frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 19 May 2022
doi: 10.3389/fmicb.2022.824217

SARS-CoV-2 Within-Host and *in vitro* Genomic Variability and Sub-Genomic RNA Levels Indicate Differences in Viral Expression Between Clinical Cohorts and *in vitro* Culture

Jessica E. Agius[1,2,3†], Jessica C. Johnson-Mackinnon[1,2,3†], Winkie Fong[1,2], Mailie Gall[3,4], Connie Lam[1], Kerri Basile[4], Jen Kok[1,4], Alicia Arnott[1,3,4], Vitali Sintchenko[1,2,3,4‡] and Rebecca J. Rockett[1,2,3‡]*

[1] Centre for Infectious Diseases and Microbiology Public Health, Westmead Hospital, Institute for Clinical Pathology and Medical Research, Westmead, NSW, Australia, [2] Faculty of Medicine and Health, Sydney Medical School, University of Sydney, Sydney, NSW, Australia, [3] Sydney Institute for Infectious Diseases, Westmead Institute for Medical Research, Westmead, NSW, Australia, [4] Centre for Infectious Diseases and Microbiology Laboratory Services, NSW Health Pathology – Institute of Clinical Pathology and Medical Research, Westmead, NSW, Australia

OPEN ACCESS

Edited by:
Helene Dutartre, UMR 5308 Centre International de Recherche en Infectiologie (CIRI), France

Reviewed by:
Pragya Dhruv Yadav, ICMR-National Institute of Virology, India
Anne Piantadosi, Emory University, United States

***Correspondence:**
Jessica C. Johnson-Mackinnon Jessica.JohnsonMackinnon@health.nsw.gov.au

† These authors have contributed equally to this work and share first authorship

‡ These authors have contributed equally to this work and share last authorship

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *29 November 2021*
Accepted: *18 April 2022*
Published: *19 May 2022*

Citation:
Agius JE, Johnson-Mackinnon JC, Fong W, Gall M, Lam C, Basile K, Kok J, Arnott A, Sintchenko V and Rockett RJ (2022) SARS-CoV-2 Within-Host and in vitro Genomic Variability and Sub-Genomic RNA Levels Indicate Differences in Viral Expression Between Clinical Cohorts and in vitro Culture. Front. Microbiol. 13:824217. doi: 10.3389/fmicb.2022.824217

Background: Low frequency intrahost single nucleotide variants (iSNVs) of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) have been increasingly recognised as predictive indicators of positive selection. Particularly as growing numbers of SARS-CoV-2 variants of interest (VOI) and concern (VOC) emerge. However, the dynamics of subgenomic RNA (sgRNA) expression and its impact on genomic diversity and infection outcome remain poorly understood. This study aims to investigate and quantify iSNVs and sgRNA expression in single and longitudinally sampled cohorts over the course of mild and severe SARS-CoV-2 infection, benchmarked against an *in vitro* infection model.

Methods: Two clinical cohorts of SARS-CoV-2 positive cases in New South Wales, Australia collected between March 2020 and August 2021 were sequenced. Longitudinal samples from cases hospitalised due to SARS-CoV-2 infection (severe) (n = 16) were analysed and compared with cases that presented with SARS-CoV-2 symptoms but were not hospitalised (mild) (n = 23). SARS-CoV-2 genomic diversity profiles were also examined from daily sampling of culture experiments for three SARS-CoV-2 variants (Lineage A, B.1.351, and B.1.617.2) cultured in VeroE6 C1008 cells (n = 33).

Results: Intrahost single nucleotide variants were detected in 83% (19/23) of the mild cohort cases and 100% (16/16) of the severe cohort cases. SNP profiles remained relatively fixed over time, with an average of 1.66 SNPs gained or lost, and an average of 4.2 and 5.9 low frequency variants per patient were detected in severe and mild infection, respectively. sgRNA was detected in 100% (25/25) of the mild genomes and

92% (24/26) of the severe genomes. Total sgRNA expressed across all genes in the mild cohort was significantly higher than that of the severe cohort. Significantly higher expression levels were detected in the spike and the nucleocapsid genes. There was significantly less sgRNA detected in the culture dilutions than the clinical cohorts.

Discussion and Conclusion: The positions and frequencies of iSNVs in the severe and mild infection cohorts were dynamic overtime, highlighting the importance of continual monitoring, particularly during community outbreaks where multiple SARS-CoV-2 variants may co-circulate. sgRNA levels can vary across patients and the overall level of sgRNA reads compared to genomic RNA can be less than 1%. The relative contribution of sgRNA to the severity of illness warrants further investigation given the level of variation between genomes. Further monitoring of sgRNAs will improve the understanding of SARS-CoV-2 evolution and the effectiveness of therapeutic and public health containment measures during the pandemic.

Keywords: within-host diversity, variants, sub-genomic RNA, SARS-CoV-2, iSNV, evolution, dynamics, COVID-19

INTRODUCTION

As the ongoing Coronavirus disease 2019 (COVID-19) pandemic unfolds across the globe, several variants of concern (VOC) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for COVID-19 have emerged globally. Given the rapid worldwide spread, and continuing functional evolution of the virus, the real-time tracking of variants has become increasingly important (Wu et al., 2020; Zhou et al., 2020).

