

1 Running title: An autoencoder-classified cluster of SARS-CoV-2 strain with two mutations in  
2 helicase

3

4 **Clusters consisting only of virus types with two**  
5 **mutations in the helicase found by Autoencoder**  
6 **analysis in Washington State, USA**

7

8 Jun Miyake\*<sup>1,2,3†</sup>, Mitsuaki Yoshino<sup>4†</sup>, Takaaki Sato<sup>1</sup>, Hirohiko Niioka<sup>5</sup>, Yasushi  
9 Sakata<sup>3</sup>, Yoshihisa Nakazawa<sup>2</sup>

10

11 <sup>1</sup> Department of Material and Life Science, Graduate School of Engineering, Osaka  
12 University, Suita, Osaka, Japan.

13 <sup>2</sup> Hitz Research Alliance Laboratory, Graduate School of Engineering, Osaka  
14 University, Suita, Osaka, Japan.

15 <sup>3</sup> Global Center for Medical Engineering and Informatics, Osaka University, Suita,  
16 Osaka, Japan.

17 <sup>4</sup> National Institute of Biomedical Innovation, Health and Nutrition, Ibaraki, Osaka,  
18 Japan.

19 <sup>5</sup> Institute for Dataability Science, Osaka University, Suita, Osaka, Japan.

20

21 \* Corresponding Author (email-address: [jun\\_miyake@bpe.es.osaka-u.ac.jp](mailto:jun_miyake@bpe.es.osaka-u.ac.jp))

22 †These authors are equally contributed.

23

24 Keywords: COVID-19, SARS-CoV-2 (nCoV-2019), Genome, Helicase, Mutant,  
25 Autoencoder, Washington State, Classification, Cluster

26

27 **Abstract**

28 **Using an autoencoder-based analysis to classify genomes of SARS-CoV-2**  
29 **coronaviruses, we found a cluster consisting only of a specific genotype**  
30 **with two mutations in the helicase. This virus genotype, called C-type**  
31 **SARS-CoV-2, was almost exclusively prevalent in the United States from**  
32 **March to July 2020. This type of virus, characterized by a pair of the**  
33 **C17747T (P504L) and A17858G (Y541C) mutations on the nsp13 gene, had**  
34 **never been highly prevalent at any other time or in any other part of the**  
35 **world. In the U.S., Washington State was the center of the epidemic, and**

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

36 **the C-type viruses, along with the viruses with wild-type helicase, seemed**  
37 **to have aroused the pandemic. In Washington State, USA, the CoViD-19**  
38 **epidemic during the first two months of the year, starting at the end of**  
39 **February 2020, was mainly caused by the type-C virus. During this period,**  
40 **the infection spread rapidly; from May onwards, the number of viruses**  
41 **with wild-type helicases became higher than that of type-C viruses, and**  
42 **no type-C viruses have been collected since early July. The involvement**  
43 **of the helicase in this COVID-19 disease was discussed.**

44

## 45 **Introduction**

46 Coronaviruses have had a major impact on human society since the 2002–2003 SARS  
47 epidemic (Coronaviridae Study Group of the International Committee on Taxonomy of  
48 Viruses. 2020). With the exception of influenza, there are few other viral infections that  
49 have been so prevalent, and it is important to elucidate the characteristics of viruses (Hua  
50 *et al.* 2020). There are questions about how host cells attack coronaviruses and whether  
51 there are other ways to defend against them besides antibodies. The coronavirus–host cell  
52 interactions such as the viral RNA-synthesizing machinery and the evasion of host innate  
53 immune responses are detailed in a review (de Welde *et al.* 2018).

54 Host organisms, including humans, have the ability to detect viral infection as  
55 "non-self-nucleic acid invasion," which is expected to trigger antiviral innate immunity  
56 and suppress viral replication. New coronaviruses have continued to mutate at a high rate  
57 (Gómez-Carballa *et al.* 2020; Jones and Manrique 2020; Nie *et al.* 2020; Rochman *et al.*  
58 2020). Since its emergence, new variants have appeared one after another, and there are  
59 now more than eight in total (Miyake *et al.* 2021). The types, viral infections, disease  
60 patterns, and mutations of the encoded proteins have not yet been comprehensively  
61 understood. We believe that significant advances in our understanding of this complex  
62 system will be made in the near future.

