Intra-host dynamic variations in SARS-CoV-2

Jiarui Li
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
Pengcheng Du
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
Lijiang Yang
Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University
Ju Zhang
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
Chuan Song
Beijing Ditan Hospital, Capital Medical University
Danying Chen
Institute of Infectious disease, Beijing Ditan Hospital, Capital Medical University
Yangzi Song
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
Nan Ding
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
Mingxi Hua
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
Kai Han
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
Rui Song
Beijing Ditan Hospital
Wen Xie
Beijing Ditan Hospital, Capital Medical University
Zhihai Chen
Beijing Ditan Hospital, Capital Medical University https://orcid.org/0000-0001-6481-4781
Xianbo Wang
Beijing Ditan Hospital, Capital Medical University
Jiangyuan Liu
Beijing Ditan Hospital, Capital Medical University
Yanli Xu
Beijing Ditan Hospital, Capital Medical University
Guiju Gao

Clinical and Research Center of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University

Qi Wang

Beijing Ditan Hospital, Capital Medical University

Lin Pu

Beijing Ditan Hospital, Capital Medical University

Lin Di

Beijing Advanced Innovation Center for Genomics, Biomedical Pioneering Innovation Center, Peking University

Jie Li

Tsinghua University

Jinglin Yue

Peking University Ditan Teaching Hospital, Beijing

Junyan Han

Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, Beijing, P. R. China

Xuesen Zhao

Institute of Infectious disease, Beijing Ditan Hospital, Capital Medical University

Yonghong Yan

Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University

Fengting Yu

Beijing Ditan Hospital, Capital Medical University

Angela R Wu

Division of Life Science and Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology

Fujie Zhang

Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
https://orcid.org/0000-0001-6386-9879

Yi Gao

Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University

Yanyi Huang

Peking University https://orcid.org/0000-0002-7297-1266

Jianbin Wang

Tsinghua University https://orcid.org/0000-0001-6725-7925

Hui Zeng

Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, Beijing, P. R. China
https://orcid.org/0000-0002-7456-6061

Chen Chen (✉ chenchen1@ccmu.edu.cn)

Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University
https://orcid.org/0000-0003-1765-8899

Article

Keywords: SARS-Cov-2, intra-host variation, spatio-temporal analysis, positive selection, fitness selection

Posted Date: March 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-144416/v1

Abstract

The mutations make uncertain to SARS-CoV-2 disease control and vaccine development. At population-level, single nucleotide polymorphism (SNPs) have displayed mutations for illustrating epidemiology, transmission, and pathogenesis of COVID-19. These mutations are to be expected by the analysis of intra-host level, which presented as intra-host variations (iSNVs). Here, we performed spatio-temporal analysis on iSNVs in 402 clinical samples from 170 patients, and observed an increase of genetic diversity along the day post symptom onset within individual patient and among subpopulations divided by gender, age, illness severity and viral shedding time, suggested a positive selection at intra-host level. The comparison of iSNVs and SNPs displayed that most of nonsynonymous mutations were not fixed suggested a purifying selection. This two-step fitness selection enforced iSNVs containing more nonsynonymous mutations, that highlight the potential characters of SARS-CoV-2 for viral infections and global transmissions.

Main

Despite a global emergence of various innovative prevention and control responses, SARS-CoV-2 continues to spread rapidly around the world [1–4]. As a positive-sense single-stranded RNA virus, a type of virus featured with high mutation rates, any unexpected mutations may change the nature history of the virus and make it even harder to control and led to reduced vaccine efficacy [5]. Till now, researchers identified large amount of mutations based on over 313 thousand published SARS-CoV-2 genomes [6,7]. Among them, two mutants, D614G [8] and a distinct lineage B.1.1.7 [9], increased rapidly in Europe, and then taken hold in worldwide including US, Canada and Australia. Therefore, the genomic changes of spreading SARS-CoV-2 should be carried out for globally monitored [10].

Viral mutations are randomly initiated, then formed as intra-host variation (iSNV), and subsequently fixed as single nucleotide polymorphisms (SNPs) that transmitted among the populations (Fig. 1). This evolutional process was driven two determinant factors: stochastic process (e.g., genetic drift or features of recent epidemiology) and the deterministic process (e.g., fitness selection) [11]. As these forces worked in tandem on SARS-CoV-2, it is often hard to differentiate when a virus mutation becomes common through fitness or by chance. Previous studies that together analyzed clinical, molecular and immunological data with SNPs in population level, have provided a clear map on epidemiology and transmission, and illustrated the potential pathogenesis [12–14].

As a large genetic mutation pool to SNPs [15], iSNVs offers large sufficient information to define the diversity and dynamics of viral evolution within individual hosts (Fig. 1). The analysis of iSNVs complementing conventional population-level SNP studies, generate a more comprehensive understanding of viral evolution, in turn, and aid clinically-relevant predictions of viral evolution that associated to infection, pandemic and immunity-evade requires studies [11]. However, few studies illustrated the genetic characters of iSNVs in COVID-19 patients.

In the present study, we quantitatively assess the SARS-CoV-2 genetic diversity and viral evolution within individual hosts, advancements in a deep viral genome sequencing [16] and an updating empirical analysis pipelines [17]. Spatio-temporal analysis of the genomic data revealed increased viral genetic diversity of SARS-CoV-2, and demonstrated a two-step fitness selection played a predominant role in the understanding of evolutionary of SARS-CoV-2.

Results

We collected 537 (183 pharyngeal, 241 sputum and 113 fecal) samples from 204 patients, which covered 34.4% of total cases in Beijing before April 30 (*Figure 2A, Table S1*). Using deep viral genome sequencing, we obtained 8.59G (IQR: 3.12G-17.38G) base pair per sample on average, of which 11.90% (IQR: 1.32%-53.92%) were mapped to the SARS-CoV-2 reference genome Wuhan-Hu-1 (accession: NC_045512.2)[18]. Samples with low viral genomic coverage were removed (see Methods) and eventually 402 (136 pharyngeal, 182 sputum and 84 fecal) samples from 170 patients underwent further analysis (*Figure 2B, Figure S1A*).

In each sample, we examined high-depth (≥100x) SARS-CoV-2 genomic sites for iSNVs, and filtered the iSNVs with stringent threshold(≥5%), which could sufficiently distinguished the true iSNVs from the sequencing errors (see Methods). These iSNVs with stringent threshold were widely distributed along the genome, and the iSNVs number in each sample was not affected by genomic coverage ratio or sequencing depth (R^2, 0.074 and 0.006, respectively, *Figure S1B, Figure S1C*). Totally, we identified 7,037 iSNVs in the 374 samples with a median density of 0.53 iSNVs/kb, which is comparable to the iSNVs previously observed in Ebola virus [17], Yellow fever virus [19] and Influenza A [20] (*Table S2*). Notably, sequencing data showed that 28 samples did not contain iSNVs compared to the reference genome (*Figure S1D*).

Uneven distribution of intra-host variations in SARS-CoV-2 genomes

We examined the locations of iSNVs along the SARS-CoV-2 genome. Compared to the lower density of these iSNVs (0.58 iSNVs/kb), we observed higher iSNVs density in 5'-UTR (1.23 iSNVs/kb) and 3'-UTR (1.07 iSNVs/kb) (*Figure 1E*). 6,790 (96.49%) iSNVs were occurred in the coding regions. Most iSNVs (4,625, 68.11%) were identified in ORF1ab, following by S gene (903 iSNVs) and N gene (459 iSNVs) (*Figure S1F*). However, after we normalized iSNVs with gene length, the highest frequency of iSNVs was obtained in ORF8 (1.02 iSNVs/kb), following by N protein (0.906 iSNVs/kb) (*Figure 2D*), which were consistent with previous identification at SNP level [21]. The analysis on allele position in codon position showed that 2,329, 2,178 and 2,283 iSNVs were occurred at the 1st, 2nd and 3rd codon position, respectively (*Figure 2E*). The Fisher-exact test on the iSNVs in codon position for each ORF displayed that ORF10 and E gene had more iSNVs at the 1st codon positions (*Figure 2E*).

We then examined the distribution of iSNVs among the patients, and found the iSNVs occurred in high frequency SNPs (hfSNPs) in the global SARS-CoV-2 genome database shared in more patients [7] (*Figure*

2F). In addition, most (63.08 %) of iSNVs have low values of minor allele frequency ranged from 0.05 to 0.20, which were rarely shared among multiple individuals (*Figure S1E*). 81.02% of iSNVs were only occurred in one patient, and 11.24% were found in two individuals. According to the neutral evolution model, we constructed a simple framework to establish the basic expectations of both allele frequencies and the genomic distance between two alleles of iSNVs in a stimulated model (*Figure 2G, Figure 2H*). Compared to the neutral evolution model, both allele frequencies and iSNVs distribution in the genome displayed an uneven distribution of these iSNVs, suggesting a purifying and positive selection on these iSNVs (*Figure 2G, Figure 2H*).

Increased genetic diversity with the disease progression

To uncover the dynamic change of viral iSNVs in COVID-19 patients, we performed spatio-temporal analysis on the iSNVs along with the epidemic period and the disease progression, as well as that from different specimens. We observed a steady increase of iSNVs along with the epidemic period among the patient population (estimated value from 0.15 to 0.83 iSNVs/kb within 97 days) (*Figure 3A*). Meanwhile, iSNVs also accumulated during the infection in individual hosts, from 0.52 iSNVs/kb on the initial day to 0.85 iSNVs/kb on 30-day post symptom onset (*Figure 3B, Table S3*). This accumulation of mutations could also be observed in 81.97% (50 out of 61) paired samples from COVID-19 patients (*Figure S2A*). No significant difference in genomic coverage was observed among different sample types or along the epidemic period (*Figure S2B*). The increased genetic diversity could be observed in all three different kinds of specimens (*Figure S2C*). Of note, while more iSNVs were occurred in feces than in sputum and pharyngeal swab samples during the early days of symptom onset, the accumulation of iSNVs in digest system is slower than that in respiratory system (*Figure S2C, Table S3*).

