

Country-wide genomic surveillance of SARS-CoV-2 strains

Kishan Kalia¹, Gayatri Saberwal¹, and Gaurav Sharma^{*1}

Kishan Kalia¹ keishna.kalia@gmail.com +91 82648 08349

Gayatri Saberwal, Ph.D.¹ gayatri@ibab.ac.in +91 87629 18331

Gaurav Sharma, Ph.D.^{*1} gauravsharma@ibab.ac.in
gaurav.amit30@gmail.com +91 76687 00190

*Corresponding author

¹ Institute of Bioinformatics and Applied Biotechnology (IBAB), Bengaluru, Karnataka, India

16 **Abstract:**

17 Genomic surveillance has enabled the identification of several SARS-CoV-2 variants, allowing the
18 formulation of appropriate public health policies. However, surveillance could be made more
19 effective. We have determined that the time taken from strain collection to genome submission for
20 over 1.7 million SARS-CoV-2 strains available at GISAID. We find that strain-wise, time lag in this
21 process ranges from one day to over a year. Country-wise, the UK has taken a median of 16 days (for
22 417,287 genomes), India took 57 days (for 15,614 genomes), whereas Qatar spent 289 days (for
23 2298 genomes). We strongly emphasize that along with increasing the number of genomes of
24 COVID-19 positive cases sequenced, their accelerated submission to GISAID should also be
25 strongly encouraged and facilitated. This will enable researchers across the globe to track the
26 spreading of variants in a timely manner; analyse their biology, epidemiology, and re-emerging
27 infections; and define effective public health policies.

28 **Introduction**

29 Genomic surveillance of the evolving SARS-CoV-2 strains is an important tool to help control the
30 raging pandemic¹. For efficient surveillance, the first major requirement is the availability of all
31 sequenced genomes on an open-access platform that is accessible by researchers worldwide, to
32 enable them to analyze how this virus is evolving and spreading. Therefore, soon after researchers
33 became aware of COVID-19, towards the end of 2019, they started depositing the sequenced
34 genomes to the Global Initiative on Sharing All Influenza Data (GISAID), a pre-existing platform for
35 all influenza viruses. As of now, GISAID is the largest open-access portal, hosting the genome
36 sequence and related epidemiological and clinical data of more than 1.7 million SARS-CoV-2
37 strains. In a mere 1.5 years, this virus has become one of the most studied organisms in history, with
38 GISAID playing a major enabling role. Thanks to ongoing genomic surveillance using this data,
39 several new variants such as B.1.1.7 (Alpha; first in the UK); B.1.351 (Beta; first in South Africa);
40 B.1.1.28 (Gamma; P.1, first in Brazil); B.1.617.2 (Delta; first in India); B.1.617.1 (Kappa; first in
41 India); P.3 (Theta; first in the Philippines); and B.1.427 and B.1.429 (Epsilon; first in the USA) have
42 been identified²⁻⁵. This information has been used worldwide to update public health policies to
43 control COVID-19 infections^{6,7}.

44 Considering the benefits of genomic surveillance^{6,8}, scientists have pressured countries to increase
45 their sequencing capacity, and this has led to several initiatives such as [COG-UK](#), [INSA-COG](#)
46 (India), [NGS-SA](#) (South Africa), [SPHERES](#) (USA), etc. However, besides increasing the fraction of
47 samples sequenced, there is another issue that scientists need to be concerned about i.e. “How soon
48 are the sequences being submitted to GISAID or any other open access platform?” Rapid submission
49 will enable the international community to analyse the variants emerging around the world quickly
50 and provide actionable information to governments.

51 **Methods:**

52 Using the latest data (as of 27 May 2021) available at GISAID, we have calculated the Collection to
53 Submission Time Lag (CSTlag) per strain. We have also calculated the median and average CSTlag
54 time for each country and continent (category 1: for all countries and category 2: for all those
55 countries who have submitted over 1000 genomes). Country population and total COVID-19 cases
56 data were obtained from Worldometer on June 02, 2021, 17:32 GMT. Based on these information,
57 we have also calculated the rate of genome sequencing normalized with total COVID-19 cases and
58 one million population per country respectively.