63 The viral helicase (nsp13 in ORF1ab), which plays a central role in the immune  
64 response of host cells, is stably maintained in various strains, and a mutant form was first  
65 isolated in Washington State, USA in February 2020 (Chan *et al.* 2020). This variant is  
66 present only in a specific genotype (Zhao *et al.* 2020), which Chan *et al.* define as

67 coevolutionary variant group 4 (CEVg4). In our previous paper (Miyake et al. 2021), we  
68 presented a method to analyze clusters of viral genomes using an autoencoder. Among  
69 the clusters classified by this method, those consisting only of genomes with mutations  
70 in the nsp13 gene on ORF-1ab (C-type SARS -CoV-2 virus genome). Since the other  
71 clusters did not show the said mutation in nsp13, it was considered that the genomes  
72 constituting the clusters had some aggregate characteristics. A closer look at the genomic  
73 variation revealed that the helicase translated and synthesized from nsp13 had double  
74 mutations, P504L and Y541C.

75 In silico analysis to evaluate protein-drug interactions suggests that helicase,  
76 the nsp13 product with the double mutation of P504L and Y541C, has an altered shape  
77 of the ATP-binding site (Ugurel et al.). In this paper, along with the analysis of the  
78 clusters, we also examined their relationship to disease during the pandemic. The ration  
79 of wild type (WT)/C SARS-CoV-2 virus ratio based on the weekly number of viruses  
80 classified by the autoencoder and compared the relationship with the number of infected  
81 people and deaths. In the case of the pandemic in Washington State, USA, we found that  
82 the type C virus was dominant in the first two months, and that the virus with wild-type  
83 helicase became overwhelmingly dominant after May 2020. We will also discuss the  
84 involvement of helicases in the current COVID-19 disease.

85

## 86 **Methods**

87 We analyzed the coronavirus epidemic in Washington State, U.S.A. from 2020/2/29 to  
88 2020/7/31. Genome sequence data of novel coronaviruses were collected from the NCBI  
89 Virus database (downloaded on February 26, 2021). Only complete genome sequences  
90 were analyzed (genes with more than two consecutive bases were excluded from the  
91 analysis). Cluster analysis of gene sequences using autoencoders was performed using  
92 the Tensor-Flow library (downgraded from V2.0 to V1.0). computers equipped with  
93 GPUs (NVIDIA Quadro P-6000) were used. The autoencoder previously classified the  
94 genotypes of the human leukocyte antigen A (HLA-A) gene (Miyake et al. 2018): a  
95 histogram of the frequency of occurrence of 1,024 5-mer words in the SARS-CoV-2  
96 ORF1ab gene was generated using a four-layer autoencoder (1,000,000 epochs with a  
97 batch size of 96) was compressed into a three-dimensional array, which were plotted in

98 three-dimensional (3D) space as (x, y, z) coordinates. As a result, 31,050 SARS-CoV-2  
99 ORF1ab genes were classified into eight categories (A1, A2, B1, B2, C, D, E1, and E2  
100 types) (Miyake et al. 2021). Epidemiological data were obtained from the Centers for  
101 Disease Control and Prevention (CDC) database.

102 Based on these autoencoder-classified SARS-CoV-2 ORF1ab genes, we  
103 examined time courses of weekly ratios of SARS-CoV-2 viral genomes with wild-type  
104 (WT) helicase (WT helicase genomes belonging to clusters A1, A2, B1, B2, D, E1 and  
105 E2) to those of C-type helicase (C-type genomes belonging to cluster C) that were  
106 collected in Washington State. Note that the C genome was collected only during weeks  
107 4-20 starting from February 1, 2020; the period starting from week 21 was excluded from  
108 the analysis.

109

## 110 **Results**

111 The ORF1ab gene was extracted from the SARS-CoV-2 genome and classified into eight  
112 clusters in three-dimensional space by autoencoder. Among them, C-type was located at  
113 the far end of the cluster spread. It was a small cluster with relatively clear boundaries.  
114 ORF1ab genes of cluster C which have C-type helicase were expanded in March (Fig.  
115 1a) and disappeared after July 3, 2020 (Fig. 1b).

116 Helicase is an essential enzyme for viral replication, but a special type of mutant  
117 was found in the United States in February to July, 2020. This mutant is characterized by  
118 the presence of mutations in the nsp13 (helicase) gene of ORF1ab that differ by two bases  
119 from the wild type (C17747T (P504L) and A17858G (Y541C)) (Chan *et al.*, 2020). C-  
120 type of ORF1ab was consisted only with the mutant genome. The mutant was not found  
121 in any other clusters. We will refer to this helicase as helicase-C in this paper. Note that  
122 there is also a helicase in which only one of the two mutations was different, but due to  
123 the very small number, it was excluded from this analysis. This cluster (category)  
124 corresponds to the CEVg4 reported by Chan *et al.* in their combinatorial analysis of  
125 genomic variation (Chan *et al.*, 2020).

126 The classification of the clusters as in Fig.1 were plotted against the time axis and  
127 the distance of the 3D data from the center for nine states in the United States (Fig. 2).

