The SARS-CoV-2 Spike Protein D614G Mutation shows Increasing Dominance and May Confer a Structural Advantage to the Furin Cleavage Domain.

Leyan Tang[1], Allison Schulkins[1], Chun-Nan Chen[1], Kurt Deshayes[2], & John S. Kenney[2]

[1]Single Cell Technology, Inc. 6280 San Ignacio Ave., Suite E, San Jose, CA 95119. [2]Antibody Solutions, 3033 Scott Blvd., Santa Clara, CA 95054. Correspondence: chunnan.chen@singlecelltechnology.com, jkenney@antibody.com.

Keywords: SARS-CoV-2, spike protein, D614G mutation, genotype distribution, furin cleavage site, secondary structure, sequence analysis, homology modeling.

Abstract

We analyzed the SARS-CoV-2 spike (S) protein amino acid sequence extracted from 11,542 viral genomic sequences submitted to the Global Initiative on Sharing All Influenza Data (GISAID) database through April 27, 2020. Consistent with prior reports, we found a major S protein mutation, D614 to G614, that was represented in 56% of all the analyzed sequences. All other mutations combined were less than 10%. After parsing the data geographically, we found most of the Chinese patient samples showed D614 (97%). By contrast, most patient samples in many European countries showed G614 (51 to 88%). In the United States, the genotypic distribution in California and Washington was similar to Asian countries, while the distribution in other US states was comparable to Europe. We observed a dramatic increase in the frequency of G614 over time in multiple regions, surpassing D614 when both were present, suggesting G614 S protein virus outcompetes D614 S protein virus. To gain insight into the consequences of the D614G mutation, homology modeling using a multi-template threading mechanism with *ab initio* structural refinement was performed for a region of the S protein (S591 to N710) spanning the D614G mutation and the S1 furin cleavage site. Molecular models of this region containing D614 or G614 revealed a major difference in secondary structure at the furin domain (RRARS, R682 to S686). The D614 model predicted a random coil structure in the furin domain whereas the G614 model predicted an alpha helix. Critical residues in the cleavage domain of G614 model were found to better align with the PDB structure of a furin inhibitor. Thus, homology modeling studies suggest a potential mechanism whereby the D614G mutation may confer a competitive advantage at the furin binding domain that may contribute to the rise of the D614G SARS-CoV-2 mutant.

Introduction

The novel coronavirus SARS-coronavirus 2 (SARS-CoV-2), a strain related to SARS-CoV that caused a pandemic in 2002-2003 [1], is the etiologic agent of COVID-19 [2]. Several factors have contributed to the rapid spread of SARS-CoV-2 throughout the world. These include a lack of pre-existing immunity, pre-symptomatic transmission, high transmissibility, and delays in responses by government agencies[3]. Of intense scientific interest is how structural features of the virus contribute to its transmission and virulence in humans.

The SARS-CoV-2 spike (S) glycoprotein is an important structural component heavily decorating the viral surface. It mediates the virus entry into host cells. The S protein contains an ectodomain, a single-pass transmembrane anchor, and an intracellular tail [4]. The ectodomain consists of a receptor-binding subunit S1 and a membrane-fusion subunit S2. The S1 subunit binds to angiotensin-converting enzyme 2 (ACE2) on the host cell surface. After proteolytic cleavage within the S1 subunit, the S2 subunit undergoes a structural change enabling viral and host membrane fusion followed by virus entry into host cells [5].

Cleavage of the S1 subunit by host proteases, including furin and transmembrane protease, serine 2 (TMPRSS2), is essential for virus infection [5–8]. The S1 subunit of SARS-CoV-2 virus has a polybasic furin cleavage domain absent in related human and bat SARS-CoV viruses [9], including bat RaTG-13, the coronavirus with the highest S protein homology to SARS-CoV-2 [10]. Furin cleavage plays a key role in the process of infection of SARS-CoV-2, MERS-CoV, and other pathogens such as Anthrax and Ebola [6, 7, 11]. Mutating the polybasic furin cleavage domain of SARS-CoV-2 reduces pseudo-virus infectivity (5). Furin is an extracellular serine endoprotease that is ubiquitously expressed on host cells in a variety of tissues, including the lung [11]. Furin provides and essential mechanism for S1 cleavage that leads to subsequent S protein processing and entry of SARS-CoV-2 into host cells. [5, 6, 12].

Given its critical role in both viral infectivity and as an antibody target, we evaluated the mutational landscape of the S protein amid the global pandemic. Our primary goal was to identify major S mutants that may impact discovery of antibody therapeutics and vaccines. Here we describe an analysis of viral genomes submitted to the Global Initiative on Sharing All Influenza Data (GISAID) [13], a public repository of viral genome sequences designed to share vital information about the virus quickly. Our effort was aided by GISAID's linking of sequences to geographic information and sampling dates. We used a database analysis suite we had initially developed to analyze and parse thousands of gene sequences of antibodies. After examining the S protein amino acid sequence from over 11,000 sequenced viral genomes, we identified a prevalent mutation, D614G, in a large clade of the global COVID-19 cases to date affirming what others have observed [14]. In addition, we found contrasting geographical distributions of the D614G mutation among different countries and regions. To seek the origins of such varying distributions of S protein genotypes, we analyzed the daily cumulative genotype counts over time by countries and US states. We observed a dramatic increase in the occurrence of the G614 over time in multiple regions, surpassing D614 when both were present.

To gain insight into the structural D614G mutation on the S protein, we applied homology modeling using a multi-template threading mechanism with *ab initio* structural refinement [15, 16] to the region where the D614 mutation resides. Using a region from AA S591 to N710, homology modeling predicted a long-distance conformational change in the furin binding domain at AA R682 to S686 induced by the substitution of G for D at AA position 614 in the model. While this result was unexpected, there are well documented observations of long-distance conformational changes in proteins induced by a single AA substitution [17–20]. The G614 model predicted an altered secondary structure in the furin cleavage domain and better alignment with a furin active site inhibitor compare to the D614 model. Taken together, our observations point to an important role for the D614G mutation in COVID-19 pathogenesis.

Results

The D614G spike mutation dominates in a rising clade.

Based on 12,845 viral genomes downloaded on April 28, 2020 from the GISAID database, a public repository of the SARS-CoV-2 viral genome and associated data [13], we identified and extracted the DNA sequence corresponding to the S protein from each viral genome. Subsequently, we analyzed the DNA sequence of S protein from each viral genome and translated into AA sequence (see Materials and Methods). Shorter or incomplete sequences and those that yielded too many stop codons were removed. The presence of many stop codons is most likely a result of lower base-calling accuracy with insertions and deletions of bases causing a frameshift. A total of 12,024 sequences were used for geographical and temporal distribution analysis. Translated sequences were compared to the S protein amino acid sequence of the Wuhan reference strain (GenBank accession number: NC_045512). We identified and removed those mutations that occurred only once or twice in the entire population of sequences, because they are most likely to reflect sequencing errors. A total of 11,542 sequences were used to compute the frequency of different mutations.

