Time-series trend of pandemic SARS-CoV-2 variants visualized using batch-learning self-organizing map for oligonucleotide compositions

Takashi Abe[1,*], Ryuki Furukawa[1], Yuki Iwasaki[2], Toshimichi Ikemura[2,*]

1. Smart Information Systems, Faculty of Engineering, Niigata University, Niigata-ken 950-2181, Japan
2. Department of Bioscience, Nagahama Institute of Bio-Science and Technology. Shiga-ken 526-0829, Japan

*** CORRESPONDENCES:**

Takashi Abe (takaabe@ie.niigata-u.ac.jp) and Toshimichi Ikemura (t_ikemura@nagahama-i-bio.ac.jp)

AUTHORS' CONTRIBUTIONS

TA and TI conceived and designed research; TA, RF, and TI performed research; TA, RF, and YI analyzed data; and all authors wrote the paper.

ABSTRACT

To confront the global threat of coronavirus disease 2019, a massive number of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome sequences have been decoded, with the results promptly released through the GISAID database. Based on variant types, eight clades have already been defined in GISAID, but the diversity can be far greater. Owing to the explosive increase in available sequences, it is important to develop new technologies that can easily grasp the whole picture of the big-sequence data and support efficient knowledge discovery. An ability to efficiently clarify the detailed time-series changes in genome-wide mutation patterns will enable us to promptly identify and characterize dangerous variants that rapidly increase their population frequency. Here, we collectively analyzed over 150,000 SARS-CoV-2 genomes to understand their overall features and time-dependent changes using a batch-learning self-organizing map (BLSOM) for oligonucleotide composition, which is an unsupervised machine learning method. BLSOM can separate clades defined by GISAID with high precision, and each clade is subdivided into clusters, which shows a differential increase/decrease pattern based on geographic region and time. This allowed us to identify prevalent strains in each region and to show the commonality and diversity of the prevalent strains. Comprehensive characterization of the oligonucleotide composition of SARS-CoV-2 and elucidation of time-series trends of the population frequency of variants can clarify the viral adaptation processes after invasion into the human population and the time-dependent trend of prevalent epidemic strains across various regions, such as continents.

KEYWORDS

COVID-19, SARS-CoV-2, Oligonucleotide composition, Batch-Learning Self-Organizing Map (BLSOM), Unsupervised explainable machine learning, Time-series trend

INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread rampantly worldwide since it was first reported in December 2019, and its momentum is still ongoing (WHO. 2020). To address the SARS-CoV-2 pandemic in detail, genome sequencing has been performed on a global scale and published by GISAID (Elbe et al. 2017), the SARS-CoV-2 genome database, having more than 780,000 viral sequences as of March 2021 (https://www.gisaid.org/). SARS-CoV-2 is an RNA virus with a fast evolutionary rate that has already been classified into eight clades by GISAID, and epidemics caused by new variant have been known to occur (Benvenuto et al. 2020; Gorbalenya et al. 2020; Sun et al. 2020; Hu et al. 2021; Kirby 2021; Wang et al. 2021). Because the number of registered genome sequences is increasing explosively, it has become difficult to cope with the current and future situation using only the conventional phylogenetic tree method based on multiple sequence alignment, which requires an enormous amount of computation time for a massive number of sequences. Therefore, it is imperative to develop a sequence alignment-free method that will enable us to easily grasp the whole picture of the big-sequence data and support efficient knowledge discovery from it.

By focusing on the frequency of short oligonucleotides (e.g., tetra- and penta-nucleotides) in a large number of genomic fragments (e.g., 10 kb) derived from a wide variety of species, we have developed an unsupervised explainable AI (batch-learning self-organizing map; BLSOM), which enables separation (self-organization) of the genomic sequences by species and phylogeny and explains the causes that contribute to this separation (Abe et al. 2003). In the analysis of genomic fragments of a wide range of microbial genomes, over 5 million sequences can be separated by phylogenetic groups with high accuracy (Abe et al. 2020).

In a prior analysis of all influenza A strains, viral genomes were separated (self-organized) by host animals based only on the similarity of the oligonucleotide composition, although no host information was provided during BLSOM learning (Iwasaki et al. 2011). On a single map, all viral sequences could be separated, and notably, BLSOM is an explainable AI that can explain diagnostic oligonucleotides, which contribute to host-dependent clustering. When studying the 2009 swine-derived flu pandemic (H1N1/2009), we could detect directional time-series changes in oligonucleotide composition because of possible adaptations to the new host, namely humans (Iwasaki et al. 2011), showing that near-future prediction was possible, albeit partially (Iwasaki et al. 2013).

We have previously revealed lineage-specific oligonucleotide compositions for a wide range of virus lineages and established a method to identify and classify viral-derived sequences in tick intestinal metagenomic sequences (Qiu et al. 2019). In the case of SARS-CoV-2, we analyzed time-series changes in mono- and oligo-nucleotide compositions and found their time-dependent directional changes that are thought to be adaptive for growth in humans, which allowed us to predict candidates of advantageous mutations for growth in human cells (Ikemura et al. 2020; Wada, Wada & Ikemura. 2020; Iwasaki Abe & Ikemura. 2021). Furthermore, we recently performed BLSOM analysis on di- to penta-nucleotide compositions in approximately 150,000 SARS-CoV-2 genomes. Because the accuracy of separation by clade increased as the oligonucleotide length increased, in this report, we present the BLSOM results for the pentanucleotide composition. BLSOM could serve as a powerful tool for elucidating comprehensive characterization of the oligonucleotide composition of SARS-CoV-2 and time-series trends of prevalent epidemic strains across various regions, such as continents.

METHODS

SARS-CoV-2 genome sequences

The full-length genome sequences of SARS-CoV-2 were downloaded from the GISAID database on November 4, 2020. The total number of sequences was 170,190. From these sequences, those with a length of more than 27 kb after removing the polyA-tail sequences were selected.

Oligonucleotide frequency and odds ratio

Pentanucleotide frequencies and odds ratios were used in the present study. The pentanucleotide odd ratios (observed/expected values) were calculated using the formula $P_{VWXYZ} = f_{VWXYZ}/f_V f_W f_X f_Y f_Z$, where f_V, f_W, f_X, f_Y and f_Z denote the frequencies of mononucleotides V, W, X, Y and Z, respectively, and f_{VWXYZ} denotes the frequency of pentanucleotide VWXYZ (Karlin et al. 1998).

