Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Infection, Genetics and Evolution 90 (2021) 104760

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

SARS-CoV-2 mutation 614G creates an elastase cleavage site enhancing its spread in high AAT-deficient regions

Chandrika Bhattacharyya [a,1], Chitrarpita Das [a,1], Arnab Ghosh [a,1], Animesh K. Singh [a], Souvik Mukherjee [a], Partha P. Majumder [a,b], Analabha Basu [a], Nidhan K. Biswas [a,*]

[a] National Institute of Biomedical Genomics, Kalyani 741251, India
[b] Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Keywords:
SARS-CoV-2
614G subtype
Neutrophil Elastase
α1-antitrypsin deficiency
SERPINA1

ABSTRACT

SARS-CoV-2 was first reported from China. Within three months, it evolved to 10 additional subtypes. Two evolved subtypes (A2 and A2a) carry a non-synonymous Spike protein mutation (D614G). We conducted phylodynamic analysis of over 70,000 SARS-CoV-2 coronaviruses worldwide, sequenced until July2020, and found that the mutant subtype (614G) outcompeted the pre-existing type (614D), significantly faster in Europe and North-America than in East Asia. Bioinformatically and computationally, we identified a novel neutrophil elastase (ELANE) cleavage site introduced in the G-mutant, near the S1-S2 junction of the Spike protein. We hypothesised that elevation of neutrophil elastase level at the site of infection will enhance the activation of Spike protein thus facilitating host cell entry for 614G, but not the 614D, subtype. The level of neutrophil elastase in the lung is modulated by its inhibitor α1-antitrypsin (AAT). AAT prevents lung tissue damage by elastase. However, many individuals exhibit genotype-dependent deficiency of AAT. AAT deficiency eases host-cell entry of the 614G virus, by retarding inhibition of neutrophil elastase and consequently enhancing activation of the Spike protein. AAT deficiency is highly prevalent in European and North-American populations, but much less so in East Asia. Therefore, the 614G subtype is able to infect and spread more easily in populations of the former regions than in the latter region. Our analyses provide a molecular biological and evolutionary model for the higher observed virulence of the 614G subtype, in terms of causing higher morbidity in the host (higher infectivity and higher viral load), than the non-mutant 614D subtype.

1. Introduction

The COVID-19 pandemic caused by the coronavirus SARS-CoV-2 has been a major threat to humans (Dong et al., 2020). Coronaviruses infect humans with varying degrees of severity and lethality (Verity et al., 2020). Four of these viruses (NL63, 229E, OC43, and HKU1) cause mild respiratory problems in humans and three others (MERS-CoV, SARS-CoV and the newly emerged SARS-CoV-2) can cause severe respiratory syndromes (Chen and Li, 2020; Gorbalenya et al., 2020; Zhu et al., 2020). SARS-CoV-2 infection was first reported from Wuhan, China, on 24th December 2019 (Zhu et al., 2020).

SARS-CoV-2 virus is a single stranded (+) sense RNA virus with a genome length of about 30Kb. The virus binds to host cell surface receptors using spike protein to mediate fusion of the viral envelope with cell membrane. The receptor binding domain (RBD) located on the head of the S1 domain of the viral Spike (S) protein attaches with the angiotensin converting enzyme 2 (ACE2), that is expressed in large quantities in specific cell types (pneumocytes) of the human lung and other tissues (Lamers et al., 2020; Muus et al., 2020; Wang et al., 2020b; Ziegler et al., 2020).

RNA sequence analysis has shown that SARS-CoV-2 has acquired mutations and diversified as it spread geographically (Biswas and Majumder, 2020; Korber et al., 2020; Van Dorp et al., 2020). Most mutations are deleterious and viruses that acquire mutations are usually eliminated (Grubaugh et al., 2020). Random fluctuations in the frequency of viral subtypes occur; however, a virus with a mutation that provides selective advantage, usually manifested by higher transmission efficiency, is expected to rapidly rise to a high frequency (Bush et al., 1999). The ancestral type (O; with amino acid Aspartic Acid [D] at the 614th position of spike protein) was first reported from China in late

* Corresponding author at: National Institute of Biomedical Genomics, P.O.: N.S.S., Kalyani 741251, West Bengal, India.
E-mail address: nkb1@nibmg.ac.in (N.K. Biswas).
[1] These authors contributed equally.

https://doi.org/10.1016/j.meegid.2021.104760
Received 10 December 2020; Received in revised form 1 February 2021; Accepted 3 February 2021
Available online 5 February 2021

December 2019. Four weeks later (on 24th January 2020), a mutant type 614G (with Glycine [G] at the 614th position of the Spike protein; possessed by subtypes A2 and A2a), was reported from China. This mutant spread rapidly and widely across Europe and North-America [GISAID: https://www.gisaid.org/ and Nextstrain: https://nextstrain.org/] outcompeting the ancestral type - 614D (Biswas and Majumder, 2020; Gudbjartsson et al., 2020; Korber et al., 2020).

Clinical studies from two independent regions; Sheffield, England (614G, *n*=314; 614D, *n*=133) and Washington, USA (614G, n=407; 614D, n= 401), on COVID-19 patients, reported about 3-fold increase in viral load for individuals infected with SARS-CoV-2 614G variant over 614D (Korber et al., 2020; Wagner et al., 2020). Recent functional studies using multiple types of cell lines, including a human lung cell line, also demonstrated that SARS-CoV-2 with 614G mutation shows higher infectivity than 614D (Daniloski et al., 2020; Ozono et al., 2020; Zhang et al., 2020). No consensus view on the mechanism of higher infectivity of the mutant virus has yet emerged (Daniloski et al., 2020; Ozono et al., 2020; Zhang et al., 2020). To infect human cells, the viral spike protein of SARS-CoV-2 anchors to membrane bound host-ACE2 and gets cleaved by host proteases that facilitates membrane fusion; specifically the type II transmembrane serine protease (TMPRSS2) of the host cleaves the Spike protein near the S1-S2 junction (Shirato et al., 2018; Andersen et al., 2020; Hoffmann et al., 2020; Walls et al., 2020).

Variations in nucleotide sequences of *ACE2* and *TMPRSS2*, the two host genes whose products are indispensable for the entry of the coronavirus into host cells (Xia et al., 2019; Hoffmann et al., 2020; Matsuyama et al., 2020; Wang et al., 2020a; Walls et al., 2020), can alter the expression and functionality of these proteins. Amino acid altering variants in *ACE2* occur at non-polymorphic frequencies (Table S1) in most human populations. Our analysis found that the polymorphic variants in and around *ACE2* gene have no impact on its expression in lung tissues (GTEX portal). Even though some recent studies (Benetti et al., 2020; Cao et al., 2020; Stawiski et al., 2020) have attempted to implicate these variants with susceptibility to infection by SARS-CoV-2, there is no convincing evidence yet that the non-polymorphic variants in *ACE2* can modulate susceptibility to infection. However, *TMPRSS2,* the product of which is involved in the proteolytic cleavage of both *ACE2* and spike proteins of SARS-CoV-2 leading to internalization of the virion in the host cell, harbours many variants that exhibit considerable variation in frequencies among human populations. The three key host genes that modulate viral entry into human are *ACE2*, *TMPRSS2* and *FURIN* (a furin cleavage site is present at the S1/S2 boundary of SARS-CoV-2 Spike protein, which is cleaved during biosynthesis and is a novel feature that sets SARS-CoV2 apart from other SARS coronaviruses). We mined human genome databases in an attempt to correlate population frequencies of variants in and around these genes with the spread of 614G subtype; but, we were unable to discover any convincing relationship (data not shown). The rate of spread of the 614G (primarily A2a) has been non-uniform across geographical regions. Unfortunately, disaggregated country-wise data of frequencies of viral subtypes by

Table 1
Number of SARS-CoV-2 RNA sequences per month for all countries in three broad regions.

Broad Region	Country	Dec-19	Jan-20	Feb-20	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Total
	Singapore	0	10	34	231	300	78	86	66	805
	China	16	322	285	159	1	0	2	5	790
	South Korea	0	7	111	92	15	178	212	49	664
	Japan	0	9	109	286	136	7	0	0	547
	Thailand	0	23	9	176	14	0	0	0	222
East Asia	Hong Kong	0	24	58	49	4	2	1	50	188
	Taiwan	0	7	11	94	10	0	0	0	122
	Malaysia	0	6	9	49	23	16	0	0	103
	Vietnam	0	4	1	73	8	0	0	0	86
	Indonesia	0	0	0	11	4	0	0	2	17
	Philippines	0	0	0	8	0	0	2	0	10
	Cambodia	0	1	0	0	0	0	0	0	1
	United Kingdom	0	2	107	10146	16707	5314	1392	236	33904
	Spain	0	0	16	1781	694	100	64	15	2670
	Portugal	0	0	0	1229	225	75	59	0	1588
	Netherlands	0	0	14	753	473	190	0	0	1430
	Switzerland	0	0	25	290	229	57	69	272	942
	Belgium	0	0	2	560	253	75	15	30	935
	Denmark	0	0	2	617	90	2	0	0	711
	Sweden	0	1	6	344	117	91	34	0	593
	Iceland	0	0	1	555	0	0	0	0	556
	France	0	5	14	329	97	1	11	0	457
	Austria	0	0	7	344	71	2	0	0	424
	Germany	0	5	15	148	98	64	15	0	345
Europe	Russia	0	0	0	108	166	53	4	0	331
	Finland	0	1	1	61	163	41	0	0	267
	Luxembourg	0	0	2	166	85	13	0	0	266
	Italy	0	3	32	116	38	3	0	5	197
	Turkey	0	0	0	61	41	78	5	0	185
	Norway	0	0	7	33	20	6	42	28	136
	Greece	0	0	1	103	14	4	0	0	122
	Latvia	0	0	0	30	14	9	9	37	99
	Poland	0	0	0	38	33	19	7	0	97
	Ireland	0	0	0	12	0	8	2	56	78
	Hungary	0	0	0	45	16	0	0	0	61
	Czech Republic	0	0	1	46	4	0	0	0	51
	Slovenia	0	0	0	4	1	0	0	0	5
	Slovakia	0	0	0	4	0	0	0	0	4
	USA	0	12	81	6788	5773	2650	1800	441	17545
	Canada	0	4	10	294	228	1	0	0	537
North America	Panama	0	0	7	145	78	0	1	4	235
	Mexico	0	0	2	30	37	12	0	0	81
	Costa Rica	0	0	0	44	4	0	13	5	66