59 **Results:**

60 Our statistical analysis (Figure 1 and S1; Table S1 and S2) for 1,718,035 SARS-CoV-2 strains
61 submitted to GISAID has determined that the Collection to Submission Time Lag (CSTlag) per
62 strain ranges from 1 day to over a year. Examining the median CSTlag values for countries that have
63 sequenced >1000 SARS-CoV-2 genomes, we note that the CSTlag from the UK is the shortest i.e.,
64 16 days for over 417,000 genomes. For the rest of Europe, the lag is 25 days for over 590,000
65 genomes. The USA has spent almost 26 days for over 498,000 genomes, whereas for Canada it is 88
66 days for over 44,000 genomes. Amongst the Oceania countries, the CSTlag for New Zealand is 40
67 days for over 1000 genomes, whereas for Australia, it is 51 days for over 17,000 genomes. In Asia,
68 the median lag is 72 days, for over 89,000 genomes, with Singapore having the shortest lag of 26
69 days for 2405 genomes, and Qatar the longest lag of 289 days for 2298 genomes. India's CSTlag is
70 57 days for 15,614 genomes whereas Japan, which has sequenced the most genomes in Asia, has
71 taken 79 days for over 37,000 genomes. For South American countries, the median lag is 61 days for
72 over 18,000 genomes, whereas countries in Africa have taken 50 days for over 7000 genomes (Table
73 S2).

74 Coming to the rate of sequencing, top-performing countries Iceland, Australia, New Zealand, and
75 Denmark have sequenced approximately 77%, 59%, 39%, and 35% of their positive cases
76 respectively (Table S1). The USA and UK have sequenced over 400,000 genomes each, which is
77 9.2% and 1.5% of their respective positive cases. India, being the second-largest country based on
78 both total population and known COVID-19 cases, has sequenced a mere 0.05% of the reported
79 cases. On average, African, Asian, and South American countries have sequenced a mere 0.36%,
80 0.21%, and 0.07% of their total COVID-19 cases, whereas this number is 1.9%, 1.4%, and 37% of
81 European, North American, and Oceania countries. Population-wise, most of the European countries,
82 the USA, Israel (Asia) and Reunion (Africa) have sequenced samples from over 1000 people per
83 million population (ppmp). Amongst countries with over 100 million population, including Brazil
84 (50 ppmp), India (11 ppmp), Indonesia (6 ppmp), Nigeria (4 ppmp), and Pakistan (1 ppmp), only the
85 USA (1,497 ppmp) and Japan (297 ppmp) have sequenced over 100 ppmp. Cumulatively, African,
86 Asian, and South American countries have carried out the genome sequencing of only 14, 21, and 49
87 ppmp, whereas this number is 1198, 948, and 607 ppmp for European, North American, and Oceania
88 countries (Table S2).

89 **Discussion:**

90 In terms of the delay in sequence submissions, there may be several reasons for this. The speed of
91 sequence submission to GISAID is based on (i) the time taken from sample collection from a
92 COVID-19 patient to RNA isolation in the lab and its dispatch to the sequencing centre and (ii) the
93 time from RNA sample arrival at the sequencing centre to the uploading of the sequence. Countries
94 with a shorter median CSTlag are more likely to have strong public health systems allowing efficient
95 sample and metadata collection, and smoother coordination between the sample collection centre, the
96 RNA isolation lab, and the sequencing lab. Countries without such a strong system would be at a
97 disadvantage and may face additional logistical problems in sample/metadata collection and shipping
98 because of lockdown-related restrictions. Several countries might have a shortage of labs that can
99 handle COVID-19 samples, or might have an overly centralized system, wherein only a few labs are
100 authorized to handle such samples, causing a delay in sequencing and submission. A paucity of funds
101 or restrictions on importing reagents and equipments would also add to the delay. The use of older
102 sequencing technologies that are low-throughput and more expensive per sample would complicate
103 matters further.