BLSOM

Kohonen’s self-organizing map (SOM), an unsupervised neural network algorithm, is a powerful tool for clustering and visualizing high-dimensional complex data on a two-dimensional map (Kohonen, 1990; Kohonen et al., 1996). We modified the conventional SOM for genome informatics on the basis of batch learning, aiming to make the learning process and the resulting map independent of the order of data input (Kanaya et al. 2001; Abe et al. 2003). The newly developed SOM, BLSOM, is suitable for high-performance parallel computing and, therefore, for big data analysis. The initial weight vectors were defined using principal component analysis (PCA), based on the variance-covariance matrix, rather than by using random values. The weight vectors (w_{ij}) were arranged in a two-dimensional lattice denoted by i (= 0, 1,…, I-1) and j (= 0, 1,…, J-1) and were set and updated as described previously (Kanaya et al. 2001; Abe et al. 2003). A BLSOM program suitable for PC cluster systems is available on our website (http://bioinfo.ie.niigata-u.ac.jp/?BLSOM).

RESULTS and DISCUSSION

BLSOM for pentanucleotide composition and their odds ratio

It should be mentioned here that SARS-CoV-2 genomes have changed their mononucleotide composition during the course of the epidemic in humans, reducing C and increasing U, regardless of clade (Mercatelli et al. 2020; Wada, Wada & Ikemura. 2020; Iwasaki Abe & Ikemura 2021), a process which is thought to be caused by the APOBEC family enzymes (Mangeat et al. 2003; Simmonds 2020). Considering this clade-independent tendency, we performed BLSOM analysis of not only the pentanucleotide composition but also their odds ratio, which can reduce the effects caused by changes in the mononucleotide composition. Additionally, to check the robustness of sequence accuracy, we used datasets with different sequence accuracies: 167,905 sequences with less than 10% unknown nucleotides other than ATGCs in the genome sequence and 130,753 sequences with less than 1% unknown nucleotides; for each sequence dataset, the number of cases by region and clade is shown in Table 1.

First, we constructed BLSOM for sequences with less than 10% unknown nucleotides, using the pentanucleotide composition and their odds ratios (Figure 1A and B). BLSOM utilizes unsupervised machine learning, and the genome sequences are clustered (self-organized) on a two-dimensional plane, based only on the difference in the vector data in a 1024 ($=4^5$)-dimensional space. Lattice points that include sequences from more than one clade are indicated in black, those that contain no genomic sequences are indicated by blank, and those containing sequences from a single clade are indicated in the color representing the clade. The odds ratio (Figure 1B) gave more accurate separations (a smaller percentage of black grid points), possibly by excluding effects owing to the clade-independent time-series change in the mononucleotide composition (Iwasaki Abe & Ikemura. 2021), which affected all SARS-CoV-2 clades. Even for the sequences with low-sequence accuracy, clade-dependent separation occurs, allowing us to understand characteristics of the oligonucleotide composition that are specific to each clade; thus, oligonucleotide-BLSOM is thought to be a robust method. However, it is clear that BLSOMs for sequences with less than 1% unknown nucleotides (Figure 1C and D) gave more accurate separation than those listed in Figure 1A and B, and the highest resolution was obtained for the BLSOM for the odds ratio (Figure 1D).

Clades have been defined by the statistical distribution of phylogenetic distances in tree construction based on multiple sequence alignments (Han et al. 2019; Tang et al. 2020), whereas BLSOM is a sequence alignment-free analysis that is suitable for the analysis of massive data. Because sequences at different locations on BLSOM have different oligonucleotide compositions, clustering according to clades means that sequences belonging to different clades have different oligonucleotide combinations, that is, differential combinations of mutations.

3D display of the data for different continents

Using BLSOM (Figure 1D) for the pentanucleotide odds ratio, Figure 1E examines the classification according to four continents (Asia, Europe, North America, and Oceania) that have large numbers of sequences. Here, the lattice points containing sequences of different continents are displayed in black, and those containing only sequences of a single continent are displayed in the color specifying each continent. Although not as clear as clade-dependent separations, regional differences have been observed, which should reflect differential shares of prevalent variants among continents. However, it is apparently difficult to obtain sufficient information from the results shown in Figure 1E alone. BLSOM is equipped with various visualization tools for analysis results; therefore, we next show the number of sequences belonging to each lattice point with a 3D display.

Again, using the BLSOM shown in Figure 1D, Figure 2 shows the number of sequences belonging to each lattice point for each clade in each continent as a vertical bar, which is colored by continent, as shown in Figure 1E. Looking laterally at a particular clade, each clade consists of several subclusters, each consisting of several high peaks surrounded by many low peaks. Different subclusters observed in each clade are distinguished by numbering in each figure, but if they are located in the same zone on BLSOM, the same number is given even if they are of different continents. Looking vertically at a particular continent, sequences of different subclusters of different clades exist in different amounts, and some subclusters are only in a particular continent, that is, the prevalent variants for each continent can be visualized in an easy-to-understand manner. In Supplementary Figure S1, the data shown in Figure 2 are displayed in 2D, and referring to the quantitative results in Figure 2, we defined sequences attributed to each subcluster in each clade.

Time-series analysis

The fact that sequences belonging to one clade were clearly separated on BLSOM indicates the importance of subdivision of each clade, and the separation on BLSOM is thought to be a good indicator of this subdivision. To further examine the biological significance of the subclusters of each clade on BLSOM, we visualized the number of sequences collected in each month in each region as a vertical bar differentially colored according to clade (Figure 3). Looking laterally at a continent, the time-series quantitative changes among different clades or different subclusters of one clade are clear. Looking at the results for a particular collection month for different continents longitudinally, quantitative changes among different clades or different subclusters of one clade are again clear, depending on the continent.

Next, for each clade in each continent, we quantitatively analyzed the time-series changes in the proportion of its subclusters using a 100% stack bar graph (Figure 4). The percentage of sequences in different subclusters are distinguished by different colors, and when a total number of sequences for a certain month is more than 100, the data for that month is indicated by a thick horizontal bar.

We focused mainly on such months.

In the clade S/L/V detected in the early stage of the epidemic (December 2019– March 2020), three major subclusters of each clade were observed and distinguished by suffix numbers, and most sequences belonged to the two subclusters: S1/L1/V1 and S2/L2/V2. In Asia, many sequences belonging to S1/L1/V1 were detected in December 2019, but in Europe and other regions, S2/L2/V2 were more abundantly detected in March and April 2020 than S1/L1/V1, and the proportion became more pronounced in April than in March. In March and April in Europe, a remarkable number of sequences belonging to S3/L3/V3 were also detected, showing three different variants prevalent at the beginning of the epidemic in Europe. Far fewer than 100 sequences were detected after May; sequences belonging to S1/L1/V1 were mainly detected in Asia and those belonging to S2/L2/V2 were shown in other regions, presenting differential trends in prevalent variants among continents.