Spectrum of mutations in the SARS-CoV-2 S protein amid pandemic

Along the entire S protein amino acid sequence, we identified mutations at a total of 107 positions. We noted the S protein amino acid sequence for 4,096 (35.49%) of the viral genomes sampled are identical to the Wuhan reference strain (Fig 1A). We will refer to the sequences carrying such an S protein as “Wuhan S” for subsequent sections. Among the mutations examined, we identified one mutation, a single amino acid substitution D614G (Fig 1A), occurring in 6,477 sequences (56.12%). This mutation is well above all other mutations combined (8.39%) and surpasses the originating S protein sequence (35.49%) from the Wuhan reference strain. This mutation has been reported by others [14, 21, 22] and the viral genome sequences containing this S protein D614G mutation has been called the G clade according to the GISAID database or categorized as the A2a clade by Nextstrain.org (https://nextstrain.org/ncov/global?branchLabel=aa&c=clade_membership&d=tree,entropy&p=full). The next most prevalent mutations are P1263L (0.67%), L5F (0.51%), D936Y (0.37%), and G1124V (0.37%). A total of 102 remaining mutations comprise 6.47% of the sampled genomes in our study. These are scattered mutations of very low frequency (<0.3%), most are single amino acid substitutions, and they do not include the D614G mutation. In addition, there are also insertions and deletions (up to 4) of amino acids, but those represent a very small fraction (<0.05%) of the sequences analyzed (Supporting Information S1). The majority of the changes in the Wuhan S protein sequence are the D614G mutation. Both the Wuhan S and D614G mutant viruses appear to have stable S protein sequences.

Geographical distribution of Wuhan S vs. D614G

We examined the geographical distribution of D614G containing viral genomes and found it is distributed mainly in Europe and parts of North America, South America, Oceania, and Asia (Figs 1B, C, D, & E). The prevalent sequences found in patient samples were Wuhan S and D614G and are represented in Figure 1 as blue or red, respectively. All other sequences were combined and are represented as *Other* (gray). Many sequences represented as *Other* in Figure 1 may be due to sequencing errors.

United States.

There appears to be distinct distribution patterns between the West Coast vs. the East Coast (Fig 1B). A majority of the patient samples in three West Coast states, Washington, Oregon, and California, show Wuhan S. On the other hand, most of the samples from the rest of the country show D614G with a few scattered exceptions. Interestingly, all 25 samples from the Grand Princess cruise ship (Cruise ship, Fig 1B) show D614G, with a majority carrying Wuhan S protein sequence.

Europe

Patient samples from all of the countries we examined from this region show D614G as the dominant sequence with the exception of Wales (Fig 1C). A significant fraction (35% to 48%) of the samples from England, Wales, Netherland, Scotland, and Spain show Wuhan S.

Asia and Oceania

With Wuhan as the epicenter of the coronavirus outbreak, China ordered the lockdown of the city on January 29, 2020 followed by restrictions for the entire country. Viral sequences from most of the Chinese and all South Korean samples show the Wuhan S protein sequence (Figs 1D). Other countries including Japan, Taiwan, India, and Australia have significant number of samples showing D614G (18% to 46%, Figs 1D & E).