104 Most of the countries with a short CSTlag are industrialized nations that are likely to have strong
105 linkages between the clinical and scientific establishments. This is not always so for other countries.
106 Many countries with a longer CSTlag have a less developed public health system. They might have
107 had to establish novel collaborations and institutional arrangements to help deal with the pandemic.
108 All of this would have taken time, which would have impacted work on the ground. Some of the
109 possible causes for delay listed above are known to have been true in India, for instance, and are
110 being resolved^{9,10}.

111 Sometimes, even after rapid sequencing, genomes may not be promptly uploaded to GISAID, and
112 there may be several reasons for that. First, the importance of genomic surveillance may not have
113 been well understood, especially in the early months of the pandemic. Second, there may be a wish
114 to withhold information, in order to publish or patent first. Third, several governments may be
115 sensitive to the issue of virulent strains, in particular, being named after their countries. The WHO
116 initiative of renaming variants with Greek letters may help in resolving this issue⁵. Finally, in many
117 countries, there may be significant bureaucratization or political interference at various steps from
118 sample collection to uploading sequences to GISAID, which adds to the delay. Although one does
119 not know the extent of various problems in each country, it is likely that far more samples have been
120 sequenced than are represented in GISAID.

121 In countries with a longer CSTlag, the sequenced variants may have enough time to establish
122 themselves across the region, or – based on a significant mutation rate¹¹ – may evolve into
123 completely new variants, if quick tracking, tracing, and actions to stop transmission are not
124 undertaken. Therefore, this issue must receive urgent attention. All bottlenecks that prevent a lower
125 CSTlag must be addressed.

126 Overall, an effective genomic surveillance system requires not only sequencing a significant fraction
127 of strains from COVID-19 patients, but also rapid genome submission to open access platforms like
128 GISAID. This will enable researchers across the globe to track the evolved variants, their mutations,
129 epidemiology, and biological consequences, which will provide crucial inputs for appropriate and
130 effective public health policies.

131 **Figure Legends:**

132 **Figure 1:** Violin plot illustrating the CSTlag values for the 54 countries that have sequenced over
133 1000 genomes. The box plot inside the violin plot depicts the median CSTlag per country. Outlier
134 CSTlag entries per country are not shown in this illustration. Each country's name is color-coded
135 according to the continent. We have also mapped the relative distribution of the number of genome
136 sequences submitted by each country as a bar plot.

137 -----

138 **Supplementary Data:**

139 **Table S1:** The population of each country and the country-wise distribution of (i) total COVID-19
140 cases, (ii) genomes submitted to GISAID, (iii) rate of genome sequencing normalized with COVID-
141 19 cases, (iv) rate of genome sequencing normalized with one million population (v) average CSTlag
142 and (vi) median CSTlag values.

143 **Table S2: Continent-wise statistics.** (i) Total population of the respective continent, (ii) Total
144 COVID-19 cases reported in the respective continent, (iii) Total genomes submitted to GISAID from
145 the respective continent, (iv) rate of genome sequencing in the continent normalized with COVID-19
146 cases, (v) rate of genome sequencing in the continent normalized with one million population, (vi)
147 average CSTlag for all strains per continent, and (vii) median CSTlag values for all strains per
148 continent.

149 **Figure S1:** Violin plot illustrating the CSTlag values for the 54 countries that have sequenced over
150 1000 genomes. The box plot inside the violin plot depicts the median CSTlag per country. All
151 CSTlag entries per country are shown in this illustration. Each country's name is color-coded
152 according to the continent. We have also mapped the relative distribution of the number of genome
153 sequences submitted by each country as a bar plot.

154 -----

155 **Funding:**

156 Gaurav Sharma, Ph.D. is supported by funding from the Department of Science and Technology-
157 INSPIRE (DST-INSPIRE) program, Government of India. This work was partially supported by the
158 Department of Electronics, IT, BT, and S&T of the Government of Karnataka, India.

159

160 **Views:**

161 The views expressed in this letter are those of the authors, and not necessarily those of either funding
162 agency or any other institution.

