Two-step fitness selection for intra-host variations in SARS-CoV-2

Jiarui Li, Pengcheng Du, Lijiang Yang, Ju Zhang, Chuan Song, Danying Chen, Yangzi Song, Nan Ding, Mingxi Hua, Kai Han, Rui Song, Wen Xie, Zhihai Chen, Xianbo Wang, Jingyuan Liu, Yanli Xu, Guiju Gao, Qi Wang, Lin Pu, Lin Di, Jie Li, Jinglin Yue, Junyan Han, Xuesen Zhao, Yonghong Yan, Fengting Yu, Angela R. Wu, Fujie Zhang, Yi Qin Gao, Yanyi Huang, Jianbin Wang, Hui Zeng, Chen Chen

PII: S2211-1247(21)01709-5
DOI: https://doi.org/10.1016/j.celrep.2021.110205
Reference: CELREP 110205

To appear in: Cell Reports

Received Date:  8 June 2021
Revised Date:  10 October 2021
Accepted Date:  13 December 2021


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021
<table>
<thead>
<tr>
<th>Allele Frequency</th>
<th>Random mutations</th>
<th>iSNVs</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5%</td>
<td>5%-95%</td>
<td>&gt;95%</td>
<td></td>
</tr>
</tbody>
</table>

**Positive selection**

- Intra-host
- Inter-host
- RNA editing
- Host effects
- Negative selection

**Stochastic**

**Unstochastic**
Two-step fitness selection for intra-host variations in SARS-CoV-2

Jiarui Li¹,²#, Pengcheng Du¹,²#, Lijiang Yang³,⁴#, Ju Zhang¹,²#, Chuan Song¹,²#, Danying Chen¹,²#, Yangzi Song¹,², Nan Ding¹,², Mingxi Hua¹,², Kai Han¹,², Rui Song¹, Wen Xie¹, Zhihai Chen¹, Xianbo Wang¹, Jingyuan Liu¹, Yanli Xu¹, Guiju Gao¹, Qi Wang¹, Lin Pu¹, Lin Di⁴,⁵,⁶, Jie Li⁷, Jinglin Yue⁸, Junyan Han¹,², Xuesen Zhao¹,², Yonghong Yan¹,², Fengting Yu¹, Angela R. Wu⁹,¹⁰, Fujie Zhang¹*, Yi Qin Gao³,⁴,¹¹*, Yanyi Huang³,⁴,⁵,⁶,¹²#, Jianbin Wang⁷,¹²#, Hui Zeng¹³,¹⁴*, Chen Chen¹³,¹⁴*

1. Beijing Ditan Hospital, Capital Medical University, Beijing 100015, P. R. China
2. Beijing Key Laboratory of Emerging Infectious Diseases, Beijing 100015, P. R. China
3. Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China
4. Beijing Advanced Innovation Center for Genomics, Biomedical Pioneering Innovation Center, Peking University, Beijing 100871, China
5. School of Life Sciences, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China
6. Institute for Cell Analysis, Shenzhen Bay Laboratory, Shenzhen 518055, China
7. School of Life Sciences, Tsinghua-Peking Center for Life Sciences, Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing 100084, China
8. Peking University Ditan Teaching Hospital, Beijing 100015, China
9. Division of Life Science, Hong Kong University of Science and Technology, Hong Kong SAR, P. R. China
10. Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, P. R. China
11. Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China
12. Chinese Institute for Brain Research, Beijing 102206, China
13. Biomedical innovation center, Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, China

#These authors contributed equally to this work.

¹⁴ Lead Contact

*Corresponding authors:
Prof. Chen Chen,
Biomedical innovation center, Beijing Shijitan Hospital, Capital Medical University, No.8 Jingshundongjie, Beijing 100015, China
Email: chenchen1@ccmu.edu.cn

Prof. Hui Zeng,
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University and Beijing Key Laboratory of Emerging Infectious Diseases, No.8 Jingshundongjie, Beijing 100015, China
Email: zenghui@ccmu.edu.cn

Prof. Jianbin Wang
School of Life Sciences, Tsinghua-Peking Center for Life Sciences, Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing 100084, P. R. China
Email: jianbinwang@tsinghua.edu.cn

Prof. Yanyi Huang
Biomedical Pioneering Innovation Center, Beijing Advanced Innovation Center for Genomics, College of Chemistry and Molecular Engineering, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, P. R. China
Email: yanyi@pku.edu.cn

Prof. Yi Qin Gao
Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Beijing Advanced Innovation Center for Genomics, Biomedical Pioneering Innovation Center, Peking University, Beijing 100871, China.
Email: gaoyq@pku.edu.cn

Prof. Fujie Zhang
Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, No.8 Jingshundongjie, Beijing 100015, China
E-mail: treatment@chinaaids.cn
SUMMARY

Spontaneous mutations introduce uncertainty into COVID-19 control procedures and vaccine development. Here, we perform spatio-temporal analysis on intra-host single-nucleotide variations (iSNVs) in 402 clinical samples from 170 patients, which reveals an increase in genetic diversity over time post-symptoms onset within individual patients. Nonsynonymous mutations are over-represented in the pool of iSNVs, but underrepresent at the single nucleotide polymorphism (SNP) level, suggesting a two-step fitness selection process: a large number of nonsynonymous substitutions are generated within the host (positive selection), and these substitutions tend to be unfixed as SNPs in population (negative selection). Dynamic iSNVs changes in subpopulations of different gender, age, illness severity and viral shedding time displayed a varied fitness selection process among populations. Taken together, our study highlights iSNVs provide a mutational pool shaping the virus rapid global evolution.

Keywords: intra-host variation; SARS-CoV-2; positive selection; dynamic genetic divergences
INTRODUCTION

Despite the global emergence of various innovative prevention and control responses, SARS-CoV-2 continues to spread rapidly around the world (Coronaviridae Study Group of the International Committee on Taxonomy of, 2020). Like other single-stranded RNA viruses, SARS-CoV-2 has a high mutation rate, and unexpected mutations may change the virus and make it harder to control, leading to reduced vaccine efficacy (Li et al., 2020). Researchers have identified numerous mutations in SARS-CoV-2 based on more than 5,000,000 published genomes (Shu and McCauley, 2017), and these mutations result in a rapid increase in PANGOLIN (phylogenetic assignment of named global outbreak lineages) (Rambaut et al., 2020), including B.1.1.7 (Alpha), B.1.617.2 (Delta) and more recently B.1.1.529 (Omicron) (Jhun et al., 2021; Gao et al., 2021). Therefore, tracking SARS-CoV-2 mutations in lineages has been a global monitoring priority.

Viral mutations randomly initiate in a small fraction of viruses within a single infected host, and generate intra-host single-nucleotide variations (iSNVs). Some iSNVs may be subsequently fixed within hosts and transmitted among populations, or transmitted as an unfixed form, and finally generate genetically diverse populations. As the virus spreads, the iSNVs are fixed through both a stochastic process (e.g. genetic drift) (McCrone et al., 2018) and a deterministic process (e.g. fitness selection) (Forni et al., 2016). With high-throughput sequencing, we can explore this process and measure genetic diversity at the population and within-host levels.

Previous population-level studies using clinical, molecular, and immunological data related to single nucleotide polymorphisms (SNPs) have provided significant insight into epidemiology and disease transmission, as well as aspects of pathogenesis (Gudbjartsson et al., 2020; Tang et al., 2020).

As a larger genetic mutation pool compared to SNPs, iSNVs provide additional information defining the diversity and dynamics of viral evolution within individual...
hosts (Holmes et al., 2016). Analysis of iSNVs complements conventional population-
level SNP studies and facilitates a more comprehensive understanding of viral
evolution, which aids clinically relevant predictions of viral evolution associated with
infection, pandemics, and immune evasion (McCrone et al., 2018). However, few
studies have explored the genetic characteristics of iSNVs in patients with COVID-19
(Lythgoe et al., 2021).