country are not available for an extended period, say from January to July 2020, for most regions. These data, for many countries, are either truncated at the beginning of this time period because of delayed start of viral sequencing or truncated at the end because of stoppage of viral sequencing for unclear reasons (Table 1). However, even aggregated data are revealing of non-uniform rate of spread. The increase in frequency of the 614G (primarily A2a) subtype viruses have been rapid in Europe and North-America, but significantly less so in East Asia (Biswas and Majumder, 2020). We hypothesized that host genomics – in view of large genomic differences between individuals of European and North-American ancestry and individuals of East Asian ancestry (Abdulla et al., 2009) – plays a role in determining the rate of increase in frequency of the 614G subtype virus, spatially. Conceptually, we posit that some host genomic backgrounds strongly favor infection by 614G subtypes while some other backgrounds do not favour infection as strongly. In this study, we have sought to test this hypothesis and have focused on host genetic factors that may regulate viral entry into host cells, other than variants in *ACE2*, *FURIN* and *TMPRSS2* genes that failed to provide satisfactory explanation in our analyses mentioned earlier. Although the meaning of the term 'virulence' is contextual, if we use the working definition proposed by Geoghegan and Holmes (2018) that virulence is the extent of harm caused by a pathogen to an infected host both in terms of morbidity and mortality (Geoghegan and Holmes, 2018), it is evident that 614G is more virulent than 614D at least in terms of morbidity. Compared to the 614D subtype, 614G is more infective – and hence has spread more widely – and, when it infects a human host it exhibits a higher viral load (Korber et al., 2020; Wagner et al., 2020). The overarching goal of this study is to develop a molecular and population genetic model for the observed greater virulence of the evolved subtype 614G and of its non-uniform spread among Caucasian and non-Caucasian populations.

2. Methods

2.1. Phylodynamic analysis of SARS-CoV-2 RNA sequences

One cardinal feature of the COVID-19 pandemic is our ability to monitor the spread and evolution of the SARS-CoV-2 virus almost in real time; since RNA sequences of the virus are being deposited every day in large numbers from all global regions to public databases. In partial fulfilment of the objective of this study, we have analysed the recent publicly available data. We downloaded all SARS-CoV-2 RNA sequences (n=82644) excluding low coverage (>5% N in the 29.9 Kb of each RNA sequence) on 20th Aug 2020, 8:50 AM, from the GISAID database (Shu and McCauley, 2017).

To analyze SARS-CoV-2 sequence data, we used the community standard nextstrain/ncov (Hadfield et al., 2018) (github.com/nextstrain/ncov) pipeline developed specifically for spatial and temporal tracking of pathogens. Nextstrain/ncov (Hadfield et al., 2018), is an open-source pipeline for phylodynamic analysis (Volz et al., 2013), including subsampling, alignment, phylogenetic inference, temporal dating of ancestral nodes and discrete trait geographic reconstruction as well as interactive data visualization. It comprises augur (Hadfield et al., 2018) (github.com/nextstrain/augur) a modular bioinformatics tool used for data analysis, and auspice (Hadfield et al., 2018) (github.com/nextstrain/auspice) a web-based visualization tool for phylogenomic and phylogeographic data.

Each downloaded fasta file was preprocessed to remove duplicate samples based on the identifier from the fasta headers (hCoV-19/<Country>/<Identifier>/<Year>). Out of the 82644 sequences, 78616 sequences passed default QC criteria of Nextstrain pipeline. These 78616 sequences were aligned using MAFFT (Katoh and Standley, 2013). These estimates were further refined using RAxML. Augur also estimates the frequency-trajectories of mutations, genotypes and clades of a phylogenetic tree, which is used by the auspice package to visually represent phylogenetic tree, geographic transmission and entropy (genetic diversity). Of the 78616 global sequences, Augur was able to assign clade information to 78203 sequences, from which 633 sequences were filtered out for lack of appropriate clade-defining mutations, sequences from August and non-human host discovered during manual curation. Remaining 77570 sequences were curated for further analysis.

In order to build phylogenetic time tree, we have performed subsampling by allowing maximum of 75 sequences per country per month per year. We estimated timescale and branch lengths of a reconstructed phylogenetic tree using IQ-TREE (Nguyen et al., 2014) (as implemented in augur) considering hCoV-19/Wuhan/WH01/2019 as ancestral; (https://www.gisaid.org/) with a global subset of 11260 sequences. We repeated the random sampling of tree building multiple times to convince ourselves that there is no difference in overall inference (data not shown).

We used the date of viral sample collection for all phylodynamic analysis to study epidemiological and evolutionary patterns. The initial sampling for sequencing was sparse, possibly non-representative and unstable. Details of the phylogenetic tree of the viral sequences and defining mutations of various clades are provided with supplementary table 3. Various population genetic summary statistics for ancestral 614D and the derived 614G clades were obtained.

2.2. Estimation of growth rate of relative frequency of 614G subtype virus

We have fitted the following sigmoidal curve to model the increase in the frequency of the 614G subtype in a population over time.

$$f(t) = \frac{a}{1 + e^{\frac{(b-t)}{c}}}$$

where, "a", "b" and "c' are the parameters of the equation and f(t): denotes the 614G subtype frequency at time t. The interpretation of the parameters are as follows: "a" denote the asymptote or the value to which f(t) asymptotically converges, "b" represents the time at which the frequency (f(t)) reaches 50% of the asymptote and inverse of "c" denotes the slope of the tangent at "b" indicating the rate of growth (Zullinger et al., 1984). We assign the asymptote, "a" to be 100 arguing that the 614G subtype has a selective advantage over the 614D. However, we also assume that the rate of growth of the 614G subtype will be different for different regions of the world (East Asia, Europe and North-America). Hence, we fitted the data for the different regions separately to estimate the values of "b" and "c" for each of the 3 regions (East Asia, Europe and North-America); a larger value of b indicates a longer time to reach 50% frequency in a population and the inverse of c is the quantitative estimate of the rate of increase. We have fitted the model using time series moving average (order = 3) of 614G subtype frequencies over time. The values of "b" were estimated separately for East Asia, Europe and North-America.

To find whether there is any significant difference in estimated "b" for different regions, we hypothesised that there exist no systematic stochastic variations in the frequency of 614G subtype in different regions; the observed variations is only because of chance factor.

To test our null hypothesis, we have resampled the month wise 614G subtype frequencies (pooled from all regions) into two random populations and calculated the absolute difference between estimated "b" values after fitting our model. We repeated this experiment 1 million times. The number of times, among the one million, when the difference between the "b" values in the two random populations exceeded our observed difference; provided us an empirical p value for the test of equality of the "b" for different regions.

2.3. In-Silico prediction of cleavage sites on SARS-CoV-2 Spike protein

Functional studies showed that by synthetically introducing mutations near the S1-S2 junction of SARS-CoV, additional proteolytic cleavage sites are generated that enhance viral membrane fusion by

several fold (Belouzard et al., 2009). We downloaded the sequence of SARS-Cov-2 spike (S) protein (QHD43416), which is 1273 amino acids long, from https://zhanglab.ccmb.med.umich.edu/COVID-19/ (Roy et al., 2010). We used PROSPER (Song et al., 2012) (https://prosper.erc.monash.edu.au/) to predict proteolytic cleavage sites based on i) local amino acid sequence profile, ii) predicted secondary structure, iii) solvent accessibility and iv) predicted native disorder. In particular, we identified potential protease substrate sites in the amino acid sequence of SARS-CoV-2 spike (S) protein. The predicted protease cleavage sites were also verified by another protease cut site prediction tool PROSPERous (Song et al., 2018). The predicted cleavage site was further verified to have mass spectrometry supportive evidence from MEROPS database (Rawlings et al., 2017).

2.4. Tissue specific expression and genomic variation in ACE2, TMPRSS2 and FURIN in global populations

Using GTEx data (https://gtexportal.org/), we identified regulatory eQTLs that are significantly associated with *ACE2*, *TMPRSS2* and *FURIN* gene expression in various human tissues. Genotype data, for all significant eQTLs as well as the data on all variants, were extracted from 1000Genomes dataset (Auton et al., 2015) with the hg19 chromosomal coordinates as reference. Initial data to assess population genomic diversity were downloaded from 1000Genomes project that included representative populations from Europe (CEU, TSI, FIN, GBR, IBS), admixed Hispanic speakers from America (MXL, PUR, CLM) and East Asia (CHB, CHD, JPT). Functional annotation of each identified variant was done using Annovar (Wang et al., 2010). We ranked the eQTLs according to their impact on the tissue-specific expression level for the gene of interest and selected variants with high impact. For all these eQTLs and non-silent variants, we calculated the genetic distance, F_{st} (Weir and Cockerham, 1984), between pairs of major continental populations and regional subpopulations, using PLINK v1.9 (Chang et al., 2015) (https://www.cog-genomics.org/plink/).

