Reporting two SARS-CoV-2 strains based on a unique trinucleotide-bloc mutation and their potential pathogenic difference

Mustak Ibn Ayub[1*]

[1.] Department of Genetic Engineering Biotechnology, University of Dhaka
*Correspondence: miayub@du.ac.bd, Phone: +8801860015272

Non-technical summary

Through an extensive analysis of the SARS-CoV-2 whole-genome sequences, here I am reporting two strains of the virus, designated as SARS-CoV-2a and SARS-CoV-2g which can be differentiated based on a unique 3 nucleotide (the building blocks of virus genome) change in the SARS-CoV-2. From literature review and computational analysis, I have characterized these strains. This bloc mutation is located in the 28881-28883 region on the reference genome map of SARS-CoV-2.

Remarkably, SARS-CoV-2a seems to be prevalent in areas/countries with relatively low COVID-19 cases (such as Portugal, Netherlands, Belgium) whereas in highly affected countries/areas (USA, Spain, France, and Germany) SARS-CoV-2g predominates. Within a country, such as in Italy, Abruzzo has very low COVID-19 cases and high presence of SARS-CoV-2a.

This is a crucial observation and can be further explored through retro-and -prospective pan-national genetic and epidemiological studies. Monitoring the dynamics of these two strains might be invaluable to manage the COVID-19 pandemic and this can be achieved by sequencing only a small region of the virus genome encompassing 28881-28883 nucleotide bloc. The two strains, SARS-CoV-2g has got GGG in those positions which is as same as the reference sequence, so should be considered as 'wild type'. Whereas, in SARS-CoV-2a the sequence has mutated into AAC. This is a unique event where three nucleotides are changing as a bloc in SARS-CoV-2. Most importantly, this bloc mutation affects the nucleocapsid (N) protein of the virus. N protein is crucial for virus replication. Literature review suggests that the (GGG>AAC) mutation would negatively affect the N protein and thus reduce its infectivity which can explain why in areas where SARS-CoV-2a is predominant, COVID-19 cases are lower.

doi:10.20944/preprints202004.0337.v1

Reporting two SARS-CoV-2 strains based on a unique trinucleotide-bloc mutation and their potential pathogenic difference

Mustak Ibn Ayub[1*]

[1.] Department of Genetic Engineering Biotechnology, University of Dhaka
*Correspondence: miayub@du.ac.bd, Phone: +8801860015272

Abstract: SARS-CoV-2, the novel coronavirus behind COVID-19 pandemic is acquiring new mutations in its genome. Although some mutations provide benefits to the virus against human immune response, a number of them may result in their reduced pathogenicity and virulence. By analyzing more than 3000 high-coverage, complete genome sequences deposited in the GISAID database, here I report a unique 28881-28883:GGG>AAC trinucleotide-bloc mutation in the SARS-CoV-2 genome that results in two sub-strains, described here as SARS-CoV-2g (28881-28883:GGG genotype) and SARS-CoV-2a (28881-28883:AAC genotype). Computational analysis and literature review suggest that this bloc mutation would bring 203-204:RG(arginine-glycine)>KR(lysine-arginine) amino acid changes in the nucleocapsid (N) protein affecting the SR (serine-arginine)-rich motif of the protein, a critical region for the transcription of viral RNA and replication of the virus. Thus, 28881-28883:GGG>AAC bloc-mutation is expected to modulate the pathogenicity of the SARS-CoV-2. Remarkably, SARS-CoV-2g and SARS-CoV-2a strains can be linked with the heterogeneity of COVID-19 cases across different regions within and between countries by analyzing existing data. Sequence analysis suggests that severely affected cities, such as Milan, Lombardy, New York, Paris have the predominant presence of SARS-CoV-2g strains, whereas less affected places like Abruzzo, Lyon, Valencia have a relatively higher presence of SARS-CoV-2a, an indication that the latter strain may contribute to the reduced cases of COVID-19. A similar relationship is observed when Netherlands, Portugal are compared with Spain, France and Germany. These analyses suggest that the SARS-CoV-2 has already evolved into a less infective SARS-CoV-2a affecting COVID-19 cases in different regions. The time a country or region needs to acquire SARS-CoV-2a strains may be indicative to the time it would need to overcome the peak of the COVID-19 cases. To confirm these assumptions, prompt retrospective and prospective epidemiological studies should be conducted in different countries to understand the course of pathogenicity of the SARS-CoV-2a and SARS-CoV-2g. Potential drugs can be designed targeting 28881-28883 region of the N protein to modulate virus pathogenicity.

Key words: SARS-CoV-2, COVID-19, strains of SARS-CoV-2

Introduction:

SARS-CoV-2 is a positive-stranded RNA virus and has already infected about two million people around the globe. With a genome size of ~30000 bases and very high infectivity, the virus has already amassed numerous changes in its genome and acquiring more.

The genome organization of SARS-CoV-2 is similar to other coronaviruses [1]. It has Open Reading Frames (ORFs) common to all beta-coronavirus (Figure-1) which includes ORF1ab responsible for most the enzymatic proteins, the surface glycoproteins (S), the envelope proteins (E), the membrane proteins (M) and the nucleocapsid proteins (N). There are also several nonstructural proteins expressed mostly from ORF3a, ORF6a, ORFF7a and ORF8a. The reference genome of the SARS-CoV-2 also includes ORF10a as part of its genome as shown in figure S1, Table-1.

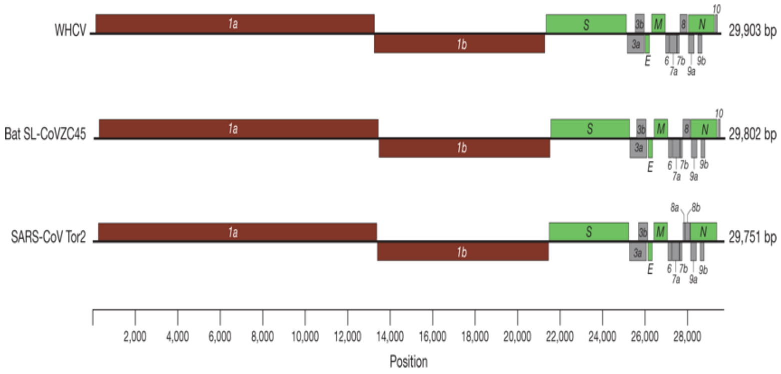

Figure-1: The genome of the SARS-CoV-2 has a very similar architecture with other beta coronaviruses. Image collected from [1]

Table-1: Size and span of the ORFs in SARS-CoV-2 according to the NCBI reference genome sequence.

