1" and the third largest haplotype "2", respectively. The earliest or latest entry of each haplotype is surrounded by a square or a circle of the corresponding color. The sampled prefectures and dates of the two haplotypes appear to overlap in a wide range.

The second largest haplotype (1,319 entries) was "1.265.2", obtaining C26464T and T23659C substitutions based on haplotype "1", with the earliest entry sampled on March 13, 2021, in Osaka, and the latest entries sampled on July 20, 2021, in Kanagawa and Ibaraki. The fourth-largest haplotype (587 entries) was "2.112.5", obtaining A5871G and C2710T substitutions based on haplotype "2", with the earliest entry sampled on June 29, 2021, in Tokyo and the latest entry sampled on September 10, 2021, in Osaka (Table 4 and Figure S3). The top 4 largest haplotypes were sampled in various prefectures as shown in Figure 5 and Figure S3.

In addition, there were 3 haplotypes showing prefectural preference: such as for the haplotype "1.265.2.145.3" (obtaining C15783T, C18501T, C26464T, G17019T, and T23659C), 63.5% (216 out of 340 entries) was sampled in Hokkaido; for the haplotype "1.265.2.454" (obtaining C26464T, G17019T, G25177T, and T23659C), 74.1% (100 out of 135 entries) was sampled in Nagasaki; and for the haplotype "1.265.2.483.1.4.x.1" (obtaining C11518T, C2048A, C20719T, C26464T, G17019T, G28079T, T23659C, and T27428C), 89.8% (79 out of 88 entries) was sampled in Fukushima. Among all the 11,788 haplotypes of the fourth wave, 15.8% (1,856 haplotypes) showed the largest sample size in Tokyo. The latest entry of the "Domestic" alpha variant was EPI_ISL_5727361, which was sampled in Tokyo on September 22, 2021. It belongs to the haplotype "1.193.x.1.x.13.x.23.x.1.2",

obtaining 12 SNVs compared to the haplotype “1” (40 SNVs compared to the Wuhan-Hu-1 reference) (Table 4 and Figure S4).

5th wave in Japan - Delta

The fifth wave starting from April 2021 was caused by the Delta variant (PANGO ID as B.1.617.2). A sub-lineage of Delta, AY.29 (also known as B.1.617.2.29), was dominant from July to September of 2022 and was reported to be a Japan-specific Delta variant (Abe and Arita, 2021; Koyama et al., 2022). The AY.29 is defined by the two nucleotide substitutions: C5239T (a synonymous substitution in ORF1ab) and T5514C (a nonsynonymous substitution causing V1750A in ORF1ab) (Abe and Arita, 2021). A sub-lineage of AY.29 called AY.29.1 was further identified, which obtained a nonsynonymous substitution G22081T (Q173H in S protein) compared to AY.29. Another sub-lineage of AY.29 called AY.29.2 was found and characterized by a nonsynonymous substitution A22803G (Q414R in S protein) (GitHub, 2021) (Table 5 and Figure 6).

Table 5. The representative haplotypes of the fifth wave that showed distinctive features

Haplotype	Main feature	Obtained SNVs compared to “1”	Date	Location
“1” (Reference)	Largest	-	From 2021-05-18 To 2021-10-13	Kanagawa Hyogo
“1.501.7”	Second largest	C5365T, C28170T	From 2021-06-04 To 2021-09-21	Kanagawa Miyagi
“1.784”	Third largest and the original AY.29.1	G22081T	From 2021-07-12 To 2021-10-05	Tokyo/Saitama Tokyo
“1.501.7.212.x.7.1”	The original AY.29.2	A22803G, G25552T, C5365T, C25701T, C11986A, C28170T	From 2021-08-04 To 2021-09-17	Tokyo Saitama
“1.262.5.1”	Prefectural preference	A28492G, A7886G, C15352T	From 2021-07-29 To 2021-08-31	Mostly sampled in Okayama
“1.501.7.174”	Prefectural preference	A8843G, C28170T, C5365T	From 2021-08-04 To 2021-09-13	Mostly sampled in Shizuoka

“1.501.7.638.2.4.1”	Prefectural preference	C17678T, C24374T, C28170T, C28724T, C5365T, C7420T	From 2021-07-26 To 2021-08-26	Mostly sampled in Tottori
“1.554.53”	Regional preference	C3923A, G11083T	From 2021-07-14 To 2021-09-15	Kinki region
“1.501.7.212.25”	Regional preference	C11986A, C28170T, C5365T, G3340T	From 2021-07-11 To 2021-09-08	
“1.554.80.47”	Regional preference	C3923A, G26230T, T26987C	From 2021-07-15 To 2021-09-06	Kanto region
“1.554.80”	Regional preference	C3923A, T26987C	From 2021-07-12 To 2021-09-11	
“1.501.7.1119”	Regional preference	C28170T, C5365T, dG21987A	From 2021-06-11 To 2021-09-01	
“1.207.75”	Regional preference	C12462T, G25565T	From 2021-07-21 To 2021-09-07	Kyushu region
“i.259”	Regional preference	C11776T, C26622T, C28170T, C28720T, C5365T, CC14407TT, G22785T, G28895T	From 2021-08-08 To 2021-09-09	
“1.501.7.278.3.1.x.2”	Regional preference	A10323G, A14351G, A28699T, C1627T, C28170T, C346T, C5365T	From 2021-11-08 To 2022-01-15	
“1.501.7.439.37.1.2”	Regional preference	C17012T, C21005T, C25207T, C28170T, C5365T, G21974C	From 2021-08-18 To 2021-09-24	Tohoku region
“1.501.7.291.2.17.2”	Regional preference	C13680T, C16887T, C28170T, C5365T, G22381T, T12430C	From 2021-08-03 To 2021-09-13	

“1.240.2.166”	Regional preference	C1376T, C1437T, T29302C	From 2021-08-06 To 2021-09-22	Chubu region
“q.2”	Includes the entry with the latest date	16 SNVs	From 2022-03-10 To 2022-03-10	Tokyo Tokyo

The genome isolated at the Airport Quarantine Station of Japan, EPI_ISL_1927416, was reported as the ancestor of the AY.29 lineage (Abe and Arita, 2021; Koyama et al., 2022), which we reconfirmed by comparing nucleotide substitution patterns (Table S2) as well as phylogenetic analysis using more recent GISAID data (Figure S5).

