Identification of a low pathogenicity clade of SARS-CoV-2

Takahiko Koyama, Laxmi Parida

IBM TJ Watson Research Center, Yorktown Heights, NY 10598, USA

Correspondence to Takahiko Koyama (email: tkoyama@us.ibm.com, postal address: 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA)

Abstract

Number of confirmed cases of COVID-19 caused by SARS-CoV-2 exceeded 5 million as of May 21, 2020. Global average of the case fatality rate of COVID-19 is about 7% so far. There exist variations in case fatality rates among countries. Particularly, Singapore and Qatar have exceptionally low case fatality rates with 0.1% while France's rate is almost 20%. Since no magic bullet treatment for COVID-19 exists, we investigated SARS-CoV-2 strains specific to Singapore in this study to identify a clade with low pathogenicity. Variant analysis revealed that a clade with variants ORF1ab L3606F, A4489V, S2015R, T2016K, and N P13L is common in Singapore. Based on our analysis of variants and historical case statistics, the clade is dominant in a recent surge. Therefore, we suggest that low case fatality rate of Singapore possibly is attributed to the clade. Although contribution of each variant to the low pathogenicity is not clear, L3606F alone does not accomplish such low pathogenicity from the comparison with case fatality data from Japan, where L3606F is dominant. Further investigation is necessary to conclude to validate this finding.

Keywords:

COVID-19; SARS-CoV-2; variant; low pathogenicity; Singapore

Introduction

Coronavirus disease 2019 (COVID-19) caused by a beta coronavirus strain, SARS-CoV-2, was first discovered in Wuhan, China in late 2019. The virus quickly spread throughout the world in early 2020. With slight delay from China, European countries were seriously affected. Especially, United Kingdom, Italy and Spain were hit hard. As of May 21, 2020, over 5 million confirmed cases worldwide have been reported and the death toll has exceeded 330,000. The United States is leading in both number of cases and death toll. Social distancing and other measures have proved to be effective and cases in United States appear to have subsided. Now, cases in Russia, Brazil and India are increasing.

It is reported that the virus can transmit before the on-set of the symptoms[1]. In serious cases, patients develop acute respiratory distress syndrome (ARDS)[2,3]. About 20% of confirmed cases are serious and 5% are critical[4]. Older patients with comorbidities such as hypertensions, obesity and diabetes have high fatality rate[5].

As the viral genome accumulates mutations, low or high pathogenicity strain might emerge. Quick identification of high and low pathogenicity strains is critical to make policies for lockdown or resumption of economic activities. In this study, we focus on identification of low pathogenicity strains from genomes and case fatality statistics.

Methods

A scattered plot of number of confirmed cases vs number of deaths by countries is drawn to elucidate case fatality rate as shown in Figure 1. The data was downloaded from Our World in Data (https://ourworldindata.org/coronavirus) on May 14th and cumulative data on May 13th was used for the plot. For each country, color was assigned to indicate case fatality rate.

Figure 2 shows the result of the variant analysis performed for genome samples from Singapore. Genomes were downloaded from GISAID[6] on May 1st, 2020 and analyzed using the method[7]. First, genomes were aligned to the reference genome NC_045512 using EMBOSS needle[8]. Then, differences with the reference genome were extracted and annotated into different variant types such as missense, synonymous, non-coding and deletions.

To estimate the number of cases of L3606F/P13L/A4489V/T2016K/S2015R in Singapore, we used the collection dates of genomes and historical numbers of confirmed cases. For each date starting January 15th, we first estimate the proportion of the clade and other clades using kernel density estimation with Gaussian kernel as shown in Figure 3 A. Then, we estimate the number of confirmed cases from the proportion and the number of confirmed cases for each date between January 15th and May 13th as illustrated as Figure 3 B.

Using genomes with ORF1ab L3606F variant, phylogenetic analysis was performed using BEAST[9] as shown in Figure 4. First, alignment of 1071 genomes was performed by MAFFT[10] followed by BEAST analysis using collection dates with the Hasegawa-Kishino-Yano mutation model[11], strict clock. Each genome was colored by country, and variant statuses of P13L, A4489V, T2016K, and S2015R were shown in the binary heatmap next to the phylogenetic tree.