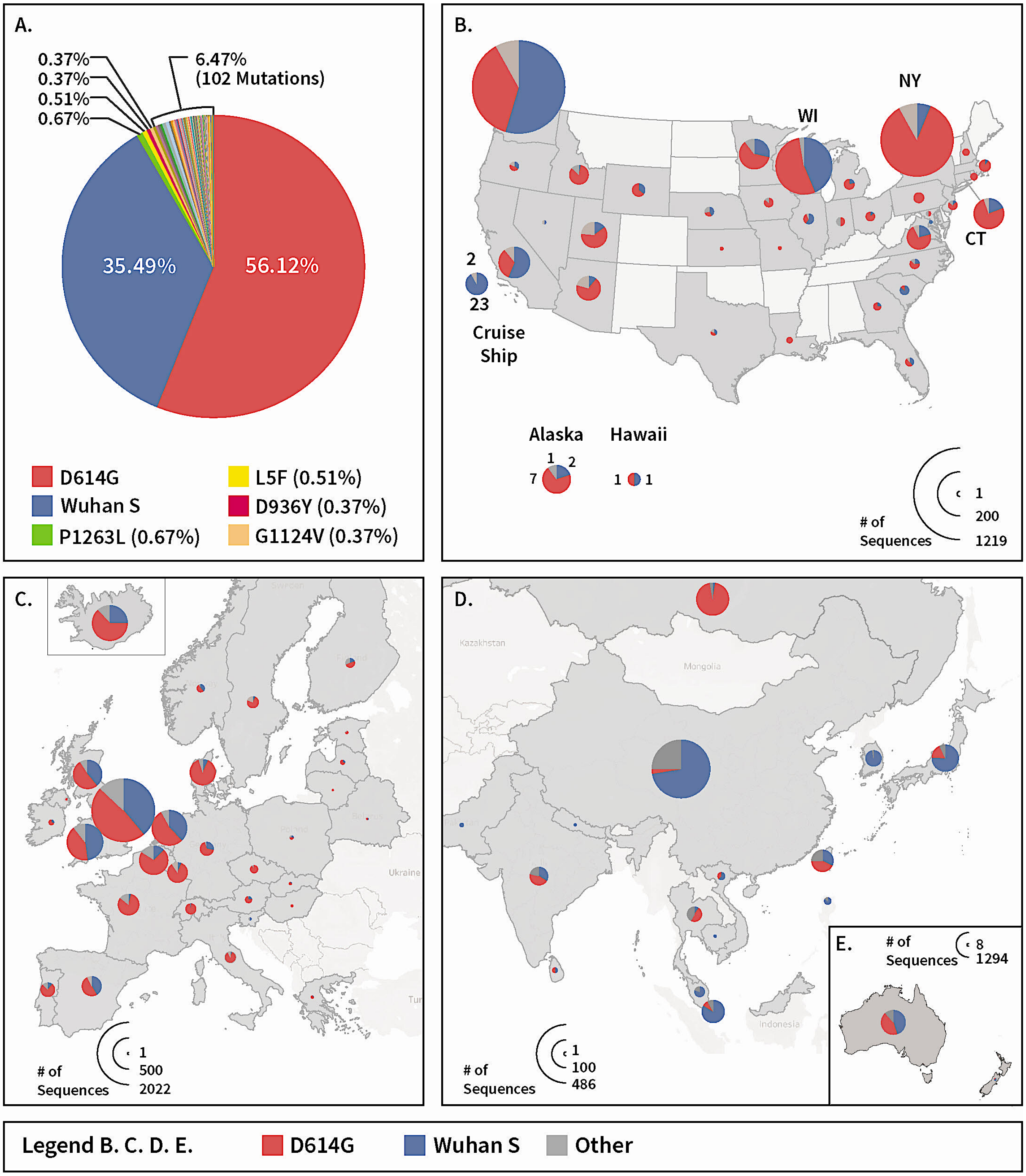


Fig 1. S protein mutations, their frequencies, and geographical representations between D614G and Wuhan S in different parts of the globe. A. Frequencies of the S protein mutations in the SARS-CoV-2 genomes sampled globally from GISAID as of April 28, 2020. Mutations with the frequency >0.3% are identified. **B.** Frequency of D614G (red), Wuhan S (blue), and other mutations (gray) in the United States and the Grand Princess cruise ship (Cruise ship). The diameter of each circle scales with the number of sequences sampled for each location. The scale is shown by the inset half circles. **C**. Frequency of D614G, Wuhan S, and other mutations in various countries in Europe. **D.** Frequency of D614G, Wuhan S, and other mutations in various countries in Asia. **E.** Frequency of D614G, Wuhan S, and other mutations in various countries in Oceania. Regions with less than 1,000 sequences were not represented in B, C, D, & E.

The temporal progression of G614 shows a rapid rise in March in various regions but with different dynamics

Europe, Asia, and Oceania

We studied the distribution of Wuhan S vs. D614G virus over time and found that the first Chinese patient who showed the D614G mutation was sampled on January 24, 2020 (virus name: hCoV-19/Zhejiang/HZ103/2020, GISAID accession ID: EPI_ISL_422425, see Supporting Information S5) but the mutation spread to very few patients in China. The second Chinese patient who showed the D614G mutation was sampled on January 28, 2020 (virus name: hCoV-19/Shanghai/SH0014/2020, GISAID accession ID: EPI_ISL_ 416327). This is the same day of the first sampling of the D614G mutation in Europe from a German patient (virus name: hCoV19/Germany/ BavPat1/2020, GISAID accession ID: EPI_ISL_406862). The DNA sequence encoding the S protein from both the second Chinese and the German patient were the same, though small scattered segments of DNA sequence from the second Chinese patient were missing base-calls.

The spread of the D614G virus was most pronounced in Europe. Note that sample collection started to pick up only in March in Europe (Fig 2, day 60 on the x axis) and at the same time, sampling waned in China, potentially influencing the reporting of the G614 virus in China. The rise of D614G virus is most obvious in Europe (Fig 2). D614G became the predominant S protein sequence sampled over time in Belgium, France, Luxembourg, Iceland, and Denmark (62.8% to 93.1%). Whereas in England, Scotland, Spain, and Netherlands, the difference between Wuhan S and D614G is less significant though D614G still ultimately becomes the majority (54.7% to 63.8%). In Wales, the rate of increase and the number of sequences at the last sampling of Wuhan S and D614G were equivalent. Russia seems to be unique with almost all samples showing D614G and very few samples with Wuhan S (<3%).

As noted above, sample collection started much earlier for China and the Wuhan S protein sequence was established early and dominated (Fig 2). However, there is very little sample collection after day 60. Japan started sampling early on, approximately one week after China, and only the Wuhan S genotype was observed in the initial sampling. Interestingly, sample collection in Japan stopped at day 44 but when resumed on day 66, most of the new sequences were D614G.

The United States and Canada

Temporal analysis was performed for Canada and the US states with at least 50 sequences submitted (Figs 2 and 3). East Coast states such as Connecticut, New York, and Virginia all show a rise in D614G to a higher frequency (66.1% to 85.8%) compared to Wuhan S once sample collection started in March (Fig 3). Similar patterns were observed for Arizona and Utah (50 to 68.3%). Minnesota and Wisconsin also show a D614G dominance, but with a significant frequency of Wuhan S sequences (28.6% to 43.7%). However, the period of collection is short for Minnesota (<20 days). Two West Coast states, Washington and California, have majority Wuhan S sequences over D614G (64.1% to 70.5%). However, Wuhan S appeared early, 46 and 42 days for Washington and California, respectively, before D614G was detected on the West Coast. Similarly, Canada showed more Wuhan S cases (Fig 2), but the first occurrence of Wuhan S was 37 days before D614G was detected.

