Title:

8266 SARS-CoV-2 Genomic Assemblies from Asymptomatic Carriers in Japan

Authors and Affiliations:

Hajime Ohyanagi[1†], Junko S Takeuchi[1†], Yuichi Kawanishi[1], Saya Ohi[1], Teiichiro Shiino[1], Moto Kimura[1], Yusuke Takahashi[2], Shigeru Yoshida[2], Minoru Kato[3], Yukumasa Kazuyama[3], Masato Ikeda[3], and Wataru Sugiura[1*]

[1]Center for Clinical Sciences, Japan Institute for Health Security, Tokyo, Japan

[2]Department of Clinical Laboratory Science, Health Sciences University of Hokkaido, Hokkaido, Japan

[3]SB Coronavirus Inspection Center Corp., Tokyo, Japan

[†]These authors equally contributed

*Corresponding author: sugiura.w@jihs.go.jp

Abstract:

In the context of public health, asymptomatic carriers of respiratory infectious diseases are considered a hidden yet critical factor in the transmission of infection and represent a key target for efforts to mitigate disease spread. To contribute a unique genomic resource to the SARS-CoV-2 research community, we collected SARS-CoV-2–positive samples from asymptomatic individuals at the SB Coronavirus Inspection Center Corp. during the COVID-19 pandemic in Japan and conducted a comprehensive analysis of their viral genomes. Using Illumina COVIDSeq technology, we successfully generated 8,266 SARS-CoV-2 genome assemblies, all of which have been made publicly available to facilitate further research. In this report, we summarize our efforts to collect SARS-

CoV-2–positive samples from asymptomatic individuals and highlight the key features and accessibility of this genomic dataset.

Background & Summary:

We operate the SB Coronavirus Inspection Center Corp. (SBCVIC) [1] , which has a key distinguishing feature compared to other medical facilities that conduct testing for general health concerns: our center collects samples from asymptomatic individuals**.** Consequently, the positive samples obtained at our facility are considered to be from asymptomatic carriers of SARS-CoV-2. Epidemiologically, asymptomatic carriers are known to have similar transmission capabilities to symptomatic individuals[2-5]. Therefore, their existence represents a hidden yet critical factor in controlling the spread of the pandemic[6,7]. From a genomic perspective, samples from asymptomatic carriers are also a valuable resource for investigating genetic factors in the viral genome that may be associated with the presence or absence of clinical symptoms in COVID-19 patients[8].

During the pandemic period in Japan[9,10], spanning from July 2020 to January 2023, the country experienced eight epidemic waves of SARS-CoV-2 infection. These included waves dominated by the Alpha (the 4th wave), Delta (the 5th wave), and Omicron sublineages BA.1, BA.2, and BA.5 (from the 6th wave onward) (**Figure 1A, B**). Over the course of the study (July 27, 2020 – January 16, 2023), we tested a total of 4,573,575 samples, among which 18,475 tested positive for SARS-CoV-2 using reverse transcription real-time quantitative polymerase chain reaction (RT-qPCR) (Ct values of $\leq$ 40) (**Figure 2** and **Table 1**). The overall average positive rate among asymptomatic individuals was 0.40%. The positive rates varied by wave: 0.05% during the 3rd wave,

0.13% in the 4th wave (Alpha), 0.23% in the 5th wave (Delta), 4.5% in the 6th wave, 8.2% in the 7th wave (BA.1/BA.2), and 5.2% in the 8th wave (BA.5) (**Figure 2**). The numbers of newly confirmed cases in Japan correlates well with the infection rate among asymptomatic individuals (**Figure 1A**).

Starting October 9, 2020 (approximately six weeks after the initiation of the study), we began collecting background information on positive cases from 45 prefectures across Japan, including Shiga, Tokyo, Nagasaki, Osaka, Hokkaido, Fukuoka, Chiba, Gifu, Kanagawa, Saitama, Hyogo, Aichi, Kyoto, Miyagi, Gumma, Hiroshima, Nagano, Kumamoto, Ibaraki, Shizuoka, Tochigi, Fukushima, Yamanashi, Nara, Miyazaki, Okinawa, Mie, Ehime, Iwate, Okayama, Kagawa, Yamagata, Saga, Aomori, Akita, Niigata, Toyama, Tottori, Ishikawa, Yamaguchi, Oita, Kagoshima, Fukui, Shimane, and Kochi (**Figure 3**). Among these cases, 51.9% were male, with a median age of 36 years (interquartile range [IQR]: 27–44) (**Table 1**).

	Number (%)
Tested individuals	4,573,575
Positive cases	18,475
Gender	
Male	6,408 (51.9%)
Female	5,944 (48.1%)
Unknown	6,123
Age (Years)	36 [27 - 44]

Table 1. Statistics of the study participants. Data are presented as median [interquartile range] for continuous measures and n (%) for categorical measures.

