Identification of novel missense mutations in a large number of recent SARS-CoV-2 genome sequences

Hugh Y. Cai[1*], Kimberly K. Cai[2], Julang Li[3*]

1. Animal Health Laboratory, University of Guelph, Guelph, Ontario, Canada
2. Department of Family Medicine, McMaster University, Hamilton and Forbes Park Medical Centre, Cambridge, Ontario, Canada
3. Department of Animal Biosciences, University of Guelph, Guelph, Ontario, Canada

Abstract. SARS-CoV-2 infection has spread to over 200 countries since it was first reported in December of 2019. Significant country-specific variations in infection and mortality rate have been noted. We performed a sequence analysis of 474 SARS-CoV-2 genomes submitted to GenBank up to April 11 and identified 5 recently emerged mutations in many the isolates (up to 40%). This finding was verified on a larger scale using the GISAID database with 8,008 SARS-CoV-2 sequences. Our analysis highlights 5 frequent new mutations that have emerged since late February 2020. These mutations are: one each missense (non-synonymous) mutation in orf1ab (C1059T), orf3 (G25563T) and orf8 (C27964T), one in 5'UTR (C241T), one in a non-coding region (G29553A). The final mutation (G29553A) was found to be almost exclusive to the US isolates. The first 3 mutations are non-synonymous, leading to amino acid substitutions in the viral protein sequence. Except for C241T, all the novel mutations identified are absent in the isolates from Italy and Spain. Although the clinical significance of these mutations is currently unclear, the findings lay the foundation for further study into the impact of SARS-CoV-2 mutations on disease incidence, severity, and host immune response. In addition, it may also provide insights into vaccine development and serological response detection for the virus.

Key words: COVID-19, SARS-CoV-2, virus, mutation, polymorphism, genome sequencing

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a RNA coronavirus, is the pathogen of coronavirus disease 2019 (COVID-19). Since it was first reported to the WHO in December 2019, it has spread to 213 countries, areas, or territories, causing 2,356,414 confirmed infections and 160,120 death worldwide (WHO April 20, 2020). The COVID-19 outbreak happened firstly in China with 84,237 confirmed cases and 4,642 deaths, then seriously hit Italy with 178,972 confirmed cases and 23,660 deaths, Spain with 195,944 confirmed cases and 20,453 deaths, then more recently the US with 723,605 confirmed a cases and 34,203 deaths, and other countries (WHO, April 20, 2020). Although country-specific differences in public health response have had a large impact on infection rate control, it is currently unclear as to whether evolution of the virus itself has also contributed to variations in infection and mortality rate.

RNA viruses possess a high mutation rate, ranging from 10^{-4}–10^{-6} mutations per round of genome replication.[1] Over 45 mutations have been described since the first SARS-CoV-2 sequence was identified in Jan 2020.[2, 3, 4, 5] However, these previous studies were based on the analysis of ~ 160 SARS-CoV-2 sequences available until mid-February 2020.[2, 3, 4, 5] By mid-April, > 550 SARS-CoV-2 sequences had been deposited in GenBank, and over 8,200 in the GISAID database. The geographic sources of the sequences have changed significantly. To provide a most recent view of the genetic variation of SARS-CoV-2, we retrieved 474 complete or close-to-complete genomes (>29,100 nt) from the National Center for Biotechnology Information (NCBI) to search for novel and high-frequency mutations. GenBank SARS-CoV-2 genomes were compared with those of the GISAID hCoV-19 database, consisting of 8,008 SARS-CoV-2s complete or close-to-complete genomes (>29,100 nt). We discovered that many SARS-CoV-2 isolates possessed mutations that were not described previously.