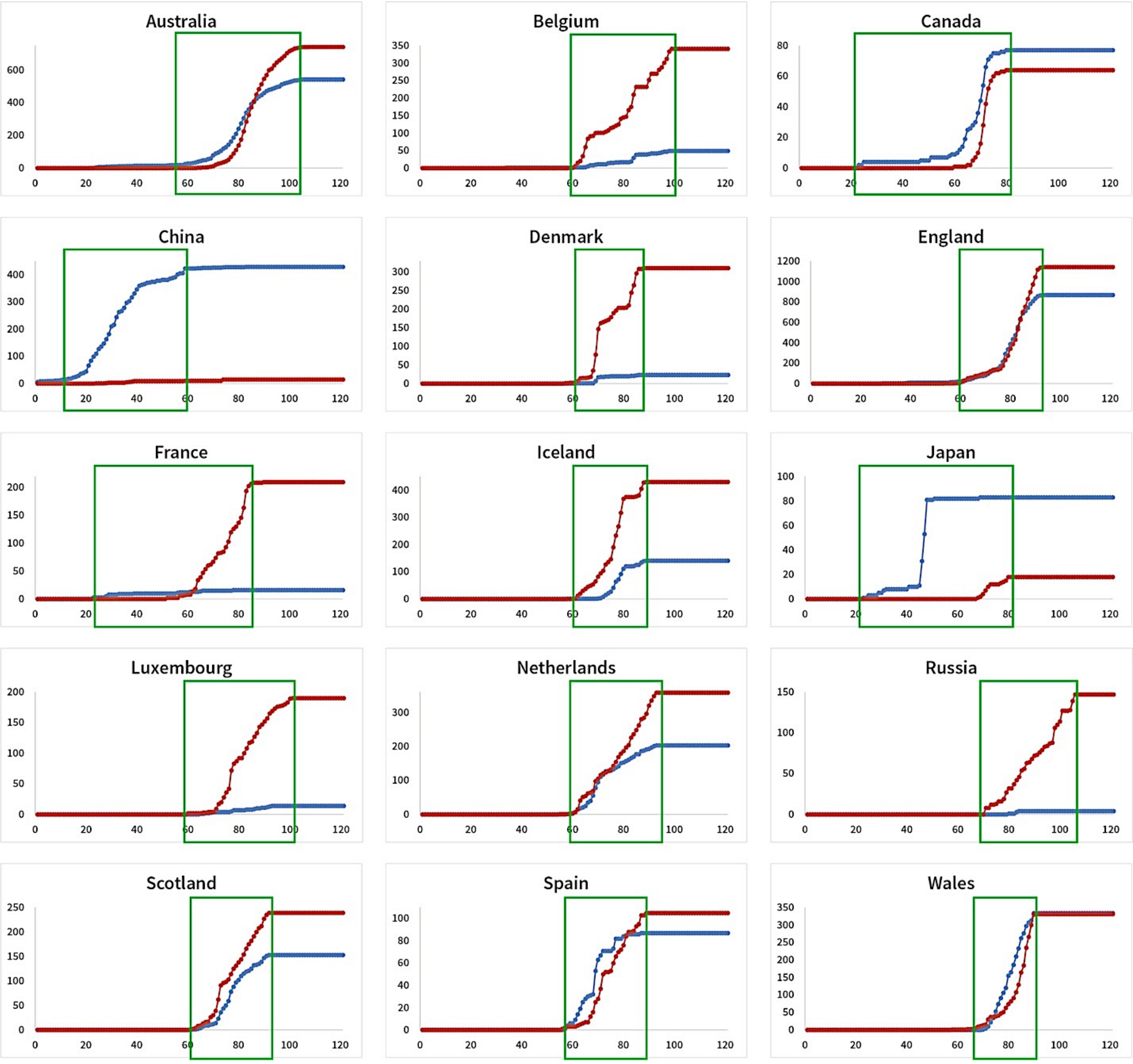


Fig 2. Daily cumulative genotype counts for countries with at least 100 sequences. Red is D614G and blue is Wuhan S. On the X axis, day 1 is January 1, 2020 and day 60 is February 29, 2020. The Y axis is the cumulative number of sequences for the specified S protein genotype. The green box highlights the period when sequenced samples were collected.

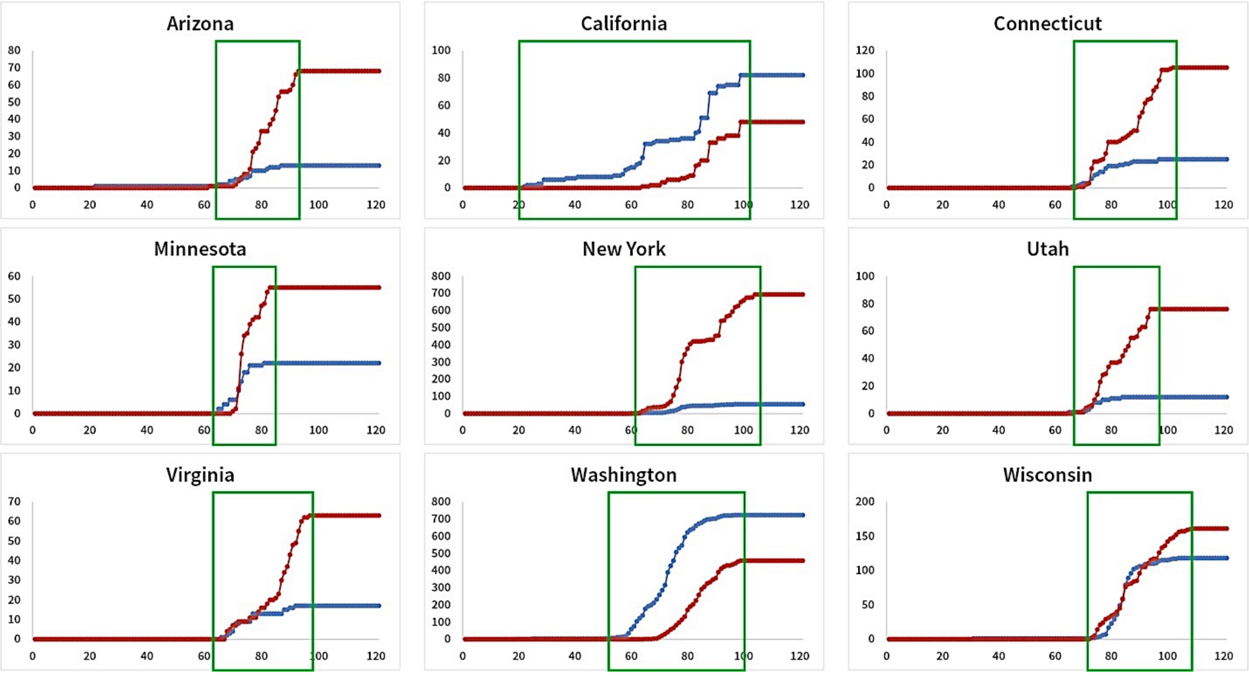
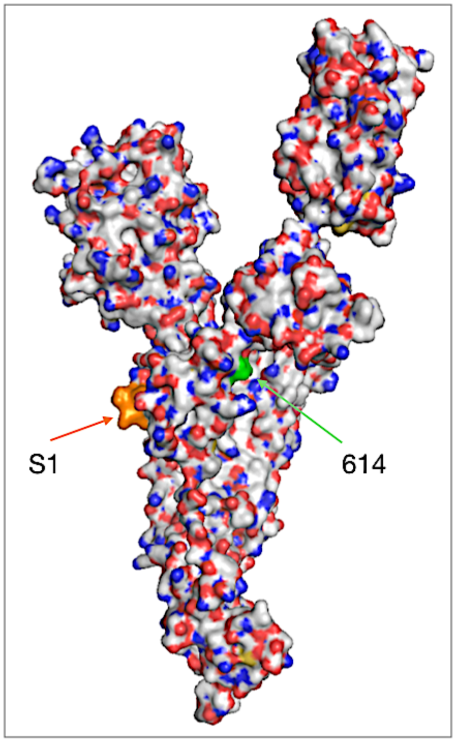


Fig 3. Daily cumulative genotype counts for US states with at least 50 sequences. Red is D614G and blue is Wuhan S. On the X axis, day 1 is January 1, 2020 and day 60 is February 29, 2020. The Y axis is the cumulative number of sequences for the specified S protein genotype. The green box highlights the period when sequenced samples were collected.

The D614G mutation offers a competitive structural advantage

The observed rise in dominance of sequences containing the D614G mutation over the Wuhan S sequence prompted an examination of the potential consequences of the D614G mutation.



The D614G mutation is proximal to the S1 cleavage domain.