In total, we obtained 18,475 SARS-CoV-2-positive samples from asymptomatic individuals. We performed whole-genome sequencing on residual saliva samples and assembled viral genome sequences that met quality thresholds for downstream analysis. Ultimately, we successfully generated 8,266 genome assemblies from these cases,

which are archived in the SBCVIC genomic dataset and categorized into two groups (**Table 2** and **Table 3**).

For technical validation, we evaluated the distribution of consensus genome sequence lengths and examined the lineage composition of the SARS-CoV-2 genomes. We also assessed the presence of known mutations to confirm the reliability of our sequencing pipeline for allele typing. In particular, we focused on two adjacent, co-occurring mutations in the nucleocapsid (N) protein R203K/G204R[11], which are associated with increased viral infectivity, fitness, and virulence. For further details, see the **Technical Validation** section.

Methods:

SARS-CoV-2 Testing (Screening of Asymptomatic Cases)

Following previously described protocols[1,12], we conducted SARS-CoV-2 testing and genome sequencing. The details are summarized below. The SBCVIC provided routine workplace-based SARS-CoV-2 screenings upon company request, as well as voluntary screenings requested by local governments. As a result, most positive cases were identified in asymptomatic individuals [1,12]. All patients gave informed consent via the opt-out method. Between July 27, 2020, and January 16, 2023, we tested approximately 4.6 million asymptomatic individuals in Japan. However, individuals who declined participation in the study were excluded from this count. The study protocol was approved by the Ethics Committee of the National Center for Global Health and Medicine (NCGM), Japan (approval number: NCGM-G-003678-00). Each participant self-collected approximately 2 mL of saliva using the ZEESAN Saliva RNA Sample Collection Kit (MD-ZSV-001; Zeesan Biotech, Fujian, China), which contained

1 mL of a guanidine-based viral inactivation buffer. The sample was mailed to the inspection center as a UN 3373 Biological Substance Category B. On the day of arrival, SARS-CoV-2 RT-qPCR testing was performed using the SARS-CoV-2 Direct Detection RT-qPCR Kit (Takara Bio, Shiga, Japan). Kits RC30JW and RD003 were used until April 4 and from April 5, 2021, respectively, according to the manufacturer's protocol. Because the guanidine-based viral inactivation buffer can inhibit RT-PCR reactions[13], we modified the pretreatment process. Specifically, 16 µL of the saliva/buffer mixture was combined with 16 µL of water and 4 µL of pretreatment reagent (Solution A) to dilute the inhibitor. The resulting 36 µL mixture was incubated at room temperature for 5 minutes, followed by 5 minutes at 95°C. A 2.5 µL aliquot of this treated mixture was used as the RT-qPCR template. Test results (positive or negative) were determined based on the protocol, with Ct values ≤ 40 considered positive.

Viral Genome Sequencing

Samples from SARS-CoV-2-positive individuals were subjected to genome sequencing. The sequencing execution rate was nearly 100% for samples with Ct values ≤ 30 across all periods. However, the rate decreased for higher Ct values (30–40), particularly from 2022 onward, due to operational constraints (see **Table 2** and **Table 3**). For RNA extraction, 50 µL of viral nucleic acid was extracted from 200 µL of saliva samples using MagMAX Viral/Pathogen Nucleic Acid Isolation Kit (Themo Fisher Scientific, MA, USA) with MVP_Saliva_200_Flex_V1 protocol for KingFisher Flex System (Themo Fisher Scientific). The eluted RNA (20 µL) was treated with RQ1 RNase-Free DNase (Promega, WI, USA). Subsequent cDNA synthesis, target amplification using the ARTIC primer set [14], and library preparation were performed according to the Illumina COVIDseq Test Reference Guide (Illumina Inc., CA, USA).

Sequencing was conducted on an Illumina NextSeq 2000 system. Sequencing data were analyzed using the Illumina DRAGEN COVID Lineage Pipeline. Consensus sequences in FASTA format were generated, collected and archived (see **Data Records**). The ARTIC primer sets and DRAGEN versions used for each sample are documented in the archive. Sequences ≥ 29,000 nucleotides in length (Category 1 in **Table 2** and **Table 3**) were analyzed for lineage assignment using Phylogenetic Assignment of Named Global Outbreak Lineages (PANGOLIN) version 4.3.1 (pangolin-data 1.21) [15].

Control Dataset

To analyze the distribution of SARS-CoV-2 lineages across Japan during the study period (July 27, 2020 – January 16, 2023), we utilized sequences registered in the GISAID EpiCov database (https://gisaid.org) as of February 6, 2023. The filtering criteria were:

- Host: Human
- Location: Asia / Japan
- Collection date: July 27, 2020 – January 16, 2023
- Sequence length: ≥ 29,000 nt
- Passage history: Original

The resulting 539,504 sequences were analyzed using PANGOLIN version 4.3.1 (pangolin-data 1.21) for lineage classification (see **Figure 1B**).