3. Results

3.1. Temporal and geographical spread of subtypes of SARS-CoV-2

The first set of RNA sequences collected from 16 infected individuals from Wuhan, China, had high sequence identity, and was named the O subtype. The O subtype spread to other provinces (e.g., Guangdong, Jiangxi) of China and also to nearby countries, e.g., Thailand (with the first submission to GISAID - Nonthaburi/61/2020 - on 8th January 2020), within two weeks of its first reporting from Wuhan. We performed phylodynamic analysis and showed that the virus evolved to B and B2 in the first two weeks of January and later to B1, B4, A2, A2a and A3 (Fig. S1a). These subtypes rapidly spread worldwide to many countries of East Asia (10 countries), Europe (6 countries) and North-America (2 countries) by the end of January. The distribution of viral clades as reported from different countries of East Asia, Europe and North-America until the end of July is summarized in (Table 2). Analysis of data on 77570 sequences generated globally till 31st July showed that mutations were distributed over 24493 nucleotide sites. Among these sites, a subset of 11 early high-frequency sites (of which 8 were coding [ORF8 -L84S, ORF1a - V378I, ORF1a - L3606F, ORF1a - A3220V, ORF3a - G251V, ORF1a - L3606F, S - D614G, ORF1b - P314L]) enabled defining the phylogenetic clade structure of the viral sequences (Fig. S2, Table 2) (Biswas and Majumder, 2020). By 31st March 2020, SARS-CoV-2 had evolved into 10 major clades; five of which had attained frequencies higher than 5%; A2a=63.80%, O=10.25%, B=6.51%, B1=5.04% and A1a=12.01% (Table 3). The A2a clade with the highest frequency is defined by nucleotide changes at two sites that are in complete non-random association (linkage disequilibrium): D614G in the Spike glycoprotein and P314L in Orf1b polyprotein (also known as, RdRp: P323L). Of the 10 clades, only 2 clades (a minor A2 subtype and the major A2a subtype) harbour amino acid G at 614th position of spike protein; the remaining 8 clades (ancestral O and evolved clades B, B1, B2, B4, A3, A6, and A1a) have D, the ancestral subtype (Fig. S1b). We have mined the GISAID database to discover that the most frequent 614G subtype (comprising A2 and A2a) arose in China in mid-January 2020 (Inferred date: January 7th 2020 CI: 16th Dec 2019 - 7th January 2020; first 614G subtype sequence was deposited in GISAID from Zhejiang [Zhejiang/HZ103/2020] on 24th January) and spread to multiple locations (e.g., Shanghai, Beijing) within the next few weeks. The earliest evidence of 614G subtype in Europe was from Germany around the end of January (Germany/BavPat1/2020 sequence deposited on 28th January). The viral landscape at that time (end of January) comprised predominantly viruses with 614D; 99.07% of submitted viral sequences from East Asia, 70.59% from Europe and 100% from North-America. Extremely rapid and dramatic shifts in the viral landscape took place within next four weeks (i.e. up until end of February). The frequency of 614G subtype rose from 29.41% to 52.22% in Europe and from 0% to 9.48% in North-America. However, this shift was slow in East Asia where the 614G frequency remained low; in China, <2% and in other East Asian countries, 1%; (Fig. 1, Fig. S3). The 614G subtype continued to rise in frequency replacing the previously frequent 614D subtype. By March, 614G became the dominant subtype in Europe (67.27%) and North-America (68.69%); the rise in East Asia was less dramatic (24.08%) (Fig. 1, Fig. S3). After March, data submissions to GISAID became extremely skewed and infrequent from many countries, including China. The rapid increase of the frequency of 614G subtype over 614D in Europe and North-America, where it quickly reached frequency of close to 100% replacing all other subtypes remains same until July, details are provided in Table S2.

To quantitatively estimate the parameters of growth of the relative frequency of the 614G subtype, we have modelled the increase in the frequency of the 614G subtype as a non-linear sigmoidal function (described in Methods Section 2.2). We fitted the moving averages of 614G frequency (Table S3) for the three regions (East Asia, Europe and North-America) to estimate the parameters of growth of the 614G subtype separately for the three regions. We assume that the 614G will have a selective advantage and will eventually reach a 100% frequency in all regions (East Asia, Europe and North-America).We have estimated that, in order to reach 50% relative frequency, the 614G subtype took significantly (sampling-resampling test, $p < 0.01$ [details in Methods Section 2.2]) longer time in East Asia (5.5 months) compared to Europe (2.15 months) as well as North-America (2.83 months) [the estimates of the parameters of the sigmoidal and the sampling-resampling p-values are tabulated in Table 4, Fig. 2]. Because of unavailability of comparable disaggregated data, we have used aggregated data to show that although the 614G subtype outcompeted the 614D in three regions, the rate of growth of the 614G subtype over the 614D is significantly higher in Europe and North-America.

3.2. The explosive increase of 614G subtype cannot be explained by early founding

A possible explanation for higher frequency of the 614G variant over 614D is that early founding events resulted in higher frequencies; without natural selection favouring 614G to outcompete 614D. To examine this possibility, we tabulated (Table S4) the frequencies of viruses of the two variant types during each week of collection for the period December, 2019 to July 31, 2020, in each of the three regions under consideration (Europe, North-America and East Asia). Table S4 shows that even though there were differences in the date of introduction of 614G in the three regions, within less than a month after introduction, the frequency of the 614G variant started rising faster than that of 614D. Further, the 614G rapidly climbed to a very high frequency in Europe and North-America (Fig. S3, Table S5); not attributable to earlier founding. Such rapid increase was not observed in East Asia after the arrival of 614G in that region. There is also no evidence of a large and

sudden influx of 614G sequences in any of the geographical regions under consideration.

3.3. D614G mutation generates a novel neutrophil elastase cleavage site

The non-synonymous D614G mutation is located between S1-RBD and S1/S2 junctions of the SARS-CoV-2 spike (S) protein (Fig. 3b). The amino acid position 614 is monomorphic for D (Aspartic acid) residue in SARS-CoVs obtained from bat, civet, pangolin and human (Fig. 3a). A previous study indicated that the introduction of new proteolytic sites at and around the S1-S2 junction substantially increases SARS-CoV fusion with the cell membrane (Belouzard et al., 2009). Using PROSPER (Song et al., 2012), we identified a novel neutrophil elastase (*ELANE*) – a serine protease – cleavage site (with support vector regression score of 1.07) at the 614G spanning region of the spike protein; also supported by mass spectrometry evidence from the MEROPS database (https://www.ebi.ac.uk/merops/). The position 614 on the Spike protein is the nearest substrate site for neutrophil elastase to cleave at 615-616, on S1 subunit of Spike protein (Fig. 3c; Table 5). To enable viral entry, the Spike glycoprotein must be cleaved by host

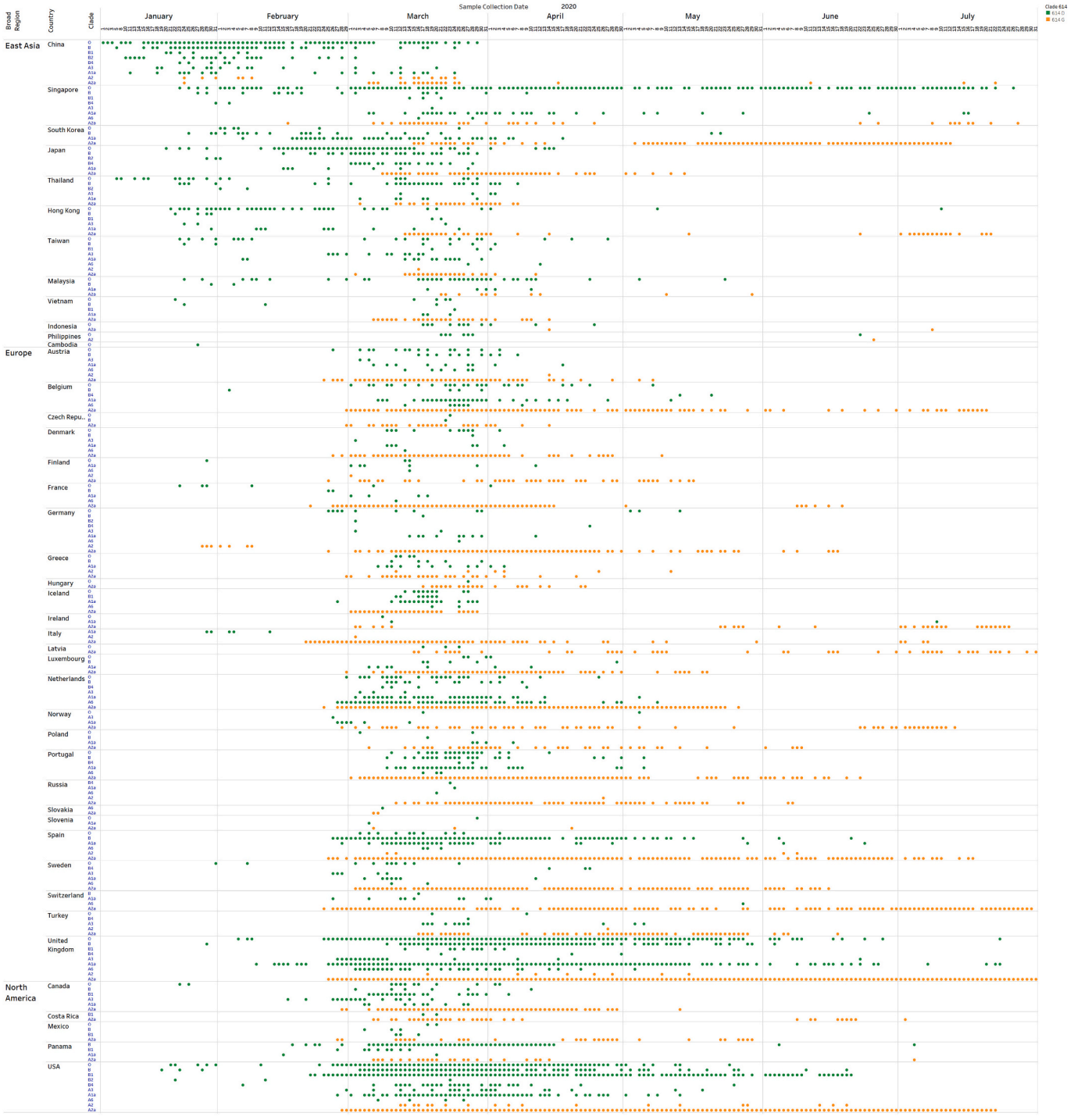

Fig. 1. Day-wise pattern of collection of SARS-CoV-2 isolates during Dec 2019 to July 31st, 2020, from different countries classified by subtype. The points are color coded based on the presence of D (Green) or G residue (Orange) at the 614th amino acid position of SARS-CoV-2 spike (S) protein.