Results and discussion

Identification of 5 novel mutations

From the 474 sequences available in GenBank, a group of 100 SARS-CoV-2 genomes were found to have a nucleotide (nt 25563) mutated from G to T (G25563T). The mutation was exclusive to the US isolate sequences collected since March 2020 in the GenBank (downloaded April 11, 2020). The new mutants accounts for 21.1% (100/474) of all full genome sequences submitted to GenBank, or 27.9% (100/358) of the US full genome sequences in GenBank. Most of the G25563T isolates (94/100) co-possessed a C1059T mutation. Moreover, 16 of the G25563T isolates had an additional C27964T mutation, which accounts for 3.4% (16/474) of all full genome GenBank sequences, or 4.5% (16/354) of the US full genome sequences in GenBank. Among all 474 full genome sequences in GenBank, 48 collected from the US in March 2020 have a G29553A mutation. In addition, a mutation (C241T) was found in 30.8% (109/354) US isolates collected mostly in March 2020. The GenBank accessions of the isolates that we found with the novel mutations are shown in supplement Table S1. Of the 5 mutations described above, 3 mutations are substitution mutations in the coding regions, which resulted in amino acid sequence changes (missense mutation; non-synonymous mutations). They are C1059T causing amino acid 265 mutation from T to I (T265I) in orf1ab, G25563T (Q57H) in orf3a, C27964T (S24L) in orf8. The G29553A mutation is in a noncoding region upstream of orf10; the C241T mutation is at the 5' untranslated (5'UTR) region. These mutations have not been described previously, to our knowledge, and were found only in the isolates submitted mostly in and after March 2020 (including a few isolates in late February; Table 1). The representative images of the 5 mutations are shown in supplement Figure S1.

Proposed classification of the new SARS-CoV-2 isolates

Recently, the SARS-CoV-2 isolates have been classified into 3 clusters (groups), namely group A, B and C, based on 3 mutations[2]. The original isolates without mutation collected in Dec 2019 from China were classified as group A; the isolates with C8782T/Y and T28144C mutations were labeled as group B (mutated from group A); when group B isolates mutated with G26144T, the mutated isolates were labeled as group C. The isolates with the 3 nonsynonymous (missense) mutations identified in our study did not fall in the category of group A, B, C, since they had many mutations on top of group A, but did not have marker mutations C8782T/Y and T28144C (group B), nor G26144T (group C). To be consistent with the recent cluster (group) classification[2], we classified the isolates with novel amino acid changes as follow: C1059T(T265I) and G25563T(Q57H) usually co-existed, they are group D; the ones with the C27964T (S24L) change are in group E.

The emerging geographic locations of group D and E SARS-CoV-2 isolates

The earliest SARS-CoV-2 sequences were collected from China in December 2019 (Table 1). Of the 19 early identified sequences, 12 were group A, 2 were group B, and 5 were group C. These data suggest that most of the isolates in the early stage of outbreak were group A. In addition, it also revealed that mutations to group B and C existed as early as December 2019. Similarly, Taiwan and India collected group A and B isolates in January 2020. In addition, Iran, Japan, Pakistan, Viet Nam, and Australia had collected only group A isolates in January 2020 (Table 1). By the time the outbreak spread to Spain in February and March 2020, all isolates collected in GenBank belonged to group B and C. In the US, the SARS-CoV-2 isolates collected in the early stage (January 2020) were group A and B, each accounting for about 50% of the isolates; 9 of 17 group A and 8 of 17 group B, respectively. However, in March, the percentage of group A isolates dropped dramatically to 5.7% (17/300); isolates in group B and their variants in group C together accounted for 62% (179/300) of the isolates submitted from the country (Table 1). More strikingly, ~ 1/3 of the US Mar-2020 isolates have at least 2 mutations identified in the current study. From the GenBank SARS-CoV-2 database (Table 1), we can see that the virus started mainly as group A, with a portion of variants mutated into group B and group C in December 2019. Thereafter, most isolates were group B and C. Then new mutants of groups D & E started to emerge, accounting for approaching 40% of the US isolates in March 2020.

Although a fairly representative snapshot, the GenBank information is obviously not a complete picture. As of April 13, 2020, 8,126 sequences were available in the GISAID hCoV-19(SARS-CoV-2) database. To validate our findings on GenBank, we retrieved all complete or near-complete genomes (>29,160 nt) from the GISAID hCoV-19 database. We analyzed these 8,008 with the focus on the new mutations (Table 2)