For clade G, which started the epidemic in Europe in February, we defined five subclusters. In February, roughly equal amounts of sequences belonging to G1 and G2 were detected in Europe and North America, but as the epidemic progressed, those belonging to G2 were mainly detected in Europe, whereas those belonging to both G1 and G2 were prevalent in North America. In Asia, only sequences belonging to G1 are detected; in Oceania, those belonging to G2 accounted for about 10% in the early stage, but afterward, those belonging to Oceania-specific G5 accounted for the majority.

For GH, we defined seven subclusters, including GH1 and GH2, which dominated in North America and Europe, respectively. In North America, in addition to GH1, several months contain approximately 20% of the sequences belonging to GH3, GH5, and GH6. In Asia, only GH1 has been detected. In Oceania, only GH4 and GH7, which were specific to this region, were detected; initially, GH4 was dominant, but after July, GH7 was primarily detected.

For GR, we defined five subclusters, including GR1 and GR2, which dominated in North America and Europe, respectively. Moreover, in Europe, GR1 was detected to the same extent as GR2 in

February, but as the epidemic progressed, GR2 began to predominate. In North America, the occupancy of GR1 and GR2 varied to some extent depending on the collection month. In Asia, GR1 was mainly detected, and in Oceania, only region-specific subclusters have been detected.

These temporospatial changes in subclusters show that the subcluster is the separation (self-organization) that reflects biological significance and is fundamental information for understanding the overall picture of the SARS-CoV-2 variants.

CONCLUSION and PERSPECTICES

Based on the phylogenetic tree construction by multiple sequence alignments, GISAID has defined seven clades of SARS-CoV-2, giving a total of eight if clade O corresponding to others is included. However, these classifications are clearly inadequate to understand the current status of SARS-CoV-2 because this RNA virus evolves at a high speed. Using only the oligonucleotide composition of many genomic sequences, the unsupervised machine learning, BLSOM, could separate viral sequences according to not only clades but also subclusters within each clade. The separation (self-organization) that AI can accomplish without any hypothesis or model is thought to be a classification from a new perspective. BLSOM is equipped with various tools that allow us to visualize the analysis results in an easily understandable way and to visualize differences in the number of subcluster sequences among continents (Figure 2) and their time-series changes (Figure 3), i.e., the distinct variations in the resulting subclusters depending on the region and the collection time.

Herein, we focused on pentanucleotide composition, but similar separations were obtained for other lengths of oligonucleotides (Ikemura et al. 2020). BLSOM is an explanatory AI that can clarify combinatorial patterns of oligonucleotides that contribute to the separation according to clades and their subclusters. BLSOM is a powerful method for elucidating comprehensive characterization of the oligonucleotide composition in a massive number of SARS-CoV-2 genome sequences. Next, it will be important to know the relationship between the strains isolated in clades and their subclusters and the causative mutations. When it comes to oligonucleotides as long as 15-mers, most are only present in one copy in the viral genome; therefore, changes in 15-mer sequences can be directly linked to mutations, and we have already started analysis from this perspective (Ikemura et al. 2020). The implementation of time-series oligonucleotide analysis of variants with rapidly expanding intra-population frequencies has enabled the identification of candidates for advantageous mutations for viral infection and growth in human cells (Wada, Wada & Ikemura. 2020).

Phylogenetic methods based on sequence alignment have been widely used in evolutionary studies (Hadfield et al. 2018; Kumar et al. 2018), and these methods are undoubtedly essential for studying the phylogenetic relationships between different viral species and variations in the same virus at the single-nucleotide level. In contrast, AI can analyze a massive number of SARS-CoV-2 sequences at once without difficulty, potentially reaching a level of one million in the near future. The AI method for oligonucleotide composition has become increasingly important as a complement to the phylogenetic tree construction method in preparing for future outbreaks of various infectious RNA viruses.

ACKNOWLEDGEMENTS

We gratefully acknowledge the authors submitting their sequences from GISAID's Database.

FUNDING INFORMATION

This research was supported by AMED Grant Number JP20he0622033h0001, and by JST, CREST Grant Number JPMJCR20H1, and by KAKENHI Grant Number 18K07151 and 20H03140.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this paper.

TABLE

Table 1. Number of SARS-CoV-2 genome sequences with less than 10% (A) and less than 1% (B) unknown nucleotides used in this study.

(A) Number of sequences with less than 10% unknown nucleotides

Clade＼Continent	Asia	Europe	North America	Oceania	Africa	South America	Unknown	Total
S	794	1,860	3,449	664	110	74	0	6,951
L	823	3,196	600	65	4	11	0	4,699
V	247	4,687	402	253	13	23	0	5,625
G	979	20,928	6,568	1,106	1,141	461	0	31,183
GH	2,058	10,325	23,916	964	232	176	0	37,671
GR	2,657	42,888	5,251	11,135	1,632	1,129	0	64,692
GV	3	12,229	3	14	0	0	0	12,249
O	2,220	1,127	553	531	60	25	0	4,516
Non-human host	35	247	19	0	1	4	13	319
#Total	9,816	97,487	40,761	14,732	3,193	1,903	13	167,905

(B) Number of sequences with less than 1% unknown nucleotides

Clade＼Continent	Asia	Europe	North America	Oceania	Africa	South America	Unknown	Total
S	731	1,047	3,056	466	71	58	0	5,429
L	760	1,964	549	49	2	10	0	3,334
V	228	3,036	366	207	10	17	0	3,864
G	877	15,200	5,071	858	634	300	0	22,940
GH	1,923	8,365	19,014	717	191	150	0	30,360
GR	2,425	32,518	4,549	9,166	1,180	871	0	50,709
GV	3	10,712	3	11	0	0	0	10,729
O	1,824	522	349	415	30	9	0	3,149
Non-human host	30	176	19	0	1	0	13	239
#Total	8,801	73,540	32,976	11,889	2,119	1,415	13	130,753

Unknown: genome sequences for which continent was not registered

FIGURE LEGENDS

Figure 1. BLSOM for pentanucleotide usage. (A) Pentanucleotide composition and (B) their odds ratio for sequences with less than 10% unknown nucleotides. (C) Pentanucleotide composition and (D) their odds ratio for sequences with less than 1% unknown nucleotides. Lattice points that include sequences from more than one clade are indicated in black, those that contain no genomic sequences are indicated by blank, and those containing sequences from a single clade are indicated in color as follows: S (🟫), L (🟪), V (🟩), G (🟪), GH (🟦), GR (🟪), GV (🟧), O (🟦), non-human host (🟨). (E) Distribution of sequences by continent on the BLSOM with the pentanucleotide odds ratio. Lattice points that include sequences from more than one continent are indicated in black, those that contain no genomic sequences are indicated by blank, and those containing sequences from a single continent are indicated in color as follows: Asia (🟧), Europe (🟥), North America (🟩), Oceania (🟦).

