Genomic and epidemiological analysis of SARS-CoV-2 viruses in Sri Lanka

Chandima Jeewandara[1], Deshni Jayathilaka [1], Diyanath Ranasinghe[1], Nienyun Sharon Hsu[2,3], Dinuka Ariyaratne[1], Tibutius Thanesh Jayadas[1], Deshan Madushanka[1], Benjamin B. Lindsey[2], Laksiri Gomes[1], Matthew D. Parker[3], Ananda Wijewickrama[4], Malika Karunaratne[4], Graham S. Ogg[1,5], Thushan I. de Silva[2], Gathsaurie Neelika Malavige[1,5*]

1. University of Sri Jayawardenepura, Nugegoda, Sri Lanka
2. Department of Infection, Immunity and Cardiovascular Diseases, Medical School, University of Sheffield, Sheffield, UK.
3. Sheffield Bioinformatics Core, The University of Sheffield, Sheffield, UK
4. National Institute of Infectious Diseases, Angoda, Sri Lanka
5. University of Oxford, Oxford, United Kingdom

Correspondence should be addressed to:

Prof. Neelika Malavige DPhil (Oxon), FRCP (Lond), FRCPath (UK)

AICBU, Department of Immunology and Molecular Medicine, Faculty of Medical Sciences,

University of Sri Jayewardenepura, Sri Lanka

Tel +94 (0) 772443193; Fax: +94 (0) 112802026

Email: gathsaurie.malavige@ndm.ox.ac.uk

Abstract word count: 145

Text word count: 2251

Article summary line: Following repeated successful control of small outbreaks early in the pandemic caused by multiple lineages, the subsequent spread of a single lineage established in Sri Lanka (B.1.411) appears to be the cause of the large outbreak that started in early October 2020.

Running title: genomic sequencing of SARS-CoV-2 in Sri Lanka

Keywords: SARS-CoV-2, genomic sequencing, pangolin lineages, B.1.411, B.1.1.7, molecular epidemiology

Abstract

Since identification of the first Sri Lankan individual with the SARS-CoV-2 in early March 2020, small clusters that occurred were largely contained until the current extensive outbreak that started in early October 2020. In order to understand the molecular epidemiology of SARS-CoV-2 in Sri Lanka, we carried out genomic sequencing overlaid on available epidemiological data. The B.1.411 lineage was most prevalent, which was established in Sri Lanka and caused outbreaks throughout the country. The estimated time of the most recent common ancestor of this lineage was 10th August 2020 (95% lower and upper bounds 6th July to 7th September), suggesting cryptic transmission may have occurred, prior to a large epidemic starting in October 2020. Returning travellers were identified with infections caused by lineage B.1.258 , as well as the more transmissible B.1.1.7 lineage. Ongoing genomic surveillance in Sri Lanka is vital as vaccine roll-out increases.

Introduction

The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has emerged as the leading cause of mortality in several countries in the world. As of the 20th of February 2021, 111.3 million cases and 2.5 million deaths have been reported worldwide (1). Due to the emergence of variants of concern, the World Health Organization has recommended whole genomic sequencing of the SARS-CoV-2 viruses within countries regularly and systematically for early identification of such variants (2).

The first patient infected with SARS-CoV-2 in Sri Lanka was reported on the 27th January 2020, who was a foreign national, with the first Sri Lankan patient reported on the 10th of March 2020 (3). In the following six months (March to September), the spread of the virus was largely contained with only 3111 reported cases, of which 38.8% were imported (4). However, there was a surge in the number of cases with discovery of a new cluster in early October 2020 in a clothing factory in the district adjacent to Colombo (Gampaha). This was followed by rapid spread of SARS-CoV-2 within the Colombo Municipality region (CMC), fish markets and subsequently to the whole country. This outbreak continues to evolve, with infections now being reported in all regions of the country. As of 4th of February 2021 a total of 66,409 cases have been reported with 332 deaths (319 within the current large outbreak) (5). We carried out SARS-CoV-2 sequencing from isolates collected throughout the different phases of the pandemic in order to determine the molecular epidemiology of SARS-CoV-2 in Sri Lanka, including current

circulation of viruses with mutations that may confer greater transmissibility and/or threaten the efficacy of vaccines.

Methods

Patients

Whole genome sequencing was carried out on diagnostic sputum or nasopharyngeal swabs from 240 patients with COVID-19, where cycle threshold (Ct) values of the quantitative SARS-CoV-2 real-time PCR was <30. Ethical approval for the study was obtained by the Ethics Review Committee of the University of Sri Jayewardenepura.