For the haplotype analysis, we first obtained 94,788 entries of “Domestic” Delta (PANGO ID as AY.29, AY.29.1, and AY.29.2). After removing genomes containing undetermined nucleotides (N) or mixed nucleotides, 80,817 genomes remained. Based on the mutation patterns compared with the ancestral genome EPI_ISL_1927416, 33,917 haplotypes were obtained for the “Domestic” Delta. The haplotype “1” having the largest number (2,709 entries) was found to be the original AY.29 with C5239T and T5514C substitutions compared to EPI_ISL_1927416. We chose this haplotype as the reference for the fifth wave. The earliest and the latest entries of haplotype “1” were sampled on May 18, 2021, in Kanagawa and on October 13, 2021, in Hyogo, respectively (Table 5 and Figure 6).

The second largest haplotype was “1.501.7” (2,448 entries), obtaining C5365T and C28170T substitution based on the haplotype “1”, with the earliest entries sampled on June 4, 2021, in Kanagawa and the latest entry sampled on September 21, 2021, in Miyagi. The third largest haplotype was the original haplotype of AY.29.1, which obtained G22081T (in S protein; Q173H) based on the haplotype “1”. The original haplotype of AY.29.1 lineage was named as “1.784” (738 entries), with the earliest entries sampled on July 12, 2021, in Tokyo and in Saitama, and the latest entry EPI_ISL_7046216 sampled on October 5, 2021, in Tokyo (Figure 6). For the AY.29.2 lineage, we found that there are 5 more SNVs fixed in this lineage compared to AY.29 other than the reported A22803G, such as G25552T, C5365T, C25701T, C11986A, and C28170T. Based on these fixed SNVs we named the haplotype of the original AY.29.2 as “1.501.7.212.x.7.1”, with the earliest entry sampled on August 4, 2021, in Tokyo and the latest entry sampled on September 17, 2021, in Saitama (Table 5 and Figure 6).

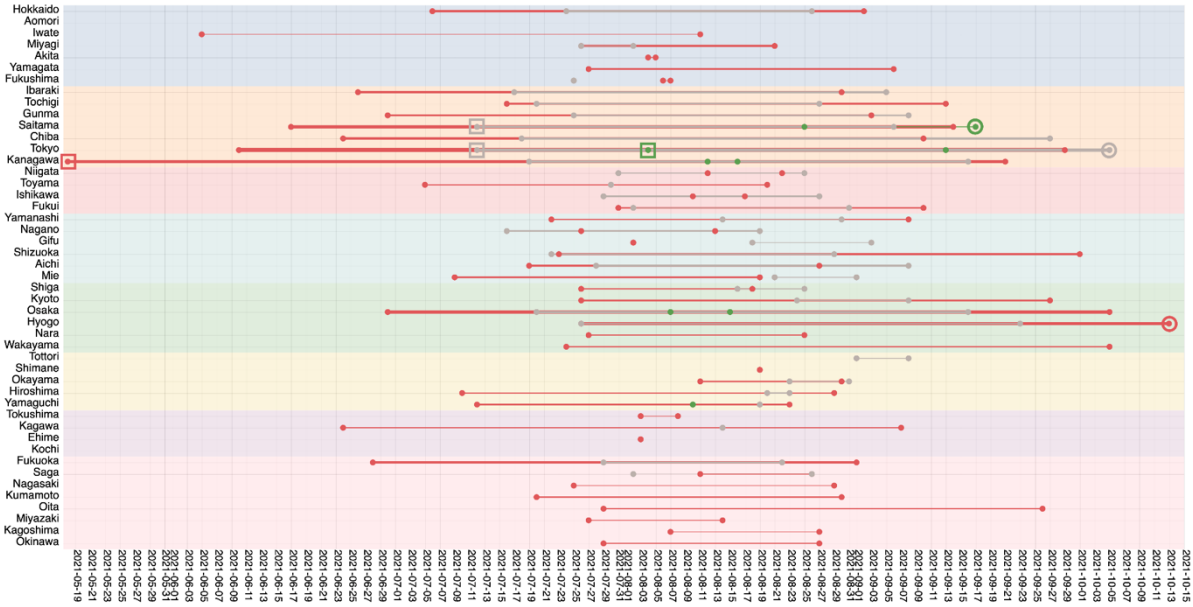

Figure 6. HaploGraph visualization for the original haplotypes of "AY.29", "AY.29.1", and "AY.29.2" lineages of the fifth epidemic wave. Dots and lines in red, grey, and green indicate the original haplotypes of AY.29 (Haplotype "1"), AY.29.1 (Haplotype "1.784"), and AY.29.2 (Haplotype "1.501.7.212.x.7.1"), respectively. The earliest or latest entry of each haplotype is surrounded by a square or a circle of the corresponding color.

Eleven of the 115 haplotypes with more than 50 entries showed more than 80% occurrence in particular regions of Japan (Figure S6). Among the 115 haplotypes, three showed prefectural preference (more than 90% was sampled in one prefecture): such as the haplotype "1.262.5.1" (obtaining A28492G, A7886G, and C15352T) in Okayama (64 out of 69 entries); the haplotype "1.501.7.174" (obtaining A8843G, C28170T, and C5365T) in Shizuoka (57 out of 58 entries); and the haplotype "1.501.7.638.2.4.1" (obtaining C17678T, C24374T, C28170T, C28724T, C5365T, and C7420T) in Tottori (49 out of 54 entries). Among all 33,917 haplotypes of the fifth wave, 21.26% showed the largest sample size in Tokyo. The latest entry of the "Domestic" delta variant was EPI_ISL_12036503, sampled in Tokyo on March 10, 2022. It belongs to the haplotype "q.2", obtaining 16 SNVs compared to the haplotype "1" (54 SNVs compared to the Wuhan-Hu-1 reference genome) (Table 5 and Figure 7).

