Distinct genetic spectrums and evolution patterns of SARS-CoV-2

Sheng Liu[1,2] *, Jikui Shen[3] *, Lei Yang[4], Chang-Deng Hu[5,6], Jun Wan[1,2,7,8] †

1. Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana, USA
2. Collaborative Core for Cancer Bioinformatics (C^3B) shared by Indiana University Simon Comprehensive Cancer Center and Purdue University Center for Cancer Research, Indiana USA
3. The Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
4. Department of Pediatrics, Indiana University School of Medicine, Indianapolis, Indiana USA
5. Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, Indiana, USA
6. Purdue University Center for Cancer Research, Purdue University, West Lafayette, Indiana USA
7. The Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana, USA
8. Department of BioHealth Informatics, Indiana University School of Informatics and Computing, Indiana University – Purdue University Indianapolis, Indianapolis, Indiana, USA

* These authors have equal contributions.

† Correspondence: Jun Wan, Department of Medical and Molecular Genetics, Indiana University School of Medicine, 410 W. 10th Street. HITS 5013. Indianapolis IN 46202, USA.

Email: junwan@iu.edu

Keywords

SARS-CoV-2, genetic variants, groups of mutations and sub-group, cross-infection, evolution.

Abstract

Four signature groups of single-nucleotide variants (SNVs) were identified using two-way clustering method in about twenty thousand high quality and high coverage SARS-CoV-2 complete genome sequences. Some frequently occurred SNVs predominate but are mutually exclusively presented in patients from different countries and areas. These major SNV signatures exhibited distinguished evolution patterns over time. Although it was rare, our data indicated possible cross-infections with multiple groups of SNVs existed simultaneously in some patients, suggesting infections from different SARS-CoV-2 clades or potential re-combination of SARS-CoV-2 sequences. Interestingly nucleotide substitutions among SARS-CoV-2 genomes tend to occur at the sites where one bat RaTG13 coronavirus sequences differ from Wuhan-Hu-1 genome, indicating the tolerance of mutations on those sites or suggesting that major viral strains might exist between Wuhan-Hu-1 and RaTG13 coronavirus.

Introduction

A novel betacoronavirus SARS-CoV-2 [1] causing human coronavirus disease 2019 (COVID-19) was first reported in Wuhan, Hubei China in December 2019 [2-4]. The pandemic of SARS-CoV2 has infected more than 6 million people over 180 countries and areas around the world with a death totally up-to 370,000 as of May 26, 2020 [5]. The most vulnerable group in this COVID-19 pandemic is elderly and those with different underlying medical conditions such as malnourished, hypertensive, diabetic, cancer and cardiovascular abnormality [6]. Much effort has been devoted by scientists all over the world to understand the features of SARS-CoV2. To date, more than 30,000 SARS-CoV-2 whole genome sequences have been uploaded to the online platform GISAID [7, 8]. Analyzing these data can potentially reveal the viral transmission routes and to identify novel mutations associated with the transmission [9]. For example, researchers employed standard phylogenomics approaches and compared consensus sequences representing the dominant virus lineage within each infected host [10, 11]. Such information will be of important value for the development of vaccine, transmission monitoring and ultimately the control of the pandemic.

Like other virus, SARS-CoV-2 exhibited dynamic transmission patterns during its spreading. The genome of SARS-CoV-2 may create random mutations over time. It has been reported that SARS-CoV-2 mutated at an average speed of about one or two mutations per month [12]. However, only some of mutations were caught and corrected by the virus's error correction machinery [13]. These signature mutations may help understand the origin and evolution of SARS-CoV-2. Given 103 earlier genome sequence data, at least two clades of SARS-CoV-2 were found to be involved in the global transmission based on T -> C mutation on a singleton site at 28144 of the complete genome, which was further termed as S clade (C28144) and L clade (T28144) [14]. Evolutionary analyses suggested S clade appeared to be more related o coronaviruses in animals. Most recently, three major clusters of SNVs involved in the pandemic were found by comparing 160 SARS-CoV-2 genomes [15] with RaTG13 coronavirus [11]. However, most of these studies were based on limited numbers of SARS-CoV-2 genomes, which might lead to debating conclusions [12, 15-19]. With the availability of increased sample size and longer time of SARS-CoV-2 spreading and developing which has covered almost all countries in the world now, it is imperative to provide a comprehensive and updated analysis of the viral genetic variations.

In this study, we took advantage of the larger datasets collected by GISAID which published about 20,000 high quality SARS-CoV-2 genomes with high converge until May 26, 2020 isolated from patients in 87 countries. Our comprehensive analyses clearly revealed distinct patterns of four major group mutations prominent in different countries and areas. We attempted to uncover novel dynamic transmission and evolution paths for specific SARS-CoV-2 variants during the first several months of COVID-19 outbreak. Some cases were found to have multiple groups of

mutations simultaneously, even though the numbers of cases were not high. Comparing with four bat coronavirus genomes, we found that alternations of nucleotides on SARS-CoV-2 genome tend to occur at the same sites where bat coronavirus sequences were different from Wuhan-Hu-1. Particularly, some of nucleotide substitutions on SARS-CoV-2 were apt to be the same as RaTG13 coronavirus sequences. We further investigated the amino acid (AA) changes on Furin and RNA binding domain (RBD) of Spike protein. Our novel genome-wide discoveries shed the light and provide more detailed information of SARS-CoV-2 which has been clouding over the world.