128 Each cluster was plotted in a different color (A1, red; A2, pink; B1, green; B2, light green;  
129 C, purple; D, orange; E1, blue; E2, light blue). The vertical axis is the distance of the 3D  
130 data used in the autoencoder classification (The nucleotide sequence (29903 nt in the  
131 reference sequence) was compressed into 3 dimensions by autoencoder analysis. The  
132 genomes in the cluster were easily identifies by their positions. Weekly numbers of  
133 helicase-C genomes in the United States, Washington State, and Australia were shown in  
134 Fig. 3.

135 The usefulness of the autoencoder classification method is that it can automatically  
136 classify the genomes into clusters without knowing the differences in gene sequences  
137 beforehand, it can handle a huge number of genomes, and it is easy to understand the  
138 temporal variation. In other words, it is a way to spatially represent the "artificial  
139 intelligence recognition" of the full-length sequence of genes. The classification of  
140 clusters depends on where the boundaries are drawn, so there can be some differences in  
141 the number of genes included in some cases.

142 The country subcategories of helicase-C SARS-CoV-2 genomes were 1,460  
143 (97.6%) in the United States, 32 (2.1%) in Australia, 2 in the United Kingdom, 1 in Peru,  
144 and 1 in Puerto Rico (country data were taken from NCBI Virus Database). Australia and  
145 less countries were not examined in this study due to small sample size.

146 In the U.S., helicase-C has been detected in areas where the SARS-CoV-2  
147 outbreak has caused infections, such as California, New York, and Washington State. In  
148 the U.S., the state of Washington had the highest number of database-registered helicase-  
149 C genomes at 68.9% (67.4% worldwide) throughout the period of February to June. This  
150 is followed by California (10.5%). Other states were Minnesota (2.8%), Utah (2.8%),  
151 Florida (1.5%), and Massachusetts (1.4%). The portion of the infections and the high  
152 mortality rate in Washington State is remarkable. Washington State was thought to be the  
153 single source of the outbreak and that it spread to other states.

154 In Washington State, helicase-C SARS-CoV-2 genomes were initially  
155 predominant in the four–eight weeks from February 1, 2020. The WT/C-type helicase  
156 ratio logarithmically increased with a kink at nine weeks when the dominance of WT  
157 helicase genomes was initiated (Fig. 4a). For comparison, weekly numbers of cumulative

158 COVID-19 cases and deaths in Washington State (obtained from the Center for Disease  
159 Control and Prevention (CDC) webpage) were shown (Fig. 4b). Weekly numbers of  
160 cumulative cases also logarithmically increased with a kink at nine weeks when the  
161 growth rate markedly decreased. The cumulative number of deaths showed a one-week  
162 lagged kink (week 10) in the logarithmic increase, with a marked decrease in the late  
163 phase. The coincidence of sinks in WT/C helicase ratios, cases, and deaths in Washington  
164 State with a 1-2 week lag suggests that the change in dominance of the SARS-CoV-2  
165 helicase-C genome to the WT helicase-C genome may be responsible for the decrease in  
166 cases and deaths.

167

## 168 **Discussion**

169 In the cluster analysis using the autoencoder, the genome of SARS-CoV-2 with helicase  
170 C could be extracted as one cluster. The other genomes with helicase WT were distributed  
171 in several clusters. These results indicate that the autoencoder can identify the ORF1ab  
172 gene of SARS-CoV with only two mutations in the helicase. Although the genomic  
173 sequence of the helicase is highly conserved in nature, mutations in such a highly  
174 conserved sequence may form specific clusters that are clearly separated by autoencoder  
175 analysis. The ability of autoencoders to extract genomes with significant mutations in  
176 specific gene regions as separate clusters will be beneficial not only for helicases but also  
177 for a variety of other application.

178 Type-C SARS-CoV-2 virus (hereafter referred to as "type-C virus") was first  
179 discovered in Washington State on February 20 (Chan et al. 2020). In June (after week  
180 21), the number of type-C virus collections has decreased significantly. The last virus was  
181 obtained in Florida on July 2. In the early stages of the pandemic (4-8 weeks counting  
182 from February 1, 2020), type-C virus initially predominated and may have contributed to  
183 the number of COVID-19 cases. By weeks 10-20, the WT helicase genome predominated.  
184 Looking at the semilog plot in Fig. 4, it is understood that the ratio of deaths/infected  
185 cases was considerably larger in the early phase than in the later phase. The inflection  
186 point of the number of infected and deaths was off by almost a week. Assuming a one-  
187 week gap between infection and outcome/death, the ratio of deaths/infected may have  
188 been temporarily more than twice as high in the early phase (4-8 weeks) as in the late

189 phase. In the late phase (10-20 weeks), this ratio was consistently less than 6%. In the  
190 later stages, this ratio was consistently less than 6%. If this large difference is dependent  
191 on the type of virus, it should be fully considered as a preparation for future viral diseases.  
192 The emergence, disappearance, and pathogenesis of type-C virus A detailed analysis of  
193 the mechanisms, mortality dependence, and other factors may be useful as a key to  
194 developing effective treatments and antiviral measures.