Severe acute respiratory syndrome coronavirus 2 genomes, approximately 30,000 bases in length are organised in a series of open reading frames (ORFs) consisting of four structural and 16 non-structural proteins (Arya et al., 2021). The 5′ end contains a leader sequence followed by the 5′ UTR and two large polyproteins (ORF1a and ORF1b) which encode all the non-structural proteins. These two ORFs are followed by the structural and accessory proteins, which include the spike protein (S), ORF3a, envelope protein (E), membrane protein (M), ORFs 6, 7a, 7b, 8, nucleocapsid (N), and ORF10 capped off by the 3′ UTR and poly-A tail (Naqvi et al., 2020; Nomburg et al., 2020). Following cytoplasmic entry into a host cell, the 1a and 1b large polyproteins are directly translated from genomic RNA (gRNA), while the remaining structural proteins are translated from sgRNA intermediaries (Sola et al., 2015; Song et al., 2019). The subgenomic RNA or sgRNA transcripts are produced through a complex mechanism involving discontinuous or "paused" transcription, followed by an RNA-dependant RNA polymerase (RdRp) template switch during negative-strand RNA synthesis (Parker et al., 2021). The resulting nested set of negative sense RNAs serve as templates for the transcription of positive strands, forming mRNAs for translation of distinct proteins. sgRNAs contain a common leader sequence (65–90 nt) derived from the 5′ untranslated region, in addition to a transcription regulatory sequence (TRS) immediately adjacent to the 5′ ORF of the structural and accessory genes, responsible for the pausing of virus transcription during negative strand synthesis (Sola et al., 2015).

The sgRNA of SARS-CoV-2 encode the structural proteins S, E, M, and N, in addition to the several accessory proteins 3a, 6, 7a, 7b, 8, and 10 (Davidson et al., 2020; Kim et al., 2020). It is not well understood role sgRNA expression plays during infections. There have been reports that the detection of sgRNAs in clinical samples indicates an active viral infection, and expression levels correlate strongly to the severity of symptoms (Wong et al., 2021). However, others report that sgRNA expression is not a reliable indicator of viral replication (Alexandersen et al., 2020). Overall, understanding of the role of sgRNA during infections and quantification of sgRNA expression during SARS-CoV-2 infections is limited.

Compared to DNA viruses, the replication of RNA viruses is typically associated with a high error rate due to the lack of sufficient proofreading activities during genome replication (Domingo and Holland, 1997). However, coronaviruses employ a highly conserved proofreading exoribonuclease encoded by non-structural protein 14 (nsp14) which enhances the fidelity of RNA synthesis (Graepel et al., 2017). Despite this mechanism, the mutation rate of SARS-CoV-2 is 1–2 mutations per month and is generally higher than DNA viruses (Smith et al., 2014; Day et al., 2020; Nakagawa and Miyazawa, 2020). Additionally, coronaviruses have the propensity to recombine and generate extensive and diverse recombination products, particularly within the spike region of the genome (Wells et al., 2021). At an inter-host level, newly emerging viruses acquire adaptive mutations to enhance replication, modulate the host response, and facilitate effective transmission. However, the intra or within-host variability of RNA viruses is associated with the quasi-species concept, leading to multiple diverse circulating quasi-species of varying frequencies linked through mutation (Karamitros et al., 2020; Ramazzotti et al., 2020). The quasi-species collectively contribute functional characteristics at the population level, and in combination with the genetic profile of the host, can influence viral phenotype and adaptive capabilities (Stone et al., 2006). Since most of the immune escape and adaptive mechanisms of SARS-CoV-2 involve intra-cellular interactions, it is expected that SARS-CoV-2 evolves through intra-host selective

pressure (Kumar et al., 2020), highlighting the capacity for the development of genetically different SARS-CoV-2 viruses within the same host.

Higher within-host diversity of viral RNA pathogens can be associated with increasing viral virulence and antigenic variability (Stone et al., 2006), exacerbated disease severity and clinical outcome, immune escape (Nowak et al., 1991), and drug resistance (Johnson et al., 2008). Given these effects, the real-time monitoring of SARS-CoV-2 variants at the within-host level is important. Monitoring within-host diversity *via* the detection of intrahost single nucleotide variants (iSNVs) can inform genomic epidemiology (Lythgoe et al., 2021), and provide early indications of diagnostic PCR dropouts (Sapoval et al., 2020). The ability to predict mutations under positive selection, particularly functionally important and emerging mutations informs public health surveillance and the design of therapeutics (Popa et al., 2020; Tonkin-Hill et al., 2021; Wang et al., 2021).

Currently, variant analyses for SARS-CoV-2 focus primarily on mutations occurring at the consensus-level (single consensus sequence for each infected person), which represent the dominant variants within infected individuals (Lythgoe et al., 2021). However, genomic epidemiology of SARS-CoV-2 has revealed the capacity for viral mutations to emerge within an individual host (Al Khatib et al., 2020; Karamitros et al., 2020; Lythgoe et al., 2021; Wang et al., 2021), and so an understanding of the complete underlying within-host diversity at the population-level proves imperative.