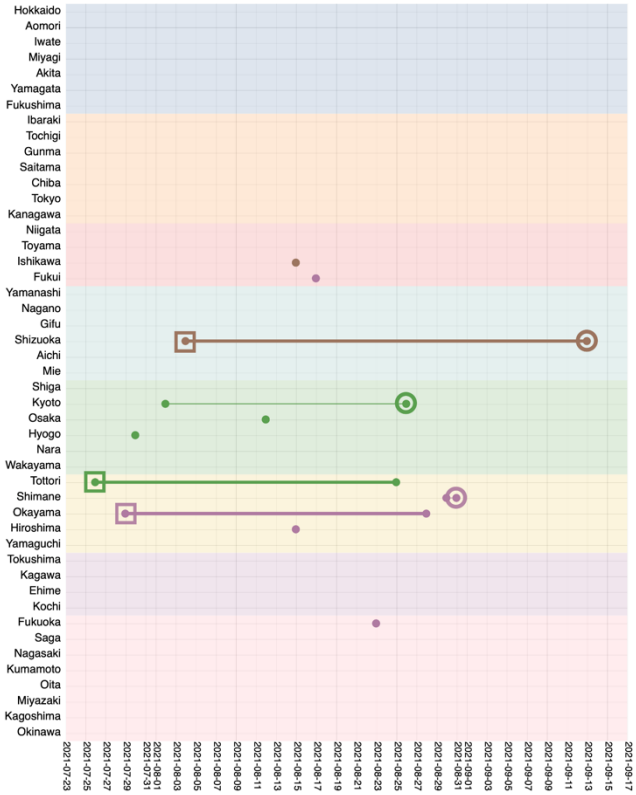

Figure 7. HaploGraph visualization for the haplotypes “1.262.5.1”, “1.501.7.174”, and “1.501.7.638.2.4.1” of the fifth epidemic wave that showed prefectural preference. Dots and lines in purple, brown, and green indicate the haplotypes “1.262.5.1”, “1.501.7.174”, and “1.501.7.638.2.4.1”, respectively. The earliest or latest entry of each haplotype is surrounded by a square or a circle of the corresponding color. The haplotype “1.262.5.1” was mostly sampled in Okayama, the “1.501.7.174” was mostly sampled in Shizuoka, and the “1.501.7.638.2.4.1” was mostly sampled in Tottori.

6th wave in Japan - Omicron

The sixth wave starting in December 2021 was caused by the Omicron variant (PANGO ID as B.1.1.529.* or BA.*). A major characteristic of the sixth wave was the large variety of multiple Omicron sub-lineages transmitted from overseas, rather than one sub-lineage becoming dominant, which had been the case for the second to the fifth waves. This makes tracing the haplotypes for the sixth wave difficult, and we did not include the Omicron haplotype dynamics in the “HaploGraph” viewer.

The weekly trends of the percentage of Omicron sub-lineages in Japan from December 2021 to May 2022 are shown in Figure 8. The total number of Omicron entries reached more than 100 from the 50th week of 2021 and it continued till the 19th week of 2022.

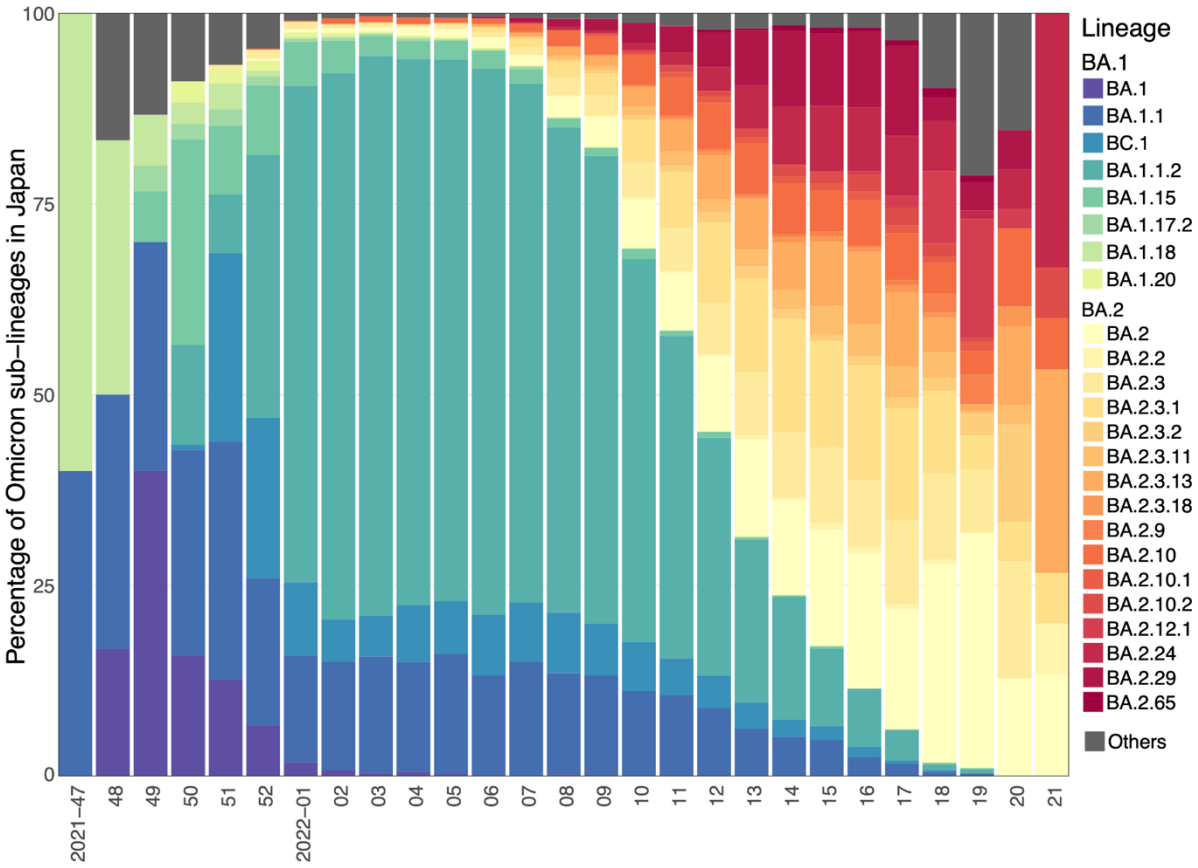

Figure 8. Weekly trends of the Omicron sub-lineages in Japan. The X-axis represents the time span from the 47th week of 2021 to the 21st week of 2022. The Y-axis represents the percentage of Omicron sub-lineages in Japan. Each color represents an Omicron sub-lineage which is described on the right. Note that the Omicron sub-lineages having less than 100 entries are all filtered into the “Others” category.