Results

As seen in Figure 1, Singapore and Qatar have exceptionally low case fatality rates with numbers of confirmed cases exceeding 10,000. While France has case fatality rate of almost 20%, these two countries have less than 0.1% of case fatality rates. Since ethnicity background is quite different, it is not easy to directly compare these numbers. However, we assume that not all but part of the difference attributes to the strain differences between them. Because Qatar has only two genomes available, we focus our analysis on Singapore, which has 71 genomes.

Variant analysis revealed that L3606F/P13L/A4489V/T2016K/S2015R is common in recent surge in Singapore as shown in Figure 2. Interestingly, we estimated that this clade is dominant in Singapore. In our estimate, 10721 out of 11001 or 97% of confirmed cases in Singapore belongs to the clade. Even if assuming all the deaths after April 1st attributes to the clade, the case fatality rate is 0.1%.

The phylogeny analysis is shown in Figure 5. P13L, A4489V, Y789Y and T2016K emerged suddenly among Singapore and Australian samples. Some Singapore samples do not have S2015R variant as well as samples in Australia. Since L3606F/P13L/A4489V/T2016K clade is observed in Australia earlier, L3606F/P13L/A4489V/T2016K/S2015R clade might have evolved from an Australian strain.

Discussion

Low case fatality rate in Singapore might attribute to differences in sampling or false positives in RT-PCR tests[12]. However, country like Korea is also performing comprehensive testing and still difference in case fatality rates is clear. Also, underreporting deaths can be another possibility.

Ethnic background can vary in Singapore and many migrant workers are infected.

Smoking habit is another factor needs to be taken into an account. Although Singapore has strict smoking policy, male smoking rate is 28% in recent survey[13]. Therefore, smoking habits play little role in low case fatality rate in Singapore.

Sampling of genomes might not be uniform enough to represent entire country.

Contributions from each variant is not clear. Whether S2015R contributes to the low pathogenicity or not can be revealed by analyzing Australian cases of L3606F/P13L/A4489V/T2016K clade.

We have used case fatality is surrogate to pathogenicity in this study. However, Singapore has low serious/critical cases as well. Therefore, this clade is not limited to just low fatality rate but overall acute pathogenicity. Long term pathogenicity needs to be evaluated carefully.

Further investigation is necessary to validate the finding. Once it is validated, live vaccines can be developed from the strain and regions where the strain is prevalent, economic activities can be resumed.

List of Abbreviations

COVID-19	Coronavirus disease 2019
GISAID	Global Initiative on Sharing All Influenza Data
PCR	Polymerase Chain Reaction
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2

Funding

The authors received no specific funding for this work.

Acknowledgements

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu™ Database, Genbank, and NGDC Genome Warehouse, and the National Microbiology Data Center (NMDC) on which this research is based. The list of genomes is provided in Supplemental Table 1.

Supplemental Materials

Supplemental Table 1: Genomes Used in Analysis (supplemental_table_1.tsv)

References

1 He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, doi:10.1038/s41591-020-0869-5 (2020).

2 Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA*, doi:10.1001/jama.2020.1585 (2020).

3 Guan, W.-j. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*, doi:10.1056/NEJMoa2002032 (2020).

4 Wu, Z. & McGoogan, J. M. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* **323**, 1239-1242, doi:10.1001/jama.2020.2648 (2020).

5 Richardson, S. *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA*, doi:10.1001/jama.2020.6775 (2020).

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

7 Koyama T., P. D., Parida L. Variant analysis of COVID-19 genomes. *Bull World Health Organ.*, doi:http://dx.doi.org/10.2471/BLT.20.253591 (2020).

8 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443-453, doi:https://doi.org/10.1016/0022-2836(70)90057-4 (1970).

9 Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* **15**, e1006650, doi:10.1371/journal.pcbi.1006650 (2019).

10 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066, doi:10.1093/nar/gkf436 (2002).

11 Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160-174, doi:10.1007/BF02101694 (1985).

12 Htun, H. L. *et al.* Responding to the COVID-19 outbreak in Singapore: Staff Protection and Staff Temperature and Sickness Surveillance Systems. *Clinical Infectious Diseases*, doi:10.1093/cid/ciaa468 (2020).