195 All C-type viruses contain a variant of the helicase, helicase-C, which is highly  
196 suspected as a cause of serious disease. Helicase is a protein that unwinds nucleic acids  
197 and is essential for viral genome replication. Adedeji *et al.* found that helicase activity  
198 was enhanced in the presence of nsp12 in the SARS-CoV epidemic of 2002-2003  
199 (Adedeji *et al.* .2012). Kindler *et al.* (Kindler *et al.* 2017; Ancar *et al.* 2020) and Deng *et*  
200 *al.* (Deng *et al.* 2017; Deng and Baker 2018) considered the following scheme. The viral  
201 RNA forms a double-stranded structure at the poly-U sequence. If the virus can use RNA-  
202 degrading enzymes (EndoU) to cleave its own RNA-poly U sequence, the immune response of  
203 the host cell is inactivated. The structure and interactions of the proteins discussed here are being  
204 studied. Pillon *et al.* published the four-dimensional protein structure of SARS-CoV-2 nsp15  
205 using Cryo-EM (Pillon *et al.* 2021). Perry *et al.* also studied that nsp 15 protein is surrounded by  
206 nsp14, 16, nsp7, 8, and 12 proteins, as well as nsp13 (helicase) using Cryo-EM (Perry *et al.* 2021).

207 Several researchers have pointed out that helicases are important targets for  
208 viral clearance (Gurung 2020; Habtemariam *et al.* 2020; White *et al.* 2020). Possible  
209 strategies for drug discovery include the application of existing drugs (Li and De Clercq  
210 2020; Pandey *et al.* 2020) and RNAi (Liu *et al.* 2020). Existing compounds that have the  
211 effect of inhibiting the RNA or protein molecules of helicases could also be used.  
212 Computer modeling studies are also underway (Gurung 2020; White *et al.* 2020). It is  
213 expected that various efforts will be made in drug development.

214 Various vaccines have been developed and used now. The biochemical  
215 reactions involved in viral replication in host cells may be used as a countermeasure. If it  
216 is possible to inhibit the intracellular replication mechanism common to various viruses,  
217 it may be an effective countermeasure.

218

219 **Acknowledgements**

220 We should like to express our sincere thanks to Dr. Tomonori Kimura of NIBIOHN for  
221 support and discussions. Thanks are to Hayao Nakamura, Yuta Nitada and Shunsuke  
222 Baba for programming and computer operation. This work was supported partially by  
223 Global Center for Medical Engineering and Informatics, Osaka University, Hitz Research  
224 Alliance Laboratory, Nihon Unisys, Ltd. and Japan Agency for Medical Research and  
225 Development (Grant Number 20bm0804008h0004.PI. Prof. S. Miyagawa).

226

## 227 **References**

228 Adedeji AO, Marchand B, Te Velhuis AJ, Snijder EJ, Weiss S, Eoff RL, Singh K, Sarafianos  
229 SG. Mechanism of nucleic acid unwinding by SARS-CoV helicase. *PLoS One*. 7:e36521  
230 (2012). doi: 10.1371/journal.pone.0036521.

231 Ancar R, Li Y, Kindler E, Cooper DA, Ransom M, Thiel V, Weiss SR, Hesselberth JR, Barton  
232 DJ. Physiologic RNA targets and refined sequence specificity of coronavirus EndoU. *RNA*.  
233 26:1976–1999 (2020). doi: 10.1261/rna.076604.120.

234 Centers for Disease Control and Prevention (CDC, Official Updates Coronavirus - COVID-19 in  
235 United States): [https://coronavirus.dc.gov/?gclid=Cj0KCCQiA5vb-BRCRARIsAJBKc6L-  
236 50jCtxFFaFZe05Os84e-Mb6dLiWHBepI9bJdHXF2rABJmLpyyIaAmxBEALw\\_wcB](https://coronavirus.dc.gov/?gclid=Cj0KCCQiA5vb-BRCRARIsAJBKc6L-50jCtxFFaFZe05Os84e-Mb6dLiWHBepI9bJdHXF2rABJmLpyyIaAmxBEALw_wcB).

237 Chan AP, Choi Y, Schork NJ. Conserved genomic terminals of SARS-CoV-2 as coevolving  
238 functional elements and potential therapeutic targets. *mSphere* 5: e00754-20 (2020). doi:  
239 10.1128/mSphere.00754-20.

