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

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
BioAider: an efficient tool for viral genome analysis and its application in tracing SARS-CoV-2 transmission
Zhi-Jian Zhou¹#, Ye Qiu¹#, Ying Pu², Xun Huang², Xing-Yi Ge¹*

¹Hunan Provincial Key Laboratory of Medical Virology, Institute of Pathogen Biology and Immunology, College of Biology, Hunan University, Changsha, China
²Department of Hospital Infection Control Center, Xiangya Hospital of Central South University, Changsha, China

Running title: BioAider software and analysis of SARS-CoV-2 genome

#These authors contributed equally to this work.

*Correspondence to: Xing-Yi Ge, College of Biology, Hunan University, 35 Tianma Road, Yuelu District, Changsha, Hunan, China, xyge@hnu.edu.cn

Highlights:
- BioAider is an efficient tool for high-throughput analysis of viral genomes.
- BioAider monitors viral variation that facilitates epidemic control of COVID-19.
- 14 substitution hotspots in SARS-CoV-2 genome indicates viral polymorphism.
- NSP13-Y541C was found to be a crucial substitution key viral replication.
- The unique SRXX repeats on N protein suggests the animal origin of SARS-CoV-2.

Abstract
The novel human coronavirus (SARS-CoV-2) causes the coronavirus disease 2019 (COVID-19) pandemic worldwide. Control of COVID-19 pandemic is vital for public health and is the prerequisite to maintain social stability. However, the origin and transmission route of SARS-CoV-2 is unclear, bringing huge difficult to virus control.
Monitoring viral variation and screening functional mutation sites are crucial to prevention and control of infectious diseases. In this study, we developed a user-friendly software, named BioAider, for quick sequence annotation and mutation analysis on large-scale genome-sequencing data. Herein, we detected 14 substitution hotspots within 3,240 SARS-CoV-2 genome sequences, including 3 groups of potentially linked substitution. NSP13-Y541C was crucial substitution which might affect the unwinding activity of helicase. In particular, we discovered a SR-rich region of SARS-CoV-2 distinct from SARS-CoV, indicating more complex replication mechanism and unique N-M interaction of SARS-CoV-2. Interestingly, the quantity of SRXX repeat fragments in SARS-CoV-2 provided further evidence of its animal origin. Overall, we developed an efficient tool for rapid identification of viral genome mutations and could facilitate the viral genomic study. Using this tool, we have found critical clues for the transmission route of SARS-CoV-2 which would provide theoretical support for the epidemic control of pathogenic coronaviruses.

Key words: BioAider, SARS-CoV-2, COVID-19, sustainable development, mutant hotspots, SR-rich region

1. Introduction

The severe pandemic of Coronavirus Disease 2019 (COVID-19) is caused by a 2019 novel coronavirus (2019-nCoV) which was first detected and characterized at the end of 2019 (Wu et al., 2020; Zhu et al., 2020). Later, the name of this virus was suggested as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by the International
Committee on Taxonomy of Viruses (ICTV) (Coronaviridae Study Group of the International Committee on Taxonomy of, 2020). SARS-CoV-2 is a new pathogen and showing high infectiousness, fast spread, partial asymptomatic infection, and other new features (Sanche et al., 2020; Y. Wang et al., 2020). At present, the virus has been detected in the most countries and regions around the world, threatening public health, normal social life, and economy. Effective control and dealing with the COVID-19 could reduce the impact of the pandemic on the economy which is important for the sustainable development of society (Rahman et al., 2020). SARS-CoV-2 has infected more than ten million people with about 600,000 deaths worldwide as of July 19th, 2020 (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/). Compared with rural areas, the spread of coronavirus in urban megacities is more serious (V, R, & Haghighat, 2020). Current research shows that the virus is likely to spread through aerosol transmission, especially in the crowded place and enclosed space (Berlanga et al., 2020; Y. Zhang et al., 2019). Most patients infected with SARS-CoV-2 mainly present lower respiratory tract infections, accompanied by fever, dry cough, dyspnea (Huang et al., 2020). So far, there is no specific drug to treat SARS-CoV-2 infection and the vaccine against SARS-CoV-2 is still on the way. Therefore, blocking the transmission of the virus is of great importance for disease prevention and pandemic control (Megahed & Ghoneim, 2020). However, most key features of SARS-CoV-2, such as its origin and intermediate hosts, are still unclear, leading to difficulty in the control of virus spread. Considering the complicated interspecies and community transmission of SARS-CoV-2, the prevention and control of the epidemic requires the
joint efforts of multiple industries (Xu, Luo, Yu, & Cao, 2020).

Coronaviruses are enveloped virus with a non-segmented positive sense RNA genome, the full-length genome is the largest among RNA viruses (Rota et al., 2003). The first ORF encodes a polyprotein 1ab (pp1ab, ORF1ab) and approximately occupied the first two-thirds of the genome, the remainder are structural and non-structural proteins (Song et al., 2019). The pp1ab usually is hydrolyzed into 16 non-structural proteins (NSP1 ~ NSP16) by one or two papain-like protease (PLPs) in NSP3 and 3C-like protease (3CL\textsuperscript{pro}) in NSP5, to promote viral RNA replication and transcription (Fehr & Perlman, 2015; Ziebuhr, 2005). For these non-structural proteins in pp1ab, NSP2 can inhibit two host proteins of PHB1 and PHB2, which play an important role in viral infection (Cornillez-Ty, Liao, Yates, Kuhn, & Buchmeier, 2009). The PLPs of NSP3 hydrolyze the N-terminal of pp1ab to NSP1 ~ NSP4, 3CL\textsuperscript{pro} of NSP5 binds 11 conserved Q-S dipeptide sites in pp1ab to produce 12 mature non-structural proteins (NSP5 ~ NSP16) (Fehr & Perlman, 2015; Lei, Kusov, & Hilgenfeld, 2018). NSP12 is a key component, also known as the RNA-dependent RNA polymerase (RdRp) and is crucial in the replication and transcription cycle of coronaviruses (Subissi et al., 2014). Currently, RdRp is considered to be the main target of antiviral drugs for SARS-CoV-2 (Gao et al., 2020). NSP13 owns NTPase/Helicase activity and can unwind double-stranded RNA and DNA helix (Jang, Lee, Yeo, Jeong, & Kim, 2008). Due to the conservation of NSP13 is necessity in all coronaviruses species, it is considered an ideal target for the development of antiviral drugs for SARS-CoV (Jia et al., 2019; Shum & Tanner, 2008). The main four structural proteins of coronaviruses are the spike protein
(S), small envelope protein (E), membrane protein (M) and nucleocapsid protein (N) (Cui, Li, & Shi, 2019). S protein interacts with molecular receptors to mediate cell membrane fusion, allowing viruses to enter host cells (Li, 2016). E and M proteins are involved in the assembly of the virus, which is related to the formation and release of the viral envelope, N protein plays an important role in virus replication and pathogenesis (Li, 2016; Schoeman & Fielding, 2019).