ORF name	Span on the genome	Size (nt)
ORF1ab	(226-21555)	21290
S	(21563-25384)	3822
ORF3a/b	25,393-26,220	828
E	26,245-26,472	228
M	26,523-27,191	669
ORF6	27,202-27,387	116
ORF7a	27,394-27,759	366
ORF7b	27,756-27,887	132
ORF8	27,894-28,259	366
N	28,274-29,533	1260
ORF10	29,558-29,674	117

Notably, whole-genome sequencing of the SARS-CoV-2 and deposition to the public databases has been progressing with an unprecedented pace during this outbreak. Up until April 10, 2020, more than 3500 high-coverage, complete genome sequences of SARS-CoV-2 have been submitted to GISAID (Global Initiative on Sharing All Influenza Data) maintained by MPII (Max Planck Institute for Informatics).

After a careful analysis of the whole genome sequences in the GISAID database, this study has established that a unique trinucleotide-bloc mutation, 28881-28883:GGG >AAC might have occurred in recent time giving rise to a new subtype of SARS-CoV-2 with potential impacts on the course of the COVID-19 pandemic. This bloc mutation is mapped within the nucleocapsid (N) gene according to the SARS-CoV-2 reference genome. N protein plays a critical role to assemble coronavirus RNA genome and create a shell around the enclosed nucleic acid. It also interacts with the viral membrane protein during viral assembly, assists in RNA synthesis, folding and virus budding. The protein also affects host cell responses to the viral infection, including cell cycle regulation and immune responses modulation [2].

The 28881-28883:GGG >AAC mutation affects the SR (serine-arginine)-rich domain of the N protein. Previously in SARS-Cov-1 the closest neighbor to SARS-CoV-2, it has been shown that experimentally introduced deletion in the SSRSSSRSRGNSR region of the SR-rich motif significantly reduces the infectious virions [3]. The 28881-28883:GGG >AAC mutation affects the location adjacent to the aforementioned region, and so is expected to impact the pathogenicity of the SARS-CoV-2 in a similar manner. This assumption is remarkably supported from the analysis conducted by combining sequence information from GISAID database and COVID-19 cases in different regions around the globe from live trackers. From this exercise, it has become evident that regions with low/moderate cases of COVID-19 have the prevalence of 28881-28883:AAC genotype (SARS-CoV-2a), whereas the highly affected regions predominantly have 28881-28883:GGG genotype (SARS-CoV-2g).

History of previous infections suggests the evolution of viruses with different pathogenicity acquired through mutations [4] [5]. Although hundreds of mutations have been reported in the SARS-CoV-2 genome to date, the trinucleotide bloc mutation reported and characterized in this study have unique features with potential impact on the pathogenicity of the virus.

The results suggest that by monitoring the prevalence of the SARS-CoV-2a and SARS-CoV-2g strains, countries may track the course of COVID-19 pandemic. Potential drugs can be designed to target SR-rich motif of the N protein to curb the pathogenicity of the SARS-CoV-2. However, some assumptions need to be confirmed with more retrospective and prospective research. Special attention should be given to trace back the COVID-19 affected human samples from where the SARS-CoV-2 sequences were obtained and follow up with their clinical outcome.

Results:

28881-28883:GGG>AAC change is a unique event resulting in two sub-strains of the SARS-CoV-2 described here as SARS-CoV-2g and SARS-CoV-2a:

In all 3000 complete genome sequences of SARS-CoV-2 analyzed in this study, there was not a single occasion where a bloc of tri-nucleotide has changed except the GGG>AAC in the 28881-28883 location of the genome. All other changes are mostly single nucleotide polymorphism (SNPs). This observation suggests that GGG>AAC change has occurred at the same time or at a short span of time. Such changes would be expected to have significant impacts on the virus life cycle and pathogenicity as discussed later.

28881-28883:GGG>AAC mutation is accompanied by three other mutations, such as 241:C>T, 3037:C>T and 14408:C>T (Figure-2); but the opposite is not always true. This implies that the 241:C>T, 3037:C>T and 14408:C>T mutations precede the 28881-28883:GGG>AAC mutation. Among them, 14408:C>T brings 323:P>L changes in the RNA polymerase [6] of the SARS-CoV-2, which may contribute in the 28881-28883:GGG>AAC change in the virus genome. Further investigation is necessary to understand this course of events.

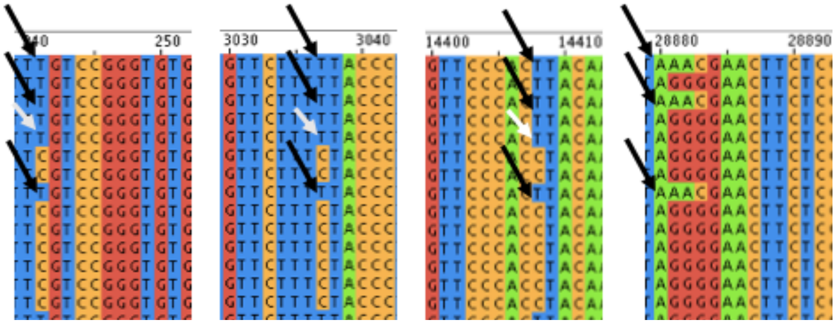

Figure-2: GGG>AAC change in the SARS-CoV-2 is always accompanied by three other C>T mutations in 241, 3037 and 14408 positions of the virus genome as indicated by black arrows. However, C>T change in these positions does not always mean the presence in GGG>AAC as indicated by white arrows. All positions are aligned to the NCBI SARS-CoV-2 reference sequence.

SARS-CoV-2a is a relatively new strain and has a distinct mutation profile compared to SARS-CoV-2g:

The 28881-28883:AAC genotype and resulting SARS-CoV-2a strain is found in samples collected in relatively recent times, mostly from March onward. All the sequences from Wuhan, the first epicenter of COVID-19 have 28881-28883:GGG genotype and so is the reference genome of the SARS-CoV-2. Although one SARS-CoV-2a affected person was reported in Italy on January [7], an analysis on the sequences deposited from Japan gives a good snapshot of its recent origin. 11 out of 95 sequences deposited there by April 7 were SARS-CoV-2a and all were collected between March 11-20, 2020. Among the 84 SARS-CoV-2g (GGG genotype), only 5 were collected in March and all other samples were collected in late January-mid February (Figure-S2). This trend is visible in other sequences deposited in the GISAID database as of April 10, 2020.

The recent SARS-CoV-2a genome looks relatively pristine compared to the SARS-CoV-2g, mostly because of the mutually exclusive mutations in them. It appears that some SARS-CoV-2 have completed a cycle of mutations to arrive at the level of SARS-CoV-2a by changing some base positions in their genome while leaving other positions untouched. SARS-CoV-2a has its own version of the leader sequence, RNA pol and nsp3 because of the complete transition in 241:C>T, 3037:C>T and 14408:C>T respectively.