Bioinformatics

The reference genome used was the Wuhan WIV04 strain (hCoV-19/Wuhan/WIV04/2019, EPI_ISL_402124), downloaded from GISAID. A multiple

sequence alignment, including the WIV04 reference, was generated in FASTA format using MAFFT [16]. The WIV04 sequence was then removed from the alignment, which was analyzed using AliView [17] and custom Python scripts to calculate allele frequencies at the R203K/G204R sites in the N protein.

Data Records:

Of the 18,475 SARS-CoV-2-positive samples, 14,201 were subjected to viral genome sequencing. By the end of December 2021, sequencing was conducted for 100% of samples with Ct ≤30 and 93.92% of those with Ct value between 30–40. From 2022 onward, the coverage remained high Ct ≤30 samples (for 99.83%) but decreased to 62.71% for Ct 30–40 samples (**Table 2**).

Year/Group		Positive Samples		Sequenced Samples		Genomes Determined (Category 1 and 2)		Genomes Determined (Category 1)	
- 2021	Ct ≤30	513	**2,025**	513	**1,933**	346	**683**	263	**444**
	Ct 30-40	1,512		1,420		337		181	
2022 -	Ct ≤30	5,258	**16,450**	5,249	**12,268**	4,389	**7,583**	2,473	**2,473**
	Ct 30-40	11,192		7,019		3,194		0	
		total	**18,475**	total	**14,201**	total	**8,266**	total	**2,917**
- 2021	Ct ≤30	100 %	**100** %	100.00 %	**95.5** %	67.45 %	**33.7** %	51.27 %	**21.9** %
	Ct 30-40	100 %		93.92 %		22.29 %		11.97 %	
2022 -	Ct ≤30	100 %	**100** %	99.83 %	**74.6** %	83.47 %	**46.1** %	47.03 %	**15.0** %
	Ct 30-40	100 %		62.71 %		28.54 %		0 %	

Table 2. Summary of viral genome sequencing. The upper section shows actual sample counts; the lower section shows corresponding percentages relative to the total number of positive samples. For the definition of Categories 1 and 2, see **Table 3**.

Raw reads and consensus genome sequences for 8,266 SARS-CoV-2 samples were submitted and are available for download from the International Nucleotide Sequence Database Collaboration (INSDC) [18]. The corresponding BioSample and other accession numbers are summarized in **Table 3**. According to operational criteria (based on genome length and Ct values), sequences were categorized into two groups (see **Table 3)**. Assembled genome sequences were also submitted to GISAID [19-21], under the accession IDs listed.

Category and Criteria	Number of Sequences	BioSample* Accession	Sequence Accession	SRA Run Accession	SRA Experiment Accession	GISAID ID
Category 1: Genome length ≥ 29,000 nucleotides and Ct value ≤ 40 (for - 2021) ≤30 (for 2022 -)	2917	SAMD00738313 - SAMD00741057, SAMD00741082, SAMD00786981 - SAMD00787151	BS007837 - BS010581, BS010689, BS016048 - BS016218	DRR600457 - DRR603201, DRR603222, DRR608552 - DRR608722	DRX581004 - DRX583748, DRX583769, DRX589099 - DRX589269	EPI_ISL_19576276 - EPI_ISL_19582591, EPI_ISL_19582637, EPI_ISL_19575435 - EPI_ISL_19575605
Category 2: Complete Genome not meeting Category 1 criteria and with ≤ 20% unidentified nucleotides	5349	SAMD00741058 - SAMD00741081, SAMD00741083 - SAMD00746626**	BS010669 - BS010688, BS010690 - BS016018	DRR603202 - DRR603221, DRR603223 - DRR608551	DRX583749 - DRX583768, DRX583770 - DRX589098	EPI_ISL_19582592 - EPI_ISL_19582606, EPI_ISL_19582632 - EPI_ISL_19582636, EPI_ISL_19582638 - EPI_ISL_19585391, EPI_ISL_19579847 - EPI_ISL_19580766, EPI_ISL_19578927 - EPI_ISL_19579846, EPI_ISL_19574700 - EPI_ISL_19575434

Table 3. Submission of genomic sequences to public repositories (INSDC/DDBJ and GISAID)

* BioSample accessions provide metadata describing the biological source materials used to generate sequencing data.

** 219 BioSample data did not meet Category 2 criteria; sequence data for these were not submitted.

Technical Validation:

Statistics of Consensus Genome Sequences

We first examined the length distribution of the 8,266 complete consensus genome sequences across both Category 1 and Category 2 (**Figure 4**). A sufficient number (n = 2,917) of nearly full-length genome sequences (≥ 29,000 nt; Category 1) were included and exclusively used for SARS-CoV-2 lineage assignment via PANGOLIN, as well as for subsequent analyses described below.