Based on seven conserved replicase domains in pp1ab (polyprotein-1ab) for the classification of coronavirus, SARS-CoV-2 belongs to the Sarbecovirus subgenus of the genus Betacoronavirus in the subfamily Orthocoronavirinae together with human SARS-CoV and bat various SARSr-CoVs (SARS-related CoVs), and SARS-CoV-2 is highly similar to bat coronavirus Bat-CoV-RaTG13 genetically (International Committee on Taxonomy of Viruses Executive, 2020; P. Zhou et al., 2020; Zhu et al., 2020). Similar to SARS-CoV, SARS-CoV-2 uses the angiotensin converting enzyme II (ACE2) as receptor, besides, the serine protease TMPRSS2 plays an important role in activating spike (S) protein (Hoffmann et al., 2020). Some novel features of SARS-CoV-2 has been revealed, like a furin protease cleavage site in the Spike which may be related to viral transmissibility (Q. Wang et al., 2020). Recently, a novel bat-derived coronavirus RmYN02 with the similar furin insertion in the S was found which emphasized the bat origin of the virus (H. Zhou et al., 2020).

As a member of RNA virus, the RNA-dependent RNA polymerase (RdRp) encoded by coronavirus lacks the proofreading capability, which leading to high mis-incorporation rate during replication even with the help of exoribonuclease like ExoN
Together with the genome recombination events, it could cause viral diversity and promote the transmission and disease of coronavirus (Hu et al., 2017). At the same time, in order to better adapt to the host, the virus usually mutates continuously under the selection pressure of the host. In particular, some non-synonymous substitution sites with high frequency are often experience strong positive selection (Pond et al., 2006). Considering the continuous increase of infected people and the high variability of the virus, it is important to pay attention to the genomic changes of SARS-CoV-2. Recently, 7 substitution hotspots in SARS-CoV-2, ORF1ab-G10818T (ORF1ab-L3606F), ORF1ab-C8517T, ORF3a-G752T (ORF3a-G251V) S-A1841G (D614G), G171T (Q57H), ORF8-T251C (ORF8-L84S) and N-GGG608_609_610AAC (N-RG203_204KR) have been reported, (Capobianchi et al., 2020; Ceraolo & Giorgi, 2020; Issa, Merhi, Panossian, Salloum, & Tokajian, 2020; C. Wang et al., 2020). In addition, the non-synonymous substitution of ORF3a-G251V and ORF8-L84S both cause the amino acid (aa) polarity changing, which may affect the conformation of the protein and lead to function altering (Ceraolo & Giorgi, 2020).

Despite these recent discoveries, in order to deal with the continuous variation of SARS-CoV-2 and the accumulation of sequencing data, methods for quick and efficient analysis of the mutations feature and substitutions hotspots of SARS-CoV-2 genome are urgently required. However, tools for sequence alignment were common, such as Clustal, Muscle or MAFFT, but sequence variation analysis downstream was scarce. Although the analysis of a large number of variant sequences could be accomplished
well through scripts, it is difficult for biological or clinical expert without bioinformatics and programming skills. Some tools such as MutPred can analyze variations by entering amino acid sequence, but unable to handle synonymous substitutions and consecutive nucleotide mutations like dinucleotide and trinucleotide substitution (Pejaver et al., 2017). In comparative genomic studies, it is very important to identify whether amino acid properties have changed when dealing with non-synonymous substitutions, which can help to identify some important mutations. In this respect, we have developed an interactive analysis tool, Bioinformatics Aider V1.0 (BioAider V1.0). BioAider showed high efficiency and convenient in gene annotation and mutation screening on multiple genome-sequencing data.

In this research, we collected 3240 complete genome sequences of SARS-CoV-2 from 64 different countries and regions, sampling time from December 24, 2019 to April 1, 2020. We conducted a detailed genome mutation analysis using BioAider. We identified 14 substitution hotspots and 3 groups of possible linkage substitutions. Especially, we found distinctive polymorphism on SR-rich region of N protein in SARS-CoV-2 and related coronaviruses in other animals. Our work provides a new tool for recognition of the variation and evolution of SARS-CoV-2, contributes to research the replication and pathogenic mechanism of SARS-CoV-2, and gives further evidence for the animal origin of SARS-CoV-2.

2. Methods and materials

2.1 Working principle of fast annotation and mutation analysis in BioAider
BioAider provides a convenient function for annotation on homologous sequence. First, import the aligned complete genome sequence set, and adjust the reference sequence for gene extraction to the forefront of the sequence set. Paste the related gene information of reference sequence in the input box, gene name, start string and end string, which separated by comma. Then BioAider can batch extract genes and annotation from a large number of sequences. Note that, the length of start string or end string is not limited in, but it is required to be unique in the reference sequence.

Second, for mutation analysis, BioAider will scan all the codons in the aligned sequence, using the first sequence in the data set as a reference sequence. BioAider can identify five different mutation types (synonymous, non-synonymous, insert, deletion and termination) based on the standard codon method. It can distinguish the changes in properties of amino acid when dealing non-synonymous substitutions. Then, it will locate the position of the corresponding base in the mutated codon, so it can identify multiple types of mutation at the same site, and also well discriminate dinucleotide and trinucleotide substitution. Next, it summarizes all the mutation sites with corresponding frequency and strains, and if user chooses to generate the frequency distribution of synonymous or non-synonymous sites, BioAider will directly generate the results in vector format image. Of note, the frequency distribution map does not include those sites that are both synonymous and non-synonymous substitution, because such sites cannot determine the common substitution frequency.

2.2 Mutation analysis of SARS-CoV-2
The 3240 complete genome sequences of SARS-CoV-2 with relatively higher quality of sequencing were downloaded from GISAID (https://www.gisaid.org/), and the reference genome sequence of SARS-CoV-2 (NC_045512.2) for ORF annotation was from GenBank (https://www.ncbi.nlm.nih.gov/genbank). All the viral strains used in this study were listed in Table S1. Multiple sequence alignment of genomic sequences of SARS-CoV-2 were accomplished using MAFFT v7.407 (Katoh & Standley, 2013; D. Zhang et al., 2020). The annotation and extraction of codon genes of these 3240 SARS-CoV-2 genome sequences using BioAider V1.0. We extracted 11 continuously coding genes based on the annotation information of NC_045512.2, including ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N and ORF10. Then we used MUSCLE program in MEGA v7.0.14 to align these coding genes based on codons method (Kumar, Stecher, & Tamura, 2016).

We combined these 11 continuously coding genes to tandem sequence in BioAider and used it to represent the complete genome sequence of SARS-CoV-2 for subsequent analysis. Among all the early sampled virus strains (Dec 24, 2019 to Dec30, 2019), genome of Wuhan/IVDC-HB-01/2019 (GISAID ID: EPI_ISL_402119, sampled in Dec30, 2019) occupied most frequently (65 strains sequence were same with IVDC-HB-01/2019) in 3240 sequenced strains and appears in multiple regions, therefore, the strain of EPI_ISL_402119 was used the reference sequence for genome variation analysis of SARS-CoV-2 in BioAider V1.0. For these possible linkage substitution hotspots, we extracted corresponding viral sequences which contained these sites using BioAider.
2.3 Structure prediction of protein

The tertiary protein structure was downloaded from the Protein Data Bank (PDB, http://www.rcsb.org), and protein homology modeling was calculated by online tools SWISS-MODEL then using Chimera 1.10.2 to compare protein models (Pettersen et al., 2004; Waterhouse et al., 2018).