13 Amul, G. G. H. & Pang, T. The State of Tobacco Control in ASEAN: Framing the Implementation of the FCTC from a Health Systems Perspective. *Asia & the Pacific Policy Studies* **5**, 47-64, doi:10.1002/app5.218 (2018).

Figure Legend

Figure 1. Plot of number of confirmed cases vs number of deaths by COVID-19 by each country. Dark red indicates high case fatality rate while yellow indicates low case fatality rate as seen in Singapore and Qatar.

Figure 2. Variant analysis of Singapore samples. There are various clades in Singapore. L3606F/P13L/A4489V/T2016K/S2015R forms the largest clades.

Figure 3. A. Histogram of L3606F/P13L/A4489V/T2016K/S2015R clade cases and all other cases in Singapore. The probabilities were created by density estimation with Gaussian kernel. B. Histogram of historical cases data in Singapore and estimate of number of cases of L3606F/P13L/A4489V/T2016K/S2015R clade cases and all other cases for each day between January 15th and May 13th, 2020.

Figure 4. Phylogeny of genomes in L3606F clade. Each genome was colored by country and variant statuses of P13L, A4489V, T2016K, S2015R were shown in the binary heatmap. L3606F/P13L/A4489V/T2016K/S2015R clade in Singapore forms a distinct cluster.

