HHS Public Access

Author manuscript

J Phys Chem Lett. Author manuscript; available in PMC 2021 May 26.

Published in final edited form as:

J Phys Chem Lett. 2020 December 03; 11(23): 10007–10015. doi:10.1021/acs.jpclett.0c02765.

Decoding Asymptomatic COVID-19 Infection and Transmission

Rui Wang,
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States

Jiahui Chen,
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States

Yuta Hozumi,
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States

Changchuan Yin,
Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, Illinois 60607, United States

Guo-Wei Wei
Department of Mathematics, Department of Electrical and Computer Engineering, and Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States

Abstract

One of the major challenges in controlling the coronavirus disease 2019 (COVID-19) outbreak is its asymptomatic transmission. The pathogenicity and virulence of asymptomatic COVID-19 remain mysterious. On the basis of the genotyping of 75775 SARS-CoV-2 genome isolates, we reveal that asymptomatic infection is linked to SARS-CoV-2 11083G>T mutation (i.e., L37F at nonstructure protein 6 (NSP6)). By analyzing the distribution of 11083G>T in various countries, we unveil that 11083G>T may correlate with the hypotoxicity of SARS-CoV-2. Moreover, we show a global decaying tendency of the 11083G>T mutation ratio indicating that 11083G>T hinders the SARS-CoV-2 transmission capacity. Artificial intelligence, sequence alignment, and network analysis are applied to show that NSP6 mutation L37F may have compromised the virus's ability to undermine the innate cellular defense against viral infection via autophagy regulation.

Corresponding Author Guo-Wei Wei − *Department of Mathematics, Department of Electrical and Computer Engineering, and Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States;* weig@msu.edu.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.0c02765.

Geographical collection data (XLSX)

Country and state supporting figures (ZIP)

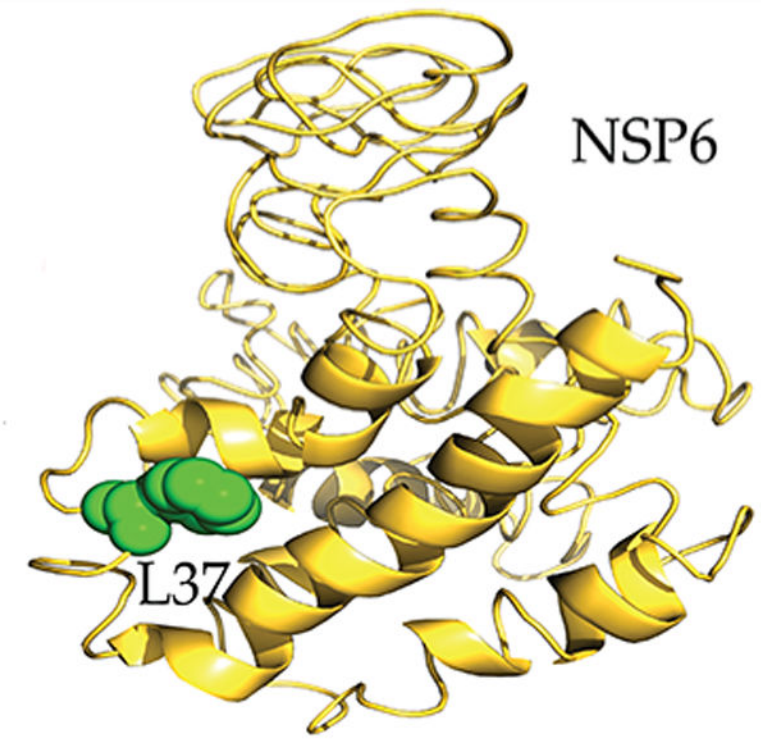
Additional analysis, materials and methods, and supplementary table and figure information (PDF)

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.0c02765

The authors declare no competing financial interest.

This assessment is in good agreement with our genotyping of the SARS-CoV-2 evolution and transmission across various countries and regions over the past few months.

Graphical Abstract



The ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has spread to more than 215 countries and territories with 40 015 371 positive cases and 1 112 545 fatalities as of October 19, 2020.[26] Unlike SARS-CoV that mainly infects the lower respiratory tract, SARS-CoV-2 is observed with a high level of shedding in the upper respiratory tract.[27] A wide variety of COVID-19 symptoms have been reported, including fever and chills, body and muscle aches, headache, nasal congestion, dry cough, fatigue, sore throat, and loss of taste or smell, 2–14 days after exposure to the virus.[27] Severe symptoms such as a high fever, severe cough, and shortness of breath indicate the onset of pneumonia. Less common gastrointestinal symptoms such as diarrhea, nausea, and vomiting have been listed by the Centers for Disease Control and Prevention. Currently, symptom-based testing, contact tracing, isolation, and quarantine are the main strategies for controlling and combating COVID-19. However, viable viruses have also been isolated from asymptomatic cases.[21] Asymptomatic and presymptomatic cases can play an important role in transmitting coronavirus.[13] Elusive asymptomatic transmission is regarded as the Achilles' heel of current strategies for controlling COVID-19.[14] Early studies with 235 cases of influenza virus infection indicate that the duration of influenza viral shedding was shorter and decayed faster and quantitative viral loads were lower in asymptomatic than in symptomatic cases.[18]

The epidemiological and virological characteristics of COVID-19 asymptomatic pathogenicity remain a mystery.

SARS-CoV-2 is a nonsegmented positive-sense RNA virus that belongs to the β-coronavirus genus, coronaviridae family, and Nidovirales order. Many RNA viruses, such as the flu virus, are prone to mutations due to the lack of proofreading in their genetic evolutions. Viral mutations are driven by a variety of factors, including the replication mechanism, polymerase fidelity, access to proofreading or postreplicative repair, sequence context, cellular environment, and host immune responses or gene editing.[23] Benefitting from an error-correction mechanism common to the Nidovirales order, the replication of coronavirus is regulated by bifunctional enzyme nonstructure protein 14 (NSP14).[11] Therefore, SARS-CoV-2 maintains relatively high accuracy in virus replication and transcription compared to the flu virus. Nonetheless, SARS-CoV-2 has had 19 126 single mutations compared with the genome collected on January 5, 2020.[25,28] Although the impacts of mutations on SARS-CoV-2 transmission and pathogenicity[2,12] and the COVID-19 diagnosis, vaccine, and medicine have been studied,[25] little is known about the connection between viral evolution and asymptomatic transmission.