3. Results

3.1 Function overview and examples of BioAider

BioAider V1.0 was developed based on Python 3.7 and R 3.5, and used PyQt5 for interface packaging. BioAider puts complex algorithms and fault-tolerant processing mechanisms inside, presenting users a very simple interface and prompts are added to interface controls. The function of BioAider are divided into three main sections (Fig 1). The SeqTools was used for common sequence processing, such as Split Sequence Fragment, Combine Gene, and sequence Fast Annotation (Fig 2A). Similar analysis, including two functions of sequence identity matrix and remove high similar sequence by specify the threshold. Especially, BioAider owns the ability to acquire the identity of nucleotide and amino acid simultaneously, and provides optional features of gap compression in Sequence Identity Matrix. For the Mutation Tools, three nested functions are applied for genome variation analysis. The Site Counter can summary the type, count and proportion of nucleotide (or amino acids) at each site for the aligned sequence datasets, the Site Scree is used for extract the sequence of the specified site
In particular, in addition to the highly summarized analysis results, the function of Mutation Analysis also provides high-quality vector graphics of synonymous or non-synonymous substitution frequency distribution for publication (Fig 2C).

3.2 Mutation sites in SARS-CoV-2 genomes

In this study, ‘substitution’ refers to synonymous or non-synonymous substitution, and ‘mutation’ includes all the genomic variance. The frequency of mutation (or substitution) refers to the number of mutant strains compared to the reference strain sequence in this study.

Compared with the early viral sampling strains (EPI_ISL_402119) in GISAID database, a total of 2152 mutation sites (regardless of insertions or deletions) were identified in 3239 SARS-CoV-2 sequenced whole genomes, accounting for 7.36% of SARS-CoV-2 full-length gene (Table 1). The number of synonymous and non-synonymous substitution sites was counted as 784 (2.68%) and 1335 (4.57%), respectively. Among these non-synonymous substitution sites, 738 resulted in changes in amino acid properties. Besides, 12 sites contained both synonymous and non-synonymous substitution depending on the strains, and a total of 21 termination mutation sites were also detected. In all the coding genes, three with the of the most mutation sites were ORF1ab, spike (S) and nucleocapsid (N) gene, the number of mutations sites were 1400, 296 and 168, respectively, accounting for 6.58%, 7.75% and 13.36% in their corresponding gene length. The gene with the highest proportion of
mutation sites was ORF10, containing 16 mutation sites were in 114 bases which occupied 14.04% of ORF10. All the mutation sites and viral strains of SARS-CoV-2 detected by BioAider in this study were summarized in Tables S2 and S3.

3.3 Substitution frequency distribution of synonymous or nonsynonymous sites
To assess the overall substitution frequency of these mutated sites, we divided the substitution frequency into six different groups, and drew the frequency spectra of 2119 substitution sites (784 synonymous and 1335 nonsynonymous) of 3239 sequenced strains (Fig 3). The result shows that the substitution sites of non-synonymous were always over synonymous except for the fifth group. Besides, a large number of sites with substitution frequencies between 1 and 5, and more than half of the substitution sites only are observed in a single strain. We also found 60 sites owning a substitution frequency greater than 20, including 40 non-synonymous substitution sites and 20 synonymous substitution sites. The substitution frequency distribution of synonymous or nonsynonymous sites for each codon gene are in Fig S1.

3.4 Substitution hotspots in SARS-CoV-2 genomes
Totally, 2119 substitution sites were detected in 3239 sequenced strains, most of them with a lower frequency. We defined the site with substitution frequency over 200 as the substitution hotspot, thus far, 14 substitution hotspots were identified which distributed in ORF1ab, S, ORF3a, ORF8 and N gene (Fig 4). Among these substitution hotspots, 10 sites were non-synonymous and 4 were synonymous. As for these non-synonymous
substitution hotspots, 6 substitution sites caused a change on polarity or chargeability of amino acid. Especially, there was a trinucleotide substitution hotspot on the N gene and caused two amino acid changes.

3.5 Spatial distribution characteristic of SARS-CoV-2 in each substitution hotspot

According to the substitution hotspots, we divided two geographical regions, China and outside of China, and studied the distribution differences of mutant and referential type in geographical area at each substitution hotspots, respectively (Fig 5). As the result showing, except substitution hotspots of ORF1ab-10818 and ORF3a-752, mutant type and referential type in the 12 other substitution hotspots showed significant spatial distribution differences (p<0.01, chi-square test) between China and outside of China. Furthermore, we found in substitution hotspots of ORF1ab-8517 and ORF8-251C, the mutant type (ORF1ab-8517T or ORF8-251C) owned a higher ratio and were more prevalent in China than outside of China, opposite to other 10 substitution hotspots.

3.6 Linkage among these substitution hotspots

Make further efforts, we found some hotspots contains substitutions with similar patterns and frequency (Table 2), indicating potential connection among these substitution hotspots, and then we found 3 groups of possible linkage substitution. In order to facilitate research, we used gene names and base position to mark them, these linkage substitution hotspots were most likely ORF1ab-2772 & ORF1ab-14144 & S-1841, ORF1ab-8517& ORF8-251 and ORF1ab-17483 & ORF1ab-17594 & ORF1ab-
To test our hypothesis, we compared the number of strains between referential type and mutant type which synchronously mutated at the possible linkage substitution hotspots (Table 3). For each combination of possible linkage substitution hotspots, we found mutant type and referential type in accounts for more than 98% of the population, it implied that the genetic variants in these combined substitution hotspots were not independent, but linkage.

3.7 Vital substitution hotspots in SARS-CoV-2 NSP13

Given 541<sup>th</sup> aa was one of known key sites of NSP13 for binding to nucleic acids in SARS-CoV, the non-synonymous substitution hotspot of ORF1ab-Y5865C (NSP13-Y541C) in NSP13 probably affects the function of NSP13. We conducted protein model prediction of SARS-CoV-2 NSP13 by homology protein modeling. The result showed that closest protein model of NSP13 to SARS-CoV-2 was from SARS-CoV (PBD ID: 6JYT, 99.83% amino acid identity). As the result shown, the tertiary structure of NSP13 were almost completely overlapping between SARS-CoV-2 and SARS-CoV (Fig 6A). Besides, the 541<sup>th</sup> aa was relatively conservative in SARS-CoV-2 related animal coronaviruses (Fig 6B).

3.8 Non-synonymous substitutions sites near furin cleavage region

Two non-synonymous substitution sites near the furin cleavage region of PRRA on S protein were identified (Fig 7 and Table 2). There were 7 mutant strains with S-G2025C
(S-Q675H) and one with S-A2024G (S-Q675R) among 3239 sequenced strains. 675th aa on S protein was the sixth amino acid upstream of PRRA, and we found S-Q675H and S-Q675R made original polar amino acids from no-charged to positively charged.

3.9 Distinctive polymorphism in SR-rich fragment of SARS-CoV-2

We found a continuous variable area with non-synonymous substitutions from 183th to 204th aa on the N protein in SARS-CoV-2 (Fig 8), and there were 5 variably sites with non-synonymous substitutions frequency over 20, including a trinucleotide substitution which led to two consecutive aa substitutions of R203K and G204R (Table 2). Notably, there were 2 strains substituted on the 196th and 201th codons of SARS-CoV-2, too, but they were synonymous substitutions which did not cause amino acid substitutions (Table S2).