Figure 2. 3D display of viral classification by clade and continent. The Z-axis corresponds to the number of sequences attributed to each lattice point. Results for all continents are shown in the ALL panel for each clade. In clades G, GH, GR and GV, lattice points where less than 5 sequences exist are not shown. The vertical bars for individual continents are distinguished by the following colors: Asia (🟧), Europe (🟥), North America (🟩), Oceania (🟦). Different subclusters are given suffix numbers.

Figure 3. 3D display of temporospatial changes. The Z-axis corresponds to the number of sequences attributed to each lattice point. Results for all collection months are shown in the ALL panel for each continent. The vertical bars for individual clades are distinguished by the following colors: S (🟥), L (■), V (🟩), G (■), GH (🟦), GR (🟪), GV (🟧).

Figure 4. Analysis of 100% stack bar graph for time-series transition in each continent for each subcluster in clades S (A), L (B), V (C), G (D), GH (E) and GR (F). The colors of each subcluster are indicated at the bottom of each figure. The results for months with more than 100 sequences are shown as thick horizontal bars. The number of sequences used in this analysis is given in Supplementary Table S1.

SUPPLEMENTARY DATA

Supplementary Figure S1. 2D display of the classification by clade and continent shown in Figure 2. Each subcluster territory is circled by a dotted line. In clades G, GH, GR and GV, lattice points where less than 5 sequences exist are not shown. The sequences belonging to each territory defined here are used for the analysis in Figure 4.

Supplementary Table S1. Sequence number of subdivided clusters in clade for each month by continent.

REFERENCES

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T. 2003. Informatics for unveiling hidden genome signatures. Genome research 13: 693-702. DOI: http://dx.doi.org/10.1101/gr.634603

Abe T, Akazawa Y, Toyoda A, Niki H, Baba T. 2020. Batch-Learning Self-Organizing Map Identifies Horizontal Gene Transfer Candidates and Their Origins in Entire Genomes. Frontiers in microbiology 11: 1486. DOI: http://dx.doi.org/10.3389/fmicb.2020.01486

Benvenuto D, Giovanetti M, Salemi M, Prosperi M, De Flora C, Junior Alcantara LC, Angeletti S, Ciccozzi M. 2020. The global spread of 2019-nCoV: a molecular evolutionary analysis. Pathogens and global health 114: 64-67. DOI: http://dx.doi.org/10.1080/20477724.2020.1725339

Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global challenges (Hoboken, NJ) 1: 33-46. DOI: http://dx.doi.org/10.1002/gch2.1018

Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM, Neuman BW et al. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature microbiology 5: 536-544. DOI: http://dx.doi.org/10.1038/s41564-020-0695-z

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics (Oxford, England) 34: 4121-4123. DOI: 10.1093/bioinformatics/bty407

Han AX, Parker E, Scholer F, Maurer-Stroh S, Russell CA. 2019. Phylogenetic Clustering by Linear Integer Programming (PhyCLIP). Molecular Biology and Evolution 36: 1580-1595. DOI: http://dx.doi.org/10.1093/molbev/msz053

Hu B, Guo H, Zhou P, Shi ZL. 2021. Characteristics of SARS-CoV-2 and COVID-19. Nature reviews Microbiology 19: 141-154. DOI: http://dx.doi.org/10.1038/s41579-020-00459-7

Ikemura T, Wada K, Wada Y, Iwasaki Y, Abe T. 2020. Unsupervised explainable AI for simultaneous molecular evolutionary study of forty thousand SARS-CoV-2 genomes. bioRxiv. DOI:10.1101/2020.10.11.335406: 2020.2010.2011.335406.

Iwasaki Y, Abe T, Wada K, Itoh M, Ikemura T. 2011. Prediction of directional changes of influenza A virus genome sequences with emphasis on pandemic H1N1/09 as a model case. DNA research 18: 125-136. DOI: http://dx.doi.org/10.1093/dnares/dsr005

Iwasaki Y, Abe T, Wada Y, Wada K, Ikemura T. 2013. Novel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains. BMC infectious diseases 13: 386. DOI: http://dx.doi.org/10.1186/1471-2334-13-386

Iwasaki Y, Abe T, Ikemura T. 2021. Human cell-dependent, directional, time-dependent changes in the mono- and oligonucleotide compositions of SARS-CoV-2 genomes. bioRxiv.

DOI:10.1101/2021.01.05.425508: 2021.2001.2005.425508.

Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nishi T, Mori H, Ikemura T. 2001. Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome. Gene 276: 89-99. DOI: 10.1016/s0378-1119(01)00673-4

Karlin S, Campbell AM, Mrazek J. 1998. Comparative DNA analysis across diverse genomes. Annual review of genetics 32: 185-225.

Kirby T. 2021. New variant of SARS-CoV-2 in UK causes surge of COVID-19. The Lancet Respiratory medicine 9: e20-e21. DOI: http://dx.doi.org/10.1016/s2213-2600(21)00005-9

Kohonen T. 1990. The self-organizing map. Proceedings of the IEEE 78: 1464-1480. DOI: 10.1109/5.58325

Kohonen T, Oja E, Simula O, Visa A, Kangas J. 1996. Engineering applications of the self-organizing map. Proceedings of the IEEE 84: 1358-1384. DOI: 10.1109/5.537105

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35: 1547-1549. DOI: 10.1093/molbev/msy096

Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D. 2003. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature 424: 99-103. DOI:

http://dx.doi.org/10.1038/nature01709

Mercatelli D, Giorgi FM. 2020. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Frontiers in microbiology 11: 1800. DOI: http://dx.doi.org/10.3389/fmicb.2020.01800

Qiu Y, Abe T, Nakao R, Satoh K, Sugimoto C. 2019. Viral population analysis of the taiga tick, Ixodes persulcatus, by using Batch Learning Self-Organizing Maps and BLAST search. The Journal of veterinary medical science 81: 401-410. DOI: http://dx.doi.org/10.1292/jvms.18-0483

Simmonds P. 2020. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere 5. DOI: http://dx.doi.org/10.1128/mSphere.00408-20

