Preliminary report on SARS-CoV-2 Spike mutation T478K

Simone Di Giacomo[1,$], Daniele Mercatelli[1,$], Amir Rakhimov[1] and Federico M. Giorgi[1,*]

[1] Department of Pharmacy and Biotechnology, University of Bologna, Via Selmi 3, 40138, Bologna, Italy
[*] Correspondence: federico.giorgi@unibo.it; Tel.: +390512094521
[$] Equal contribution

Abstract

Several SARS-CoV-2 variants have emerged, posing a renewed threat to COVID-19 containment and to vaccine and drug efficacy. In this study, we analyzed more than 820,000 SARS-CoV-2 genomic sequences deposited up to March 26, 2021 and identified a novel T478K mutation located on the SARS-CoV-2 Spike protein. The mutation is structurally located in the region of interaction with human receptor ACE2 and was detected in 4,214 distinct cases. We show that T478K has appeared and risen in frequency since January 2021, predominantly in Mexico and North America, but we could also detect it in several European countries.

Keywords: SARS-CoV-2; COVID-19, Genomic Surveillance; Spike; T478K; S:T478K; Spike:T478K

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the etiological cause of Coronavirus Disease 19 (COVID-19) is responsible for the most severe pandemic outbreak of the current century [1]. Naturally, it is the object of unprecedented scientific scrutiny, with more than 800,000 SARS-CoV-2 genomic sequences having been generated and publicly shared since December 2019. This avalanche of data was made possible thanks to the efforts of thousands of contributing laboratories across the World, and collected by the GISAID initiative database [2]. This currently allows to run nearly real-time operations of genomic surveillance, by scrutinizing the evolution of the virus temporally and geographically [3]. In the first 26 months since the appearance of SARS-CoV-2, genomic surveillance has proven itself fundamental in tracking viral outbreaks [4] and in identifying potential new variants of clinical concern. One of these is the variant B.1.1.7 [5], characterized by 18 mutations over the reference genomic sequence (NCBI entry NC_045512.2, most notably a mutation A23063T, causing an aminoacidic change N501Y in the viral Spike protein interaction domain with human receptor Angiotensin Converting Enzyme 2 (ACE2) [6]. The interaction with ACE2, a surface protein expressed in human respiratory epithelial cells, is one of the key mechanisms for viral entry

in the host, and it is a molecular mechanism directly connected with host specificity, early transmissibility [7] and higher viral infectivity [8].

N501Y is only one of the 9 Spike mutations of variant B.1.1.7, also characterized by mutations in polyprotein ORF1a, proteins ORF8 and Nucleocapsid (N) [9]. Another mutation in the Spike protein, D614G, has arisen in early 2020 and is currently present in more than 90% of all circulating SARS-CoV-2s; this mutation is not located in the interaction domain with ACE2, but it has been associated with increased entry efficiency with human host cells [10]. Virtually all variants of concern contain mutations in the Spike protein, such as variant B.1.351 (Spike mutations K417N, E484K, N501Y, D614G and A701V) and variants B.1.427/B.1.429 (Spike mutations S13I, W152C, L452R and D614G). A recently published clinical study [11] has shown a decreased vaccine efficacy against the variant B.1.351, testifying the need to track and monitor all SARS-CoV-2 mutations, with a particular accent on those affecting the Spike protein sequence.

In this short Communication, we will show a report on a novel SARS-CoV-2 Spike mutation, T478K, which is also located at the interface of the Spike/ACE2 interaction, and it is worryingly rising in prevalence among SARS-CoV-2 sequences collected since the beginning of 2021.