The variable area was rich in Ser (S) and Arg (R) and contains SRXX repeat fragments. We compared the region among SARS-CoV-2, SARS-CoV-2 related coronavirus, SARS-CoV and SARSr-CoV in this area (Fig 8), and found that it was relatively conserved in SARS-CoV-2 related coronavirus, SARSr-CoV and SARS-CoV, but showed distinctive polymorphic in SARS-CoV-2. Interestingly, we found 4 SRXX repeat fragments in most strains of SARS-CoV-2, SARS-CoV-2 related coronavirus strains, but SARSr-CoV and SARS-CoV lacked the third SRXX repeat fragments due to one amino acid substitution. We found the number of SRXX repeat fragments well reflects the evolutionary relationship among SARS-CoV-2, SARS-CoV-2 related coronavirus, SARSr-CoV and SARS-CoV. Especially, the substitution of two
continuously amino acids on the last SRXX repeat fragments (203\textsuperscript{th} and 204\textsuperscript{th} in SARS-CoV-2) was exclusively observed in SARS-CoV-2.

To explore the polymorphism of SR-rich region (aa 183-204) on SARS-CoV-2 N protein, we intercepted the amino acids in this region. After culling some strains sequence containing degenerate bases in the region that could not be translated normally, we screened the SR-rich regions in the remaining 3233 strains. The results showed the SR-rich regions in SARS-CoV-2 can be divided into 23 different types (Table 4), including the reference type. The two main types were the reference type (2561 strains) and the mutant type of N-R203K\_G204R (494 strains). The majority of the 22 mutant types harbored only single-amino-acid substitution compared to the reference type, indicating that this region was relatively conserved among most SARS-CoV-2 strains. However, we could still find some distinctive polymorphism among different strains. The sampling date showed that most mutations existed in the 3233 recently sequenced strains, implying a constant evolution of the SR-rich region in SARS-CoV-2 genome.

4. Discussion
COVID-19 pandemic poses a great challenge for the healthy cities and societies, its prevention and control requires cross-disciplinary cooperation and research, including biology, industry, medicine and computer information technology (Xu et al., 2020). Computational biology, big data and artificial intelligence are playing a more important role in daily monitoring, prevention and treatment of infectious diseases, contributing to maintaining the sustainable development of society (Liu & Li, 2020; Zhou, He, Cai,
Wang, & Su, 2019). With the continuous variation of SARS-CoV-2, program tools for high-throughput analyses on big sequencing data of genome are of great importance for genomic and evolutionary studies. More importantly, simplicity of operation and high summary of analysis results could save a lot of time for researchers.

Some graphic-interface bioinformatics programs, such as MEGA, DNAMAN and Geneious, are commonly used to browse or align the sequences of genes and genomes, but they are not suitable for the mutation analysis of large-scale datasets (Kearse et al., 2012; Kumar et al., 2016). For instance, SNP-sites can screen the mutant sites in datasets of DNA sequences by sequence alignment with the reference sequence, but it cannot tell the types of those mutations (e.g. synonymous mutation or non-synonymous mutation) which are important for viral genome analyses (Page et al., 2016). In addition, SNP-sites is a command-line program which is not friendly to users not familiar with bioinformatic programs. Another program named MutPred is capable of analyzing amino acid sequences but not nucleic acid sequences, thus it cannot reveal the variance on viral genomes (Pejaver et al., 2017).

Here, we present BioAider, an efficient graphical tool for mutation analysis of viral genome and could processes nucleic acid and amino acid sequence analysis in parallel. After several simple steps of operation including dataset input and parameter setup, BioAider will process the analyses automatically and outputs comprehensive and visualized graphs presenting all the information about the mutations in the input genomes, such as mutation type list (e.g. synonymous mutation, non-synonymous mutation, insertion, deletion or nonsense mutation), mutation frequency and changes in
amino acid properties. Meanwhile, BioAider is designed for high-throughput analyses, and this interface makes it more efficient experience for the handling of big data. BioAider greatly improve the efficiency for genome variation analysis of SARS-CoV-2, contributing to the study of SARS-CoV-2 and the development of healthy smart cities.

In addition, BioAider harbors a built-in database of all 25 amino acid codon systems (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?) and can be used for the analyses of genomes from other organisms such as bacteria and archaea. The expansion and optimization of BioAider in the future will endow the software with more functions.

In this study, we conducted high-resolution analysis of mutations within SARS-CoV-2 genome based on 3240 sequenced strains using BioAider. 2152 mutation sites among different strains were detected which accounted for 7.36% of the complete genome. These mutation sites included 1,335 non-synonymous substitution sites, implying abundant variations in the genome of SARS-CoV-2. However, more than half of the mutation sites were only observed in a single strain of SARS-CoV-2 which was difficult to be distinguished from sequencing errors. Considering there were only 60 sites with substitution frequency above 20. Therefore, we speculated that there were no large-scale mutations in the sequencing strains we analyzed.

According to the spatial distribution analysis between strains of mutant types and referential type regarding the substitution hotspots, 12 of 14 sites showed significant distribution differences in China and outside of China, indicating different virus prevalence in the two regions and may evolve in different directions in the future. We also noticed mostly strains sampled in China were before Mar 2020, this was due to the
epidemic of China has been roughly controlled in March. Considering these facts, we speculate that virus differentiation in and outside of China may attribute to human intervention. Besides, these non-synonymous substitution hotspots should be given more attention, they were most likely related to the phenotype of the virus. The latest study has discovered the strain with substitution of S-D614G could increases the infectivity of SARS-CoV-2 (Korber et al., 2020).

In SARS-CoV NSP13, the 541\textsuperscript{th} aa was critical for the protein function and double mutations of S539A/Y541A showed higher unwinding activity for nucleic acids binding (Jia et al., 2019). The amino acid identity of NSP13 between SARS-CoV-2 and SARS-CoV was 99.83\% and their tertiary structure was almost completely overlapping, indicating that NSP13 541\textsuperscript{th} aa is also a vital site for the function of SARS-CoV-2 NSP13. Notably, the mutant strain of NSP13-541C first appeared in the United States (sampled on Feb 20, 2020), and the United States was the country with the largest number and extremely high proportion of this mutant strains (316, 78\% of all the 405 mutant strains with NSP13-Y541C). We also noted that there were a total of 711 strains sampled in the United States on Feb 20, 2020 and beyond, strains with NSP13-Y541C accounted for almost half of these strains in the United States. This fact indicates that NSP13-Y541C has gone through less negative selection pressure and even this variation may be beneficial. Therefore, we speculate that NSP13-Y541C improves the unwinding activity of NSP13 and promotes the replication of SARS-CoV-2, contributing to their rapid spread in the United States. However, further studies are required to reveal the detailed effect of substitution in 541\textsuperscript{th} aa of SARS-CoV-2 NSP13.
and whether the possible linkage substitution site of 504\textsuperscript{th} aa plays a synergistic role in this process.