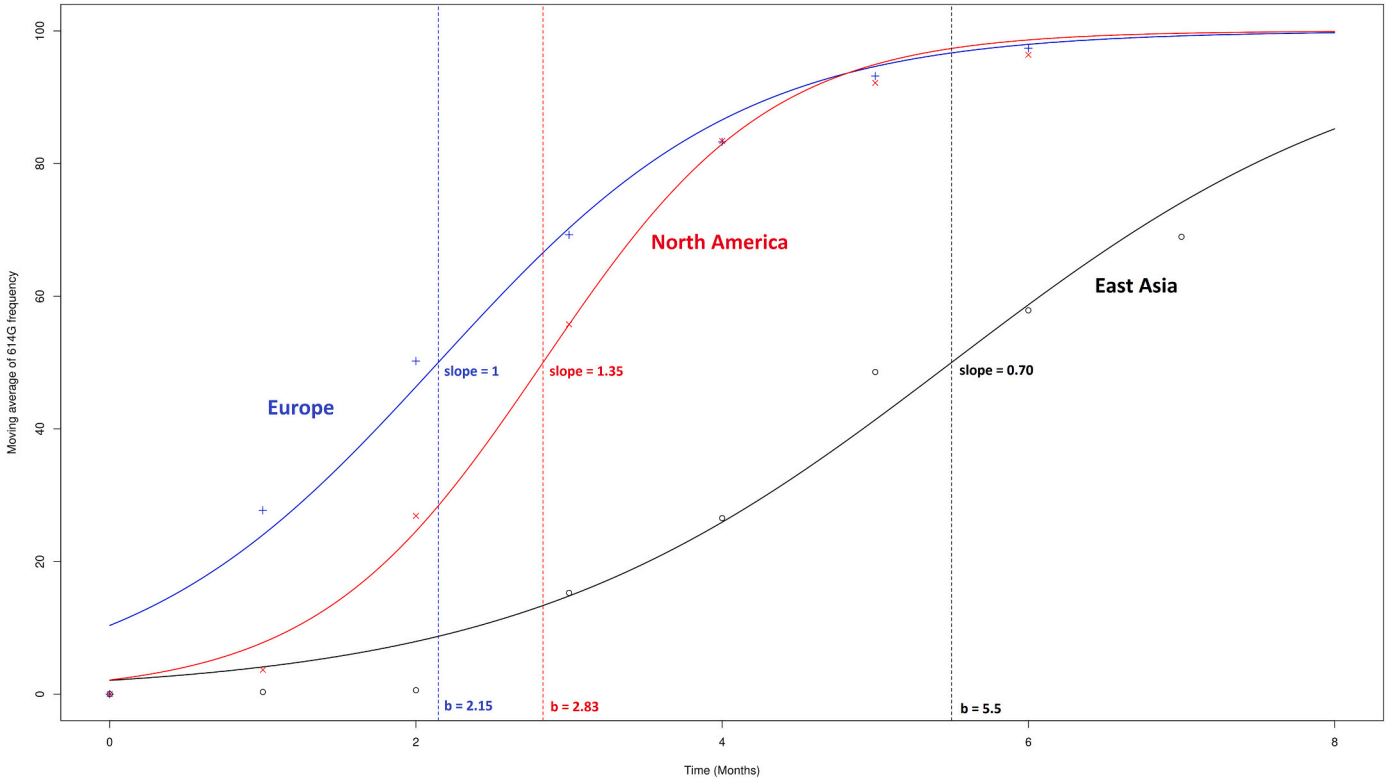

Fig. 2. Modelling of SARS-CoV-2 spike 614G subtype relative frequency with time for different regions (East Asia, Europe and North-America). Each point represents the time series moving averages (order = 3) of 614G subtype frequency over months in different regions. To reach 50% relative frequency, the 614G subtype took significantly less time in Europe (2.15 months) and North-America (2.83 months) as compared to East Asia (5.5 months).

proteases to enable fusion of the viral envelope with the host cell membrane (Hoffmann et al., 2020; Shang et al., 2020). An additional cleavage site enables processing of Spike protein by both of these two host serine proteases, TMPRSS2 (Hoffmann et al., 2020; Matsuyama et al., 2020; Wang et al., 2020a; Walls et al., 2020) and neutrophil elastase (Matsuyama et al., 2005; Belouzard et al., 2010; Hu et al., 2020), and thus likely enhances the possibility of entry of SARS-CoV-2 into the host cell. This predictive inference derives support from the evidence that the spike protein with the 614G variant is cleaved by neutrophil elastase 4-fold more efficiently than 614D that is cleaved only by TMPRSS2 (Hu et al., 2020). Since cleavage leads to activation of the Spike protein, the activation of spike protein of 614G subtype virus is expected to be greater when there is a higher level of active neutrophil elastase. Higher level of neutrophil elastase can accrue if there is a higher number of neutrophils. If data support that there is higher number of neutrophils among Caucasians (Europeans and North-Americans) than among non-Caucasians (East Asians), then the rapid spread of 614G among Caucasians can be easily explained. However, studies have shown that there is no significant difference in baseline neutrophil levels among Caucasians and non-Caucasians (Bain et al., 1984; Tajuddin et al., 2016). We noted that the expression level of neutrophil elastase is regulated by α1-antitrypsin, an elastase inhibitor encoded by the human *SERPINA1* gene (Dau et al., 2015; Strnad et al., 2020).

3.4. Non-uniform geographical spread of 614G is likely to have been determined by differences in prevalence of α1-antitrypsin deficiency

SERPINA1, that encodes the neutrophil elastase inhibitor α1-antitrypsin (AAT), is known to be multi-allelic; with alleles denoted as M, S and Z; and many rare alleles (Brantly et al., 1988; Crystal et al., 1989; Hutchison, 1998; Strnad et al., 2020). The serum level of AAT in an individual is dependent on the individual's genotype; MM individuals have the highest and normal (100%) level. MS, SS, MZ, SZ, and ZZ – the five major deficiency genotypes – express ~80%, 60%, 55%, 40%, and 15% of AAT, respectively (Blanco et al., 2017). Thus, the prevalence of AAT deficiency varies among populations (Burrows et al., 2000; Dau et al., 2015; Borel et al., 2018) in proportion to the population frequencies of the alleles S and Z. The primary role of circulating AAT, which is synthesized in the liver (Gómez-Mariano et al., 2020), is to protect the lung tissue against damage by neutrophil elastase (Dau et al., 2015; Strnad et al., 2020). Some of the AAT deficient genotypes (e.g., S-Z: [SS, SZ and ZZ]) are susceptible to COPD (Dahl et al., 2005). We have re-analysed data on S-Z frequency in various populations (Blanco et al., 2017; Shapira et al., 2020) and found the S-Z deficient genotype is more abundant in European (18.7-75.9 per 1000 individuals; median = 31.20 per 1000 individuals) and North-American (23.4-32.1/1k; median = 29.00/1k) countries as compared to East Asian countries (0-19.9/1k; median = 2.25/1k), described in Table S6. Considering all deficient genotypes, AAT deficiency is common in populations of Europe and North-America (9-16%) and rare (~2%) among East Asians (Hutchison, 1998; De Serres, 2002; Crowther et al., 2004; De Serres et al., 2010) (Table S7). The frequency of the major deficiency allele Z is very high among Caucasians compared to non-Caucasians of East and Southeast Asia (Blanco et al., 2017). In sum, (a) higher prevalence of deficiency of the neutrophil elastase inhibitor AAT in Caucasian compared to non-Caucasian populations will lead to a higher level of neutrophil elastase; (b) the higher level of neutrophil elastase will activate the spike protein of the subtype 614G, but not of 614D, in greater amounts; and (c) as a consequence the 614G subtype will spread more efficiently in Caucasian populations of Europe and North-America than in non-Caucasian populations of East Asia (Fig. 4).

4. Discussion

Within four months of its first appearance, SARS-CoV-2 spread

Table 2
The distribution of viral clades as reported from different countries of East Asia, Europe and North-America until the end of July.

Broad Region	Country	O	B	B1	B2	B4	A1a	A3	A6	A2	A2a	A2/A2a*	Total
East Asia	Singapore	601	19	3	0	2	48	1	2	0	128	1	805
	China	474	168	10	33	8	21	21	0	15	40	0	790
	South Korea	6	43	0	0	0	135	0	0	0	480	0	664
	Japan	147	64	0	4	42	9	0	0	0	281	0	547
	Thailand	30	106	0	2	0	8	4	0	0	72	0	222
	Hong Kong	76	5	2	0	0	13	4	0	0	88	0	188
	Taiwan	22	7	4	0	0	14	13	2	1	59	0	122
	Malaysia	83	3	0	0	0	5	0	0	0	12	0	103
	Vietnam	6	4	2	0	0	3	0	0	0	71	0	86
	Indonesia	14	0	0	0	0	0	0	0	0	2	1	17
	Philippines	9	0	0	0	0	0	0	0	1	0	0	10
	Cambodia	1	0	0	0	0	0	0	0	0	0	0	1
Europe	United Kingdom	2374	362	20	0	5	4010	50	86	9	26906	82	33904
	Spain	16	911	0	0	0	76	0	3	5	1656	3	2670
	Portugal	24	39	0	0	2	84	0	3	0	1436	0	1588
	Netherlands	30	13	0	0	5	71	3	161	0	1147	0	1430
	Switzerland	0	1	0	0	0	11	0	1	0	926	3	942
	Belgium	39	4	0	0	4	47	0	16	0	825	0	935
	Denmark	28	2	0	0	0	8	1	1	0	670	1	711
	Sweden	15	0	0	0	3	8	8	4	0	555	0	593
	Iceland	24	0	16	0	0	92	0	3	0	421	0	556
	France	8	2	0	0	0	4	0	1	0	442	0	457
	Austria	17	14	0	0	0	11	2	6	1	372	1	424
	Germany	27	1	0	1	1	11	2	1	9	292	0	345
	Russia	0	0	0	0	2	1	0	1	1	326	0	331
	Finland	3	0	0	0	0	6	0	1	1	256	0	267
	Luxembourg	4	5	0	0	0	7	0	0	0	250	0	266
	Italy	0	0	0	0	0	6	0	0	1	190	0	197
	Turkey	2	0	0	0	1	0	24	0	1	157	0	185
	Norway	2	0	0	0	0	8	1	0	0	125	0	136
	Greece	4	6	0	0	0	15	0	0	6	80	11	122
	Latvia	5	0	0	0	0	0	0	0	0	94	0	99
	Poland	2	1	0	0	0	4	0	0	0	90	0	97
	Ireland	1	0	0	0	0	3	0	0	0	73	1	78
	Hungary	1	0	0	0	0	0	0	0	0	60	0	61
	Czech Republic	1	1	0	0	0	0	0	0	0	49	0	51
	Slovenia	1	0	0	0	0	1	0	0	0	3	0	5
	Slovakia	0	0	0	0	0	0	0	1	0	3	0	4
North America	USA	476	376	1867	3	58	256	26	5	22	14345	111	17545
	Canada	16	7	49	0	0	11	20	0	0	433	1	537
	Panama	0	181	10	0	0	2	0	0	0	42	0	235
	Mexico	2	4	7	0	0	0	0	0	0	67	1	81
	Costa Rica	0	0	2	0	0	0	0	0	0	64	0	66
Total		4591	2349	1992	43	133	5009	180	298	73	53588	217	68473