In the present study, we quantitatively assessed the genetic diversity of SARS-CoV-
2 and viral evolution within individual hosts using deep viral genome sequencing
(Chen et al., 2020) and empirical analysis pipelines (Ni et al., 2016). Spatio-temporal
analysis of genomic data revealed that within-host variation was not randomly
distributed throughout the genome, and such variation increased the viral genetic
diversity of SARS-CoV-2, suggesting a role for selection. Non-synonymous mutations
were over-represented in the pool of iSNVs (Mutation allele frequency (MuAF): >5%
& <95%) but underrepresented in the set of SNPs (MuAF: >95%), indicating a two-
step fitness selection process. We also explored the effects of iSNVs (unfixed
mutation) on clinical characteristics and the binding affinity of the spike (S) protein,
which might account for the observed directional selection process. Our results
suggest that it is important to simultaneously study the within- and between-host
dynamics of emerging viruses to understand their evolutionary histories and thus
direct efforts at developing effective methods of prevention and control.
RESULTS

Identification of intra-host variations in SARS-CoV-2 genomes

We collected 537 (183 pharyngeal, 241 sputum and 113 faecal) samples from 204 patients, covering 34.4% of total cases diagnosed in Beijing before April 30, 2020 (Figure 1A, Table S1). Using targeted capture viral genome sequencing, we obtained 8.59G (inter-quartile range [IQR]: 3.12G-17.38G) base pairs 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) (Wu et al., 2020). Samples with low viral genomic coverage were removed (see Methods), and eventually 402 (136 pharyngeal swab, 182 sputum and 84 faecal) samples from 170 patients were selected for further analysis (Figure S1A). The sequencing depth was 28,720x across the genome (IQR: 4,815-46,343, Figure S1B). Among these patients, 81 had at least two samples collected from different sample types or time points (Figure 1B).

For each sample, we performed a high-depth (100x as the minimum depth) search of SARS-CoV-2 genomic sites for iSNVs, which were filtered using a stringent threshold (≥5%) that sufficiently distinguished true iSNVs from sequencing errors (see Methods and Supplementary). We validated the sample processing pipeline with technical replicates in 62 samples with independent library preparation. Under the 5% cut-off for mutation allele frequency (MuAF), we identified 450 reproducible iSNVs out of 498 iSNVs in the first experiments. The iSNVs meeting this stringent threshold were widely distributed throughout the genome, and the number of iSNVs in each sample was not affected by genomic coverage, sequencing depth ($R^2$ for iSNV number and genomic coverage, 0.074; $R^2$ for iSNV number and sequencing depth, 0.006) (Figure S1C). In total, we identified 7,037 iSNVs in 374 samples with a median density of 0.53 iSNVs/kb, which is comparable to the number of iSNVs previously reported for SARS-CoV-2 (Lythgoe et al., 2021) and in other virus, such as Ebola virus (Ni et al., 2016), yellow fever virus (Chen et al., 2018) and influenza A
(Debbink et al., 2017) (Table S2). About 93% of samples (374/402) harboured at least one iSNV in comparison with the reference genome (Table S3).

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

We examined the locations of iSNVs along the SARS-CoV-2 genome and found an overall relatively low density of iSNVs (0.58 iSNVs/kb), which is comparable to previous report (Lythgoe et al., 2021; Popa et al., 2020). Higher iSNV density was observed in the 5’-UTR (1.23 iSNVs/kb) and 3’-UTR (1.07 iSNVs/kb) (Figure 1C). We found 6,790 (96.49%) iSNVs in coding regions, which account for 97.85% of the whole genome (29261/29903). Most iSNVs (4,625, 68.11%) were identified in ORF1ab, followed by the S gene (903 iSNVs) and N gene (459 iSNVs) (Figure S1D). However, after we normalized iSNVs for gene length, the highest frequency of iSNVs was found in ORF8 (1.02 iSNVs/kb), followed by the N gene (0.906 iSNVs/kb) (Figure 1D). These results were consistent with a previous study of SARS-CoV-2 at the SNP level (Zhang et al., 2020). Thus, the imbalance of mutations among genes might occur at iSNV level, and be maintained in the fixation process. In addition, analysis of codon position revealed 2,329, 2,178 and 2,283 iSNVs at the 1st, 2nd and 3rd codon positions, respectively (Figure 1E). Fisher’s exact test revealed that ORF10 and the E gene had a significantly greater number of iSNVs at the 1st codon position in comparison with the other codon positions in coordinate genes (Figure 1E).

We next examined the distribution of iSNVs among the patients. Among the 4690 iSNV sites, 81.02% were only observed in one patient, and 18.98% were found in at least two individuals, which is comparable to the previous report (Lythgoe et al., 2021). There were 16 highly recurrent iSNVs, which shared in at least 15 patients. Among them, 12 sites were found to overlap with high frequency SNPs (hfSNPs) sites, which had been previously defined in the global SARS-CoV-2 genome database 2019nCoVR, based on more than 5% of patients (Song et al., 2020) (Figure 1F). This phenomenon was also described in a previous report (Tonkin-Hill et al., 2021), and
may be due to convergent positive selection or mutational hotspots. We constructed a simple framework to calculate the distribution of genomic distance between pairs of alleles of iSNVs, and the fitted density line shows significant difference compared to the stochastically generated mutations (Figure 1G, Kolmogorov-Smirnov test, \( p\)-value<0.001), suggesting non-stochasticity of iSNV distribution. Moreover, based on a modification of the stochastic birth-death model (Bozic et al., 2016), we predicted allele frequencies to describe how mutations accumulate as cells expand. Although the distribution was similar to the neutral model (Figure 1H), iSNV numbers in the high frequency alleles were greater than expected values in the neutral evolution model, suggesting a potential positive selection of iSNV sites (Figure S1E).

**Genetic diversity increases as the disease progresses**

To uncover dynamic changes in viral iSNVs in patients with COVID-19, we performed spatio-temporal analysis of iSNVs along the epidemic period and disease progression using different specimens. First, we observed a steady increase in the number of iSNVs over time during the epidemic (estimated value from 0.15 to 0.83 iSNVs/kb within 97 days) (Figure 2A). Similar increases in iSNV numbers were observed in the other viral genomes that cause acute and chronic infections that we examined (YFV, ZIKV, HIV and HCV) (Chen et al., 2018; Oliveira et al., 2018; Barton et al., 2016; Skums et al., 2015). Additionally, iSNVs also accumulated during the infection in individual hosts, from 0.52 iSNVs/kb on day 1 to 0.85 iSNVs/kb 30 days post-symptoms onset (Figure 2B, Table S3). The accumulation was still detectable even after the normalization with viral RNA Ct values.

However, spatial analysis of the three types of samples, representing different body sites/locations, showed increased genetic diversity in all three specimens along disease progression (Figure S2A). We identified more iSNV number on the initial day and lower accumulation rate along the disease process in faecal samples, comparing to
those of the pharyngeal swab and sputum samples (Figure S2A, p-value=0.006, ANOVA test).

Positive selection process for iSNVs

To test whether the observed accumulation of genetic diversity was caused by fitness selection, we explored the dynamic change of nonsynonymous and synonymous mutations with the disease progression in the S, N, ORF1ab and other genes. In comparison with neutral synonymous mutations, accumulation of mutations in nonsynonymous regions was more rapid in all genes, and the S gene displayed the highest accumulation rate (Figure 2C). Along the genome, we found 5,197 iSNVs were nonsynonymous mutations, whereas only 1,593 iSNVs were synonymous mutations. The ratio of nonsynonymous to synonymous variants in all patients was 3.26 (mean ratio of 3.16 observed per patient). The ratio of nonsynonymous to synonymous iSNVs diverges between genes: the ratio in the S gene (ratio = 5.31) was significantly higher than that of the other parts of the viral genome (Figure 2D). The mean values of the minor allele frequencies of nonsynonymous and synonymous iSNVs were 0.189 and 0.195, respectively (Figure S2B). With a simple substitution model, we used the ratio of Ka/Ks to measure if the genes in SARS-CoV-2 genome were under selection pressure. The ratio of Ka/Ks in the S gene increased from 1.01 to 2.46 as the disease progressed, indicating that positive selection occurred with the disease progression, at least for the S gene (Figure 2E).

In addition, we use two data sets from IEDB (Dhanda et al., 2019) (experimentally confirmed and predicted epitope regions) (Shrock et al., 2020) to evaluate the fraction of nonsynonymous/synonymous mutations within and outside epitope regions. The fraction of nonsynonymous mutations in the epitope regions was significantly higher than expected (27.57% vs. 25.74%, p-value = 0.016, Fisher exact test, Figure S2C). Accordingly, nonsynonymous mutations outside of epitope regions were significantly less frequent (Figure S2C). Further correlation analysis between the fraction of
nonsynonymous sites in predicted epitope regions and the time of symptoms onset revealed that the number of nonsynonymous sites in the S gene increased as the disease progressed (*Figure S2D*).