This work reports the first association of a SARS-CoV-2 single nucleotide polymorphism (SNP) variant and COVID-19 asymptomatic cases based on the genotyping of 75 775 SARS-CoV-2 genome isolates. We reveal a significant correlation between asymptomatic infections and SARS-CoV-2 single mutation 11083G>T-(L37F) on NSP6. NSP6 is a common protein of α- and β-coronaviruses located at the endoplasmic reticulum (ER).[6] As a multiple-spanning transmembrane protein, coronavirus NSP6 participates in viral autophagic regulation. Autophagy degrades alien components to provide an innate defense against viral infection and promote cell death and morbidity.[22] In response to extreme cases of starvation, autophagy generates autophagosomes to transfer long-lived proteins and unnecessary or dysfunctional components to lysosomes for orderly degradation. Studies show that NSP6 undermines the capability of autophagosomes to transport viral components to lysosomes for degradation by rendering smaller-diameter autophagosomes in nutrient-rich media, thus enhancing viral replication.[6] Additionally, NSP6 proteins induce a larger number of autophagosomes per cell compared with starvation to facilitate the assembly of replicase proteins.[3,6] Although further studies are required to understand the molecular mechanism of the NSP6 regulation of autophagy, it is clear that coronavirus NSP6 is extremely important to viral protein folding, viral assembly, and replication. It is of paramount importance to understand how NSP6 mutation L37F leads to COVID-19 asymptomatic transmission and reduced virulence. By predicting the effect of single mutations on the protein folding free energy, the overall stability of the 3D structure of macromolecules will be captured.[16]

We show that NSP6 is a relatively conservative protein and the region around residues is quite conservative in several SARS-CoV-2-related genomes to maintain the crucial regulative function of NSP6. Using artificial intelligence (AI), topological data analysis (TDA), and a variety of network models, we further demonstrate that mutation L37F disrupts the folding stability of NSP6. We uncover the correlation between NSP6 mutation L37F and weakened SARS-CoV-2 virulence. We also analyze the global transmission and

find the decay tendency of NSP6 mutation L37F. We demonstrate that the evolutionary dynamics of L37F is age- and gender-independent.

We analyze 75775 SARS-CoV-2 complete genome sequences deposited in the GISAID (https://www.gisaid.org/) database up to October 19, 2020. Among them, 9912 samples have patient status information recorded as asymptomatic, symptomatic, hospitalized, intensive care unit (ICU), deceased, and so on. In particular, 537 samples are labeled with asymptomatic (76) and symptomatic (461) cases. By genotyping 537 genome samples, we find that mutation 11083G>T is significantly relevant to asymptomatic infection with a p-value being much smaller than 0.05. Mutation 11083G>T changes leucine (L) residue at position 37 of NSP6 to phenylalanine (F), denoted as L37F.

In the following sections, we investigate the relationship between mutation 11083G>T-(L37F) and asymptomatic infections using both gene-specific analysis and protein-specific analysis. The global evolution and transmission pathway of the mutation are studied as well.

Our customized data set generated from genotyping 75 775 SARS-CoV-2 complete genome sequences downloaded from the GISAID database is summarized in Table 1. Here, patient status is referred to as incomplete records of asymptomatic, symptomatic, hospitalized, deceased, and gender. Additionally, our data set contains data collection dates and locations of genome isolates. The detailed information on our data set is available in the Supporting Information.

We first analyze 537 sequences that have either asymptomatic or symptomatic labels. Here, 466 sequences are collected in Japan. The statistical analysis shows that the Pearson correlation coefficient between asymptomatic records and 10083G>T is 0.61. Moreover, the p-value is 5.42×10^{-56}, which reveals the significant relevance between asymptomatic and single mutation 11083G>T-(L37F).

Additionally, we apply Fisher's exact test. A 2×2 contingency table is given in Table 2. The odds ratio and p-value are 33.39 and 8.45×10^{-35}, which confirm the significant relevance between asymptomatic and single mutation 11083G>T-(L37F). To be noted, a total of 457 unique SNPs can be detected among 537 sequences that have either asymptomatic or symptomatic labels. Therefore, multiple hypothesis testing is also carried out by using the Benjamini−Hochberg procedure. Assume that the Benjamini−Hochberg procedure controls the false discovery rate (FDR) at level α. If we set $\alpha = 0.03$, then corresponding p-values are illustrated in Figure 1. (The mutation that has a p-value larger than 0.05 will not be illustrated in the figure.) We can see that 25 mutations pass the test. Among them, 241C>T, 3037C>T, 14408C>T, 23403A>G, 11083G>T, 28881G>A, 28882G>A, and 28883G>C have a frequency greater than 5000. However, except for 11083G>T, the other mutations that pass the multiple hypothesis test appear in only three sequences with asymptomatic labels. Therefore, we say that 11083G>T is the only mutation that has significant relevance to being asymptomatic.

Next, we split our genotyping data set of 75 775 sequences into different countries and extract those records with the 11083G>T mutation. Table 3 summarizes the total number of sequences related to 11083G>T-(L37F), denoted as N_{L37F}, the total number of sequences

N_S, the 11083G>T-(L37F) ratio, the number of total cases, the number of total deaths, and the death ratio of 20 countries up to October 19, 2020.