Materials and Methods

We downloaded all publicly available SARS-CoV-2 genomic sequences from the GISAID database on March 26, 2021. This yielded 826,572 samples, annotated with features such as collection date, region of origin, age, and sex of the infected patient. Only viruses collected from human hosts were kept for further processing, discarding e.g., environmental samples or viruses obtained from other mammals. We compared all these sequences with the SARS-CoV-2 reference genome NC_045512.2, which provided 16,751,677 mutation events; these nucleotide mutations were then converted in corresponding cumulative effects on protein sequence using the Coronapp pipeline [12]. The 3D rendering of the location of S:T478K in the SARS-CoV-2 Spike / Human ACE2 complex was based on the crystal structure from [13], deposited in the Protein Data Bank [14] entry 6VW1. All statistical analysis, algorithms and plotting were implemented with the R software [15].

Results

In total, we could detect the Spike:T478K (S:T478K) mutation in 4,214 distinct patients. The majority of these mutations (3,640 samples, Figure 1A) are associated to variant B.1.1.222, first detected in Mexico; S:T478K is present in 65.0% of B.1.1.222 cases. The remaining S:T478K events are distributed in small numbers (N<100) in other lineages phylogenetically not derived from B.1.1.222, supporting the hypothesis that this mutation has arisen more than once in distinct events. S:T478K is also present in 70% of the B.1.486 lineage samples, however this clade has currently only 10 reported cases and is therefore of no concern yet.

S:T478K does not seem to be significantly associated with patient age (one way ANOVA test p>0.1, Figure 1B), nor with patient sex (Figure 1C). The geographic distribution of S:T478K (Figure 1D) shows a noticeable prevalence in Mexico, where it constitutes 38.1% (1,203 distinct cases) of all sequenced SARS-CoV-2 genomes. We could detect S:T478K mutations in 2,536 samples from the United States of America, totaling 1.3% of all genomes generated in the country. The S:T478K is therefore primarily present in North America, but it has been detected also in European countries such as Germany, Turkey and Switzerland (Figure 1E).

One of the reasons of concern about S:T478K is that it is rapidly growing over time, both in number of detected samples (Figure 2A) and in prevalence, calculated as number of cases over total number of sequenced genomes (Figure 2B). We detected this grow starting at the beginning of 2021, and S:T478K is, at the time of writing (March 26, 2021) characterizing almost 2.0% of all sequenced SARS-CoV-2. As a comparison, we show the growth observed for Spike mutations S:N501Y, which rose in November 2020 (Figure 2C), and S:D614G, which exponentially grew in frequency starting February 2020 (Figure 2D).

The location of S:T478K is within the interaction domain with the human receptor ACE2, roughly encompassing amino acids 350 to 550 of the SARS-CoV-2 Spike protein. In particular, the position of S:T478K is on the interface with ACE2, as shown by crystal structures of the complex (Figure 2E).

S:T478K is frequently co-occurring with three other Spike mutations located outside the canonical ACE2 interaction region. One is D614G (99.8% co-occurrence), one of the founding events of SARS-CoV-2 lineage B.*, currently the most diffused Worldwide. The other two are P681H and T732A, with 95.2% and 91.4% co-occurrence with S:T478K, respectively (Table 1). We could detect S:T478K in co-presence with other Spike mutations as well, but currently all at much lower frequencies (<5%). The Spike S:T478K mutation is frequently co-existing also with mutations in other proteins, such as

the diffused two-aa Nucleocapsid mutation N:RG203KR, and mutations in Non-Structural Proteins (NSPs) derived from the polyprotein encoded ORF1 (Open Reading Frame), which include for example the viral RNA-dependent RNA polymerase NSP12 (Table 2).