Last, we investigated the dynamics of intra-host evolution in 268 samples from 61 longitudinally sampled patients. The interval durations between the first and the last samples collected were more than 5 days. None of the patients were subject to antibodies and immunosuppressant treatment. Although the mutation patterns over time were varied across individual patients, most patients (45/61) showed an increased mutational diversity (*Figure S2E*). We estimated the accumulation rate for each individual using a linear model of iSNV number and post-symptom onset time. Consistent with rapidly accumulated iSNVs in nonsynonymous sites, 84.44% of the patients (38/45) displayed higher accumulation rates in nonsynonymous sites than synonymous sites (*Figure S2F*). A small number of stable iSNV sites (81/3629) appeared recurrently across time points. In addition, we identified 255 recurrent iSNVs in different time points. Among them, 143 iSNVs showed increased allele frequencies, and the remaining 112 iSNVs had decreased allele frequencies at the latter time points (*Figure S2G*), providing stronger evidence of potential positive selection within-host.

**RNA editing in regions of increased genetic diversity** RNA-editing enzymes can mutagenize single-stranded RNA and DNA molecules and provide a defence mechanism against virus. Two families of RNA-editing enzyme have been demonstrated to contribute to the mutational spectrum of SARS-CoV-2 (Di Giorgio et al., 2020). The APOBEC deaminase (apolipo protein B mRNA editing catalytic polypeptide-like deaminase) deaminates cytosines into uracils (C-to-U, including C-to-U and G-to-A), while ADAR deaminase (RNA-specific adenosine deaminase) deaminates adenines into inosines (A-to-I, including A-to-G and U-to-C) (Di Giorgio et al., 2020).
We measured all mutational types for all iSNVs, and the top five iSNV mutation types were ranked as follows (most to least common): U-to-C, C-to-U, G-to-U, A-to-G, and G-to-A (*Figure 3A*); four of them might be introduced by APOBEC and ADAR. Unlike the A-to-I RNA editing signal in the human transcriptome, we did not observe obvious depletion of G bases in position -1 (*Figure S3A*). To evaluate the dynamic change of RNA-editing levels along with disease progress, we calculated the correlations between the minor allele frequencies of iSNVs and the time after symptoms onset. Consistently, the minor allele frequencies of C-to-U and G-to-A mutations due to APOBEC-mediated RNA editing were slightly increased in vivo. In contrast, the frequencies of other mutational types, including ADAR-mediated RNA editing in vivo, did not increase with the duration of infection. This identification was consistent with previous studies which found that coronavirus infection induced APOBEC activity, but not ADAR activity. We next compared the mutational types and their levels of accumulation for both nonsynonymous and synonymous mutations. All four substitutions mediated by APOBECs/ADARs were more abundant in nonsynonymous mutations (*Figure S3B*). In addition, synonymous mutations accumulated more rapidly in comparison with nonsynonymous mutations (*Figure 3B*). These results suggested that RNA editing mediated by APOBECs was also impacted by SARS-CoV-2 infection, especially the rates of C-to-U and G-to-A mutations.

**Influence of possible host effects on genetic diversity**

Exacerbated inflammatory response has been observed in severe and critical patients (Hadjadj et al., 2020), and different patterns of immunity have been reported in patients of different gender (Takahashi et al., 2020) and age (Zheng et al., 2020) groups. Therefore, we evaluated the influence of host effects on viral mutation, and measured dynamic changes in the number of iSNVs within groups of patients based on gender, age, illness severity, and viral shedding time (*Table 1*). Each sample was
recalibrated based on symptoms onset date. Increased genetic diversity of iSNVs was observed in all groups (Figure 4A), suggesting that accumulation of iSNVs occurred in all populations rather than a specific population. We also observed different slopes and initial values in the fitness linear model between iSNVs number and time post-symptoms onset in these patient groups (Figure 4B). A higher accumulation rate was observed in middle-aged patients (15–65 year) compared to elderly patients ($p$-value =0.037, ANOVA test), especially in patients whose viral shedding time was more than 6 weeks ($p$-value =0.041, ANOVA test, Figure S4A). We also observed a significant increase in nonsynonymous sites in mid-age compare to elderly patients ($p$-value =0.035, ANOVA test), while the synonymous accumulation rate did not differ among different age groups (Figure S4B, Figure S4C). The different iSNV accumulation rate suggests the presence of a different fitness selection process during the initial infection stage and subsequent infection stages after symptoms onset (Figure 4A).

To further investigate specific mutations that could be influenced by host effects we compared the proportions of patients with and without recurrently mutated sites (Figure 4C). We constructed a matrix of patients with/without iSNVs based on 52 iSNV sites shared by more than six patients among a group of 170 patients. These patients were categorized into four independent classes: gender, age, illness severity and viral shedding time (Table 1). Each iSNV site was subjected to an independent Fisher’s exact test. Patients with severe disease, elderly patients, patients with long viral shedding time, and male patients preferentially showed significant enrichment in 4, 5, 4, and 8 iSNVs, respectively, compared with patients with mild/moderate disease, children/middle-aged patients, patients with short viral shedding time (<6 weeks), and female patients (Figure 4C). These iSNVs were distributed in ORF1ab, S, N, ORF6 and ORF8. Among the 52 recurrent iSNV sites, we identified 27 sites that preferentially occurred in patients with severe disease, of which 12 overlapped with previously reported hfSNP sites in the public database 2019nCoVR. In contrast, the
25 iSNVs that preferentially occurred in patients with moderate disease did not overlap with hfSNPs sites (Figure 4D, \(p\)-value < 0.001, Fisher’s exact test). This enrichment of hfSNPs sites was not observed in any other categories that were stratified based on gender, age, and viral shedding time, indicating a non-stochastic process.

Uneven purifying selection processes from iSNVs to SNPs

To identify the mutations fixed from iSNVs to SNPs, we compared the genomic sites of iSNVs and SNPs in the 2019nCoVR database (Song et al., 2020; Zhao et al., 2020a). Among 7,037 iSNVs, 15.59% of iSNVs had already been identified as SNPs before our observation period (May, 2020), and 11.28% of iSNVs were fixed from May, 2020 to December, 2020, while the remaining iSNVs (73.13%) were not fixed as SNPs (Figure 5A). Nonsynonymous iSNVs displayed a lower fixation rate in comparison with that of synonymous iSNVs (20.92% vs. 41.12%, \(p\)-value < 0.001, Fisher’s exact test; Figure 5B). This finding is supported by a model in which nonsynonymous iSNVs occur at a high frequency within an individual due to positive selection or incomplete purifying selection, but are less likely to become fixed in the population due to purifying selection. Next, we performed Fisher’s exact test to compare the proportion of fixed mutations in each gene, and S and ORF1ab were found to have lower fixation rates (21.04% and 20.56%, respectively, Figure 5C) at nonsynonymous and synonymous sites (Figure S5A). Consistently, the nonsynonymous-to-synonymous ratios of iSNVs in ORF1ab, N and S (excluding D614G) were greater than those estimated for the identified SNPs, consistent with uneven purifying selection of these genes (Figure 5D). With disease progression, iSNV fixation rates in nonsynonymous and synonymous sites were stable in the population (Figure 5E, Figure S5B), indicating a process of similar purifying selection as the disease progressed. We also observed that mutation might have occurred in iSNVs before they were fixed as SNPs. For example, accumulation of C7051T alleles was observed in our study before May, whereas the first C7051T SNP
was not reported until June, 2020 (Figure 5F). Although sampling bias might have
also limited our observation of SNPs in the early stages of the epidemic period, our
observation of mutations at the iSNV level might have increased our detection
sensitivity for mutations before they were fixed. These results indicate that iSNVs
provide a complementary resource of genetic information to illuminate the
evolutionary history of SARS-CoV-2.