Among the potential linkage substitution hotspots identified in this study, ORF1ab-8517 and ORF8-251 were observed in a recent research (Ceraolo & Giorgi, 2020; Tang et al., 2020). Among the triple linkage substitution hotspots of ORF1ab-277, ORF1ab-14144 and S-1841, ORF1ab-C2772T belongs to synonymous substitution, while ORF1ab-C14144T caused the aa change of NSP12-P323L in interface region which is a bridge section connecting NiRNA (nidovirus RdRp-associated nucleo-tidyltransferase) and Fingers of RdRp (Gao et al., 2020). The 614\textsuperscript{th} aa in S protein was located in the subdomain (SD) region downstream of the receptor-binding domain (RBD) on S1. We found that mutant strains with NSP12-323L & S-614G were popular in the world and owned more than half the frequency (52\%) in the population, indicating that this mutant was dominant in SARS-CoV-2. However, the functional impact of linkage substitution in these two sites is still unclear at present, and whether NSP12-323 and S-614 are related on specific function needed more experimental data to verify.

Previous studies have reported that SR-rich region in SARS-CoV is crucial for N protein multimerization and the interaction with membrane (M) protein (He, Dobie, et al., 2004; He, Leeson, et al., 2004). In the SR-rich region of N protein, there was only 2-aa difference between SARS-CoV-2 (referential type) and SARS-CoV. Therefore, the similar function of SR-rich region may exist in SARS-CoV-2, although there is no relevant research reported at present. Compared to SARS-CoV, SARS-CoV-2 harbors
one more SRXX repeat fragments in the 193\textsuperscript{th}-196\textsuperscript{th} aa of N protein. Previous research reported that the SR-rich region of SARS-CoV in 184\textsuperscript{th}-196\textsuperscript{th} aa was crucial for N protein multimerization and the deletion of this region completely would make N protein abolish the self-multimerization (He, Dobie, et al., 2004). Besides, SRXX repeat fragments may play an important role in SARS-CoV infection (Luo et al., 2005). We noted that SARS-CoV could be regarded as a deletion of SRXX in SR-rich region compared to SARS-CoV-2, and the transmission rate of SARS-CoV-2 was higher than SARS-CoV, thus, whether SR-rich region of SARS-CoV-2 plays an important role in this process is worth to explore. However, the research about SR-rich region in coronavirus was still limited at present. Especially, unlike SARS-CoV, the region in SARS-CoV-2 is still constantly evolving, implying that the N protein of SARS-CoV-2 may employ a more flexible replication mechanism and even interaction with M protein. A study reported that the N protein of SARS-CoV can specifically bind to heterogeneous nuclear ribonucleoproteins (hnRNPs) A1 and plays an important role in RNA replication and transcription, especially, the key binding region is in the SR-rich region of SARS-CoV N protein (aa 161-210), the interaction between human hnRNP A1 and SARS-N protein may be the key to SARS-CoV replication and transcription (Luo et al., 2005). Whether such a similar mechanism exists in SARS-CoV-2 is still unclear. However, in human cells, there are more than 20 hnRNPs has been discovered, and SR-rich region of SARS-CoV-2 showing distinctive polymorphism, whether the SARS-CoV-2 can bind to one of these hnRNPs needs more verification. It might provide a new hint in understanding the process of SARS-CoV-2 replication in human
cells. Besides, the potential phenotype changes related to nonsynonymous substitution hotspots on third SRXX repeat fragments of SARS-CoV-2 may be also worth attention.

In the hotspot screening, we have found several substitutions critical for viral replication potentially. For instance, the ORF1ab-C794T was located on NSP2 and NSP2 was reported to inhibit the host protein PHB1 and PHB2 which benefited viral replication, indicating that ORF1ab-C794T may affect viral replication (Cornillez-Ty et al., 2009). The ORF3a-G171T (Q57H) was located at the transmembranous domain of the 3a protein (Fig S2). 3a protein was reported to form ion channels on host cell membrane and enhance the membrane permeability which benefited SARS-CoV life cycle (Minakshi, Padhan, Rehman, Hassan, & Ahmad, 2014). Thus, ORF3a-G171T (Q57H) may affect the formation of ion channels and subsequently influence the viral replication. As for ORF3a-G752T (G251V), a recent analysis predicted that it was in an important functional domain and might be related to virulence, infectivity, ion channel formation and virus release (Issa et al., 2020). We also found several important sites for virus entry, though they showed lower substitution frequency in our research data. For instance, S-Q675H and S-Q675R near furin cleavage region possibly influence the cleavage of RRAR, a critical step for virus entry (Fig 7). We also detected a strain with the mutation of S-R408I located in the RBD (Table S2) which was reported to play an important role in virus-receptor binding by a recent study (Saha, Banerjee, Tripathi, Srivastava, & Ray, 2020).

5. Conclusion
BioAider greatly simplifies annotation and genome variation analysis of large-scale sequences data. We initially revealed the variation characteristics of SARS-CoV-2 and predicted the possible impact for some non-synonymous substitution hotspots based on BioAider, contributing to study the formation mechanism and evolution of SARS-CoV-2. The distinctive polymorphism SR-rich region of N protein in SARS-CoV-2 provides a new clue to establish anti-virus strategy for viral replication. In this study, we could not include all the sequences of SARS-CoV-2 due to the rapid increase of the genomic data. More researches or real time analysis with relevant clinical data may contribute to the viral epidemiology of SARS-CoV-2 and treatment of COVID-19. The BioAider software could help doctors or those who are not familiar with bioinformatics to analyze the viral genomes, which is conducive to prevention and control of the COVID-19 or similar viral infectious diseases.

**Declaration of interests**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgements**

This work was jointly funded by National Natural Science Foundation of China (grant number 32041001 and 81902070), National Key Research and Development Program of China (grant number 2017YFD0500104), the Provincial Natural Science Foundation of Hunan Province (grant number 2019JJ20004, 2019JJ50035, and 2020SK3001), and the Emergency Project of Prevention and Control for COVID-19 of
Central South University (grant number 160260003).

**Data accessibility**

BioAider is available from https://github.com/ZhijianZhou01/BioAider/releases, and the detailed user manual is available from https://github.com/ZhijianZhou01/BioAider.

**References**

Berlanga, F. A., Liu, L., Nielsen, P. V., Jensen, R. L., Costa, A., Olmedo, I., & de Adana, M. R. (2020). Influence of the geometry of the airways on the characterization of exhalation flows. Comparison between two different airway complexity levels performing two different breathing functions. *Sustainable Cities and Society, 53*, 101874. doi:10.1016/j.scs.2019.101874

Capobianchi, M. R., Rueca, M., Messina, F., Giombini, E., Carletti, F., Colavita, F., . . . Bartolini, B. (2020). Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. *Clin Microbiol Infect*. doi:10.1016/j.cmi.2020.03.025

Ceraolo, C., & Giorgi, F. M. (2020). Genomic variance of the 2019-nCoV coronavirus. *J Med Virol, 92*(5), 522-528. doi:10.1002/jmv.25700

Cornillez-Ty, C. T., Liao, L., Yates, J. R., 3rd, Kuhn, P., & Buchmeier, M. J. (2009). Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. *J Virol, 83*(19), 10314-10318. doi:10.1128/JVI.00842-09

Coronaviridae Study Group of the International Committee on Taxonomy of, V. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-
nCoV and naming it SARS-CoV-2. *Nat Microbiol, 5*(4), 536-544. doi:10.1038/s41564-020-0695-z