Results

Genetic variants of SARS-CoV-2

We downloaded and analyzed 19,411 SARS-CoV-2 complete genome sequences after excluding low-coverage ones from The Global Initiative on Sharing Avian Influenza Data (GISAID) database (https://www.gisaid.org/) [7, 8] as of May 26, 2020. Using Wuhan-Hu-1 (NCBI Reference Sequence: NC_045512.2, GISAID ID: EPI_ISL_402125) as reference genome, we found total 9,986 nucleotide sites with single nucleotide variations (SNVs) among all samples. Majority of SNVs had very low occurrence frequency as shown in Figure 1A, suggesting a high chance of random mutations. Four nucleotide substitutions were identified in over 70% of genome sequences: A23403G, C3037T, C14408T, and C241T. They distributed at distinct SARS-CoV-2 genome locations, on the gene body of Spike, ORF1a, and ORF1ab, and upstream of ORF1ab, respectively. Additionally, there were other 52 unique SNVs arose from larger than 1% of populations (n > 194). It is interesting that some of these frequent SNVs co-occurred with overlapping ratio larger than 0.9 (see methods) which were connected by blue lines in Figure 1A. They may appear across different proteins. For example, A23403G and C14408T were found simultaneously on 72% samples, while A23403G changes an aspartate to a glycine on Spike and C14408T converts a proline to a leucine on ORF1ab.

Among total 56 frequent SNVs, 29 mutations are nonsynonymous variants (Figure 1B). Some of them have been discussed separately by previous studies [11, 14, 20] or marked as elements in clades G, S, and V from the GISAID report [7, 8]. Here, two-way clustering was performed with distance function of one minus Jaccard index on 56 frequent SNVs and about 20,000 samples (Figure 1B). It is clear to see four major groups of SNVs covering approximately 96% of samples, including A (C14408T/A23403G, occurring on 14,055 samples), B (T28144C on 2088 samples), C (G11083T/G26144T on 2243 samples), and D (G1440A/G2891A on 347 samples). The geographical locations of infected patients bearing these special mutations were very different.

Thirty-seven countries and areas with virus sample numbers larger than 50 were chosen to investigate the geographical distributions of these SNVs. Group A, represented by two nonsynonymous mutations, A23403G and/or C14408T, represents totally 72% of samples in the study, including about 82% of samples from Europe and 67% of North America. The top three countries with the highest ratio in group A (Figure 2A) were Russia (90%), Switzerland (97%), and Denmark (96%). Group B was distinguished by nonsynonymous mutation T28144C which results in substitution of a leucine by a serine on ORF8. It was projected in Thailand (42%), China (35%), Spain (32%) and some other Asian countries/areas (Figure 2B). Group C was featured by one synonymous C14805T and two nonsynonymous SNVs, G11083T and G26144T (Figure 1B), which substituted a leucine with a phenylalanine on ORF1ab and a glycine with a valine on ORF3a, respectively. This group existed in many Asian and European countries and areas, e.g. Hong Kong, Japan, Singapore, England, Iceland, Turkey etc. (Figure 2C), as reported previously [7, 8, 11, 14, 20]. Group D includes two nonsynonymous SNVs, G1440A and G2891A, both of which change the amino acid sequences on ORF1ab. It confirms the clade D, previously defined by Guan et al. [20] based on smaller set of patients. G1440A led to the amino acid change,

G212D on nonstructural protein 2 (nsp2), while G2891A caused A58T on nonstructural protein 3 (nsp3). D-group was mainly found in several European countries, e.g. Wales (17%), Germany (12%), and Belgium (5%) (Figure 2D).

Besides featured SNVs in each major group, some SNVs co-occurred with the same signature SNVs but covered smaller populations. Importantly, many non-major SNVs were mutually exclusively presented with each other in different countries and areas (Figure 1B). For instance, about 20 mutations co-occurred with A23403G and C14408T in the group A but composed sub-types of A (Figure 1B). Taken as an example, both G25563T and C1059T were found in group A. However, two separable sets of cases were associated with the combinations of G25563T and C1059T (Figure 2A), sub-cluster A1 including both G25563T and C1059T, and sub-cluster A2 including G25563T but excluding C1059T. They represented different strains that were found in distinct populations from varied countries. Specifically, A1 occurred on 23% of all worldwidely collected SAR-CoV-2 genomes, particularly in 71% of Denmark, 54% of Israel, and 52% of USA, whereas A2 was found in only 5% population level, which were mostly discovered in Saudi Arabia (71%), Turkey (38%), and Columbia (33%). Another sub-cluster, A3, had consecutive mutations at positions 28881-28883 on SARS-CoV-2 complete genome, leading a nonframeshift substitution on ORF9. A3 occupied in 23% of worldwide cases, represented by three major countries and areas, Russia (67%), Norther Ireland (62%), and Greece (62%). Some sub-clusters of mutations were found in around 5% or lesser worldwide cases, but they were significantly over-represented in several major countries and areas. For instance, A4 with synonymous mutation C15324T was detected in an African country, Democratic Republic of the Congo (D.R.C.). Approximately 66% of D.R.C. samples had mutation C15324T, coming together with 35% of Switzerland and 29% of Luxembourg.

Patients from one country might have different main groups or sub-types of mutations. A synonymous A20268G in cluster A5 (Figure 2A) was sampled in Spain (45%), Scotland (22%), and Iceland (20%). It is interesting that other 32% of Spain had another distinguished nonsynonymous mutation T28144C in group B (Figure 2B). Most of Spain samples (20% of total) in group B also had a unique mutation, C28863T, substituting a serine with a leucine on ORF9, co-occurring with T28144C. A large population of Australia samples were found in group B as well. But they came with additional diverse mutually exclusive mutations, e.g. either C18060T/A17858G/C17747T (8%), or C28863T (7%), or G28077C (4%). Similar scenarios were observed in USA, where approximately 21% of samples encompassed T28144C with C18060T/A17858G/C17747T, while another 3% was recognized with a different nonsynonymous mutation G28077C in the same main group B.