Molecular functions of S protein variations before purifying selection

The S protein drives cellular binding and entry through receptors and acts as a major
determinant of the host range, cell type, tissue tropism, and pathogenesis of
coronaviruses (Li, 2013). Therefore, we analysed 21 nonsynonymous sites of 606
iSNV sites identified in the coding region of the S protein, which caused 20 amino
acids changes: (1) nine were detected outside the receptor-binding domain (RBD)
region in more than six individuals, including substitution of two amino acids change
in three linked iSNVs (R246E and S247N caused by A22298G, G22299A and
G22302A, Figure 6A) ; (2) eleven resided within the RBD or S1/S2 cleavage sites in
more than two individuals, including seven iSNVs located in the receptor-binding
motif (RBM). Since few of these mutations had been reported in SARS-CoV-2, we
compared the mutation sites of seven iSNVs in the RBM of SARS-CoV-2 with the
consensus sequences of SARS-CoV-2-like coronaviruses in other animals (bat and
pangolin) to explore their potential molecular functions. All of these sites were
heterogeneous, suggesting that mutations in these protein sites may not be random
mutations (Figure 6A). In addition, patients with these mutations had no contact
history except for two patients with the G485V mutation (Table S4). No iSNVs had
emerged in patients at the first time point of genome sequencing (Figure S6A, Table
S4), and no evidence indicated that these iSNVs were linked in the genome. Thus,
these results indicate that iSNVs in the RBD seem to be generated by independent
viral evolution.
To elucidate the effects of these mutations at the molecular level, we first used the SARS-CoV-2 pseudovirus infection assay in HEK293T cells to assess the viral entry efficacy of 20 of the 22 S protein mutants, excluding S50L and M731V, which we failed to obtain in the mutation assay. Compared with the reference strain, 18 of the 20 tested mutants displayed a decreased (fold change <0.25, p-value < 0.05, t-test) or comparable (fold change <4 or >0.25) viral entry efficacy; only mutants R685G and D614G exhibited a similar level of increased viral entry efficacy (Figure 6B, Table S5, fold change >4, p-value < 0.05, t-test). Since the L518V mutation is far away from the binding interface between the S protein and hACE2/CB6, and four mutants (N422K, E471K, G485V and Y505C) displayed a significant decrease (fold change <0.25) in viral entry efficacy, the other five RBD mutants (V407L, L452Q, V483F, Q493H and Q498H) were tested for sensitivity to CB6 (Table S6). Wild-type (WT) and D614G viruses were included as controls. Modest differences between these RBD variants and the reference strain (≤4-fold) were observed with regard to susceptibility to CB6 (Table S5). Some variants, including V483F, Q493H and Q498H, were even more sensitive than the reference strain to CB6, which indicated that CB6 antibodies could block viral entry despite the RBD mutations.

Next, we focused on mutations within the RBM by simulating the binding of the corresponding mutants to human angiotensin-converting enzyme 2 (hACE2) (Wang et al., 2020) and CB6 neutralizing antibodies (Shi et al., 2020) using molecular dynamics simulations (Figure 6C). Simulation of the L518V mutation was not conducted for reasons described above. For comparison, we also simulated the binding of the WT RBD to hACE2 and CB6. The Cα root-mean-square deviation (Cα RMSD) values of the complexes of different mutant RBDs bound to hACE2 varied within a range close to the Cα RMSD value of the corresponding complex with WT RBD, suggesting that the nine mutations did not induce dramatic conformational changes (Figure S6B).
Further exploration of structural characteristics revealed that different mutations affect the binding of the RBD to hACE2 in various ways. For example, residue 422 was mutated to change its uncharged N side chain to a positively charged K, producing the N422K mutant. Consequently, a much stronger hydrogen bond (salt bridge) was formed between K422 and E406. In addition, strong repulsion between K422 and R403, as well as attraction between K422 and Y453, caused breakage of the hydrogen bonds between Y505 and E37, as well as those between Y453 and H34. As a result, both the number of hydrogen bonds and the contact area formed between the mutant RBDs and hACE2 were reduced (Figure 6D). Regarding the G485V mutation, V has a relatively bulky side chain compared to G; in order to reposition this bulky side chain, the G485V mutation led to the expulsion of F486 from the hydrophobic pocket formed by residues including F28, L79, M82 and Y83 of hACE2. As a result, both the contact area and the number of hydrogen bonds formed between the mutant RBD and hACE2 were reduced. Importantly, the hydrogen bond formed between Y505 of the S protein and E37 of hACE2, which is important for binding, was lost (Figure 6E). As a possible consequence, the binding of mutants N422K/G485V to hACE2 was weakened. Binding free energy calculations using the MM/GBSA method also indicated weakened binding of N422K/G485 to hACE2 (Figure 6C).

Changes in hydrogen bonding and contact area were also investigated for other mutants (Figure S6B). Most of the mutants with iSNVs displayed decreased viral entry ability due to fewer hydrogen bonds and/or reduced contact area between mutant RBDs and hACE2. We also analysed simulation results for complexes between mutant RBDs and the CB6 antibody (Figure S6C). Most of the mutants exhibited greatly reduced contact area compared with that of the WT sequence. Therefore, decreased binding affinity was observed due to fewer hydrogen bonds, consistent with greater binding free energy in the MM/GBSA calculations. Therefore, both the observation of mutations in the pseudoviral infection assays and computation...
of their interaction energies indicated weakened viral infection for most iSNVs identified in the S protein.

**DISCUSSION**

The ongoing SARS-CoV-2 pandemic is of worldwide concern. SARS-CoV-2 mutations arise naturally as the virus replicates, and the resulting SNPs may impact selection and transmission. Within a year of confirmation of the first case of COVID-19, thousands of mutations involving SNPs had been identified, among which only a very small minority caused changes in SARS-CoV-2 infectivity and immune evasion. Mutations in the SARS-CoV-2 genome can be used to explore signs of selection that accumulate during viral evolution (Sanjuan and Domingo-Calap, 2016; Xue and Bloom, 2020). In the present study, we demonstrated that intra-host variation in SARS-CoV-2 in a set of more than 400 clinical samples was not randomly distributed throughout the genome, suggesting a role for selection. Compared with synonymous mutations, non-synonymous mutations were over-represented in iSNVs, but under-represented in SNPs (population level). Molecular functional analysis of the effects of mutations in the S protein, as well as correlational analysis between mutations and clinical characteristics, suggested that mutations increasing disease severity or benefitting immune escape were rare, which could explain the low number of nonsynonymous mutations in populations.

Our sequencing approach provided evidence of two-step fitness selection for intra-host variations in SARS-CoV-2. The first step of selection occurs after randomised mutations are generated, and positive selection (e.g. patient-derived fitness selection) mediates accumulation of nonsynonymous iSNVs, indicating that genetic diversity increases with the disease progression of COVID-19. Positive selection results in two characteristics of iSNVs: (i) an increase in genetic diversity along with the COVID-19 disease progression. Both the RNA-editing and host immunity may affect this process. Our results show that the rate of iSNV accumulation is significantly higher in middle-
aged individuals than in the elderly. Evidence exists that elderly patients are under higher immune pressure (Zheng et al., 2020), which may be associated with the accumulation rate of iSNVs. (ii) An uneven distribution of iSNVs among patients and genomes. The recurrent iSNVs in distinct individuals also imply that these sites were under selection pressure, including positive selection and incomplete purifying selection. The tendency of more iSNVs with allele frequency increase in one individual indicates that these sites were under positive selection. Moreover, these mutations preferentially accumulate in nonsynonymous mutation sites, which might affect key features of the virus, including infectivity, virulence, and immunogenicity. There are several other studies on other virus, such as Ebola virus (Ladner et al., 2015), Lassa virus (Andersen et al., 2015), and influenza virus (Illingworth et al., 2014), support the existence of intra-host positive selection. High rates of mutation accumulation over short time periods in SARS-CoV-2 have been reported in previous studies of immunodeficient or immunosuppressed patients chronically infected with SARS-CoV-2 (Avanzato et al., 2020). In addition, several recent publications on the SARS-CoV-2 genome identified signals of positive selection (Velazquez-Salinas et al., 2020) and conservation within the gene encoding the S protein.

The second step in the selection process is purifying selection, which accounts for the reduction in the number of nonsynonymous mutations in the transition from iSNVs to SNPs. There are three implications of this: i) Frequency-dependent interactions and clonal interference may act as important forces driving the selection process; ii) Whether the progeny virus can establish productive infections may affect the spreading of mutations in a population (Lythgoe et al., 2021); iii) The transmission bottleneck is the determining factor for productive infections, both among hosts and between cells in a single individual. Since patients sampled in Beijing were mostly imported and/or sporadic cases, it was difficult to measure purifying selection from iSNVs to SNPs in the present study. These current findings provide insight into the
emergence of mutations worldwide, as well as direction for efforts aimed at controlling the global COVID-19 pandemic.

As nonsynonymous mutations accumulated in individuals, we also observed nonsynonymous mutations in the S protein. Although both the in vitro experiment and molecular dynamic simulation proved that the viral entry efficacy of high-frequency iSNV mutations tend to be weakened, it remains unclear whether these weakened viruses persist for lengthy periods of time in patients with long COVID-19 infections. However, it should be noted that mutations in populations are expected under the current situation. Changes in circumstances, such as widespread adoption of antibody therapy, will alter the selection pressure on each mutation (Zhou et al., 2020). Newly emerging mutations, such as N501Y, E484K, and L452R, show their potential immune escape from the pressure of monoclonal antibodies and vaccines (Li et al., 2021; Motozono et al., 2021). Strong selection pressure arising from antibodies and vaccines might lead to rapid remodelling of virus genetics through direct selection or genetic drift. Therefore, more mutations of SARS-CoV-2 will likely be observed as iSNVs, along with the R685G mutation in cleavage sites, which increases viral entry efficiency, at least in some cell types. These mutations may enhance the persistence of SARS-CoV-2 in multiple organs (Plante et al., 2020), although such SNPs have not yet been identified in public databases.