A representation of the monomer of SARS-CoV-2 spike protein taken from PDB ID: 6VSB [4] is shown in Figure 4. The full Cryo-EM structure of the trimer has only medium resolution. It is important to note that the PDB:6VSB structure is missing aa 673-686. Thus, the structure of the furin cleavage domain, 682-RRARS-686, is not known. However, based upon the location of amino acids bracketing the furin cleavage domain the D614G mutation would be approximately 24 angstroms from the S1 cleavage domain in this representation.

Fig 4. Surface representation of the SARS-Cov2-S protein. Space filling representation showing the general location of the furin cleavage domain (orange) and D614 (green).

The D614G mutation is predicted to induce a conformational change in the Furin cleavage domain.

The protein region from AA S591 to N710 was modeled using I-TASSER [15, 16] multiple template threading methodology (see Materials and Methods). Modeling was performed using SARS-CoV spike protein PDB ID: 5X58 [23] as the template. This structure is the highest resolution PDB structure with the highest homology to SARS-CoV-2. The AA region of 591 to 710 was chosen for analysis since it contains both the site of the D614G mutation and the furin cleavage site with defined secondary structure in the vicinity of the 2 sites. The exact same threading method was used for both the Wuhan S and D614G mutation sequences with the only change in the sequence of the 2 models being D614 to G614, referred to below as D614 and G614, respectively. The Template Modeling Scores (TM-Score) for the D614 and G614 models were 0.59 and 0.58, respectively. A model with a TM-Score > 0.5 indicates a model of correct topology [24].

Figure 5A shows a ribbon representation of the AA 591 to 710 region of the D614 and G614 models. Upon inspection of the 2 structures, we observe that the five residues in the furin polybasic cleavage domain (RRARS, R682– S686) are 22 angstroms from AA614 in these models. Comparison of the two models clearly demonstrates the furin cleavage domain has the largest change in secondary structure within AA 591 to 710. The G614 model shows a more compact alpha helical structure whereas the D614 model has a random coil. This finding is supported by analysis using the Kabsch and Sander algorithm [25] to define the secondary structure of proteins (KSDSSP) shown in Figure 5B. KSDSSP uses the coordinates of the backbone atoms of a protein to determine which residues are in alpha helices and beta strands based on hydrogen bonding interactions. An alpha helix in the furin cleavage domain is predicted for G614 but is absent for D614. Thus, I-TASSER and KSDSSP both predict an alpha helical structure in the furin cleavage domain for G614 that was absent for D614. Additionally, I-TASSER and KSDSSP predict no other major changes in secondary structure. It is compelling that modeling showed the greatest conformation change induced by the D614G mutation in a functionally relevant location (e.g. furin polybasic peptide domain). Given that the cryo-EM structure lacks a 13 amino acid stretch from 673-686 of the S protein, our models are the first representation of the furin cleavage domain of the D614 and G614 spike proteins of SARS-CoV-2. Exchange of the acidic side chain of aspartate at position 614 for the hydrogen atom of glycine creates changes in the hydrogen bonding network forming an alpha helical structure that is 22 angstroms from AA 614.

The D614G mutation changes the orientation of critical residues in the furin cleavage domain.

The minimal sequence requirement for furin catalysis of a cleavage site is RXXR*X where * indicates the site of amide hydrolysis C-terminal to the recognition site [11]. For the S1 domain of SARS CoV-2 it has been demonstrated that furin cleaves between R685 and S686, at the C-terminus of the consensus furin recognition domain 682-RRAR*S -686 [6]. Catalysis is accomplished by placement of arginines at the P1 (R685) and P4 (R682) positions of the substrate into the corresponding binding pockets of furin [26].

The cleavage domains for D614 and G614 were aligned using Pymol with the structure of an active site inhibitor (Meta-guanidinomethyl-phenylacetyl-RVR). A high resolution crystal structure of the inhibitor bound into the furin catalytic domain PDB ID: 5XJH [26] was used for the alignment. Comparing the alignments of the D614 vs. G614 cleavage domain to the inhibitor (Fig 6) it is observed that the orientation of the G614 P1 and P4 residues is closer to the orientation of the P1 and P4 residues of the inhibitor structure compared to D614. The closer a substrate is to the transition state orientation, the lower the energetic requirement for cleavage [26, 27]. Therefore, the cleavage of the G614-containing S protein is proposed to be energetically favored over the D614-containing S protein.