Viral RNA Extraction

Viral RNA was extracted using QIAamp viral RNA mini kit (Qiagen, USA), SpinStarTM Viral Nucleic Acid Extraction kit 1.0 (ADT Biotech, Malaysia) or FastGene RNA Viral Kit, (Nippon Genetics, Germany) according to manufacturer's instructions. Presence of ORF1ab gene and N gene of SARS-CoV-2 was detected with Novel Coronavirus (2019-Ncov) Nucleic Acid Diagnostic Kit (Sansure Biotec) and S gene was detected by Taqpath COVID-19 RT PCR kit (Applied biosystems) by real time RT PCR in ABI 7500 real time PCR system (Applied Biosystems, USA).

Library Preparation and Next Generation sequencing (NGS)

Library preparation was attempted using either the TruSeq Stranded Total RNA Library Prep Gold (Illumina, San Diego, USA) or the The AmpliSeq for Illumina SARS-CoV-2 Community Panel, in combination with AmpliSeq for Illumina library prep, index, and accessories (Illumina, San Diego, USA). Shotgun metagenomic sequencing workflow was used for four initial

samples, while for the remainder (n= 236) a targeted RNA/cDNA amplicon assay was used (Supplementary methods).

Phylogenetic Analysis of SARS-CoV-2 sequences

Sequences with ≥10x median read depth and ≥86% genome coverage were taken forward for further analysis. GISAID accession numbers, location, and the clades of sequences used are shown in Supplementary table 2. 240 sequences from Sri Lanka were combined with 3200 representative global sequences. The representative sample of global sequences was obtained in two steps using all available data on GISAID up until 12th March 2021. The first step included randomly selecting one sequence per country per epi week. This was then followed by a random sampling of the remaining sequences to generate a sample of 3200 sequences. All sequences were then aligned to the SARS-CoV-2 reference strain MN908947.3 using MAFFT version 7.477 (Katoh_2002). We masked alignment positions that have been previously flagged as problematic (https://github.com/W-L/ProblematicSites_SARS-CoV-2) and manually removed obvious sequencing errors and potential homoplasic positions. A maximum likelihood tree was constructed using IQ-TREE2 version 2.1.2 (Minh_2020) with the GTR+G model of nucleotide substitution (Tavaré_1986, Yang_1994) and 1000 replicates of ultrafast bootstrapping (-B 1000) and SH-aLRT branch test (-alrt 1000).

Molecular clock phylogenetic analysis was undertaken using sequences from Sri Lanka. The alignment and maximum likelihood tree construction were performed using MAFFT and IQ-TREE2 as described above. TreeTime (6) was used to infer a molecular clock phylogeny using a strict evolutionary rate of $1.1 \times 10{-}3$ substitutions/site/year (estimated by Duchene et al.(7)) and a standard deviation of 0.00004. The tree was rerooted with least-squares criteria in TreeTime.

Eight samples were excluded from the analysis due to their inconsistent temporal signal. Lineages were assigned using Pangolin (version v2.3.8, lineages version 2021-04-23). Phylogenetic tree visualizations were produced using R (v3.5.3), R/ape, R/ggtree, R/ggplot2, R/ggtreeExtra, R/dplyr, R/phytools, R/tidytree. A proportional symbol map of Sri Lanka was plotted with GPS coordinates of the sampling locations of B.1.411 sequences using R (v4.0.1), R/maps, R/ggplot2, R/ggrepel, R/cowplot and R/dplyr. Each sampling location was indicated by a coloured bubble proportionate to the number of sequences sampled within. Colombo district was zoomed into a sub map (longitude: 79.80 -79.98, latitude: 6.80 - 6.98) in order to visualize the suburbs as Colombo had the highest sampling density.

Results

Of six samples collected in March 2020 from returning travellers and their contacts (period A, figure 1: four from Colombo district, two from Kalutara district), two belonged to lineage B.4, two to B 1.1, one to B.1 and one to B (Figure 2 and Supplementary table 1). During early April, SARS-CoV-2 spread within closed community clusters in the CMC region (period B, Figure 1). Two viruses from these clusters belonged to lineages B.4 and B.1 (Supplementary table 1). Period C in figure 1 was thought to be due to an outbreak initiated by the returning workforce from the Middle East. Sequence was obtained from only one virus, which belonged to lineage B. Again, due to detection of infected patients at the airport and mandatory quarantine of all individuals for at least 14 days, cases appeared not to spill over to the community. However, there was a sudden surge in the number of cases in mid-July in a drug rehabilitation centre (DRC), in the North Central Province (period D, figure 1). The origin of this outbreak was not

known and three lineage B.1 sequences obtained formed a cluster separate from earlier B.1 sequences from Sri Lanka. This outbreak was also subsequently controlled.