During the period shown in Figure 8, 134 sub-lineages of Omicron were detected in Japan. The top five of them were BA.1.1.2 (53.9 %), BA.1.1 (12.0 %), BC.1 (also as BA.1.1.1.1) (5.9 %), BA.2 (4.8 %) and BA.2.3.1 (4.0 %), respectively. In addition, other sub-lineages accounted for 19.4% of the total. There were 6 sub-lineages (BA.2.3, BA.2.10, BA.2.29, BA.1.15, BA.2.3.13, and BA.2.24, in the ascending order of the number of entries) in which more than 1000 entries were detected. There were also 13 sub-lineages in which more than 100 entries were detected. Three of the top five sub-lineages, BA.1.1.2, BA.1.1, and BA.2, were first detected as Japan quarantine strains. Most of the other Omicron sub-lineages had also been transmitted from overseas and were spreading in Japan. In contrast, the remaining two of the top 5, BC.1 (also as BA.1.1.1.1) and BA.2.3.1, obtained the unique mutations that were specific to Japan and were registered as Japan lineages in PANGO nomenclature. The BC.1 has a nonsynonymous substitution A22259G (in S protein; I233V) compared to its parental lineage BA.1.1.1, and the earliest entry was sampled in Osaka on

December 18, 2021. The BA.2.3.1 has a nonsynonymous substitution C23591G (in S protein; Q677E) compared to its parental lineage BA.2.3, and the earliest entry was sampled in Ibaraki on January 10, 2022.

The weekly trends of the Omicron sub-lineages showed that the largest sub-lineage BA.1.1.2 became predominant after the 52nd week of 2021 and peaked in the 3rd week of 2022 when it accounted for 73.4% of all entries. From the 12th week of 2022, the entry number of BA.2 and its sub-lineages became larger than that of BA.1 and its sub-lineages. Thus, during the sixth wave in Japan, there was a shift from the BA.1 and its sub-lineages to the BA.2 and its sub-lineages.

Discussion

The high sequence quality of SARS-CoV-2 genomes enabled us to analyze their nucleotide substitutions and to trace the spread of COVID-19. Here we described the dynamics of SARS-CoV-2 haplotypes in Japan using SARS-CoV-2 HaploGraph, a new web resource for visualization. For the HaploGraph, it is important noting the followings: 1) we used the GISAID database, which restricts users from providing detailed information on nucleotide substitutions which are derived from the database; 2) not all genomes of the SARS-CoV-2 that spread in Japan were sampled, sequenced, and stored in GISAID; 3) quality of genomes could differ depending on the sequencing laboratories; and 4) recombination of variants were not considered. For the first point, according to the GISAID term of use, the data stored in the GISAID should not be displayed or accessed through a separate portal or a network of institutions (https://www.gisaid.org/registration/terms-of-use/; Arita 2021). This policy makes it difficult to provide details on the mutations of the SARS-CoV-2 variants on our SARS-CoV-2 HaploGraph browser. Indeed, the HaploGraph browser is based on the processed statistics of GISAID information that is available at the GitHub website, detail information of which cannot be provided. For the second point, the sampling of SARS-CoV-2 genomes could also be biased depending on the area and dates. Therefore, when and where a variant was detected for the first time is not necessarily the location and date of the first emergence of the variant. On the other hand, the number of COVID-19 patients and the number of sequenced SARS-CoV-2 genomes were well correlated for the periods of the 1st to the 5th waves in Japan (Figure 9; Pearson's correlation coefficient, r = 0.962, P-value < 0.001). This observation suggests that haplotypes visualized by "HaploGraph" could roughly represent the transmission of SARS-CoV-2 in Japan, although more than 90% of SARS-CoV-2 genomes sampled from COVID-19 patients were not sequenced (Figure 9). For the third point, in this study, we used the "complete genomes" — genomes without any undetermined and mixed nucleotides— for the identification of all SNVs in each genome. Nonetheless, the sequencing quality of Japan on SARS-CoV-2 genomes is considerably high compared to that of the overseas. For instance, the proportion of the sequences that contain 0 or 10 undefined nucleotides reached 89.4% or 91.1% in the entries sequenced in Japan, while it was 36.0% or 48.0% in the entries sequenced in other countries. This was one important factor that made this study possible with reliable results. For the fourth point, some variants may emerge though recombination events. It is difficult to distinguish whether mutation(s) emerged by recombination or parallel evolution when the number of mutations is limited. Indeed, many recombination events are reported for the SARS-CoV-2 genomes, particularly for the Omicron variants (Ou et al., 2022). Since various nucleotide substitutions accumulated in the genomes of Omicron variants, recombination of Omicron variants could be easily detected. Before the Omicron strain arose, one of the few

examples of recombinant strains reported was the XC variant, a recombination of the Delta and Alpha mutants detected in Japan (Sekizuka et al., 2022). The lack of consideration of recombination is one of the limitations of this study.