Compared to the SARS-CoV-2g counterpart, SARS-CoV-2a has very few changes in its genome. An analysis done on 214 SARS-CoV-2a and 1013 SARS-CoV-2g sequences from different countries show their vivid difference. In this exercise, it was checked whether at any position of the genome there is more than 5% change among the sequences. The result summarized in Figure- 3 shows that the SARS-CoV-2a has only 3 positions with such changes whereas SARS-CoV-2g has 17 such positions (Supplementary Table-ST1). The positions which have changed in less than 10% cases generally are country-specific, except the 26144:G>T which has been found in sequences from various countries.

doi:10.20944/preprints202004.0337.v1

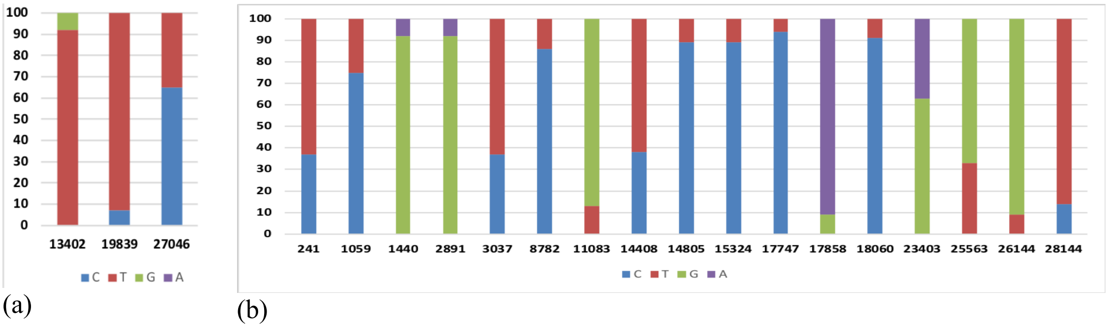

Figure-3: Frequency of change in base positions in SARS-CoV-2a and SARS-CoV-2g. (a) Only three positions showed more than 5% change in SARS-CoV-2a. (b) 17 positions have experienced changes in SARS-CoV-2g. The X-axis shows base positions on reference genome, Y-axis shows % occurrence of A, T, C, G nucleotides at specific positions.

Among these mutations, particularly interesting are the 25563:G>T and 26144:G>T mutations that affect ORF3a and are mutually exclusive in SARS-CoV-2a and SARS-CoV-2g: This was considered important as ORF3a protein modulate the immune responses, including 'cytokine storm' in the host [8]. All SARS-CoV-2a are free of those mutations, whereas, in SARS-CoV-2g strain, these mutations are frequent. Interestingly these two mutations are also mutually exclusive, i.e., all the SARS-CoV-2g with 25563:G>T mutations are free from 26144:G>T mutations and vice versa.

This pattern of mutational exclusiveness requires more elaborate analysis to trace the evolution of the SARS-CoV-2 strains, as they hold important clues on their pathogenicity.

Impacts of 28881-28883:GGG>AAC mutation on the pathogenicity of the SARS-CoV-2

According to the NCBI reference genome, 28881-28883:GGG>AAC bloc results in two amino acid 203-204:RG>KR changes in the nucleocapsid (N) protein of the SARS-CoV-2. Looking at the surrounding sequence of these amino acids (Figure-4), it appears that the mutation will discontinue a serine-arginine (S-R) dipeptide by introducing a lysine in-between them.

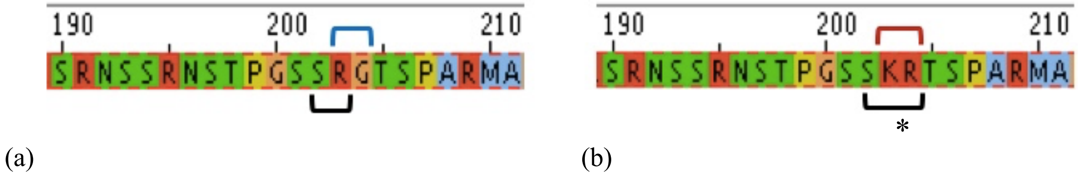

Figure-4: Impacts of 203-204:RG>KR mutation in the N protein. (a) The wild type N protein with intact S-R dipeptide and (b) the mutated N protein which has the S-R dipeptide disrupted with the insertion of lysine in-between them. Blue and red bars on the top indicates the wild type and mutated amino acids respectively. The bottom black bar in (b) with asterisk indicates the disrupted S-R dipeptide.

According to the NCBI Reference Sequence: YP_009724397.2 of the SARS-CoV-2 nucleocapsid (N) protein, the changes in the mutated N protein is expected to have impacts on its structure and function. Lysine is a basic and polar hydrophilic charged (+) amino acid. Its inclusion in the motif should have impacts on the overall characteristics of the protein as reported before [9]. Especially, the serine-arginine dipeptide disruption may hamper the phosphorylation of the SR-rich domain- crucial for the cellular localization and translation inhibitory function of the N protein [10].

Previous experimental work by deleting part of the SR domain in SARS-CoV-1 has shown reduced pathogenicity in the virus [3]. So the disruption discussed above should have some negative impact on the mutated N protein in SARS-CoV-2 if not complete loss of its function. A computation analysis shows that RG>KR mutation would change the length and arrangements of the alpha-helix of the Nucleocapsid protein (Figure-5). Laboratory experiments can confirm these predictions.

A multiple sequence alignment analysis and clustering based on neighbor-joining algorithm show that changing the amino acids at 203-205: RG > KR of the N protein put the SARS-CoV-2 as the only neighbor to a bat alpha-corona virus, whereas the wild type N protein clusters with several other viruses including MERS-CoV (Figure-5).

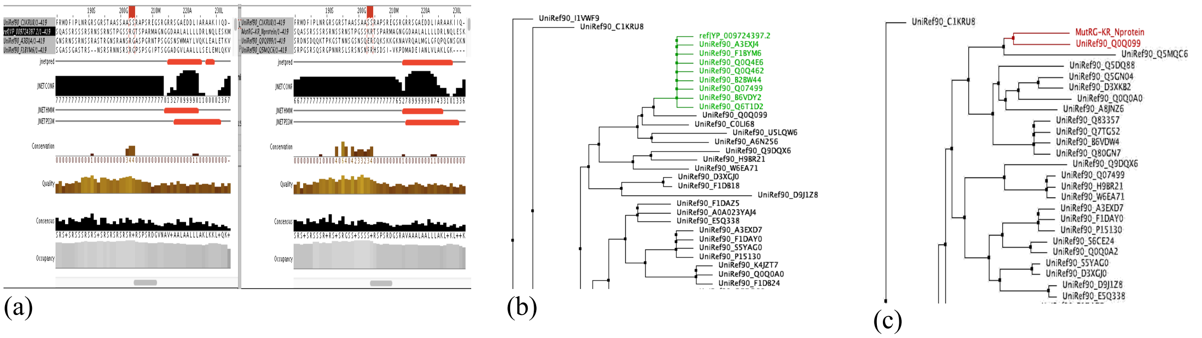

Figure-5: Impact of the RG>KR mutation on N protein. (a) The amino acid mutation at 203-204 position of the N protein changes the size and arrangement of the alpha-helix of the protein as indicated by the red bars. The left panel in the image is the reference N protein whereas the right panel represents the mutated N protein. Conservations, quality and consensus has been indicated in the bottom layers of the image. (b) Neighbor-joining clustering based on RG dipeptide after multiple sequence alignments shows seven protein (green) as the neighbor of the wild type N protein. (C) Clustering based on the mutated KR dipeptide has found only one immediate neighbor (red) to the mutated N protein indicating its rarity.