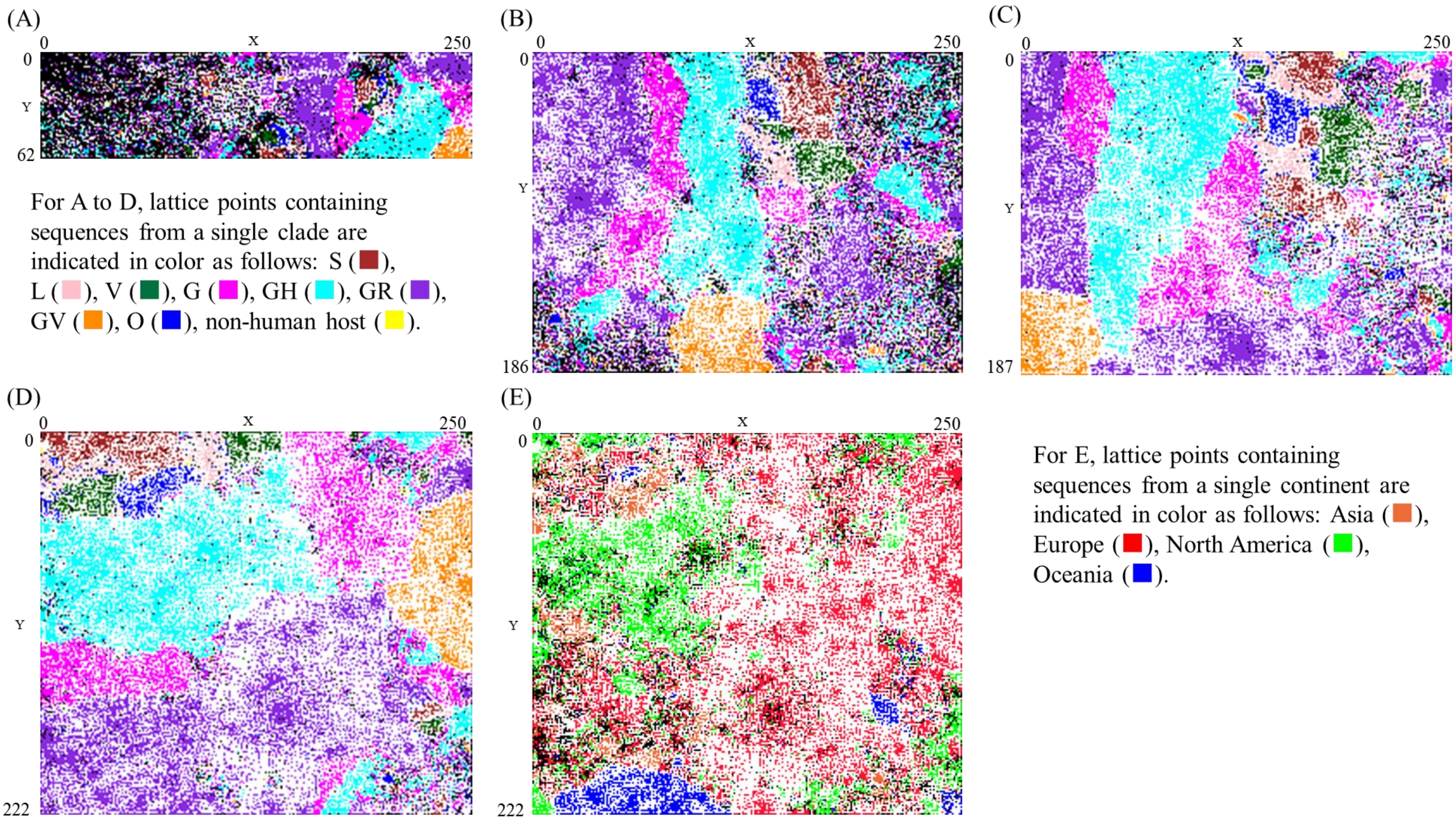
Sun J, He WT, Wang L, Lai A, Ji X, Zhai X, Li G, Suchard MA, Tian J, Zhou J et al. 2020. COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. Trends in molecular medicine 26: 483-495. DOI: http://dx.doi.org/10.1016/j.molmed.2020.02.008

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z et al. 2020. On the origin and continuing evolution of SARS-CoV-2. National Science Review 7: 1012-1023 %@ 2095-5138. DOI: http://dx.doi.org/10.1093/nsr/nwaa036

Wada K, Wada Y, Ikemura T. 2020. Time-series analyses of directional sequence changes in SARS-CoV-2 genomes and an efficient search method for candidates for advantageous mutations for growth in human cells. Gene: X 5: 100038. DOI: http://dx.doi.org/10.1016/j.gene.2020.100038

Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei GW. 2021. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Communications biology 4: 228. DOI: http://dx.doi.org/10.1038/s42003-021-01754-6

World Health Organization. 2020. Coronavirus Disease (COVID-2019). Situation Reports. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019



For A to D, lattice points containing sequences from a single clade are indicated in color as follows: S (🟫), L (■), V (🟩), G (■), GH (■), GR (🟪), GV (🟧), O (🟦), non-human host (🟨).

For E, lattice points containing sequences from a single continent are indicated in color as follows: Asia (🟧), Europe (🟥), North America (🟩), Oceania (🟦).