Many factors affect the death ratio of COVID-19, such as the number of confirmed cases, the level of the healthcare system, the pandemic response policy, and the medical and biochemical mechanism. Table 3 suggests that the 11083G>T-(L37F) ratio may be one of the factors affecting the death ratio. For example, Singapore has the second-highest 11083G>T-(L37F) mutation ratio as of 0.673. A piece of recent news reported that half of the COVID-19 cases in Singapore are symptomless,[15] which matches our finding that mutation 11083G>T-(L37F) is relevant to asymptomatic infections. Moreover, Singapore has the lowest death ratio as listed in Table 3, suggesting that single mutation 11083G>T-(L37F) may have weakened the SARS-CoV-2 virulence. A similar deduction can be obtained from the records of Malaysia, Turkey, Jordan, Norway, and South Korea. The relatively high 11083G>T-(L37F) mutation ratio (greater than 0.135) with a correspondingly low death ratio (less than 2.70%) further validates that 11083G>T-(L37F) may be relevant to the asymptomatic-induced hypotoxicity of SARS-CoV-2. Moreover, eight out of nine countries whose mutation ratios are less than 0.100 have a death ratio greater than 2.9%.

The death ratios in the United Kingdom, Belgium, France, and Spain are even higher than 3%, which supports our assumption that mutation 11083G>T-(L37F) may weaken the SARS-CoV-2 virulence.

Figure 2 illustrates the number of complete SARS-CoV-2 sequences in our data set with 11083G>T-(L37F) detected versus the number of complete SARS-CoV-2 sequences without 11083G>T-(L37F) detected every 14 days in China, Japan, Singapore, India, the United Kingdom, Malaysia, the United States, Australia, and Spain as well as in two states in the United States: New York and Washington. The blue and red bars represent the 11083G>T-(L37F) counts and other mutation counts every 14-day period, respectively. Similar bar plots for all countries/regions involving this specific mutation can be found in the Supporting Information. Singapore and Malaysia's plots show that 11083G>T was widely found after mid-March, which is consistent with the report saying that at least half of Singapore's newly discovered COVID-19 cases show no symptoms.[15] Almost all of the cases collected after the beginning of April have the 11083G>T mutation with asymptomatic infections recorded, which may be the most robust evidence to associate the SARS-CoV-2 11083G>T or L37F mutation on NSP6 with the asymptomatic infections. India is one of the countries that also have a large number of 11083G>T mutation cases. It was reported on June 1, 2020 that 80% of all COVID-19 patients in India were asymptomatic or showed mild symptoms,[8] which supports our deduction that 11083G>T correlates with asymptomatic infection. Moreover, one can see that mutation 11083G>T in China, the United Kingdom, and Australia is not as abundant as in Singapore, Japan, and South Korea. Furthermore, as discussed before, asymptomatic infection may be associated with the hypotoxicity of SARS-CoV-2 as well. Note that in Figure 2 China, the United States, and Spain have relatively lower occurrences of 11083G>T, which may contribute to their relatively high death ratios.

However, this specific mutation is not the only factor in determining the death ratio. The number of total infected cases, diagnostic testing, medical and health conditions, age

structure, and nursing-home population are also critical factors. For example, Japan has a relatively high ratio of 11083G>T-(L37F), while the death ratio is 1.77%. Similarly, the United Kingdom has the highest death ratio, but its ratio of mutation 11083G>T-(L37F) is not in the last echelon. One of the possible reasons is that the healthcare system was heavily broken by a sudden increase in infected cases. The lack of proper healthcare attention to a large number of mid/severe COVID-19 cases resulted in a high mortality ratio. Note that although Russia has the lowest 11083G>T-(L37F) mutation ratio lists in Table 3, its 1.72% death ratio caused by COVID-19 is also relatively low. One possible reason is Russia's relatively low median age.

In total, 65 genome samples are related to 11083G>T among 1512 samples that were collected before May 12 in New York. After that, fewer cases were collected in New York, and the 11083G>T mutation ratio decreased. Also, the state of Washington had a small proportion of the 11083G>T mutation before May 26, 2020. From May 26 to June 09, 52 out of 695 genome sequences had mutation 11083G>T, which may indicate that the asymptomatic infection may be becoming increasingly prevalent in Washington.

Globally, the ratio of samples with the 11083G>T mutation over all samples, shown in the middle chart of the last row in Figure 2, decays over time, suggesting that the mutation is unfavorable to viral transmission. Two abnormal peaks appearing on February 18 and June 9 in the ratio were due to the L37F mutation counts from Japan and Washington, respectively. The relatively small numbers of total genome sequences on these dates induce the jumps. Unfortunately, the number of genome sequences decays rapidly after April as shown in Figure 2, while globally the number of SARS-CoV-2 infections increases steadily according to the WHO's daily situation reports (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports).

Furthermore, we track the global transmission pathways of the single mutation 11083G>T-(L37F) to understand its spread dynamics. We found that the first genome containing single mutation 11083G>T-(L37F) was sequenced in Chongqing, China on January 18, 2020, which can be considered to be the ancestor of 11083G>T in our data set. However, the complete genome sequences released on the GISAID do not include all of the infected cases. Mutation 11083G>T was also detected in a commutation record [8782C>T, 11083G>T, 28144T>C] in Yunnan, China 1 day earlier than Chongqing's sequence, which indicates that the true ancestor of 11083G>T might have occurred earlier than January 18, 2020. Note that the first SARS-CoV-2 sequence released in GISAID was collected in Wuhan, China on December 24, 2019. China's COVID-19 epicenter, Wuhan, implemented a lockdown on January 23, 2020, which explains why China has a relatively low ratio of the 11083G>T-(L37F) mutation.

The first confirmed case reported by the Singapore government was on January 23, 2020. Interestingly, the first genome sequence related to mutation 11083G>T-(L37F) was found in Singapore on January 29, 2020, only a few days after the first confirmed cases. Therefore, the transmission of COVID-19 in Singapore in the initial stage contains a large proportion of the11083G>T-(L37F) mutation. The prevalence of the 11083G>T-(L37F) mutation in Singapore's SARS-CoV-2 infections explains its low death ratio.