Distributions of SARS-CoV-2a and SARS-CoV-2g within and among countries and their potential impacts on COVI-19 cases: This study started from the observation that although Italy has the third largest reported COVID-19 cases in the world (as of April 11), Abruzzo of Italy has much fewer cases compared to Lombardy. As of April 9, 2020, some 54802 COVID-19 cases have been reported in Lombardy whereas in Abruzzo the number is 1931. By looking at the region-specific sequences from GISAID database, it was found that out of 30 sequences deposited from Italy by April 11, 2020 total 13 sequences came as 28881-28883:AAC (SARC-CoV-2a) and most strikingly 10 of them are from Abruzzo (Figure-6). This is ~77% of the total number of sequences (N=13) that came from that region. From the Lombardy region, 4 sequences were deposited, all of them are 28881-28883:GGG (SARS-CoV-2g). Although the numbers of sequence are not too high to reach any conclusion, it gave an impetus to check whether the difference in the prevalence of SARS-CoV-2a might be linked with the lower number of COVID-19 cases.

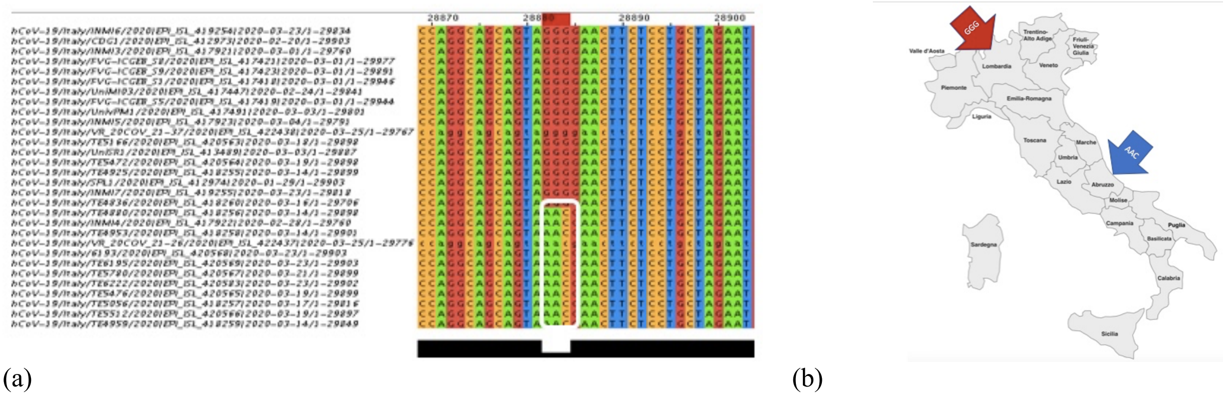

Figure-6: Prevalence of two strains of SARS-CoV-2 in Italy. (a) According to the deposited sequence from Italy, it was found that Abruzzo has particularly the SARS-CoV-2a strain (indicated by the white box in the aligned sequences, whereas other regions including Lombardy might be enriched in SARS-CoV-2g (b) The regions marked on the map of Italy indicate the presence of the 28881-28883:AAC and 28881-28883:GGG strains of SARS-CoV-2.

When other COVID-19 high vs. low regions were analyzed within and among countries and compared with their SARS-CoV-2 sequence entries in the GISAID database, a trend was observed that there is an inverse relationship between the reported number of COVID-19 cases and the relative abundance of SARS-CoV-2a strain. As of April 9, 2020 Belgium, Netherlands, Portugal have 31%, 50% and 60% SARS-CoV-2a, whereas Spain has only ~4% (N=83) and France ~3% (N=150). Deposited sequences from Germany, France, Belgium and Netherlands came from different part of those countries. Sequences from Portugal were deposited from a few numbers of laboratories located at different places.

In case of UK, 26% of strains showed SARS-CoV-2a genotype, but location information could not be confirmed for them. When European countries with more than 50 submitted sequences (as of April 9, 2020) were analyzed and then compared their reported COVID-19 cases, it appeared that the countries with a relatively higher prevalence of SARS-CoV-2a have lower cases of COVID-19 (Figure-7). This assumption should be taken with caution as there must be many factors responsible for the differences in COVID-19 cases in different countries, including their testing and reporting policy. However, the persistent observation of SARS-CoV-2a prevalence in countries and regions with low COVID-19 cases warrants an immediate molecular and epidemiolocal research around the world to check the impacts of the two SARS-CoV-2 strains on the course of COVID-19 pandemic.

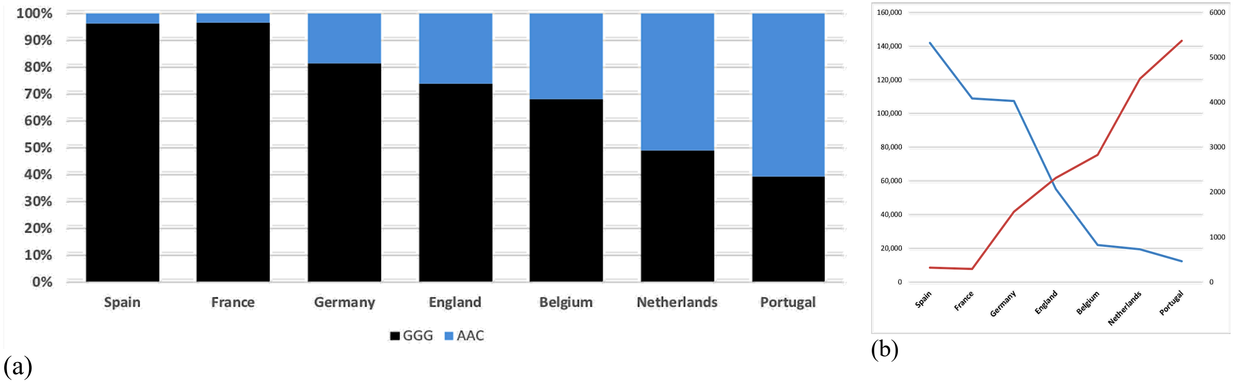

Figure-7: Top sequence depositors in GISAID database give an indication about the impact of the relative prevalence of SARS-CoV-2a and COVID-19 cases. (a) As indicated by the blue layers, Spain and France have very low relative presence of SARS-CoV-2a, whereas Portugal has more SARS-Cov2a than SARS-CoV-2g followed by Netherlands, Belgium ad England. (b) A comparison with the COVID-19 cases show that countries with higher prevalence of SARS-CoV-2a, such has Portugal has the lowest number of COVID-19 cases among these countries . The pattern matches in Netherlands, Belgium, UK, Germany, France and Spain. In the graph blue line indicates the number of COVID-19 cases as of April 9, 2020, and the red line indicates the normalized value of the relative prevalence of SARS-CoV-2a in the deposited sequence in GISAID database.