Figure 1

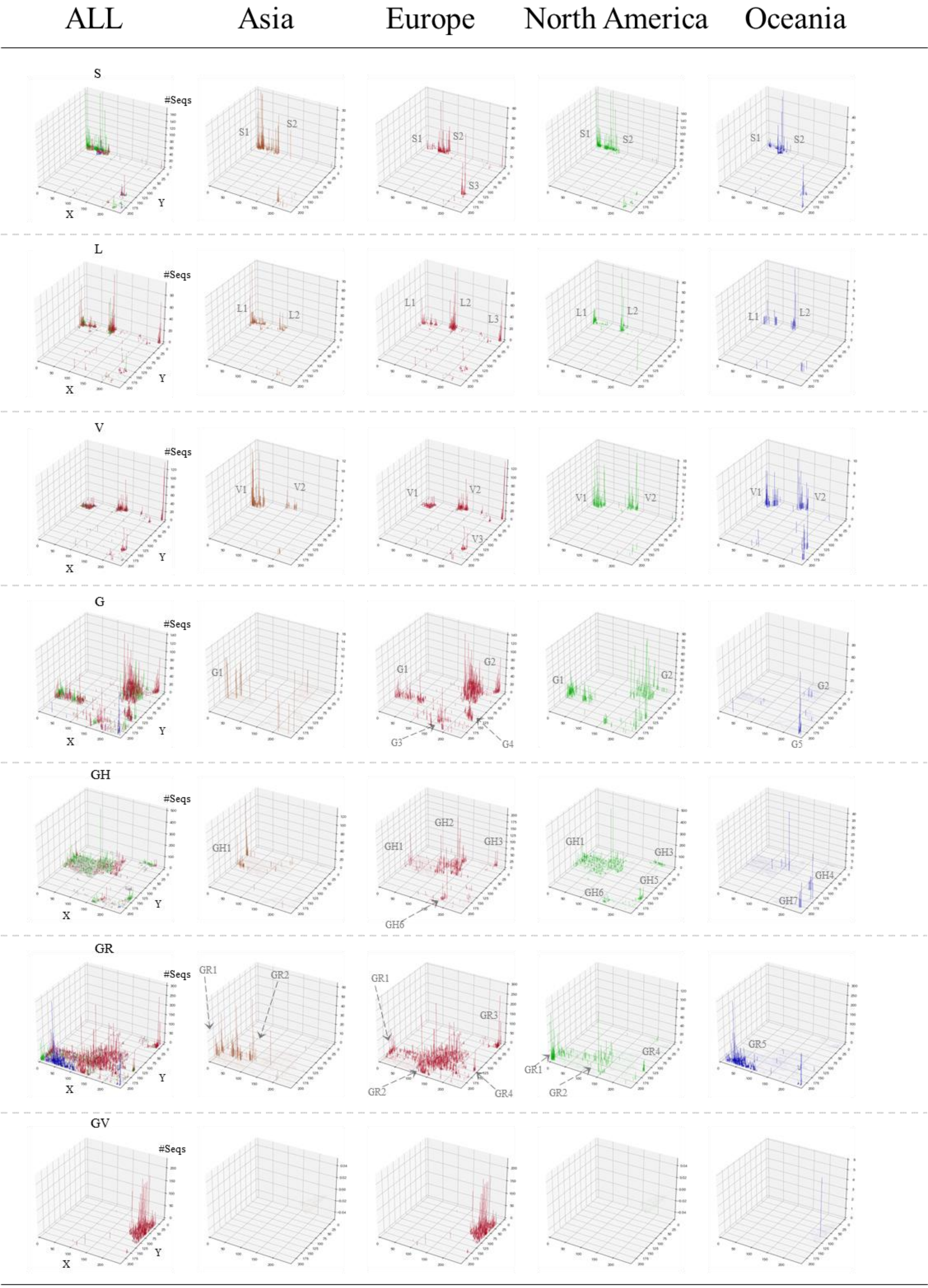

ALL
Asia
Europe
North America
Oceania
S
#Seqs
X
Y
S1
S2
S1
S2
S3
S1
S2
S1
S2
L
#Seqs
X
Y
L1
L2
L1
L2
L3
L1
L2
L1
L2
V
#Seqs
X
Y
V1
V2
V1
V2
V3
V1
V2
V1
V2
G
#Seqs
X
Y
G1
G1
G2
G3
G4
G1
G2
G2
G5
GH
#Seqs
X
Y
GH1
GH1
GH2
GH3
GH6
GH1
GH3
GH5
GH6
GH4
GH7
GR
#Seqs
X
Y
GR1
GR2
GR1
GR3
GR2
GR4
GR1
GR2
GR4
GR5
GV
#Seqs
X
Y

Figure 2

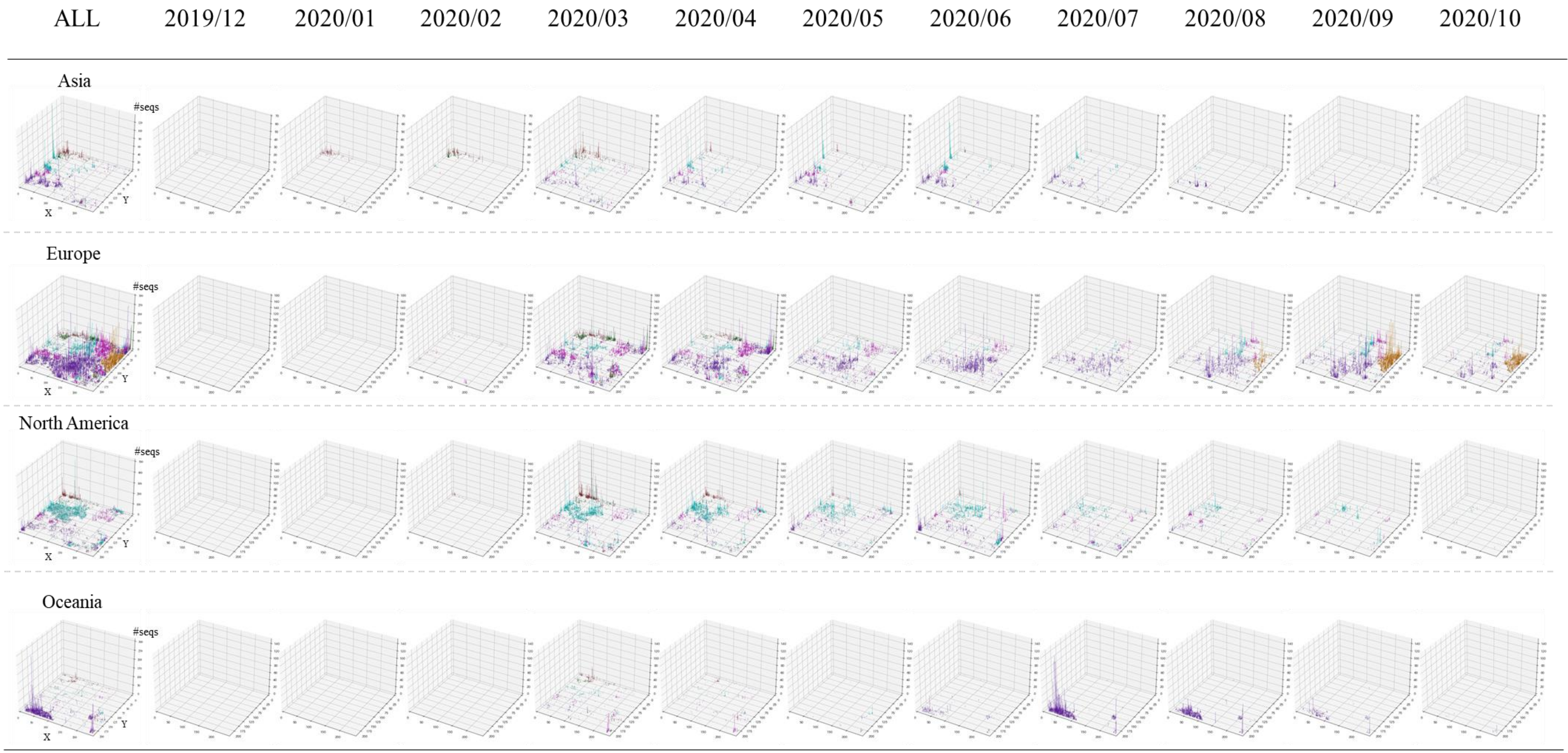

ALL
2019/12
2020/01
2020/02
2020/03
2020/04
2020/05
2020/06
2020/07
2020/08
2020/09
2020/10
Asia
Europe
North America
Oceania
#seqs
X
Y