Next, we measured dynamic changes of the genetic diversity in ORFs with the disease progression, and observed a rapid accumulation of mutation in the S, N, ORF1ab and other genes. The accumulation in nonsynonymous is more rapid in all genes, in comparison to synonymous mutations that respected as the neutrality (*Figure 3C*), and *S* gene displayed the highest accumulation rate (*Figure 3C*). To test whether the accumulation of genetic diversity was caused by their fitness selections, we computed the iSNVs number of nonsynonymous and synonymous divergency among genes, and found that 5,197 iSNVs displayed as nonsynonymous mutations as and higher than that in synonymous (1,593). The ratio of nonsynonymous to synonymous variants was 3.26. The ratio of nonsynonymous to synonymous of iSNVs in S gene (ratio=5.31) was significantly higher than the other part of viral genome (*Figure 3D*). The mean values of minor allele frequencies of non-synonymous and synonymous iSNVs were 0.189 and 0.195, respectively (*Figure S2B*). The ratio (Ka/Ks) between the number of nonsynonymous substitutions per nonsynonymous site (Ka) and the number per of synonymous substitutions per synonymous site (Ks) in S gene was also increased from 1.01 to 2.46 with the disease progression. Assuming that Ks represent a neutral expectation, Ka/Ks values of *S* gene over than 1 in the late of disease progression, indicated signatures of positive selection with disease progression at least in S genes (*Figure 3E*). In addition, the fraction of nonsynonymous mutations in predict epitope region [22] was significantly higher than expected (27.57% *vs.* 25.74%, *p-value*=0.016), while nonsynonymous mutations out of epitopes regions was

significantly lower (*Figure 3F*). The further correlation analysis between the fraction of nonsynonymous sites and the time of symptoms onsets, displayed nonsynonymous increased genes along with disease progression (*Figure S2E*), especially in S gene (*Figure 3G*).

RNA editing in the increase genetic diversity

Host-dependent RNA editing of APOBEC and ADARs has been acknowledged to cause the bias of mutational type in SARS-CoV-2 [23], thus, we measured all mutational types on all iSNVs. The top five mutation types of iSNVs were U-to-C, followed by C-to-U, G-to-U, A-to-G, and G-to-A (*Figure 4A*). Among them, four were the common substitutions caused by APOBECs/ADARs [23] (*Figure 4A*). In addition, unlike the A-to-I RNA editing signal in human transcriptome, we did not observe an obvious depletion of G bases in position -1 that presents at A-to-I edited positions (*Figure S3A*). We also calculated the correlation between the minor allele frequencies of iSNVs and the time post symptom onset. Previous study has showed an induction of APOBECs triggered by coronavirus infection but ADAR did not [23]. Consistently, the minor allele frequencies of C-to-U and G-to-A mediated by APOBEC-mediated RNA editing slightly increased *in vivo*. By contrast, the other mutational type including ADAR-mediated RNA editing *in vivo* did not increased along the infection time of patients (*Figure 4B, Figure S3B*). These results suggested RNA-editing of APOBECs were also impacted by the infectious of SARS-CoV-2, which turned back to change the mutation rates of C-to-U and A-to-G in SARS-CoV-2.

Influence of patient-derived immune-selection on the genetic diversity

To evaluate the influence of immune-selection on viral mutation, we measured the dynamic changes of the number of iSNVs within patient that stratified based on gender, age, illness severity and viral shedding time (*Figure S4A, Table 1*). Each sample was recalibrated based on symptom onset date (*Figure 5A*). The increased genetic diversity of iSNVs could be observed in all groups (*Figure 5A*). Female patients, mid-age (15-65y) patients, the patients with mild symptoms and the patients within 4-6 weeks viral shedding duration accumulated iSNVs much faster than other patients in each corresponding group (*Figure 5A*). Different slops and initial values were observed in these patient-groups, suggested a different fitness selection at the initial infection and the following infection stage along the day post symptom onset.

To further investigate the potential mutations influenced by patient-derived immune-selection, we compared the proportions population with or without iSNVs that frequently occurred in these patients. Among the 52 iSNVs sites that were highly occurred frequency in the population (shared by more than six patients), we identified 4, 5, 4 and 8 iSNVs significantly existed in server, old, long viral shedding time and male patients, compared to mild/moderate, young, short viral shedding time(<14 days) and female patients(*Figure 5B*). These iSNVs were distributed in ORF1ab, S, N, ORF6 and ORF8. Intriguingly, among the 52 high frequent iSNVs sites, we identified 27 iSNVs preferred occurred in severe patients; and 12 of them were hfSNPs sites. In contrast, the 25 iSNVs that showed preference in moderate patient did not overlapped with hfSNPs (*Figure 5C*, *p-value* < 0.001). This enrichment of hfSNPs were not observe in any other categories that stratified based on gender, age, and viral shedding time (*Figure S4B*), indicating that

the propagation of these hfSNPs might under patient-derived immune-selection. Taken together, the accumulation and the different distribution of iSNVs in each group, hinted the possibility of fitness selections along the symptom onset.

Uneven purifying selection process from iSNVs to SNPs

To identify the mutations purified from iSNVs to SNPs, we compared the genomic site of iSNVs and SNPs in public database [6,7]. Among 7037 iSNVs, 15.59% of iSNVs had been identified as SNPs before our observation period (May, 2020), and 11.28% of iSNVs was fixed from May, 2020 to December, 2020, and the rest iSNVs (73.13%) were still not fixed as SNPs (*Figure 6A*). Nonsynonymous iSNVs displayed a lower fixation rate than synonymous iSNVs (20.92% *vs.* 41.12%, *p-value* <0.001; *Figure 6B*), which is supported the model that nonsynonymous iSNVs rise to high frequency with an individual due to positive selection, but are less likely to become fixed in the population due to purifying selection. Then, we performed Fisher's exact test to compare the proportion of fixed mutations in each gene, and S and ORF1ab had a lower fixation rates (21.04% and 20.56%, *Figure 6C*), either in nonsynonymous or synonymous sites (*Figure S5A*). Consistently, the nonsynonymous-to-synonymous ratio of iSNVs in ORF1ab, N and S (excluding D614G) were over than that estimated in identified SNPs, presenting as an uneven purifying selection in these genes. As disease progress, iSNVs fixation rates in nonsynonymous and synonymous sites were stable in the population (*Figure 6D*, *Figure S5B*), indicating a similar purifying selection with disease progress. Interestingly, we also observed that mutation in iSNVs might be earlier detected before they fixed in SNPs. For example, accumulative frequencies on C7051T alleles had been observed in our studies before May, whereas the first C7051T SNP was reported until June, 2020 (*Figure 6F*). All these data indicate iSNVs provide prospective and complement genetic information to illustrate the SARS-CoV-2 evolution.

Molecular function on variations in S protein before purifying selection

S protein drives the cellular entry binding on receptors and acts as a major determinant of host range, cell, tissue tropism, and pathogenesis of coronaviruses [24]. Therefore, we further analyzed 21 of total 606 iSNVs sites identified in the coding region of S protein that caused 20 amino acids changes. 1) Nine were detected outside of RBD more than six individuals; including three linked iSNVs (A22298G, G22299A, and G22302A) with the substitutions of two amino acids were located ahead of RBD of protein (R246E and S247N, *Figure 7A*). 2) Eleven resided within the receptor-binding domain (RBD) or S1/S2 cleavage sites more than two individuals, including seven iSNVs were located in the receptor-binding motif (RBM). We compared the mutation sites of seven iSNVs in RBM in SARS-CoV-2 to the consensus sequencing in other animal (Bat and Pangolin) SARS-CoV-2-like coronavirus, and found that all these sites were heterogeneous (*Figure 7A*). The patients with these mutations have no contact history except for two patients with G485V (*Table S4*); no iSNVs emerged in the first time point of genome sequencing in patients with multiple time point (*Figure S6A, Table S4*); and no evidence supported these iSNVs were linked in the genome. Taken together, these data indicated that iSNVs in RBD seem to be generated by independently viral evolution.

To elucidate the effects of these mutations at the molecular level, we first used the SARS-COV-2 pseudovirus infection assay to assess the viral entry efficacy of 20 of the 22 S-protein mutants except S50L and M731V. Compared to reference strain, 18 of 20 tested mutants displayed a decreased (fold-change <0.25) or comparable efficacy (fold-change <4 or fold-change >0.25) of viral entry; only mutants R685G and D614G exhibited a similar level of increased viral entry efficacy (*Figure 7B*, fold-change >4). Since the residue L518V is far away from the binding interface between S protein and hACE2/CB6 and four mutants (N422K, E471K, G485V and Y505C) displayed a significant decrease (fold-change < 0.25) of viral entry efficacy, the other five mutants in RBD (V407L, L452Q, V483F, Q493H and Q498H) were test for the sensitivity to CB6 (*Table S5*). Wide type and D614G was included as control. Modest differences between these RBD variants and reference strain (within 4-fold) were observed in their susceptibility to CB6 (*Table S5*). It is worth mentioning that some variants including V483F, Q493H and Q498H were even more sensitive to CB6 compared with reference strain, which illustrate that CB6 antibody could still block the entry of virus with the existence of RBD mutation.