O = [Ancestral Clade], B = [ORF8 - L84S (T28144C)], B1 = [ORF8 - L84S, nt - C18060T], B2 = [ORF8 - L84S, nt - C29095T]
B4 = [ORF8 - L84S, N - S202N], A3 = [ORF1a - V378I, ORF1a - L3606F], A6 = [nt - T514C], A7 = [ORF1a - A3220V]
A1a = [ORF3a - G251V, ORF1a - L3606F], A2 = [S - D614G], A2a = [S - D614G, ORF1b - P314L]

Table 3
Count and proportions of 10 SARS-CoV-2 phylogenetic clades

Sl No	Clade	Count (till March 31st)	Proportion (till March 31st)	Count (till July 31st)	Proportion (till July 31st)
1	O	2859	10.25	4591	6.70
2	B2	43	0.15	43	0.06
3	B4	104	0.37	133	0.19
4	A3	165	0.59	180	0.26
5	A6	267	0.96	298	0.44
6	B	1815	6.51	2349	3.43
7	B1	1406	5.04	1992	2.91
8	A1a	3351	12.01	5009	7.32
9	A2	41	0.15	73	0.11
10	A2a	17795	63.80	53588	78.26
11	A2/A2a	48	0.17	217	0.32
Total		27894	100	68473	100

Table 4
Estimation of coefficients for SARS-CoV-2 614G frequency growth model.

Region	Coefficient	Estimate	p value
East Asia	b	5.4978	3.99E-08
East Asia	c	1.4273	1.36E-04
Europe	b	2.1455	1.78E-05
Europe	c	0.9941	4.98E-04
North America	b	2.8299	7.72E-08
North America	c	0.7398	3.27E-05
Model	$f(t) = \frac{a}{1 + e^{\frac{(b-t)}{c}}}$		a = 100
Resampling test for estimated b			
	p value	H[0]: beta (b) of spike 614G frequency growth curve is same for all regions	
EAS-EUR	0.000653		
EAS-AMR	0.005976		
EUR-AMR	0.35849		

rapidly to more than 200 countries in different continents. SARS-CoV-2 evolved from an ancestral subtype (O) to 9 major derived subtypes, with characteristic sets of mutations. One non-synonymous (Aspartic acid to Glycine) mutation at amino acid position 614 (D614G) of the Spike protein of the coronavirus stands out. Viruses (classified as A2 and A2a clades only, of the 10 clades) carrying the derived allele (G) spread

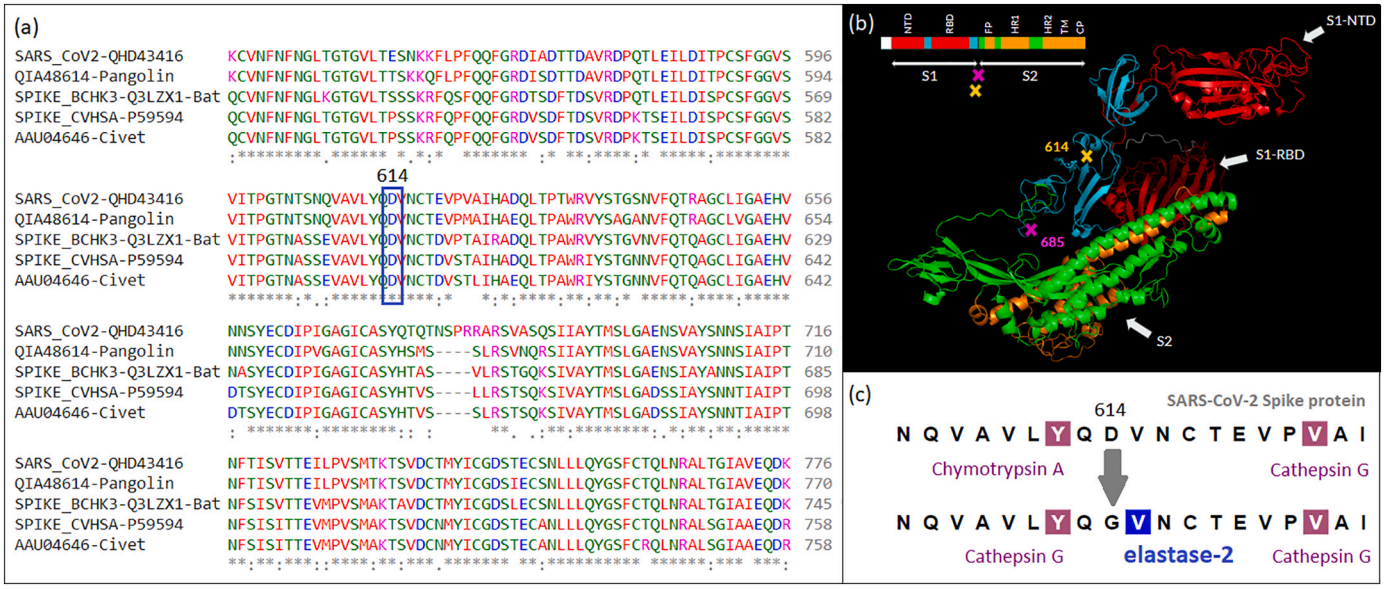

Fig. 3. (a) Multiple alignment of spike (S) protein of corona viruses from Bat, Pangolin, Civet and Human SARS-CoV with the Human SARS-CoV-2 revealed >70% sequence identity. The 614th amino acid position in the S protein was found to be conserved among species, until the occurrence of the D614G mutation. (b) Domain structure of SARS-CoV-2 S protein; amino acid residues 614 and 685 (S1-S2 junction) are marked on the 3D protein structure. (c) An additional neutrophil elastase (elastase -2) cleavage site around S1-S2 junction was introduced due to the D614G mutation in SARS-CoV-2; the Glycine at 614th position is predicted to be the nearest substrate site for neutrophil elastase to perform proteolytic cleavage in the adjacent residue.

Table 5
Prediction of protease cut site in SARS-CoV-2 S protein around 614 AA position.

1) SARS-CoV-2 S protein 614G imputed sequence (TSNQVAVLYQ[G]VNCTEVPVAI)

Merops ID	Protease name	Position	P4-P4' site	N-fragment (kDa)	C-fragment (kDa)	Cleavage score
M10.004	*matrix metallopeptidase-9*	7	QVAV|LYQG	0.69	1.6	1.14
M10.004	*matrix metallopeptidase-9*	20	VPVA|I	2.18	0.11	1.13
M10.004	*matrix metallopeptidase-9*	8	VAVL|YQGV	0.81	1.48	0.99
S01.131	Neutrophil elastase (elastase-2) [a]	12	YQGV|NCTE	1.37	0.92	1.07
S01.133	*cathepsin G*	9	AVLY|QGVN	0.97	1.32	1.02
S01.133	*cathepsin G*	19	EVPV|AI	2.11	0.18	0.94

2) SARS-CoV-2 S protein 614D sequence (TSNQVAVLYQ[D]VNCTEVPVAI)

Merops ID	Protease name	Position	P4-P4' site	N-fragment (kDa)	C-fragment (kDa)	Cleavage score
M10.004	*matrix metallopeptidase-9*	7	QVAV|LYQD	0.69	1.54	1.14
M10.004	*matrix metallopeptidase-9*	20	VPVA|I	2.13	0.11	1.13
M10.004	*matrix metallopeptidase-9*	8	VAVL|YQDV	0.81	1.43	0.99
S01.001	chymotrypsin A (cattle-type)	9	AVLY|QDVN	0.97	1.26	0.94
S01.133	*cathepsin G*	9	AVLY|QDVN	0.97	1.26	1.13
S01.133	*cathepsin G*	19	EVPV|AI	2.05	0.18	0.94

[a] Specific to Spike 614G subtype

geographically with extreme rapidity (Biswas and Majumder, 2020; Korber et al., 2020), but non-uniformly, outcompeting the clades that possessed the ancestral allele D. The rapid and non-uniform geographical spread of 614G, as we have shown here, cannot be fully explained by early founding effect. We, therefore, sought to evaluate whether the observed geographical distribution of the 614G subtype can be explained, at least in part, by (a) the intrinsic advantage of the 614G subtype over the 614D, and (b) the allele-frequency distribution of specific gene variants in human host populations. In other words, using the broader concept of 'virulence' (Geoghegan and Holmes, 2018) to include morbidity, and not just mortality, here we have provided a molecular genetic model of greater virulence of the evolved subtype 614G – higher ability to infect a human host resulting in higher mean viral load – compared to 614D. In addition, we have provided a population genetic model for the non-uniform geographical spread of the 614G subtype among Caucasian populations of Europe and North-America and non-Caucasian populations of East Asia.