Mutation 11083G>T-(L37F) first appeared in the United States on January 22, 2020 in Arizona in a 26-year-old male with comutations [8782C>T, 11083G>T, 28144T>C, 29095C>T], which is the descendant of [8782C>T, 11083G>T, 28144T>C] found in China. Since then, more records related to 11083G>T have appeared in the United States. At this stage, 728 genome isolates from the United States in our data set are relevant to 11083G>T. France is the next country that found the 11083G>T mutation at an early stage. Both of them detected this mutation on January 29, 2020. Although the 11083G>T mutation has arrived in the United States and the United Kingdom at a very early stage, different subtypes of SARS-CoV-2 spread vastly due to their large population, which makes the 11083G>T mutation nondominant.

In Jordan, the earliest sequence released on GISAID was on March 16, 2020, with the 11083G>T mutation, which indicates the high 11083G>T ratio in Jordan. Note that multiple reasons may lead to a low death ratio of 1.06% in Jordan. One may be due to the hypotoxicity of SARS-CoV-2 caused by mutation 11083G>T. The other is that the healthcare system is in good condition due to the small number of infected patients.

The first confirmed COVID-19 cases reported by the Vietnamese government were on January 16, 2020. However, the first genome isolates related to mutation 11083G>T in our data set were collected on March 17, 2020. 11083G>T was found in Vietnam 2 months after the first confirmed cases, which resulted in a low 11083G>T mutation ratio. Surprisingly, none of the patients died in Vietnam before June. As a lower-middle-income country with nearly 100 million population and a much less advanced healthcare system than other countries such as the United Kingdom and France, Vietnam was successful in tackling COVID-19, which has attracted attention. They immediately took action to ban the entry of foreign nationals and implement strict social distancing rules, which is the essential factor in keeping the coronavirus death toll at zero.

Figure 3(a) generated by Highcharts illustrates the distribution and proportion of the 11083G>T-(L37F) mutation in the United States. Red and blue represent the 11083G>T-(L37F) mutation and non-11083G>T mutations, respectively. The color of the dominant type of mutation decides the base color of each state. One can see that Alaska, Texas, Oregon, and Ohio have higher proportion of 11083G>T than do the other states.

Figure 3(c) created by using Highcharts illustrates the distribution and proportion of the 11083G>T-(L37F) mutation in the world. Mutation 11083G>T-(L37F) dominated in only a few countries such as Japan, Singapore, Pakistan, Jamaica, Kenya, Belarus, and Brunei. Most of the other countries have a small proportion of mutation 11083G>T-(L37F). Detailed information for all countries/regions is available in the Supporting Information.

Figure 3(b) displays the potential age and gender disparities in mutation 11083G>T-(L37F). Among 75 775 complete genome sequences, 24 762 cases have age labels and 25 306 cases (male cases, 13 427; female cases, 11 879) have gender labels, which are used for our age-related and gender-related analysis, respectively. Overall, male cases have a slightly higher ratio of the 11083G>T-(L37F) mutation than female cases. In terms of the age distribution, the ratio is significantly higher in the 99–109 year age group. However, only 22 cases fall

into this group. Therefore, we conclude that mutation 11083G>T-(L37F) is quite evenly distributed over age and gender groups, implying that the mutation does not exhibit host-dependent behavior.

Figure 4 depicts the sequence and structural information on NSP6. Figure 4(a) is generated by an online server,[20] Figure 4(b) is plotted by PyMol,[7] and Figure 4(c) was created in ref 1. As shown in Figure 4(a), NSP6 is a multispan membrane protein. In mutation 11083G>T-(L37F), both leucine and phenylalanine are α-amino acids and nonpolar. However, phenylalanine has a benzyl ring on the side chain, making the secondary structure of NSP6 more compact. As shown in Figure 4(a,c), from NSP6 protein residue position 34 to 37, the residues are FFFL. Therefore, there will be four continuous phenylalanine amino acid FFFFs after the L37F mutation. As a result, mutation L37F can stiffen the NSP6 structure by the aromatic−aromatic, hydrophobic, or π-stacking interactions. Such an additional interaction can significantly change the NSP6 function.[3]

Figure 4(c) presents the alignment of the SARS-CoV-2 NSP6 sequence with those of the other four coronaviruses, including that of SARS-CoV. The sequence identities between SARS-CoV-2 NSP6 and the NSP6s of SARS-CoV,[19] RaTG13,[30] ZC45,[17] and BM48–31[9] are 87.2, 99.7, 97.9, and 83.8%, respectively. Note that the overall sequence identity between SARS-CoV-2 and SARS-CoV is 83.8%. Therefore, NSP6 is relatively conservative. Additionally, NSP6 position 37 is in a relatively conservative environment among different species. Position 37 for all other species is valine, which is also hydrophobic and extremely similar to leucine. Two amino acids differ by one $-CH_2-$ unit. A large region from 25 to 45 has only three mutations among five species, indicating that the region is potentially important to protein function.

The three-dimensional (3D) structure of NSP6 is displayed in Figure 4(b). To further understand the impact of mutation L37F on NSP6, we carry out several theoretical analyses using AI,[4] TDA,[5] the flexibility and rigidity index (FRI),[29] and a large number of other network theory models.[10] Our results are summarized in Table 4. The protein folding stability change following mutations is defined by $\Delta\Delta G = \Delta G_m - \Delta G_w$, where ΔG_w is the free-energy change of the wild type and ΔG_m is the free-energy change of the mutant type.[4] The folding stability change ΔG measures the difference between folded and unfolded states. Thus, a negative $\Delta\Delta G$ indicates the mutation causing destabilization and vice versa. We use a TDA-based deep learning method[4] to compute $\Delta\Delta G$ in this work. As shown in Table 4, a negative folding stability change $\Delta\Delta G$ is found for the mutation, indicating a destabilizing mutation L37F for NSP6. Physically, a large hydrophobic residue (F) at the lipid bilayer and solvent interface reduces the NSP6 stability. Considering the sequence alignment in Figure 4(c), we have also carried out a similar calculation for the mutation of L37 V: $\Delta\Delta G = -0.74$ kcal/mol. Therefore, leucine is favored at residue position 37 for SARS-CoV-2 NSP6.

