Virus Evolution, 2022, **8(2)**, 1–11

DOI: https://doi.org/10.1093/ve/veac106
Advance access publication date: 29 November 2022
Research Article

SARS-CoV-2 intra-host single-nucleotide variants associated with disease severity

Yi Zhang,[1,†] Ning Jiang,[1,2,†] Weiqiang Qi,[3,†] Tao Li,[3,†] Yumeng Zhang,[1] Jing Wu,[1] Haocheng Zhang,[1] Mingzhe Zhou,[1] Peng Cui,[1] Tong Yu,[1] Zhangfan Fu,[1] Yang Zhou,[1] Ke Lin,[1] Hongyu Wang,[1] Tongqing Wei,[2] Zhaoqin Zhu,[3,*,‡] Jingwen Ai,[1,*,‡,§] Chao Qiu,[1,*,4,‡] and Wenhong Zhang[1,2,*,‡,**]

[1]Department of Infectious Diseases, National Clinical Research Center for Aging and Medicine, Shanghai Key Laboratory of Infectious Diseases and Biosafety Emergency Response, Huashan Hospital, Fudan University, Shanghai 200040, China, [2]State Key Laboratory of Genetic Engineering and Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai, China, [3]Shanghai Public Health Clinical Center, Shanghai 200051, China and [4]Institutes of Biomedical Sciences, Shanghai Medical College, Fudan University, Shanghai, China

[†]These authors contributed equally (Y. Zhang, N. Jiang, W. Qi, and T. Li).

[‡]These corresponding authors contributed equally (W. Zhang, C. Qiu, J. Ai, and Z, Zhu).

[§]https://orcid.org/0000-0002-5152-1557

[**]https://orcid.org/0000-0002-9165-3212

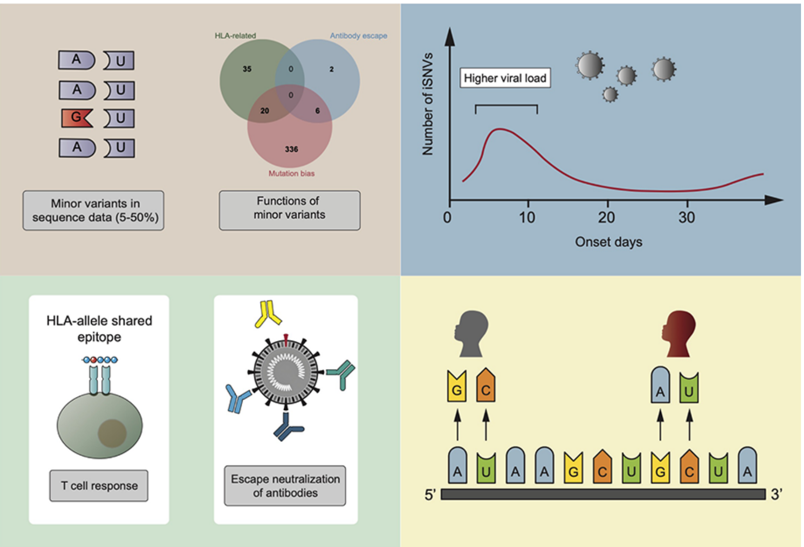
*Corresponding authors: E-mail: zhangwenhong@fudan.edu.cn; qiuchao@fudan.edu.cn; jingwenai1990@126.com; zhaoqinzhu@163.com

Abstract

Variants of severe acute respiratory syndrome coronavirus 2 frequently arise within infected individuals. Here, we explored the level and pattern of intra-host viral diversity in association with disease severity. Then, we analyzed information underlying these nucleotide changes to infer the impetus including mutational signatures and immune selection from neutralizing antibody or T-cell recognition. From 23 January to 31 March 2020, a set of cross-sectional samples were collected from individuals with homogeneous founder virus regardless of disease severity. Intra-host single-nucleotide variants (iSNVs) were enumerated using deep sequencing. Human leukocyte antigen (HLA) alleles were genotyped by Sanger sequencing. Medical records were collected and reviewed by attending physicians. A total of 836 iSNVs (3–106 per sample) were identified and distributed in a highly individualized pattern. The number of iSNVs paced with infection duration peaked within days and declined thereafter. These iSNVs did not stochastically arise due to a strong bias toward C > U/G > A and U > C/A > G substitutions in reciprocal proportion with escalating disease severity. Eight nonsynonymous iSNVs in the receptor-binding domain could escape from neutralization, and eighteen iSNVs were significantly associated with specific HLA alleles. The level and pattern of iSNVs reflect the *in vivo* viral–host interaction and the disease pathogenesis.

Key words: COVID-19; SARS-CoV-2 variants; intra-host; immune response; HLA; neutralization antibody.

Graphical Abstract



Introduction

Coronavirus disease 2019 (COVID-19) is a newly emerging infectious disease caused by a positive single-stranded ribonucleic acid (RNA) virus—severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Huang et al. 2020; Lu et al. 2020; Wu et al. 2020; Zhou et al. 2020). Variants of SARS-CoV-2 escape from the immune system of human hosts, causing cracks in vaccination efforts to reach herd immunity (Cele et al. 2021; Garcia-Beltran et al. 2021; Hacisuleyman et al. 2021; Li et al. 2021; Emary et al. 2021a). However, the majority of variants that develop within hosts are eventually eliminated or not transmitted, and their significance, thus far, has been largely overlooked. Some mutations in the SARS-CoV-2 genome are selected by immune pressure as a consequence of viral–host interactions. Decoding these viral mutation profiles in individual cases might uncover the details of disease-associated *in vivo* viral–host interactions, identify regions important for immune recognition, and forecast future immune-resistant variants.

Once the founder SARS-CoV-2 infects the human host, there are several paths to generate and select mutations (Dolan, Whitfield, and Andino 2018). First, viral mutations may arise stochastically through rapid replication within the host (Su et al. 2016). Second, the host-adaptive mutations might change the viral transmissibility and lethality (Korber et al. 2020; Challen et al. 2021; Davies et al. 2021; Tegally et al. 2021) and produce immune escape variants with the selective advantage of decreased sensitivity to naturally acquired or vaccination-induced immunity (Collier et al. 2021; Madhi et al. 2021; Supasa et al. 2021; Wang et al. 2021; Wu et al. 2021; Emary et al. 2021b). Third, mutations compromised in fitness of viral replication could revert when jumping into infected individuals where the selective pressure no longer exists (Allen et al. 2004; Davenport et al. 2008; Liu et al. 2021).

With deep sequencing technologies, it is feasible to sequence each sample at a depth of coverage adequate to enumerate the rare frequency variants in quasispecies. When the variants exist, multiple bases will be called at the same location on the aligned sequence reads, which are termed intra-host single-nucleotide variants (iSNVs). The iSNVs have been successfully employed in epidemiological chain investigations, gauging transmission bottleneck sizes and estimating infection time (Giorgi et al. 2013; Puller, Neher, and Albert 2017; Popa et al. 2020; Lythgoe et al. 2021; Valesano et al. 2021). However, the lack of medical records causes challenges in field studies and usually makes it difficult to categorize cases via disease classification. Therefore, determining whether the iSNV profile is related to disease severity remains unexplored and inconclusive.

Shanghai was one of the first affected metropolises to have successfully contained the initial outbreak of SARS-CoV-2 by March 2020 through the implementation of public health interventions. The active surveillance and quarantine effectively prevented the epidemic from spreading into the local community and also provided an opportunity to capture the asymptomatic and presymptomatic cases and follow them up throughout the latter course of infection.