Group C including G11083T and G26144T existed in many Asian and European countries and areas (Figure 2C), such as Hong Kong, Japan, Singapore, England, Iceland, and Turkey, as reported previously [11]. However, different countries and areas were distinguished by extra mutations in the same main group C. For example, over 80% of Hong Kong samples were detected with nonsynonymous G26144T on ORF3a, while Singapore had over 80% samples with nonsynonymous mutation C28311T on ORF9.

Interestingly, a synonymous mutation C14805T existed in both group B and C (Figure 1B), covering over 8% of worldwide samples. Although majority (85%) of C14805T was a signature mutation in group C (G11083T/G26144T) in over 10% patients of Europe and North America, C14805T also co-occurred with T9477A/G25979T/C28863T in group B, which were mostly found in Spain (20%) and Australia (7%) (Figure 2B).

Currently, it lacks sufficient evidences to make a conclusive statement about the origins of all SARS-CoV-2 mutations. But time-annotated data collections can still explore geographical evolution patterning of specific SNVs, albeit the limited number of high quality and high coverage sequenced viral genomes. For example, only three cases with mutations T28144C and C18060T

(one sub-type in group B) reported in Washington State of USA in January 2020, in addition to seven cases in China and additional one in Singapore at almost same time (Figure 3A). It is notable that T28144C and C18060T co-occurred with additional de novo nonsynonymous mutations C17747T/A17858G on ORF1ab on 52 cases in February 2020, 51 from Washington State of USA and one from Grand Princess Cruise. No such case was detected in other countries and areas. One month later, this group of signature mutations spread over many states of USA, particularly west coast of USA, and other countries and areas of different continents, including Australia, Canada, Iceland, Mexico, England, and Taiwan etc.

Over half of American patients had been sampled with mutations C14408T/A23403G and C1059T on SARS-CoV-2 genome (Figure 2A). In January 2020, only one case with both C14408T and A23403G was found from China in our dataset (Figure 3B). The first case in USA was reported in New Hampshire at east coast with co-occurrence of C1059T, in addition to five in France, one in Belgium and one in Senegal. The numbers of such cases boosted up in USA and other countries in March 2020, including 354 in Denmark, 163 in Australia, and tens in France, England, etc. In the USA, approximately 800 cases were found on east coast of USA, while over 400 cases were identified on west coast as well.

The mutations of group A (C14408T/A23403G) indicated at least two strains of SARS-CoV-2 distinguishable on the sites of Spike and ORF1ab. One observed from Wuhan-Hu-1was named as DP with an aspartate on 614 of Spike and a proline on 4715 of ORF1ab, while another potential one, GL, had a glycine on the site of 614 on Spike and a leucine on the site of 4715 on ORF1ab instead. The ratio of GL strain in all USA cases increased dramatically from 6% in February to 96% in May 2020 (Figure 3C). The similar growing trend was observed in most of other countries, regardless when this group of mutations were first present (Figure 3C). In general, 96% of samples from all these countries had strain GL in May 2020 compared to only 3% in February (Figure 3D), suggesting that the GL strain of SARS-CoV-2 might become much more stable and prevailing than the other strain DP from Wuhan-Hu-1 after 6-month evolution and transmission.

Different groups of mutations also exhibited distinguished evolution patterns in countries selected in the study (Figures 3C and D). Taking B-group SNVs for example, we found that the ratio decreased over time from 33% in January 2020 to 0% in May in these countries, indicating that at least two strains existed at the early of COVID-19 pandemic. However, strains including variant at 28144 other than Wuhan-Hu-1 disappeared after 6-months of transmission. Only the strain that has the same nucleotide T28144 as Wuhan-Hu-1 finally became the most stabilized strain found in the host. The similar patterns were observed for groups C and D as well, even though a sudden increasing was found in March 2020 due to unknown reasons. For instance, Germany had a high ratio, 47.8% (11 out of 23), of samples with SNVs G1440A and G2891A in February, while 25% of (96 out of 384) Wales were sampled the same mutations in March 2020.

Cross-infections of patients

Four main groups of mutations mutually exclusively occurred in most of 20,000 patients in our study, indicating a unique viral strain in the host. However, as reported in early of March 2020, a patient hospitalized in Iceland infected by two SARS-CoV-2 subtypes simultaneously [21]. One strain of the SARS-nCoV-2 coronavirus was more aggressive, according to Reykjavik Grapevine newspaper citing CEO of CODE Genetics biopharmaceutical company Kari Stefansson. The second strain is a mutation from the original version of the coronavirus that appeared in Wuhan, China. This was regarded as the first known case of double infection. Gudbjartsson et. al. [22] reported that patient T25 carries both the A2a1a strain and the A2a1a+25958 strain. As shown in Figure 4, 11 genomes bore both A-group and B-group mutations, while 194 genomes had group variants of A and C, and both B-group and C-group were involved in 27 genomes. Strikingly, one

patient from Spain was detected with three groups of mutations simultaneously, A, B and C. The SARS-CoV-2 genomes with novel D-group mutations were not overlapped with either A or B. But 14 out of 333 genomes embraced group C as well.

Comparison of variants between SARS-CoV-2 and bat coronavirus sequences

Bats have been thought to be reservoir species for many infectious diseases, including SARS-CoV-2. In order to understand the possible association between SNVs among SARS-CoV-2 genome sequences from patients and bat coronavirus sequences, we aligned four bat coronavirus sequences to Wuhan-Hu-1 complete genome. The ratios of variants between Wuhan-Hu-1 and bats were 3.9% (RatG13), 11.7% (bat-SL-CoVZC45), 12.0% (bat-SL-CoVZXC21), and 5.8% (RmYN02). As previously described, 9,986 out of 29903 nts (33.4%) on SARS-CoV-2 genome underwent variations among about 20,000 samples. Interestingly, all four bats' coronavirus in our study showed significantly (the least p = 5.8e-46) elevated ratios of SARS-CoV-2 SNVs on the sites where bat's sequences differed from SARS-CoV-2 compared to that on SARS-CoV-2 complete genome (Figure 5A), suggesting that the sites where SARS-CoV-2 differ from bats might have higher tolerance for sequence mutations. Among them, RatG13 coronavirus covered the higher ratio (53.3%) than bat-SL-CoVZC45 (44.9%), bat-SL-CoVZXC21 (44.9%), and 5.8% RmYN02 (49.7%).