Sri Lanka then had a period of approximately two months during August and September, where no locally acquired infections were reported (3). In early October, an outbreak occurred in a clothing factory that heralded the ongoing wave of infections. This was soon followed by widespread transmission in many factories, fish markets island wide, in the highly populated CMC region in Colombo and subsequently to many areas in the country (period E, figure 1). We sequenced 228 samples taken during October 2020 to early March 2021, to determine how the outbreak evolved and establish if there were any new introductions to fuel ongoing infections.192 of these 228 viruses were classified into a novel lineage B.1.411, which appears to have arisen in Sri Lanka and successfully established the largest outbreak of SARS-CoV-2 in Sri Lanka to date. M666I mutation in NSP12 was the most frequent mutation in the B.1.411 lineage which was observed in all the sequences of this lineage (192/192). However, the predominant mutation on the S protein was the D614G, which was seen in 190 of B.1.411 samples followed by H1159Y, seen in 185 samples (Supplementary table 2). Importantly, one genome sequenced in mid-February had the E484K mutation along with the other B.1.411 defining mutations. The other mutations that were seen in this lineage are T116I in NSP2, L37F in NSP6, P323L and M666I in NSP12 and T205I in the N protein (Supplementary table 2). A phylogenetic tree of all sequences in our analysis is shown in Figure 2.

Based on TimeTree analysis (Figure 3), the estimated time of the most recent common ancestor (tMRCA) of the B.1.411 lineage was the 10th of August 2020 (95% lower and upper bounds 6th July to 7th September). The distribution of B.1.411 cases throughout the country is shown in

Figure 4. Twenty infections sampled between December 2020 and early March 2021 belonged to the B.1.1.7 lineage. The first of these was an imported infection on the 1st of February 2020, with a further 16 imported cases between January and March, followed by three from community samples with no known travel links.

Eleven imported infections sampled in Sri Lankan quarantine centres were classified into UK dominant B.1.258 (n=8) and B.1.1.25 lineages (n=1), a South American dominant B.1.420 lineage (n=1) and the B.1.351 lineage first described in South Africa (n=1) (Supplementary table 2). Importantly, the spike N439K mutation was seen in 4/11 of those from quarantine centres which is the first detection of this mutation in Sri Lanka. The ΔH69/V70 deletion co-occurred in three sequences with N439K and were classified as B.1.258 lineage viruses.

Discussion

We report the first description of SARS-CoV-2 molecular epidemiology in Sri Lanka from March 2020 to early March 2021. The virus strains identified in March 2020 belonged to clades B.1, B.2, B 1.1 and B.4, demonstrating that SARS-CoV-2 strains were introduced to Sri Lanka from multiple locations (8, 9).

Sri Lanka underwent a national lockdown very early in the pandemic on the 20th of March 2020, when only 66 patients with SARS-CoV-2 were confirmed. This lockdown, which continued until mid-May, managed to contain the outbreak and prevent community transmission, except within isolated community clusters. A further contained outbreak occurred in mid-July within a drug rehabilitation centre (DRC). Sequencing of a limited number of these samples showed that this outbreak was due to viruses belonging to lineage B.1 but which were distinct to the former B.1

samples. The outbreak in the DRC was also subsequently controlled and Sri Lanka did not report any cases of locally acquired infection during the months of August and September. Small numbers of reported cases were from imported infections only.

A large outbreak was abruptly discovered in early October after a clothing factory employee presented with pneumonia caused by SARS-CoV-2, which was followed by the emergence of a large second wave. We report that these were due to a lineage first described in samples from Sri Lanka, B.1.411, that dispersed throughout the country. A molecular clock analysis revealed that this lineage most likely emerged in August and therefore, it is possible that the virus was circulating in the community for several weeks before leading to the large outbreak that started in October. This highlights the potential for cryptic community transmission leading to a national epidemic wave even in the face of strict quarantine rules for returning travellers.

The, B.1.411 Sri Lankan lineage has a unique spike mutation H1159Y in the C terminal region, which was seen in 185/192 viruses belonging to this lineage. The significance of this mutation is unknown. Also, the P323L mutation in NSP12 region, which is known to have co-evolved with D614G mutation was seen in 172/192 of the B.1.411 genomes (10). Even though there is no direct correlation between P323L mutation and infectivity, given the fact that this mutation is widespread and almost 100% co-existent with D614G some argue that this mutation could contribute to the enhanced viral replication and infectivity seen in D614G dominant strains (10). Most importantly, one B.1.411 genome carrying the E484K mutation on the spike protein was detected from the community in mid-February, 2021, demonstrating the potential for this lineage to evolve mutations that may evade antibody responses. Even though E484K mutation is predominantly seen in B.1.351 and P.1 lineages, recent evidence indicates introduction of this mutation into other lineages such as B.1.1.7 and B.1.243 (11). The other more frequent mutations

were T166I in NSP2, L37F in NSP6, and T205I in N-protein. The L37F mutation in NSP6 is thought to render the NSP6 protein less stable and therefore, compromise the function of NSP6 (12). The other mutations have been frequently reported in many other SARS-CoV-2 lineages(13) while the other changes that were detected in the amino acids have not been associated with increased or reduced virulence.