We have performed molecular population genetic analysis using publicly available data of RNA sequences of SARS-CoV-2 isolates to correlate pathogen evolution with viral transmission. We have shown that – the 614G subtype which arose in East Asia in January 2020, spread with extreme rapidity throughout the European and North-American continents. The spread of 614G subtype has been so explosive, that in 10 weeks (of February and March 2020) over 64.11% of globally infected individuals were identified to carry the 614G subtype starting from only 1.95% in January. The 614G subtype frequency further arose to 78.69% worldwide by end of July. We were able to show that the 614G subtype outcompeting the ancestral 614D subtype viruses significantly faster in Europe and North-America as compared to East Asia.

The D614G mutation is near the S1-S2 junction on SARS-CoV-2 Spike (S) protein. SARS-CoV-2-pseudotyped lentiviral particles with this

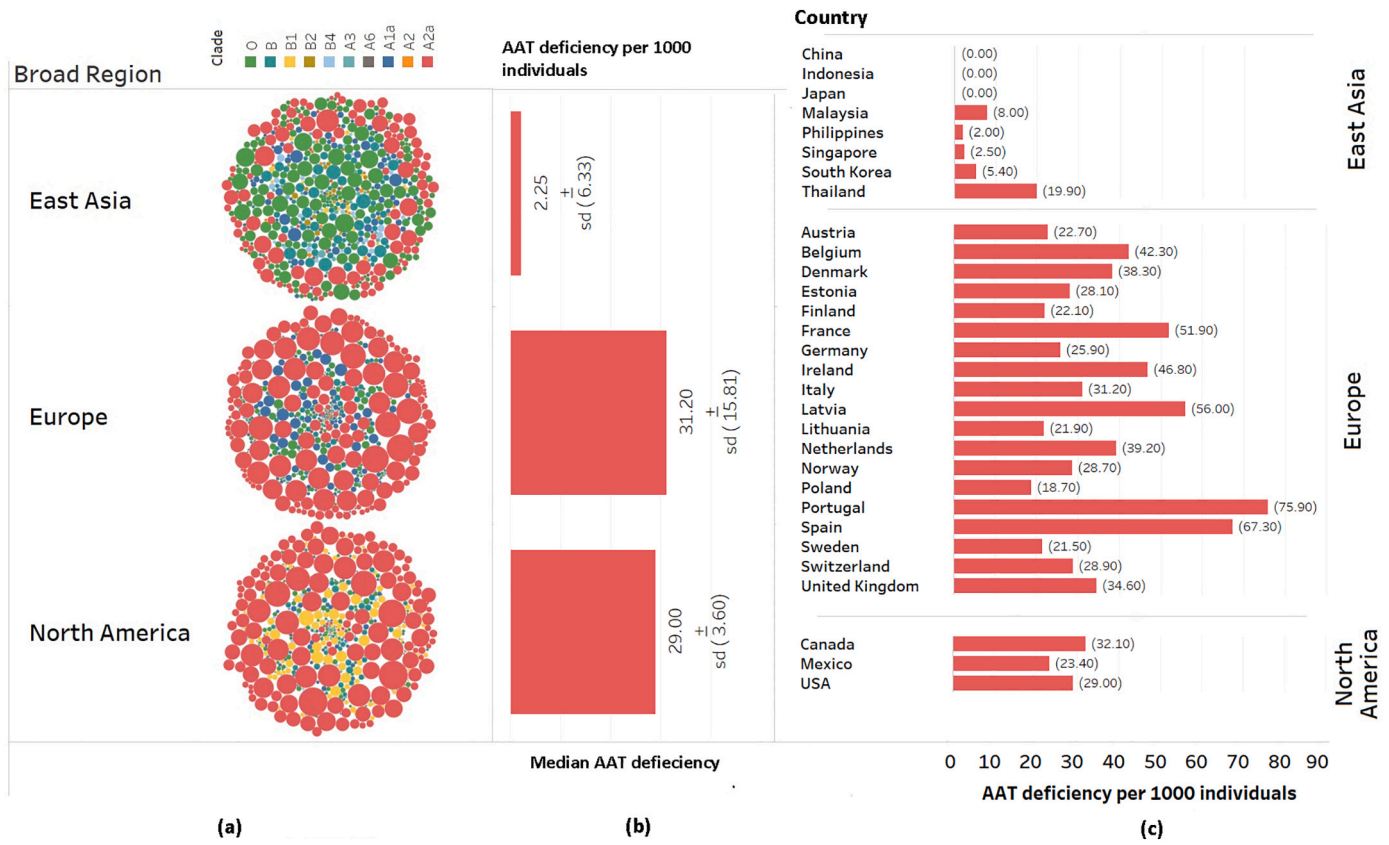

Fig. 4. (a) SARS-CoV-2 subtype distribution in different populations (East Asian, European and North-American). Circle size is proportional to the number of sequences submitted on a particular date and the circle colour distinguishes the subtypes. The positions of circles from the centre towards the outer periphery correspond to the date of submission from January to July [more recent submissions are towards the periphery]. (b) The horizontal bar represents median alpha-1 antitrypsin deficiency for different continental populations. (c) Thin horizontal bars represent alpha-1 antitrypsin deficiency per 1000 individuals of individual countries in three broad regions.

variant 614G were shown to infect multiple human cell types more efficiently (~3.5 fold), compared to 614D variant (Ozono et al., 2020). Presence of Glycine at 614th amino acid position of Spike protein was speculated to result in higher stability of the protein (Daniloski et al., 2020; Ozono et al., 2020; Zhang et al., 2020). Here, we have shown that SARS-CoV-2 virus acquired an additional neutrophil specific cleavage site in the Spike protein during the first month of the transmission. This additional cleavage site allows 614G viruses to gain entry into human cell lines more easily in presence of elastase (Hu et al., 2020). Further, the neutrophil elastase inhibitor Sivelestat Sodium significantly decreased 614G spike protein cleavage in presence of neutrophil elastase (Hu et al., 2020).

We noted that in spite of the advantage gained by 614G mutant viruses for entry into host cells, the spread of this mutant virus was non-uniform across geographical regions; explosive spread among individuals of European and North-American ancestry, but slower spread among individuals of East Asian ancestry. We, therefore, explored host genomic differences to explain the difference in geographical spread of the mutant virus. Due to presence of an additional neutrophil elastase cleavage site, the 614G subtype virus predicted to have selective advantages over ancestral 614D subtype. We therefore expected there will be difference in availability of neutrophil elastase among Europeans, North-Americans and East Asians which might result in observed differential acceleration. As shown by multiple epidemiological studies, there is no significant differences in baseline neutrophil levels among Europeans, North-Americans and East Asians (Bain et al., 1984; Tajuddin et al., 2016). On the other hand, α1-antitrypsin (AAT) which is the main inhibitor of neutrophil elastase (Dau et al., 2015; Strnad et al., 2020) found to have multiple deficient alleles in higher proportion of European (9.36%) and North-American (16.31%) populations as compared to less than 2% among East Asians (Hutchison, 1998; De Serres, 2002; Crowther et al., 2004; De Serres et al., 2010). The main function of AAT is protect lung from inflammation and tissue damage by neutrophil elastase (Dau et al., 2015; Strnad et al., 2020). Several studies have shown that AAT deficiency among populations of European ancestry is a major contributor of diseases like COPD which is caused by lung inflammation by neutrophil elastase (Dahl et al., 2005). During acute-phase response, AAT levels can increase by 100% in persons with a normal genotype (MM), but the rise is markedly attenuated in persons with deficiency alleles (Strnad et al., 2020). As a large proportion of European and North-American population suffers from AAT deficiency, the availability of active neutrophil elastase at the site of infection is expected to be higher which contributed faster growth rate of 614G subtype virus, compared to the 614D subtype, that carries the additional elastase specific cleavage site. We emphasize that this finding along with other social factors may explain the differential geographical/ethnic spread of 614G. Based on the inferences of our analyses and additional inferences derived from studies by others, we have proposed that the evolved SARS-CoV-2 subtype 614G is more virulent in terms of causing greater morbidity in the host because the D to G mutation creates an additional cleavage site that enhances entry of the 614G subtype into the host cell and consequently enhances its ability to infect. However, this subtype's ability to infect humans is not uniform and depends on the genotypes at specific loci of the *SERPINA1* gene in the host as summarised in Fig. 5. We have correlated the extent of spread of the subtype across populations with the population frequency of relevant genotypes, that causes a deficiency of the protein – a1-antitrypsin – encoded by *SERPINA1*. Thus, we have provided a holistic molecular biological and