Then, we focused on the mutations within RBM and simulated the bonding of the corresponding mutants bound to human Angiotensin-converting enzyme 2, hACE2 [25] and neutralizing antibodies CB6 [26] using molecular dynamics simulations (*Figure 7C*). The simulation of the mutation L518V were not conducted as previously described. For comparison, we also simulated and inspected the binding of WT RBD to hACE2 and CB6, respectively. Cα root-mean-square deviation (Cα RMSD) of the complex of different mutant RBD bound to hACE2 varied within the similar ranges of the corresponding complex of WT RBD, suggesting that the 9 mutations may not induce dramatic conformational changes (*Figure S6B*). Then we looked into more structural details and found that different mutations could affect the binding to hACE2 in varied ways. For example, the residue 422 was mutated from the uncharged side chain N to the positive charged K in mutant N422K. Therefore, a much stronger hydrogen bond was formed between K422 and E406. In addition, strong repulsion between K422 and R403 and attraction between K422 and Y453 induced the broken of the hydrogen bonding between Y505 and E37 and Y453 and H34. As a result, both the numbers of hydrogen bond and contact area formed between the mutant RBD and hACE2 were reduced apparently (*Figure 7D*). As for the mutation G485V, since V has a relatively bulky side chain comparing to G, in order to reposition the bulky side chain, G485V mutation led to the escaping of F486 from the hydrophobic pocket formed by the residues including F28, L79, M82 and Y83 of hACE2. As a result, not only the contact area but also the numbers of hydrogen bond formed between mutant RBD and hACE2 was reduced. Especially, the hydrogen bond formed between Y505 of S protein and E37 of hACE2 which is important in binding was lost (*Figure 7E*). As a possible consequence, the binding of mutants N422K/G485V to hACE2 would be weakened. The binding free energy calculations using MM/GBSA method also indicated that the weakened binding of N422K/G485 and hACE2 (*Figure 7C*). The hydrogen bond and contact area changes of the other mutants were also inspected and discussed (*Figure S6B*). These data illustrated most of the mutants with observed iSNVs in study decreased the viral entry ability by the reduce of hydrogen bond and/or contact area formed between mutant RBD and hACE2. We also analyzed simulation results of the complex with the mutant RBD and the antibody CB6 (*Figure S6C*). Most of the mutants reduced the contact area a lot compared to WT, such that, the decreased binding

affinities were observed which was manifested by the reduced number of hydrogen bond and higher binding free energy obtained in MM/GBSA calculations. Therefore, both the observation of mutations in pseudoviral infection assay and computation of their interaction displayed a weaken trends of viral infection in most of iSNVs identified in S protein.

Discussion

The ongoing pandemic of SARS-CoV-2 has raised worldwide alert. Mutations of SARS-CoV-2 arise naturally as the virus replicates, and presented as SNPs under the selection and transmission. Within one year after the first case of COVID-19 was confirmed, thousands of mutations on SNPs have already been identified, among which only a very small minority caused the changes in SARS-CoV-2 infectivity and immune evasion. The mutations that cropped up in SARS-CoV-2 genome can be used to unearth the signs of selections that accumulated during viral evolution [27,28]. In this study, we illustrated intra-host variations of SARS-CoV-2 and demonstrated the two-step fitness selections. The first step of selection occurs after randomized mutations were generated, and a positive selection (might be patient-derived fitness selection) mediates an accumulation of iSNVs that indicated by increased genetic diversity along with COVID-19 disease progression. The second step selection is purifying selection, which accounts for the reduction of nonsynonymous mutations from iSNVs to SNPs. This two-step fitness model highlights the SARS-CoV-2 evolution at within-host scale to better understand the substrate for global pandemic.

The positive selection resulted two characters of iSNVs: i) the increase of genetic diversity along with COVID-19 progress, and ii) an uneven distribution of iSNVs among patients and genome. Under the selection, the former forces virus presents as more mutants that changed amino acids, which might affect key features of the virus including the infectivity, virulence and immunogenicity. Studies of Ebola virus [29], Lassa virus [30] and influenza virus [11] have also compared the proportions of nonsynonymous to synonymous, also displayed intra-host positive selection. High rates of mutation accumulation over short time periods in SARS-CoV-2 have been reported previously in studies of immunodeficient or immunosuppressed patients who are chronically infected with SARS-CoV-2 [31-33]. The latter result of positive selection displayed the different pressure of positive selection among patients and genomic regions. The spatio-temporal analysis on the viral mutations within hosts to monitor dynamics of genomic changes along with the effects of age, gender as well as viral shedding time and illness severity for better understanding viral nature history and the selective pressure in different patients, to trace viral spreading, and guide the design of vaccines. Several recent publications around the SARS-CoV-2 genome have found signals of positive selection [34-36] and conservation within the gene encoding spike glycoprotein. Because the detection of viral mutations in most studies still relied on consensus genomic sequences, which was after the purifying process, this fitness selection along the day post symptom onset has not been observed at SNP level. Instead of relying solely on the dominant viral sequences in the patients, we included low frequency viral mutations within hosts, the iSNVs, which might be better to reflect the dynamic changes of selective pressure for more efficient viral spreading or immune evasion.

As nonsynonymous mutation accumulated in individuals, the observation of differences between within- and inter- patients suggested a strong purifying selection from iSNVs to SNPs, especially in S and ORF1ab. This purifying selection process is also observed in Lassa virus [30], Ebola virus [29] and dengue virus[37]. In addition, we assessed the viral entry efficacy of high frequencies mutations iSNVs suggested a weaken trend of viral evolution. However, we remain unclear whether these weaken virus could persists in patients with long COVID-19 [38]. However, it should be note that all these raised mutations observed in population are expected under the current situation. Once circumstance was changed, such as antibody therapy used, the selection pressures on each mutation should be varied. For example, the selection arising from antibody might be different and strong. These strong selections might rapidly remodel multiple virus genetics through direct selection or genetic drift. Therefore, more potential mutations of SARS-CoV-2 that presented as iSNVs should be noticed, as well as the mutation R685G in cleavage sites increasing the viral entry efficiencies, at least in some cells and might direct results of SARS-CoV-2 persisted in multiple organs [38], although it was never occurred as SNPs in any patient in public database.

The big concern about SARS-CoV-2 evolution is whether serious mutants that changed the viral infections or virulence, and whether these mutants would spread worldwide in future. Recently, the emergence and spread of mutants D614G [8] and the strain from a distinct phylogenetic cluster lineage B.1.1.7 [9] has been considered as a pandemic threaten to public health[39,40]. These mutants were identified benefited by the global genomic monitor program. Consistent with our results, these mutations have been detected as iSNVs in previous samples [41]. Taken considering that iSNVs emerged more nonsynonymous mutations by positive selection, genomic monitor at within-host scale might be an important complement of current genomic monitors. Controlling or eliminating infectious agents at their sources of transmission should also consider to control the viral mutations to reduce their further adaption in future.

In the urgent time of vaccine development and the treatment strategy selection, we might not have enough time to wait for the mutations to fix in the population to prove their functions [42]. Early knowledge of potential evolution would benefit vaccine design [5,43]. The associations between these emerged mutations and the illness severity and treatments should be carefully considered. More genomic data at intra-host level should be explored to illustrate the genetic selections in whole genome scale, and their potential effects on illness severity, clinical outcomes as well as the susceptibility to different populations.

Methods

Patients and clinical cohort

Our study included all COVID-19 confirmed and admitted patients between January 20 and April 30, 2020 in Beijing Ditan Hospital, where received and confirmed the first case in Beijing, China. Totally, we enrolled 204 cases, which occupied 34.4% confirmed patients in Beijing. All patients were confirmed by the RT-PCR test in pharyngeal swabs and administrated in Beijing Ditan Hospital. All patients are treated and

managed in the ward after being diagnosed. Standardized electronic medical records were applied to collect basic demographics and epidemiological features, medical histories, and clinical information. Patients were diagnosed and discharged according to the 7th guideline for the diagnosis and treatment of COVID-19 from the National Health Commission of the People's Republic of China, where they meet all following criteria: 1) afebrile (body temperature ≤37°C) for more than three days; 2) resolution in respiratory symptoms; 3) substantially improved acute exudative lesions on chest computed tomography (CT) images; 4) two consecutively negative RT-PCR results in respiratory specimens (at least 24 hours apart). Server patients included critical illness in this study. To calculate the sampling time to symptom onset, we applied all days to the same time scale. For the patient with symptom, we set initial symptom day as 0, and for the asymptotic patient the original time is the day with first positive RT-PCR test. The date of converting to negative were defined as the date when all specimens (including pharyngeal swab, sputum and feces specimen) turned to negative. Viral shedding time were calculated from the initial day to the date converting to negative.

The internal use of samples for diagnostic workflow optimization was agreed under the medical ethical rules of each of the participating partners and approved by the Review Board of Beijing Ditan Hospital (Beijing, China) and the Ethics Committee of State Key Laboratory of Pathogen and Biosecurity (KT2020-006-01).

Laboratory procedure

Clinical samples including pharyngeal swabs, sputum or feces specimens for the RT-PCR test were collected multiply according to the instruction of the infection prevention and control measures in Chinese guidance on infection prevention and control in healthcare settings. RNA was extracted as previously described in P2+ and/or P3 laboratory [44]. Viral RNA was extracted using the QIAamp® Viral RNA Mini Kit according to the manufacturer's instructions, except that carrier RNA was omitted to facilitate downstream high-throughput sequencing analysis. Real-time RT-PCR assays were recommended for the nucleic acid test and determined by RT-PCR targeting the open reading frame 1ab (ORF1ab) region and nucleoprotein (N) gene of SARS-CoV-2, as described elsewhere [2]. A cycle threshold value less than or equal to 37 in at least one gene was interpreted as positive for SARS-CoV-2, according to Chinese national guidelines. The Ct values in RT-PCR of these samples were available and ranged from 12.00 to 37.52.