In theory, 9,986 nucleotide variants among 20,000 SARS-CoV-2 genomes can potentially turn to be any one of three nucleotides other than original ones from Wuhan-Hu-1. But it turned out that SARS-CoV-2 SNVs had the same mutated nucleotides as RaTG13 coronavirus does on 430 out of 1156 (37.2%) sites where RaTG13 coronavirus sequence differed from Wuhan-Hu-1 (Figure 5B), including C29095T [11] and seven high frequent SNVs identified from major groups, e.g. C2416T and C3037T from group A, C8782T, C18060T, C24034T, and T28144C from group B, and C23929T in group C. The ratio for RaTG13 coronavirus was much higher than the ratios (around 24%) observed in other three bat coronavirus sequences.

SNVs on Spike protein

Similar to SARS coronavirus, SARS-CoV-2 entered human cells through its high-affinity receptor-binding domain (RBD) and its proteolytic proteases to human protein ACE2 which enables an efficient cell entry [23]. Group A had a significant mutation at A23403G leading to AA change, D614G, on Spike protein. This mutation was discussed as clade G [7, 8] covering over 70% of total sequenced genomes (96% in May) in our study. In addition, Cryo-EM-based structural analysis [24] revealed that 5 key amino acids within 434-507 of Spike protein contributed most to the binding activity. This was also confirmed by several recent cryo-EM structural studies [25-28]. The key AAs of SARS-CoV-2 RBD are: L455, G482, V483, E484, G485, F486, Q493, S494 and N501. Interestingly, we identified several nonsynonymous SNVs of SARS-CoV-2 on L455, V483, G485 and S494 from sequenced sample, for instance, G22927T (L455F), G23009T (V483F), T23010C (V483A), G23105A (G485S), and T23042C (S494P). Among them, 28 viral genomes had mutation T23010C (V483A), all of which were sampled in USA, including 26 from Washington State.

Furin is responsible for the proteolytic cleavage. In SARS-CoV-2, 15-nt CCTCGGCGGGCACGT encodes five AAs: PRRAR (681-685), locating at 23603-23617 of Wuhan-Hu-1 complete genome. Furin cleavage site bears a RXXR pattern [29, 30]. R685 makes an ideal Furin proteolytic cleavage site [31]. Twenty-eight SARS-CoV-2 genomes were detected with mutations in the region, including 10 from England, 7 from USA and 3 from Switzerland. The earliest case was a patient from Hangzhou, China. Nonsynonymous SNVs, C23604T, was most frequent among all others, causing

the mutation of P681L. Other AA mutations included P681H/P, R682Q/W, R683P/Q, and A684V/T/E.

Discussion

In this study, we comprehensively analyzed 19,411 SARS-CoV-2 whole-genome sequences in GISAID viral portal until May 26 2020, as well as four bat genome sequences. We explored the mutation patterns that distinguish patients from different countries and areas. Some of frequent SNVs reported previously were identified and discussed individually. We used bioinformatics approaches to systematically explore four mutually exclusive major groups of SNVs on SARS-CoV-2 genome. These mutations were detected in populations from different geographical locations. One group consisting of two nonsynonymous mutations, G1440A and G2891A, was discovered in several European countries, e.g. Wales and Germany. Both nonsynonymous mutations change the amino acid sequences on ORF1ab. G1440A led to the AA change of G392D on ORF1ab, G212D on nonstructural protein 2 (nsp2), while G2891A caused A876T on ORF1ab, A58T on nonstructural protein 3 (nsp3). Interestingly, the structures of nsp2 and nsp3 predicted by I-Tasser web site [32] showed that both nsp2 G212 and nsp3 A58 appear to be exposed to the solvent. Nsp2 G212D falls on the region which was homologous to the endosome-associated protein similar to the avian infectious bronchitis virus (PDB 3ld1) which plays a key role in the viral pathogenicity [33]. Nsp3 A58T locates at N-terminal ubiquitin-like domain that plays an important role in viral replication. As visualized by Pymol [34], changes of G212D and A58T add clashes between residues 212 and ASN183, between residues 59 and ILE62, respectively (Supplemental Figure 1). Predictions by PROVEAN [35] suggest that these changes are neutral, indicating they are stabilizing variants given the fact that over 300 patients have been detected with these two variants. These results could provide some insights of possible new functions of some important proteins and their therapeutic potential comparing with other coronavirus.

Distinct time-course evolution patterns were observed for four major groups of mutations. The viral strains different from Wuhan-Hu-1 may gradually replace the earlier one detected from Wuhan-Hu-1, e.g. GL with mutations C14408T and A23403G. Or the strains same as Wuhan-Hu-1 at SNV sites may become dominant after several month evolutions, e.g. groups B-D (Figures 3C and D). It is hard to make a solid conclusion about this kind of aberrant emergence of new strains based on current evidence, particularly due to the lack of enough numbers of high-quality sequenced samples world widely, including China, before February 2020. However, with more and more clinical data generated, evolution patterning associated with specific biological functions may be clearly uncovered. For example, several groups recently reported that A23403G mutation in Spike protein might alter the antigenic property and transmission ability due to the change of protomer interaction [36, 37].