In the GISAID hCoV-19 database, 17.7 % (1,417/8,008) and 0.6% (50/8,008) were group D and E isolates, respectively (Table 2). In addition, we identified 55.3% (4,427/8,008) with the novel mutation of C241T. Consistent with our finding from GenBank sequences, 43% of the US isolates belong to group D. In addition, group D isolates have been present widely; they account for substantial isolates submitted to GISAID hCoV-19 database in late February to March 2020: Canada (21.7%, 28/129), UK (6.4%, 175/2,726), France (53.9%, 110/204), Iceland (17.3%, 104/601), Australia (16.9%, 66/391), Netherlands (11.1%, 65/585), Belgium (12.1%, 39/322), Luxembourg (37.2%, 32/86), and Finland (40%, 16/40). It is striking to note that no group D mutation was found in any of the SARS-CoV-2 isolates submitted by Italy (44) and Spain(105), respectively, although the outbreaks in those 2 countries were severe and several weeks earlier than the countries in other parts of Europe and North America. We speculate that group D mutations occurred in late February to early March 2020. Since group D were found in multiple countries in a relatively short period of time, the mutation may have possibly emerged in multiple countries independently. Among the 8,008 genomes in the GISAID hCoV-19 database, 50 (0.6%) had the C27964T (group E) mutation, 42 from the US, 2 from Canada, and 6 from Australia. Although it is a relatively small number, this mutation is in a coding region resulting in an amino acid sequence change and is thus also worth attention. The 6 Australian group E isolates are different from those collected from the US in that they did not have the mutations of group D and C1059T. Since the Australian group E isolates are different from the ones collected in the US and Canada, they possibly evolved in Australia independently.

Group B (C8782T/Y and T28144C), and group C (C26144T) sequences were found in 29.5%, 30.5%, and 6.3% of 95 isolates collected before Feb 14, 2020.[5] However, these mutations are absent in the genomes of the US group D and E isolates, suggesting that the US group D isolates evolved directly from the ancestral strains (group A). Another interesting finding of our study was the discovery of the mutation G29553A. It was found in 1.4% (110/8,008) GISAID SARS-CoV-2 genomes from the world, or 6.9% (109/1,591) in the US SARS-CoV-2 genomes. The >100 G29553A isolates are almost exclusively, with the exception of one (Iceland), from the US. The mutation is in a noncoding region of the virus genome, although the significance of the mutation is currently unknown.

The potential impact of the emergence of group D and E SARS-CoV-2 strains.

Group D and group E defining mutations found on orf3a and orf8 respectively are regions associated with the expression of accessory proteins. Accessory proteins are not required for viral replication but may affect viral virulence and pathogenesis.[5] Orf3a is 72% conserved between SARS-CoV and SARS-CoV-2. Based on its function in SARS-CoV, it has been postulated that Orf3a is involved in cell apoptosis.[7] Mutations in Orf3a in SARS-Cov-2 have been shown to also result in loss or change of epitopes that may help the virus evade the host immune response[7]. There may be clinical implications of the missense mutations of these proteins. First, patients who have already recovered from earlier COVID-19 infection may have incomplete or reduced immunity when subsequently exposed to the newly emerging group D or group E SARS-CoV-2. Second, development of ELISA serologic testing must account for the potential epitope variability among different SARS-CoV2 groups. Accuracy of serologic testing may be adversely affected by current and emerging mutations in these accessory proteins. Further study on the biochemical and clinical impact of the Q57H substitution noted in orf3a (group D) and the S24L substitution on orf8 (group E), especially on viral virulence, and pathogenesis host immune response, are warranted. Most group D isolates also demonstrated the missense C1059T mutation in orf1ab (T265I). Orf1ab encodes a replicase that is involved in viral transcription and replication.[8] It would be important to further elucidate the role of T265I substitution in viral replication.

Global efforts to increase sequencing of SARS-CoV-2 isolates will be critical for mutation monitoring and clinical correlation. In addition to epidemiologic analysis, identifying new mutations in the SARS-CoV-2 isolates may, among other efforts, shed light on vaccine development, and help in evaluating the current molecular testing protocol. Fortunately, none of the group D and E mutations that we identified were in the PCR targets in the protocols listed in WHO website (WHO.int, access April 17, 2020).

Materials and Methods

On April 11, 2020, there were 547 SARS-Cov-2 sequences deposited in GenBank, from which, we downloaded 474 complete or near-complete genomes of 29,161 to 29,866 nucleotides (nt) (hereafter refereed as full genome), including 358 from US, 64 from China, 24 from Spain, and 27 from other countries or regions (Table 1). The SARS-CoV-2 isolate Wuhan-Hu-1 collected in December 19, 2019 and deposited in GenBank in January 2020 (GenBank accession. NC045512)[9] was used as a reference for mutation analysis. All nucleotide position labeling in our study was based on the alignment with this sequence. The SARS-CoV-2 full genome sequences (474 in total) downloaded from GenBank were multiple aligned by a bioinformatic software, Geneious v.11 (Auckland, New Zealand) using Map to a Reference Assembly function. The aligned sequences were visually examined to confirm that they were aligned properly. The variants/SNP were identified by the software automatically and verified by visual confirmation. Short fragments (30 nt) containing the novel mutations identified in our study were used as queries to blast search against the sequences downloaded from GenBank to verify the existence of the mutations.