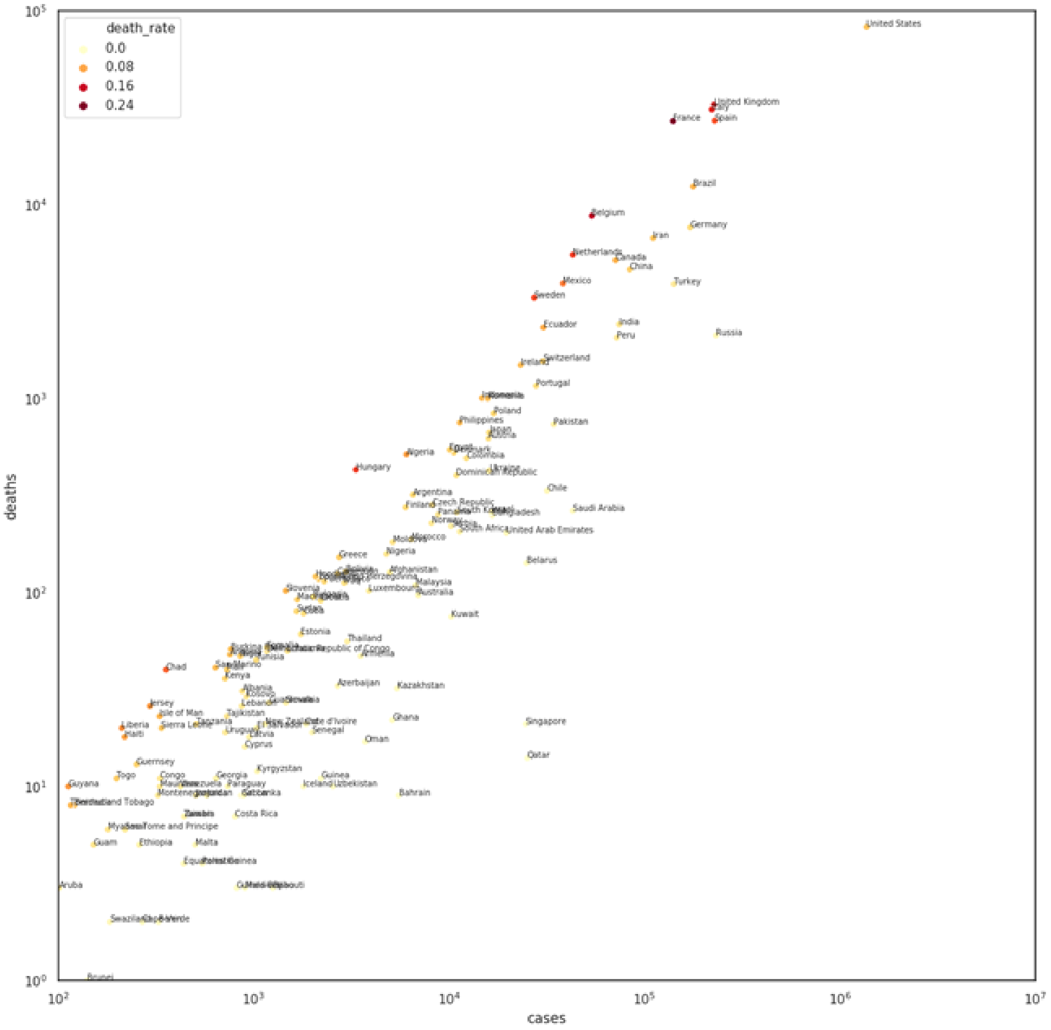

Figure *1*. Plot of number of confirmed cases vs number of deaths by COVID-19 by each country. Dark red indicates high case fatality rate while yellow indicates low case fatality rate as seen in Singapore and Qatar.

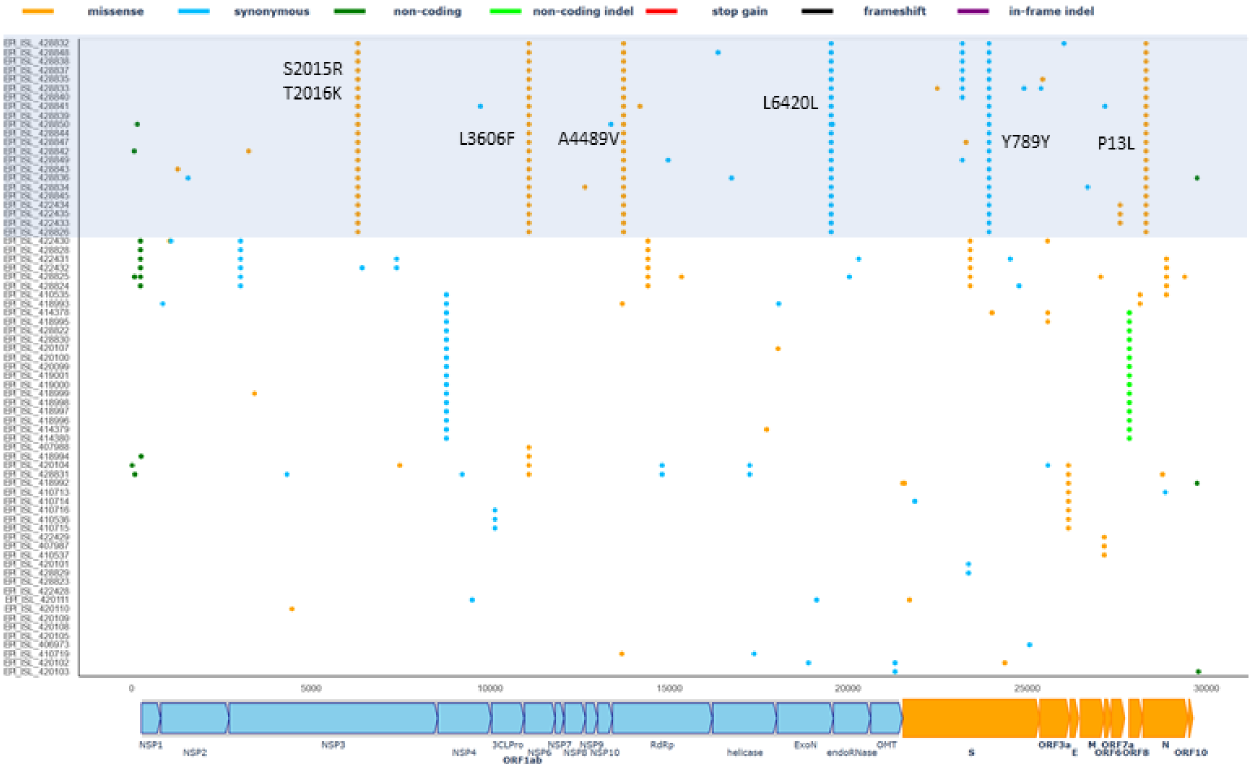

Figure *2*. Variant analysis of Singapore samples. There are various clades in Singapore. L3606F/P13L/A4489V/T2016K/S2015R forms the largest clades.