High-throughput genomic sequencing of the viral genome

We collected 183 pharyngeal swabs, 241 sputum, and 113 fecal samples for meta-transcriptomic sequencing. Viral RNA was extracted using the protocol described above. After performing rRNA removal using the MGIEasy rRNA Depletion Kit (BGI, Shenzhen, China), we used the novel Metagenomic RNA EnRichment VirAl sequencing (MINERVA) approach to obtain virus sequences [16]. Briefly, this approach uses direct tagmentation of RNA/DNA hybrids using Tn5 transposase to greatly simplify the sequencing

library construction process, allowed us to conduct rapid library preparation using low volume input RNA templates (5.4 ul) within 4 hours. The meta-transcriptome libraries further underwent the enrichment process using biotinylated RNA probes targeting the whole viral genome (iGeneTech, Beijing, China). The final viral-enriched libraries were sequenced on an Illumina NextSeq500 in 2x75bp pair-end mode.

Sequencing analysis of the viral genome

Quality control and error correction were implemented as previous reported[45]. To avoid nucleotide-specific substitution errors in each read, we removed the low-quality bases at the end of reads with a threshold of Q20 and a minimum read length requirement of 50 bp. Reads without their corresponding paired reads were disregarded. The remaining paired reads were used as clean reads.

High quality viral genomic data were selected for iSNVs analysis. We firstly mapped the clean read data to the reference genome Wuhan-Hu-1 (GenBank accession no. NC_045512.2), and obtained 4.11 Mb (QRI: 0.44-22.96 Mb) high-quality viral reads per sample. After removing the low-quality genomic data, we obtained 402 samples with enough data, and meet the following criteria: i) with sequencing depth >=1 and reference coverage >=50%, ii) and depth >=100 and reference coverage >=10%.

iSNVs calling to avoid sequencing error and contaminants

Calling of iSNV was performed as previously described but with different parameters [17]. Briefly: (1) Sequencing reads were pair-ended aligned to the reference genome sequence (GenBank accession no. NC_045512.2 [18]) using Bowtie2 v2.1.0[46] by default parameters and the alignments were reformatted using SAMtools v1.3.1[47]; (2) for each site of the SARS-CoV-2 genome, the aligned low-quality bases and indels were excluded to reduce possible false positives and the site depth and strand bias were re-calculated; (3) samples with more than 3,000 sites with a sequencing depth ≥100× were selected as candidate samples for iSNV calling.

A series of criteria were used to ensure high quality iSNVs with Q30 reads support and: (1) minor allele frequency of ≥5% (a conservative cutoff based on an error rate estimation); (2) depth of the minor allele of ≥5; and (3) strand bias of the minor allele less than tenfold or the fisher-exact test for the major allele and minor allele, and (4) to avoid the errors by reads mapping, the serial adjacent iSNVs (distance <50bp) that contains >5 iSNVs were further filtered.

PCR, sequencing errors and the potential contaminants induced in metagenomic sequencing always the big concern of next generation sequencing. To better detect these contaminants and sequencing error, 1) we firstly enrolled negative control and pharyngeal swab from the healthy to confirm whether we have false-positive of virus in the study. 2) The iSNVs patterns from each round were carefully examined, and we did not observe similar iSNVs pattern in each round. 3) We also examined the nucleotide stat in each site, we only detected 12/7037 iSNVs containing three polymorphic state. Compared to the substitution, the sequencing error might be randomized to occur nucleotide acid. 3) Sanger sequencing technology was used to validate the iSNVs result. We selected the major mutations occurred in one cluster to perform

PCR-based Sanger sequencing, and displayed with a consistent result. 4) Considering the previous research[45] by PCR amplicon that the identification of iSNV with the threshold of MuAF >0.007 could be a <0.001 false discovery rate (FDR), we elevated the threshold of MuAF to 0.05, and the result should efficient reduce FDR, and avoid all sequencing errors and potential contaminant in the study.

Normalized iSNVs and mutation rate estimation

The iSNV number for each sample were normalized to iSNVs/kb, where only the available region to identify iSNVs were calculated. The iSNVs sites were calculated along the genome by removing the duplicated iSNVs in each patient, where iSNVs in different samples of the same patients were only calculated once. We also applied linear regression to evaluate the correlations between iSNVs and the time of outbreak and infections in each patient.

We defined minor allele frequency (minor AF) as the allele frequency of the minor allele, whether reference allele or alternative allele, representing the potential substitution in each individual. Mutation allele frequency (MuAF) was defined as the ratio of alternative alleles and total alleles, representing the substitution in an outbreak. We estimated the number of mutations (m) in different mutation rate frequencies (a). According to the Bozic's model[48], the expected number of mutations above frequency was: **see equation 1 in the supplementary files section.**

In this formula, μ represented the mutation rate occurs per genome per replication cycle, b represented the virus reproduce rate and d represented the rate of virus leave the population. We applied a generalized linear model to calculate the parameter and plot the predicted lines. The SARS-CoV-2 virus reproducing and declining rate b and d was not clear. For comparison of mutation rate, we displayed the expected mutation number with different mutation rate with virus reproducing and declining rate in influenza virus [49,50].

The simulation of iSNV mutation position distribution was permutated with runif() function in R package, following the uniform distribution with iSNV number as parameters. Then we ranked the real iSNV and simulated position and calculated the genomic distance between two neighbor mutants. The genomic distances between two simulated iSNVs followed Poisson distribution, while that estimated iSNV biased from the simulation.

iSNV functional and epidemiological annotation

The iSNVs were annotated by Perl scripts and compared with the genomic annotation file of reference genome Wuhan-Hu-1(NC_045512.2) from NCBI. The public SNP files was downloaded from public database 2019nCoVR [6,7] (https://bigd.big.ac.cn/ncov/variation/annotation) on November 19, which collected the whole genome sequences from CNGBdb, GenBank, GISAID, GISAID, GWH and NMDC. To better understand the frequencies of SNPs that occurred in population, each iSNVs were compared to the levels of SNPs according to the frequency in 2019nCoVR at the same period of the sampling. Level I to III was according to previous definition in public database, whereas level I represents the frequency was

more than 0.05; level II represents the frequency was between 0.01 and 0.05; level III represents the frequency was less than 0.01. The iSNVs never appeared in public database, were named as level IV. Ka and Ks was calculated by KaKs_Calculator[51] (version 2.0, June. 2009) with MYN model[52]. The epitope region was downloaded from the IEDB database (https://www.iedb.org/) [22], with Epitopes set to “Linear Epitope” and Organism set to “SARS-CoV-2” and Host set to “Humans”.

High correlated iSNVs in genome by phasing analysis

Pairwise phasing analysis was performed for adjacent iSNVs (distance <50bp) as previous reported[19]. For a given pairwise iSNVs, reads harboring both positions were extracted from the alignment (SAM file). The reads with both sites mutated are referred to as phased reads, and those with only one site mutated are referred to as non-phased reads. The ratio of phased to non-phased reads that more than 0.9 was selected as phased iSNVs. Moreover, the phased alternative iSNVs allele frequency should >0.05. For the phased iSNVs in a same protein codon, we re-annotate the iSNVs with phased alleles.

In spite of short distance (<50 bp) in the genome, the correlation of long-distance (distance >50 bp) pairwise iSNVs were also calculated by linear adjust R-square of variation. High correlated and potential linked iSNVs should be identified in at least 3 samples, and with a high correlation of MuAFs (>0.6).

S protein structure analysis and molecular dynamics simulation

All Molecular Dynamistic (MD) simulations were performed using AMBER 20 package. The crystal structure of binary complexes between hACE2 and RBD (PDB ID codes 6LZG) [25] was used as the initial structure in MD simulation. Protein was modeled with AMBER FF14SB [53] all-atom protein force field and solvated by a truncated octahedron TIP3P [54] water box in which the boundary is at least 11 Å from any protein atoms. The solvated protein was neutralized and filled with a concentration of 0.13 M of KCl salt. In these simulations, the SHAKE [55] algorithm with a relative geometric tolerance of 10^{-5} was used to constrain all chemical bonds. Mass repartitioning was also applied, which means the mass of the heavy atom that the hydrogen is attached to is adjusted so that the total mass remains the same. Thus, all dynamics utilized a 4 fs time step. Long-range electrostatics was treated by the particle-mesh Ewald (PME) [56] method with default settings and a 9 Å direct space nonbonded cutoff was used in all simulations. The system was first subjected to 10,000 steps of minimization and then gradually heated to 300 K under constant volume conditions in 1 ns. After another 5 ns of simulations using the constant isothermal–isobaric ensemble at 1 atm and 300 K, each system was equilibrated for an additional 10 ns. The Monte Carlo baristas and a weak-coupling thermostat were used. MD simulations were extended for 300 ns with coordinates recorded every 10 ps. The same procedure was followed for simulations of the antibody CB6[26] and RBD complex (PDB ID code 7C01 and 7BWJ). Nine RBD mutants: V407L, N422K, L452Q, E471K, V483F, G485V, Q493H, Q498H, Y505C bound to either hACE2 or CB6 were also simulated. The simulations of these systems were also followed the same procedure. In the analysis of the simulation trajectories, hydrogen-bond is defined as a geometry with a cutoff length of 3.5 Å between the two heavy atoms of the hydrogen-bond donor and acceptor and an X-H· · ·Y (X and Y stand for heavy

atoms) angle cutoff of 135°. A hydrogen-bond is counted when the distance between X and Y is less than 3.5 Å and the X-H· · ·Y angle is greater than 135°.