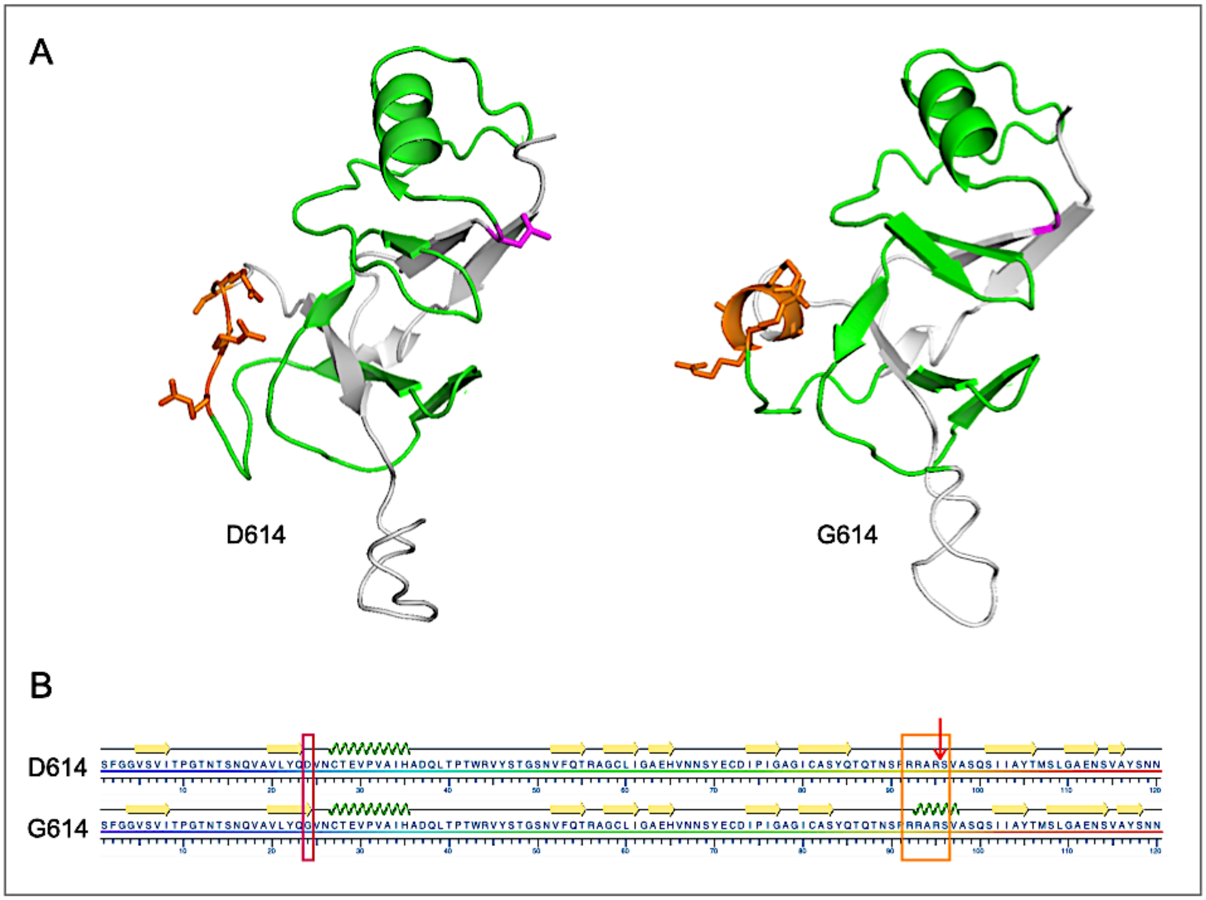

A
D614
G614
B
D614
SFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNN
G614
SFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNN
10
20
30
40
50
60
70
80
90
100
110
120

Fig 5. Structural representations of D614 and G614 SARS-CoV-2 S protein from S591 through N710. A. I-TASSER-derived models using SARS-CoV spike protein PDB ID: 5X58 as the template. The furin cleavage domain 682-RRARS-686 is shown in orange and D614 (left) or G614 (right) are shown in magenta. Residues between 614 and the cleavage domain are shown in green. **B.** Results of the KSDSSP algorithm for defining secondary structure of proteins. The D614G mutation is highlighted with a magenta box, the furin cleavage domain with a orange box and the cleavage site by the arrow. Beta sheet or alpha helix secondary structures are indicated by yellow arrows or green corkscrew, respectively.

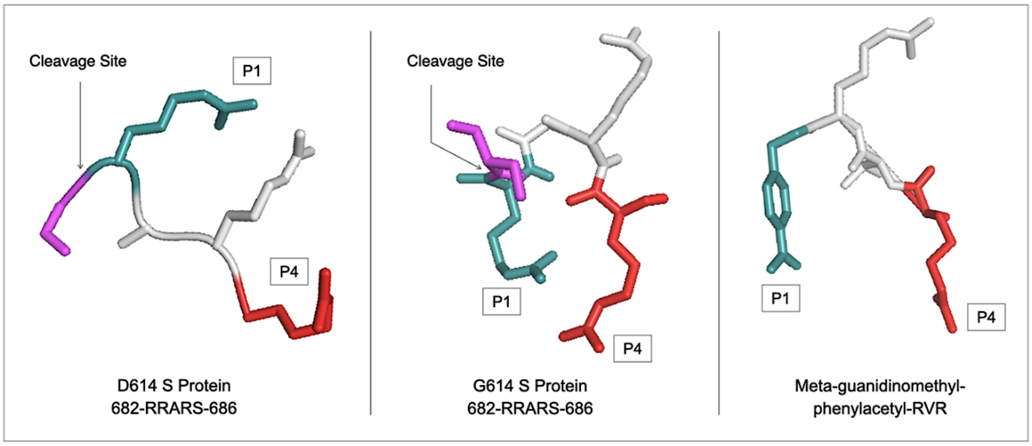


Fig 6. Alignment of models of the D614 and G614 cleavage domains with an active site inhibitor, Meta-guanidinomethyl-phenylacetyl-RVR. The P1 and P4 residues of the D614 and G614 cleavage domains and inhibitor are shown in teal and red, respectively. The cleavage site serine residue is shown in magenta.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2020 doi:10.20944/preprints202005.0407.v1

The D614G mutation aligns key residues more favorably within the active site of furin.

Using the model of furin and the inhibitor (PDB ID: 5JXH), the inhibitor aligned structures of D614 or G614, R682– S686, were substituted for the inhibitor in the furin catalytic domain using Pymol (Fig 7). G614 puts the necessary guanadino side chains, P1 (R685) and P4 (R682), into their respective enzyme pockets whereas the D614 only offers one or the other. This scenario would predict a lower energetic barrier of the G614-containing S protein for cleavage by furin. Thus, the predicted long-range conformational change induced by the D614G mutation more favorably aligns critical P1 and P4 residues into the furin binding pockets. The D614G mutation may enhance cleavage of the S1 site on the S protein furin and thereby give an advantage to the G614 mutant for S protein processing.