Moreover, because all the samples were collected at the onset of the pandemic, the SARS-CoV-2 consensus sequences of each specimen are highly homogeneous with several unique lineage-defining sites. Thus, this study takes advantage of these individual patients with similar founder virus regardless of eventual symptom severity and investigated whether the burst of its genetic variants contains the footprint imprinted by the host immune responses fencing off SARS-CoV-2, which might make a major distinction in disease severity. Our findings improve the understanding of genetic information conveyed in sequences of samples with distinct clinical outcomes that contribute to fundamental properties of SARS-CoV-2 infection and disease pathogenesis.

Methods

Patients and sample collection

A total of sixty-one COVID-19 patients were admitted to Huashan Hospital and Shanghai Public Health Clinical Center from 23 January to 31 March 2020. We included thirty-eight male and twenty-five female patients with ages ranging from 7 to 82 years. All patients had confirmed positive reverse transcription polymerase chain reaction tests targeting SARS-CoV-2. The clinical course classification was according to an expert consensus statement from the Shanghai Clinical Treatment Expert Group (Shanghai Clinical Treatment Expert Group for CoronaVirus Disease 2019, 2020). The clinical information included onset days, diagnosis time, admission time, symptoms, underlying disease, chest X-ray, laboratory examinations, and antiviral drugs. Epidemiologic data were also obtained. We extracted total viral RNA from patients' sputum or nasopharyngeal swab using the QIAamp Viral RNA Mini Kit (QIAGEN, Germany). Peripheral blood mononuclear cells (PBMCs) were prepared using whole blood. Then, the viral RNA and PBMC samples were both stored at –80°C. This study and informed consent were approved by the Ethical Committee of Huashan Hospital of Fudan University.

SARS-CoV-2 amplicon sequencing

RNA was reverse transcribed into complementary deoxyribonucleic acid (DNA) using the TaKaRa PrimeScript RT Master Mix kit (TaKaRa, Japan). Next, we performed amplicon sequencing spanning the whole genome of SARS-CoV-2. The SARS-CoV-2–specific primers with overlaps were synthesized as described (Zhang et al. 2020). The primers were divided into two pools: A and B, and each included forty-nine pairs of primers with approximately 400 base pairs (bp). Then, PCR products were purified with AMPure XP magnetic beads (Beckman, USA) and prepared for libraries according to Zhang et al. (2020). The qualified libraries were sequenced on an Illumina NovaSeq 6000 Platform (Illumina, USA) using a pair-end 150-bp strategy. Three pairs of samples were employed to carry out technical replicate experiments. The sequencing data have been deposited to the Genome Sequence Archive database of the National Genomics Data Center (https://bigd.big.ac.cn/) with submission numbers PRJCA008324 and PRJCA012126.

Sequence analysis, iSNV calling, and substitution frequency

Raw sequencing data were filtered using trim galore (v3.4) by removing reads that contained adapter sequences, low-quality and short sequences (Phred Quality Score <20 and length <75 bp), and reads with over twenty ambiguous bp (base quality <20). SARS-CoV-2 strain Wuhan-Hu-1 (accession number: NC_045512.1) was used as the reference genome for mapping reads. Bowtie2 (v 2.3.3.1) was used for mapping reads, and candidate single-nucleotide polymorphisms (SNPs) were identified using SAMtools (v 1.9). The number of mapping reads, mapping ratio, sequence coverage, and depth was generated to evaluate the quality of specimens. To minimize false discovery, analyses were conducted on samples with at least 100-fold mean coverage. The iSNV sites were determined as referenced (Ni et al. 2016; Wang et al. 2021): first, Phred Quality Scores of ≥20 and ≥200× depth were satisfied; second, (1) minor allele frequency (MAF) of ≥5 per cent, (2) at least ten reads to support the minor allele, and (3) strand bias of the minor allele and reads with major allele less than ten folds. Additionally, iSNVs located at SARS-CoV-2 primers, 20 bp upstream and downstream of the primers, were removed. Annotations were then made according to the reference genome of National Center for Biotechnology Information. The frequency of twelve types of substitutions was calculated using the ratio of iSNV numbers to corresponding bases (A, U, C, and G, separately). The mutated allele frequency (MuAF) for each candidate nucleotide site was the ratio of reads with the mutated allele of all uniquely mapped reads.

Phylogenetic analysis

Mapped to SARS-CoV-2 strain Wuhan-Hu-1, we further detected the major mutations, namely SNP of consensus sequences for each specimen. Bowtie2 (v 2.3.3.1) was used for mapping reads, and candidate SNPs were identified using SAMtools (v 1.9). We constructed phylogenetic trees using Molecular Evolutionary Genetics Analysis X with maximum-likelihood estimation and the general time reversible nucleotide substitution model. Evolutionary distances with at least 500 replicates were used for bootstrapping. The results were used to annotate the phylogenetic tree in iTOL v4.

HLA typing and association calculation

Genomic DNA was extracted from 1 million PBMCs per sample with QIAGEN DNeasy Blood and Tissue kits (QIAGEN, Germany). We applied the PCR sequence–based typing method using the HLA Class-I typing kit (Weihe Company, Jiangsu Province, China). To obtain a higher-resolution relationship between HLA type and iSNVs, we performed an association analysis. The iSNVs should meet the following criteria: the HLA allele type accounted for ≥5 per cent of total patients (≥3) and nonsynonymous or truncated iSNV alteration existed in more than 5 per cent of patients. Unadjusted Pearson's chi-squared test and odds ratio were calculated first. Then, a Bonferroni-corrected P-value of <0.05 was considered statistically significant. We calculated the association between HLA and iSNV using mixed linear regression; iSNV existence was treated as fixed effect; and age, sex, and onset days were random effects.

Statistical analysis

Pearson's chi-squared test was used to evaluate independent binomial variables. We also performed the analysis of variance when normal distribution was satisfied and Kruskal–Wallis rank tests when these assumptions could not be met. The Bonferroni correction was used to compare each pair of groups. Pearson's correlation analysis was conducted to analyze correlations. Poisson analyses were employed to analyze the correlation between iSNV and depth, and sampling date. We have further calculated the distribution of nonsynonymous/synonymous (NS/S) among types of substitutions using Poisson regression. For mutation pattern and disease severity association, we separated moderate and severe apart from analyses of all enrolled patients and added total iSNV numbers as covariates. P-values of <0.05 were considered statistically significant. Statistical analysis was performed using Stata (v 14.0) software, and figures were generated using GraphPad Prism (v 8) and RStudio (v 1.2).

Results

Demographic characteristics of patients and SARS-CoV-2 genomes

To better understand SARS-CoV-2 within-host diversity throughout COVID-19 infection, we characterized the iSNVs in cross-sectional high-quality stocked RNA specimens in different disease phases. The detailed descriptions including the sampling date and the phases of disease for each patient are shown in Fig. 1, and a summary of patient characteristics is given in Supplementary Table S1. Since active screening was employed in Shanghai, asymptomatic cases (three samples collected within 1–2 days after confirmation by quantitative polymerase chain reaction assay) and patients in the presymptomatic phase (four samples collected 1–4 days prior to symptom onset) had been identified from the group at high risk of contracting the infection. In samples collected at the symptomatic phase, five, forty-four, and five samples from mildly symptomatic, moderate, and severe or critical cases, respectively, were collected mostly within an average of 7 (ranging from 0 to 33) days of the disease progression to moderate or severe phase. Fifty-six of the sixty-one COVID-19 patients were administered the antiviral medicine, with thirty-seven, nineteen, and fourteen patients receiving chloroquine, umifenovir, and lopinavir/ritonavir, respectively, and two patients received darunavir/cobicistat. Only one patient was administered oseltamivir. Glucocorticoids were utilized by eleven patients.