Given the urgency of vaccine development and treatment strategies, it may not be advisable to wait for mutations to be fixed in populations before efforts are undertaken to demonstrate their effects (Thanh Le et al., 2020). Importantly, early knowledge of potential evolution could guide vaccine design (Li et al., 2020; Poh et al., 2020). Associations between emerging mutations and illness severity as well as treatments should be carefully considered. More genomic data at the intra-host level should be gathered to allow investigation of genetic selection at the whole-genome
level, including the potential effects on illness severity, clinical outcomes, and the susceptibility of different populations.

519 Limitation of this study

Our study described the dynamic change of iSNVs in patients. However, limited by the number of patients, we did not observe a significant difference in distinct patient groups. The relationship between immune pressure and iSNV dynamics will be an interesting and important future study.
Figures

Figure 1. Spatio-temporal analysis of genomic sequencing and iSNV loci identified in this study. (A) Dot plot of the collection time and average sequencing depth of sequenced pharyngeal (red circles), sputum (blue circles) and faecal (green circle) samples. (B) The accumulative number of cases for single or multiple samples. (C) Distribution of iSNV frequency in the genome counted using a window of 100 bp. (D) Normalised number of iSNVs in coding (red) and non-coding (dark blue) regions. (E) Proportion of iSNVs occurring in different positions of codons in each gene. The number of iSNVs in each category is marked on the corresponding bar. (F) Distribution of iSNV loci. The iSNV sites were compared with SNPs reported in the 2019nCoVR database, and sites are marked according to the level of frequency of SNPs occurring in the population (levels I, II, III and IV coloured red, green, blue and purple, respectively; level I represents the SNPs in public databases with the highest frequencies; see Methods). The y axis represent the number of samples and N represent the number of patients. (G) Distribution of genomic distances between two variants in expected and identified iSNVs. The red line represents the real data of iSNVs and the grey line represent the expected curve, which follows the Poisson distribution. (H) Density of iSNVs against the minor allele frequency. The red line was fitted using the generalised linear model and described how mutations accumulates neutral evolution. The grey histogram represents the real number of definitely allele frequency range. See also Figure S1 and Table S2.

Figure 2. Spatio-temporal analysis of iSNVs reveals increased genetic diversity. (A) Normalised iSNV number against the date of sample collection based on linear regression. (B) Normalised iSNV number against the time post-symptoms onset. (C) Normalised iSNV number causing nonsynonymous and synonymous mutations against time post-symptoms onset for the S, N and ORF1ab genes. (D) Proportion of
iSNVs causing nonsynonymous and synonymous mutations in each gene. The number of iSNVs in each category is marked on the corresponding bar. (E) Ka/Ks ratio of the S, N, Orf1ab and other genes in different symptoms onset periods. See also Figure S2.

Figure 3. iSNV distributions of different mutation types. (A) Normalised iSNV number for different mutation types. Mutations in coding (red) and non-coding (blue) regions are distinguished by colour. (B) Minor AF values of ADAR targets (A to I, causing A to G and T to C) and APOBEC targets (C to U, causing C to T and G to A) against time post-symptoms onset of patients. See also Figure S3.

Figure 4. Distribution of iSNVs among different groups of patients. (A) Correlation between normalized iSNV number and onset time for patients grouped by gender, age, illness severity and viral shedding time. (B) The estimated accumulation rate by linear model in Figure 4A. The error bar represents the 95% confidence interval of the estimate value. (C) Effects of gender, age, illness severity and viral shedding time on iSNV frequency in the population. The size of the point represents the p-value of a Fisher’s exact test comparing the patients in each population for each protein site. The * symbol in the genome site represents hfSNPs. (D) Histogram (left) shows the frequencies of iSNVs related to illness severity. The iSNVs marked with stars are significantly differentially distributed between severe and non-severe populations. Mutations are distinguished by colour (red = more severe patients, grey = less severe patients). The dashed line represents the average proportion of severe patients. The plot on the right represents the number of cases with these iSNVs. The frequencies of these iSNVs in public databases 2019nCoVR are marked with different coloured lines (levels I, II, III and IV are red, green, blue and purple, respectively). See also Figure S4.
Figure 5. Biased fixation of iSNVs in public SNP databases and local SNP datasets. (A) Number of fixed iSNVs in different time periods. (B) Number of fixed synonymous and nonsynonymous iSNVs in fixed and unfixed phases. (C) Fraction of fixed and unfixed nonsynonymous mutations in each gene. (D) Nonsynonymous/synonymous ratio of identified iSNVs and SNPs in each gene. S* represents S genes without the D614G mutation. (E) Fixation rate of iSNVs over time post-symptoms onset. (F) Example of early fixation trends of C7051T in iSNVs in our dataset (left) and in public databases (right). See also Figure S5.

Figure 6. Genetic and molecular structure analysis of iSNVs observed in S protein genes. (A) Location of iSNVs (top) and the number of samples with iSNVs. RBD and RBM regions are marked. Below are the mutations of amino acid residues in SARS-CoV-2, PCoV-GD, Bat-RaTG13, PCOV-GX and SARS-CoV at the locations corresponding to the shared iSNVs in the RBM region. SNPs present in public databases are marked by stars. PCoV-GX, pangolin CoV isolate GX-PL4; PCoV-GD, pangolin CoV isolate MP789. The GenBank accessions codes for these CoVs are: 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), and PCoV-GD (isolate MP789, MT084071.1). (B) Fold change of viral entry in T Rex 293 hACE2 cells for different iSNV mutations. The dashed lines represent fold-changes of 0.25 and 4. The * symbol indicates mutants with significantly altered viral entry efficacy based on t-tests. Data are represented as mean ± SD. The results were confirmed in at least three separate experiments. (C) Relative fold change in the number of hydrogen bonds, contact area, and binding energy in MD calculations for mutants compared with WT protein bound to hACE2 and CB6. (D) Crystal structures of the SARS-CoV-2 RBD/hACE2 complex for WT and N422K mutant proteins. The red square highlights the affected regions. (E) Crystal structures of the SARS-CoV-2
RBD/hACE2 complex for WT and G485V mutant proteins. The red square highlights the affected regions. See also Figure S6, Table S4, Table S5 and Table S6.
Table 1. Differences in iSNV count, normalized iSNV number, and number of patients with iSNVs among patient subpopulations.

See also Table S1 and Table S3

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>No. of included patients (%)</th>
<th>Median iSNV count (qu1-qu3)</th>
<th>Wilcoxon test p-value</th>
<th>Median normalized iSNVs (qu1-qu3)</th>
<th>Wilcoxon test p-value</th>
<th>Patients with iSNVs (%)</th>
<th>Fisher’s exact test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>170</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>163(95.88%)</td>
<td></td>
</tr>
<tr>
<td>Age groups (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–15</td>
<td>17 (10.00%)</td>
<td>28 (6–52)</td>
<td>0.648</td>
<td>0.637 (0.339–1.130)</td>
<td>0.243</td>
<td>17 (100%)</td>
<td></td>
</tr>
<tr>
<td>16–65</td>
<td>131 (77.06%)</td>
<td>17 (5.5–50)</td>
<td>-</td>
<td>0.552 (0.268–0.863)</td>
<td>0.243</td>
<td>124 (94.65%)</td>
<td></td>
</tr>
<tr>
<td>&gt;66</td>
<td>22 (12.94%)</td>
<td>37.5 (8.75–63.5)</td>
<td>0.224</td>
<td>0.505 (0.332–0.654)</td>
<td>0.866</td>
<td>22 (100%)</td>
<td>0.594</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>83 (48.82%)</td>
<td>15 (3.5–51)</td>
<td><strong>0.029</strong></td>
<td>0.478 (0.247–0.795)</td>
<td><strong>0.019</strong></td>
<td>78 (93.97%)</td>
<td>0.269</td>
</tr>
<tr>
<td>Male</td>
<td>87 (51.18%)</td>
<td>25 (9–55.5)</td>
<td></td>
<td>0.599 (0.342–0.962)</td>
<td></td>
<td>85 (97.70%)</td>
<td></td>
</tr>
<tr>
<td>Disease level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>39 (22.94%)</td>
<td>21 (5.5–57)</td>
<td>0.706</td>
<td>0.489 (0.269–0.812)</td>
<td>0.746</td>
<td>37 (94.87%)</td>
<td>1</td>
</tr>
<tr>
<td>Moderate</td>
<td>98 (57.65%)</td>
<td>17.5 (5–44)</td>
<td>-</td>
<td>0.567 (0.270–0.872)</td>
<td>-</td>
<td>93 (94.90%)</td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>33 (19.41%)</td>
<td>21 (11–66)</td>
<td>0.135</td>
<td>0.545 (0.356–0.932)</td>
<td>0.519</td>
<td>33 (100%)</td>
<td>0.330</td>
</tr>
<tr>
<td>Infection duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–14</td>
<td>22 (12.94%)</td>
<td>7 (3–18.25)</td>
<td><strong>0.004</strong></td>
<td>0.338 (0.182–0.576)</td>
<td>0.087</td>
<td>22 (100%)</td>
<td>0.556</td>
</tr>
<tr>
<td>15–28</td>
<td>48 (28.23%)</td>
<td>16 (6–43.25)</td>
<td>0.225</td>
<td>0.546 (0.336–0.871)</td>
<td>0.921</td>
<td>46 (95.83%)</td>
<td>1</td>
</tr>
<tr>
<td>29–42</td>
<td>43 (25.29%)</td>
<td>29 (12–90)</td>
<td>0.384</td>
<td>0.592 (0.335–0.893)</td>
<td>0.767</td>
<td>41 (95.35%)</td>
<td></td>
</tr>
<tr>
<td>&gt;43</td>
<td>57 (33.53%)</td>
<td>25 (6–68)</td>
<td></td>
<td>0.606 (0.245–0.874)</td>
<td>-</td>
<td>54 (94.74%)</td>
<td></td>
</tr>
</tbody>
</table>
STAR METHODS