240 Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species  
241 Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it  
242 SARS-CoV-2. *Nat. Microbiol*. 5:536–544 (2020). doi: 10.1038/s41564-020-0695-z.

243 de Wilde AH, Snijder EJ, Kikkert M, van Hemert MJ. Host factors in coronavirus replication,  
244 *Curr Top Microbiol Immunol*. 419:1–42 (2018). doi: 10.1007/82\_2017\_25.

245 Deng X, Hackbart M, Mettelman RC, O'Brien A, Mielech AM, Yi G, Kao CC, Baker SC.  
246 Coronavirus nonstructural protein 15 mediates evasion of dsRNA sensors and limits apoptosis  
247 in macrophages. *Proc Natl Acad Sci U S A*. 114:E4251–E4260 (2017). doi:  
248 10.1073/pnas.1618310114.

249 Deng X, Baker SC. An "Old" protein with a new story: Coronavirus endoribonuclease is  
250 important for evading host antiviral defenses. *Virology*. 517:157–163 (2018). doi:  
251 10.1016/j.virol.2017.12.024.

- 252 Gómez-Carballa A, Bello X, Pardo-Seco J, Martín-Torres F and Salas A. Mapping genome  
253 variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders.  
254 *Genome Res.* 30:1434–1448 (2020). doi: 10.1101/gr.266221.120.
- 255 Gurung AB. *In silico* structure modelling of SARS-CoV-2 Nsp13 helicase and Nsp14 and  
256 repurposing of FDA approved antiviral drugs as dual inhibitors. *Gene Rep.* 21:100860 (2020).  
257 doi: 10.1016/j.genrep.2020.100860.
- 258 Habtemariam S, Nabavi SF, Banach M, Berindan-Neagoe I, Sarkar K, Sil PC, Nabavi SM. Should  
259 we try SARS-CoV-2 helicase inhibitors for COVID-19 therapy? *Arch Med Res.* 51:733–735  
260 (2020). doi: 10.1016/j.arcmed.2020.05.024.
- 261 Hua Li, Zhe Liu, Junbo Ge, Scientific research progress of COVID-19/SARS-CoV-2 in the first  
262 five months, *J Cell Mol Med.* 24:6558–6570 (2020). doi: 10.1111/jcmm.15364.
- 263 Jones LR, Manrique JM. Quantitative phylogenomic evidence reveals a spatially structured  
264 SARS-CoV-2 diversity. *Virology* 550:70–77 (2020). doi: 10.1016/j.virol.2020.08.010.
- 265 Kindler E, Gil-Cruz C, Spanier J, Li Y, Wilhelm J, Rabouw HH, Züst R, Hwang M, V'kovski P,  
266 Stalder H, Marti S, Habjan M, Cervantes-Barragan L, Elliot R, Karl N, Gaughan C, van  
267 Kuppeveld FJ, Silverman RH, Keller M, Ludewig B, Bergmann CC, Ziebuhr J, Weiss SR,  
268 Kalinke U, Thiel V. Early endonuclease-mediated evasion of RNA sensing ensures efficient  
269 coronavirus replication. *PLoS Pathog.* 13:e1006195 (2017). doi:  
270 10.1371/journal.ppat.1006195.
- 271 Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nat Rev*  
272 *Drug Discov.* 19:149–150 (2020). doi: 10.1038/d41573-020-00016-0.
- 273 Liu C, Zhou Q, Li Y, Garner LV, Watkins SP, Carter LJ, Smoot J, Gregg AC, Daniels AD, Jervey  
274 S, Albaiu D. Research and Development on Therapeutic Agents and Vaccines for COVID-19  
275 and Related Human Coronavirus Diseases. *ACS Cent Sci.* 6:315–331 (2020). doi:  
276 10.1021/acscentsci.0c00272.
- 277 Miyake J, Kaneshita Y, Asatani S, Tagawa S, Niioka H, Hirano T. Graphical classification of  
278 DNA sequences of HLA alleles by Deep learning. *Human Cell* 31:102–105 (2018). doi:  
279 10.1007/s13577-017-0194-6.
- 280 Miyake J, Sato T, Baba S, Nakamura H, Niioka H, Hirano T, Nakazawa Y. Cluster analysis of  
281 SARS-CoV-2 gene using deep learning autoencoder: Gene profiling for mutations and  
282 transitions. *bioRxiv* 2021.03.16.435601 (2021). doi: 10.1101/2021.03.16.435601. Preprint.
- 283 NCBI Virus Database: <https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/find-data/virus>
- 284 Nie Q, Li X, Chen W, Liu D, Chen Y, Li H, Li D, Tian M, Tan W, Zai J. Phylogenetic and  
285 analyses of SARS-CoV-2. *Virus Res.* 287:198098 (2020). doi: 10.1016/j.virusres.2020.198098.