Figure 3

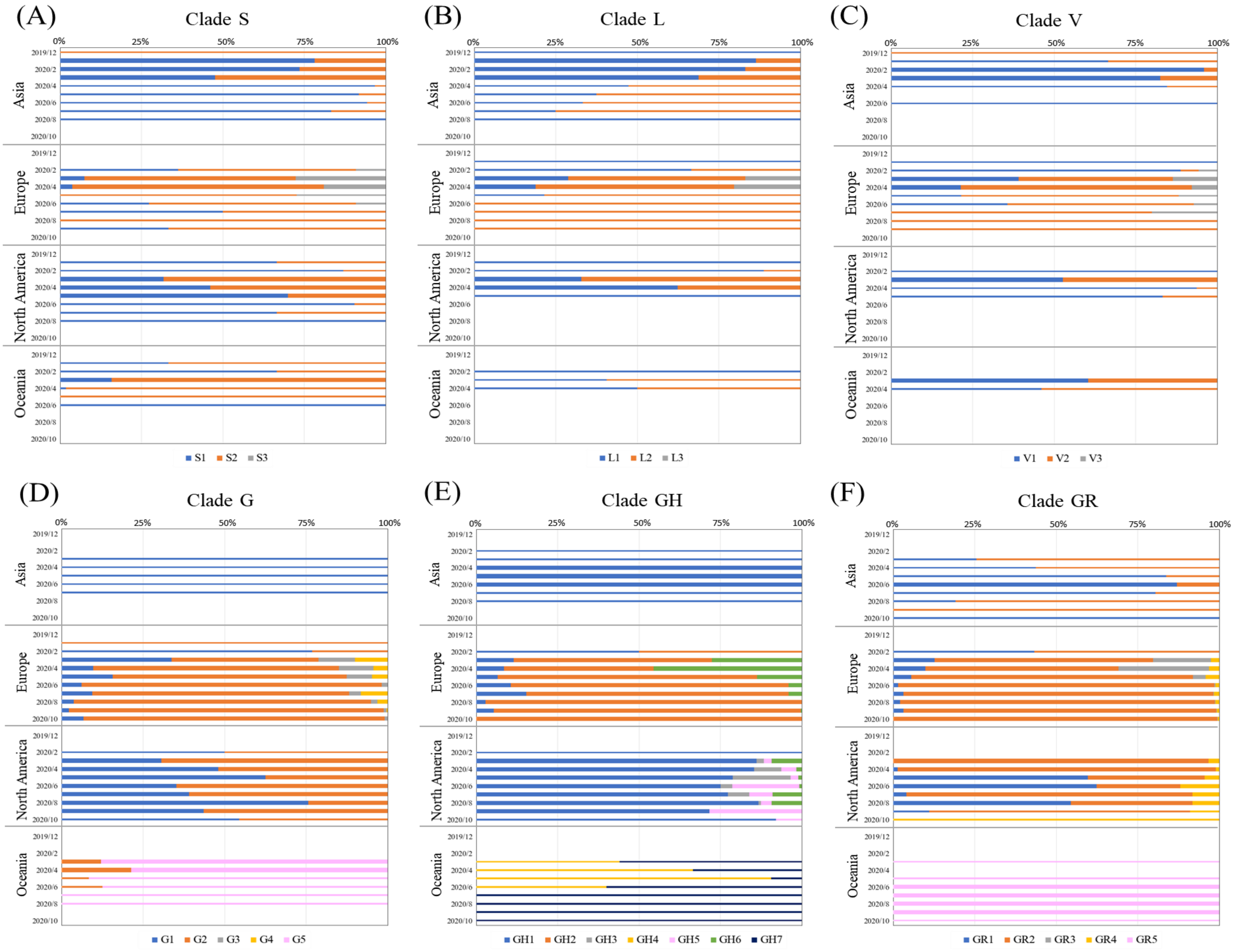
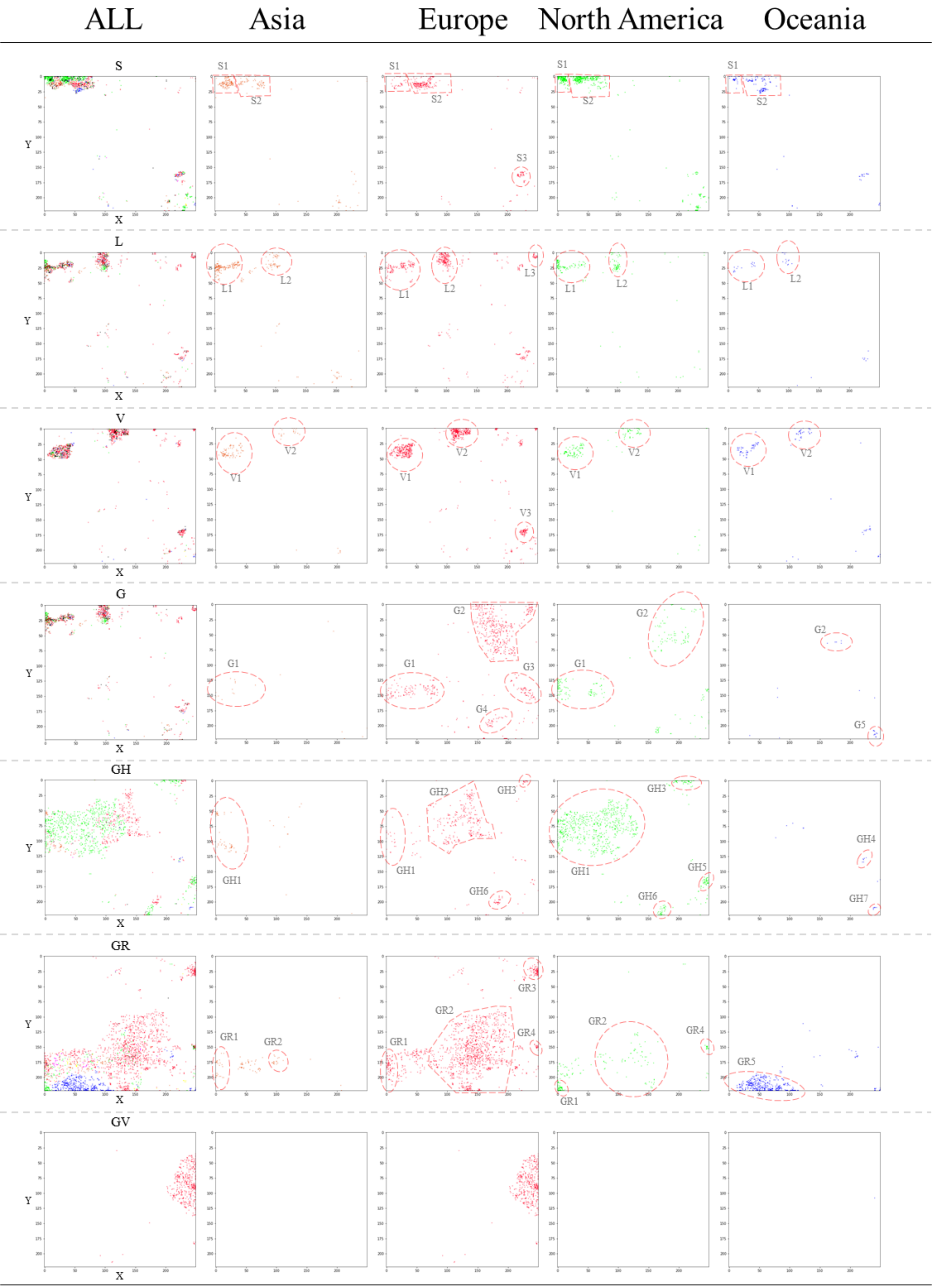