Epidemiological Distribution of SARS-CoV-2 Variants

Since the SBCVIC dataset represents a SARS-CoV-2 subpopulation within the broader Japanese population, we anticipated a comparable lineage distribution between our dataset and that of GISAID. Lineage analysis using PANGOLIN (**Figure 1B** and **1C**) revealed a similar distribution of SARS-CoV-2 variants between SBCVIC and GISAID datasets, supporting the validity of the overall SBCVIC sampling, sequencing, and analysis workflow.

Detection of R203K/G204R Mutations in the Nucleocapsid Protein

The 2,917 Category 1 genome assemblies were also used for further technical validation. As a control, we randomly selected 2,917 SARS-CoV-2 genome sequences from symptomatic individuals of Japanese origin deposited in GISAID during the same period (July 27, 2020 to January 16, 2023) (**Table 4**). To demonstrate the genotyping capability of our dataset, we focused on two well-known adjacent co-occurring mutations R203K/G204R in the N protein of SARS-CoV-2, which are known to affect viral pathogenicity[11].

Genome sequences were aligned to the SARS-CoV-2 reference genome, and the allele frequencies at the R203K/G204R sites were calculated for both the GISAID

(symptomatic) and SBCVIC (asymptomatic) groups (see **Method**). Our SBCVIC dataset exhibited a lower rate of missing allele calls (5.9%) compared to the GISAID dataset (17.8%), indicating high-quality genotyping across the majority of samples. Interestingly, the frequency of the mutant allele in the SBCVIC (asymptomatic) group was comparable to that in the GISAID (symptomatic) group (**Table 4**) (P = 0.799676, G-test), despite our initial expectation that mutation rates would differ between symptomatic and asymptomatic individuals. While we do not propose a specific hypothesis here, this allele-level analysis highlights a potential application of the SBCVIC dataset.

	Wild allele (R203/G204**)**	**Mutant allele (**203K/204R**)**	**Allele Not Determined**
GISAID (Symptomatic, n=2917)	7	2391	519 (17.8%)
SBCVIC (Asymptomatic, n=2917)	7	2739	171 (5.9%)

Table 4. Number of wild-type and mutant alleles at the nucleocapsid protein mutation sites (R203K/G204R) in symptomatic and asymptomatic populations.

Usage Note:

The dataset of SARS-CoV-2 samples collected and analyzed from asymptomatic individuals in Japan is highly distinctive. We demonstrated the validity of this SBCVIC dataset for analyzing the epidemiological distribution of SARS-CoV-2 variants and for conducting allele-level genotyping. We believe that this dataset has significant potential to provide insights into the modulation of SARS-CoV-2 virulence and to enhance our understanding of viral fitness.

Code Availability

All custom scripts used in this study are available on GitHub at https://github.com/oyasai55/calc_allele_frequencies/.

Acknowledgements

We thank Mr. Takahiro Sakurai and Mr. Hiroshi Ikawa for their support with bioinformatics analyses. We also acknowledge Mr. Tsuyoshi Saito, Ms. Natsumi Miyazaki, Mr. Atsushi Sugi, and Mr. Masanori Tamanaha for their assistance in managing sample metadata.

Ethics Approval:

All participants provided written informed consent for the collection of saliva samples, viral genome sequencing, and the use of associated demographic information. Participation in the study was voluntary, and individuals had the option to opt out via the study website. The study protocol was approved by the Institutional Review Board of the National Center for Global Health and Medicine (approval number: NCGM-G-003678-00). All procedures were conducted in accordance with the principles of the Declaration of Helsinki, as revised in 2013.

Author contributions

H.O. and J.S.T. designed and performed the analyses and wrote the manuscript. J.S.T. and Yuichi K. managed the genome sequence data from SBCVIC. S.O. conducted bioinformatics and statistical analyses. J.S.T. and S.O. conducted figure/table visualization. T.S. and W.S. supervised the entire analysis process. Moto K. directed the research contract with the SBCVIC and intellectual properly. S.Y., Minoru K. and

Yukumasa K. was responsible for RT-qPCR testing with the viral genome sequencing at SBCVIC. M.I. and W.S. conceived an inspection system by SBCVIC and acquired research funding. All authors critically reviewed and approved the final version of the manuscript.

Competing interests:

This research was supported by the SB Coronavirus Inspection Center Corp. Moto Kimura and Wataru Sugiura have received research grants from SB Coronavirus Inspection Center Corp. Masato Ikeda, Yukumasa Kazuyama, and Minoru Kato are employees of the SB Coronavirus Inspection Center Corp.

Figures:

In a separate file (pptx).