Since the emergence of the 'second wave' of SARS-CoV-2 infections in early October 2020, all repatriation from overseas was stopped for a few months and subsequently, restarted in December 2020. Along with this, viruses of many lineages were identified within the quarantine centers where overseas returnees were housed. Importantly, the B.1.1.7 variant, which has been associated with higher transmissibility (14) was initially identified within these quarantine centres, but later from the community. Four viruses carried the N439K mutation in the receptor binding motif (RBM) in the S-protein which is known to enhance the binding affinity of the S-protein to human ACE2 receptor and the resistance against several neutralizing monoclonal antibodies (15). In addition, three of those genomes showed the S:H69/V70 mutation, which often co-occurs with the RBM with amino acid replacements such as N439K (16). This also is shown to associate with increased infectivity. None of these were identified within the community.

In summary, the viruses identified in March 2020, appear to be predominantly introduced by multiple sources such as from Europe and the Middle East and these strains were responsible for the subsequent outbreaks that were seen in Sri Lanka until July. The large ongoing outbreak that started in early October, appears to be due to spread of a single virus lineage, B.1.411 until the current data availability to end of March 2021, with new lineages introduced recently from

visitors from overseas. As SARS-CoV-2 vaccine rollout commences in Sri Lanka, ongoing genomic surveillance for variants of concern will be vital.

Acknowledgement

We thank Dr. Julian Villabonas-Arenas for sharing some scripts for viral sequence analysis. We are grateful to the World Health Organization, UK Medical Research Council and the Foreign and Commonwealth Office for support.

References

1. Medicine JHUa. Coronavrus Resource Centre. CRITICAL TRENDS: TRACKING CRITICAL DATA 2020 [cited 2020; Available from: https://coronavirus.jhu.edu/

2. Organization WH. Genomic sequencing of SARS-CoV-2. A guide to implementation for maximum impact on public health 8 January 2021: World Health Organization; 2021. p. 94.

3. Epidemiology unit MoH, Sri Lanka. COVID-19, National Epidemiological Report, Sri Lanka. 2020 [cited; Available from: https://www.epid.gov.lk/web/

4. Epidemiology unit MoH, Sri Lanka. Coronavirus disease 2019 (COVID-19) - Situation Report – 31.08.2020: Epidemiology Uni; 2020 31.08.2020.

5. Worldometer. COVID-19 CORONAVIRUS OUTBREAK. 2020 [cited; Available from:

6. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018 Jan;4(1):vex042.

7. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol. 2020 Jul;6(2):veaa061.

8. Pangolin. SARS-CoV2 lineages. 2020 [cited; Lineages assigned by the software tool pangolin]. Available from: https://cov-lineages.org/descriptions.html

9. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020 Jul 15.

10. Kannan SR, Spratt AN, Quinn TP, Heng X, Lorson CL, Sonnerborg A, et al. Infectivity of SARS-CoV-2: there Is Something More than D614G? J Neuroimmune Pharmacol. 2020 Dec;15(4):574-7.

11. Wise J. Covid-19: The E484K mutation and the risks it poses. BMJ (Clinical research ed. 2021 Feb 5;372:n359.

12. Rui Wang YH, Changchuan Yin, Guo-Wei Wei. Decoding asymptomatic COVID-19 infection and transmission. Preprint; 2020 July 2020.

13. Kaushal N, Gupta Y, Goyal M, Khaiboullina SF, Baranwal M, Verma SC. Mutational Frequencies of SARS-CoV-2 Genome during the Beginning Months of the Outbreak in USA. Pathogens. 2020 Jul 13;9(7).

14. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. medRxiv. 2021:2020.12.30.20249034.

15. Thomson EC, Rosen LE, Shepherd JG, Spreafico R, da Silva Filipe A, Wojcechowskyj JA, et al. Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. Cell. 2021 Jan 28.

16. Kemp S, Meng B, Ferriera I, Datir R, Harvey W, Collier D, et al. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion H69/V70. bioRxiv. 2021:2020.12.14.422555.