A total of 105 SNPs were identified in 63 sequences. The phylogenetic tree of each viral consensus sequence demonstrated that few lineage-featured substitutions divided these samples into six GISAID lineages: L, S, V, G, GH, and GR (Supplementary Fig. S1). Thirty-six samples from domestic cases of Wuhan-related exposures and the other samples from independent overseas importation cases were collected. The L lineage ranked first in local cases, whereas most viral strains (80.0 per cent, 20/25) were classified as G, GH, and GR types for the imported cases. The distinction of samples with known time and geographical information suggested that laboratory contamination was unlikely.

Distributions and characteristics of iSNVs

High-quality deep sequencing data with a minimum of 3,500× mean read depth and genome coverage of 98.49 per cent enabled us to detect reliable iSNVs at low frequency (Supplementary Dataset 1). We established a stringent cutoff (MAF ≥5 per cent), which is far higher than the threshold error rates of PCR and sequencing. Three pairs of samples were employed to carry out technical replicate experiments, and the allele frequency correlation index R^2 was 0.9514 (Supplementary Analyses S1). Consequently, 836 iSNVs were detected through amplicon sequencing and filtered through the analysis pipeline; their frequency ranged from 5.02 to 50.00 per cent per site (Fig. 2A). The genomic depth and sampling date had low impacts on iSNV/kb numbers, and the R^2 values of Poisson regression were less than 0.38 and 0.31, indicating that the degree of fitting of Poisson regression was marginal (Supplementary Fig. S2). The fitting curve depicted the trend of the number of iSNVs with infection duration, which increased within days after infection and then declined after the peak until the eradication of the virus (Fig. 2B). Furthermore, in the consecutive samples of a patient (P10), the patterns of iSNV of each day were completely different, which suggests a short life of iSNVs and a highly dynamic process of quasispecies composition (Supplementary Fig. S3).

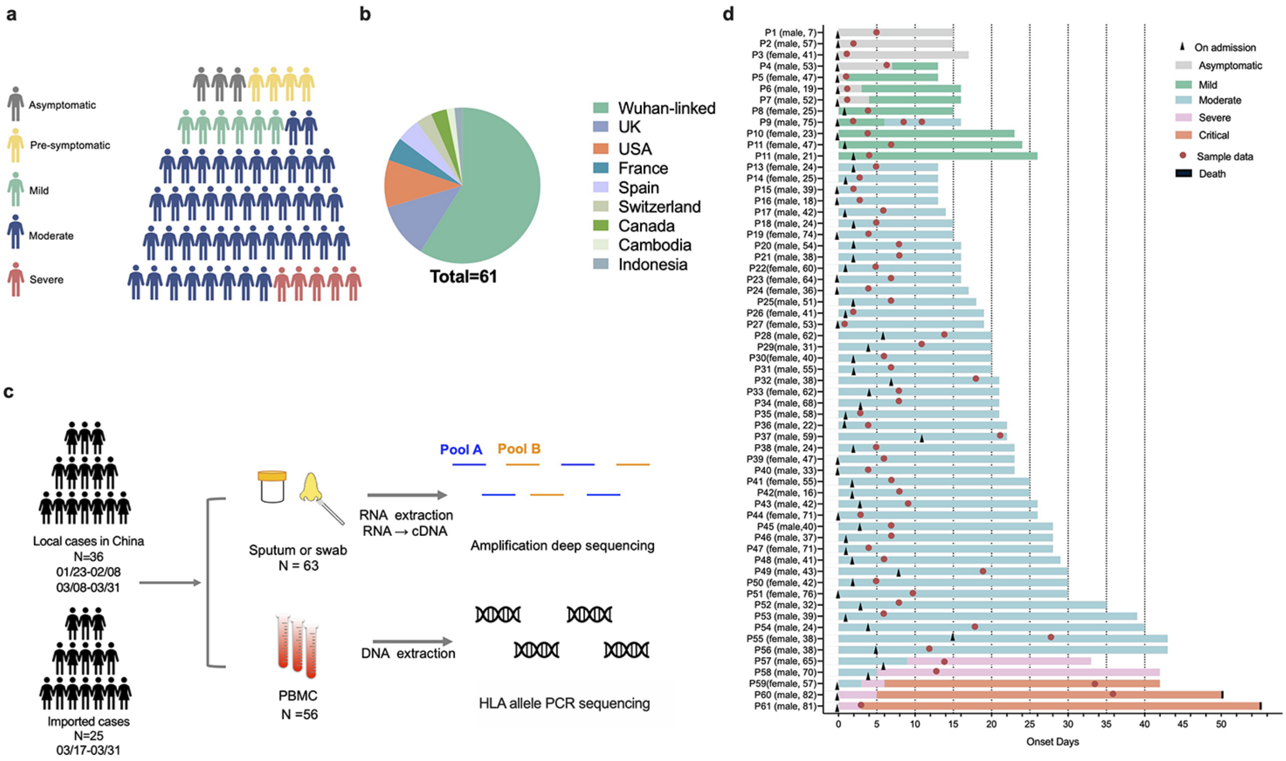


Figure 1. The flowchart and COVID-19 case distribution in this study. (A) The five types of different clinical presentations during collection in this study. Asymptomatic, presymptomatic, mild, moderated, and severe/critical patients were shown using gray, green, blue, and red, respectively. (B) The epidemiological information of enrolled COVID-19 patients. We showed the places they came from as a pie chart. (C) The flowchart of the design of this study. Sputum or nasopharyngeal swabs were sent for amplification deep sequencing. At the same time, PBMC samples were prepared for PCR to identify the HLA allele type. (D) The duration of onset days and hospitalization of enrolled patients. The circle showed the sample date of patients, and the triangles meant the date of on admission. The stripe showed death in Patient (Pt) 60 and Pt 61.

Of these samples, 24.2 iSNVs were detected on average (ranging from 3 to 106 iSNVs), and only one patient had no iSNV. Most samples (90.48 per cent, 57/63) yielded less than 60 iSNVs, while two samples presented more than 100 iSNVs (Fig. 2C). The iSNVs showed a highly individualized pattern (752/836, 78 per cent), with 184 (22 per cent) occurring repeatedly in several samples, while eighteen (2.15 per cent) were identified in over ten samples (Fig. 2D).

Next, we studied the distribution of iSNVs across the viral genome (Fig. 2E) and found that 97.6 per cent (816/836) were located in coding regions; the highest densities of normalized iSNV/kb were in open reading frames (*ORF*)*8a*, *ORF10*, and *ORF3a* (Fig. 2F) (54.64, 51.28, and 50.72 iSNV/kb).