Discussion

In this short communication, we report the distribution of the Spike mutation S:T478K and its recent growth in prevalence in the SARS-CoV-2 population. While there is currently no report of association of this variant with clinical features, S:T478K's rapid growth may indicate an increased adaption of SARS-CoV-2 variants carrying it, particularly lineage B.1.1.222. The distribution of this mutation is higher in North America [16], but we could detect it also in several European countries. The location of S:T478K in the interaction complex with human ACE2 may affect the affinity with human cells and therefore influence viral infectivity. An *in silico* molecular dynamics study on the protein structure of Spike has predicted that the T478K mutation, substituting a non-charged amino acid (Threonine) with a positive one (Lysine) may significantly alter the electrostatic surface of the protein, and therefore the interaction with ACE2, drugs or antibodies [17], and that the effect can be increased if combined by other co-occurring Spike mutations (see Table 1). Another experiment showed that T478K and T478R mutants were enriched when SARS-CoV-2 viral cultures were tested against weak neutralizing antibodies [18], highlighting, at least *in vitro*, a possible genetic route the virus can follow to escape immune recognition. Everything considered, we believe that the continued genetical and clinical monitoring of S:T478K and other Spike mutations is of paramount importance to better understand COVID-19 and be able to better counteract its future developments.

	Number of cases	Percentage
S:D614G	4209	99.79
S:P681H	4017	95.23
S:T732A	3856	91.42
S:L5F	136	3.22
S:D111N	70	1.66
S:T95I	57	1.35
S:T716I	46	1.09
S:E154A	41	0.97
S:T20I	33	0.78

S:N211D	30	0.71

Table 1. Ten SARS-CoV-2 Spike mutations most frequently co-occurring with S:T478K. We report the number of genomes where both mutations are present, and the percentage over total number of samples where S:7478K is reported.

	Number of cases	Percentage
NSP12b:P314L	4140	98.15
NSP6:I49V	4021	95.33
NSP3:P141S	4006	94.97
NSP4:T492I	3996	94.74
N:RG203KR	3980	94.36
NSP9:T35I	3793	89.92
ORF8:L4P	598	14.18
NSP2:T44I	504	11.95
N:Q418H	424	10.05
N:V350F	255	6.05

Table 2. Ten SARS-CoV-2 Non-Spike mutations most frequently co-occurring with S:T478K. We report the number of genomes where both mutations are present, and the percentage over total number of samples where S:7478K is reported. NSP: Non-Structural Protein; N: Nucleocapsid protein; ORF: Open Reading Frame.

Figure Legends

Figure 1. **A**: Prevalence of Spike mutation T478K in PANGO lineages. The top 10 lineages are reported, sorted by number of S:T478K samples over total number of lineage samples. The discrete number of S:T478K is reported on top of each bar. **B**: Frequency of sequenced SARS-CoV-2 genomes carrying the S:T478K mutation, divided into 5-years age ranges. **C**: Pie chart showing the distribution of S:T478K by patient sex (for patients whose sex was not reported, we indicated "Unknown"). **D**: Number of S:T478K samples over total samples sequenced from each country. The 10 countries with higher frequency (in percentage) are shown. Discrete numbers of S:T478K are reported on top of each bar. **E**: Geographic global projection of S:T478K cases detected in each country. The color scale indicates number of SARS-CoV-2 genomes carrying the S:T478K mutation, in logarithm-10 scale.

Figure 2. **A**: Number of sequenced SARS-CoV-2 genomes carrying S:T478K mutation over time, measured weekly. **B**: Prevalence over time of S:T478K in the SARS-CoV-2 population, measured as the number of S:T478K genomes over the total number of sequenced genomes. **C**: Prevalence over time of S:N501Y in the SARS-CoV-2 population. **D**: Prevalence over time of S:D614G in the SARS-CoV-2 population. **E**: 3D representation of the SARS-CoV-2 Spike / Human ACE2 interacting complex, derived from the crystal structure from [13].

Supplementary Material: Supplementary file S1: the SARS-CoV-2 genome annotation coordinates, in GFF3 format, used in this study, and based on NCBI reference genome sequence NC_045512.2.

Author Contributions: Conceptualization, SDG and FMG. Funding acquisition, FMG. Writing – original draft preparation, FMG. Writing – review and editing, DM and SDG. Methodology, AR and FMG. Validation, DM, SDG and FMG. Software, AR and FMG. All authors contributed to the study and approved the final version of the manuscript.