Package of pseudoparticles bearing spike protein from SARS-CoV-2 and variants

The codon-optimized S gene of SARS-CoV-2 (NC_045512.2) were constructed into pSecTag2/Hygro A plasmid and using as a template to generate S mutants following site-directed mutagenesis PCR. The various spike protein pseudovirus bearing luciferase reporter gene were packaged as reported previously [57]. In briefly, 24 h prior to transfection, 6×10^5 293T cells were plated per well in 6 well plates. All transfections used 4 μg plasmid DNA with 6 μl TurboFect transfection reagent (Thermo Fisher) in 400 μL Opti-MEM (Gibco). Single-cycle HIV-1 vectors pseudotyped with SARS-CoV-2 Spike protein, either reference or mutants, were produced by transfection of either HIV-1 pNL4-3 Δenv Δvpr luciferase reporter plasmid (pNL4-3.Luc.R-E-) in combination with the indicated Spike expression plasmids, at a ratio of 4:1. Viral supernatant was harvested at 48 h and 72 h post transfection, spun down by centrifugation to remove cell debris and filtered through a 0.45 μm filter unit (Sartorius). Lenti-X p24 rapid titre kit (Takara Bio) was used to quantify the viral titres following the manufacturer's instructions.

Generation of cell lines expressing hACE2 and Virus infectivity assay

T Rex 293 hACE2 cell line, which express human ACE2 in tetracycline-dependent manner, was established previously [58]. In brief, 5×10^5 Flp-IN T Rex cells were plated per well in a 6 well plate. The next day cells were cotransfected with a pcDNA5/FRT-derived human ACE2 expression plasmid and pOG44 (Invitrogen) at a molar ratio of 1:1. Two days after transfection, cells were trypsinized and reseeded at less than 25% confluence. The hACE2 cDNA-integrated cells were selected with 250 μg/ml hygromycin and 5 μg/ml blasticidin. Two weeks later, separate colonies appeared, and the pool of such cells was expanded to generate cell lines that express hACE2 (T Rex 293 hACE2) upon the addition of tetracycline into the culture medium at the concentration of 1 μg/ml.

T Rex 293 hACE2 cells transfected with pCAGGS-TMPRSS2 plasmid were seeded in a 96-well plate at a concentration of 2×10^4 cells per well and cultured for 12 h upon the addition of tetracycline (1 μg/ml). Using the HIV-1 p24 antigen quantification, we normalized the pseudotyped virus particles to the same amount (10 ng of p24). After normalization, 100 μl of the pseudotyped virus with 10-fold dilution was added to wells in 96-well T Rex 293 hACE2/TMPRSS2 culture plate. The plates were then incubated at 37°C in a humidified atmosphere with 5% CO_2. The culture medium containing 2% FBS was refreshed after 12 h and incubated for an additional 48 h. Assays were developed with a luciferase assay system (Promega), and the relative light units (RLU) were read on a Promega GloMax Luminometer. Three to seven independent experiments were conducted with triplicate samples.

CB6 mAbs neutralizing assay

For the neutralization assay, T Rex 293 hACE2 cells transfected with pCAGGS-TMPRSS2 plasmid were seeded in a 96-well plate at a concentration of 2×10^4 cells per well and cultured for 12 h upon the

addition of tetracycline (1 μg/ml).100 μl supernatant containing pseudoviruses was incubated with equal volume of five-fold serially diluted antibodies for 1 h at 37 °C. CB6 mAb were tested in the concentrations ranging from 0.64 ng/ml to 10.00 μg/ml. The mixtures of pseudoviruses and CB6 mAb were then added to T Rex 293 hACE2/TMPRSS2 cells in 96-well plate with two replicates. After a 12 h incubation, the medium was replaced with DMEM containing 2% FBS and the samples were incubated for an additional 48 h at 37 °C. Luciferase activity was measured using a GloMax 96 Microplate luminometer (Promega). The titers of CB6 mAb were calculated as 50% inhibitory concentration (IC50) using GraphPad Prism 6.0.

Statistical analysis of iSNVs and patients' clinical characteristics

Continuous variables were expressed as median and interquartile ranges (IQR) as appropriate. Categorical variables were summarized as the numbers and corresponding percentages in each category. The distribution of codon positions was compared with the normal distribution using Kolmogorov-Smirnov test. The correlation between iSNVs and the effects of age, gender, illness severity and viral shedding time were Kruskal-Wallis test. Only the sites on which iSNVs were occurred in over six patients were included. The correlation between the proportion distribution of the population with iSNVs and hfSNPs in public database were calculated by Fisher's exact test.

Additional Information:

Supplementary Information is available for this paper.

Data availability

The sequencing data has been submitted to the National Genomics Data Center, China National Center for Bioinformation (Accession number: HRA000181, HRA000349).

Declarations

Acknowledgments

We thank all health care workers involved in the diagnosis and treatment of the COVID-19 patients in Beijing Ditan Hospital, Capital Medical University. This research was supported by Beijing Hospital Authority (Grant No. DFL20191801), National Natural Science Foundation of China (92053202, 22050003 and 21873007).

Author Contributions

J.L., P.D. and N.D. performed the sequence data analysis. D.C and X.Z performed the pseudoviral assays infection experiment. L.Y., and Y. G. preformed the molecular dynamics analysis. F.Z., J.Z., Y.S., R.S., W.X., Z.C., X.W., J. L., Y.X., G.G, Q.W., L.P, F.Y. collected and analyzed the clinic data. C.S., K.H., L.D., J.L., J.Y., M.H., J.H., Y.Y. performed most of the experiments. A.W., Y.G., Y.H., J.W. and G.G provided intellectual input and helped to interpret the data. J.L., P.D., L.Y., D.C, C.C., Y.G., and H.Z. wrote the manuscript. All of the

authors discussed the results and commented on the manuscript. C.C., Z.H., J.W., Y.H., Y.G and G.G. supervised the study.

Declaration of Interests:

The authors declare no competing interests.

References

1. Guan, W. J. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* **382**, 1708–1720, doi:10.1056/NEJMoa2002032 (2020).
2. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727–733, doi:10.1056/NEJMoa2001017 (2020).
3. Coronaviridae Study Group of the International Committee on Taxonomy of, V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5**, 536–544, doi:10.1038/s41564-020-0695-z (2020).
4. https://www.who.int/westernpacific/emergencies/covid-19.
5. Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell*, doi:10.1016/j.cell.2020.07.012 (2020).
6. Zhao, W. M. *et al.* The 2019 novel coronavirus resource. *Yi Chuan* **42**, 212–221, doi:10.16288/j.yczz.20-030 (2020).
7. Song, S. *et al.* The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. *Genomics, Proteomics & Bioinformatics* (2020).
8. Korber, B. *et al.* 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 (2020).
9. Santos, J. C. & Passos, G. A. The high infectivity of SARS-CoV-2 B.1.1.7 is associated with increased interaction force between Spike-ACE2 caused by the viral N501Y mutation. *bioRxiv*, 2020.2012.2029.424708, doi:10.1101/2020.12.29.424708 (2021).
10. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, doi:10.2807/1560-7917.ES.2017.22.13.30494 (2017).
11. McCrone, J. T. *et al.* Stochastic processes constrain the within and between host evolution of influenza virus. *Elife* **7**, doi:10.7554/eLife.35962 (2018).
12. Gudbjartsson, D. F. *et al.* Spread of SARS-CoV-2 in the Icelandic population. *N Engl J Med* (2020).
13. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*, doi:10.1038/s41564-020-0770-5 (2020).
14. Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *Nati Sci Rev* **7**, 1012–1023, doi:10.1093/nsr/nwaa036 (2020).

15. Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. *Nature* **538**, 193–200, doi:10.1038/nature19790 (2016).
16. Chen, C. *et al.* MINERVA: A Facile Strategy for SARS-CoV-2 Whole-Genome Deep Sequencing of Clinical Samples. *Mol Cell* **80**, 1123–1134 e1124, doi:10.1016/j.molcel.2020.11.030 (2020).
17. Ni, M. *et al.* Intra-host dynamics of Ebola virus during 2014. *Nat Microbiol* **1**, 16151, doi:10.1038/nmicrobiol.2016.151 (2016).
18. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269, doi:10.1038/s41586-020-2008-3 (2020).
19. Chen, C. *et al.* Phylogenomic analysis unravels evolution of yellow fever virus within hosts. *PLoS Negl Trop Dis* **12**, e0006738, doi:10.1371/journal.pntd.0006738 (2018).
20. Debbink, K. *et al.* Vaccination has minimal impact on the intrahost diversity of H3N2 influenza viruses. *PLoS Pathog* **13**, e1006194, doi:10.1371/journal.ppat.1006194 (2017).
21. Zhang, X. *et al.* Viral and host factors related to the clinical outcome of COVID-19. *Nature* **583**, 437–440, doi:10.1038/s41586-020-2355-0 (2020).
22. Dhanda, S. K. *et al.* IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic Acids Res* **47**, W502-W506, doi:10.1093/nar/gkz452 (2019).
23. Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci Adv* **6**, eabb5813, doi:10.1126/sciadv.abb5813 (2020).
24. Li, F. Receptor recognition and cross-species infections of SARS coronavirus. *Antiviral Res* **100**, 246–254, doi:10.1016/j.antiviral.2013.08.014 (2013).
25. Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **181**, 894–904 e899, doi:10.1016/j.cell.2020.03.045 (2020).
26. Shi, R. *et al.* A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature*, doi:10.1038/s41586-020-2381-y (2020).
27. Sanjuan, R. & Domingo-Calap, P. Mechanisms of viral mutation. *Cell Mol Life Sci* **73**, 4433–4448, doi:10.1007/s00018-016-2299-6 (2016).
28. Xue, K. S. & Bloom, J. D. Linking influenza virus evolution within and between human hosts. *Virus Evol* **6**, veaa010, doi:10.1093/ve/veaa010 (2020).
29. Ladner, J. T. *et al.* Evolution and Spread of Ebola Virus in Liberia, 2014–2015. *Cell Host Microbe* **18**, 659–669, doi:10.1016/j.chom.2015.11.008 (2015).
30. Andersen, K. G. *et al.* Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. *Cell* **162**, 738–750, doi:10.1016/j.cell.2015.07.020 (2015).
31. Choi, J. Y. COVID-19 in South Korea. *Postgrad Med J* **96**, 399–402, doi:10.1136/postgradmedj-2020-137738 (2020).
32. Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183**, 1901–1912 e1909,

doi:10.1016/j.cell.2020.10.049 (2020).