It is worth reporting here that SARS-CoV-2a are also present in other European nations with low COVID-19 cases such as Finland, Austria, Denmark, Iceland, Estonia.
When checked in South America, Brazil showed a sizable presence of SARS-CoV-2a. Sequences from Argentina and Chile indicate the presence of SARS-CoV-2a in those countries. More data will be needed from that region to be assured whether the presence of SARS-CoV-2a strains may be linked with the relatively lower reported cases of COVID-19 in South America, or it is just an artifact.

The most severely affected region in North America, New York has predominantly SARS-CoV-2g. Only recently, some SARS-CoV-2a sequences have been observed in sequences from samples isolated in New York. As of April 9, only 5% sequences are of SARS-CoV-2a and 95% are of SARS-CoV-2g (N=145). If the assumptions made above are true, then the more infectious SARS-CoV-2g might be behind the high COVID-19 cases in New York and until SARS-CoV-2a takes the upper-hand, the trend would continue. However, 241:C>T, 3037:C>T and 14408:C>T changes were present in the 89% of the 145 deposited sequences by April 9, 2020. As these mutations work as the precursor for the 28881-28883:GGG>AAC, it is expected that the GGG>AAC change will increase in the future giving rise to more SARS-CoV-2a and lowering the COVID-19 cases.

Both Australia and Canada also have a presence of SARS-CoV-2a strains in some regions (Figure- S3).
In Asia, Japan shows 11 of the submitted 95 sequences as SARS-CoV-2a. Vietnam, India, Thailand, Singapore have got SARS-CoV-2a strains according to the submitted sequence. However, sequences from South Korea, Malaysia, Nepal did not reflect the strain as of April 7, 2020. More sequences from different regions of these countries might be necessary to get a complete picture. China alone has deposited 250 whole sequence data as of April 9. However, most of these sequences are from samples collected in February. Samples collected at a later date should be screened for AAC genotype as discussed before.

A big caveat in the proposed link of SARS-CoV-2a with lower numbers of COVID-19 cases in some countries is their difference in testing rates. Different countries do test at different rates. However, the number of COVID-19 cases are not always related to the test rate. Germany has the highest rate of testing (16/1000 people) followed by Austria (13.3/1000 people) [11]. However, the COVID-19 cases in Austria are just 13560, as of April 11, 2020 [12]. Austria has deposited only 18 sequences in the GISAID database (as of April 11, 2020) and 3 of them are SARS-CoV-2a. More sequencing of virus genome (or least the N protein) can help understand the picture. A properly designed pan-national study will be able to help understand the actual scenario after considering the confounding factors such as healthcare provision, gender, age distributions, economic condition, environment, nutrition, control measures of the country etc.

Discussion:

Hundreds of mutations have been reported in SARS-CoV-2 so far and the tally is increasing as more sequences being deposited in the public databases. It is often a challenge to make practical use of those sequences (and mutation) data. This study reports for the first time the rise and probable impacts of two strains SARS-CoV-2a and SARS-CoV-2 from the original SARS-CoV-2 strain after analyzing available sequence and COVID-19 case data. The mutually exclusive nature of these two strains may work as anchors to follow them both retro-and-prospectively.

The uniqueness of the trinucleotide mutations (28881-2883:GGG>AAC) makes it a highly potential candidate to follow the trend of the COVID-19 pandemic across regions caused by SARS-CoV-2. The molecular analysis presented in this paper has set the ground to assume that SARS-CoV-2a is linked with lower cases of infection because of the mutated SR-motif important for viral replication. However, this needs to be confirmed by i) further laboratory experiment on the particular location on the SR motif and ii) epidemiological research by matching the sequence data from different countries with their COVID-19 patients. Factors that may contribute to the GGG>AAC conversion should also be investigated. Demography, nutritional status, geographical location, environmental factors may play roles for this conversion as empirically SARS-CoV-2g (GGG) strains seem to be predominant in the megacities.

This study could explain the COVID-19 cases in different courtiers from where reliable data were obtainable. However, an explanation for the fatality difference still remains elusive. In a comparison between Lombardy and Abruzzo, it appears that the lethality is less in SARS-CoV-2a infected areas. This remains true when different regions of Netherlands were compared. However, when a country-wise comparison is made, the picture is not clear-cut. Notably, Germany (with a low prevalence of SARS-CoV-2a strains and higher COVID-19 cases) has much lower fatality compared to Netherlands or Brazil, both of which have a higher presence of SARS-CoV-2a. An obvious explanation is the difference in the healthcare provisions, age distributions and other local and policy differences in different countries.

Nevertheless, based on the information on the two strains of SARS-CoV-2, the fatality can be discussed from molecular perspective too. Among the mutations differences between the two strains as discussed above, it is particularly important to note that the ORF3a gene in the SARS-CoV-2a strain remains unmutated compared the SARS-CoV-2g where in many cases either 25563:G>A or 26144:G>A mutations are present in a mutually exclusive manner. It is already known that ORF3a plays a critical role to induce over reaction from inflammatory cytokines which often leads to the 'cytokine storms' [13], one of the most important reasons behind the fatality from COVID-19.

The complete absence of 25563:G>T and 26144:C>T mutations in the SARS-CoV-2a indicates that this strain will express an active ORF3a protein whereas more than 40% SARS-CoV-2g strains might be mutated for this gene (~33% 25563:G>T and ~9% 26144:G>T) (Figure-3).

This implies that SARS-CoV-2a, although will have less infectivity because of the mutated N protein, this strain might be more lethal than those SARS-CoV-2g with ORF3a mutations. This explanation is supported by the sequence data from Germany where 45% (N=52) strains are mutated for 25563:G>T and 6% (N=52) for 26144:G>T. This extrapolation should be considered with caution as there might be other attenuating mutations and confounding factors.

However, if 28881-28883:GGG>AAC is a decisive change that makes the SARS-CoV-2a less pathogenic compared to the SARS-CoV-2g, then 203-204:RG>KR positions of the N protein should be targeted to design drugs to affect the replication of the virus and thus reduce the pathogenicity of SARS-CoV-2 infection. Mathematical models to predict the course of the COVID-19 pandemic should consider the impact of 28881-2883:GGG>AAC mutation in the SARS-CoV-2 genome to better understand the course of the infection and guide nations' preparedness. For nations with no elaborate facilities for whole-genome sequencing, RT-PCR based testing should be recommended by targeting 28881-28883 region. This will give diagnostic information on COVID-19 together with the information on the two sub-strains: SARS-CoV-2a and SARS-CoV-2g in an infected person. This will allow gathering valuable information about the prevalence of these two strains are prevalent in those countries.