Figure legends:

Figure 1. Monthly epidemiological distribution of SARS-CoV-2 variants in Japan (July 27, 2020 - January 16, 2023). (A) Monthly number of newly confirmed COVID-19 cases in Japan, based on open data from the Ministry of Health, Labour and Welfare (https://covid19.mhlw.go.jp/en/, accessed May 1, 2023)[22]. (B) All domestic sequences registered in the Global Initiative on Sharing All Influenza Data (GISAID) EpiCoV database as of February 6, 2023 (n = 539,504). (C) Positive cases detected in asymptomatic individuals in this study (n = 2,917). SARS-CoV-2 lineage analysis was conducted using the Phylogenetic Assignment of Named Global Outbreak Lineages (PANGOLIN) version 4.3.1. Data visualization was performed using R version 4.3.1.

Figure 2. Number of tested (gray) and SARS-CoV-2-positive (blue) cases in this study. Between July 27, 2020, and January 16, 2023, a total of 4,573,575 saliva samples were tested, of which 18,475 were positive by RT-qPCR testing. Individuals who opted out of the study were excluded from this count.

Figure 3. Geographic distribution of sample collection. This figure was generated using the R package *NipponMap* (https://cran.r-project.org/web/packages/NipponMap)[23,24]. Color intensity indicates the number of samples collected, with darker shades representing higher counts. The top five prefectures with the highest number of samples were Shiga (n = 6,082), Tokyo (n = 3,118), Nagasaki (n = 521), Osaka (n = 429), and Hokkaido (n = 401).

Figure 4. Distribution of SARS-CoV-2 genome consensus sequence lengths obtained in this study (n = 8,266). A histogram shows the distribution of sequence lengths (X-axis) and their frequencies (Y-axis). Category 1 sequences (n = 2,917) are shown in dark purple, and Category 2 sequences (n = 5,349) are shown in green. The dashed line represents the threshold for near-complete genome sequences (29,000 nt).

References:

1 Shiino, T. *et al.* Molecular epidemiology of SARS-CoV-2 genome sentinel surveillance in commercial COVID-19 testing sites targeting asymptomatic individuals during Japan's seventh epidemic wave. *Sci Rep* **14**, 20950 (2024). https://doi.org/10.1038/s41598-024-71953-8

2 Buitrago-Garcia, D. *et al.* Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic

review and meta-analysis. *PLoS Med* **17**, e1003346 (2020). https://doi.org/10.1371/journal.pmed.1003346

3 Buitrago-Garcia, D. *et al.* Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: Update of a living systematic review and meta-analysis. *PLoS Med* **19**, e1003987 (2022). https://doi.org/10.1371/journal.pmed.1003987

4 Casey-Bryars, M. *et al.* Presymptomatic transmission of SARS-CoV-2 infection: a secondary analysis using published data. *BMJ Open* **11**, e041240 (2021). https://doi.org/10.1136/bmjopen-2020-041240

5 Yang, Q. *et al.* Just 2% of SARS-CoV-2-positive individuals carry 90% of the virus circulating in communities. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org/10.1073/pnas.2104547118

6 Davis, J. T. *et al.* Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. *Nature* **600**, 127-132 (2021). https://doi.org/10.1038/s41586-021-04130-w

7 Nabeshima, T. *et al.* COVID-19 cryptic transmission and genetic information blackouts: Need for effective surveillance policy to better understand disease burden. *Lancet Reg Health West Pac* **7**, 100104 (2021). https://doi.org/10.1016/j.lanwpc.2021.100104

8 Ferdinand, A. S. *et al.* An implementation science approach to evaluating pathogen whole genome sequencing in public health. *Genome Med* **13**, 121 (2021). https://doi.org/10.1186/s13073-021-00934-7

9 Kitahara, K., Nishikawa, Y., Yokoyama, H., Kikuchi, Y. & Sakoi, M. An overview of the reclassification of COVID-19 of the Infectious Diseases Control Law in Japan. *Glob Health Med* **5**, 70-74 (2023). https://doi.org/10.35772/ghm.2023.01023

10 Takeuchi, J. S. *et al.* Large-scale screening of SARS-CoV-2 variants in Tokyo, Japan: A 3-year and 9-month longitudinal survey. *Glob Health Med* **7**, 151-160 (2025). https://doi.org/10.35772/ghm.2025.01004

11 Wu, H. *et al.* Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host Microbe* **29**, 1788-1801 e1786 (2021). https://doi.org/10.1016/j.chom.2021.11.005

12 Terada-Hirashima, J. *et al.* Investigation of the use of PCR testing prior to ship boarding to prevent the spread of SARS-CoV-2 from urban areas to less-populated remote islands. *Glob Health Med* **4**, 174-179 (2022). https://doi.org/10.35772/ghm.2022.01008

13 Katsuno, T. *et al. Diagnostic accuracy of direct reverse transcription-polymerase chain reaction using guanidine-based and guanidine-free inactivators for SARS-CoV-2 detection in saliva samples* (Avairable at SSRN (preprint), 2023).