Funding: This research was funded by the Italian Ministry of University and Research, Montalcini grant.

Data Availability Statement: All data supporting this study is available on the GISAID portal https://www.gisaid.org/

Acknowledgments: We are very grateful to the GISAID Initiative and all its data contributors, i.e. the Authors from the originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fontanet, A.; Autran, B.; Lina, B.; Kieny, M.P.; Karim, S.S.A.; Sridhar, D. SARS-CoV-2 Variants and Ending the COVID-19 Pandemic. *Lancet Lond. Engl.* **2021**, *397*, 952–954, doi:10.1016/S0140-6736(21)00370-6.

2. Shu, Y.; McCauley, J. GISAID: Global Initiative on Sharing All Influenza Data - from Vision to Reality. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **2017**, *22*, doi:10.2807/1560-7917.ES.2017.22.13.30494.

3. Mercatelli, D.; Giorgi, F.M. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Front. Microbiol.* **2020**, *11*, 1800, doi:10.3389/fmicb.2020.01800.

4. Mercatelli, D.; Holding, A.N.; Giorgi, F.M. Web Tools to Fight Pandemics: The COVID-19 Experience. *Brief. Bioinform.* **2021**, *22*, 690–700, doi:10.1093/bib/bbaa261.

5. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. *Nat. Microbiol.* **2020**, *5*, 1403–1407, doi:10.1038/s41564-020-0770-5.

6. Ortuso, F.; Mercatelli, D.; Guzzi, P.H.; Giorgi, F.M. Structural Genetics of Circulating Variants Affecting the SARS-CoV-2 Spike/Human ACE2 Complex. *J. Biomol. Struct. Dyn.* **2021**, 1–11, doi:10.1080/07391102.2021.1886175.

7. Leung, K.; Shum, M.H.; Leung, G.M.; Lam, T.T.; Wu, J.T. Early Transmissibility Assessment of the N501Y Mutant Strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **2021**, *26*, doi:10.2807/1560-7917.ES.2020.26.1.2002106.

8. Khan, A.; Zia, T.; Suleman, M.; Khan, T.; Ali, S.S.; Abbasi, A.A.; Mohammad, A.; Wei, D.-Q. Higher Infectivity of the SARS-CoV-2 New Variants Is Associated with K417N/T, E484K, and N501Y Mutants: An Insight from Structural Data. *J. Cell. Physiol.* **2021**, doi:10.1002/jcp.30367.

9. Galloway, S.E.; Paul, P.; MacCannell, D.R.; Johansson, M.A.; Brooks, J.T.; MacNeil, A.; Slayton, R.B.; Tong, S.; Silk, B.J.; Armstrong, G.L.; et al. Emergence of SARS-CoV-2 B.1.1.7 Lineage - United States, December 29, 2020-January 12, 2021. *MMWR Morb. Mortal. Wkly. Rep.* **2021**, *70*, 95–99, doi:10.15585/mmwr.mm7003e2.

10. Ozono, S.; Zhang, Y.; Ode, H.; Sano, K.; Tan, T.S.; Imai, K.; Miyoshi, K.; Kishigami, S.; Ueno, T.; Iwatani, Y.; et al. SARS-CoV-2 D614G Spike Mutation Increases Entry Efficiency with Enhanced ACE2-Binding Affinity. *Nat. Commun.* **2021**, *12*, 848, doi:10.1038/s41467-021-21118-2.

11. Madhi, S.A.; Baillie, V.; Cutland, C.L.; Voysey, M.; Koen, A.L.; Fairlie, L.; Padayachee, S.D.; Dheda, K.; Barnabas, S.L.; Bhorat, Q.E.; et al. Efficacy of the ChAdOx1 NCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* **2021**, doi:10.1056/NEJMoa2102214.