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Prof. Chen Chen (chenchen1@ccmu.edu.cn).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

• The sequencing data has been submitted to the National Genomics Data Center, China National Center for Bioinformation and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.

• Original pseudo-virus assay and other experimental data have been deposited at Mendeley and are publicly available as of the date of publication. Additional supplemental items are available from Mendeley Data at http://dx.doi.org/10.17632/rmr4bcsgrp.1.

• The original code is available in Zenodo at http://dx.doi.org/10.5281/zenodo.5703845.

• Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patients and clinical cohort

Our study included all confirmed and admitted patients with at Beijing Ditan Hospital, Beijing, China, where the first case in Beijing was recorded, between January 20 and April 30, 2020. We enrolled 204 cases, accounting for 34.4% of
confirmed patients in Beijing, containing 102 female and 102 male, 22 children (<15 y), 226 mid-age (>15y and <65y) and 29 elders (>65y) (Table S1). All patients were confirmed by RT-PCR tests on pharyngeal swab samples at Beijing Ditan Hospital. All patients were treated and managed in the ward after diagnosis. Standardised electronic medical records were employed to collect basic demographic and epidemiological information, medical histories, and clinical data. 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 (https://www.chinadaily.com.cn/pdf/2020/1.Clinical.Protocols.for.the.Diagnosis.and.Treatment.of.COVID-19.V7.pdf). Patients were classified into three severity categories: mild, moderate and severe (traditionally defined as severe and critical illness). Briefly, mild cases were those with mild clinical symptoms, for whom imaging showed no sign of pneumonia. Moderate cases were those showing fever and respiratory symptoms with radiological findings of pneumonia. Severe cases included adult cases meeting any of the following criteria: (1) respiratory distress (R30 breaths/min), (2) oxygen saturation percentage of 93% at rest, (3) arterial partial pressure of oxygen (PaO2)/fraction of inspired oxygen (FiO2) percentage of 300 mmHg.

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

**Epidemic analysis**

Through detailed epidemiological investigation, we identified 26 infection clusters involving 79 patients (Supplementary Data Table S7). No super-spreader events were identified.
Time point labels for each patient were defined based on clinical characteristics. For patients with symptoms, we set the time of symptoms onset as day 0. For asymptotic patients, day 0 was set as the day of the first positive RT-PCR test. The date of negative transition was defined as the date when all specimens (including pharyngeal swab, sputum and faecal specimens) yielded negative results. Viral shedding time was calculated from day 0 to the date of negative transition.

Samples from the same individual whose samples collected duration between the first and last time points more than 5 days were defined as longitudinal samples. Among the 165 patients whose viral shedding time more than 5 days, we selected 61 patients with longitudinal samples.

METHOD DETAILS

Laboratory procedures

Clinical samples, including pharyngeal swabs, sputum and faecal specimens for RT-PCR tests, were collected in replicate according to the instructions of the infection prevention and control measures in the Chinese guidance on infection prevention and control in healthcare settings. RNA was extracted using previously described methods appropriate for P2+ and/or P3 laboratories (Chan et al., 2020). Viral RNA was extracted using a QIAamp Viral RNA Mini Kit according to the manufacturer’s instructions, except that carrier RNA was omitted to facilitate downstream high-throughput sequencing analysis. DNA was removed from the samples via DNase I (NEB) digestion. The resulting total RNA was concentrated using an RNA Clean & Concentrator-5 Kit (Zymo Research), and its quality was assessed by a Fragment Analyzer Automated CE System (AATI). Quantification was performed by Qubit 2.0 (Invitrogen). Diagnostic real-time RT-PCR assays targeting the open reading frame 1ab (ORF1ab) region and nucleoprotein (N) gene of SARS-CoV-2 were performed as described elsewhere (Zhu et al., 2020). A cycle threshold (Ct) value ≤37 in at least
one gene was interpreted as a positive test for SARS-CoV-2 according to Chinese national guidelines. The Ct values of the tested samples ranged from 12.00 to 37.52.

**High-throughput genomic sequencing of the viral genome**

We collected 183 pharyngeal swab samples, 241 sputum samples, and 113 faecal samples for meta-transcriptomic sequencing. Viral RNA was extracted using the protocol described above. After performing ribosomal RNA (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 (Chen et al., 2020). This approach uses direct tagmentation of RNA/cDNA hybrids with Tn5 transposase to greatly simplify the sequencing library construction process, allowing us to conduct rapid library preparation using low volume input RNA templates (5.4 µl) within 4 h. The step-by-step protocol is available in a previous paper (Chen et al., 2020). Briefly, 2.7 ml RNA (after rRNA and DNA removal) was used for standard SHERRY reverse transcription, with the following modifications: (1) 10 pmol random decamers (N10) were added to improve coverage and (2) initial concentrations of dNTPs and oligo-dT (T30VN) were increased to 25 mM and 100 mM, respectively. For the 5.4 ml and 10.8 ml input volumes, the entire reaction was simply scaled up 2- and 4-fold, respectively. The RNA/cDNA hybrid was tagmented in TD reaction buffer (10 mM Tris-HCl pH 7.6, 5 mM MgCl₂, 10% DMF) supplemented with 3.4% PEG8000 (VWR Life Science), 1 mM ATP (NEB), and 1 U/ml RNase inhibitor (TaKaRa). The reaction was incubated at 55°C for 30 min. A 20 ml volume of tagmentation product was mixed with 20.4 ml Q5 High-Fidelity 2X Master Mix (NEB) and 0.4 ml SuperScript II reverse transcriptase, after which it was incubated at 42°C for 15 min to fill the gaps, followed by 70°C for 15 min to inactivate SuperScript II reverse transcriptase. Next, index PCR was performed by adding 4 ml 10 mM unique dual index primers and 4 ml Q5 High-Fidelity 2X Master Mix, and thermal cycling was performed at 98°C for 30
s, followed by 18 cycles at 98°C for 20 s, 60°C for 20 s, and 72°C for 2 min, with a final step at 72°C for 5 min. The PCR product was then purified using 0.8× VAHTS DNA Clean Beads (Vazyme).

The meta-transcriptome libraries were further enriched using biotinylated RNA/cDNA probes targeting the whole viral genome (iGeneTech, Beijing, China). The library was first quantified for the N gene using quantitative PCR after a 1:200 dilution. Next, 8–16 libraries were pooled together based on qPCR results. The pooled library was further processed with a TargetSeq One Cov Kit (iGeneTech) following the manufacturer’s instructions. The iGeneTech Blocker was replaced by IDT xGen Universal Blockers (NXT). The final viral-enriched libraries were sequenced on an Illumina NextSeq500 platform in 2×75 bp paired-end mode.

**Sequencing analysis of the viral genome**

Quality control and error correction were implemented as previously reported (Ni et al., 2018). To avoid nucleotide-specific substitution errors in each read, we removed low-quality bases at the ends 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 iSNV analysis. We firstly mapped clean read data to the Wuhan-Hu-1 reference genome (GenBank accession no. NC_045512.2) using Bowtie2 v2.1.0 (Langmead and Salzberg, 2012) with default parameters and obtained 4.11 Mb (QRI: 0.44–22.96 Mb) high-quality viral reads per sample. After removing low-quality genomic data, we obtained 402 samples with sufficient data that met the following criteria: (i) sequencing depth ≥1 and reference coverage ≥50%, and (ii) depth ≥100 and reference coverage ≥10%.