163

164 **Conflict of interest:**

165 The authors declare that they have no conflicts of interest.

166

167 **References:**

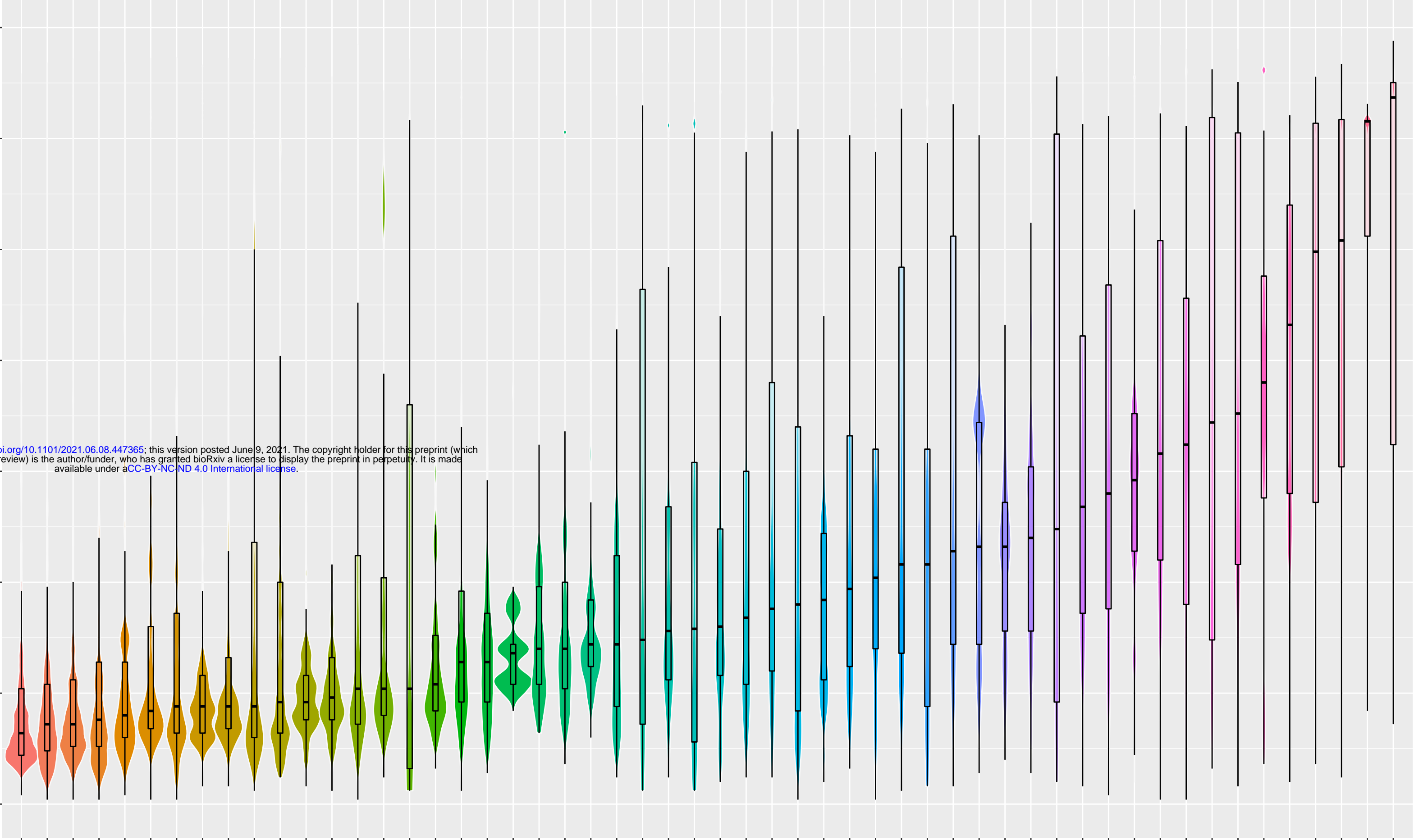
- 168 1. Lo, S. W. & Jamrozny, D. *Nat. Rev. Microbiol.* **18**, 478 (2020).
169 2. The Lancet. *The Lancet* vol. 397 445 (2021).
170 3. Burki, T. *Lancet (London, England)* **397**, 462 (2021).
171 4. Abdool Karim, S. S. & de Oliveira, T. *N. Engl. J. Med.* NEJMc2100362 (2021)
172 doi:10.1056/NEJMc2100362.
173 5. WHO. *WHO Activities* <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/> (2021).
174 6. Cyranoski, D. *Nature* vol. 589 337–338 (2021).
175 7. Grubaugh, N. D., Hodcroft, E. B., Fauver, J. R., Phelan, A. L. & Cevik, M. *Cell* **184**, 1127–1132
176 (2021).
177 8. Editorial. *Nat. Biotechnol.* **39**, 527–527 (2021).
178 9. Office of the Principal Scientific Advisor, Government of India.
179 <https://sites.google.com/view/corona-appeal/home> (2021).
180 10. Agrawal, A. *Nature* **594**, 9 (2021).
181 11. Ascoli, C. A. *Nature Biotechnology* vol. 39 274–275 (2021).