Among transitions, the normalized ratio of substitutions to bases (kb) showed that C > U mutations occurred in more than 2 per cent of C bases (Supplementary Table S2), significantly higher than the other three types of transitions (P < 0.001, odds ratio = 2.68). Among the transversion, the frequency of G > U/C > A (215/491 transversion, 43.79 per cent) was higher than others (P < 0.001, odds ratio = 1.36). We noted that the nonsynonymous substitutions in G > U were more enriched than in other types (P < 0.0001) (Supplementary Table S2). The ratio of nonsynonymous to synonymous substitutions was then calculated (Fig. 2G). Furthermore, the MuAF in *ORF8* was significantly higher than iSNVs in other regions (P < 0.001, Supplementary Fig. S4).

iSNV mutational signatures related to disease severity

Analyses of the transitions C > U/G > A and U > C/A > G (≥3 iSNVs) in patients with different disease manifestations were then conducted to explore the relationship between mutation bias and disease severity. The ratio demonstrated that the frequency of C > U/G > A or U > C/A > G in the sum of these four types was associated with disease severity (Fig. 3A). The mean normalized proportion of C > U/G > A substitutions in asymptomatic or mild disease cases was significantly higher than that of patients experiencing moderate and severe disease cases (45.10 vs. 17.68 per cent, P = 0.007). Conversely, the opposite trend of U > C/A > G substitutions was observed (82.32 per cent in moderate and severe disease cases vs. 54.90 per cent in asymptomatic and milder cases, P = 0.007). Besides, the frequency of substitutions on the negative strand (A > G and C > U) was observed to be significantly lower than that on the positive strand (G > A and U > C) (P < 0.0001), which is consistent with SARS-CoV-2 that is a single-stranded positive-sense RNA virus. We found that the total iSNV number has no significant effect on C > U/G > A or U > C/A > G proportions. After separating moderate and severe disease cases, the mutation pattern could still be observed (Supplementary Analyses S2).

Then, the frequency of C > U/G > A and A > G/U > C substitutions in different coding sequence (CDS) regions was calculated. The results here were normalized using the denominator of bases in each ORF region. The A > G/U > C substitutions were enriched

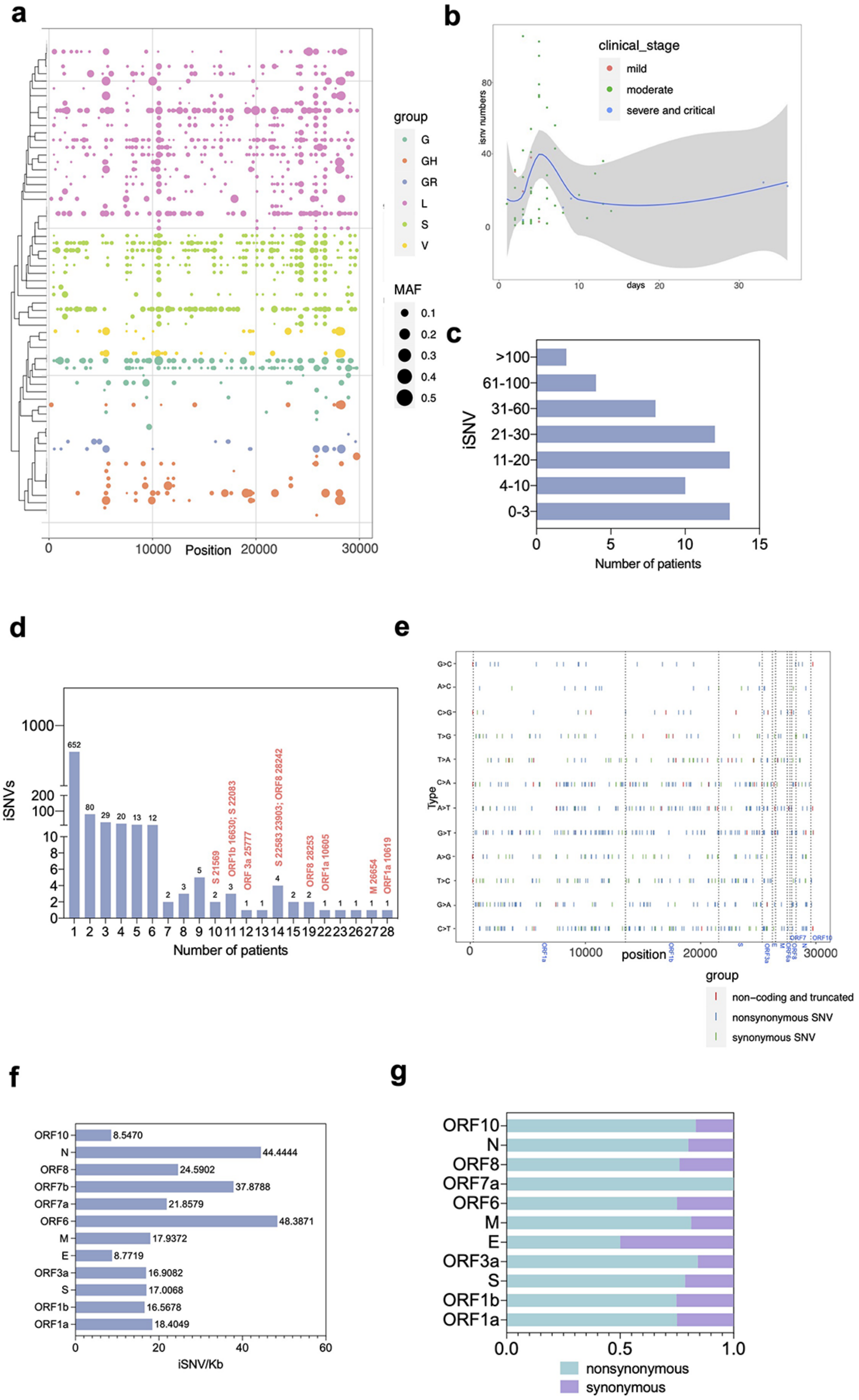


Figure 2. Distributions and characteristics of iSNVs. (A) Distribution of iSNVs among genome in enrolled samples. The circle size represented MAF, and showed the L, S, V, G, GR, and GH lineages, respectively. The genome position was shown above the scatter diagram. (B) Mapping of iSNVs along the disease course. The dots represent mild, moderate, severe, and critical stagers when sampling. The *x*- and *y*-axes showed disease onset days and numbers of iSNVs, respectively. (C) The numbers showed the number of iSNVs owned by each sample. (D) The bar diagram indicated the co-occurring iSNVs in populations. The iSNVs with detection in more than ten samples have been labeled, and the specific number of iSNVs was indicated above the bar. (E) The different substitutions of iSNV sites along the genome. The *x*- and *y*-axes represented genome size and substitution types, respectively. (F) The normalized iSNV/kb in different CDS regions. (G) The annotations (nonsynonymous, synonymous, noncoding, and truncation) in twelve types of substitutions. (H) The ratio of NS/S in indifferent CDS regions.