This work further recommends more active efforts to look into the genomes of the SARS-CoV-2 with closer pan-national collaboration to understand the transitions and distributions of the SARS-CoV-2a and SARS-CoV-2g strains for better understanding and management of COVID-19. Experimental and epidemiolocal research together with genome information will be key to make use of the analysis and assumptions presented in this paper.

Acknowledgement: I am particularly grateful to Dr. Bahar Tunçgenç, Dr. S.M Mahbubur Rashid and Dr. Mohammad Ali for their critical and valuable comments on the analysis and the manuscript preparation. I affectionately remember my niece Ananya Mohioshi and three nephews Abu Talha Shahid, Masrur Taz, and Ehsan Sajid who, despite being at their juvenile age, were curious about my work and listened to my explanations with enthusiasm. I dedicate this work to all the frontline healthcare professionals and other government and non-government agencies who are working hard to help people during this difficult time of COVID-19 epidemic.

Conflict of interest: I have no conflict of interest.

Reference:

1. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
2. McBride, R., van Zyl, M. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6**, 2991–3018 (2014).
3. Tylor, S. *et al.* The SR-rich motif in SARS-CoV nucleocapsid protein is important for virus replication. *Can. J. Microbiol.* **55**, 254–260 (2009).
4. Dyer, O. Two strains of the SARS virus sequenced. ;. *BMJ.* **326(7397):**, 999 (2003).
5. Marra, M. A. *et al.* The genome sequence of the SARS-associated coronavirus. *Science (80-.).* **300**, 1399–1404 (2003).
6. Yin, C. *Genotyping coronavirus SARS-CoV-2: methods and implications*.

7. Stefanelli, P. *et al.* Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. *Eurosurveillance* **25**, 2000305 (2020).
8. Menachery, V. D. *et al.* MERS-CoV accessory orfs play key role for infection and pathogenesis. *MBio* **8**, (2017).
9. Liu, C. I., Hsu, K. Y. & Ruaan, R. C. Hydrophobic contribution of amino acids in peptides measured by hydrophobic interaction chromatography. *J. Phys. Chem. B* **110**, 9148–9154 (2006).
10. Peng, T.-Y., Lee, K.-R. & Tarn, W.-Y. Phosphorylation of the arginine/serine dipeptide-rich motif of the severe acute respiratory syndrome coronavirus nucleocapsid protein modulates its multimerization, translation inhibitory activity and cellular localization. *FEBS J.* **275**, 4152–4163 (2008).
11. • Testing rate for COVID-19 select countries worldwide 2020 | Statista. Available at: https://www.statista.com/statistics/1104645/covid19-testing-rate-select-countries-worldwide/. (Accessed: 11th April 2020)
12. • Coronavirus cases worldwide 2020 | Statista. Available at: https://www.statista.com/statistics/1043366/novel-coronavirus-2019ncov-cases-worldwide-by-country/. (Accessed: 11th April 2020)
13. Siu, K. *et al.* Severe acute respiratory syndrome Coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB J.* **33**, 8865–8877 (2019).
14. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
15. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539–539 (2011).

Procedure:

Data collection:

This study has been conducted based on an analysis of the whole genome sequences of the SARS-CoV-2 from GISAID database. Not only the country-wise information was considered, but the study also took the advantage of region-based sequences deposited in the database. COVID-19 trackers, such as Microsoft Bing and Statista website were frequently used to get information about COVID-19 cases in the regions from where virus genome sequences have been deposited.

Firstly, the study looked at the sequences deposited from Italy, one of the worst affected countries where the death toll is very high. In Italy, Lombardy has experienced most cases and deaths from COVID-19, a big contrast with Abruzzo, which has a very low number of COVID-19 cases and deaths. When the region specific-sequences deposited in the GISAID database were examined, it was found that SARS-CoV-2 from Abruzzo stands out compared to other regions of Italy, especially of Lombardy. The most striking difference was the change in a bloc of three nucleotides at 28881-28883 location where a GGG>AAC change has occurred. Sequences from Abruzzo are predominantly 28881-28883:AAC whereas from Lombardy, those are 28881-28883:GGG
The study then expanded to look at more than 3000 whole-genome sequences from various regions around the globe and found a relationship between the presence of AAC strain in a region with the number of the COVID-19 cases there.

Data of COVID-19 cases and deaths were collected from Statista website, Bing COVID-19 tracker and whenever necessary from local government websites.
All reference sequences were used from NCBI virus database.

Analysis: The sequences downloaded from GISAID were analyses using Jalview [14]. Sequence alignment was performed using Clustal Omega [15]. Clusters based on various sequence features were built based on neighbor-joining algorithm from Jalveiw.
For protein structure prediction, JPred secondary structure prediction service was used. PyMOL software was used to see the view of the amino acids at specific positions.

Supplementary figures and table

Figure-S1: Genome arrangement of SARS-CoV-2. Adapted from the NCBI reference genome.

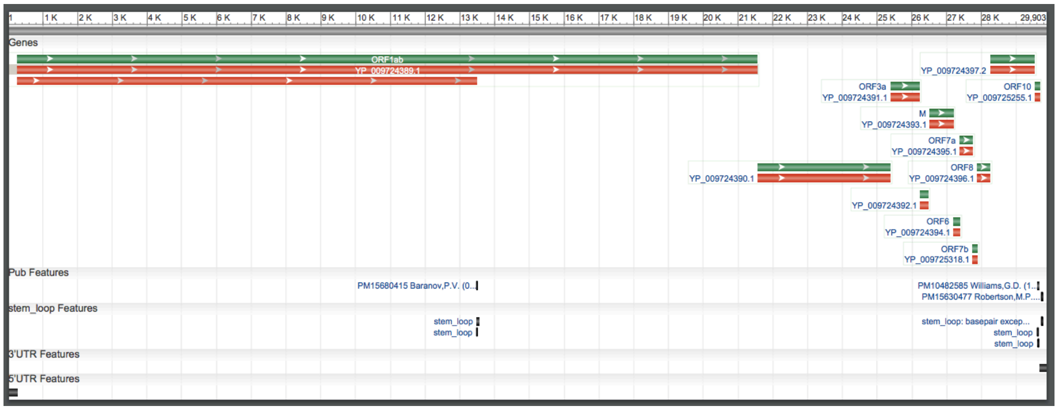

Figure-S2: Sequences from Japan indicate the relatively recent origin of SARS-CoV-2a.