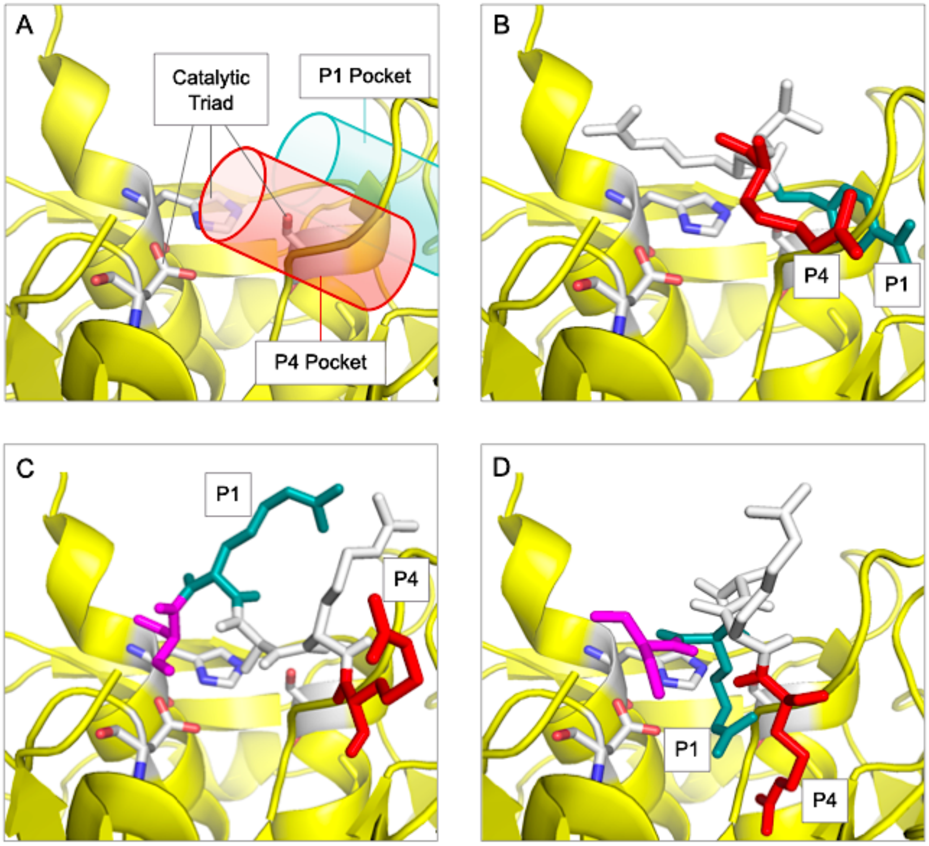


Fig 7 Alignment of D614 and G614 in the catalytic domain of furin. A. The catalytic domain of furin is shown in yellow (panels **A-D**) with the locations of the P1 (green) and P4 (red) binding pockets shown as cylinders. **B.** Alignments of Inhibitor, Meta-guanidinomethyl-phenylacetyl-RVR. **C.** D614 682-RRARS-686. **D.** G614 682-RRARS-686.

Discussion

Consistent with reports from others, we have identified a major mutation in the S protein gene at position 614. Interestingly, this mutant S protein appeared early in China (January 24th, 2020), and 4 days later appeared in Germany, but it was not widespread in the data until March where massive infection hit regions like Europe and the Northeastern United States. It is very likely that the virus with the D614G mutation that appeared in the first Chinese and German patients are related. The virus that infected the German patient coincided with the beginning of the spread throughout Europe.

We have observed different frequencies of D614G mutation vs. Wuhan S in various parts of the world. While we have analyzed over 11,000 sequences, there are many factors that could affect the precision of our analysis. These factors include, but are not limited to, availability of testing, submission of sequences, implementation of quarantine practices, and sampling time periods. In our calculations of the daily cumulative genotype counts, we excluded those regions with less than 100 sequences and US states with less than 50 sequences. Nonetheless, clear patterns were observed in the frequencies of the Wuhan S vs. D614G S protein as time progressed.

The Wuhan reference sequence was first uploaded early in China and quickly established itself (Fig 2). Wuhan S also appeared early in Japan, California, Washington, and Canada. As of the last sample sequenced in the dataset for these areas, Wuhan S remained the predominant genotype. Frequencies for the Wuhan S and D614G sequences rose to equivalent levels in Wales at which point sampling ceased.

Since collection slowed down after March 1 in China, when D614G began appearing in most countries, it is difficult to conclude whether D614G may be on the rise in China. Japan started collection early on day 20 where Wuhan S rose as the dominant strain. Japan essentially stopped sampling around day 47 but began collecting samples for sequencing again on day 67. The new sampling shows a greater increase in the D614G mutant compared to Wuhan S.

When the virus appeared in Europe, the D614G genotype spread very quickly and rapidly dominated that area. Wuhan S was also found in all European countries at significant levels. However, in multiple countries with similar sampling periods, and where D614G and Wuhan S arose at similar times, D614G became the dominant genotype. In France and Australia, where Wuhan S appeared at least 30 days before D614G, D614G became the predominant genotype at the end of the sampling period.

With regard to the United States, our data is generally consistent with reports [28] that the virus was introduced on the East Coast from Europe and the West Coast from China. The earliest sequenced samples in California were predominantly Wuhan S, while the earliest sequenced samples in New York were predominantly D614G. Over time, both sequences appeared in the data across the United States. At the end of the sampling period, California and Washington were predominantly Wuhan S. However, Wuhan S had a much earlier start than D614G. All 25 S protein sequences sampled from the Grand Princess cruise ship showed D614 and most were identical to Wuhan S. The relatively short time of transmission and occurrence within an enclosed space, suggest a founder effect. By contrast, in 7 out of the 9 states evaluated where D614G and Wuhan S appeared at approximately the same time (Arizona, Minnesota, Virginia, Connecticut, New York, Wisconsin, Utah), D614G became the predominant genotype at the end of the sampling period.

The observed dominance and dramatic increase over time of the D614G mutation points to a competitive advantage. While the impact of the mutation on the mortality and morbidity of COVID-19 virus has yet to be determined, our analysis suggests a functional advantage offered by this single amino acid mutation in the SARS-CoV-2 S protein. Since the S protein is critical to virus infection, we sought to determine the impact of the D614G mutation on the S protein structure and function.

SARS-CoV-2 S protein contains a furin cleavage domain that is critical for infectivity[6-8]. The available structural detail of the SARS-CoV-2 S protein lacks resolution of the secondary structure of the furin cleavage domain. We used threading techniques to create a prediction of the secondary structure of the region that contains the mutation site and the furin cleavage domain. Two models were created, D614 and G614, and a provocative result was observed in which the only significant change in structure is seen at the furin cleavage domain. This observation was confirmed by KSDSSP analysis. The D614G mutation creates an alpha helical structure at the furin cleavage domain. While the D614G mutation is 68 amino acid N-terminal of the cleavage, distal changes in protein structure by a single amino acid change have been observed with other proteins [17–20]. The non-conservative AA change from aspartate to glycine for the D614G mutation is certain to change the hydrogen bonding and packing around the region of the mutation.

In our analysis, G614 simultaneously arranges the P1 arginine (R685) and P4 arginine (R682) with better proximity to their respective binding pockets. Admittedly, furin is a promiscuous enzyme with >1,000 protein substrates and while the orientation of P1 and P4 arginine's are critical for cleavage, this does not exclude the involvement of adjacent amino acids [32]. Our models predicted the cleavage domain resides in a flexible, solvent exposed loop. Furin cleavage domains are generally believed to reside in unstructured loops [39]. However, there are no PDB structures of furin and a cleavable protein substrate. We compared our models to a high-resolution PDB structure for furin and an active site inhibitor, Meta-guanidinomethyl-phenylacetyl-RVR [26]. Our alignments of the D614 and G614 cleavage domains with the active site inhibitor would predict that G614 will act as a better substrate for furin cleavage. An increase in S protein cleavage would lead to more rapid membrane fusion and cell entry, especially if furin cleavage is rate determining in the membrane fusion process.

Our homology modeling results have revealed a new potential mechanism for how the D614G mutation could lead to a more competitive SARS-CoV-2 virus that has gained dominance. However, it does not preclude other proposed mechanisms [22, 30], also based upon structure modeling, that could act in concert with the one we propose. It is quite possible that the D614G mutation may influence several critical processes so that the overall change in rate of infection is greater than is possible with a modification of a single event. Our observed rise of the D614G mutation and its effect on the furin cleavage domain of the S protein suggests the mutation will have an important role in COVID-19 epidemiology and the design of therapeutic interventions.