To verify our findings on a larger scale, 8,126 hCoV-19 (SARS-CoV-2) sequences from GISAID (Global Initiative on Sharing All Influenza Data) website (https://www.gisaid.org) were downloaded and analyzed with the same methods as described above for the GenBank sequences.

Acknowledgement

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GenBank and GISAID’s hCov-19 Database on which this research is based.

As we were completing this manuscript, a manuscript by Yao et al was posted as a preprint on April 19, 2020, in medRxiv (https://www.medrxiv.org/content/about-medrxiv), which described the C241T mutation.

Data availability

All sequence data used in this study were available from the GenBank and GISAID. GenBank accessions of the isolates with novel mutations identified in this study can be found in Supplement Table S1.

Author information

Affiliations

HYC, Animal Health Laboratory, University of Guelph, Guelph, Ontario, Canada
KC, Department of Family Medicine, McMaster University, Hamilton and Forbes Park Medical Centre, Cambridge, Ontario, Canada
JL, Department of Animal Biosciences, University of Guelph, Guelph, Ontario, Canada

Contributions

HYC and JL conceived the study. HYC collected and analyzed the data, HYC KC, JL co-interpret the data and wrote the article. All authors reviewed and commented to the final version.

Corresponding authors

Correspondence to Hugh Y. Cai or Julang Li

Funding Support:

Natural Sciences and Engineering Research Council of Canada (NSERC), Food from Thought and OMAFRA.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Jenkins, G., Rambaut, A., Pybus, O., Holmes E.C. Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. J Mol Evol 54, 156–165 (2002)
2. Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A. 2020 Apr 8. pii: 202004999. doi: 10.1073/pnas.2004999117. [Epub ahead of print] (2020).
3. Pachetti, M., Marini, B., Benedetti, F. Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R.C., Zella, D. & Ippodrino, I. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNApolymerase variant. J Translational Med DOI: 10.21203/rs.3.rs-20304/v1(2020).
4. Phan, T. Genetic diversity and evolution of SARS-CoV-2. Infect Genet Evol. 2020 Feb 21;81:104260. doi: 10.1016/j.meegid.2020.104260. [Epub ahead of print] (2020).
5. Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T. & Zhang, Z. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 2020 Mar 13. doi: 10.1002/jmv.25762. [Epub ahead of print] (2020).
6. Liu, D. X., Fung, T. S., Chong, K. K.-L., Shukla, A. & Hilgenfeld, R. Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Research*, *109*, 97–109. doi: 10.1016/j.antiviral.2014.06.013\(2014)
7. Issa, E., Merhi, G., Panossian, B., Salloum, T., & Tokajian, S. SARS-CoV-2 and ORF3a: Non-Synonymous Mutations and Polyproline Regions. *MSystems*. doi: 10.1101/2020.03.27.012013 (2020)
8. Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Zheng, M., Chen, L., & Li, H. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharmaceutica Sinica B. doi: 10.1016/j.apsb.2020.02.008 (2020).
9. 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., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu,Y., Wang, Q.M., Zheng, J.J., Zu, L., Holmes, E.C., Zhang, & Y.Z. A new coronavirus associated with human respiratory disease in China. Nature 2020. DOI: 10.1038/s41586-020-2008-3 (2020).

Table 1. Novel mutations identified in GenBank SARS-CoV-2 genomes as of April 11, 2020

Country and date of collection	Number of isolates	C241T (5'UTR)	C1059T (T to I) orf1ab Group D$^{\&}$	G25563T (Q to H) orf3a Group D$^{\&}$	C27964T (S to L) orf8 Group E	G29553A (noncoding)	Group A	T28144C C8782T/Y orf8 Group B^	G26144T Orf3a Group C^
Reference NC_045512 Wuhan-Hu-1~ (Dec-2019)	-	0	0	0	0	0	0	0	0
US (Mar-2020)	300	112	94	100	16	48	17	179	7
US (Feb-2020)	41	4	0	0	0	0	32	5*	0
US (Jan-2020)	17	0	0	0	0	0	9	8	0
China (Feb-Mar-2020)	3	0	0	0	0	0	3	0	0
China (Jan-2020)	43	0	0	0	0	0	23	14	6
China (Dec-19)	19	0	0	0	0	0	14	2	5
Twain, China (Jan-Feb-2020)	3	0	0	0	0	0	2	1	0
Australia (Jan-2020)	1	0	0	0	0	0	1	0	0
Brazil (Feb-2020)	1	0	0	0	0	0	1	0	0
Colombia (Mar-2020)	1	0	0	0	0	0	0	1	0
Finland (Jan-2020)	1	0	0	0	0	0	1	0	0
India (Jan-2020)	2	0	0	0	0	0	1	1	0
Iran (Mar-2020)	1	0	0	0	0	0	1	0	0