14 *Artic Network SARS-CoV-2*, <https://artic.network/ncov-2019> (

15 O'Toole, A. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* **7**, veab064 (2021). https://doi.org/10.1093/ve/veab064

16 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013). https://doi.org/10.1093/molbev/mst010

17 Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276-3278 (2014). https://doi.org/10.1093/bioinformatics/btu531

18 Karsch-Mizrachi, I. *et al.* The international nucleotide sequence database collaboration (INSDC): enhancing global participation. *Nucleic Acids Res* (2024). https://doi.org/10.1093/nar/gkae1058

19 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46 (2017). https://doi.org/10.1002/gch2.1018

20 Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Wkly* **3**, 1049-1051 (2021). https://doi.org/10.46234/ccdcw2021.255

21 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22** (2017). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

22 The Ministry of Health, L. a. W., Japan. *Visualizing the data: information on COVID-19 infections.*, <https://covid19.mhlw.go.jp/en/> (2023).

23 NipponMap: Japanese Map Data and Functions. v. 0.2 (2017).

24 R: A language and environment for statistical computing. (2021).

Figure 1

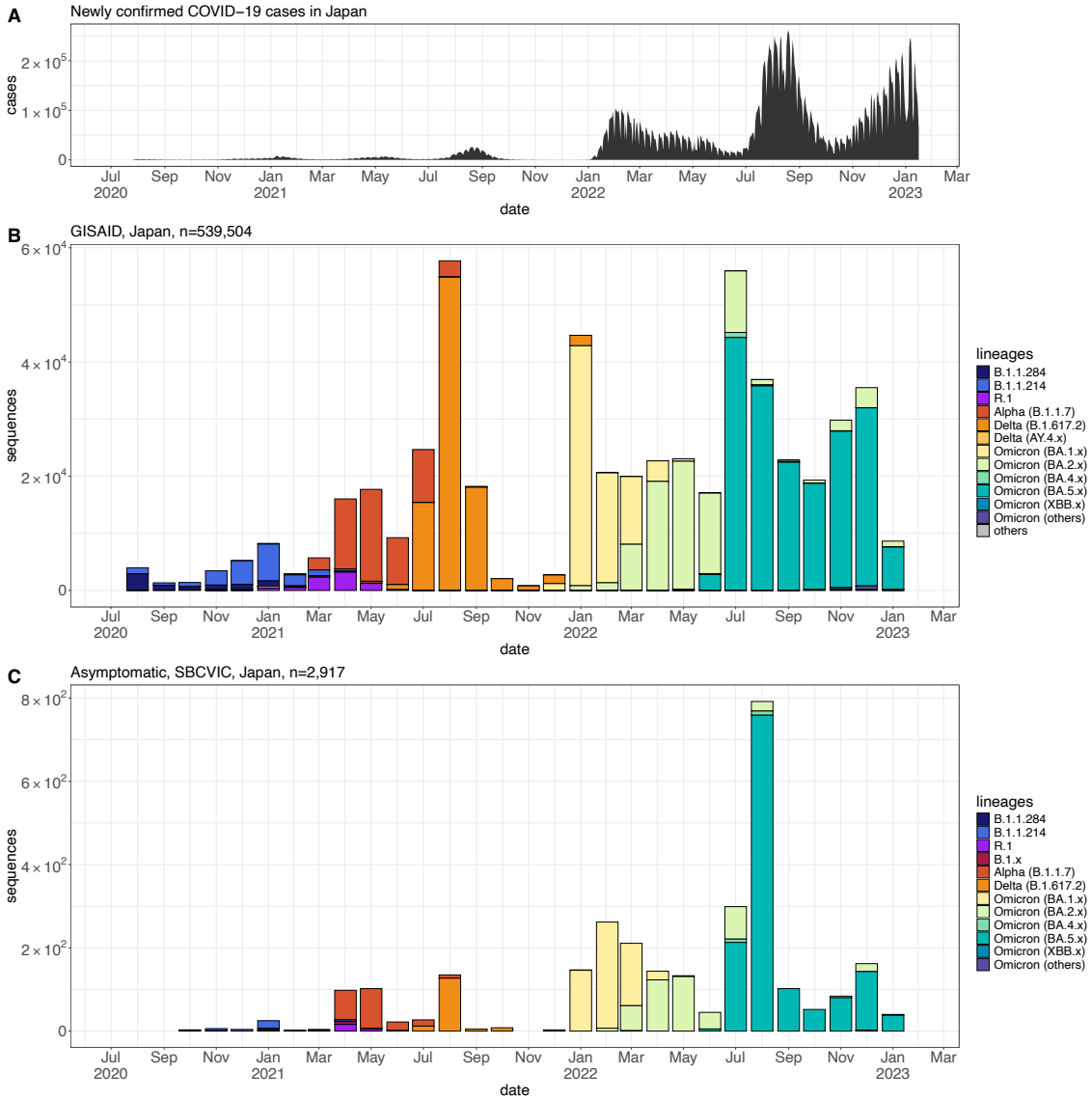

A
Newly confirmed COVID−19 cases in Japan
cases
2 × 10^5
1 × 10^5
0
Jul 2020
Sep
Nov
Jan 2021
Mar
May
Jul
Sep
Nov
Jan 2022
Mar
May
Jul
Sep
Nov
Jan 2023
Mar
date
B
GISAID, Japan, n=539,504
sequences
6 × 10^4
4 × 10^4
2 × 10^4
0
date
lineages
B.1.1.284
B.1.1.214
R.1
Alpha (B.1.1.7)
Delta (B.1.617.2)
Delta (AY.4.x)
Omicron (BA.1.x)
Omicron (BA.2.x)
Omicron (BA.4.x)
Omicron (BA.5.x)
Omicron (XBB.x)
Omicron (others)
others
C
Asymptomatic, SBCVIC, Japan, n=2,917
sequences
8 × 10^2
6 × 10^2
4 × 10^2
2 × 10^2
0
date
lineages
B.1.1.284
B.1.1.214
R.1
B.1.x
Alpha (B.1.1.7)
Delta (B.1.617.2)
Omicron (BA.1.x)
Omicron (BA.2.x)
Omicron (BA.4.x)
Omicron (BA.5.x)
Omicron (XBB.x)
Omicron (others)

Figure 2

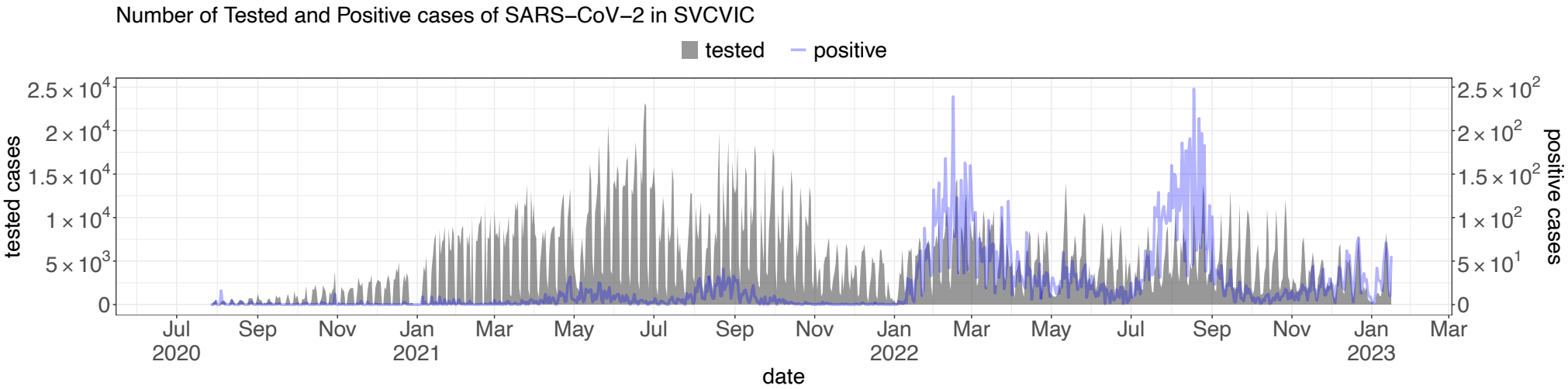

Number of Tested and Positive cases of SARS−CoV−2 in SVCVIC
tested
positive
tested cases
positive cases
2.5 × 10^4
2 × 10^4
1.5 × 10^4
1 × 10^4
5 × 10^3
0
2.5 × 10^2
2 × 10^2
1.5 × 10^2
1 × 10^2
5 × 10^1
0
Jul 2020
Sep
Nov
Jan 2021
Mar
May
Jul
Sep
Nov
Jan 2022
Mar
May
Jul
Sep
Nov
Jan 2023
Mar
date