In general, four main groups of mutations were mutually exclusively presented. The bioinformatics and computational analyses exhibited a few hundred patients who were identified to carry multiple major groups of SNVs at the same time. Without clear clues that homologous recombination could occur in these viruses, we just defined such overlaps as cross-infections based on our observations and current knowledge. There are several scenarios about these cross-infections. One possibility is that two or three strains co-existed and prevailed in the population of the same region during the periods. Alternatively, the patients could be infected with one strain first then another one later, suggesting that primary infection did not yield immunity against the subsequent infection from a different strain. Of course, the percentage of cross-infection cases was very low when compared with approximately 20,000 samples in this study. It might be the consequences of the quarantine and lockdown policy enforced after the spread of COVID-19, while social distancing and wearing face mask are considered effective approaches in reducing the chance of cross infection [38-41] which reduced the likelihood that people met patients with different SARS-CoV-2 strains at the same time.

We further compared SNVs among SARS-CoV-2 genomes on human patients to bat coronavirus nucleotides different from Wuhan-Hu-1. It is interesting that SARS-CoV-2 SNVs, particularly those high-frequent mutations, tend to occur at the same sites where bats sequences varied from Wuhan-Hu-1, suggesting the high tolerance of these sites for genetic mutations, or potentials of SARS-CoV-2 turning to a wild-type pathogenic phenotype. RaTG13 coronavirus was most similar to SARS-CoV-2 from perspective of sequences, but it held the highest ratio of SARS-CoV-2 variants which became the same nucleotides as bats' coronavirus sequences at the same sites. This suggests that some strains of SARS-CoV-2 deviated from Wuhan-Hu-1 might be more similar to coronavirus in RaTG13 than in other bats presented in this paper, even though we don't have more evidence to show the exact connections between them. Our results may shed the light to search intermediate host and further understand the mechanisms of interspecies transmission in future.

In summary, we attempted to uncover fundamental views of SARS-CoV-2 mutations which may help us understand functional consequences due to the viral genetic instability. Our efforts in exploring the patterns of SARS-CoV-2 transmission and evolution in different geographical locations can be helpful to fight against the pandemic. Our findings may provide useful insights on SARS-CoV-2 replication, pathogenicity, and even possible implications for drug discovery, antibody design or vaccine development after being incorporated with other new published studies, e.g. interaction maps between SARS-CoV-2 proteins and human proteins [42].

Acknowledgments

We are grateful to scientists and researchers for depositing whole genomic sequences of Novel Pneumonia Coronavirus (SARS-CoV-2/hCoV-19/2019-nCoV) at the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu™. Thanks to GISAID database for allowing us to access the sequences for non-commercial scientific research. This research was partially funded by grants from the National Institutes of Health (P30CA082709) and Walther Cancer Foundation.

References

1. Viruses, C.S.G.o.t.I.C.o.T.o., *The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2.* Nat Microbiol, 2020. **5**(4): p. 536-544.
2. Wu, F., et al., *A new coronavirus associated with human respiratory disease in China.* Nature, 2020. **579**(7798): p. 265-269.
3. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020. **579**(7798): p. 270-273.
4. Zhu, N., et al., *A Novel Coronavirus from Patients with Pneumonia in China, 2019.* N Engl J Med, 2020. **382**(8): p. 727-733.
5. Dong, E., H. Du, and L. Gardner, *An interactive web-based dashboard to track COVID-19 in real time.* Lancet Infect Dis, 2020. **20**(5): p. 533-534.
6. Guzik, T.J., et al., *COVID-19 and the cardiovascular system: implications for risk assessment, diagnosis, and treatment options.* Cardiovasc Res, 2020.
7. Elbe, S. and G. Buckland-Merrett, *Data, disease and diplomacy: GISAID's innovative contribution to global health.* Glob Chall, 2017. **1**(1): p. 33-46.
8. Shu, Y. and J. McCauley, *GISAID: Global initiative on sharing all influenza data - from vision to reality.* Euro Surveill, 2017. **22**(13).

9. Zhang, L., et al., *Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission.* Medrxiv, 2020.
10. Lai, A., et al., *Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2.* J Med Virol, 2020.
11. Forster, P., et al., *Phylogenetic network analysis of SARS-CoV-2 genomes.* Proc Natl Acad Sci U S A, 2020. **117**(17): p. 9241-9243.
12. Kupferschmidt, K., *https://www.sciencemag.org/news/2020/03/mutations-can-reveal-how-coronavirus-moves-they-re-easy-overinterpret*, in *Science News*. 2020.
13. Kupferschmidt, K. and J. Cohen, *Race to find COVID-19 treatments accelerates.* Science, 2020. **367**(6485): p. 1412-1413.
14. Tang, X., et al., *On the origin and continuing evolution of SARS-CoV-2.* National Science Review, 2020: p. nwaa036.
15. Sanchez-Pacheco, S.J., et al., *Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary.* Proc Natl Acad Sci U S A, 2020. **117**(23): p. 12518012519.
16. Forster, P., et al., *Reply to Sanchez-Pacheco et al., Chookajorn, and Mavian et al.: Explaining phylogenetic network analysis of SARS-CoV-2 genomes.* Proc Natl Acad Sci U S A, 2020. **117**(23): p. 12524-12525.
17. Chookajorn, T., *Evolving COVID-19 conundrum and its impact.* Proc Natl Acad Sci U S A, 2020. **117**(23): p. 12520-12521.
18. Rambaut, A., et al., *A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology*. 2020, bioRxiv.
19. Mavian, C., et al., *Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. .* Proc Natl Acad Sci U S A, 2020. **117**(23): p. 12522-12523.
20. Guan, Q., et al., *The genomic variation landscape of globally-circulating clades of SARS-CoV-2 defines a genetic barcoding scheme.* biorxiv, 2020.
21. *https://www.mbs.news/a/2020/03/icelandic-man-reportedly-caught-two-coronavirus-subtypes-simultaneously.html*, in *MBS news*. 2020.
22. Gudbjartsson, D.F., et al., *Spread of SARS-CoV-2 in the Icelandic Population.* N Engl J Med, 2020.
23. Hoffmann, M., et al., *SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor.* Cell, 2020. **181**(2): p. 271-280.e8.
24. Wrapp, D., et al., *Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.* Science, 2020. **367**(6483): p. 1260-1263.
25. Shang, J., et al., *Cell entry mechanisms of SARS-CoV-2.* Proc Natl Acad Sci U S A, 2020. **117**(21): p. 11727-11734.
26. Wang, Q., et al., *Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2.* Cell, 2020. **181**(4): p. 894-904.e9.
27. Shang, J., et al., *Structural basis of receptor recognition by SARS-CoV-2.* Nature, 2020. **581**(7807): p. 221-224.
28. Yuan, M., et al., *A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV.* Science, 2020. **368**(6491): p. 630-633.
29. Molloy, S.S., et al., *Human furin is a calcium-dependent serine endoprotease that recognizes the sequence Arg-X-X-Arg and efficiently cleaves anthrax toxin protective antigen.* J Biol Chem, 1992. **267**(23): p. 16396-16402.