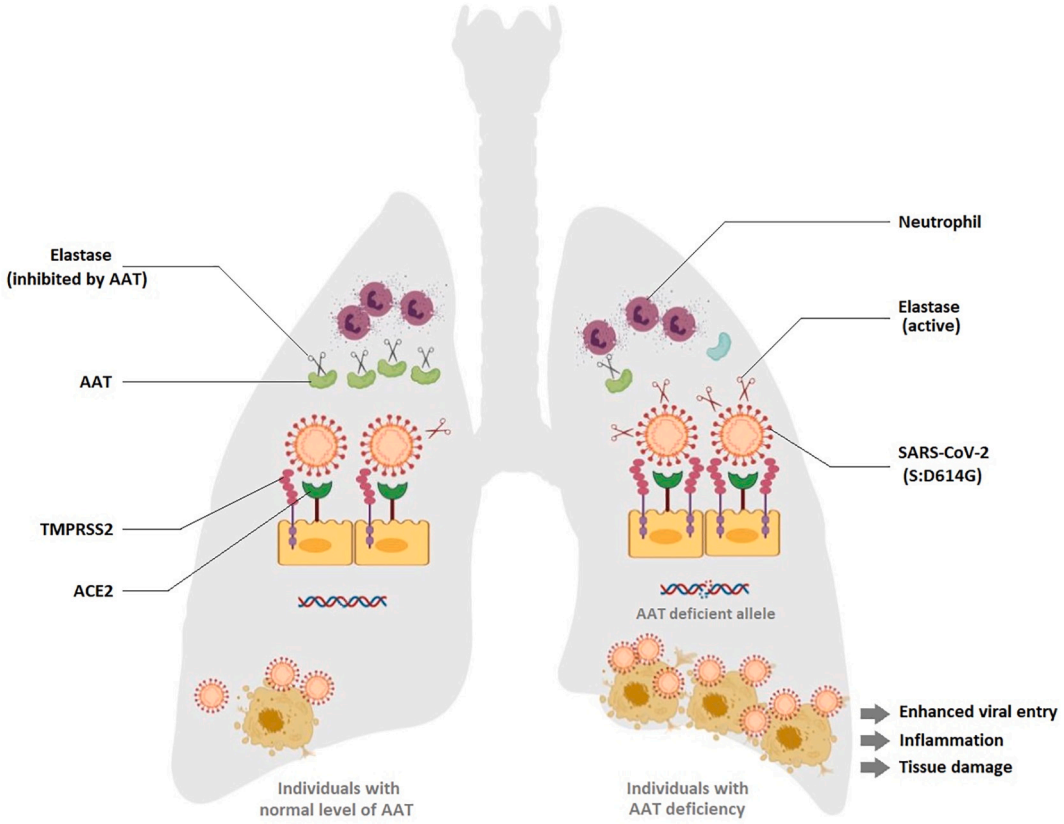

Fig. 5. A composite diagram depicting the major findings of this study
The SARS-CoV-2 614G subtype virus harbours an additional mutation induced spike protein cleavage site that is specific for neutrophil elastase. The 614G subtype virus is expected to get advantage in presence of active elastase. The schematic image of the human lung in this figure is partitioned into two components – (i) left – individuals with wildtype *SERPINA1* genotypes and normal levels of AAT, (ii) right - individuals with variant *SERPINA1* genotypes and AAT deficiency. Individuals with AAT deficiency will have higher level of active neutrophil elastase and will be more susceptible to infection by 614G subtype virus.

evolutionary model of virulence and spread of the evolved 614G subtype of the novel coronavirus SARS-CoV-2. In line with our findings, it was proposed in a recent review paper, AAT deficient allele frequency may have an impact of SARS-CoV-2 infection and mortality (Yang et al., 2021). Our findings are indicative of the possibility of using elastase inhibitors or AAT supplementation, already under consideration for treatment of chronic obstructive pulmonary disease and in COVID-specific clinical trial (NCT04495101), as therapy to prevent infection by the 614G subtype of SARS-CoV-2 (Ohbayashi, 2002; Hu et al., 2020; Mohamed et al., 2020; Németh et al., 2020; Strnad et al., 2020).

5. Conclusions

An additional neutrophil elastase cleavage site in Spike protein of SARS-CoV-2 was introduced by D614G mutation. Therefore, elevation of neutrophil elastase level at the site of infection will enhance the activation of Spike protein thus facilitating host cell entry for 614G, but not the 614D, subtype. The level of neutrophil elastase in the lung is modulated by its inhibitor α1-antitrypsin (AAT). AAT prevents lung tissue damage by elastase. However, many individuals exhibit genotype-dependent deficiency of AAT. AAT deficiency eases host-cell entry of the 614G virus, by retarding inhibition of neutrophil elastase and consequently enhancing activation of the Spike protein. AAT deficiency is highly prevalent in European and North-American populations, but much less so in East Asia. Therefore, the 614G subtype is able to infect and spread more easily in populations of the former regions than in the latter region. Our analyses provide a molecular biological and evolutionary model for the higher observed virulence of the 614G subtype, in terms of causing higher morbidity in the host (higher infectivity and higher viral load), than the non-mutant 614D subtype. Our work opens up possibility for considerations of AAT supplements in prevention of SARS-CoV-2 Spike 614G subtype virus.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2021.104760.

Availability of data and materials

The data underlying this article are available in 1000 Genomes (Data mined from http://grch37.ensembl.org/Homo_sapiens/Info/Index), GoNL (http://www.nlgenome.nl/), dbSNP(https://www.ncbi.nlm.nih.gov/snp/), SweFreq (https://swefreq.nbis.se/dataset/SweGen), deCode (https://www.decode.com/), alMENA (http://clingen.igib.res.in/almena/), jMorp (https://jmorp.megabank.tohoku.ac.jp/202001/), GnomAD (https://gnomad.broadinstitute.org/) and GenomeAsia 100K (https://genomeasia100k.org/).

Funding

This study was funded by a grant from National Supercomputing Mission, MeiTy, India.

Contributions

N.K.B conceived the study. P.P.M, A.B, S.M and N.K.B formulated the study design. C.B, C.D, A.G., A.K.S, and N.K.B analyzed data. All authors were involved in manuscript writing. A.B, N.K.B and P.P.M. edited the final manuscript. All authors read and approved the contents of the manuscript.

Declaration of Competing Interest

Authors would like to declare that we do not have any conflicts of interest in connection with the submitted work.

Acknowledgments

All viral sequence data related to SARS-CoV-2 was available from Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database and NIH Genbank resource. We acknowledge all the contributions of novel coronavirus sequencing data by members of the broader scientific community and GISAID for making everything accessible. The credits for sample origins and sequence submitting lab details are provided in Dataset 1. We are also grateful to Prof. Sharmila Sengupta, Dr. Amlan Das and Dr. Indranil Banerjee for providing some valuable comments on the manuscript. PPM acknowledges support to his National Science Chair by the Science & Engineering Board, Govt. of India.

References:

Abdulla, M.A., et al., 2009. Mapping human genetic diversity in Asia. Science 326 (5959), 1541–1545. https://doi.org/10.1126/science.1177074. American Association for the Advancement of Science.

Andersen, K.G., et al., 2020. The proximal origin of SARS-CoV-2. Nat. Med. 26 (4), 450–452. https://doi.org/10.1038/s41591-020-0820-9.

Auton, A., et al., 2015. A global reference for human genetic variation. Nature 526 (7571), 68–74. https://doi.org/10.1038/nature15393.

Bain, B., Seed, M., Godsland, I., 1984. Normal values for peripheral blood white cell counts in women of four different ethnic origins. J. Clin. Pathol. 37 (2), 188–193. https://doi.org/10.1136/jcp.37.2.188.

Belouzard, S., Chu, V.C., Whittaker, G.R., 2009. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proc. Natl. Acad. Sci. U. S. A. 106 (14), 5871–5876. https://doi.org/10.1073/pnas.0809524106. National Academy of Sciences.

Belouzard, S., Madu, I., Whittaker, G.R., 2010. Elastase-mediated activation of the severe acute respiratory syndrome coronavirus spike protein at discrete sites within the S2 domain. J. Biol. Chem. 285 (30), 22758–22763. https://doi.org/10.1074/jbc.M110.103275, 2010/05/27. American Society for Biochemistry and Molecular Biology.

Benetti, E., et al., 2020. ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population. Eur. J. Hum. Genet. 28 (11), 1602–1614. https://doi.org/10.1038/s41431-020-0691-z.

Biswas, N., Majumder, P., 2020. Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type. Indian J. Med. Res. 151 (5), 450–458. https://doi.org/10.4103/ijmr.ijmr_1125_20.

Blanco, I., et al., 2017. Alpha-1 antitrypsin Pi*SZ genotype: estimated prevalence and number of SZ subjects worldwide. Int. J. Chronic Obstruct. Pulmonary Dis. 12, 1683–1694. https://doi.org/10.2147/COPD.S137852. Dove Medical Press.

Borel, F., et al., 2018. Editing out five Serpina1 paralogs to create a mouse model of genetic emphysema. Proc. Natl. Acad. Sci. 115 (11) https://doi.org/10.1073/pnas.1713689115, 2788 LP – 2793.

Brantly, M., Nukiwa, T., Crystal, R.G., 1988. Molecular basis of alpha-1-antitrypsin deficiency. Am. J. Med. United States 84 (6A), 13–31. https://doi.org/10.1016/0002-9343(88)90154-4.

Burrows, J.A.J., Willis, L.K., Perlmutter, D.H., 2000. Chemical chaperones mediate increased secretion of mutant α1-antitrypsin (α1-AT) Z: A potential pharmacological strategy for prevention of liver injury and emphysema in α1-AT deficiency. Proc. Natl. Acad. Sci. 97 (4) https://doi.org/10.1073/pnas.97.4.1796, 1796 LP – 1801.

Bush, R.M., et al., 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol. Biol. Evol. United States 16 (11), 1457–1465. https://doi.org/10.1093/oxfordjournals.molbev.a026057.

Cao, Y., et al., 2020. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell Discov. 6 (1), 11. https://doi.org/10.1038/s41421-020-0147-1.

Chang, C.C., et al., 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience 4 (1), 7. https://doi.org/10.1186/s13742-015-0047-8.

Chen, Y., Li, L., 2020. SARS-CoV-2: virus dynamics and host response. Lancet Infect. Dis. 20 (5), 515–516. https://doi.org/10.1016/s1473-3099(20)30235-8.

Crowther, D.C., et al., 2004. Practical genetics: alpha-1-antitrypsin deficiency and the serpinopathies. Eur. J. Hum. Genet. 12 (3), 167–172. https://doi.org/10.1038/sj.ejhg.5201127.

Crystal, R.G., et al., 1989. The alpha 1-antitrypsin gene and its mutations. Clinical consequences and strategies for therapy. Chest 95 (1), 196–208. https://doi.org/10.1378/chest.95.1.196. United States.

Dahl, M., et al., 2005. The protease inhibitor PI*S allele and COPD: a meta-analysis. Eur. Respiratory J. Engl. 26 (1), 67–76. https://doi.org/10.1183/09031936.05.00135704.