Next, the FRI rigidity index R_η measures the geometric compactness and topological connectivity of the network consisting of C_α at each residue and the heavy atoms involved in the mutation.[29] The scale parameter η determines the range of pairwise interactions. In this work, the parameter η is set to 8 Å. As shown in Table 4, the increase in the FRI rigidity index R_η is found, indicating the increase in the rigidity of the first α helix of NSP6. It is

known that protein flexibility is required to allow it to function through molecular interactions within the cell, among cells, and even between organisms.[24] The quadruplet phenylalanine resulting from mutation L37F can significantly reduce NSP6's interactions with ER.

Moreover, seven network models[10] are utilized to analyze the L37F mutation in NSP6 as shown in Table 4. We consider the network of heavy atoms at residue 37 and C_α atoms in NSP6. The connections are allowed between any pair of atoms within a cutoff distance of 8.0 Å. The mutation-induced decrease in the edge density d is due to the factor that the mutation increases the edge on the surface of the protein. The mutation that induced the decrease in the average path length $\langle L \rangle$ is due to the mutation that increases the edge at shorter distances. The mutation that induced the decrease in average betweenness centrality $\langle C_b \rangle$ is due to the increase in the crowd of vertices by the mutation residue. The mutation that induced the increase in average eigencentrality $\langle C_e \rangle$ due to the increase in the number of connected components will enlarge the maximal eigenvalues. The mutation that induced the increase in the average subgraph centrality $\langle C_s \rangle$ is due to the decrease in the average participating rate of each edge. The mutation that induced the decrease in average communicability $\langle M \rangle$ is due to the decrease in the neighbor edge participation rate of each edge. Finally, the mutation that induced the increase in the average communicability angle $\langle \Theta \rangle$ is due to the increase in the alignments among different edges. Together with the FRI rigidity index and the protein folding stability changes, these network assessments show that NSP6 becomes unstable and less functional after the L37F mutation.[3]

While asymptomatic infections of severe acute respiratory syndrome 2 (SARS-CoV-2) have been reported worldwide in the past few months, little is known about the formation mechanism and virological characteristics of asymptomatic infections. This work parses 75 775 complete genome isolates of SARS-CoV-2 and for the first time reveals the relationship between asymptomatic cases and a single nucleotide mutation 11083G>T-(L37F) on NSP6. After genotyping 537 sequences with asymptomatic and symptomatic records, we find that 11083G>T is significantly relevant to the asymptomatic infection. By analyzing the distribution of 11083G>T in various countries, we unveil that the 11083G>T mutation may be correlated with the asymptomatic infection and the hypotoxicity of SARS-CoV-2. Moreover, we track the dynamics of the 11083G>T mutation ratio globally and discover its decaying tendency, indicating that the 11083G>T mutation hinders SARS-CoV-2 transmission. Furthermore, the analysis of the distribution of 11083G>T for different ages and genders unveils that the 11083G>T-driven mutation does not exhibit host-dependent behavior. The protein-specific analysis is also taken into consideration. The 11083G>T mutation leads the leucine (L) residue at position 37 of NSP6, changing to phenylalanine (F). By employing the graph network analysis and topology-based mutation predictor, we find that 11083G>T-(L37F) destabilizes the structure of NSP6 proteinwise. As one of the most conservative proteins of SARS-CoV-2, NSP6 involves viral autophagic regulation. Therefore, this destabilized mutation may result in the inefficiency of NSP6 in participating in viral protein folding, viral assembly, and the replication cycle, which underpins our deduction that 11083G>T-(L37F) may be relevant to the asymptomatic infections and weaken the SARS-CoV-2 virulence. The present work indicates that NSP6 can be an effective therapeutic target.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant GM126189, NSF grants DMS-1721024, DMS-1761320, and IIS1900473, the Michigan Economic Development Corporation, George Mason University award PD45722, Bristol-Myers Squibb, and Pfizer. The authors thank The IBM TJ Watson Research Center, The COVID-19 High Performance Computing Consortium, NVIDIA, and MSU HPCC for computational assistance.

REFERENCES

(1). Bah T Inkscape: Guide to a Vector Drawing Program; Prentice Hall Press, 2007.

(2). Becerra-Flores M; Cardozo T SARS-CoV-2 viral spike g614 mutation exhibits higher case fatality rate. Int. J. Clin. Pract 2020, 74, e13525. [PubMed: 32374903]

(3). Benvenuto D; Angeletti S; Giovanetti M; Bianchi M; Pascarella S; Cauda R; Ciccozzi M; Cassone A Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy. J. Infect 2020, 81, e24.

(4). Cang Z; Wei G-W Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. Bioinformatics 2017, 33 (22), 3549–3557. [PubMed: 29036440]

(5). Carlsson G Topology and data. Bulletin of the American Mathematical Society 2009, 46 (2), 255–308.

(6). Cottam EM; Whelband MC; Wileman T Coronavirus NSP6 restricts autophagosome expansion. Autophagy 2014, 10 (8), 1426–1441. [PubMed: 24991833]

(7). DeLano WL; et al. Pymol: An open-source molecular graphics tool. CCP4 Newsletter on Protein Crystallography 2002, 40 (1), 82–92.

(8). Desk W Asymptomatic infections: India's next big coronavirus challenge. The WEEK, 2020.

(9). Drexler JF; Gloza-Rausch F; Glende J; Corman VM; Muth D; Goettsche M; Seebens A; Niedrig M; Pfefferle S; Yordanov S; et al. Genomic characterization of severe acute respiratory syndrome-related coronavirus in european bats and classification of coronaviruses based on partial rna-dependent RNA polymerase gene sequences. J. Virol 2010, 84 (21), 11336–11349. [PubMed: 20686038]

(10). Estrada E Topological analysis of SARS-CoV-2 main protease. Chaos 2020, 30 (6), No. 061102. [PubMed: 32611087]

(11). Ferron F; Subissi L; De Morais ATS; Le NTT; Sevajol M; Gluais L; Decroly E; Vonrhein C; Bricogne G; Canard B; et al. Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. Proc. Natl. Acad. Sci. U. S. A 2018, 115 (2), E162–E171. [PubMed: 29279395]

(12). Forster P; Forster L; Renfrew C; Forster M Phylogenetic network analysis of SARS-CoV-2 genomes. Proc. Natl. Acad. Sci. U. S. A 2020, 117 (17), 9241–9243. [PubMed: 32269081]

(13). Furukawa N; Brooks J; Sobel J Evidence supporting transmission of severe acute respiratory syndrome coronavirus 2 while presymptomatic or asymptomatic. Emerging Infect. Dis 26(7), 2020, DOI: 10.3201/eid2607.201595.