doi:10.20944/preprints202005.0466.v1

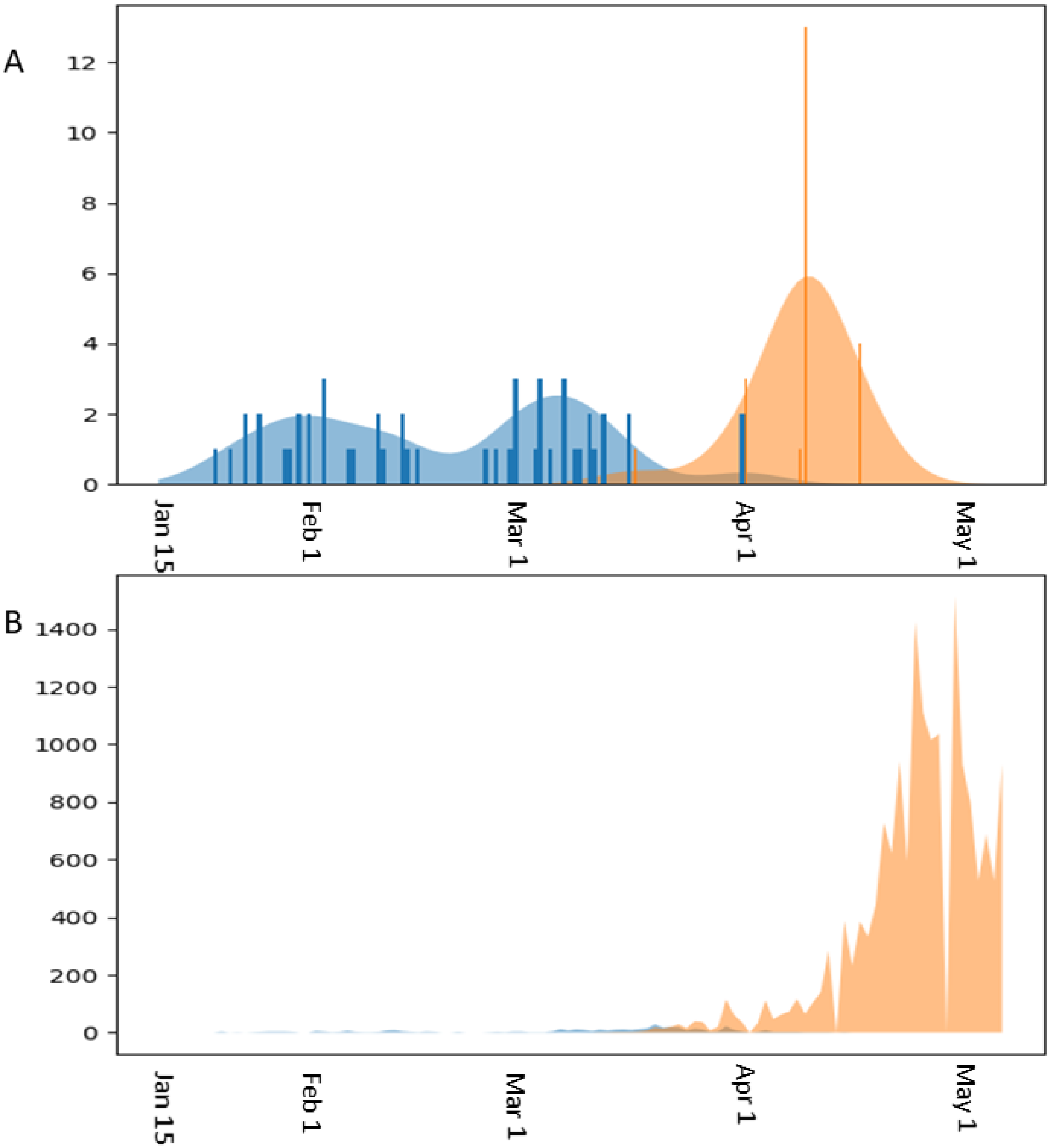

Figure *3*. A. Histogram of L3606F/P13L/A4489V/T2016K/S2015R clade cases and all other cases in Singapore. The probabilities were created by density estimation with Gaussian kernel. B. Histogram of historical cases data in Singapore and estimate of number of cases of L3606F/P13L/A4489V/T2016K/S2015R clade cases and all other cases for each day between January 15th and May 13th, 2020.