Israel (Feb-Mar-2020)	3	1	0	0	0	0	2	0	0
Italy (Jan-2020)	1	0	0	0	0	0	1	0	0
Japan (Jan-Feb-2020)	6	0	0	0	0	0	6	0	0
Nepal (Jan-2020)	1	0	0	0	0	0	1	0	0
Pakistan (Mar-2020)	2	1	0	0	0	0	1	0	0
Peru (Mar-2020)	1	1	0	0	0	0	1	0	0
Spain (Feb-Mar-2020)	24	6	0	0	0	0	6	17[#]	1[$]
Sweden (Feb-2020)	1	0	0	0	0	0	1	0	0
Viet Nam (Jan-2020)	2	0	0	0	0	0	2	0	0
Total	474	122	94	100	16	48	126	206	18
%		25.7	19.8	21.1	3.4	10.1	26.6	43.5	3.8

& C1059T and G25563T both are marker mutations of group D, and mostly coexists.
* Four of the group B isolates have C241T mutation
# six of the group B isolates have C241T mutation
$ this isolate did not have the mutation of B group
% these 7 C isolates have no mutation of the B group
^ Previously described[2, 5]

Table 2. Novel mutations identified in GISAID hCoV-19 (SARS-CoV-2) genomes as of April 13, 2020

Country	Number of sequences	C241T (5'UTR)	C1059T (T to I) orf1ab	G25563T (Q to H) orf3a Group D	C27964T (S to L) orf8 Group E	G29553A (Noncoding)
US	1,591	795	606	692	42	109
UK	2,726	1,489	100	175	0	0
Viet Nam	8	3	0	0	0	0
Thailand	5	4	2	2	0	0
Taiwan, China	40	19	11	8	0	0
Switzerland	52	51	0	0	0	0
Spain	105	47	0	0	0	0
South Africa	19	4	0	0	0	0
Slovenia	3	5	0	0	0	0
Slovakia	4		0	2	0	0
Singapore	45	3	1	1	0	0
Senegal	23	19	7	10	0	0
Saudi	3	2	0	2	0	0
Russia	4	4	1	2	0	0
Portugal	101	86	2	4	0	0
Peru	2	2	0	0	0	0
Panama	1	1	0	0	0	0
Pakistan	2	1	0	0	0	0
Norway	29	18	8	8	0	0
Nigeria	1	1	0	0	0	0
Netherlands	585	365	52	65	0	0
Mexico	12	8	0	0	0	0
Luxembourg	86	80	32	32	0	0
Lithuania	1	1	0	0	0	0
Latvia	5	3	0	0	0	0
Kuwait	8	4	0	0	0	0

Japan	102	18	0	5	0	0
Italy	44	37	0	0	0	0
Israel	2	1	0	0	0	0
Ireland	13	10	0	0	0	0
India	33	17	0	0	0	0
Iceland	601	421	95	104	0	1
Hungary	3	3	0	0	0	0
Greece	3	2	0	0	0	0
Ghana	15	8	3	3	0	0
Germany	64	40	21	21	0	0
Georgia	13	7	2	2	0	0
France	204	186	62	110	0	0
Finland	40	32	7	16	0	0
Estonia	4	4	1	1	0	0
Denmark	9	9	1	1	0	0
DRC	42	37	0	7	0	0
Czech	4	3	1	1	0	0
Colombia	2	1	0	0	0	0
China	230	13	2	0	0	0
Chile	7	1	0	0	0	0
Canada	129	54	14	28	2	0
Brazil	36	29	1	2	0	0
Belgium	322	286	32	39	0	0
Belarus	2	1	0	0	0	0
Austria	21	16	4	4	0	0
Australia	391	170	54	66	6	0
Argentina	3	3	0	1	0	0
Algeria	3	3	3	3	0	0
Total	8,008	4,427	1,125	1,417	50	110
%		55.3	14.0	17.7	0.6	1.4