12. Mercatelli, D.; Triboli, L.; Fornasari, E.; Ray, F.; Giorgi, F.M. Coronapp: A Web Application to Annotate and Monitor SARS-CoV-2 Mutations. *J. Med. Virol.* **2020**, doi:10.1002/jmv.26678.

13. Shang, J.; Ye, G.; Shi, K.; Wan, Y.; Luo, C.; Aihara, H.; Geng, Q.; Auerbach, A.; Li, F. Structural Basis of Receptor Recognition by SARS-CoV-2. *Nature* **2020**, *581*, 221–224, doi:10.1038/s41586-020-2179-y.

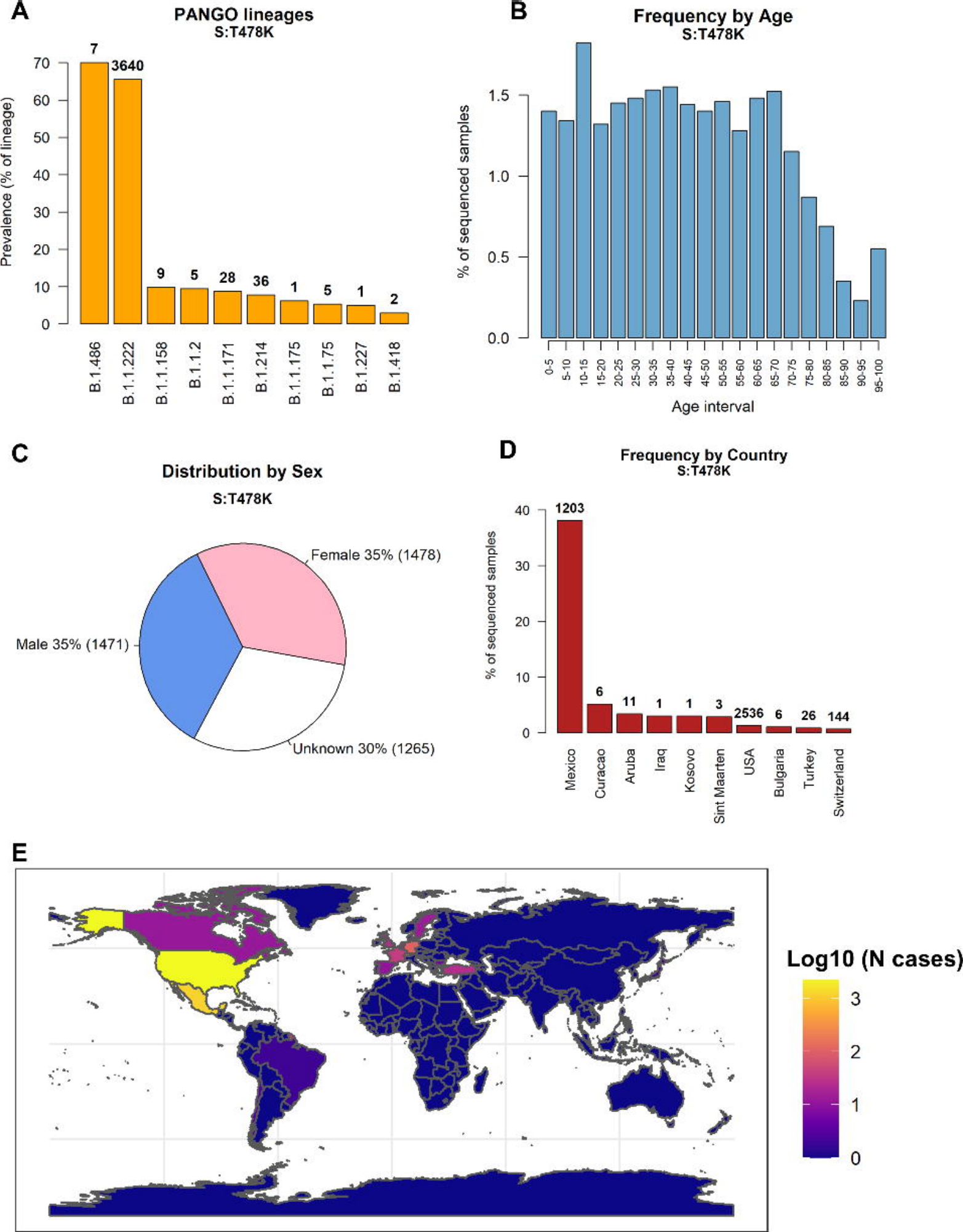
14. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242, doi:10.1093/nar/28.1.235.