**iSNV calling to avoid sequencing errors and contaminants**
Calling of iSNVs was performed as previously described, but with different parameters (Ni et al., 2016). Briefly, (1) sequencing reads were paired-end aligned to the reference genome sequence and reformatted using SAMtools v1.3.1 (Li et al., 2009); (2) for each site of the SARS-CoV-2 genome, aligned low-quality bases and indels were excluded to reduce the potential for false-positive results, 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. Several criteria were used to ensure high-quality iSNVs: (1) minor allele frequency ≥5% (a conservative cut-off based on error rate estimation), (2) depth of the minor allele ≥5, and (3) strand bias of the minor allele less than 10-fold or no significance detected located in one strand (Fisher’s exact test for the major allele and minor allele). To avoid errors from read mapping, serial adjacent iSNVs (distance <50 bp) containing >5 iSNVs were also filtered. (4) PCR duplicate were removed with Picard MarkDuplicates, and all iSNV sites should be observed in the dataset after removing PCR duplicate.

Sequencing errors, and potential contaminants from metagenomic sequencing are significant concerns associated with next-generation sequencing. To effectively detect contaminants and sequencing errors, we performed several procedures and tests. (1) Negative control samples from healthy cases were tested to determine the potential for false-positive results. (2) The iSNV patterns from each round were carefully examined, and similar iSNV patterns were not observed in different rounds. (3) The nucleotide statistics for each site showed that we detected only 12/7037 iSNVs containing three polymorphic states, indicating that few sequencing errors occurred.

**Validation of iSNVs by sample re-sequencing and PCR validation**

To validate the reproducibility of the iSNVs data, we prepare two independent libraries in 62 of 402 samples. Among them, 55 samples met the criteria for iSNV calling (sequencing depth >100x, and >10% coverage). Under the 5% cut-off for
mutation allele frequency (MuAF), we identified 498 iSNVs and 525 iSNVs in the
first and second experiments, respectively. Over 90% (450) of iSNVs identified in the
first round of sequencing were also observed in the second round (Figure S7A). The
MuAFs in the two experiments were similar and highly correlated (R-square, 0.89;
Figure S7B).

We also estimated the proportion of shared iSNVs under MuAF cutoff thresholds
of 0.1%, 1%, 5%, 10% and 20% (Supplementary Data Table S8). We set the iSNVs
identified in the second round but not in the first round as false positive mutations.
The false positive rate is defined as the ratio of number of false positive mutations and
iSNV numbers identified in the first round.

We also test if the reproducibility is affected by Ct value. We plot the false positive
rate against the Ct value and time post-symptoms onset. Result showed that our
validation on mutations was robust in reproducibility with viral load and disease
process (Figure S7C, Figure S7D).

Ten PCR primer-pairs were designed for 12 genomic sites (337, 3429, 6195/6268,
7468, 10074, 18512, 17119/17318, 17812, 21711 and 23731). The primer sets are
listed in the Supplementary Data Table S9. Limited by the efficiency of primers and
the quality of the samples, we obtained positive PCR results for 10 genomic sites for 8
patients (16 samples). Further Sanger sequencing indicated that 7 genomic sites had
mutants or iSNVs in these samples.

Correlation of iSNV density and Ct value
To test if the difference of observed number of iSNVs were due to false positive in
low viral load samples, we plot the iSNV numbers to Ct value and found that iSNV
richness slightly decreased with higher viral load, consistent with the previous report
(Popa et al., 2020) (Figure S7E). We corrected the iSNV per KB by Ct value with
linear model (Figure S7F), and observed the same increase trend over time after
correction (Figure S7G).
Negative control sample in sequencing

To determine the potential false positive result, we sequenced 95 negative control samples from 31 individuals, including 40 pharyngeal swab, 15 sputum and 40 faeces. We obtained 1502 reads mapped to the SARS-CoV-2 genome in median for the negative control sample. All of the negative control sample did not meet the crania of complete genome (consensus genome >50% and deep sequencing genome >10%) (Figure S7H). The reads mapped to the SARS-CoV-2 genome showed significant difference in the negative control vs. high and low coverage samples (Figure S7I).

Normalised iSNVs and mutation rate estimation

The number of iSNVs for each sample was normalised by the number of iSNVs per kb, with only the region used to identify iSNVs used for calculation. The iSNV sites were identified throughout the genome by removing duplicated iSNVs in each patient, ensuring that iSNVs in different samples from each patient were only counted once. We also applied linear regression to evaluate the correlation between iSNVs and the time after the initial case reported, as well as the correlation between iSNVs and individual infection time of each patient.

We defined minor allele frequency (minor AF) as the allele frequency of the minor allele, whether a reference allele or alternative allele, representing a 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) based on different mutation rate frequencies (a). According to the Bozic model (Bozic et al., 2016), the expected number of mutations was calculated using the following formula:

\[ m = \frac{u(1 - a)}{(1 - \frac{b}{d})a} \]
where $\mu$ represents the mutation rate per genome per replication cycle, $b$ represents
the virus reproduction rate, and $d$ represents the rate of virus decline in the population.

We applied a generalised linear model to calculate the parameters and plotted the
predicted values. The SARS-CoV-2 virus reproduction and decline rates $b$ and $d$ were
not clear at the time of the study. Therefore, for comparisons of mutation rates, the
numbers of expected mutations with different mutation rates were calculated using the
virus reproduction and decline rates reported for influenza virus (Beauchemin and
Handel, 2011; Xue et al., 2018).

Simulation of the iSNV mutation position distribution was performed with the runif
function in the R using a uniform distribution with the number of iSNVs as
parameters. We then ranked the actual iSNVs and simulated positions, after which we
calculated the genomic distance between pairs of neighbouring mutants. The genomic
distance between pairs of simulated iSNVs followed the Poisson distribution.

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 the NCBI. Public SNP
files were downloaded from the 2019nCoVR public database (Song et al., 2020; Zhao
et al., 2020a) (https://bigd.big.ac.cn/ncov/variation/annotation) on November 19 2020,
yielding whole genome sequences from CNGBdb, GenBank, GISAID, GW and
NMDC. To better understand the frequencies of SNPs that occurred in populations,
each iSNV was compared with the levels of SNPs according to the frequency in
2019nCoVR during the same period of sampling. Levels I to III were set according to
previous definitions in public databases, where level I represents a frequency >0.05,
level II represents a frequency between 0.01 and 0.05, and level III represents a
frequency <0.01. The iSNVs not present in public databases were assigned to level
IV. Ka and Ks were calculated by KaKs_Calculator (Wang et al., 2010) (version 2.0,
June 2009 2020) using the MYN model (Zhang et al., 2006), taking the rates of
transitional and transversional substitution as well as codon frequency bias into consideration. The Ka/Ks ratio has an expected value of 1 for neutral evolution.

**Epitope region prediction**

Predicted epitope regions were downloaded from the IEDB database (https://www.iedb.org/) (Dhanda et al., 2019), with ‘Epitopes’ set to ‘Linear Epitope’, ‘Organism’ set to ‘SARS-CoV-2’ and ‘Host’ set to ‘Human’. Detailed epitope regions and sequences are provided in the Supplementary data (Table S10). We also explored the experimental epitope regions with the same parameters as the predicted epitopes, but with two different parameters: ‘Epitopes’ set to ‘Any’ and ‘Assay’ set to ‘B Cell Assays: neutralisation | biological activity’.

**Identifying highly correlated iSNVs in the genome by phasing analysis**

Pairwise phasing analysis was performed for adjacent iSNVs (distance <50 bp) as previously reported (Chen et al., 2018). For a given pairwise iSNV, reads harbouring both positions were extracted from the alignment (SAM file). Reads with both sites mutated were designated as phased reads, and those with only one site mutated were designated as non-phased reads. Reads with a ratio of phased to non-phased reads >0.9 were selected as phased iSNVs. Moreover, the phased alternative iSNV allele frequency was required to be >0.05. For phased iSNVs in a given protein codon, we re-annotated the iSNVs with phased alleles.

Despite the short distance (<50 bp) in the genome, the correlations between long-distance (>50 bp) pairwise iSNVs were calculated by the linear adjusted R-square of variation. iSNVs were considered to be highly correlated and potentially linked if they were identified in at least three samples and possessed highly correlated MuAFs (>0.6).