30. Shiryaev, S.A., et al., *High-resolution analysis and functional mapping of cleavage sites and substrate proteins of furin in the human proteomes* PLoS One, 2013. **8**(1): p. e54290.
31. Coutard, B., et al., *The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade.* Antiviral Res, 2020. **176**: p. 104742.
32. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nat Methods, 2015. **12**(1): p. 7-8.
33. Angeletti, S., et al., *COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis.* Med Virol, 2020.
34. Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 2.4*. 2015.
35. Choi, Y. and A.P. Chan, *PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels.* Bioinformatics, 2015. **31**(16): p. 2745-2747.
36. Becerra-Flores, M. and T. Cardozo, *SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate.* Int J Clin Pract, 2020: p. e13525.
37. Korber, B., et al., *Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2.* Biorxiv, 2020.
38. Wang, Y., et al., *Reduction of secondary transmission of SARS-CoV-2 in households by face mask use, disinfection and social distancing: a cohort study in Beijing, China.* BMJ Glob Health, 2020. **5**(5).
39. West, R., et al., *Applying principles of behaviour change to reduce SARS-CoV-2 transmission.* Nat Hum Behav, 2020. **4**(5): p. 451-459.
40. Eikenberry, S.E., et al., *To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic.* Infect Dis Model, 2020. **5**: p. 293-308.
41. Cheng, V.C., et al., *The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2.* J Infect, 2020.
42. Gordon, D.E., et al., *A SARS-CoV-2 protein interaction map reveals targets for drug repurposing.* Nature, 2020.

Figures

Figure 1. SNVs on 19,411 SARS-CoV-2 complete genomes. (A) Circos plot shows distribution, frequency and co-occurrences of SNV2. From outer to inner circle: coronavirus genome location (nt), gene annotation, occurrence ratios of SNVs at the site (log_{10} scale, red bars), and connections with high co-occurrence rates (> 0.9) represented by blue lines. The darker the blue, the higher co-occurrence rates. (B) Fifty-six high-frequent SNVs were detected (in purple) in over nineteen thousand patients worldwide. Four major clusters of SNVs can be formed to represent patients from different geographical locations.

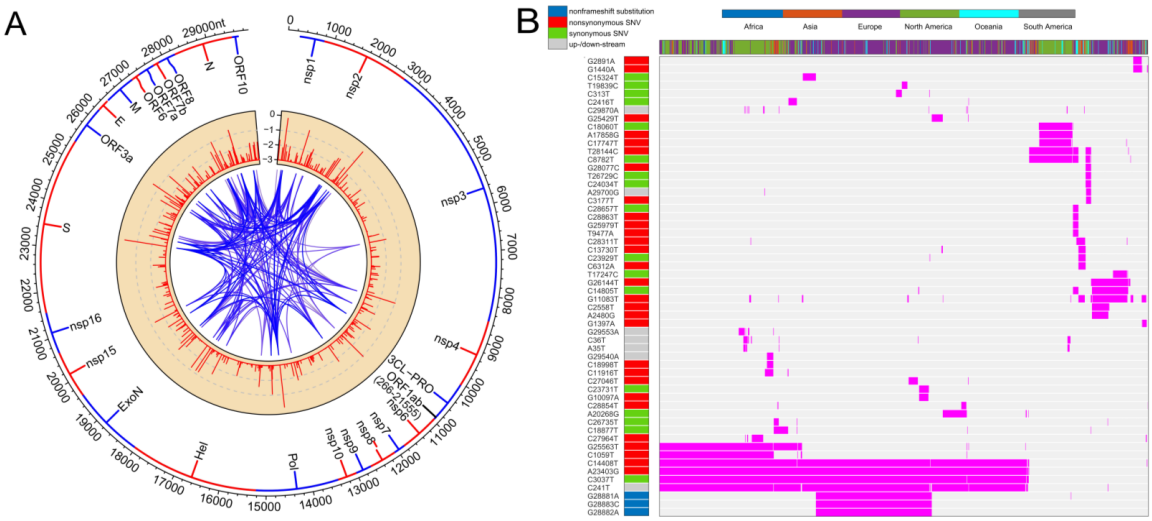

Figure 2. Frequencies of signature SNVs in top three countries and areas. Four groups and their sub-groups had distinct representing countries and areas.