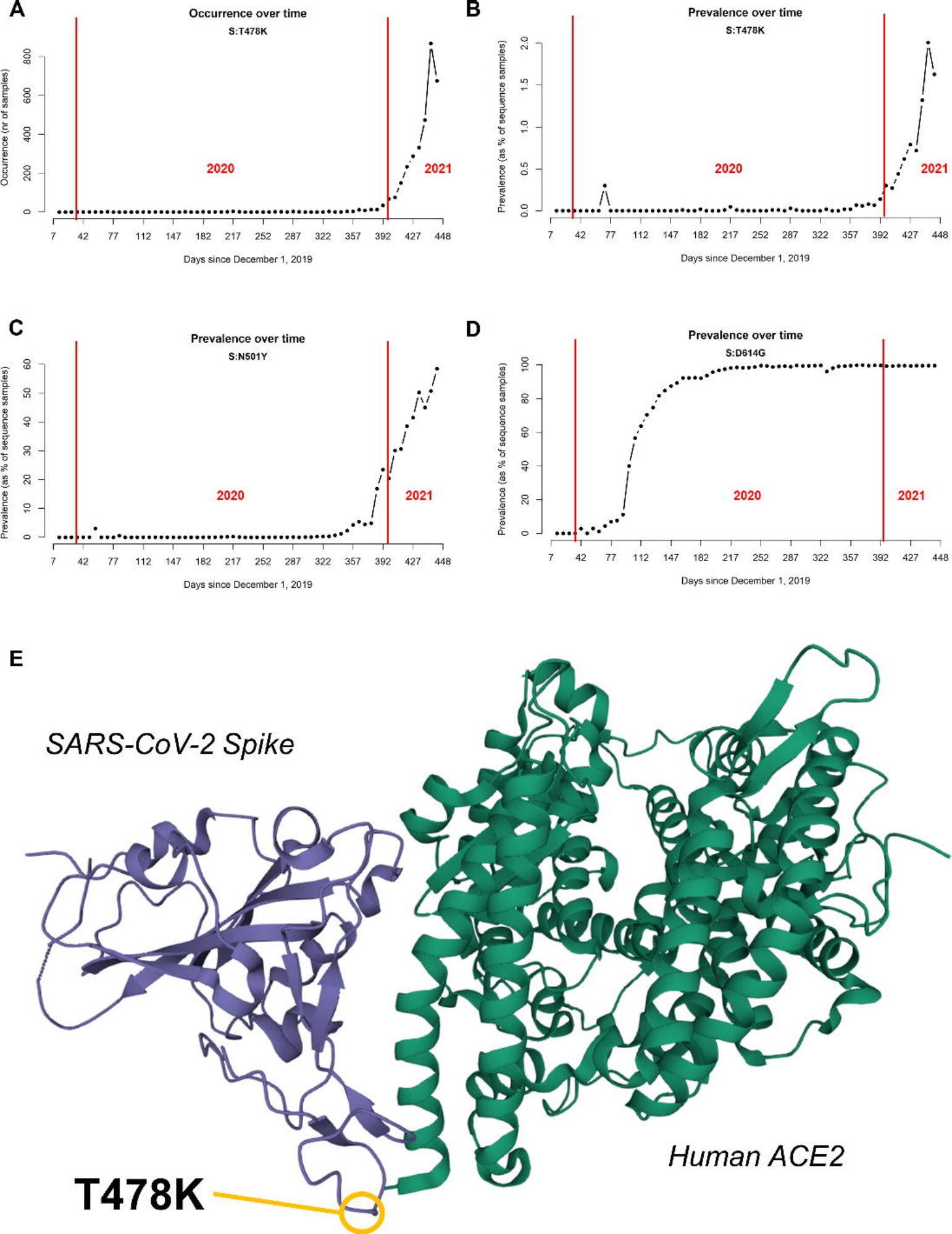
15. Crawley, M.J. *The R Book*; John Wiley & Sons, 2012; ISBN 978-1-118-44896-0.