(14). Gandhi M; Yokoe DS; Havlir DV Asymptomatic transmission, the achilles heel of current strategies to control COVID-19, 2020.

(15). Geddie J Exclusive: Half of singapore's new COVID-19 cases are symptomless, taskforce head says. Reuters, 2020.

(16). Getov I; Petukh M; Alexov E SAAFEC: predicting the effect of single point mutations on protein folding free energy using a knowledge-modified MM/PBSA approach. Int. J. Mol. Sci 2016, 17 (4), 512. [PubMed: 27070572]

(17). Hu D; Zhu C; Ai L; He T; Wang Y; Ye F; Yang L; Ding C; Zhu X; Lv R; et al. Genomic characterization and infectivity of a novel SARS-like coronavirus in chinese bats. Emerging Microbes Infect 2018, 7 (1), 1–10.

(18). Ip DK; Lau LL; Leung NH; Fang VJ; Chan K-H; Chu DK; Leung GM; Peiris JM; Uyeki TM; Cowling BJ Viral shedding and transmission potential of asymptomatic and paucisymptomatic influenza virus infections in the community. Clin. Infect. Dis 2016, 64 (6), 736–742.

(19). Lee N; Hui D; Wu A; Chan P; Cameron P; Joynt GM; Ahuja A; Yung MY; Leung C; To K; et al. A major outbreak of severe acute respiratory syndrome in Hong Kong. N. Engl. J. Med 2003, 348 (20), 1986–1994. [PubMed: 12682352]

(20). Omasits U; Ahrens CH; Müller S; Wollscheid B Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics 2014, 30 (6), 884–886. [PubMed: 24162465]

(21). Oran DP; Topol EJ Prevalence of asymptomatic SARS-CoV-2 Infection: A Narrative Review. Ann. Int. Med, 2020.

(22). Ouyang L; Shi Z; Zhao S; Wang F-T; Zhou T-T; Liu B; Bao J-K Programmed cell death pathways in cancer: a review of apoptosis, autophagy and programmed necrosis. Cell Proliferation 2012, 45 (6), 487–498. [PubMed: 23030059]

(23). Sanjuań R; Domingo-Calap P Mechanisms of viral mutation. Cell. Mol. Life Sci 2016, 73 (23), 4433–4448. [PubMed: 27392606]

(24). Teilum K; Olsen JG; Kragelund BB Functional aspects of protein flexibility. Cell. Mol. Life Sci 2009, 66 (14), 2231. [PubMed: 19308324]

(25). Wang R; Hozumi Y; Yin C; Wei G-W Decoding SARS-CoV-2 transmission, evolution, and ramification on COVID-19 diagnosis, vaccine, and medicine. J. Chem. Inf. Model 2020, DOI: 10.1021/acs.jcim.0c00501.

(26). WHO. Coronavirus Disease 2019 (COVID-19) Situation Report 151, 2020.

(27). Wölfel R; Corman VM; Guggemos W; Seilmaier M; Zange S; Müller MA; Niemeyer D; Jones TC; Vollmar P; Rothe C; et al. Virological assessment of hospitalized patients with COVID-2019. Nature 2020, 581 (7809), 465–469. [PubMed: 32235945]

(28). Wu F; Zhao S; Yu B; Chen Y-M; Wang W; Song Z-G; Hu Y; Tao Z-W; Tian J-H; Pei Y-Y; et al. A new coronavirus associated with human respiratory disease in china. Nature 2020, 579 (7798), 265–269. [PubMed: 32015508]

(29). Xia K; Opron K; Wei G-W Multiscale multiphysics and multidomain modelsFlexibility and rigidity. J. Chem. Phys 2013, 139 (19), 194109. [PubMed: 24320318]

(30). Zhou P; Yang X-L; Wang X-G; Hu B; Zhang L; Zhang W; Si H-R; Zhu Y; Li B; Huang C-L; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020, 579 (7798), 270–273. [PubMed: 32015507]

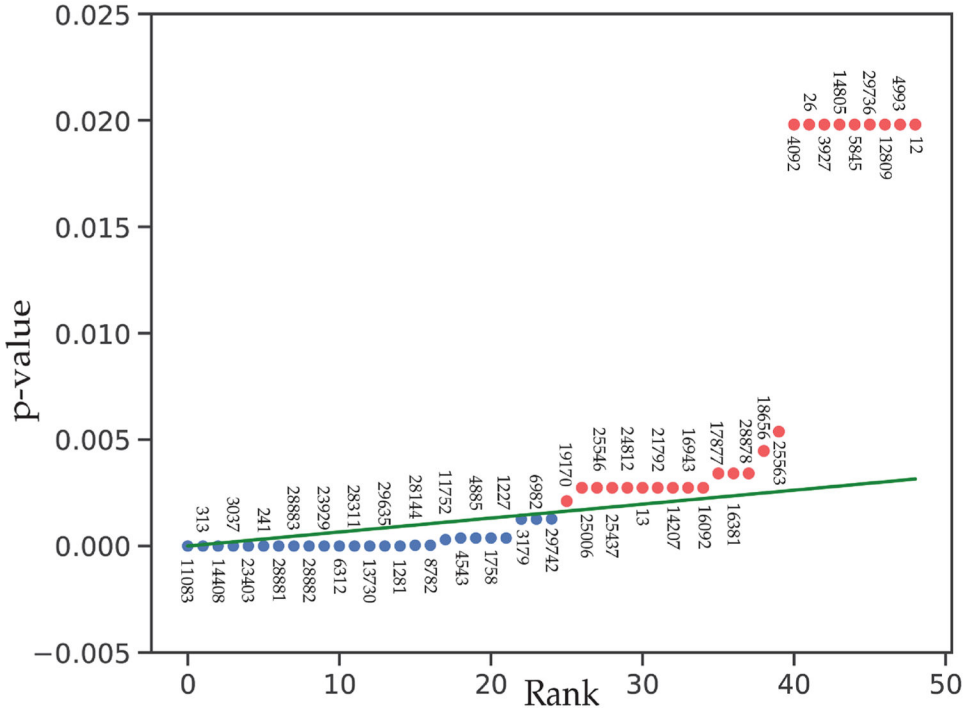


Figure 1.