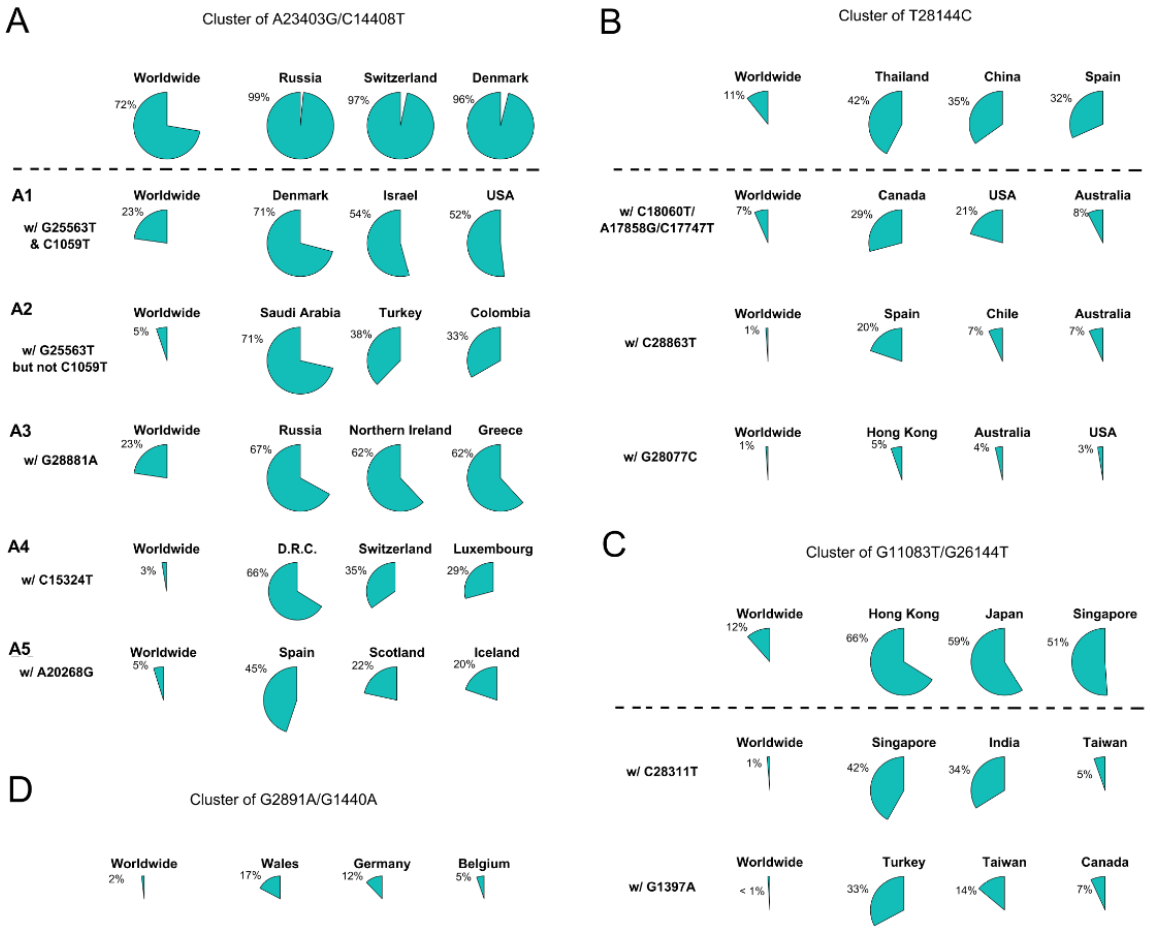

Figure 3. SNVs migration and evolution patterns over time. (A) SNVs of T28144C with C18060T and additional C17747T/A17858G spread in USA and other countries and areas from Jan to March of 2020. (B) SNVs of C14408T and A23403G with C1059T spread in USA and other countries and areas from January to March of 2020. (C) The ratios of four main groups of SNVs, A-D in Figures 1 and 2, varied in different countries/areas with time development. (D) Average ratios of groups A-D SNVs show distinct patterns from January to May of 2020.

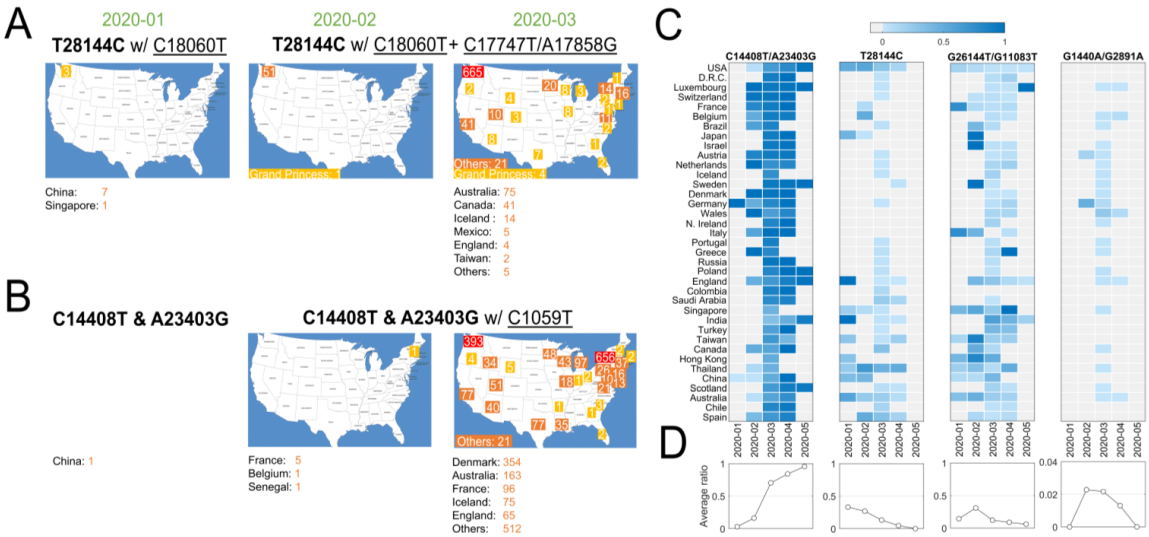

Figure 4. Number of samples carrying major groups of SNVs. Over 200 viral samples encompassed multiple groups of mutations simultaneously.