bioRxiv preprint doi: <https://doi.org/10.1101/2021.06.08.447365>; this version posted June 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Compact area

Collection to Submission Time Lag (CSTlag)

525
150
125
100
75
50
25
0



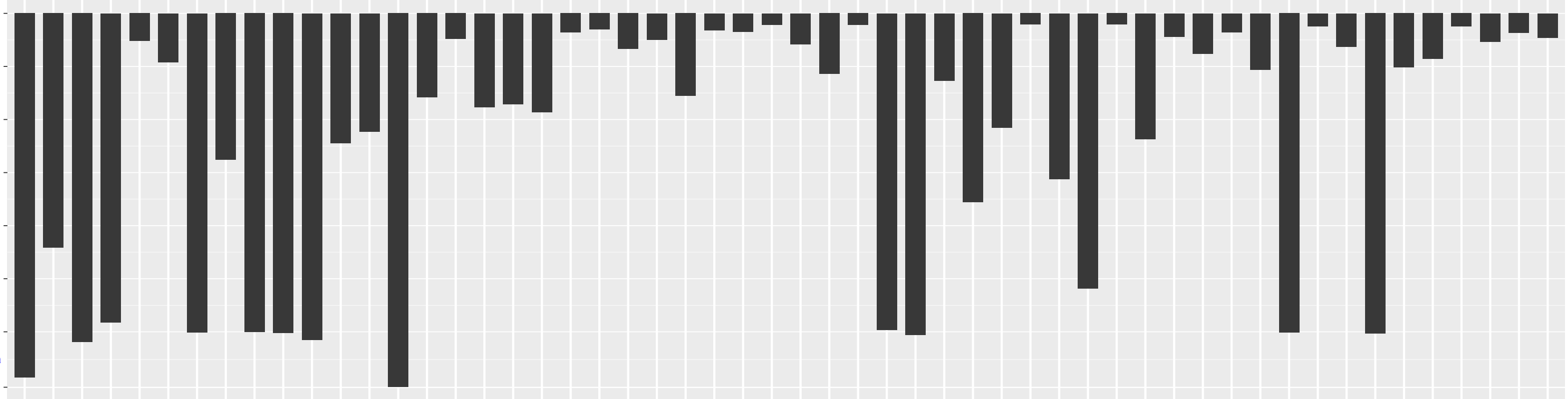
Continent Name

- Europe
- Asia
- North America
- South America
- Oceania
- Africa

United Kingdom
Belgium
Germany
Italy
Slovakia
Turkey
France
Ireland
Netherlands
Switzerland
Denmark
Poland
Slovenia
USA
Portugal
Singapore
Lithuania
Mexico
Norway
Estonia
Croatia
Latvia
Bulgaria
Luxembourg
Bangladesh
Greece
New Zealand
Czech Republic
South Korea
Colombia
Spain
Sweden
South Africa
Australia
Brazil
Thailand
India
Austria
Reunion
Israel
Chile
Russia
Indonesia
Finland
Japan
China
Argentina
Canada
Iceland
Philippines
Peru
Hong Kong
United Arab Emirates
Qatar

Genomes submitted to GISAID (May 27, 2021)

0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000



Compact area