Visualization of the Benjamini−Hochberg procedure. The *y* axis is the *p*-value of the mutation, and the *x* axis shows the rank of its corresponding *p*-values. The slope of the line is α/m, where m is the total number of independent hypotheses. Here, $m = 457$ and $\alpha = 0.03$. The blue dots represent the mutations that pass the test, and the red dots represent the mutations that do not pass the test.

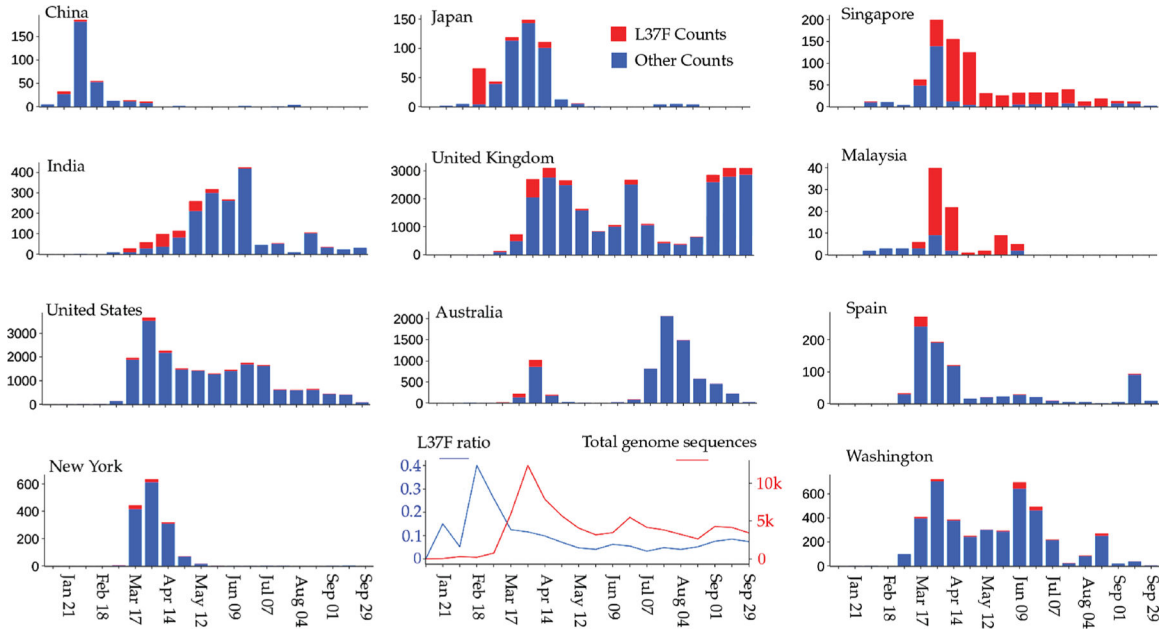


Figure 2.

Time evolution of SARS-CoV-2 mutation 11083G>T-(L37F) in nine countries and two states (i.e., New York and Washington) from the United States. The bar plots show the frequency of cases with and without the 11083G>T-(L37F) mutation. Each bar width covers a 14-day period. The blue line plot illustrates the evolution of the L37F mutation ratio computed as the count of genome sequences having the L37F mutation over the total count of genome sequences in each time period. The red line plot shows the evolution of the total count of genome sequences.

Figure 3.

(a) Pie-chart plot of the distribution of genome samples with and without the 11083G>T-(L37F) mutation in the United States. The red and blue colors represent the 11083G>T-(L37F) mutation and non-11083G>T mutations, respectively. The color of the dominant mutation type decides the base color of each state. (b) Bar plot of the age and gender distributions of the ratio of the number of samples having mutation L37F over the total number of samples having age and/or gender labels. The straight lines over all age groups are the group average ratios. (The average ratio from the total data set having age labels overlaps with the average ratio from the data set having male labels.) (c) Pie-chart plot of the distribution of genome samples with and without the 11083G>T-(L37F) mutation in the world. The color of the dominant mutation decides the base color of each country.