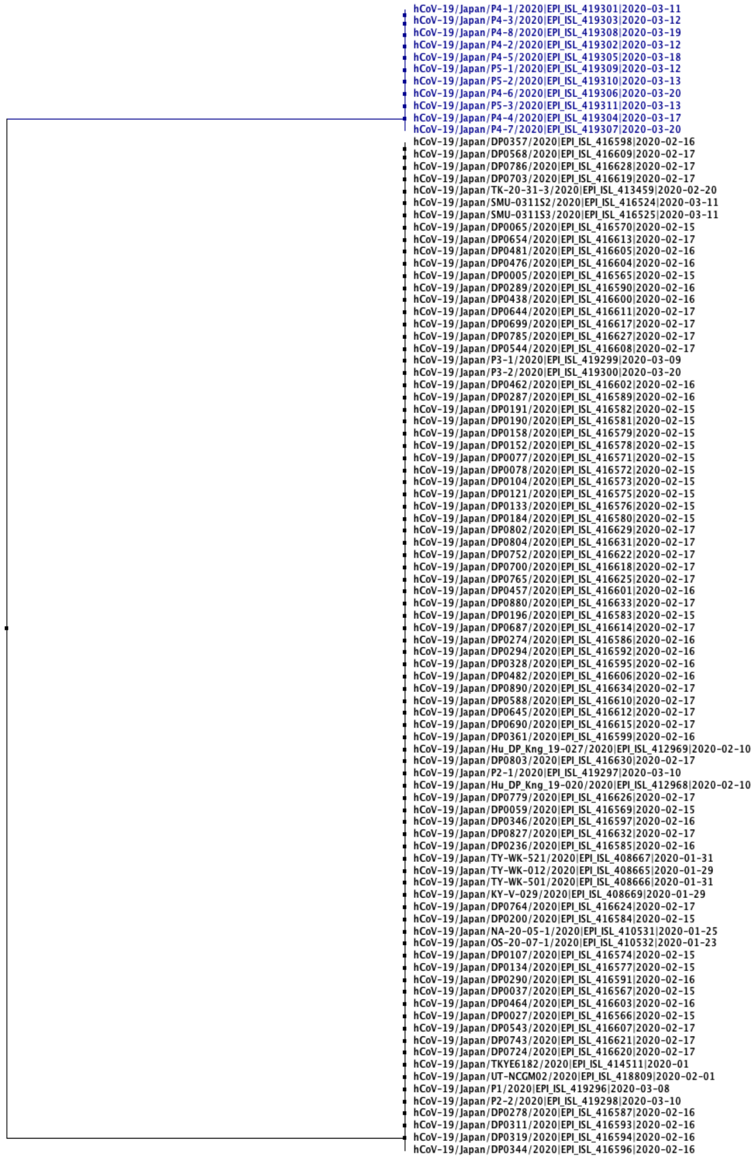
hCoV-19/Japan/P4-1/2020|EPI_ISL_419301|2020-03-11
hCoV-19/Japan/P4-3/2020|EPI_ISL_419303|2020-03-12
hCoV-19/Japan/P4-8/2020|EPI_ISL_419308|2020-03-19
hCoV-19/Japan/P4-2/2020|EPI_ISL_419302|2020-03-12
hCoV-19/Japan/P4-5/2020|EPI_ISL_419305|2020-03-18
hCoV-19/Japan/P5-1/2020|EPI_ISL_419309|2020-03-12
hCoV-19/Japan/P5-2/2020|EPI_ISL_419310|2020-03-13
hCoV-19/Japan/P4-6/2020|EPI_ISL_419306|2020-03-20
hCoV-19/Japan/P5-3/2020|EPI_ISL_419311|2020-03-13
hCoV-19/Japan/P4-4/2020|EPI_ISL_419304|2020-03-17
hCoV-19/Japan/P4-7/2020|EPI_ISL_419307|2020-03-20
hCoV-19/Japan/DP0357/2020|EPI_ISL_416598|2020-02-16
hCoV-19/Japan/DP0568/2020|EPI_ISL_416609|2020-02-17
hCoV-19/Japan/DP0786/2020|EPI_ISL_416628|2020-02-17
hCoV-19/Japan/DP0703/2020|EPI_ISL_416619|2020-02-17
hCoV-19/Japan/TK-20-31-3/2020|EPI_ISL_413459|2020-02-20
hCoV-19/Japan/SMU-0311S2/2020|EPI_ISL_416524|2020-03-11
hCoV-19/Japan/SMU-0311S3/2020|EPI_ISL_416525|2020-03-11
hCoV-19/Japan/DP0065/2020|EPI_ISL_416570|2020-02-15
hCoV-19/Japan/DP0654/2020|EPI_ISL_416613|2020-02-17
hCoV-19/Japan/DP0481/2020|EPI_ISL_416605|2020-02-16
hCoV-19/Japan/DP0476/2020|EPI_ISL_416604|2020-02-16
hCoV-19/Japan/DP0005/2020|EPI_ISL_416565|2020-02-15
hCoV-19/Japan/DP0289/2020|EPI_ISL_416590|2020-02-16
hCoV-19/Japan/DP0438/2020|EPI_ISL_416600|2020-02-16
hCoV-19/Japan/DP0644/2020|EPI_ISL_416611|2020-02-17
hCoV-19/Japan/DP0699/2020|EPI_ISL_416617|2020-02-17
hCoV-19/Japan/DP0785/2020|EPI_ISL_416627|2020-02-17
hCoV-19/Japan/DP0544/2020|EPI_ISL_416608|2020-02-17
hCoV-19/Japan/P3-1/2020|EPI_ISL_419299|2020-03-09
hCoV-19/Japan/P3-2/2020|EPI_ISL_419300|2020-03-20
hCoV-19/Japan/DP0462/2020|EPI_ISL_416602|2020-02-16
hCoV-19/Japan/DP0287/2020|EPI_ISL_416589|2020-02-16
hCoV-19/Japan/DP0191/2020|EPI_ISL_416582|2020-02-15
hCoV-19/Japan/DP0190/2020|EPI_ISL_416581|2020-02-15
hCoV-19/Japan/DP0158/2020|EPI_ISL_416579|2020-02-15
hCoV-19/Japan/DP0152/2020|EPI_ISL_416578|2020-02-15
hCoV-19/Japan/DP0077/2020|EPI_ISL_416571|2020-02-15
hCoV-19/Japan/DP0078/2020|EPI_ISL_416572|2020-02-15
hCoV-19/Japan/DP0104/2020|EPI_ISL_416573|2020-02-15
hCoV-19/Japan/DP0121/2020|EPI_ISL_416575|2020-02-15
hCoV-19/Japan/DP0133/2020|EPI_ISL_416576|2020-02-15
hCoV-19/Japan/DP0184/2020|EPI_ISL_416580|2020-02-15
hCoV-19/Japan/DP0802/2020|EPI_ISL_416629|2020-02-17
hCoV-19/Japan/DP0804/2020|EPI_ISL_416631|2020-02-17
hCoV-19/Japan/DP0752/2020|EPI_ISL_416622|2020-02-17
hCoV-19/Japan/DP0700/2020|EPI_ISL_416618|2020-02-17
hCoV-19/Japan/DP0765/2020|EPI_ISL_416625|2020-02-17
hCoV-19/Japan/DP0457/2020|EPI_ISL_416601|2020-02-16
hCoV-19/Japan/DP0880/2020|EPI_ISL_416633|2020-02-17
hCoV-19/Japan/DP0196/2020|EPI_ISL_416583|2020-02-15
hCoV-19/Japan/DP0687/2020|EPI_ISL_416614|2020-02-17
hCoV-19/Japan/DP0274/2020|EPI_ISL_416586|2020-02-16
hCoV-19/Japan/DP0294/2020|EPI_ISL_416592|2020-02-16
hCoV-19/Japan/DP0328/2020|EPI_ISL_416595|2020-02-16
hCoV-19/Japan/DP0482/2020|EPI_ISL_416606|2020-02-16
hCoV-19/Japan/DP0890/2020|EPI_ISL_416634|2020-02-17
hCoV-19/Japan/DP0588/2020|EPI_ISL_416610|2020-02-17
hCoV-19/Japan/DP0645/2020|EPI_ISL_416612|2020-02-17
hCoV-19/Japan/DP0690/2020|EPI_ISL_416615|2020-02-17
hCoV-19/Japan/DP0361/2020|EPI_ISL_416599|2020-02-16
hCoV-19/Japan/Hu_DP_Kng_19-027/2020|EPI_ISL_412969|2020-02-10
hCoV-19/Japan/DP0803/2020|EPI_ISL_416630|2020-02-17
hCoV-19/Japan/P2-1/2020|EPI_ISL_419297|2020-03-10
hCoV-19/Japan/Hu_DP_Kng_19-020/2020|EPI_ISL_412968|2020-02-10
hCoV-19/Japan/DP0779/2020|EPI_ISL_416626|2020-02-17
hCoV-19/Japan/DP0059/2020|EPI_ISL_416569|2020-02-15
hCoV-19/Japan/DP0346/2020|EPI_ISL_416597|2020-02-16
hCoV-19/Japan/DP0827/2020|EPI_ISL_416632|2020-02-17
hCoV-19/Japan/DP0236/2020|EPI_ISL_416585|2020-02-16
hCoV-19/Japan/TY-WK-521/2020|EPI_ISL_408667|2020-01-31
hCoV-19/Japan/TY-WK-012/2020|EPI_ISL_408665|2020-01-29
hCoV-19/Japan/TY-WK-501/2020|EPI_ISL_408666|2020-01-31
hCoV-19/Japan/KY-V-029/2020|EPI_ISL_408669|2020-01-29
hCoV-19/Japan/DP0764/2020|EPI_ISL_416624|2020-02-17
hCoV-19/Japan/DP0200/2020|EPI_ISL_416584|2020-02-15
hCoV-19/Japan/NA-20-05-1/2020|EPI_ISL_410531|2020-01-25
hCoV-19/Japan/OS-20-07-1/2020|EPI_ISL_410532|2020-01-23
hCoV-19/Japan/DP0107/2020|EPI_ISL_416574|2020-02-15
hCoV-19/Japan/DP0134/2020|EPI_ISL_416577|2020-02-15
hCoV-19/Japan/DP0290/2020|EPI_ISL_416591|2020-02-16
hCoV-19/Japan/DP0037/2020|EPI_ISL_416567|2020-02-15
hCoV-19/Japan/DP0464/2020|EPI_ISL_416603|2020-02-16
hCoV-19/Japan/DP0027/2020|EPI_ISL_416566|2020-02-15
hCoV-19/Japan/DP0543/2020|EPI_ISL_416607|2020-02-17
hCoV-19/Japan/DP0743/2020|EPI_ISL_416621|2020-02-17
hCoV-19/Japan/DP0724/2020|EPI_ISL_416620|2020-02-17
hCoV-19/Japan/TKYE6182/2020|EPI_ISL_414511|2020-01
hCoV-19/Japan/UT-NCGM02/2020|EPI_ISL_418809|2020-02-01
hCoV-19/Japan/P1/2020|EPI_ISL_419296|2020-03-08
hCoV-19/Japan/P2-2/2020|EPI_ISL_419298|2020-03-10
hCoV-19/Japan/DP0278/2020|EPI_ISL_416587|2020-02-16
hCoV-19/Japan/DP0311/2020|EPI_ISL_416593|2020-02-16
hCoV-19/Japan/DP0319/2020|EPI_ISL_416594|2020-02-16
hCoV-19/Japan/DP0344/2020|EPI_ISL_416596|2020-02-16