Figure 3

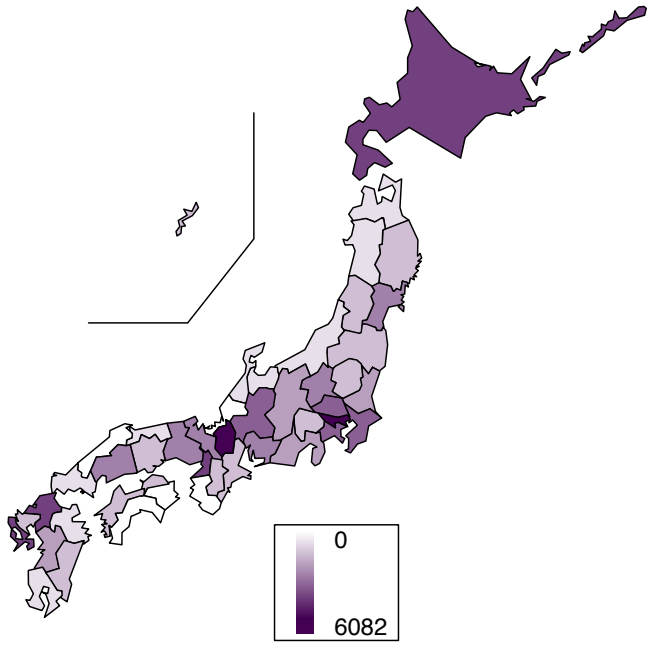
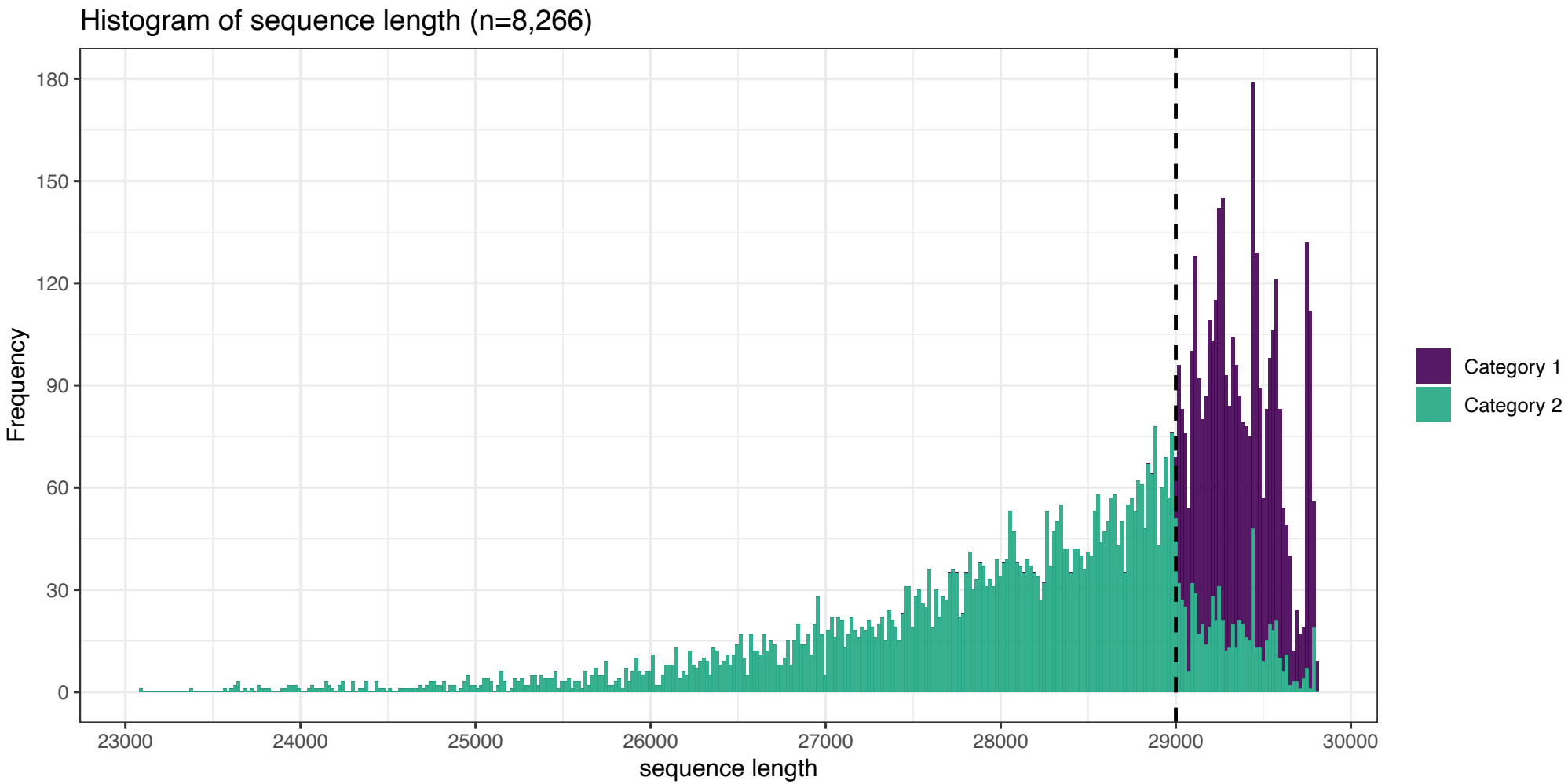


Figure 4


Histogram of sequence length (n=8,266)
Frequency
180
150
120
90
60
30
0
23000
24000
25000
26000
27000
28000
29000
30000
sequence length
Category 1
Category 2