Daniloski, Z., Guo, X., Sanjana, N.E., 2020. The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types. bioRxiv. https://doi.org/10.1101/2020.06.14.151357. Cold Spring Harbor Laboratory.

Dau, T., et al., 2015. Autoprocessing of neutrophil elastase near its active site reduces the efficiency of natural and synthetic elastase inhibitors. Nat. Commun. 6 (1), 6722. https://doi.org/10.1038/ncomms7722.

De Serres, F.J., 2002. Worldwide racial and ethnic distribution of alpha1-antitrypsin deficiency: summary of an analysis of published genetic epidemiologic surveys. Chest 122 (5), 1818–1829. https://doi.org/10.1378/chest.122.5.1818. United States.

De Serres, F.J., Blanco, I., Fernández-Bustillo, E., 2010. Ethnic differences in alpha-1 antitrypsin deficiency in the United States of America. Ther. Adv. Respiratory Dis. Engl. 4 (2), 63–70. https://doi.org/10.1177/1753465810365158.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20 (5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1. Elsevier.

Geoghegan, J.L., Holmes, E.C., 2018. The phylogenomics of evolving virus virulence. Nat. Rev. Genet. 19 (12), 756–769. https://doi.org/10.1038/s41576-018-0055-5.

Gómez-Mariano, G., et al., 2020. Liver organoids reproduce alpha-1 antitrypsin deficiency-related liver disease. Hepatol. Int. 14 (1), 127–137. https://doi.org/10.1007/s12072-019-10007-y.

Gorbalenya, A.E., et al., 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. 5 (4), 536–544. https://doi.org/10.1038/s41564-020-0695-z.

Grubaugh, N.D., Petrone, M.E., Holmes, E.C., 2020. We shouldn't worry when a virus mutates during disease outbreaks. Nat. Microbiol. 5 (4), 529–530. https://doi.org/10.1038/s41564-020-0690-4.

Gudbjartsson, D.F., et al., 2020. Spread of SARS-CoV-2 in the Icelandic population. N. Engl. J. Med. 382 (24), 2302–2315. https://doi.org/10.1056/NEJMoa2006100. Massachusetts Medical Society.

Hadfield, J., et al., 2018. NextStrain: Real-time tracking of pathogen evolution. Bioinformatics 34 (23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407.

Hoffmann, M., et al., 2020. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 181 (2), 271–280 e8. https://doi.org/10.1016/j.cell.2020.02.052.

Hu, J., et al., 2020. The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity and decreases neutralization sensitivity to individual convalescent sera. bioRxiv. https://doi.org/10.1101/2020.06.20.161323. Cold Spring Harbor Laboratory.

Hutchison, D.C.S., 1998. α1-Antitrypsin deficiency in Europe: geographical distribution of Pi types S and Z. Respir. Med. 92 (3), 367–377. https://doi.org/10.1016/S0954-6111(98)90278-5.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780. https://doi.org/10.1093/molbev/mst010.

Korber, B., et al., 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 182 (4), 812–827, 2020/07/03. Cell Press. e19. https://doi.org/10.1016/j.cell.2020.06.043.

Lamers, M.M., et al., 2020. SARS-CoV-2 productively infects human gut enterocytes. Science 369 (6499), 50–54. https://doi.org/10.1126/science.abc1669. American Association for the Advancement of Science.

Matsuyama, S., et al., 2005. Protease-mediated enhancement of severe acute respiratory syndrome coronavirus infection. Proc. Natl. Acad. Sci. U. S. A. 102 (35), 12543–12547. https://doi.org/10.1073/pnas.0503203102. National Academy of Sciences.

Matsuyama, S., et al., 2020. Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. Proc. Natl. Acad. Sci. 117 (13), 7001–7003. https://doi.org/10.1073/pnas.2002589117. National Academy of Sciences.

Mohamed, M.M.A., El-Shimy, I.A., Hadi, M.A., 2020. Neutrophil Elastase Inhibitors: A potential prophylactic treatment option for SARS-CoV-2-induced respiratory complications? Crit. Care 24 (1), 311. https://doi.org/10.1186/s13054-020-03023-0.

Muus, C., et al., 2020. Integrated analyses of single-cell atlases reveal age, gender, and smoking status associations with cell type-specific expression of mediators of SARS-CoV-2 viral entry and highlights inflammatory programs in putative target cells. bioRxiv. https://doi.org/10.1101/2020.04.19.049254. Cold Spring Harbor Laboratory.

Németh, T., Sperandio, M., Mócsai, A., 2020. Neutrophils as emerging therapeutic targets. Nat. Rev. Drug Discov. 19 (4), 253–275. https://doi.org/10.1038/s41573-019-0054-z.

Nguyen, L.-T., et al., 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32 (1), 268–274. https://doi.org/10.1093/molbev/msu300.

Ohbayashi, H., 2002. Neutrophil elastase inhibitors as treatment for COPD. Expert Opinion Investig. Drugs. Engl. 11 (7), 965–980. https://doi.org/10.1517/13543784.11.7.965.

Ozono, S., et al., 2020. Naturally mutated spike proteins of SARS-CoV-2 variants show differential levels of cell entry. bioRxiv. https://doi.org/10.1101/2020.06.15.151779. Cold Spring Harbor Laboratory.

Rawlings, N.D., et al., 2017. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. 46 (D1), D624–D632. https://doi.org/10.1093/nar/gkx1134.

Roy, A., Kucukural, A., Zhang, Y., 2010. I-TASSER: A unified platform for automated protein structure and function prediction. Nat. Protoc. 5 (4), 725–738. https://doi.org/10.1038/nprot.2010.5.

Shang, J., et al., 2020. Cell entry mechanisms of SARS-CoV-2. Proc. Natl. Acad. Sci. 117 (21), 11727–11734. https://doi.org/10.1073/pnas.2003138117. National Academy of Sciences.

Shapira, G., Shomron, N., Gurwitz, D., 2020. Ethnic differences in alpha-1 antitrypsin deficiency allele frequencies may partially explain national differences in COVID-19 fatality rates. FASEB J. https://doi.org/10.1096/fj.202002097. John Wiley & Sons, Ltd, n/a(n/a).

Shirato, K., Kawase, M., Matsuyama, S., 2018. Wild-type human coronaviruses prefer cell-surface TMPRSS2 to endosomal cathepsins for cell entry. Virology 517, 9–15. https://doi.org/10.1016/j.virol.2017.11.012.

Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 e30494.

Song, J., et al., 2012. PROSPER: An Integrated Feature-Based Tool for Predicting Protease Substrate Cleavage Sites. PLoS One 7 (11). https://doi.org/10.1371/journal.pone.0050300. Public Library of Science. e50300.

Song, J., et al., 2018. PROSPERous: High-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinformatics 34 (4), 684–687. https://doi.org/10.1093/bioinformatics/btx670.

Stawiski, E.W., et al., 2020. Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility. bioRxiv. https://doi.org/10.1101/2020.04.07.024752. Cold Spring Harbor Laboratory, p. 2020.04.07.024752.

Strnad, P., McElvaney, N.G., Lomas, D.A., 2020. Alpha1-Antitrypsin Deficiency. N. Engl. J. Med. 382 (15), 1443–1455. https://doi.org/10.1056/NEJMra1910234. Massachusetts Medical Society.

Tajuddin, S.M., et al., 2016. Large-Scale Exome-wide association analysis identifies loci for white blood cell traits and pleiotropy with immune-mediated diseases. Am. J. Hum. Genet. 99 (1), 22–39. https://doi.org/10.1016/j.ajhg.2016.05.003.

Van Dorp, L., et al., 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. 83, 104351. https://doi.org/10.1016/j.meegid.2020.104351.

Verity, R., et al., 2020. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect. Dis. 20 (6), 669–677. https://doi.org/10.1016/S1473-3099(20)30243-7.

Volz, E.M., Koelle, K., Bedford, T., 2013. Viral Phylodynamics. PLoS Comput. Biol. 9 (3) https://doi.org/10.1371/journal.pcbi.1002947. Public Library of Science. e1002947.

Wagner, C., et al., 2020. Comparing viral load and clinical outcomes in Washington State across D614G mutation in spike protein of SARS-CoV-2. Available at. https://github.com/blab/ncov-D614G.

Walls, A.C., et al., 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181 (2), 281–292 e6. https://doi.org/10.1016/j.cell.2020.02.058.

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38 (16) https://doi.org/10.1093/nar/gkq603 e164–e164.

Wang, Q., et al., 2020a. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell 181 (4), 894–904 e9. https://doi.org/10.1016/j.cell.2020.03.045.

Wang, W., et al., 2020b. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA 323 (18), 1843–1844. https://doi.org/10.1001/jama.2020.3786.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38 (6), 1358–1370. https://doi.org/10.2307/2408641. Society for the Study of Evolution, Wiley.

Xia, S., et al., 2019. A pan-coronavirus fusion inhibitor targeting the HR1 domain of human coronavirus spike. Sci. Adv. 5 (4) https://doi.org/10.1126/sciadv.aav4580. American Association for the Advancement of Science. eaav4580.

Yang, C., et al., 2021. 'α1-Antitrypsin deficiency and the risk of COVID-19: an urgent call to action. Lancet Respir. Med. https://doi.org/10.1016/S2213-2600(21)00018-7. Elsevier.

Zhang, L., et al., 2020. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat. Commun. 11 (1), 6013. https://doi.org/10.1038/s41467-020-19808-4.

Zhu, N., et al., 2020. A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. 382 (8), 727–733. https://doi.org/10.1056/NEJMoa2001017. Massachusetts Medical Society.

Ziegler, C.G.K., et al., 2020. SARS-CoV-2 Receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. Cell 181 (5), 1016–1035 e19. https://doi.org/10.1016/j.cell.2020.04.035.

Zullinger, E.M., et al., 1984. Fitting Sigmoidal Equations to Mammalian Growth Curves. J. Mammal. 65 (4), 607–636. https://doi.org/10.2307/1380844.