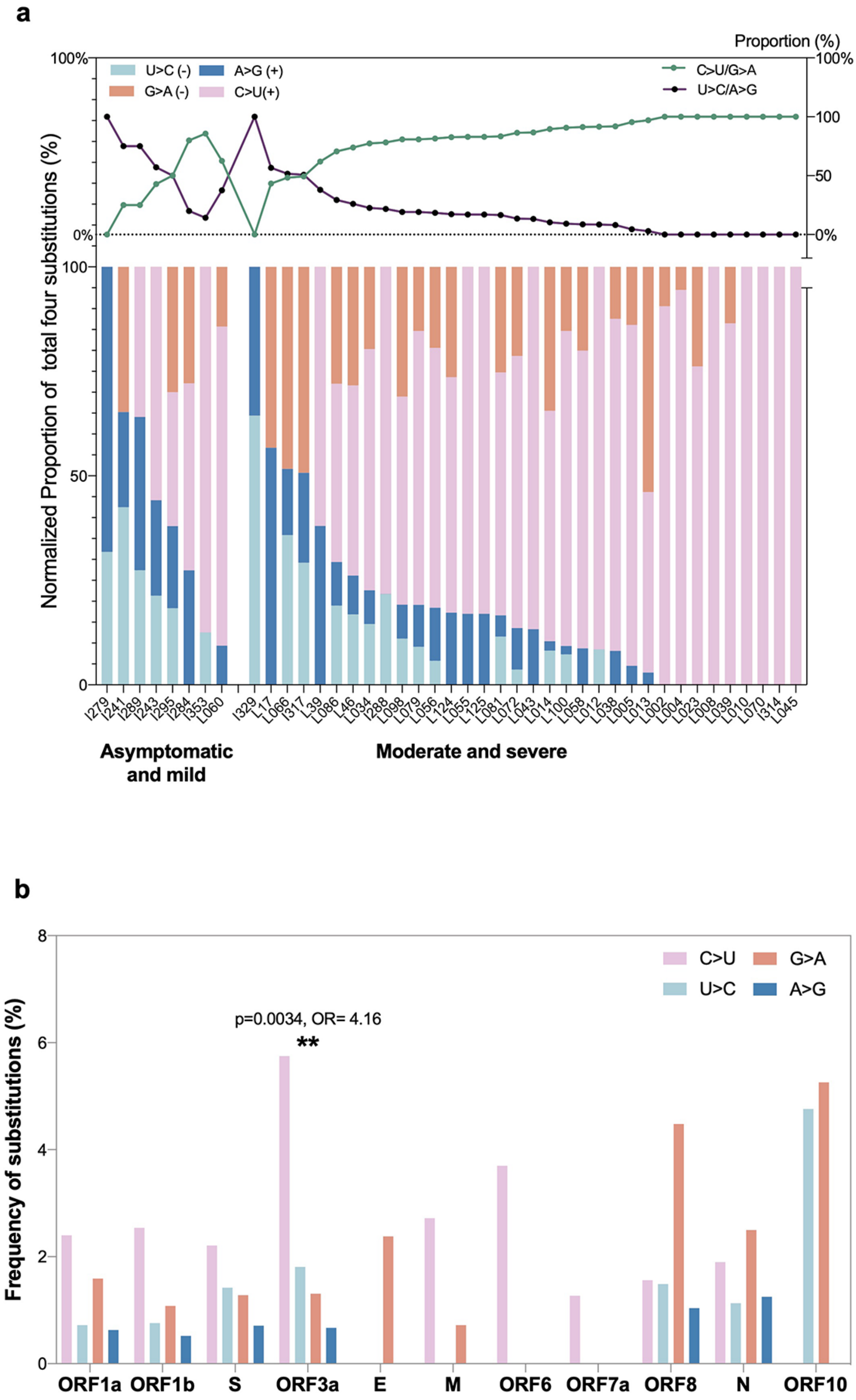


Figure 3. iSNV mutation C > U/G > A and U > C/A > G within patients. (A) The stripes showed the ratio of substitution C > U/G > A and the ratio of U > C/A > G in the sum of these four types. The line graph presented C > U/G > A minus U > C/A > G substitutions in each sample. (B) The plot presented the frequency of U > C/A > G and C > U/G > A substitutions in accordingly A, U, C, and G bases.

Table1. Potential escape from neutralization antibody in spike protein.

iSNV position in S gene	Amino acid in S protein	Average MAF	Frequency in populations	Significance	Target region of neutralization antibody	Reported in references
23016	G485V	0.066	3.28% (2/61)	Next to E484 mutation; next to REGN10933 and CoV-2832 escape mutation site F486	RBM	Puller, Neher, and Albert (2017) and Shanghai Clinical Treatment Expert Group for CoronaVirus Disease (2019, 2020)
22983	Q474R	0.100	1.64% (1/61)	Next to LY-CoV016 escape mutation site A475	RBD	Puller, Neher, and Albert (2017)
22992	S477N	0.101	1.64% (1/61)	Located at S477 mutation site	RBM	Puller, Neher, and Albert (2017)
23051	F497L	0.068	1.64% (1/61)	Next to CoV2-2499 escape mutation site 496	RBM	Shanghai Clinical Treatment Expert Group for CoronaVirus Disease (2019, 2020)
22689	T376I	0.052	1.64% (1/61)	Located at CoV2-2082 and CoV2-2094 escape mutation site 376	Core RBD	Shanghai Clinical Treatment Expert Group for CoronaVirus Disease (2019, 2020)
23076	Y505C	0.078	1.64% (1/61)	Located at P2C-1F11 epitope targeted site 505	RBD	Zhang et al. (2020)
23030	F490L	0.106	1.64% (1/61)	Located at CoV2-2479 and CoV2-2096 escape mutation site F490	RBD	Shanghai Clinical Treatment Expert Group for CoronaVirus Disease (2019, 2020)
22698	C379F	0.097	1.64% (1/61)	Next to CoV2-2677, CoV2-2082, CoV2-2094, and rCR3022 escape mutation site C378	Core RBD	Shanghai Clinical Treatment Expert Group for CoronaVirus Disease (2019, 2020)

in *ORF10* (Fig. 3B), while the C > U/G > A substitutions were abundant in *ORF3a*, *ORF8*, and *ORF10* (>5 per cent bases). Compared to the other three substitution types, C > U was higher in *ORF3a* ($P = 0.0034$, adjusted $P < 0.05$, odds ratio = 4.16).

iSNVs with potential to escape from neutralization antibody and T-cell recognition

A plethora of new variants that threaten to circumvent vaccines and existing natural immunity become less visible to the immune system. iSNV-induced amino acid change could escape from neutralization antibodies validated by previous studies, including those that have been proved for clinical use and in research development (Ge et al. 2021; Greaney et al. 2021; Starr et al. 2021) (Table 1). The G485V is next to E484 and F486, which are the escape mutations of REGN10933 and CoV-2832. Q474R and S477N are close to the escape mutation site A475 of LY-CoV016. For neutralization antibodies in an experimental stage, CoV2-2082 and CoV2-2094 shared an escape mutation site T376I. Furthermore, Y505C and F490L sites have been reported to escape from the P2C-1F11 epitope of several antibodies. F497L was adjacent to F496, the potential escape mutation of CoV2-2499. The mentioned mutations E484, S477, and F496 were all located at the receptor-binding motif (RBM) in receptor-binding domain (RBD), which is responsible for direct binding to Angiotensin-converting enzyme 2. All together suggest variants in spike that may affect receptor binding or neutralization by antibodies rise from iSNVs.

The other part of the adaptive immune response is cytotoxic T lymphocytes, and the rapid emergence of sequence variation within HLA-associated peptides provides clear evidence for host-driven immune selection during infection. Therefore, we examined whether common nonsynonymous mutations exist in patients carrying the same HLA allele. HLAs were highly polymorphic; HLA-A, HLA-B, and HLA-C possessed eleven, seventeen, and eleven genotypes among these patients, respectively, in the fifty-six successfully genotyped samples. Seventy-eight iSNV sites met the criteria for HLA analysis. Approximately 70.51 per cent (55/78) of nonsynonymous or frameshift iSNVs were found to be significantly correlated with specific HLA alleles (Fig. 4A). iSNV 10619 was both positively associated with HLA-C*01:02 and negatively associated with HLA-A*33:03 ($P = 0.034$, odds ratio = 3.67; $P = 0.036$, odds ratio = 0.19, respectively). Although our sample size is not large enough for statistical detection of some associations, eighteen iSNVs were characterized with corresponding HLA types following the Bonferroni correction for multiple tests, especially HLA-B*15:02, which was characterized with polymorphism at Position 6027 (odds ratio = 104, adjusted $P < 0.001$), and iSNV was specifically associated with HLA-B*15:01 (odds ratio = 104, adjusted $P < 0.001$) (Fig. 4C). The mixed linear regression considering age, sex, and onset days as covariates identified 196 pairs of HLA and iSNV with a *P*-value of <0.05 (Supplementary Dataset 2), and the above iSNVs and corresponding HLA with obvious *P*-value corrected by Bonferroni analyses had a high coefficient and a significant *P*-value as well.

iSNV existence in presymptomatic patients and correlation with clinical indicators

iSNVs were detected in samples of three asymptomatic and four presymptomatic patients (Fig. 5A), illustrating immune pressure before symptom onset. For adaptive immune response, we also identified several HLA-related iSNVs (adjusted $P < 0.05$) in asymptomatic patients as well, including nonsynonymous polymorphism at Positions 10056 and 10481 to be correlated with HLA-B*35:01. Surprisingly, the suspected immune escape–causing iSNV was recorded in a presymptomatic patient.