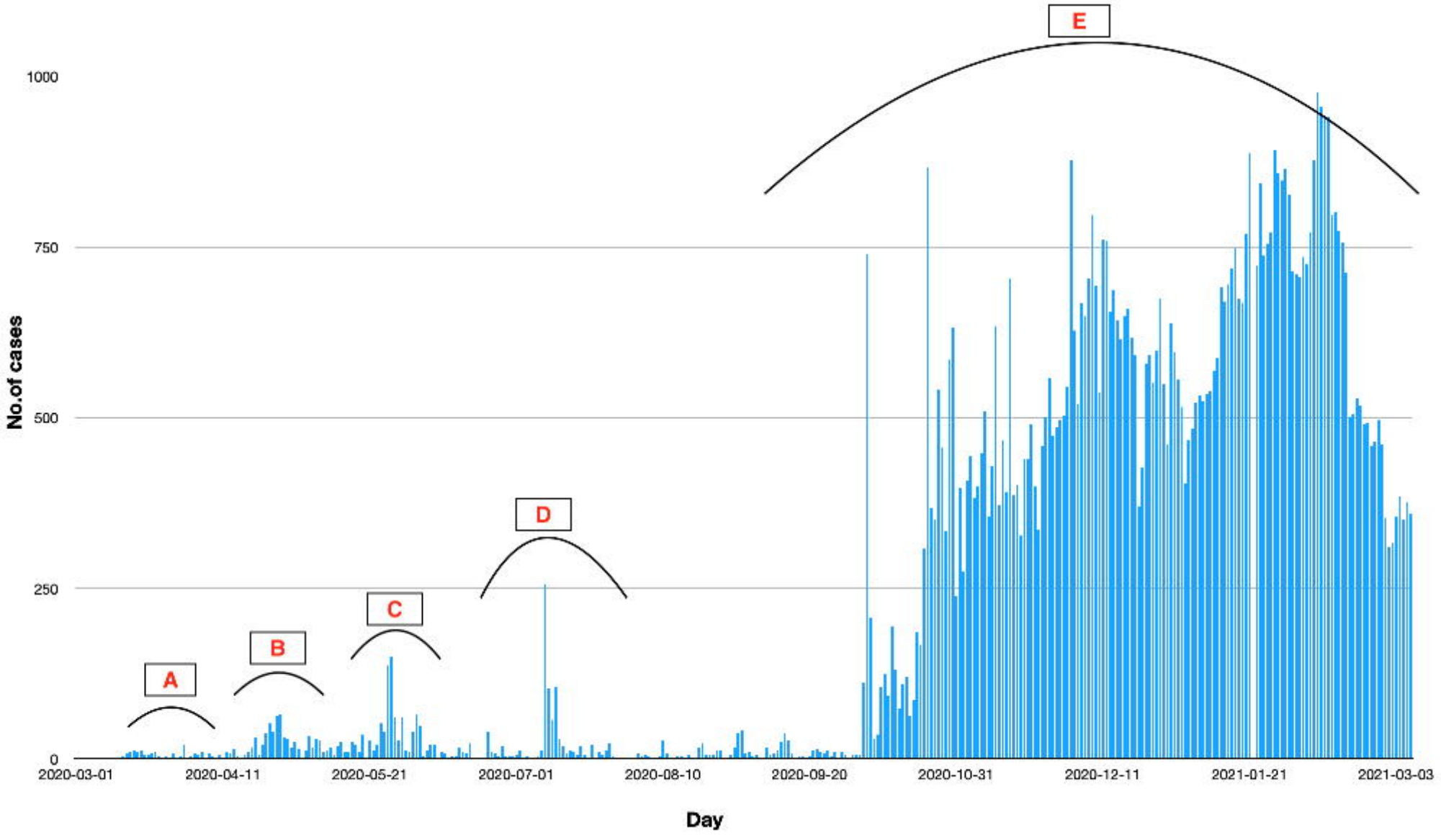
Figure legends

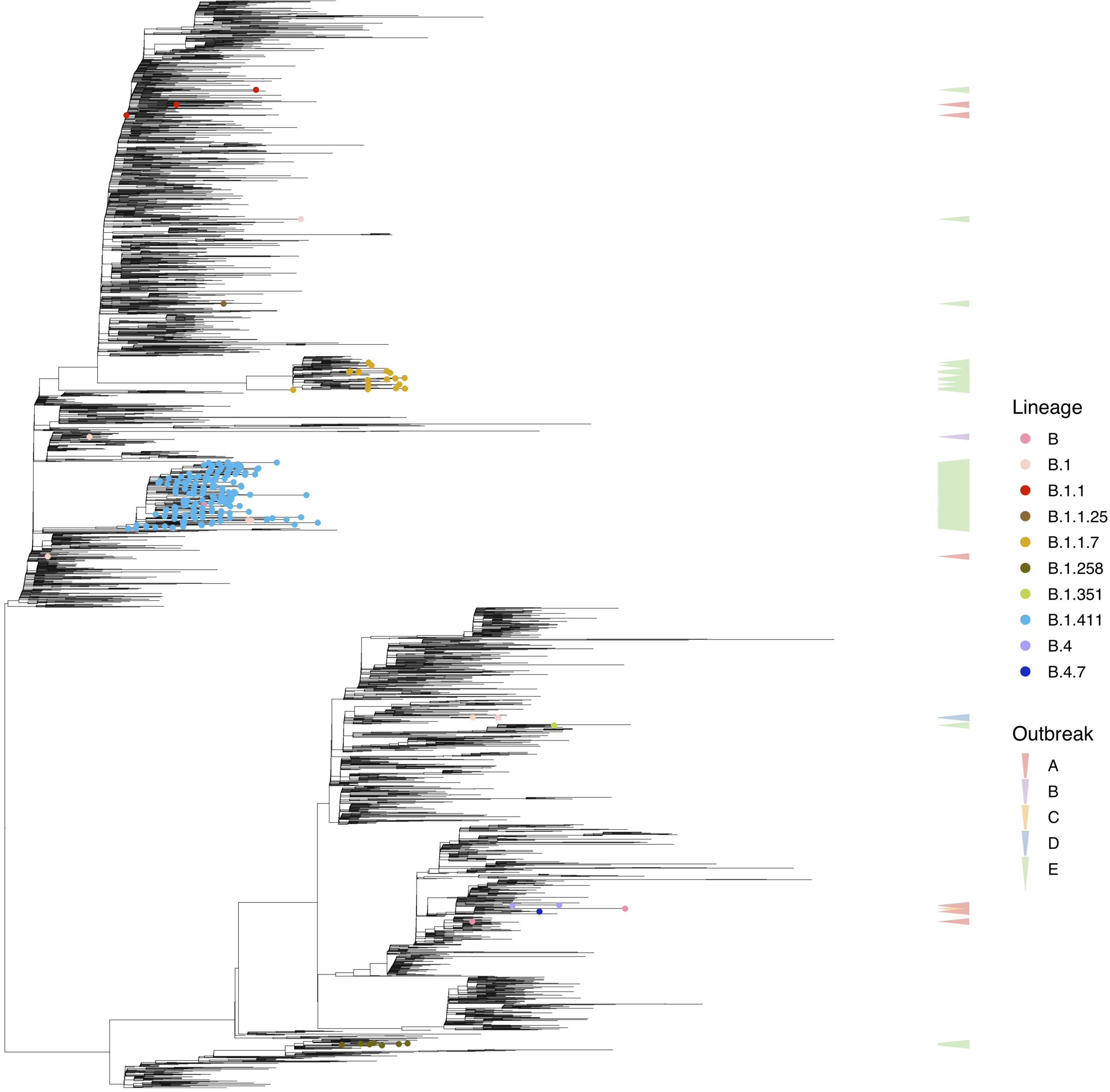
Figure 1: Epidemiological curve of COVID-19 patients reported in Sri Lanka from 10th March 2020 to 3rd March 2021 (Data from Epidemiology Unit, Ministry of Health, Sri Lanka)(3).