(A)
Clade S
(B)
Clade L
(C)
Clade V
(D)
Clade G
(E)
Clade GH
(F)
Clade GR
0%
25%
50%
75%
100%
Asia
Europe
North America
Oceania
2019/12
2020/2
2020/4
2020/6
2020/8
2020/10
S1 S2 S3
L1 L2 L3
V1 V2 V3
G1 G2 G3 G4 G5
GH1 GH2 GH3 GH4 GH5 GH6 GH7
GR1 GR2 GR3 GR4 GR5

Figure 4



Supplementary Figure S1

Supplementary Table 1. Sequence number of subdivided cluster in clade for each month by continent.

Continent	Asia										
Clade	Clade S										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
S1		112	95	122	59	34	17	5	1		
S2	1	31	34	134	2	3	1	1			
S3											
#Total	1	143	129	256	61	37	18	6	1		
Clade	Clade L										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
L1	21	172	169	144	9	3	3	1	3		
L2		27	34	65	10	5	6	3			
L3											
#Total	21	199	203	209	19	8	9	4	3		
Clade	Clade V										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
V1		2	98	85	11		1				
V2	1	1	4	18	2						
V3											
#Total	1	3	102	103	13		1				
Clade	Clade G										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
G1				5	23	33	29	1			
G2											
G3											
G4											
G5											
#Total				5	23	33	29	1			
Clade	Clade GH										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
GH1			4	39	142	239	310	97	6		
GH2											
GH3											
GH4											
GH5											
GH6											
GH7											
#Total			4	39	142	239	310	97	6		
Clade	Clade GR										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
GR1				21	42	72	121	74	9		4
GR2				61	54	14	18	18	38	34	
GR3											
GR4											
GR5											
#Total	0	0	0	82	96	86	139	92	47	34	4

Continent	Europe										
Clade	Clade S										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
S1			4	52	8		3	1		1	
S2			6	449	163	16	7	1	4	2	
S3			1	191	40	6	1				
#Total			11	692	211	22	11	2	4	3	
Clade	Clade L										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
L1		1	14	297	92	11					
L2			7	562	297	40	9	2	1	2	
L3				174	99						
#Total		1	21	1,033	488	51	9	2	1	2	
Clade	Clade V										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
V1		1	16	542	126	11	15				
V2			1	657	420	34	24	4	1	4	
V3			1	187	46	6	3	1			
#Total		1	18	1,386	592	51	42	5	1	4	
Clade	Clade G										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
G1			30	861	272	118	31	21	23	22	23
G2		4	9	1,152	2,121	540	465	175	568	961	316
G3				288	294	58	9	8	12	8	3
G4				253	122	36		18	19	3	
G5											
#Total		4	39	2,554	2,809	752	505	222	622	994	342
Clade	Clade GH										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
GH1			1	115	82	30	16	23	24	59	
GH2			1	606	441	363	129	121	814	1,034	295
GH3											
GH4											
GH5											
GH6				274	436	62	6	6		2	
GH7											
#Total			2	995	959	455	151	150	838	1,095	295
Clade	Clade GR										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
GR1			13	284	378	94	53	52	62	115	
GR2		0	17	1,499	2,287	1,466	3,365	1,565	2,895	3,488	1,091
GR3				397	1,072	66	9	2	3	8	2
GR4				55	117	70	40	27	34	24	4
GR5											
#Total	0	0	30	2,235	3,854	1,696	3,467	1,646	2,994	3,635	1,097

Continent	North America										
Clade	Clade S										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
S1		2	47	603	276	91	67	2	1		
S2		1	7	1,293	323	39	7	1			
S3											
#Total		3	54	1,896	599	130	74	3	1		
Clade	Clade L										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
L1		4	8	88	113	19					
L2			1	180	68						
L3											
#Total		4	9	268	181	19					
Clade	Clade V										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
V1			2	136	75	5					
V2				122	5	1					
V3											
#Total			2	258	80	6					
Clade	Clade G										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
G1			1	107	259	179	200	101	174	51	6
G2			1	242	280	107	368	157	56	66	5
G3											
G4											
G5											
#Total			2	349	539	286	568	258	230	117	11
Clade	Clade GH										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
GH1			1	2,936	1,946	1,178	1,463	477	409	394	35
GH2											
GH3				77	191	266	69	41	3		
GH4											
GH5				80	106	36	402	44	16	156	3
GH6				313	37	15	14	55	43		
GH7											
#Total			1	3,406	2,280	1,495	1,948	617	471	550	38
Clade	Clade GR										
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10
GR1			0		3	241	586	6	60	5	
GR2			0	176	213	144	240	129	41	36	
GR3											
GR4				6	2	18	112	12	9	4	1
GR5											
#Total	0	0	0	182	218	403	938	147	110	45	1

Continent	Oceania											#Total
Clade	Clade S											#Total
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10	
S1		1	2	46	1		1					1,654
S2		2	1	244	54	2						2,829
S3												239
#Total		3	3	290	55	2	1					4,722
Clade	Clade L											
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10	
L1			1	13	3							1,189
L2				19	3							1,341
L3												273
#Total			1	32	6							2,803
Clade	Clade V											#Total
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10	
V1				81	6							1,213
V2				53	7							1,359
V3												244
#Total				134	13							2,816
Clade	Clade G											#Total
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10	
G1												2,570
G2				19	25	1	1					7,639
G3												680
G4												451
G5				138	92	11	7	27	1			276
#Total				157	117	12	8	27	1			11,616
Clade	Clade GH											#Total
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10	
GH1												10,026
GH2												3,804
GH3												647
GH4				11	16	29	2					58
GH5												843
GH6												1,263
GH7				14	8	3	3	13	8	9	1	59
#Total				25	24	32	5	13	8	9	1	16,700
Clade	Clade GR											#Total
Month	2019/12	2020/1	2020/2	2020/3	2020/4	2020/5	2020/6	2020/7	2020/8	2020/9	2020/10	
GR1												2,295
GR2												18,889
GR3												1,559
GR4												535
GR5				1		5	260	3,722	1,576	314	3	5,881
#Total	0	0	0	1	0	5	260	3,722	1,576	314	3	29,159