Cui, J., Li, F., & Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol, 17*(3), 181-192. doi:10.1038/s41579-018-0118-9

Denison, M. R., Graham, R. L., Donaldson, E. F., Eckerle, L. D., & Baric, R. S. (2011). Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol, 8*(2), 270-279. doi:10.4161/ma.8.2.15013

Fehr, A. R., & Perlman, S. (2015). Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol Biol, 1282*, 1-23. doi:10.1007/978-1-4939-2438-7_1

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., . . . Rao, Z. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science, 368*(6492), 779-782. doi:10.1126/science.abb7498

He, R., Dobie, F., Ballantine, M., Leeson, A., Li, Y., Bastien, N., . . . Li, X. (2004). Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem Biophys Res Commun, 316*(2), 476-483. doi:10.1016/j.bbrc.2004.02.074

He, R., Leeson, A., Ballantine, M., Andonov, A., Baker, L., Dobie, F., . . . Li, X. (2004). Characterization of protein-protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Res, 105*(2), 121-125. doi:10.1016/j.virusres.2004.05.002

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., Erichsen, S., . . . Pohlmann, S. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell, 181*(2), 271-280 e278.
Hu, B., Zeng, L. P., Yang, X. L., Ge, X. Y., Zhang, W., Li, B., . . . Shi, Z. L. (2017). Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog*, 13(11), e1006698. doi:10.1371/journal.ppat.1006698

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., . . . Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. doi:10.1016/S0140-6736(20)30183-5

International Committee on Taxonomy of Viruses Executive, C. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat Microbiol*, 5(5), 668-674. doi:10.1038/s41564-020-0709-x

Issa, E., Merhi, G., Panossian, B., Salloum, T., & Tokajian, S. (2020). SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis. *mSystems*, 5(3), e00266-00220. doi:10.1128/mSystems.00266-20

Jang, K. J., Lee, N. R., Yeo, W. S., Jeong, Y. J., & Kim, D. E. (2008). Isolation of inhibitory RNA aptamers against severe acute respiratory syndrome (SARS) coronavirus NTPase/Helicase. *Biochem Biophys Res Commun*, 366(3), 738-744. doi:10.1016/j.bbrc.2007.12.020

Jia, Z., Yan, L., Ren, Z., Wu, L., Wang, J., Guo, J., . . . Rao, Z. (2019). Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res*, 47(12), 6538-6550. doi:10.1093/nar/gkz409

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., . . . Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647-1649. doi:10.1093/bioinformatics/bts199

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., . . . Wyles, M. D. (2020). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. doi:10.1016/j.cell.2020.06.043

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*, 33(7), 1870-1874. doi:10.1093/molbev/msw054

Lei, J., Kusov, Y., & Hilgenfeld, R. (2018). Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res*, 149, 58-74. doi:10.1016/j.antiviral.2017.11.001

Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu Rev Virol*, 3(1), 237-261. doi:10.1146/annurev-virology-110615-042301

Liu, H., & Li, Y. (2020). Smart cities for emergency management. *Nature*, 578(7796), 515. doi:10.1038/d41586-020-00523-5

Luo, H., Chen, Q., Chen, J., Chen, K., Shen, X., & Jiang, H. (2005). The nucleocapsid protein of SARS coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein A1. *FEBS Letters*, 579(12), 2623-2628.
Megahed, N. A., & Ghoneim, E. M. (2020). Antivirus-built environment: Lessons learned from Covid-19 pandemic. *Sustainable Cities and Society, 61*, 102350.
doi:10.1016/j.scs.2020.102350

Minakshi, R., Padhan, K., Rehman, S., Hassan, M. I., & Ahmad, F. (2014). The SARS Coronavirus 3a protein binds calcium in its cytoplasmic domain. *Virus Res, 191*, 180-183. doi:10.1016/j.virusres.2014.08.001

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., & Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom, 2*(4), e000056. doi:10.1099/mgen.0.000056

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H.-J., . . . Radivojac, P. (2017). MutPred2: inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv*, 134981. doi:10.1101/134981

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem, 25*(13), 1605-1612. doi:10.1002/jcc.20084

Pond, S. L., Frost, S. D., Grossman, Z., Gravenor, M. B., Richman, D. D., & Brown, A. J. (2006). Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput Biol, 2*(6), e62. doi:10.1371/journal.pcbi.0020062

Rahman, M. A., Zaman, N., Asyhari, A. T., Al-Turjman, F., Alam Bhuiyan, M. Z., & Zolkipli, M. F. (2020). Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices. *Sustainable Cities and Society, 62,*
102372. doi:10.1016/j.scs.2020.102372

Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., . . .
Bellini, W. J. (2003). Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science, 300*(5624), 1394-1399. doi:10.1126/science.1085952

Saha, P., Banerjee, A. K., Tripathi, P. P., Srivastava, A. K., & Ray, U. (2020). A virus that has gone viral: amino acid mutation in S protein of Indian isolate of Coronavirus COVID-19 might impact receptor binding, and thus, infectivity. *Biosci Rep, 40*(5). doi:10.1042/BSR20201312

Sanche, S., Lin, Y. T., Xu, C., Romero-Severson, E., Hengartner, N., & Ke, R. (2020). High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerg Infect Dis, 26*(7). doi:10.3201/eid2607.200282

Schoeman, D., & Fielding, B. C. (2019). Coronavirus envelope protein: current knowledge. *Virol J, 16*(1), 69. doi:10.1186/s12985-019-1182-0

Shum, K. T., & Tanner, J. A. (2008). Differential inhibitory activities and stabilisation of DNA aptamers against the SARS coronavirus helicase. *Chembiochem, 9*(18), 3037-3045. doi:10.1002/cbic.200800491

Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., . . . Qin, C. (2019). From SARS to MERS, Thrusting Coronaviruses into the Spotlight. *Viruses, 11*(1). doi:10.3390/v11010059

Subissi, L., Imbert, I., Ferron, F., Collet, A., Coutard, B., Decroly, E., & Canard, B. (2014). SARS-CoV ORF1b-encoded nonstructural proteins 12-16: replicative enzymes as antiviral targets. *Antiviral Res, 101*, 122-130. doi:10.1016/j.antiviral.2013.11.006
Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., . . . Lu, J. (2020). On the origin and continuing evolution of SARS-CoV-2. *National Science Review*. doi:10.1093/nsr/nwaa036

V, A. A. R., R., V., & Haghighat, F. (2020). The contribution of dry indoor built environment on the spread of Coronavirus: Data from various Indian states. *Sustainable Cities and Society*, 102371. doi:10.1016/j.scs.2020.102371

Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., & Zhang, Z. (2020). The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J Med Virol*. doi:10.1002/jmv.25762

Wang, Q., Qiu, Y., Li, J. Y., Zhou, Z. J., Liao, C. H., & Ge, X. Y. (2020). A Unique Protease Cleavage Site Predicted in the Spike Protein of the Novel Pneumonia Coronavirus (2019-nCoV) Potentially Related to Viral Transmissibility. *Virol Sin*. doi:10.1007/s12250-020-00212-7