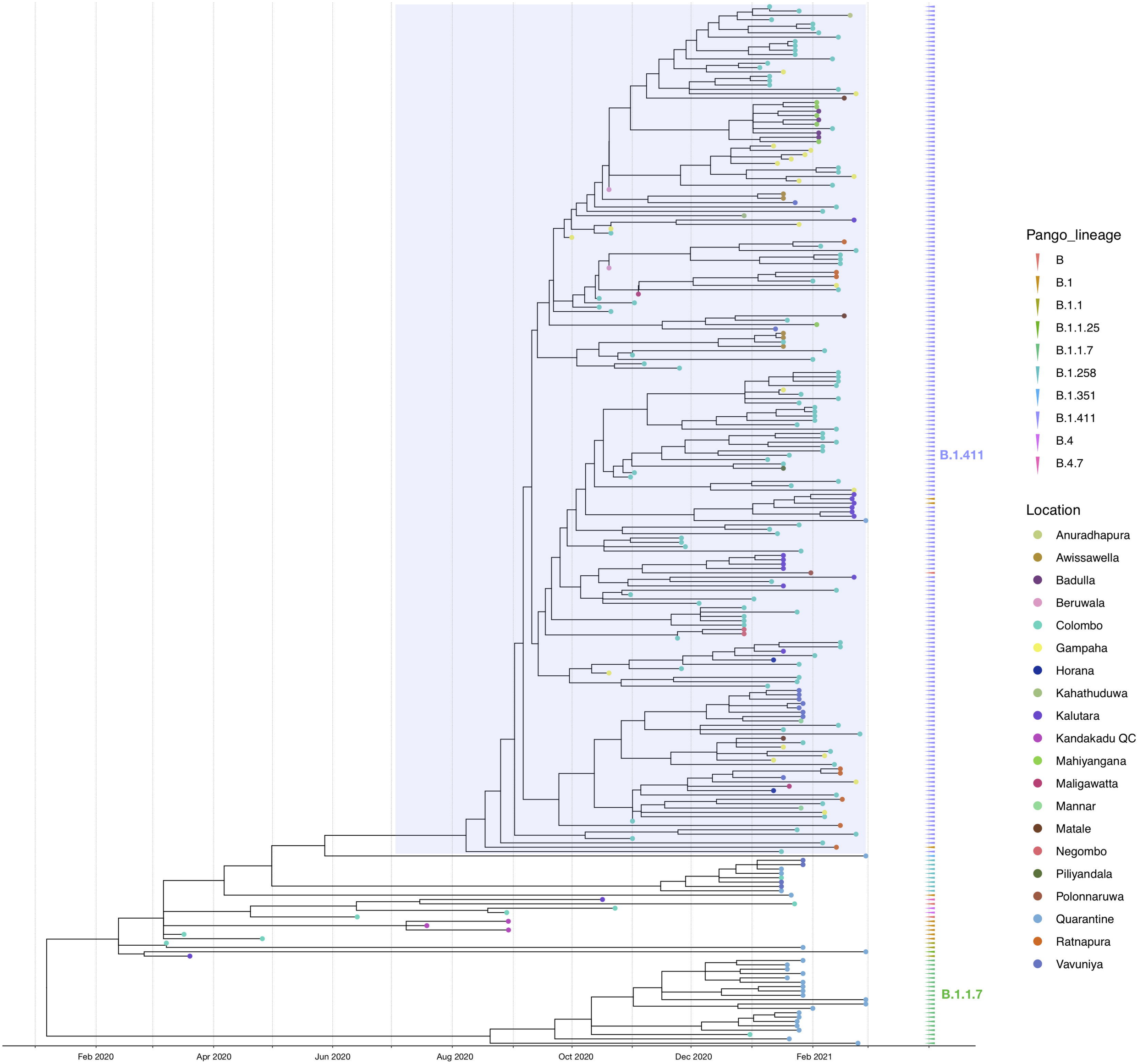
Figure 2. Phylogenetic analysis of Sri Lanka sequences. Maximum likelihood tree of 240 Sri Lanka samples and a representative sample of global sequences obtained from GISAID. Tips are coloured by Pango lineages, and the external layer on the right indicates the outbreak period associated with each Sri Lanka sample. The tree was estimated using IQTree2 (GTR maximum likelihood model and +G heterogeneity rate).