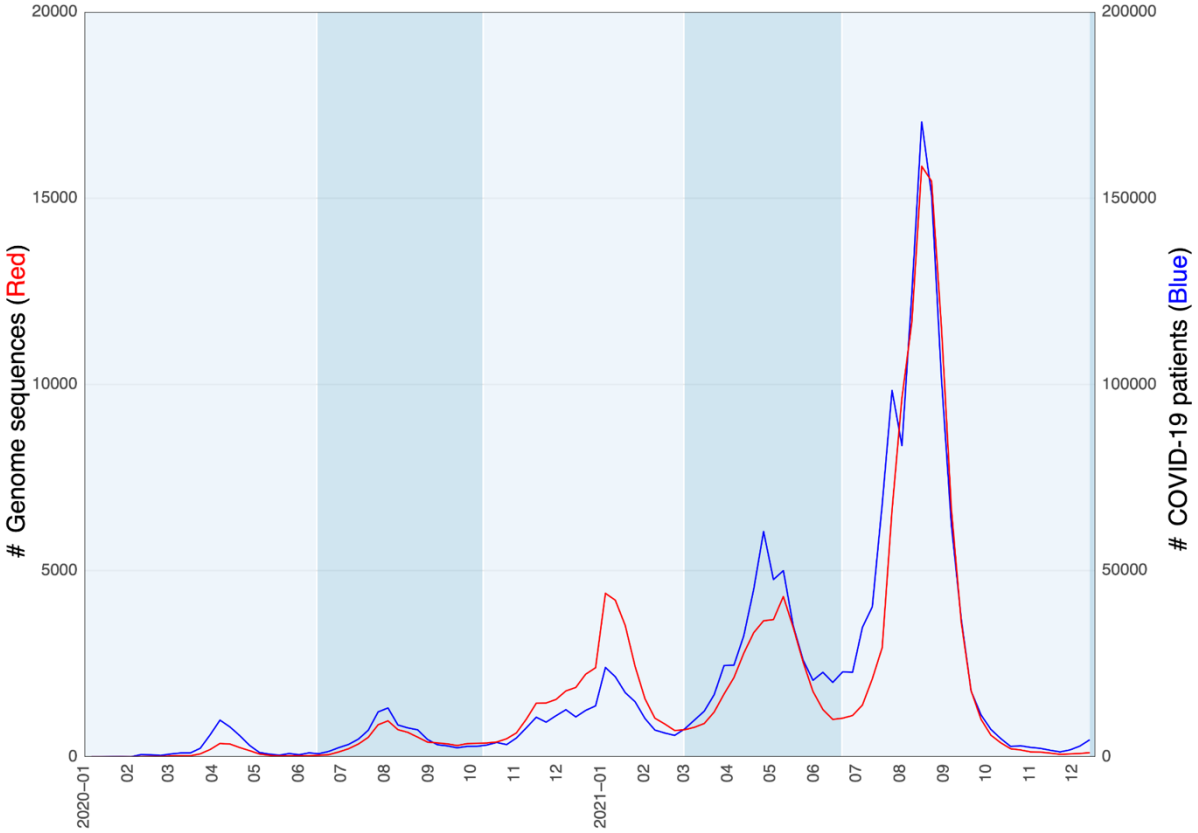

Figure 9. Weekly counts for the number of SARS-CoV-2 genomes and COVID-19 patients in Japan. The X-axis represents the time span from the first to the fifth waves in Japan (i.e., January 1, 2020, to December 16, 2021). The light blue and blue colors in the background depict the five wave periods. The left and right Y axes represent the number of genomes stored in GISAID (shown in red) and the number of COVID-19 patients (shown in blue) obtained from OurWorldInData.org, respectively. Note that the scale of the left and right Y axis is 10 times different. Those numbers were counted weekly.

In this study, we did not analyze R.1 lineage, which was not the main cause of the fourth epidemic wave of Japan, but spread on a considerably large scale (originally assigned as B.1.1.316). The R.1 lineage contains potential escape mutations in the Spike (S) protein receptor-binding domain (E484K) and N-terminal domain (W152L) (Hirotsu et al., 2021a). It was reported that the R.1 was originated from the European strain B.1.1.114 from March to April 2020 by acquiring 13 mutations (NIID, 2021b), although there was no evidence showing those mutations were obtained in Japan. The number of R.1 entries sampled in each wave period is summarized in Table S3. It shows that 87.2% of R.1 entries were sampled during the period of the fourth wave. It explains one of the reasons why the

580 percentage of the Alpha entries (PANGO ID: B.1.1.7 and Q.*) analyzed in our study was
581 relatively low in the fourth wave period compared to that of the dominant lineages in other
582 wave periods.
583 Our study clearly indicated that, although major SARS-CoV-2 lineages differed
584 depending on epidemic waves in Japan, some remaining variants existed in the next or in
585 the further epidemic waves (Tables 2 - 5). Such haplotypes were not traceable for each
586 mutation step (Tables 2 - 5). This observation suggests that the major variants did not
587 disappear after the epidemic wave and continued to remain in small numbers. This
588 observation was also suggested by other studies (Nabeshima et al., 2021; Ode et al., 2022).
589 Our haplotype analysis also showed the long-lasting transmission of haplotypes without
590 mutation. For example, the original haplotype of AY.29, which did not have a single
591 mutation, continued to circulate in Japan for as long as 5 months (Figure 6 and Table 5). It is
592 known that SARS-CoV-2 shows relatively lower mutation rate mutation partially owing to
593 nsp14 (Robson et al., 2020, Peacock et al., 2021), although amino acid replacement(s) on
594 nsp14 may affect SARS-CoV-2 mutation rates (Takada et al. 2022). Further, a substantial
595 selection of mutations associated with immune escape and binding affinity to the receptor
596 has characterized genomes of SARS-CoV-2, the trend of which is expected to continue
597 (Kistler et al. 2022; Carabelli et al. 2023). Therefore, it remains important to monitor genomic
598 mutations of SARS-CoV-2.
599 COVID-19 continues to spread worldwide, and over 14.8 million SARS-CoV-2
600 genomes were registered in GISAID as of 1/31, 2023 (https://www.gisaid.org; Khare et al.,
601 2021). However, due to the GISAID term of use (https://www.gisaid.org/registration/terms-of-
602 use/), the non-registered general public cannot use the genome sequences and the related
603 data stored in the GISAID. In addition, even for the registered researchers it is not easy to
604 use the GISAID data due to its huge size and inefficient compression (Kryukov et al., 2022).
605 Thus, the GISAID data from which the valuable information can be derived, including the
606 transmission patterns, may remain difficult to access. As shown in this study, SARS-CoV-2
607 HaploGraph allows users to understand a graphical representation of the transmission of
608 COVID-19 in Japan based on SARS-CoV-2 genome information without the need to perform
609 detailed GIASID data analysis. The methodology and platform of SARS-CoV-2 HaploGraph
610 can also be applied to future genomic studies of other infectious diseases.

Materials and Methods

Dataset

SARS-CoV-2 genome sequences and their annotation information used in this study were downloaded from the GISAID EpiCoV database (https://www.gisaid.org) as of June 4, 2022 (11,193,128 sequences) (Khare et al., 2021; Elbe and Buckland-Merrett, 2017; Shu et al., 2017). The GISAID IDs and related information used in this study was summarized in the following GISAID website: https://doi.org/10.55876/gis8.221004wz. The daily number of COVID-19 positive patients in Japan was obtained from Our World in Data (https://ourworldindata.org/) as of October 19, 2022. The period of each epidemic wave of COVID-19 in Japan was obtained from the documents of the Ministry of Health, Labour and Welfare (MHLW) of Japan (MHLW, 2022) as summarized in Table 1.

Haplotype Naming System

We obtained the variations for each SARS-CoV-2 genome stored in GISAID by aligning the genome sequences to the reference genome Wuhan-Hu-1 (GenBank ID: NC_045512) using MAFFT version 7.490 with default options (Katoh et al., 2019). Based on the pairwise alignment, mutations on different nucleotide sites were identified as substitution, insertion, or deletion. We then obtained the haplotype of each genome sequence by classifying the single nucleotide variants (SNVs) contained in the genome. Note that continuous indels are treated as one SNV in this study because they could be caused by a single mutation.