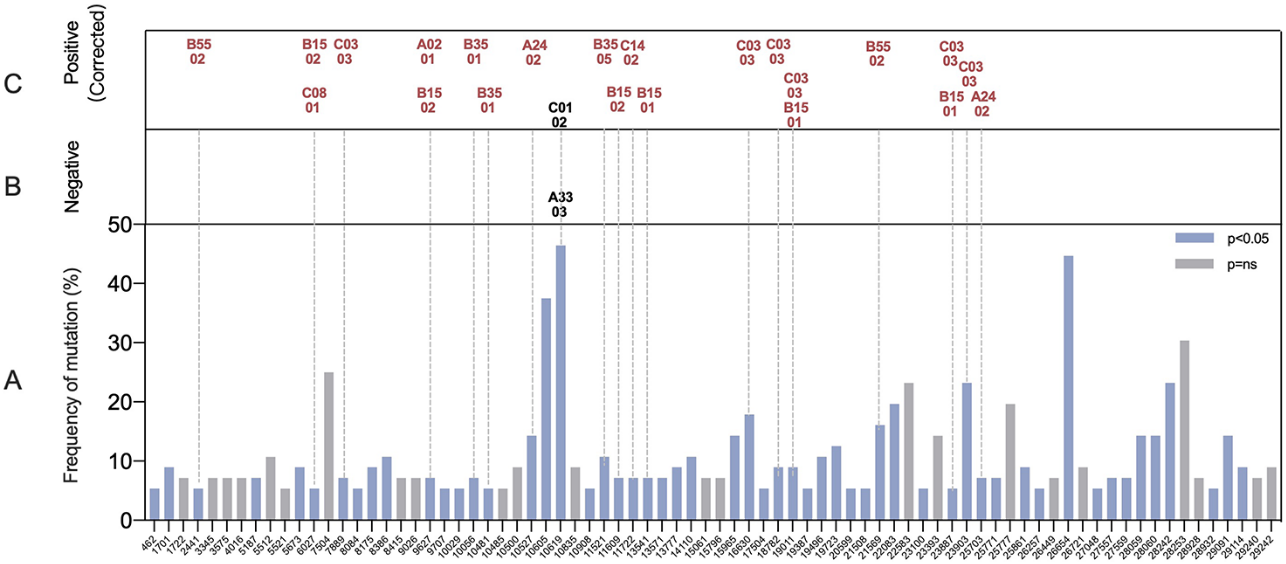


Figure 4. Map of polymorphism rate and HLA associations at nucleotide positions and affinity changes after mutation. The association between iSNV site and specific HLA allele type. (A) A total of seventy-eight nonsynonymous and truncated iSNV sites met the criteria of inclusion: the HLA allele type accounted for >5 per cent of total patients (≥3), and iSNV alteration existed in more than 5 per cent patients. The iSNVs with purple stripes were significantly correlated with the specific HLA allele type; (B) the negative relationship between iSNV 10619 and HLA-A*33:03; (C) the adjusted positive relationship (adjusted *P*-value < 0.05) between iSNV and HLA allele type.

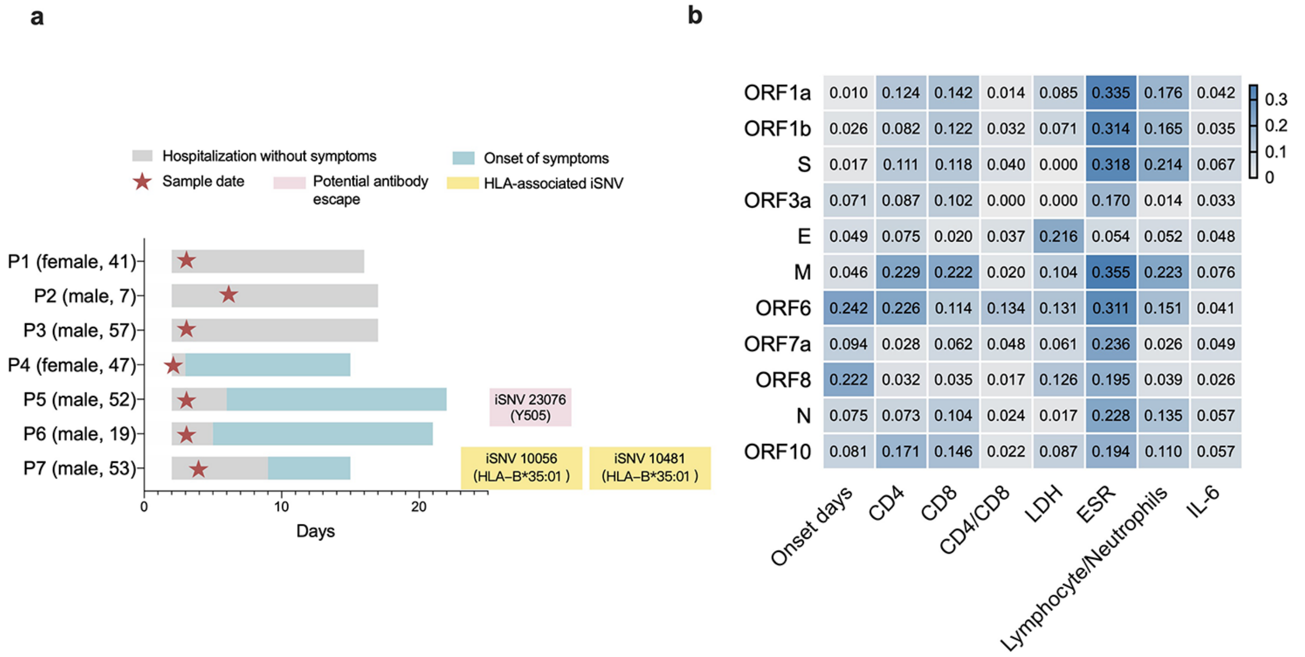


Figure 5. Correlations of iSNVs with clinical manifestations. (A) iSNVs were detected in three asymptomatic patients (P1, P2, and P3) and four presymptomatic patients (P4, P5, P6, and P7). (B) The heat map showed Pearson's correlation between clinical indicators in relationship to iSNVs. The marked numbers indicated the coefficient of correlation. *P < 0.05, **P < 0.01.

We illustrated that the iSNVs/kb in each coding region was unrelated to onset days, CD4 or CD8 cell count, CD4 to CD8 ratio, levels of lactate dehydrogenase, and interleukin-6 (Fig. 5B). No significant clinical indicator was found to be correlated with iSNV frequencies of CDS regions.

Discussion

To sketch how the virus is mutating within individuals as the disease progresses, we have profiled the within-host diversity of SARS-CoV-2 in the full spectrum of the COVID-19 disease course using the tools of in-depth sequencing and a set of cross-sectional samples collected at time points representing disease phases from presymptomatic to severe. Our study found that most iSNVs detected were in an individualized pattern, iSNVs exhibited C > U/G > A enrichment in moderated and severe patients, and some iSNVs located at the receptor-binding domain of the spike protein, which was consistent with immune selection including interferon-induced deaminase and T- and B-cell immune responses. Finally, these immune escape variants, which could be spread, were associated with geographical or demographic limiting factors, such as biased HLA genotype distribution in various countries or ethnicity.