33. Kemp, H. I., Corner, E. & Colvin, L. A. Chronic pain after COVID-19: implications for rehabilitation. *Br J Anaesth* **125**, 436–440, doi:10.1016/j.bja.2020.05.021 (2020).
34. Velazquez-Salinas, L. *et al.* Positive Selection of ORF1ab, ORF3a, and ORF8 Genes Drives the Early Evolutionary Trends of SARS-CoV-2 During the 2020 COVID-19 Pandemic. *Front Microbiol* **11**, 550674, doi:10.3389/fmicb.2020.550674 (2020).
35. Berrio, A., Gartner, V. & Wray, G. A. Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. *bioRxiv*, 2020.2009.2016.300038, doi:10.1101/2020.09.16.300038 (2020).
36. Forni, D. *et al.* Extensive Positive Selection Drives the Evolution of Nonstructural Proteins in Lineage C Betacoronaviruses. *J Virol* **90**, 3627–3639, doi:10.1128/JVI.02988-15 (2016).
37. Holmes, E. C. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *J Virol* **77**, 11296–11298, doi:10.1128/jvi.77.20.11296-11298.2003 (2003).
38. Meeting the challenge of long COVID. *Nat Med* **26**, 1803, doi:10.1038/s41591-020-01177-6 (2020).
39. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, doi:10.1038/s41586-020-2895-3 (2020).
40. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468, doi:10.1126/science.abe8499 (2020).
41. Lythgoe, K. A. *et al.* Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. *bioRxiv* (2020).
42. Thanh Le, T. *et al.* The COVID-19 vaccine development landscape. *Nat Rev Drug Discov* **19**, 305–306, doi:10.1038/d41573-020-00073-5 (2020).
43. Poh, C. M. *et al.* Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients. *Nat Commun* **11**, 2806, doi:10.1038/s41467-020-16638-2 (2020).
44. Chan, J. F. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**, 514–523, doi:10.1016/S0140-6736(20)30154-9 (2020).
45. Ni, M., Chen, C. & Liu, D. An Assessment of Amplicon-Sequencing Based Method for Viral Intrahost Analysis. *Virol Sin* **33**, 557–560, doi:10.1007/s12250-018-0052-z (2018).
46. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359, doi:10.1038/nmeth.1923 (2012).
47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, doi:10.1093/bioinformatics/btp352 (2009).
48. Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLoS Comput Biol* **12**, e1004731, doi:10.1371/journal.pcbi.1004731 (2016).
49. Xue, K. S., Moncla, L. H., Bedford, T. & Bloom, J. D. Within-Host Evolution of Human Influenza Virus. *Trends Microbiol* **26**, 781–793, doi:10.1016/j.tim.2018.02.007 (2018).

50. Beauchemin, C. A. & Handel, A. A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead. *BMC public health* **11**, S7 (2011).
51. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77–80, doi:10.1016/S1672-0229(10)60008-3 (2010).
52. Zhang, Z., Li, J. & Yu, J. Computing Ka and Ks with a consideration of unequal transitional substitutions. *BMC Evol Biol* **6**, 44, doi:10.1186/1471-2148-6-44 (2006).
53. Maier, J. A. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. **11**, 3696–3713 (2015).
54. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. J. T. J. o. c. p. Comparison of simple potential functions for simulating liquid water. **79**, 926–935 (1983).
55. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. J. J. o. c. p. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. **23**, 327–341 (1977).
56. Darden, T., York, D. & Pedersen, L. J. T. J. o. c. p. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. **98**, 10089–10092 (1993).
57. Zhao, X. *et al.* Broad and Differential Animal Angiotensin-Converting Enzyme 2 Receptor Usage by SARS-CoV-2. *J Virol* **94**, doi:10.1128/JVI.00940-20 (2020).
58. Zhao, X. *et al.* Inhibition of endoplasmic reticulum-resident glucosidases impairs severe acute respiratory syndrome coronavirus and human coronavirus NL63 spike protein-mediated entry by altering the glycan processing of angiotensin I-converting enzyme 2. *Antimicrob Agents Chemother* **59**, 206–216, doi:10.1128/AAC.03999-14 (2015).

Table

Table 1. The differences of iSNVs count , normalized iSNV number and numbers of patients with iSNV among patient subpopulation

Characteristics	Included Patients No.(%)	iSNV count median (qu1-qu3)	Wilcoxon test P-value	# of normalized iSNV median (qu1-qu3)	Wilcoxon test P-value	#of patients with iSNV (%)	Fisher-exact test P-value
Total	170						
Age groups(years)							
0-15	17(10.00%)	28(6-52)	0.648	0.637(0.339-1.130)	0.243	17(100%)	1
16-65	131(77.06%)	17(5.5-50)	-	0.552(0.268-0.863)	-	124(94.65%)	-
>66	22(12.94%)	37.5(8.75-63.5)	0.224	0.505(0.332-0.654)	0.866	22(100%)	0.594
Sex							
Female	83(48.82%)	15(3.5-51)	**0.029**	0.478(0.247-0.795)	**0.019**	78(93.97%)	0.269
Male	87(51.18%)	25(9-55.5)		0.599(0.342-0.962)		85(97.70%)	
Disease level							
Mild	39(22.94%)	21(5.5-57)	0.706	0.489(0.269-0.812)	0.746	37(94.87%)	1
Moderate	98(57.65%)	17.5(5-44)	-	0.567(0.270-0.872)	-	93(94.90%)	-
Severe	33(19.41%)	21(11-66)	0.135	0.545(0.356-0.932)	0.519	33(100%)	0.330
Infection Duration							
0-14	22(12.94%)	7(3-18.25)	**0.004**	0.338(0.182-0.576)	0.087	22(100%)	0.556
15-28	48(28.23%)	16(6-43.25)	0.225	0.546(0.336-0.871)	0.921	46(95.83%)	1
29-42	43(25.29%)	29(12-90)	0.384	0.592(0.335-0.893)	0.767	41(95.35%)	1
>43	57(33.53%)	25(6-68)	-	0.606(0.245-0.874)	-	54(94.74%)	-

Figures

Figure 1

The selection process from iSNV to SNPs. (A) The mutations were initialized with random mutations. Considering the sequencing errors, the low allele frequencies of iSNVs (commonly <5%) were ignored in iSNVs analysis; and the high allele frequencies of iSNVs were considered as SNPs (>95%). (B) Two steps of mutation fitness selection: from low frequency random mutations to iSNVs, which occurs within host and affected by RNA-editing and host immunity; and from iSNVs to SNPs, which occurs both intra- and

inter- host process, including the fitness selection and purifying selection. (C) The observed distribution of genomic distances between two random mutations should follow Poisson distribution, whereas that of iSNVs and SNPs are away from the expected distribution. The number of observed mutations was also decreased. (D) The diagram of observed mutations at random mutation level, iSNV level and SNP level.