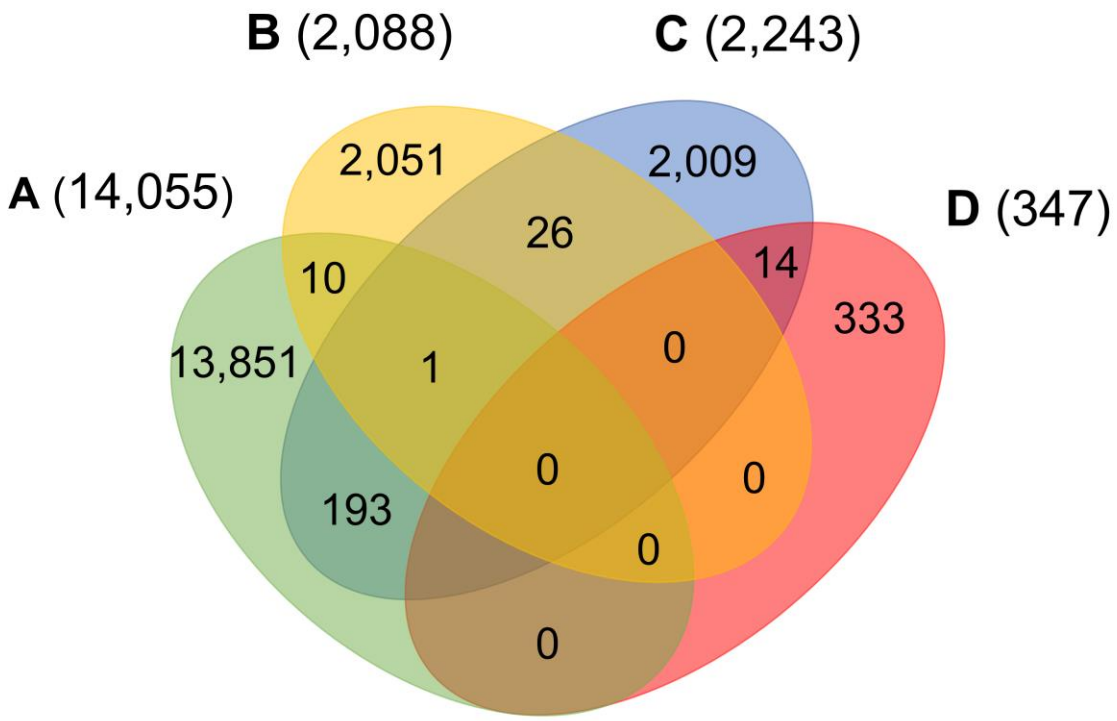
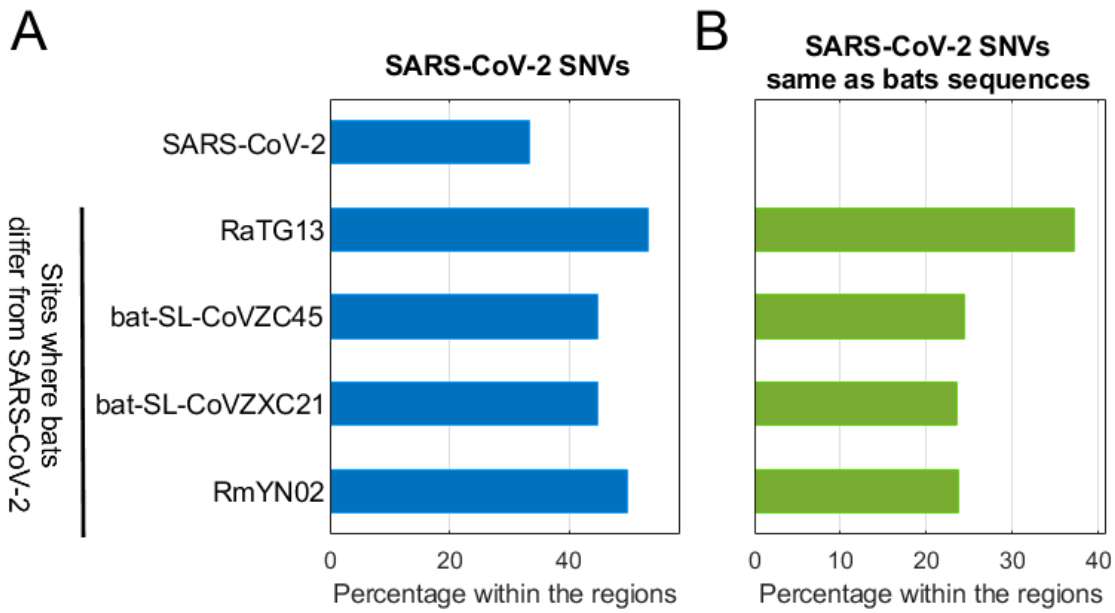
Figure 5. Comparison between nucleotides of bats (RaTG13, bat-SL-CoVZC45, bat-SL-CoVZXC21, and RmYN02) different from SARS-CoV-2 and SNVs among 19,411 SARS-CoV-2 complete genomes. (A) Percentage of SARS-CoV-2 SNVs on different regions, including SARS-CoV-2 complete genome, and sites where bats differ from SARS-CoV-2; (B) Percentage of SARS-CoV-2 SNVs which were exactly same as bats nucleotides within the same regions as shown in (A).