Figure- S3: 13 out of 128 sequences from Canada are of SARS-CoV-2a. However, 12 came from Ontario (out of 64 sequences from there) and one from British Columbia.

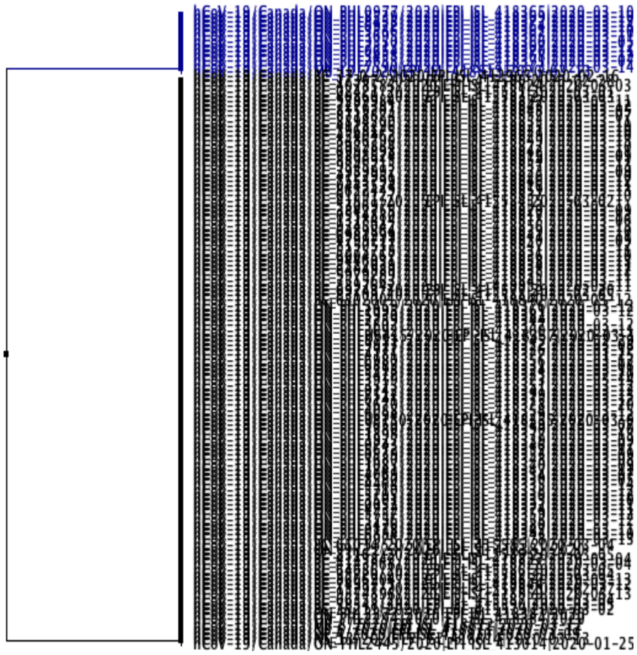
hCoV-19/Canada/ON_PHL9872/2020|EPI_ISL_418365|2020-03-10
hCoV-19/Canada/ON-PHL2445/2020|EPI_ISL_413614|2020-01-25

Table- ST1

Positions of bases that show >5% variation in the sequences of SARS-CoV-2a and SARS-CoV-2g

SAER-CoV-2a				
Position	Frequency of bases			
	C	T	G	A
13402		92	8	
19839	7	93		
27046	65	35		
SARS-CoV-2g				
Position	Frequency of bases			
	C	T	G	A
241	37	63		
1059	75	25		
1440			92	8
2891			92	8
3037	37	63		
8782	86	14		
11083		13	87	

14408	38	62		
14805	89	11		
15324	89	11		
17747	94	6		
17858			9	91
18060	91	9		
23403			63	37
25563		33	67	
26144		9	91	
28144	14	86		