Wang, Y., Tong, J., Qin, Y., Xie, T., Li, J., Li, J., . . . He, Y. (2020). Characterization of an asymptomatic cohort of SARS-COV-2 infected individuals outside of Wuhan, China. *Clin Infect Dis*. doi:10.1093/cid/ciaa629

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauroiello, G., Gumienny, R., . . . Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.*, 46(W1), W296-W303. doi:10.1093/nar/gky427

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., . . . Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798), 265-269. doi:10.1038/s41586-020-2008-3

Xu, C., Luo, X., Yu, C., & Cao, S.-J. (2020). The 2019-nCoV epidemic control strategies and
future challenges of building healthy smart cities. *Indoor and Built Environment, 29*(5), 639-644. doi:10.1177/1420326x20910408

Zhang, D., Gao, F., Jakovlić, I., Zou, H., Zhang, J., Li, W. X., & Wang, G. T. (2020). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Molecular Ecology Resources, 20*(1), 348-355. doi:10.1111/1755-0998.13096

Zhang, Y., Feng, G., Bi, Y., Cai, Y., Zhang, Z., & Cao, G. (2019). Distribution of droplet aerosols generated by mouth coughing and nose breathing in an air-conditioned room. *Sustainable Cities and Society, 51*, 101721. doi:10.1016/j.scs.2019.101721

Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., . . . Shi, W. (2020). A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr Biol*. doi:10.1016/j.cub.2020.05.023

Zhou, H., He, S., Cai, Y., Wang, M., & Su, S. (2019). Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. *Sustainable Cities and Society, 50*, 101605. doi:10.1016/j.scs.2019.101605

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., . . . Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature, 579*(7798), 270-273. doi:10.1038/s41586-020-2012-7

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., . . . Research, T. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med, 382*(8), 727-733. doi:10.1056/NEJMoa2001017
Ziebuhr, J. (2005). The Coronavirus Replicase. In L. Enjuanes (Ed.), *Coronavirus Replication and Reverse Genetics* (pp. 57-94). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fig 1. Function overview in BioAider.
| Genes | Length | Mutation sites Count/Proportion | S(p) / N(p) / S-N(p) / Termination(p) |
|-------|--------|-------------------------------|--------------------------------------|
| NSP1  | 540    | 48/8.88%                      | 24(4.44%)/24(4.44%)/0/0              |
| NSP2  | 1914   | 168/8.77%                     | 46(2.40%)/122(6.37%)/0/0             |
| NSP3  | 5835   | 384/6.58%                     | 130(2.23%)/253 (4.33%)/1(0.02%)/0    |
| NSP4  | 1500   | 99/6.60%                      | 46(3.07%)/53(3.53%)/0/0              |
| NSP5  | 918    | 42/4.58%                      | 19(2.07%)/23(2.51%)/0/0              |
| NSP6  | 870    | 71/8.16%                      | 30(3.45%)/41(4.71%)/0/0              |
| NSP7  | 249    | 19/7.63%                      | 9(3.61%)/10(4.02%)/0/0               |
| NSP8  | 594    | 27/4.55%                      | 9(1.52%)/16(2.69%)/1(0.17%)/1(0.17%)|
| NSP9  | 339    | 21/6.19%                      | 11(3.24%)/10(2.95%)/0/0              |
| NSP10 | 417    | 27/6.48%                      | 10(2.40%)/14(3.36%)/1(0.24%)/2(0.48%)|
| NSP12 | 2796   | 157/5.61%                     | 58(2.07%)/96(3.43%)/0/3(0.11%)       |
| NSP13 | 1803   | 109/6.05%                     | 42(2.33%)/67(3.72%)/0/0              |
| NSP14 | 1581   | 99/6.25%                      | 41(2.59%)/56(3.54%)/1(0.06%)/1(0.06%)|
| NSP15 | 1038   | 62/5.97%                      | 18(1.73%)/43(4.14%)/1(0.10%)/0       |
| NSP16 | 894    | 67/7.49%                      | 27(3.02%)/39(4.36)/0/1(0.11%)        |
| Count | 21288  | 1400/6.57%                    | 520(2.44%)/867(4.07%)/5(0.02%)/8(0.04) |
| S     | 3819   | 296/7.75%                     | 105(2.75%)/186(4.87%)/3(0.08%)/2(0.05%) |
| ORF3a | 825    | 94/11.39%                     | 31(3.75%)/63(7.64%)/0/0              |
| E     | 225    | 22/9.78%                      | 9(4.00%)/13(5.78%)/0/0               |
| M     | 666    | 44/6.61%                      | 28(4.21%)/16(2.40%)/0/0              |
| ORF6  | 183    | 22/12.01%                     | 4(2.18%)/16(8.74%)/0/2(1.09%)        |
| ORF7a | 363    | 40/11.02%                     | 12(3.31%)/22(6.06%)/1(0.27%)/5(1.38%) |
| ORF7b | 129    | 11/8.52%                      | 3(2.33%)/7(5.42%)/0/1(0.77%)         |
| ORF8  | 363    | 39/10.75%                     | 10(2.75%)/26(7.16%)/0/3(0.83%)       |
| N     | 1257   | 168/13.36%                    | 58(4.61%)/107(8.51%)/3(0.24%)/0      |
| ORF10 | 114    | 16/14.04%                     | 4(3.51%)/12(10.53%)/0/0              |
| Complete | 29232 | 2152/7.36%                  | 784(2.68%)/1335(4.57%)/12(0.04%)/21(0.07%) |
Table 2. Sites with nonsynonymous substitution frequency over 20 or synonymous substitution frequency over 200