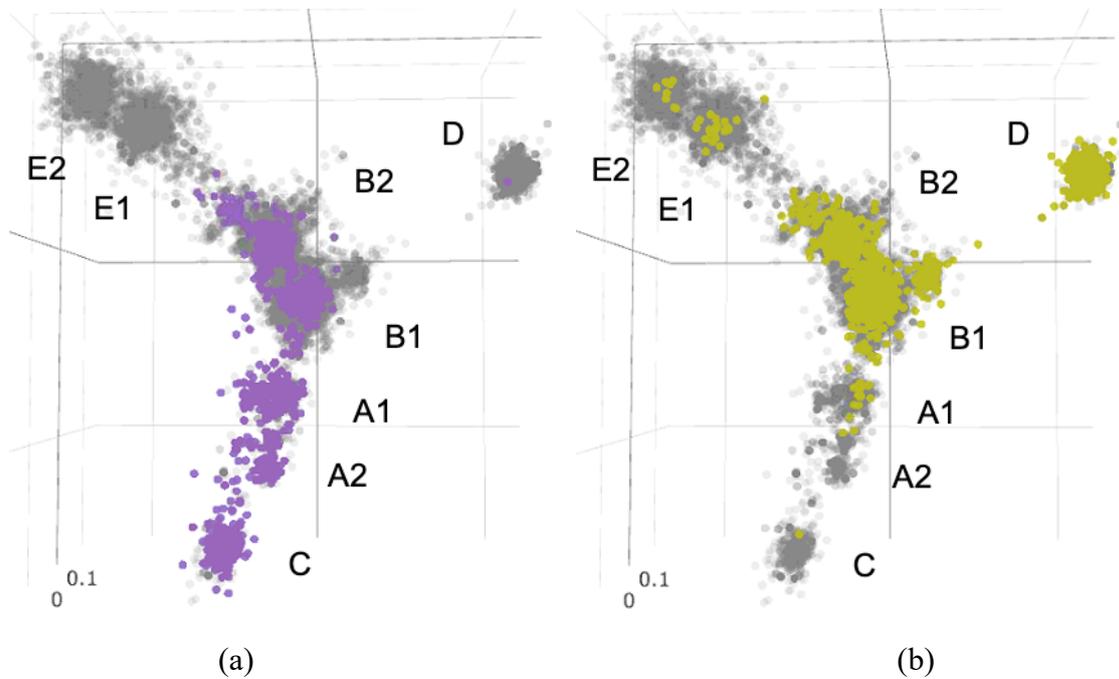
- 286 Pandey SC, Pande V, Sati D, Upreti S, Samant M. Vaccination strategies to combat novel corona  
287 virus SARS-CoV-2. *Life Sci.* 256:117956 (2020). doi: 10.1016/j.lfs.2020.117956.
- 288 Perry JK, Appleby TC, Bilello JP, Feng JY, Schmitz U, Campbell EA. An atomistic model of  
289 the coronavirus replication-transcription complex as a hexamer assembled around nsp15.  
290 *bioRxiv* 2021.06.08.447516 (2021). doi: <https://doi.org/10.1101/2021.06.08.447516>. Preprint.
- 291 Pillon MC, Frazier MN, Dillard LB, Williams JG, Kocaman S, Krahn JM, Perera L, Hayne CK,  
292 Gordon J, Stewart ZD, Sobhany M, Deterding LJ, Hsu AL, Dandey VP, Borgnia MJ, Stanley  
293 RE. Cryo-EM structures of the SARS-CoV-2 endoribonuclease Nsp15 reveal insight into  
294 nuclease specificity and dynamics, *Nature Communications* 12, Article number: 636 (2021)  
295 <https://doi.org/10.1038/s41467-020-20608-z>
- 296 Rochman ND, Wolf YI, Faure G, Zhang F, Koonin EV. Ongoing adaptive evolution and  
297 globalization of Sars-Cov-2. *bioRxiv.* 2020.10.12.336644 (2020). doi:  
298 10.1101/2020.10.12.336644. Preprint.
- 299 Ugurel OM, Mutlu O, Sariyer E, Kocer S, Ugurel E, Inci TG, Ata O, Turgut-Balik D. Evaluation  
300 of the potency of FDA-approved drugs on wild type and mutant SARS-CoV-2 helicase  
301 (Nsp13). *Int J Biol Macromol.* 163:1687–1696 (2020). doi: 10.1016/j.ijbiomac.2020.09.138.
- 302 White MA, Lin W, Cheng X. Discovery of COVID-19 inhibitors targeting the SARS-CoV-2  
303 Nsp13 helicase. *J Phys Chem Lett.* 11:9144–9151 (2020). doi: 10.1021/acs.jpcclett.0c02421.
- 304 Zhao J, Zhai X, Zhou J. Snapshot of the evolution and mutation patterns of SARS-CoV-2.  
305 *bioRxiv.* 2020.07.04.187435 (2020). doi: 10.1101/2020.07.04.187435. Preprint.
- 306 Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen  
307 HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao  
308 K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia  
309 outbreak associated with a new coronavirus of probable bat origin. *Nature* 579:270–273  
310 (2020), doi: 10.1038/s41586-020-2012-7.
- 311
- 312