Materials and methods

Spike protein sequence analysis dataset and approach

In order to investigate the transmission dynamics of SARS-CoV-2, the full genome sequences from the patient samples were retrieved from GISAID [13]. First, all the whole genome sequences (>29,000 bp) of the virus were downloaded (n=12,845). We identified and extracted the DNA sequence corresponding to the S protein from each viral genome data by NCBI BLAST and a proprietary algorithm written in Wolfram Mathematica available from Wolfram Research, Champaign, IL. Subsequently, we analyzed each S protein sequence and translated it. We removed shorter or incomplete sequences and those that yielded too many stop codons, most likely a result of lower base-calling accuracy with insertions and deletions of bases causing a frameshift. The dataset was processed, and the geographical figures were generated using the Filled Maps capabilities in Tableau available from Tableau Software, Seattle, WA. Data sets are provided in Supporting Information S1 to S5.

Molecular Modeling

The 3-D protein region from AA 591 to 710 was modeled using I-TASSER [15, 16] multiple template threading methodology in the NovaFold application software available from DNASTAR Inc., Madison, WI. The models with the highest Tm-Score for D614 or G614 were aligned using Pymol [31] available from Schroedinger, New York, NY. The Tm scores for both models were > 0.5. TM-scores for other models were below the Tm-Score threshold of 0.5. Residues surrounding the cleavage site are identified using the nomenclature of Schechter and Berger [32]. Distances between atoms were calculated using Pymol distance measurement. PDB files for 591-710 D614 and G14 are provided in Supporting Information S6 & S7.

Supporting information S1 to S7

S1 Spike protein mutation frequencies
S2 Spike genotype map data
S3 Spike genotype countries temporal progression data
S4 Spike genotype USA temporal progression data
S5 GISAID Dataset
S6 PDB file of S591-N710 for D614
S7 PDB file of S591-N710 for G614

Acknowledgements

The authors wish to than Joshua Lowitz and Roger Friedensen for reviewing the manuscript and Laetitia Eaton for help with figures. The authors, **LT**, **AS**, & **CC** were responsible for the S protein mutation analysis and **KD** & **JK** were responsible for the structural analysis of the D614G mutation.

References

1. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nature Reviews Microbiology. 2016;14:523–34.
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020. doi:10.1056/NEJMoa2001017.
3. Yuen, K., Ye, Z.-., Fung, S. *et al.* SARS-CoV-2 and COVID-19: The most important research questions. *Cell Biosci* **10,** 40 (2020). https://doi.org/10.1186/s13578-020-00404-4.
4. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. bioRxiv. 2020;:2020.02.19.956581.
5. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F. Cell entry mechanisms of SARS-CoV-2. PNAS. 2020, 202003138; DOI:10.1073/pnas.2003138117.
6. Bestle D, Heindl MR, Limburg H, van TVL, Pilgram O, Moulton H, et al. TMPRSS2 and furin are both essential for proteolytic activation and spread of SARS-CoV-2 in human airway epithelial cells and provide promising drug targets. preprint. Microbiology; 2020. doi:10.1101/2020.04.15.042085.
7. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020. doi:10.1016/j.cell.2020.02.052.
8. Braun E, Sauter D. Furin-mediated protein processing in infectious diseases and cancer. Clin Transl Immunology. 2019;8. doi:10.1002/cti2.1073.
9. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Research. 2020;176:104742.
10. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature Medicine. 2020;26:450–2.
11. Thomas G. Furin at the cutting edge: from protein traffic to embryogenesis and disease. Nat Rev Mol Cell Biol. 2002;3:753–66.

12. Hoffmann M, Hofmann-Winkler H, Pöhlmann S. Priming Time: How Cellular Proteases Arm Coronavirus Spike Proteins. In: Böttcher-Friebertshäuser E, Garten W, Klenk HD, editors. Activation of Viruses by Host Proteases. Cham: Springer International Publishing; 2018. p. 71–98. doi:10.1007/978-3-319-75474-1_4.
13. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Euro Surveill. 2017;22. doi:10.2807/1560-7917.ES.2017.22.13.30494.
14. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. arXiv:200310965 [q-bio, stat]. 2020. http://arxiv.org/abs/2003.10965. Accessed 5 May 2020.
15. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5:725–38.
16. Yang J, Zhang Y. Protein Structure and Function Prediction Using I-TASSER. Curr Protoc Bioinformatics. 2015;52:5.8.1-5.815.
17. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci USA. 2009;106:21149–54.
18. Modi T, Ozkan SB. Mutations Utilize Dynamic Allostery to Confer Resistance in TEM-1 β-lactamase. Int J Mol Sci. 2018;19.
19. Tyukhtenko S, Rajarshi G, Karageorgos I, Zvonok N, Gallagher ES, Huang H, et al. Effects of Distal Mutations on the Structure, Dynamics and Catalysis of Human Monoacylglycerol Lipase. Sci Rep. 2018;8:1719.
20. Lella M, Mahalakshmi R. Metamorphic Proteins: Emergence of Dual Protein Folds from One Primary Sequence. Biochemistry. 2017;56:2971–84.
21. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Journal of Translational Medicine. 2020;18:179.
22. Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020;:2020.04.29.069054.
23. Yuan Y, Cao D, Zhang Y, Ma J, Qi J, Wang, Q. et al. RCSB PDB - 5X58: Prefusion structure of SARS-CoV spike glycoprotein, conformation 1. 2017. doi:10.2210/pdb5x58/pdb.
24. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010;26:889–95.
25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.
26. Dahms SO, Arciniega M, Steinmetzer T, Huber R, Than ME. Structure of the unliganded form of the proprotein convertase furin suggests activation by a substrate-induced mechanism. Proc Natl Acad Sci U S A. 2016;113:11196–201.
27. Jencks WP. Catalysis in chemistry and enzymology. New York: Dover; 1987.
28. Gonzalez-Reiche AS, Hernandez MM, Sullivan M, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. medRxiv. 2020;:2020.04.08.200569
29. Shiryaev SA, Chernov AV, Golubkov VS, Thomsen ER, Chudin E, Chee MS, et al. High-resolution analysis and functional mapping of cleavage sites and substrate proteins of furin in the human proteome. PLoS ONE. 2017;8(1): e54290. https://doi.org/10.1371/journal.pone.0054290.
30. Becerra-Flores M, Cardozo T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int J Clin Pract. 2020.
31. Lill MA, Danielson ML. Computer-aided drug design platform using PyMOL. J Comput Aided Mol Des. 2011;25:13–9.
32. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. 1967. Biochem Biophys Res Commun. 2012;425:497–502.