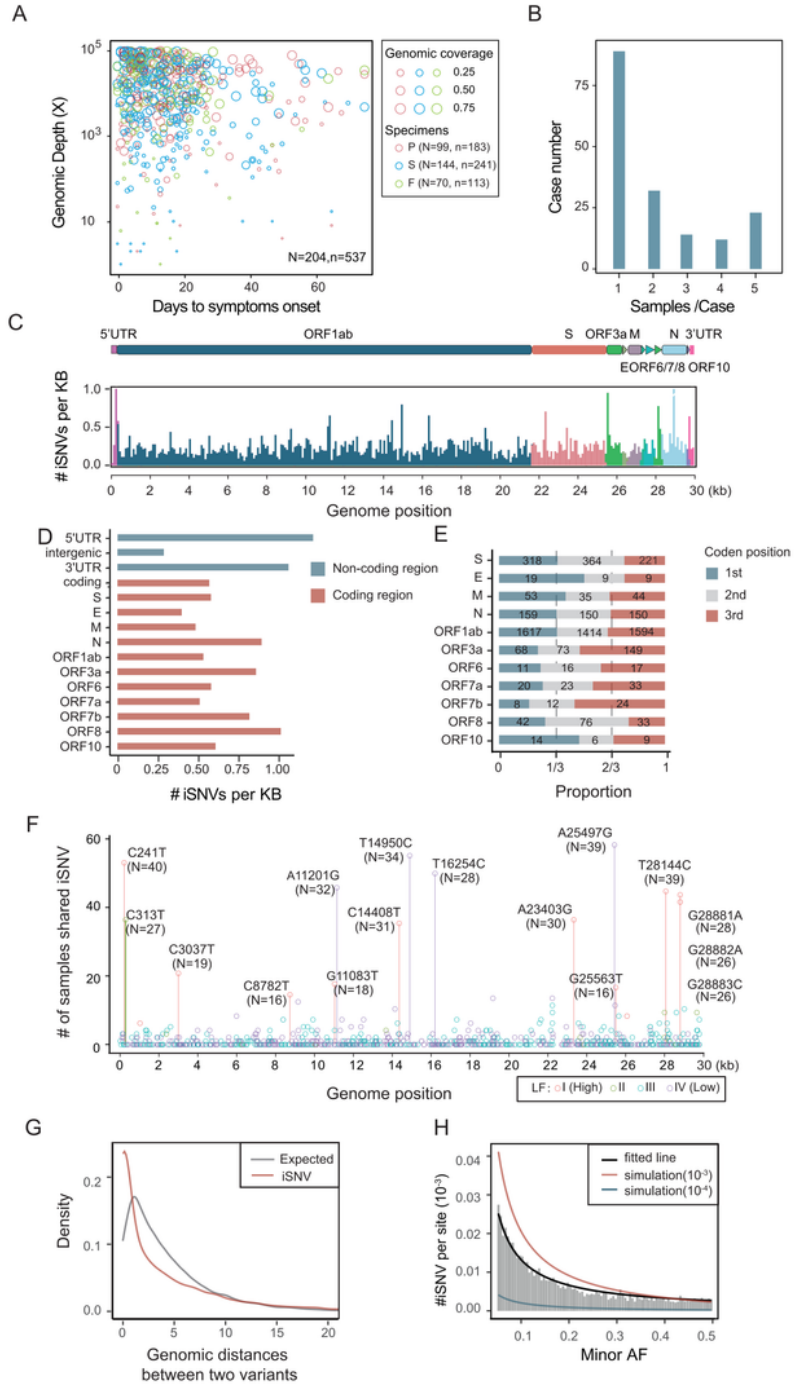

Figure 2

The spatio-temporal analysis on the genomic sequencing and iSNVs locus identified in this study. (A) The dot plot between collection time and sequencing depth of pharyngeal (red circle), sputum (blue circle) and fecal (green circle) samples sequenced. (B) The accumulative case number of case with single or multiple samples. (C) The distribution of iSNVs frequency along the genome counted by the window of 100bp. (D) The normalized iSNV number in coding (red) and non-coding regions (dark blue). (E) The proportion of iSNVs occurred in different position of the codon in each gene. The number of iSNVs in each category was marked on the corresponding bar. (F) The distribution of iSNVs locus. Compared iSNVs sites to the SNPs which were reported in NGDC database, the sites were marked according to the level of frequencies of SNPs occurred in the population (level from I to IV: red, green, blue and purple, of which level I was the highest frequency SNPs in public database, see Methods). (G) The distribution of genomic distance between two variants in expected and identified iSNVs. The expected curve follows Poisson distribution. (H) The density of iSNVs number along the iSNV minor allele frequency. The black line was fitted with generalized linear model using the minor allele frequency. The blue and red lines represent simulation curves with mutation rate 10-3 and 10-4 as previous described in influenza virus.

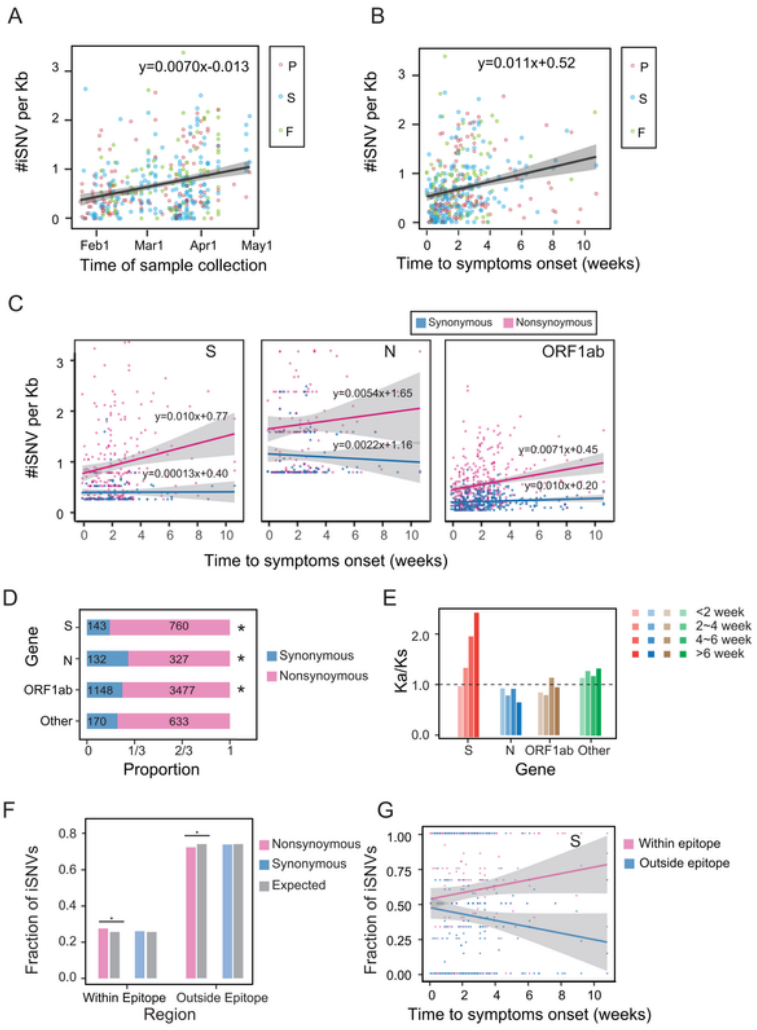

Figure 3

The spatio-temporal of iSNVs displayed an increase genetic diversity. (A) Normalized iSNV number against the absolute date of sample collection with a linear regression. (B) Normalized iSNV number against the time post symptom onset. (C) Normalized iSNV number causing nonsynonymous and synonymous mutations against the weeks post symptom onset in gene S, N and ORF1ab. (D) The proportion of iSNVs causing nonsynonymous and synonymous mutations in each gene. The number of

iSNVs in each category was marked on the corresponding bar. (E) The Ka/Ks ratio of gene S, N, Orf1ab and other genes in different symptom onset period. (F) The fraction of iSNV numbers in epitope regions. The expected value was calculated by the predict epitope region in whole protein. (G) The relative nonsynonymous count normalized by total iSNV epitope fraction in different gene region. (H) The relative nonsynonymous count normalized by total iSNV epitope fraction along with the time to symptom onset.

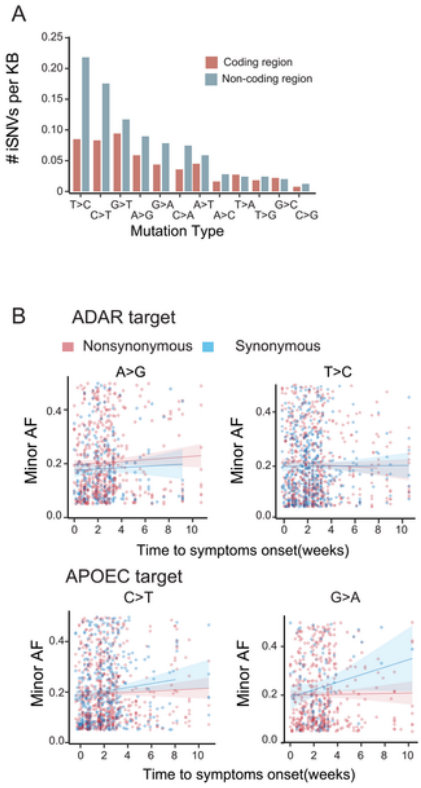

Figure 4

The iSNV distribution of different mutation type. (A) The normalized iSNV number of different mutation types. The mutations in coding (red) and non-coding (blue) regions were distinguished by color. (B) The nucleotide sequence context for the ADARs target (A -> I) and APOBECs target (C -> U). (C) The minor AF of ADARs target (A -> I, causing A -> G and T -> C) and APOBECs target (C -> U, causing C -> T and G -> A) against the days to symptoms onset of patients.