313

314 **Figures**

315

316



320 **Fig.1. Autoencoder clustered 3D view of ORF1ab gene of SARS-CoV-2 (March and**

321 **July 2020).**

322 The number of C-type ORF1ab genes expanded rapidly in March, and new appearances

323 stopped as of July 3, 2020. (a) ORF1ab gene that appeared in March 2020 (purple). (b)

324 ORF1ab gene that appeared in July 2020 (yellow-green). Shown in gray as background

325 is the cluster of ORF1ab genes for the period 19/12/2019-2021/2/16.

326

327  
328  
329  
330  
331

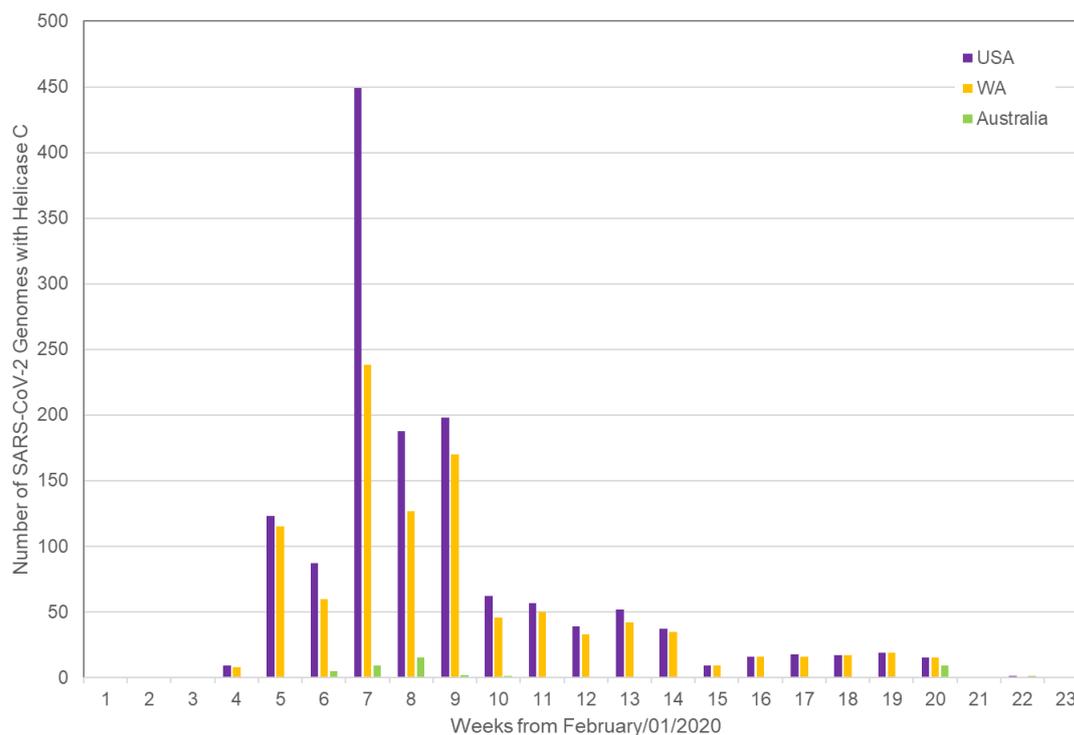


332  
333  
334  
335  
336  
337  
338  
339

**Fig.2. Evolution of clusters in nine US states.**

Eight clusters classified by the autoencoder of the SARS-CoV-2 ORF1ab gene in nine U.S. states (WA, CA, UT, FL, MA, NY, MN, WI, ME) plotted in different colors by collection date (A1, red; A2, pink; B1, green; B2, light green; C, purple; D, orange; E1, blue; E2, light blue). Cluster C with helicase C appeared in February and disappeared after July 3, 2020.