In line with previous reports, we found that iSNVs were frequently detected in the majority of samples and their number represented a clear temporal trend pacing with disease progression and the level of viral replication, similar to previous reports (Tonkin-Hill et al. 2021). At present, evidence indicates that Type I and III interferon (IFN) responses (Broggi et al. 2020; Major et al. 2020) are most important to combat the viral replication; however, the magnitude and exact kinetics of this response in severe vs mild COVID-19 patients remain controversial (Schultze and Aschenbrenner 2021). We discovered intra-host mutations that were not stochastically positioned throughout the genome and a significant bias toward C–U/G–A. Among these mutations, synonymous mutations could be involved in the regulation at the steps of the viral life cycle, such as transcription, splicing, messenger RNA transport, and translation. Compared with synonymous mutations, nonsynonymous mutations are more dominant across each of the open reading frames, indicating that all proteins are under selective pressure and adaptation to a human host. Among these viral proteins, ORF8 is less conservative due to its trivial functions (Chen et al. 2020) and is important in human host adaptation. Furthermore, *ORF8* was found to be critical to hijacking the antiviral interferon pathway (Zinzula 2021).

Concerning a preferred usage of intra-host mutations, interferon-induced restriction factors adenosine deaminase acting on RNA or apolipoprotein B mRNA editing enzyme family (Samuel 2001; Harris and Liddament 2004; Olson, Harris, and Harki 2018; Di Giorgio et al. 2020; Pfaller, George, and Samuel 2021) could exclusively deaminate an adenine or cytosine of the viral RNA, which consequently initiates the A–G/U–C or C–U/G–A transitions, imprinting in the remnant viral sequences for evading degradation. Interestingly, we found a disease phase-specific mutational pattern of milder disease–featured A–G/U–C and severe status–related C–U/G–A, which might reflect the activation of different interferon-induced pathways, and involve in SARS-CoV-2 replication, contributing to the viral diversity. Another possible explanation is the difference in the anatomic site where the virus is produced. Infection initiates at the upper respiratory tract and exacerbates once disseminating into the lower respiratory tract. The IFN expression in COVID-19 was reported to vary according to location, viral load, and disease severity. For example, mildly ill patients owned a higher expression of IFN-λ1 and IFN-λ3 in the upper airways, while IFN-λ2 and IFN-$\alpha\beta$ were abundant in the lower airways (Sposito et al. 2021). For this mutation pattern, it might lie in base oxidation, especially in the C–U/G–A transitions, and other factors including viral load with consequent RT bias could also lead to mutation bias.

This finding suggests that adequate activation of an appropriate interferon pathway is critical to restricting viral replication and prophylaxis use; the right type of interferon might prevent disease exacerbation, whereas the wrong type might accelerate disease progression. The cause-and-effect relationship between the type of RNA editing and disease severity warrants further animal experiments to investigate the role of RNA editing in the pathological process.

Along with the publicly sequenced database that has been deposited for 1 year into the SARS-CoV-2 pandemic, increasing mutations have proved to cause immune escape. With several iSNV sites located at a spike protein, closely monitoring immune escape variants through iSNVs would aid in designing immunogens and developing neutralizing antibodies for preventing and treating immune escape variants.

Although our cohort size is small, there are several HLA-associated iSNVs, whether nonsynonymous or synonymous, which become circulating in a small population. This HLA-related selection pressure favors the preservation of mutations *in vivo* and could be made evident for T-cell immune response against SARS-CoV-2 at the population level (Moore et al. 2002). The divergence of geographic defining strains could be caused by a demographically and geographically localized population, which is predominant with some types of HLA (Bhattacharya et al. 2007).

We illustrated a high level of quasispecies complexity of SARS-CoV-2 within hosts. The intra-patient virus genetic diversity was reported to increase after treatment (Kemp et al. 2021). In immunocompromised patients or severe patients who underwent a long course of disease, the complexity and mutations of the virus might increase, even becoming a new virus variant (Weigang et al. 2021; Cele et al. 2022). As an indicator of inflammatory response during infection, erythrocyte sedimentation rates were found to be more likely correlated with the number (frequencies) of iSNVs than other clinical laboratory tests, which was reported to show an increase in the patients with severe disease (Ghahramani et al. 2020; Lapic, Rogic, and Plebani 2020).

Several limitations should be noted in the interpretations of our projections. First, we failed to sequence sufficient swab samples since respiratory samples were disposed of according to biosafety regulatory guidelines, and only high-quality leftover viral RNA samples were used for our testing. Second, we only successfully enrolled one patient with consecutive samples; thus, the iSNV variations during an individual's disease course could not be observed. Third, the iSNV mutational signatures are robustly present; the detailed relationship warrants further experiments. Additionally, the number of asymptomatic patients was limited in our study; thus, the iSNV pattern or immune selection in these patients could not be addressed comprehensively.

In conclusion, we demonstrated that intra-host SARS-CoV-2 single-nucleotide variants were mostly in an individualized pattern and revealed the immune response spectrum on iSNVs associated with disease severity.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 82041010, 92169212, and 82161138018), Shanghai Key Laboratory of Infectious Diseases and Biosafety Emergency Response (grant number 20dz2260100), Key Discipline Construction Plan from Shanghai Municipal Health Commission (GWV-10.1-XK01), and Shanghai Municipal Science and Technology Major Project (HS2021SHZX001).

Conflict of interest: None declared.

References

Allen, T. M. et al. (2004) 'Selection, Transmission, and Reversion of an Antigen-Processing Cytotoxic T-lymphocyte Escape Mutation in Human Immunodeficiency Virus Type 1 Infection', *Journal of Virology*, 78: 7069–78.

Bhattacharya, T. et al. (2007) 'Founder Effects in the Assessment of HIV Polymorphisms and HLA Allele Associations', *Science*, 315: 1583–6.

Broggi, A. et al. (2020) 'Type III Interferons Disrupt the Lung Epithelial Barrier upon Viral Recognition', *Science*, 369: 706–12.

Cele, S. et al. (2021) 'Escape of SARS-CoV-2 501Y.V2 from Neutralization by Convalescent Plasma', *Nature*, 593: 142–6.

——— et al. (2022) 'SARS-CoV-2 Prolonged Infection during Advanced HIV Disease Evolves Extensive Immune Escape', *Cell Host & Microbe*, 30: 154–62.

Challen, R. et al. (2021) 'Risk of Mortality in Patients Infected with SARS-CoV-2 Variant of Concern 202012/1: Matched Cohort Study', *BMJ (Clinical Research Ed.)*, 372: n579.

Chen, S. et al. (2020) 'Extended ORF8 Gene Region Is Valuable in the Epidemiological Investigation of Severe Acute Respiratory Syndrome-Similar Coronavirus', *The Journal of Infectious Diseases*, 222: 223–33.

Collier, D. A. et al. (2021) 'Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA Vaccine-Elicited Antibodies', *Nature*, 593: 136–41.

Davenport, M. P. et al. (2008) 'Rates of HIV Immune Escape and Reversion: Implications for Vaccination', *Trends in Microbiology*, 16: 561–6.

Davies, N. G. et al. (2021) 'Increased Mortality in Community-Tested Cases of SARS-CoV-2 Lineage B.1.1.7', *Nature*, 593: 270–4.

Di Giorgio, S. et al. (2020) 'Evidence for Host-Dependent RNA Editing in the Transcriptome of SARS-CoV-2', *Science Advances*, 6: eabb5813.