16. Lasek-Nesselquist, E.; Pata, J.; Schneider, E.; George, K.S. A Tale of Three SARS-CoV-2 Variants with Independently Acquired P681H Mutations in New York State. *medRxiv* **2021**, 2021.03.10.21253285, doi:10.1101/2021.03.10.21253285.

17. Rezaei, S.; Sefidbakht, Y.; Uskoković, V. Comparative Molecular Dynamics Study of the Receptor-Binding Domains in SARS-CoV-2 and SARS-CoV and the Effects of Mutations on the Binding Affinity. *J. Biomol. Struct. Dyn.* **2020**, 1–20, doi:10.1080/07391102.2020.1860829.

18. Muecksch, F.; Weisblum, Y.; Barnes, C.O.; Schmidt, F.; Schaefer-Babajew, D.; Lorenzi, J.C.C.; Flyak, A.I.; DeLaitsch, A.T.; Huey-Tubman, K.E.; Hou, S.; et al. Development of Potency, Breadth and Resilience to Viral Escape Mutations in SARS-CoV-2 Neutralizing Antibodies. *BioRxiv Prepr. Serv. Biol.* **2021**, doi:10.1101/2021.03.07.434227.


A
PANGO lineages
S:T478K
Prevalence (% of lineage)
0
10
20
30
40
50
60
70
7
3640
9
5
28
36
1
5
1
2
B.1.486
B.1.1.222
B.1.1.158
B.1.1.2
B.1.1.171
B.1.214
B.1.1.175
B.1.1.75
B.1.227
B.1.418
B
Frequency by Age
S:T478K
% of sequenced samples
0.0
0.5
1.0
1.5
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85-90
90-95
95-100
Age interval
C
Distribution by Sex
S:T478K
Female 35% (1478)
Male 35% (1471)
Unknown 30% (1265)
D
Frequency by Country
S:T478K
% of sequenced samples
0
10
20
30
40
1203
6
11
1
1
3
2536
6
26
144
Mexico
Curacao
Aruba
Iraq
Kosovo
Sint Maarten
USA
Bulgaria
Turkey
Switzerland
E
Log10 (N cases)
0
1
2
3


A
Occurrence over time
S:T478K
Occurrence (nr of samples)
0
200
400
600
800
2020
2021
7 42 77 112 147 182 217 252 287 322 357 392 427 448
Days since December 1, 2019
B
Prevalence over time
S:T478K
Prevalence (as % of sequence samples)
0.0
0.5
1.0
1.5
2.0
2020
2021
7 42 77 112 147 182 217 252 287 322 357 392 427 448
Days since December 1, 2019
C
Prevalence over time
S:N501Y
Prevalence (as % of sequence samples)
0
10
20
30
40
50
60
2020
2021
7 42 77 112 147 182 217 252 287 322 357 392 427 448
Days since December 1, 2019
D
Prevalence over time
S:D614G
Prevalence (as % of sequence samples)
0
20
40
60
80
100
2020
2021
7 42 77 112 147 182 217 252 287 322 357 392 427 448
Days since December 1, 2019
E
SARS-CoV-2 Spike
Human ACE2
T478K