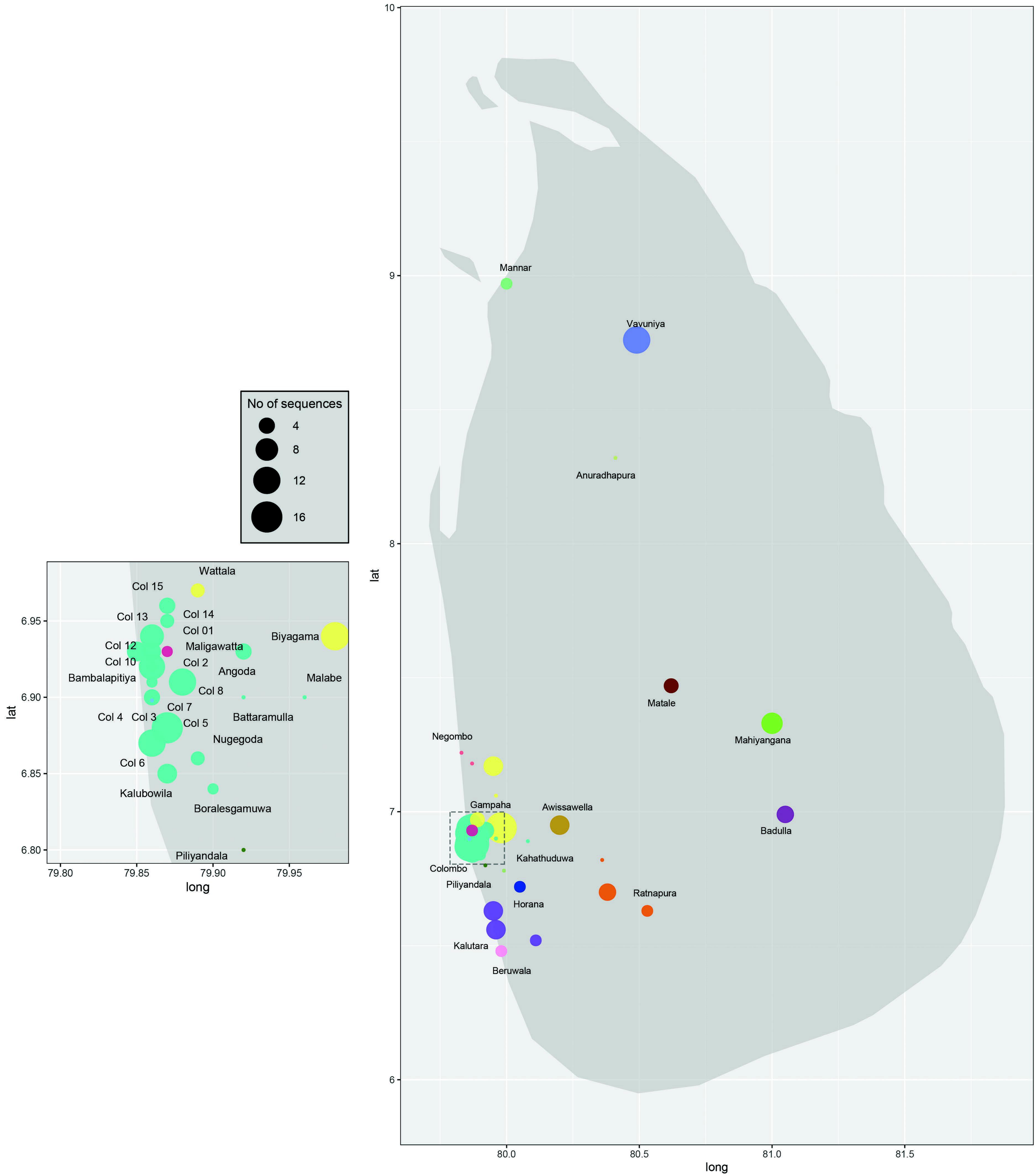
Figure 3. Molecular clock phylogeny of Sri Lanka samples. Tips are coloured by the sample location, and the external layer on the right shows the sample lineages. The molecular clock was inferred using TreeTime with an evolutionary rate of 1.1×10^{-3} substitutions/site/year and a standard deviation of 0.00004.

Figure 4. Geographical distribution of SARS-CoV-2 B.1.411 lineage infections in Sri Lanka sampled from October 2020 to March 2021. Radius of each bubble accounts for the number of sequences reported from its representing district and the colour code based on the location is adopted from the figure 3. Colombo district is expanded in on the left side panel to visualize its suburbs.


E
1000
750
500
250
0
No.of cases
D
C
B
A
2020-03-01
2020-04-11
2020-05-21
2020-07-01
2020-08-10
2020-09-20
2020-10-31
2020-12-11
2021-01-21
2021-03-03
Day


Lineage
B
B.1
B.1.1
B.1.1.25
B.1.1.7
B.1.258
B.1.351
B.1.411
B.4
B.4.7
Outbreak
A
B
C
D
E


Pango_lineage
B
B.1
B.1.1
B.1.1.25
B.1.1.7
B.1.258
B.1.351
B.1.411
B.4
B.4.7
Location
Anuradhapura
Awissawella
Badulla
Beruwala
Colombo
Gampaha
Horana
Kahathuduwa
Kalutara
Kandakadu QC
Mahiyangana
Maligawatta
Mannar
Matale
Negombo
Piliyandala
Polonnaruwa
Quarantine
Ratnapura
Vavuniya
B.1.411
B.1.1.7
Feb 2020
Apr 2020
Jun 2020
Aug 2020
Oct 2020
Dec 2020
Feb 2021


No of sequences
4
8
12
16
Wattala
Col 15
Col 13
Col 14
Col 01
Biyagama
Col 12
Maligawatta
Col 10
Col 2
Angoda
Bambalapitiya
Malabe
Col 8
Col 7
Col 4
Col 3
Col 5
Battaramulla
Nugegoda
Col 6
Kalubowila
Boralesgamuwa
Piliyandala
6.95
6.90
6.85
6.80
79.80
79.85
79.90
79.95
lat
long
Mannar
Vavuniya
Anuradhapura
Matale
Mahiyangana
Negombo
Gampaha
Awissawella
Badulla
Kahathuduwa
Colombo
Piliyandala
Ratnapura
Horana
Kalutara
Beruwala
10
9
8
7
6
80.0
80.5
81.0
81.5
lat
long