For each SARS-CoV-2 epidemic wave in Japan, the reference haplotype was determined as the sequence with an early sampling date with the minimum number of SNVs. The ID for each haplotype was assigned as follows. First, the reference haplotype was named "1" and if there were more than one reference haplotype in a lineage, we increased the number (*i.e.*, 2, 3, …). Then, when a haplotype acquires one SNV, a number is added and connected to the first number with a dot (*i.e.*, "."). For example, "1.*N*" (where *N* is any number) indicates the acquisition of one specific mutation from haplotype 1. Backward mutations when an SNV reverts back to the Wuhan-Hu-1 wild-type haplotype were also counted. In such cases, the haplotype was started by a lowercase "d" before the backward mutation. For example, a mutation “dG26849T” indicates the haplotype contains a backward mutation “G” at position 26849 (*i.e.*, the genotype is the same as that of Wuhan-Hu-1), although the reference haplotype of the lineage has the mutation “T” at the same position. In cases where there were multiple candidates for an immediate ancestor of a haplotype, we selected the oldest haplotype (the one sampled on the earliest date), and the other haplotype(s) are concatenated with a pipe (*i.e.*, “|”). If an immediate descendant with a single mutational difference was not found, a two-step forward descendant was searched. When it

was found, an "x" was used to indicate the unknown intermediate state. For example, "1.2.x.1" is a haplotype with two SNVs from haplotype 1.2. If two intermediate states were not found, the haplotype would start with a letter in alphabetical order according to the number of SNVs compared to the reference haplotype of the lineage. For example, a haplotype “c.1” has three SNVs compared to the reference haplotype and its ancestors are not traceable within 1 or 2 SNVs till the reference haplotype.

Phylogenetic analysis

Multiple alignments of SARS-CoV-2 genomes were generated using the following two methods: 1) if the number of genomes is over 50,000, we applied minimap2 version 2.24 (Li et al., 2018) to construct pairwise alignment to the Wuhan-Hu-1 reference sequence, then applied gofasta version v1.0.0 (GitHub, 2022) to build a multiple alignment based on the pairwise alignments; 2) if this is not the case, we used MAFFT version 7.490 (Katoh et al., 2019) with --6merpair and --addfragments options to construct the multiple alignments. Using the multiple alignment, we constructed a maximum-likelihood-based phylogenetic tree using IQtree version 2.1.3 (Nguyen et al., 2015) with the following options: -m GTR+I+G -B 1000. The phylogenetic tree was visualized by ggtree version 2.4.1 (Yu et al., 2020).

Haplotype visualization

To visualize the mutation and propagation of each SARS-CoV-2 haplotype, we developed the HaploGraph web application. HaploGraph is designed for tracing when and where each haplotype was firstly or lastly detected, when and where the haplotype was observed during the wave period, and for how long the haplotype was recorded. These haplotypes are distinguished by the accumulated single nucleotide variations as described in the previous section; they can be tracked from the ancestral haplotypes to the descendants. For each epidemic wave, users can choose one or more haplotypes from the list, select a region, or specify a range of dates of interest on the user interface of the HaploGraph (please see Figure 1 as well). Toggle buttons enable users to show 1) “trace” for indicating propagation of the identical haplotypes, and 2) “duration” for presenting how long the haplotype was observed. The source code of the HaploGraph is available at the GitHub repository (https://github.com/ktym/covid19/) that is implemented with the D3.js (https://d3js.org/) and Bootstrap (https://getbootstrap.com/) libraries.

Acknowledgments

We thank Koichi Kitao for his technical assistance. We gratefully acknowledge all data contributors: scientists and medical staff, their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative: doi.org/10.55876/gis8.221004wz.

Funding

This work was supported by the JST (Japan Science and Technology Agency) CREST JPMJCR20H6 (to S.N.) and JPMJCR20H1 (to T.A., K.K., S.A., M.A.), by AMED (Japan Agency for Medical Research and Development) JP20he0622043 and JP20fk0108536 (to T.K.), and by a grant from ROIS (Research Organization of Information and Systems) (to J.O.).

References:

Abe, T., and Arita, M. (2021) Genomic Surveillance in Japan of AY.29—A New Sub-lineage of SARS-CoV-2 Delta Variant with C5239T and T5514C Mutations. medRxiv doi: https://doi.org/10.1101/2021.09.20.21263869

Arima, Y., Kanou, K., Arashiro, T., K Ko, Y., Otani, K., Tsuchihashi, Y., Takahashi, T., Miyahara, R., Sunagawa, T., and Suzuki, M. (2021) Epidemiology of Coronavirus Disease 2019 in Japan: Descriptive Findings and Lessons Learned through Surveillance during the First Three Waves. JMA journal **4**, 198–206.

Arima, Y., Shimada, T., Suzuki, M., Suzuki, T., Kobayashi, Y., Tsuchihashi, Y., Nakamura, H., Matsumoto, K., Takeda, A., Kadokura, K., et al. (2020) Severe Acute Respiratory Syndrome Coronavirus 2 Infection among Returnees to Japan from Wuhan, China, 2020. Emerg. Infect. Dis. **26**, 1596–600.

Arita, M. (2021) Open access and data sharing of nucleotide sequence data. Data Science Journal, 20(1), 28. DOI: http://doi.org/10.5334/dsj-2021-028

Carabelli, A. M., Peacock, T. P., Thorne, L. G., Harvey, W. T., Hughes, J., Silva, T. I. de, Peacock, S. J., Barclay, W. S., Silva, T. I. de, Towers, G. J. et al. (2023) SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nat. Rev. Microbiol. In press.

Challen, R., Brooks-Pollock, E., Read, J. M., Dyson, L., Tsaneva-Atanasova, K., and Danon, L. (2021) Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. BMJ. **372**, n579.

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., et al. (2021) Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science **372**, eabg3055.

Davies, N. G., Jarvis, C. I., CMMID COVID-19 Working Group, Edmunds, W. J., Jewell, N. P., Diaz-Ordaz, K., and Keogh, R. H. (2021) Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. Nature **593**, 270–274.

Desingu, P. A., Nagarajan, K., and Dhama, K. (2022) Emergence of Omicron third lineage BA.3 and its importance. J. Med. Virol. **94**, 1808–1810.