**S protein structure analysis and molecular dynamics simulation**
All molecular dynamistic (MD) simulations were performed using the AMBER 20 package. The crystal structure of the binary complex of hACE2 and RBD (PDB ID 6LZG) (Wang et al., 2020) was used as the initial structure in the MD simulation. The protein was modelled with the AMBER FF14SB (Maier et al., 2015) all-atom protein force field and solvated by a truncated octahedron TIP3P (Jorgensen et al., 1983) water box, in which the boundary was at least 11 Å from any protein atoms. The solvated protein was neutralised and filled with 0.13 M KCl salt. In these simulations, the SHAKE (Ryckaert et al., 1977) algorithm with a relative geometric tolerance of $10^{-5}$ was used to constrain all chemical bonds. Mass repartitioning was applied to adjust the mass of the heavy atom to which the hydrogen was attached so that the total mass remained constant. Thus, all dynamics utilised a 4 fs time step. Long-range electrostatics were treated using the particle-mesh Ewald (PME) (Darden et al., 1993) method with default settings, and a 9 Å direct space non-bonded cutoff was used in all simulations. The system was first subjected to 10,000 steps of minimisation, after which it was gradually heated to 300 K under constant volume conditions in 1 ns. After another 5 ns of simulation using the constant isothermal-isobaric ensemble at 1 atm and 300 K, each system was equilibrated for an additional 10 ns. A Monte Carlo barostat and a weak-coupling thermostat were used. The MD simulations were performed for 300 ns with coordinates recorded every 10 ps. The same procedure was followed for simulations of the CB6 antibody (Shi et al., 2020) and the RBD complex (PDB ID 7C01 and 7BWJ). Nine RBD mutants (V407L, N422K, L452Q, E471K, V483F, G485V, Q493H, Q498H and Y505C) bound to either hACE2 or CB6 were also simulated. Simulations of these systems were performed with the procedure described above. In the analysis of simulation trajectories, hydrogen bonds were defined as a geometry with a cut-off length of 3.5 Å between the two heavy atoms of the hydrogen bonding donor and acceptor, and an X-H· · ·Y (X and Y represent heavy atoms) angle cutoff of 135°. A hydrogen bond was counted when the distance between X and Y was less than 3.5 Å and the X-H· · ·Y angle was greater than 135°.
Packaging of pseudoparticles bearing the SARS-CoV-2 spike protein and its variants

The codon-optimised S gene of SARS-CoV-2 (NC_045512.2) was incorporated into the pSecTag2/Hygro A plasmid and used as a template to generate S mutants by site-directed mutagenesis. The various S protein pseudoviruses bearing luciferase reporter genes were packaged as reported previously (Zhao et al., 2020b). In brief, 24 h prior to transfection, 293T cells were plated at a density of 6×10^5 cells 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 S protein, either reference protein 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 S protein expression plasmid at a ratio of 4:1. Viral supernatants were harvested at 48 h and 72 h post-transfection, centrifuged to remove cell debris, and filtered through a 0.45 µm filter unit (Sartorius). A Lenti-X p24 rapid titre kit (Takara) was used to quantify the viral titres following the manufacturer’s instructions.

Generation of cell lines expressing hACE2 and virus infectivity assay

The T Rex 293 hACE2 cell line, which expresses human ACE2 in a tetracycline-dependent manner, was established previously (Zhao et al., 2015). In brief, Flp-IN T Rex cells were plated at a density of 5×10^5 cells per well in a 6-well plate. The next day, the plated 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, the transfected cells were trypsinised 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 cells was expanded to generate cell lines that expressed hACE2 (T Rex
293 hACE2) upon the addition of tetracycline to the culture medium at a concentration of 1 μg/ml.

T Rex 293 hACE2 cells transfected with the 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 an HIV-1 p24 antigen quantification assay, we normalised the amount of pseudotyped virus particles (10 ng of p24). After normalisation, 100 μl of pseudotyped virus at a 10-fold dilution was added to each well of a 96-well T Rex 293 hACE2/TMPRSS2 culture plate, which was incubated at 37°C in a humidified atmosphere with 5% CO₂. The culture medium containing 2% foetal bovine serum (FBS) was refreshed after 12 h, and cells were incubated for an additional 48 h. Assays were performed using 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 mAb neutralisation assay**

For the neutralisation assay, T Rex 293 hACE2 cells transfected with the pCAGGS-TMPRSS2 plasmid were seeded in a 96-well plate at a density of 2×10^4 cells per well and cultured for 12 h upon the addition of tetracycline (1 μg/ml). A 100 μl volume of supernatant containing pseudoviruses was incubated with an equal volume of five-fold serially diluted antibodies for 1 h at 37°C. CB6 mAbs were tested at concentrations ranging from 0.64 ng/ml to 10.00 μg/ml. The mixtures of pseudoviruses and CB6 mAbs were then added to T Rex 293 hACE2/TMPRSS2 cells in a 96-well plate in duplicate. After a 12 h incubation, the medium was replaced with DMEM containing 2% FBS and samples were incubated for an additional 48 h at 37°C. Luciferase activity was measured using a GloMax 96 Microplate luminometer (Promega). The titres of CB6 mAbs were calculated as the 50% inhibitory concentration (IC50) using GraphPad Prism 6.0.
QUANTIFICATION AND STATISTICAL ANALYSIS

Continuous variables were expressed as median and interquartile range (IQR) values as appropriate. Categorical variables were summarised as the numbers and corresponding percentages in each category. The distribution of distance between adjacent iSNVs and expected Poisson model were performed by Kolmogorov-Smirnov test. The distance of mutations frequency under neutral selection and real frequency were performed by linear model. The distribution of codon positions was compared with the normal distribution using the Kolmogorov-Smirnov test. The correlations between iSNVs and age, gender, illness severity and viral shedding time were assessed by the Kruskal-Wallis test. The accumulation rate difference among different patient groups were performed by ANOVA linear model test. Only sites at which iSNVs occurred in more than six patients were included in the analysis. The significance of the correlations between the proportions of the population with iSNVs and hfSNPs in public databases was calculated by Fisher’s exact test.

SUPPLEMENTARY DATA

SupplementaryData1. Primers and predicted epitope region. Related to STAR method.

ACKNOWLEDGMENTS

We thank all health care workers involved in the diagnosis and treatment of patients with COVID-19 at Beijing Ditan Hospital, Capital Medical University. This research was supported by the Beijing Hospital Authority (No. DFL20191801), the Beijing Municipal Science & Technology Commission (No. Z201100007920017 and Z201100005420012), and the National Natural Science Foundation of China (92053202, 22050003 and 21873007).

AUTHOR CONTRIBUTIONS
J.L., P.D. and N.D. performed sequencing data analysis; D.C. and X.Z. performed pseudoviral assay infection experiments. L.Y. and Y.G. performed 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. and F.Y. collected and analysed clinical data; C.S., K.H., L.D., J.L., J.Y., M.H., J.H. and 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 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


Influenza Virus. Trends Microbiol 26, 781-793.


A

Genomic coverage
- 0.25
- 0.50
- 0.75

Specimens
- P (N = 99, n = 183)
- S (N = 144, n = 241)
- F (N = 70, n = 113)

B

Case number

Samples per case

C

#iSNVs per KB

Genome position

D

5'UTR

intergenic

3'UTR

coding

E

Codon position

4th

3rd

2nd

1st

F

# of samples shared iSNV

Genome position

G

Density

Genomic distances between two variants

H

#iSNV per site (10^-3)

MAF

Neutral model
9 mutations in S protein outside of RBD in more than 6 individuals
11 mutations changed amino acids in RBD or cleavage sites in more than 2 individuals

Viral entry in 293T cells with hACE2+TMPRSS2

C

SNP fixed before May, 2020
SNP fixed after May, 2020

Number of HB Contact Area ΔG

Fold change

Fold change

WT
P26H/R246E/S247N/D614G/A694D/M731V/I771V/I1130L
V407L/N422K/L452Q/E471K/V483F/G485V/Q493H/Q498H/Y505C/L518V/R685G

D

WT
N422K

E

WT
G485V
Highlights:

- Intra-host variations provide a mutational pool for SARS-CoV-2 rapid evolution.
- Two-step selection contributes to accumulation of intra-host variation.
- iSNVs in SARS-CoV-2 display a positive selection longitudinally in individuals.
- The transmission bottleneck removes nonsynonymous mutations from viral population.

eTOC
The Intra-host variations in SARS-CoV-2 provide a mutational pool shaping the virus rapid global evolution. Here, Li et al. illustrate dynamic changes of iSNVs in a longitudinal cohort, and explored a two-step fitness selection process of within-host viral evolution.