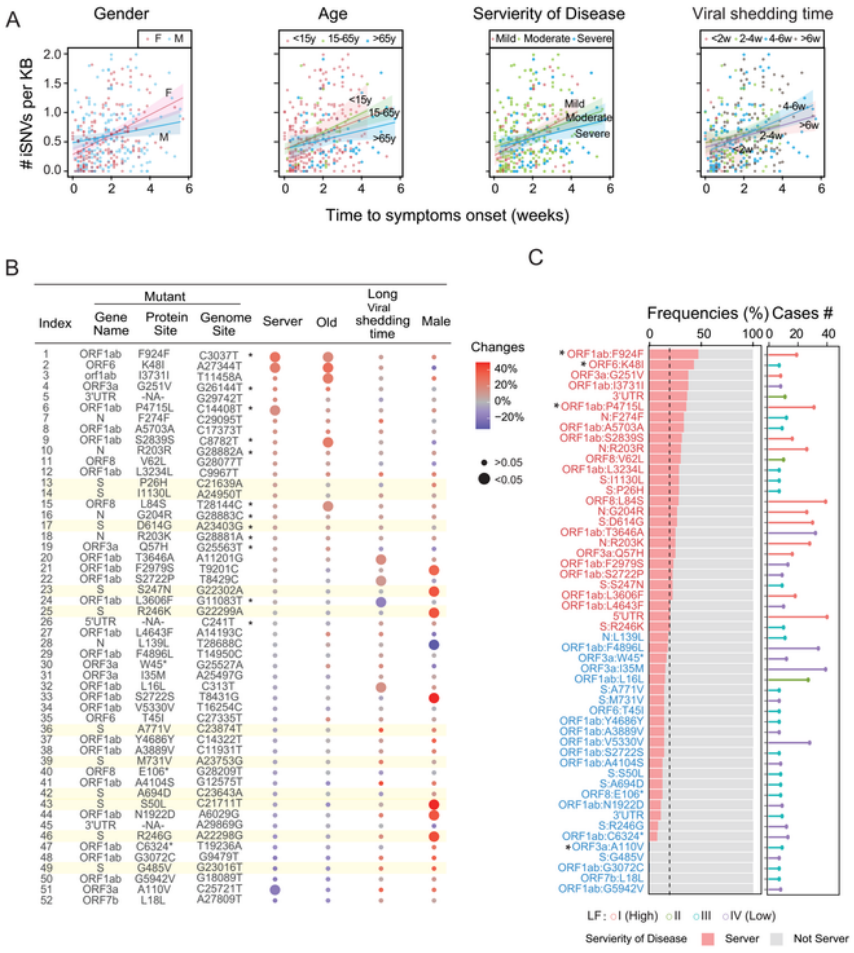

Figure 5

The distribution of iSNVs among different groups of patients. (A) The correlation of normalized iSNV number with the onset time of patients among patients grouped by gender, age, illness severity and viral shedding time. (B) The high frequency iSNVs in population impacted by gender, age, illness severity and viral shedding time. The size of the point represents the p-value of Fisher's exact test on the patients in population. * in genome site, represent the hfSNPs. (C) The histogram in the left shows the frequencies of iSNVs related to illness severity, and the iSNVs in red font are those related to severe outcomes. The iSNVs marked with star were those significantly differential distributed between server and non-server population. The mutations were distinguished by color (red: more severe patients observed, and grey: more non-severe patients observed). The dash line represents the average proportion of the severe patients. The plot on the right represents the number of cases with these iSNVs. The level of the frequencies of these iSNVs in public database were marked with different color of line (level I to IV: red, green, blue and purple).

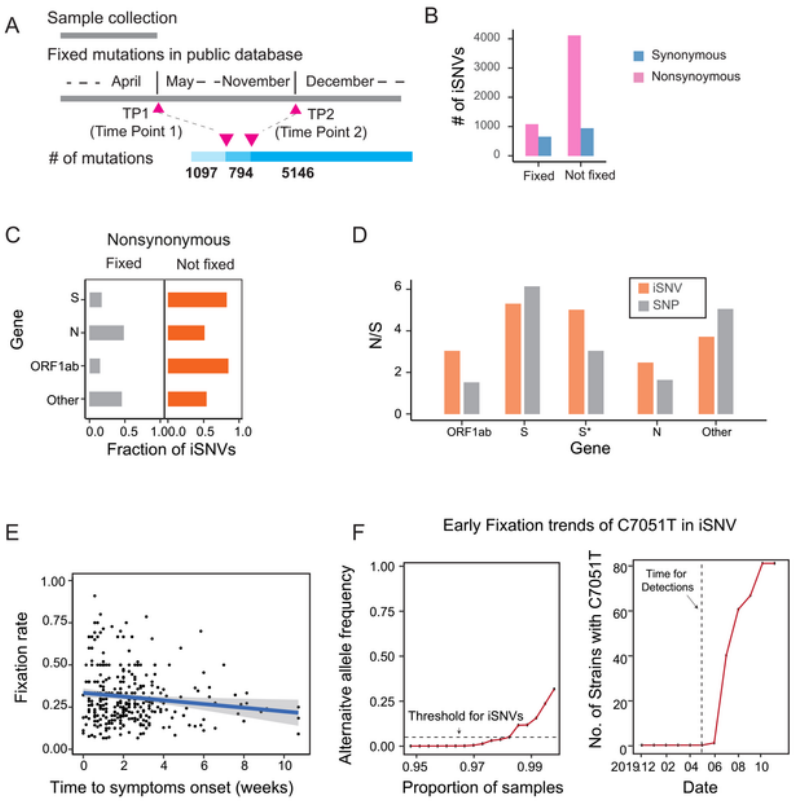

Figure 6

The biased fixation of iSNVs in public SNP database and local SNP dataset. (A) The number of fixed iSNVs in different time period. (B) The number of fixed synonymous and nonsynonymous iSNVs in fixed and not-fixed phase. (C) The fraction of fixed and not-fixed nonsynonymous mutation in each gene. (D) The Nonsynonymous/Synonymous ratio of identified iSNV and SNP in each gene. S* represents S gene

that excluding D614G mutation. (E) The fix rate of iSNVs along the time to symptoms onset. (F) An example of early fixation trends of C7051T in iSNVs in our dataset (left) and in public database (right).

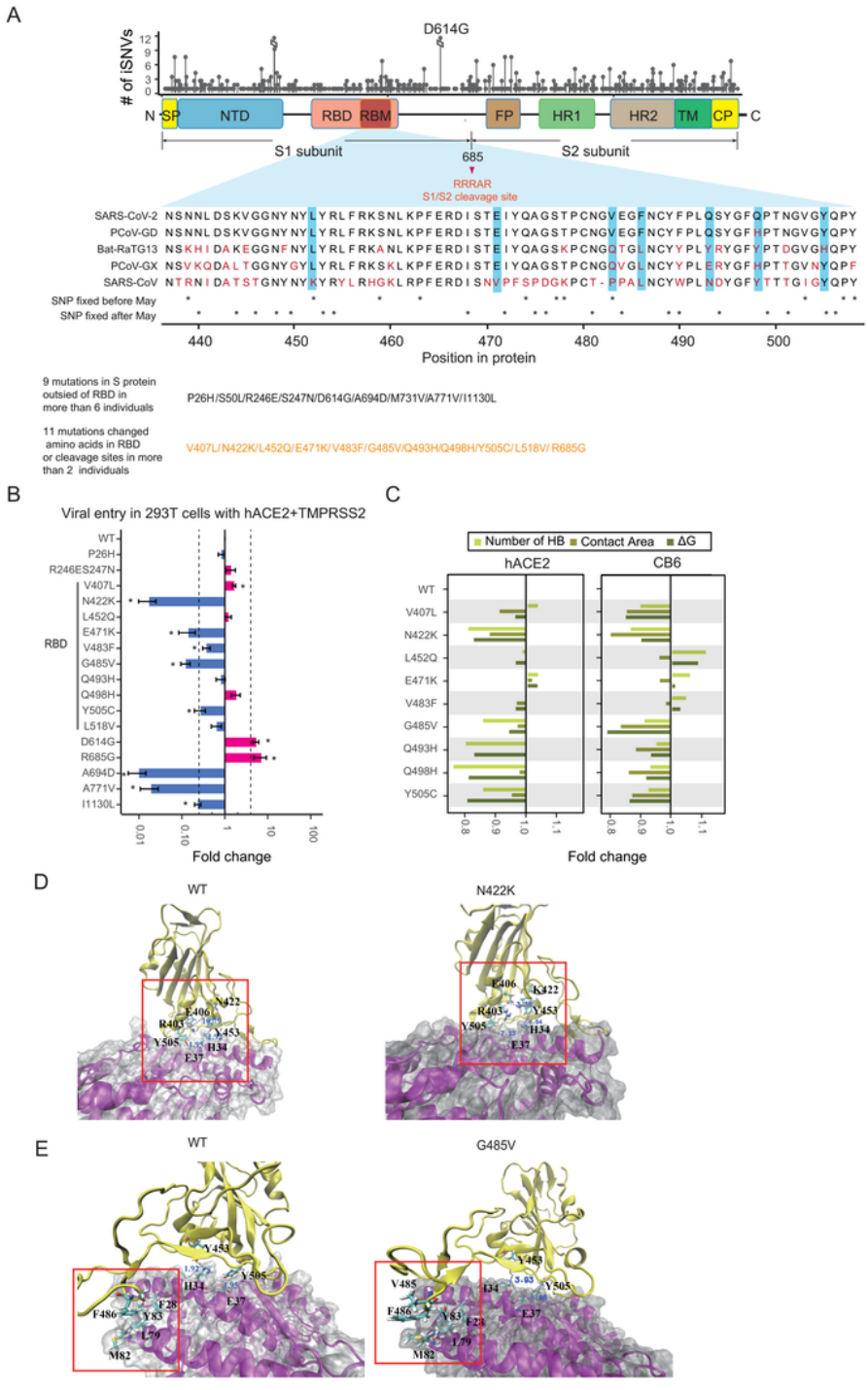

Figure 7

The genetic and molecular structure analysis on iSNVs observed in S protein. (A) On the top is the location of iSNVs and the number of samples with the iSNVs, and the RBD and RBM region are marked. Below are the mutations of amino acid residue in SARS-CoV-2, PCoV-GD, Bat-RaTG13, PCOV-GX and

SARS-CoV at the locations corresponding to the shared iSNVs in RBM region. SNPs appeared in public database were marked by stars. PCoV-GX: pangolin CoV isolate GX-PL4. PCoV-GD:pangolin CoV isolate MP789. The GenBank No. for these CoVs is: SARS-CoV-2 (isolate Wuhan-Hu-1, NC_045512.2), SARS-CoV (isolate Tor2, NC_004718.3), bat-RaTG13 (MN996532.1), PCoV-GX (isolate P4L, MT040333.1), PCoV-GD (isolate MP789, MT084071.1) (B) The fold change of viral entry in T Rex 293 hACE2 cells of different iSNV mutations. The dash line represent the fold-change of 0.25 and 4. The * signed the mutants that significantly changed viral entry efficacy by t-test.(C) The relatively fold change of number of hydrogen bond, contract area and change of binding energy in Molecular Dynamic calculation of the mutants compared to WT bound to hACE2 and CB6. (D) Crystal structures of SARS-CoV-2 RBD/hACE2 complex in WT and N422K mutation. The red square represents the effect regions. (E) Crystal structures of SARS-CoV-2 RBD/hACE2 complex in WT and G485V mutation. The red square represents the effect regions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials0104rm.pdf
- file2Supplementarymaterialsdb.docx
- equation1.docx