Earnest, R., Uddin, R., Matluk, N., Renzette, N., Turbett, S. E., Siddle, K. J., Loreth, C., Adams, G., Tomkins-Tinch, C. H., Petrone, M. E., et al. (2022) Comparative transmissibility of SARS-CoV-2 variants Delta and Alpha in New England, USA. Cell Rep. Med. **3**, 100583.

Elbe, S., and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob. Chall. **1**, 33–46.

GitHub. (2021) Available online at: https://github.com/cov-lineages/pango-designation/issues/381 (accessed July 13, 2022).

GitHub. (2022) Available online at: https://github.com/virus-evolution/gofasta (accessed July 13, 2022).

Graham, M. S., Sudre, C. H., May, A., Antonelli, M., Murray, B., Varsavsky, T., Kläser, K., Canas, L. S., Molteni, E., Modat, M., et al. (2021) Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. Lancet Public Health **6**, e335–e345.

Grint, D. J., Wing, K., Williamson, E., McDonald, H. I., Bhaskaran, K., Evans, D., Evans, S. J., Walker, A. J., Hickman, G., Nightingale, E., et al. (2021) Case fatality risk of the SARS-CoV-2 variant of concern B.1.1.7 in England, 16 November to 5 February. Euro Surveill. **26**, 2100256.

Gupta, N., Das, M., Rakshit, P., Singh, S., Abraham, P., Panda, S., and Team, N. (2021) SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. Microorganisms, **9**, 1542.

Harrison, A. G., Lin, T., and Wang, P. (2020) Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. Trends Immunol. **41**, 1100–1115.

Hirotsu, Y., and Omata, M. (2021a) Detection of R.1 lineage severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with spike protein W152L/E484K/G769V mutations in Japan. PLoS Pathog. **17**, e1009619.

Hirotsu, Y., and Omata, M. (2021b) SARS-CoV-2 B.1.1.7 lineage rapidly spreads and replaces R.1 lineage in Japan: Serial and stationary observation in a community. Infect. Genet. Evol. **95**, 105088.

Iketani, S., Liu, L., Guo, Y., Liu, L., Chan, J. F., Huang, Y., Wang, M., Luo, Y., Yu, J., Chu, H., et al. (2022) Antibody evasion properties of SARS-CoV-2 Omicron sublineages. Nature **604**, 553–556.

Ito, K., Piantham, C., and Nishiura, H. (2021) Predicted dominance of variant Delta of SARS-CoV-2 before Tokyo Olympic Games, Japan, July 2021. Euro Surveill. **26**, 2100570.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. **20**, 1160–1166.

Khailany, R. A., Safdar M., and Ozaslan M. (2020) Genomic characterization of a novel SARS-CoV-2. Gene Rep. **19**, 100682.

Khare, S., Gurry, C., Freitas, L., Schultz, M. B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R. T., Yeo, W., et al. (2021) GISAID's Role in Pandemic Response. China CDC Wkly. **3**, 1049–1051.

Kistler, K. E., Huddleston, J. and Bedford, T. (2022) Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. Cell Host Microbe **30**, 545-555.e4.

Koyama, T., Tokumasu, R., Katayama, K., Saito, A., Kudo, M., and Imoto, S. (2022) Cross-Border Transmissions of the Delta Substrain AY.29 During Tokyo Olympic and Paralympic Games. Front. Microbiol. **13**, 883849.

Kryukov K., Jin L., and Nakagawa S. (2022) Efficient compression of SARS-CoV-2 genome data using Nucleotide Archival Format (NAF). Patterns **3**, 100562.

Leung, K., Shum, M. H., Leung, G. M., Lam, T. T., and Wu, J. T. (2021) Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. Euro Surveill. **26**, 2002106.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**, 3094–3100.

Luan, B., Wang, H., and Huynh, T. (2021) Enhanced binding of the N501Y-mutated SARS-CoV-2 spike protein to the human ACE2 receptor: insights from molecular dynamics simulations. FEBS Lett. **595**, 1454–1461.

Meng, B., Abdullahi, A., Ferreira, I. A. T. M., Goonawardane, N., Saito, A., Kimura, I., Yamasoba, D., Gerber, P. P., Fatihi, S., Rathore, S., et al. (2022) Altered TMPRSS2 usage by SARS-CoV-2 Omicron impacts infectivity and fusogenicity. Nature **603**, 706–714.

Ministry of Health, Labour and Welfare (MHLW). (2020) The first patient with pneumonia associated with novel coronavirus (in Japanese). Available online at: https://www.mhlw.go.jp/stf/newpage_08906.html (accessed July 13, 2022).

Ministry of Health, Labour and Welfare (MHLW). (2021) Asymptomatic carriers of a SARS-CoV-2 infection (mutant variant) (Airport Quarantine) in Japan (in Japanese). Available online at: https://www.mhlw.go.jp/stf/newpage_22507.html (accessed July 13, 2022).

Ministry of Health, Labour and Welfare (MHLW). (2022) The 81st SARS-CoV-2 Infectious Disease Control Advisory Board (in Japanese). Available online at: https://www.mhlw.go.jp/content/10900000/000931584.pdf (accessed July 13, 2022).

Mlcochova, P., Kemp, S. A., Dhar, M. S., Papa, G., Meng, B., Ferreira, I. A. T. M., Datir, R., Collier, D. A., Albecka, A., Singh, S., et al. (2021) SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. Nature **599**, 114–119.

Morens, D. M., Breman, J. G., Calisher, C. H., Doherty, P. C., Hahn, B. H., Keusch, G. T., Kramer, L. D., LeDuc, J.W., Monath, T.P., and Taubenberger, J. K. (2020) The Origin of COVID-19 and Why It Matters. Am. J. Trop. Med. Hyg. **103**, 955–959.

Nabeshima, T., Takazono, T., Ashizawa, N., Miyazaki, T., Inoue, S., Ngwe Tun, M. M., Izumikawa, K., Mukae, H., Moi, M. L., and Morita, K. (2021) COVID-19 cryptic transmission and genetic information blackouts: Need for effective surveillance policy to better understand disease burden. Lancet Reg. Health West Pac. **7**, 100–104.

National Institute of Infectious Diseases (NIID). (2021a) Molecular epidemiological survey using genomic information of the new coronavirus SARS-CoV-2 (as of January 14, 2021). IASR Vol. 42 p61-64: March 2021 issue (in Japanese). Available online at:

https://www.niid.go.jp/niid/ja/2019-ncov/2488-idsc/iasr-news/10152-493p01.html (accessed July 13, 2022).