340



341

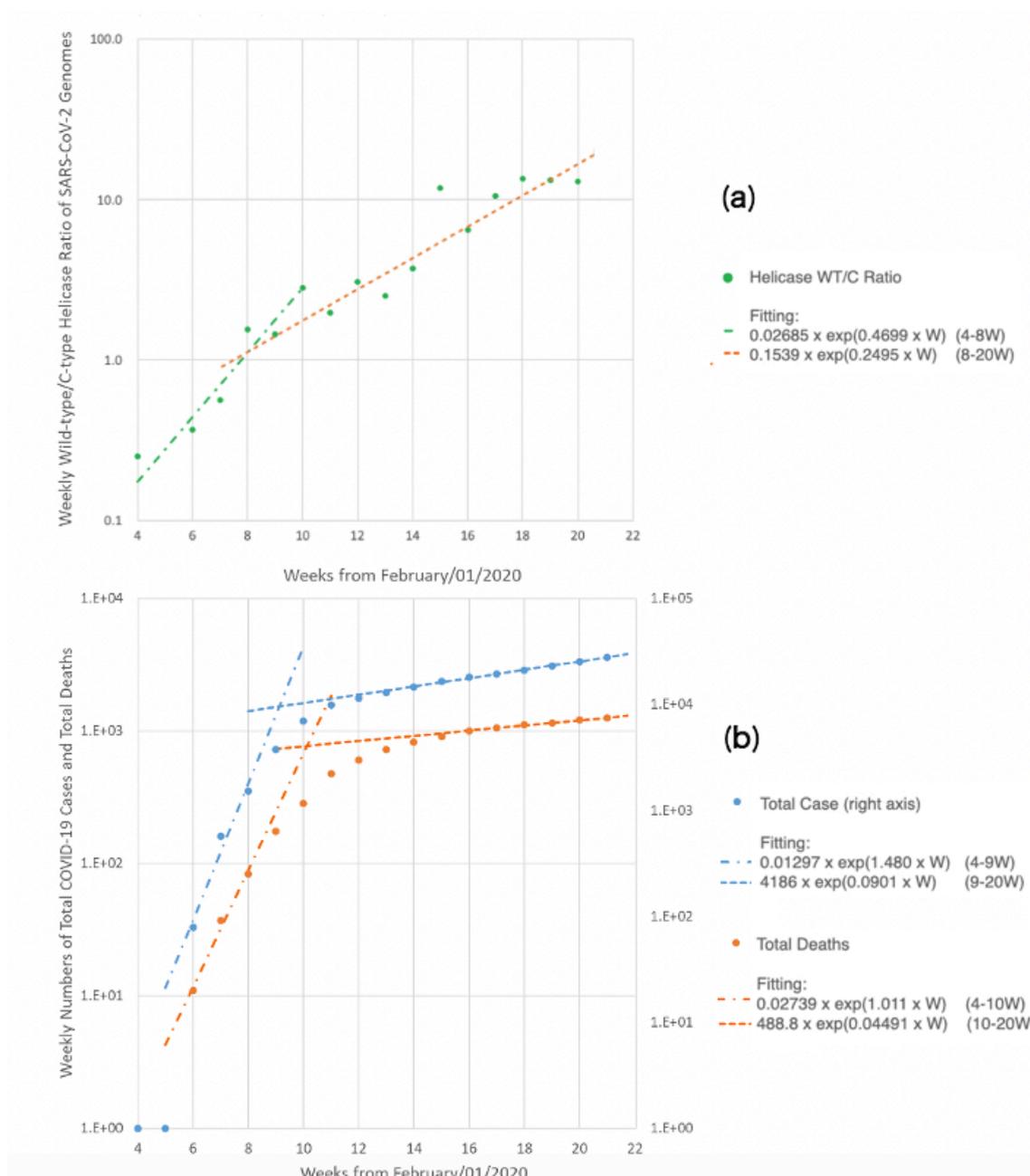
342 **Fig. 3. Transition of C-type SARS-CoV-2 in USA, Washington State and Australia.**

343 Weekly numbers of autoencoder-classified SARS-CoV-2 ORF1ab genes in USA,

344 Washington State, and Australia are shown by number of weeks from

345 February/01/2020.

346



347  
348  
349  
350  
351

**Fig. 4. Ratio of the numbers of wild-type and C-type helicase SARS-CoV-2 viruses collected in Washington State in USA.**

352 (a) The ratio of wild-type/C-type helicase SARS-CoV-2 genomes (green) showed  
353 logarithmic increases with a kink around nine weeks from February 1, 2020 in  
354 Washington State in USA. The first half corresponds to the period when the C-type  
355 helicase viruses rapidly increased, and the second half to the period when the wild-type

356 helicase viruses became predominant. Total deaths (orange) and total cases (blue) also  
357 showed logarithmic increases with kinks around ten and nine weeks, respectively.