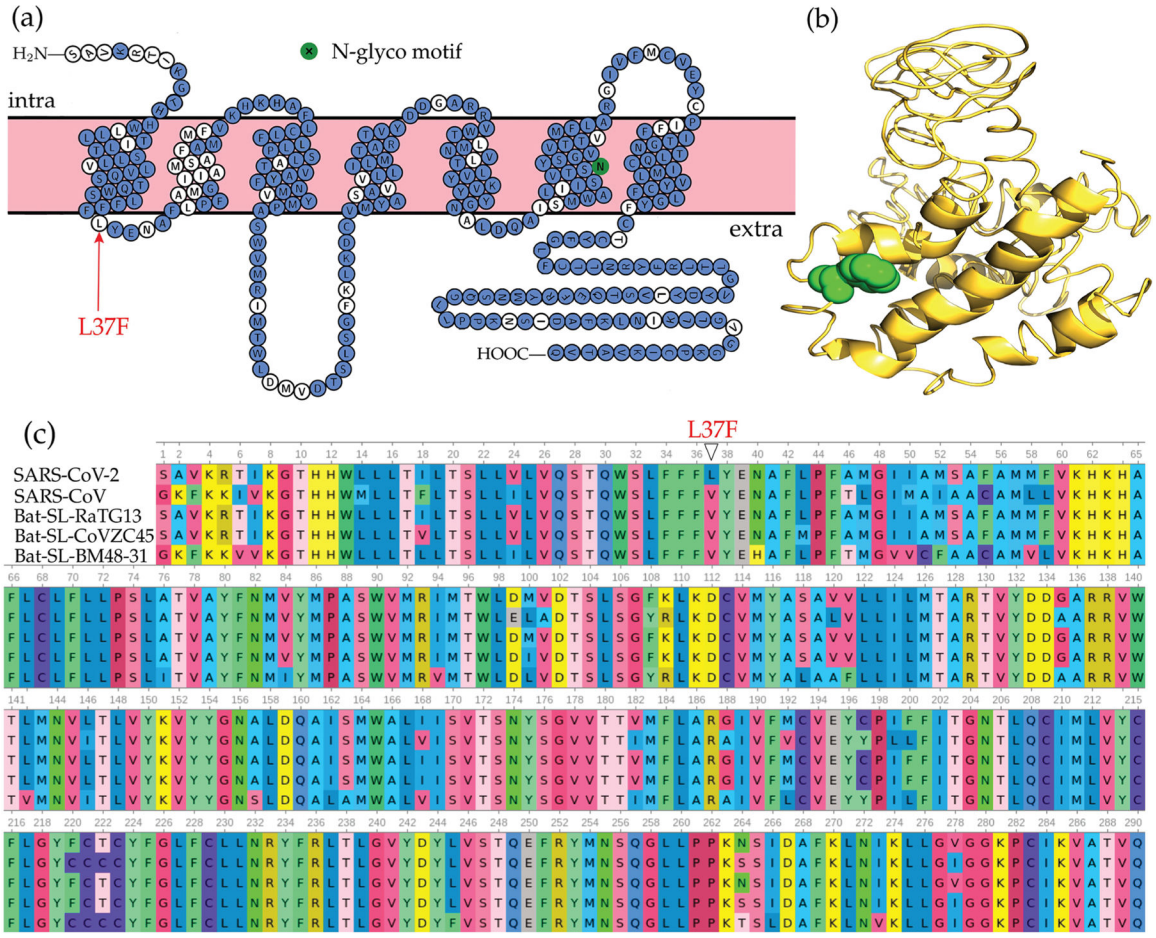


Figure 4.

(a) Visualization of SARS-CoV-2 NSP6 proteoform. The 11083G>T mutation in the genome sequence leads to the residue 37 leucine (L) mutant to phenylalanine (F). We use a red arrow to point out the mutation detected at residue 37. According to the sequence alignment results in (c), we color the conservative SARS-CoV-2 NSP6 residues blue. (b) Three-dimensional structure of the SARS-CoV-2 NSP6 protein. Green represents the mutation residue at position 37 of NSP6. (c) Sequence alignments for the NSP6 proteins of SARS-CoV-2, SARS-CoV, bat coronavirus RaTG13, bat coronavirus CoVZC45, and bat coronavirus BM48–31. The numbering is generated according to SARS-CoV-2.

Table 1.

Characteristics of the Customized Data Set

sample size	with 11803G>T	with patient status	asymptomatic	symptomatic
75 775	6052	9912	76	461

Table 2.

2 × 2 Contingency Table for Fisher's Exact Test

	asymptomatic	symptomatic
with 11803G>T	57	38
without 11803G>T	19	423

Table 3.

11083G>T-(L37F) Mutation Ratio and Death Ratio in Each Country as of October 19, 2020[a]

country/region	N_{L37F}	N_S	mutation ratio	total cases	total deaths	death ratio (%)
Singapore	555	825	0.673	57 965	28	0.05
Japan	89	528	0.169	96 534	1711	1.77
Turkey	29	126	0.230	359 784	9727	2.70
Jordan	7	22	0.318	50 750	540	1.06
India	237	1890	0.125	7 864 811	118 534	1.51
Norway	28	207	0.135	17 232	279	1.62
Australia	335	7324	0.046	27 499	905	3.29
South Korea	140	275	0.509	25 836	457	1.77
United Kingdom	2734	27 365	0.100	854 014	44 745	5.23
Canada	145	1285	0.113	211 732	9888	4.67
Vietnam	5	80	0.062	1160	35	3.02
Belgium	45	869	0.052	320 931	10 796	3.36
Malaysia	69	93	0.742	25 742	221	0.86
China	18	326	0.055	91 675	4746	5.18
France	39	984	0.040	1055942	34362	3.25
United States	728	20 101	0.036	84 031 217	222 507	0.26
Brazil	10	358	0.028	5 353 656	156 471	2.92
Spain	48	866	0.055	1 046 132	34 752	3.32
Russia	5	479	0.010	1 513 877	26 050	1.72

[a] N_{L37F} and N_S represent the total number of sequences with 11083G>T-(L37F) and the total number of sequences in each country listed in our data set, respectively.

Table 4.

Folding Stability Change and Graph Network Descriptor Consisting of Wild-Type and Mutant-Type NSP6 Calculated by a Web Server[44] [a]

mutation L37F			destabilizing
ΔΔ*G*			−0.74 kcal/mol
descriptors	**wild type**	**mutant type**	**relative change (%)**
R_8	19.73	19.83	0.50
D	0.05816	0.05779	−0.64
$\langle L \rangle$	22.7394	22.7373	−0.01
$\langle C_b \rangle$	0.00943	0.00942	−0.11
$\langle C_e \rangle$	0.03508	0.03514	0.17
$\langle C_s \rangle$	8 786 858	8 820 927	0.39
$\langle M \rangle$	2 826 948	2 822 450	−0.16
$\langle \Theta \rangle$	1.8399	1.8731	1.77

[a]The folding stability change is $\Delta\Delta G = \Delta G_m - \Delta G_w$, where ΔG_m is the mutation-type free-energy change and ΔG_w is the wild-type free-energy change.

R_8: FRI rigidity index with $\eta = 8$; *d*: edge density; $\langle L \rangle$: average path length; $\langle C_b \rangle$: average betweenness centrality; $\langle C_e \rangle$: average eigencentrality; $\langle C_s \rangle$: average subgraph centrality; $\langle M \rangle$: average communicability; $\langle \Theta \rangle$: average communicability angle.