Dolan, P. T., Whitfield, Z. J., and Andino, R. (2018) 'Mechanisms and Concepts in RNA Virus Population Dynamics and Evolution', *Annual Review of Virology*, 5: 69–92.

Emary, K. R. W. et al. (2021a) 'Efficacy of ChAdOx1 nCoV-19 (AZD1222) Vaccine against SARS-CoV-2 Variant of Concern 202012/01 (B.1.1.7): An Exploratory Analysis of a Randomised Controlled Trial', *The Lancet*, 397: 1351–62.

——— et al. (2021b) 'Efficacy of ChAdOx1 nCoV-19 (AZD1222) Vaccine against SARS-CoV-2 Variant of Concern 202012/01 (B.1.1.7): An Exploratory Analysis of a Randomised Controlled Trial', *The Lancet*, 397: 1351–62.

Garcia-Beltran, W. F. et al. (2021) 'Multiple SARS-CoV-2 Variants Escape Neutralization by Vaccine-Induced Humoral Immunity', *Cell*, 184: 2372–83.

Ge, J. et al. (2021) 'Antibody Neutralization of SARS-CoV-2 through ACE2 Receptor Mimicry', *Nature Communications*, 12: 250.

Ghahramani, S. et al. (2020) 'Laboratory Features of Severe vs. Non-severe COVID-19 Patients in Asian Populations: A Systematic Review and Meta-analysis', *European Journal of Medical Research*, 25: 30.

Giorgi, E. E. et al. (2013) 'Modeling Sequence Evolution in HIV-1 Infection with Recombination', *Journal of Theoretical Biology*, 329: 82–93.

Greaney, A. J. et al. (2021) 'Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain That Escape Antibody Recognition', *Cell Host & Microbe*, 29: 44–57 e9.

Hacisuleyman, E. et al. (2021) 'Vaccine Breakthrough Infections with SARS-CoV-2 Variants', *New England Journal of Medicine*, 384: 2212–8.

Harris, R. S., and Liddament, M. T. (2004) 'Retroviral Restriction by APOBEC Proteins', *Nature Reviews Immunology*, 4: 868–77.

Huang, C. et al. (2020) 'Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China', *The Lancet*, 395: 497–506.

Kemp, S. A. et al. (2021) 'SARS-CoV-2 Evolution during Treatment of Chronic Infection', *Nature*, 592: 277–82.

Korber, B. et al. (2020) 'Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus', *Cell*, 182: 812–27.e19.

Lapic, I., Rogic, D., and Plebani, M. (2020) 'Erythrocyte Sedimentation Rate Is Associated with Severe Coronavirus Disease 2019 (COVID-19): A Pooled Analysis', *Clinical Chemistry and Laboratory Medicine (CCLM)*, 58: 1146–8.

Li, Q. et al. (2021) 'SARS-CoV-2 501Y.V2 Variants Lack Higher Infectivity But Do Have Immune Escape', *Cell*, 184: 2362–71.

Liu, J. et al. (2021) 'Role of Mutational Reversions and Fitness Restoration in Zika Virus Spread to the Americas', *Nature Communications*, 12: 595.

Lu, R. et al. (2020) 'Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding', *The Lancet*, 395: 565–74.

Lythgoe, K. A. et al. (2021) 'SARS-CoV-2 Within-Host Diversity and Transmission', *Science*, 372: eabg082.

Madhi, S. A. et al. (2021) 'Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant', *New England Journal of Medicine*, 384: 1885–98.

Major, J. et al. (2020) 'Type I and III Interferons Disrupt Lung Epithelial Repair during Recovery from Viral Infection', *Science*, 369: 712–7.

Moore, C. B. et al. (2002) 'Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level', *Science*, 296: 1439–43.

Ni, M. et al. (2016) 'Intra-Host Dynamics of Ebola Virus during 2014', *Nature Microbiology*, 1: 16151.

Olson, M. E., Harris, R. S., and Harki, D. A. (2018) 'APOBEC Enzymes as Targets for Virus and Cancer Therapy', *Cell Chemical Biology*, 25: 36–49.

Pfaller, C. K., George, C. X., and Samuel, C. E. (2021) 'Adenosine Deaminases Acting on RNA (ADARs) and Viral Infections', *Annual Review of Virology*, 8: 239–64.

Popa, A. et al. (2020) 'Genomic Epidemiology of Superspreading Events in Austria Reveals Mutational Dynamics and Transmission Properties of SARS-CoV-2', *Science Translational Medicine*, 12: eabe2555.

Puller, V., Neher, R., and Albert, J. (2017) 'Estimating Time of HIV-1 Infection from Next-Generation Sequence Diversity', *PLOS Computational Biology*, 13: e1005775.

Samuel, C. E. (2001) 'Antiviral Actions of Interferons', *Clinical Microbiology Reviews*, 14: 778–809.

Schultze, J. L., and Aschenbrenner, A. C. (2021) 'COVID-19 and the Human Innate Immune System', *Cell*, 184: 1671–92.

Shanghai Clinical Treatment Expert Group for CoronaVirus Disease 2019 (2020) 'Comprehensive Treatment and Management of Corona Virus Disease 2019: Expert Consensus Statement from Shanghai', *Chinese Journal of Infectious Diseases*, 38: 134–8.

Sposito, B. et al. (2021) 'The Interferon Landscape along the Respiratory Tract Impacts the Severity of COVID-19', *Cell*, 184: 4953–6.

Starr, T. N. et al. (2021) 'Prospective Mapping of Viral Mutations That Escape Antibodies Used to Treat COVID-19', *Science*, 371: 850–4.

Su, S. et al. (2016) 'Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses', *Trends in Microbiology*, 24: 490–502.

Supasa, P. et al. (2021) 'Reduced Neutralization of SARS-CoV-2 B.1.1.7 Variant by Convalescent and Vaccine Sera', *Cell*, 184: 2201–11.

Tegally, H. et al. (2021) 'Detection of a SARS-CoV-2 Variant of Concern in South Africa', *Nature*, 592: 438–43.

Tonkin-Hill, G. et al. (2021) 'Patterns of Within-Host Genetic Diversity in SARS-CoV-2', *eLife*, 10: e66857.

Valesano, A. L. et al. (2021) 'Temporal Dynamics of SARS-CoV-2 Mutation Accumulation Within and Across Infected Hosts', *PLoS Pathogens*, 17: e1009499.

Wang, P. et al. (2021) ‘Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7’, *Nature*, 593: 130–5.

Wang, Y. et al. (2021) ‘Intra-host Variation and Evolutionary Dynamics of SARS-CoV-2 Populations in COVID-19 Patients’, *Genome Medicine*, 13: 30.

Weigang, S. et al. (2021) ‘Within-Host Evolution of SARS-CoV-2 in an Immunosuppressed COVID-19 Patient as a Source of Immune Escape Variants’, *Nature Communications*, 12: 6405.

Wu, F. et al. (2020) ‘A New Coronavirus Associated with Human Respiratory Disease in China’, *Nature*, 579: 265–9.

Wu, K. et al. (2021) ‘Serum Neutralizing Activity Elicited by mRNA-1273 Vaccine’, *New England Journal of Medicine*, 384: 1468–70.

Zhang, X. et al. (2020) ‘Viral and Host Factors Related to the Clinical Outcome of COVID-19’, *Nature*, 583: 437–40.

Zhou, P. et al. (2020) ‘A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin’, *Nature*, 579: 270–3.

Zinzula, L. (2021) ‘Lost in Deletion: The Enigmatic ORF8 Protein of SARS-CoV-2’, *Biochemical and Biophysical Research Communications*, 538: 116–24.