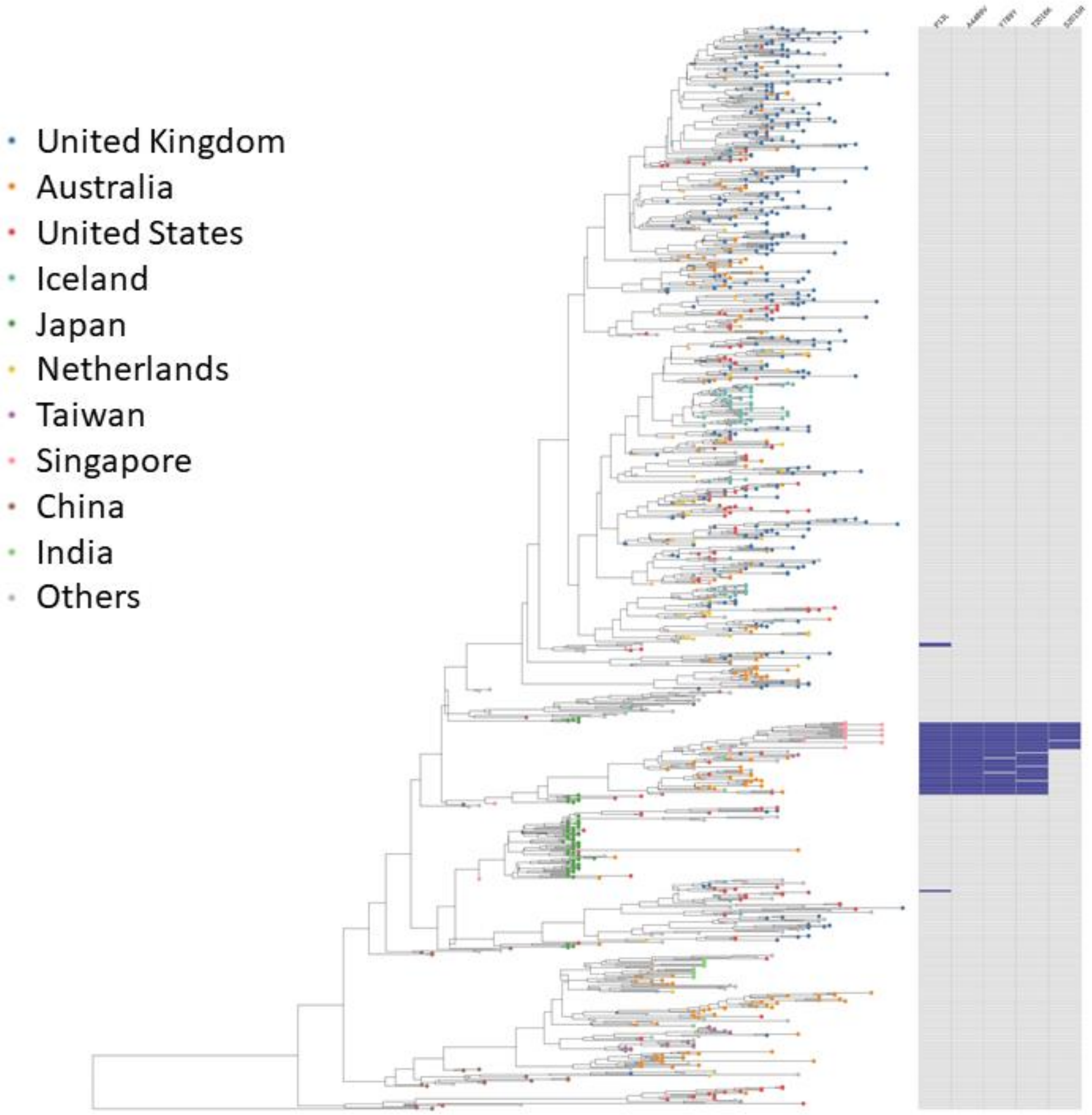

Figure *4*. Phylogeny of genomes in L3606F clade. Each genome was colored by country and variant statuses of P13L, A4489V, T2016K, S2015R were shown in the binary heatmap. L3606F/P13L/A4489V/T2016K/S2015R clade in Singapore forms a distinct cluster.