| Gene   | Substitutions | AA change | Changed in properties of aa | Frequency/Proportion | Sampling date          |
|--------|---------------|-----------|----------------------------|-----------------------|------------------------|
| T225A  | T225A         | D75E      | No                         | 29/0.89%              | 2020-01-21 to 2020-03-25 |
| C794T* | C794T*        | T265I     | (polar, none-charge) to (non-polar) | 464/14.32%           | 2020-02-21 to 2020-03-31 |
| G1132A | G1132A        | V378I     | No                         | 73/2.25%              | 2020-01-18 to 2020-03-30 |
| G1175A | G1175A        | G392D     | (polar, none-charge) to (polar, negative-charge) | 68/2.10%          | 2020-02-25 to 2020-03-31 |
| A2215G | A2215G        | I739V     | No                         | 77/2.81%              | 2020-02-09 to 2020-03-29 |
| C2293T | C2293T        | P765S     | (non-polar) to (polar, none-charge) | 91/2.81%          | 2020-02-09 to 2020-03-29 |
| G2626A | G2626A        | A876T     | (non-polar) to (polar, none-charge) | 65/2.00%          | 2020-02-25 to 2020-03-31 |
| C2912T | C2912T        | P971L     | No                         | 30/0.93%              | 2020-01-28 to 2020-03-25 |
| C8517T*| C8517T*       | P971L     | No                         | 660/20.37%            | 2020-01-05 to 2020-03-26 |
| T9212A | T9212A        | F3071Y    | (non-polar) to (polar, none-charge) | 49/1.51%           | 2020-02-01 to 2020-03-23 |
| G9832A | G9832A        | G3278S    | No                         | 26/0.80%              | 2020-03-02 to 2020-03-29 |
| A10058G| A10058G       | K3353R    | No                         | 44/1.36%              | 2020-01-30 to 2020-03-21 |
| G10818T*| G10818T*    | L3606F    | No                         | 439/13.55%            | 2020-01-17 to 2020-03-31 |
| C11651T| C11651T       | S3884L    | (polar, none-charge) to (non-polar) | 30/0.93%          | 2020-03-12 to 2020-03-23 |
| C14144T*| C14144T*    | P4715L    | No                         | 1700/52.47%           | 2020-02-01 to 2020-04-01 |
| C14541T*| C14541T*    | P4715L    | No                         | 284/8.77%             | 2020-02-09 to 2020-03-29 |
| C17483T*| C17483T*    | P5828L    | No                         | 390/12.04%            | 2020-02-20 to 2020-03-25 |
| A17594G*| A17594G*    | Y5865C    | No                         | 405/12.50%            | 2020-02-20 to 2020-03-25 |
| C17796T*| C17796T*    | L3606F    | No                         | 410/12.65%            | 2020-01-19 to 2020-03-25 |
| T18472C| T18472C       | F6158L    | No                         | 28                    | 2020-03-03 to 2020-03-25 |
| C18734T| C18734T       | A6254V    | No                         | 27                    | 2020-03-12 to 2020-03-23 |
| G19420T| G19420T       | V6474L    | No                         | 24                    | 2020-03-05 to 2020-03-20 |
| Gene   | Substitutions | AA change | Changed in properties of aa | Frequency/Proportion | Sampling date            |
|--------|---------------|-----------|----------------------------|----------------------|--------------------------|
| S      | A1841G*       | D614G     | (polar, negative-charge) to (polar, none-charge) | **1720/53.09%**     | 2020-01-28 to 2020-04-01 |
|        | G2025C8       | Q675H     | (polar, none-charge) to (polar, positive-charge)   | 7/0.22%              | 2020-03-03 to 2020-03-27  |
|        | A2024G*       | Q675R     | (polar, none-charge) to (polar, positive-charge)   | 1/0.03%              | 2020-03-04               |
|        | AG2827CC      | S943P     | (polar, none-charge) to (non-polar)               | 22/0.68%             | 2020-02-29 to 2020-03-20  |
| ORF3a  | G171T*        | Q57H      | (polar, none-charge) to (polar, positive-charge)   | **578/17.84%**       | 2020-02-21 to 2020-03-31  |
|        | C296T         | A99V      | No                                                   | 23/0.71%             | 2020-02-29 to 2020-03-31  |
|        | G587T         | G196V     | (polar, none-charge) to (non-polar)               | 48/1.48%             | 2020-02-25 to 2020-03-23  |
|        | G752T*        | G251V     | (polar, none-charge) to (non-polar)               | **299/9.23%**        | 2020-01 to 2020-03-29     |
| M      | A8G           | D3G       | (polar, negative-charge) to (polar, none-charge)   | 41/1.27%             | 2020-02 to 2020-03-30     |
|        | C524T         | T175M     | (polar, none-charge) to (non-polar)               | 117/3.61%            | 2020-03-01 to 2020-03-29  |
| ORF7a  | C242T         | S81L      | (polar, none-charge) to (non-polar)               | 25/0.77%             | 2020-03-07 to 2020-03-25  |
| ORF8   | G184C         | V62L      | No                                                   | 41/1.27%             | 2020-01-15 to 2020-03-25  |
|        | G184T         | V62L      | No                                                   | 3/0.09%              | 2020-03-18 to 2020-03-24  |
|        | T251C*        | L84S      | (non-polar) to (polar, none-charge)               | **646/19.94%**       | 2020-01-05 to 2020-03-26  |
| N      | C38T          | P13L      | No                                                   | 29/0.89%             | 2020-02-27 to 2020-03-23  |
|        | G578T         | S193I     | (polar, none-charge) to (non-polar)               | 27/0.83%             | 2020-01-21 to 2020-03-23  |
|        | C581T         | S194L     | (polar, none-charge) to (non-polar)               | 21/0.65%             | 2020-01-16 to 2020-03-21  |
|        | C590T         | S197L     | (polar, none-charge) to (non-polar)               | 49/1.51%             | 2020-02-25 to 2020-03-23  |
|        | G605A         | S202N     | No                                                   | 34/1.05%             | 2020-01-25 to 2020-03-26  |
|        | GGG608-609-610 AAC* | R203K | No                                                   | **499/15.40%**       | 2020-02-25 to 2020-04-01  |

* Represents the substitution hotspots

& These sites were near furin cleavage region of S protein though the substitution frequency was below 20.
Table 3. The number of strains with possible linkage substitution among mainly mutant and referential strain

| Combination of possible linkage substitution hotspots | Referential/Count | Mainly mutant/Count | Others |
|-----------------------------------------------------|-------------------|---------------------|--------|
| ORF1ab-2772 & ORF1ab-14144 & S-1841                  | CCA/1504(46%)     | TTG/1690(52%)       | 45(2%) |
| ORF1ab-8517 & ORF8-251                               | CT/2571(79%)      | TC/641(20%)         | 27(1%) |
| ORF1ab-17483 & ORF1ab-17594 & ORF1ab-17796          | CAC/2824(87%)     | TGT/386(12%)        | 29(1%) |
Table 4. The 23 types of SR-rich regions in N protein of SARS-CoV-2.

Note: The mutation sites are in bold comparing with the reference strain.

| Type               | SSXX-1 | SSXX-2 | SSXX-3 | SSXX-4 | First strain | Sampling date |
|--------------------|--------|--------|--------|--------|--------------|---------------|
|                    | 183    | 184    | 185    | 186    | 187          | 188           |
| Reference type     | S      | S      | R      | S      | R            | N             |
| N-R203K_G204IR     | S      | S      | S      | R      | S            | N             |
| N-S197L            | S      | S      | R      | S      | S            | R             |
| N-S202N            | S      | S      | R      | S      | R            | N             |
| N-S193I            | S      | S      | R      | S      | R            | N             |
| N-S194L            | S      | S      | R      | S      | S            | R             |
| N-S188L            | Y      | S      | S      | R      | S            | R             |
| N-S183Y            | Y      | S      | S      | R      | S            | R             |
| N-R195K            | S      | S      | S      | R      | S            | R             |
| N-R185C            | S      | S      | S      | R      | S            | S             |
| N-S190I            | S      | S      | S      | R      | S            | R             |
| N-S188P            | S      | S      | R      | S      | P            | R             |
| N-R191L & N-R203K_G204IR | S | S | R | S | R | L |
| N-G200S            | S      | S      | S      | R      | S            | R             |
| N-N192K            | S      | S      | R      | S      | R            | K             |
| N-P199S            | S      | S      | S      | R      | S            | R             |
| N-P199T            | S      | S      | S      | R      | S            | R             |
| N-P199L            | S      | S      | S      | R      | S            | R             |
| N-T198I & N-R203K_G204R | S | S | R | S | R | N |
| N-R203K            | S      | S      | S      | R      | S            | R             |
| N-S186F            | S      | S      | R      | S      | F            | R             |
| N-R191C            | S      | S      | S      | R      | C            | N             |
| N-S186Y            | S      | S      | R      | S      | Y            | R             |

Sampling date: 03-18-20-21-22-23