National Institute of Infectious Diseases (NIID). (2021b) Domestic influx of B.1.1.316 strain carrying the new coronavirus SARS-CoV-2 Spike protein E484K mutation as of February 2, 2021 (in Japanese). Available online at: https://www.niid.go.jp/niid/ja/diseases/ka/corona-virus/2019-ncov/2488-idsc/iasr-news/10188-493p02.html (accessed July 29, 2022).

National Institute of Infectious Diseases (NIID). (2022) About SARS-CoV-2 mutant strain B.1.1.529 strain – Omicron strain (in Japanese). Available online at: https://www.niid.go.jp/niid/ja/2019-ncov/2551-cepr/11029-cepr-b11529-9.html (accessed July 13, 2022).

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32**, 268–274.

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., and Agha, R. (2020) The socio-economic implications of the coronavirus pandemic (COVID-19): A review. Int. J. Surg. **78**, 185–193.

Ode, H., Nakata, Y., Nagashima, M., Hayashi, M., Yamazaki, T., Asakura, H., Suzuki, J., Kubota, M., Matsuoka, K., Matsuda, M., et al. (2022) Molecular epidemiological features of SARS-CoV-2 in Japan, 2020-1. Virus Evol. **8**, veac034.

Okumura, N., Tsuzuki, S., Saito, S., Hattori, S. I., Takeuchi, J. S., Saito, T., Ujiie, M., Hojo, M., Iwamoto, N., Sugiura, W., Mitsuya, H., and Ohmagari, N. (2022) Neutralising activity and antibody titre in 10 patients with breakthrough infections of the SARS-CoV-2 Omicron variant in Japan. J. Infect. Chemother. **28**, 1340–1343.

Ou, J., Lan, W., Wu, X., Zhao, T., Duan, B., Yang, P., Ren, Y., Quan, L., Zhao, W., Seto, D., et al. (2022) Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. Signal Transduct. Target Ther. **7,** 138.

Outbreak.info. (2023) Available online at: https://outbreak.info/ (accessed January 13, 2023)

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., and Pybus, O. G. (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. **5**, 1403–1407.

Robson, F., Khan, K. S., Le, T. K., Paris, C., Demirbag, S., Barfuss, P., Rocchi, P. and Ng, W. L. (2020) Coronavirus RNA proofreading: molecular basis and therapeutic targeting. Mol. Cell **79**, 710–727.

Saito, A., Irie, T., Suzuki, R., Maemura, T., Nasser, H., Uriu, K., Kosugi, Y., Shirakawa, K., Sadamasu, K., Kimura, I., et al. (2022) Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. Nature **602**, 300–306.

Sampath, S., Khedr, A., Qamar, S., Tekin, A., Singh, R., Green, R., and Kashyap, R. (2021) Pandemics Throughout the History. Cureus **13**, e18136

Sekizuka, T., Itokawa, K., Hashino, M., Kawano-Sugaya, T., Tanaka, R., Yatsu, K., Ohnishi, A., Goto, K., Tsukagoshi, H., Ehara, H., et al. (2020a) A Genome Epidemiological Study of SARS-CoV-2 Introduction into Japan. mSphere **5**, e00786 –20.

Sekizuka, T., Itokawa, K., Kageyama, T., Saito, S., Takayama, I., Asanuma, H., Nao, N., Tanaka, R., Hashino, M., Takahashi, T., et al. (2020b) Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. Proc. Natl. Acad. Sci. USA. **117**, 20198–20201.

Sekizuka, T., Itokawa, K., Saito, M., Shimatani, M., Matsuyama, S., Hasegawa, H., Saito, T., and Kuroda, M. (2022) Genome Recombination between the Delta and Alpha Variants of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Jpn. J. Infect. Dis. **75**, 415–418.

Shu, Y., and McCauley, J. (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. **22**, 30494.

Smallwood, N., Harrex, W., Rees, M., Willis, K., and Bennett, C. M. (2022) COVID-19 infection and the broader impacts of the pandemic on healthcare workers. Respirology **27**, 411–426.

Suzuki, R., Yamasoba, D., Kimura, I., Wang, L., Kishimoto, M., Ito, J., Morioka, Y., Nao, N., Nasser, H., Uriu, K., et al. (2022) Attenuated fusogenicity and pathogenicity of SARS-CoV-2 Omicron variant. Nature **603**, 700–705.

Syed, A. M., Taha, T. Y., Tabata, T., Chen, I. P., Ciling, A., Khalid, M. M., Sreekumar, B., Chen, P. Y., Hayashi, J. M., Soczek, K. M., et al. (2021) Rapid assessment of SARS-CoV-2-evolved variants using virus-like particles. Science **374**, 1626–1632.

Takada, K., Ueda, M.T., Shichinohe, S., Kida, Y., Ono, C., Matsuura, Y., Watanabe, T., Nakagawa, S. (2022) Genomic diversity of SARS-CoV-2 can be accelerated by mutations in the nsp14 gene. iScience **26**, 106210.

Tanaka, H., Hirayama, A., Nagai, H., Shirai, C., Takahashi, Y., Shinomiya, H., Taniguchi, C., and Ogata, T. (2021) Increased Transmissibility of the SARS-CoV-2 Alpha Variant in a Japanese Population. Int. J. Environ. Res. Public Health **18**, 7752.

Tokyo Metropolitan Institute of Public Health (TMIPH). (2022) Whole genome analysis of new coronavirus detected in Tokyo (Omicron BA strain) (in Japanese). Available online at: https://www.tmiph.metro.tokyo.lg.jp/lb_virus/sars2ngstree/ (accessed July 13, 2022).

Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., et al. (2021) Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature **593**, 266–269.

World Health Organization (WHO). (2022b) WHO Coronavirus (COVID-19) Dashboard. Available online at: https://covid19.who.int/ (accessed July 13, 2022).

World Health Organization (WHO). (2022a) Tracking SARS-CoV-2 variants. Available online at: https://www.who.int/activities/tracking-SARS-CoV-2-variants (accessed July 13, 2022).

Yamagishi, T., Kamiy, a. H., Kakimoto, K., Suzuki, M., and Wakita, T. (2020) Descriptive study of COVID-19 outbreak among passengers and crew on Diamond Princess cruise ship, Yokohama Port, Japan, 20 January to 9 February 2020. Euro Surveill. **25**, 2000272.

Yu, G. (2020) Using ggtree to Visualize Data on Tree-Like Structures. Curr Protoc Bioinformatics **69**, e96.