This study aimed to investigate the consistency and timing of iSNV detection over the course of clinical and *in vitro* SARS-CoV-2 infections using longitudinally collected specimens from the same patient. We examined iSNV profiles shared by SARS-CoV-2 lineages during an epidemiological characterised outbreak in Sydney, Australia. We investigated if these iSNVs were more frequently detected in severe illness, and if they develop over the time course of COVID-19 disease. We also measured changes in sgRNA to investigate if sgRNA is associated with increased genomic diversity during the course of individual infections.

MATERIALS AND METHODS

Clinical Specimens

Clinical specimens were collected within a time span from 5 days prior to the onset of COVID-19 symptoms to 23 days post symptom onset. If the date of symptom onset was unknown, the date of sample collection from the first positive specimen was considered the date of symptom onset. A total of 90 clinical specimens RT-qPCR positive for SARS-CoV-2 were examined. The majority of specimens were from the upper respiratory tract with nasopharyngeal swabs (n = 78), lower respiratory tract samples included bronchoalveolar lavages (n = 10), and sputum (n = 2) representing SARS-CoV-2 cases diagnosed in NSW, Australia between March 2020 and August 2021 (**Supplementary Figure 1**). The cohorts consisted of cases admitted to the intensive care unit ± intubation (classified as severe disease) (n = 19 cases, 48 specimens), and mild cases which recovered as outpatients (classified as mild disease) (n = 32, 42 specimens). Nasopharyngeal swabs in Universal Transport Media (UTM) which were RT-qPCR negative for SARS-CoV-2 (n = 4) and collected in the study period were also included as negative controls. All specimens were de-identified and stored at −80°C. Following genome and variant level quality filtering the final cohorts consisted of 16 severe cases (n = 26 swabs) and 23 mild cases (n = 25 swabs) (**Supplementary Figure 1**).

Ethics Statement

Governance and human ethics approval for clinical metadata and use of specimens from cases positive for SARS-CoV-2 in New South Wales was obtained by Western Sydney Local Health District Human Research Ethics Committee (2020/ETH02426 and 2020/ETH02282).

Cultured Isolates

Daily sampling was conducted to determine the fifty percent tissue culture infective dose ($TCID_{50}$) of SARS-CoV-2 viral stocks from three lineages of SARS-CoV-2 (Lineage A – referred to as Wuhan, Beta – B.1.351, and Delta – B.1.617.2). Briefly, clinical samples confirmed to be SARS-CoV-2 positive by RT-qPCR (n = 3) were sequenced to determine the infecting SARS-CoV-2 lineage before being used for inoculation. CostarÒ 24-well clear tissue culture-treated multiple well plates (CorningÒ, Corning, NY, United States) were seeded at 40% confluence with Vero C1008 cells (Vero 76, clone E6, Vero E6) (ATCC® CRL-1586™) in Dulbecco's minimal essential medium (DMEM, Lonza Bioscience, Alpharetta, GA, United States), and supplemented with 9% foetal bovine serum (FBS, HyClone, Cytiva, Sydney, NSW, Australia). Culture media was changed within 12 h and contained 1% FBS, and 1% antimicrobials including amphotericin B deoxycholate (25 μg/mL), penicillin (10,000 U/mL), and streptomycin (10,000 μg/mL) (Lonza, Basel, Switzerland) to inhibit microbial overgrowth. The plates were inoculated with 200 μL of serially diluted virus stock (1×10^{-2} to 1×10^{-6}) in triplicate. Cells were incubated at 37°C in 5% CO_2 for 4 days (days 0–3), and were sealed with AeraSealä Film (Excel Scientific, Inc., Victorville, CA, United States) to minimise evaporation, spillage, and well-to-well cross-contamination. Visual inspection for the presence of cytopathic effect (CPE) was undertaken daily and 100 μL of supernatant was sampled from a single well to quantify viral replication every 24 h. Mycoplasma testing was routinely conducted to exclude contamination of the culture media and cell line. The presence of CPE along with confirmation and quantification of SARS-CoV-2 viral load was performed by RNA extraction and RT-qPCR of culture supernatant daily (days 0–3) after viral load quantification RNA extracts were stored at −80°C or immediately used to prepare libraries for sequencing. All culture samples were identified *via* the following naming convention: <lineage> -<day><dose> (i.e., A-D1-03; Lineage A, sampling day one, dilution 1×10^{-3}). All SARS-CoV-2 culture was performed under level 3 biosafety conditions within

a NSWHP physical containment level 4 laboratory (PC4) accredited facility.

RNA Extraction

Total RNA was isolated from mild clinical samples using the RNeasy Mini Kit (Qiagen) with minor modifications. A total volume of 200 μL of UTM/culture supernatant was added to 600 μL of RLT buffer and vortexed briefly. Next, 800 μL of 70% ethanol was added and mixed well by pipetting. The solution was then loaded on the RNeasy column in successive aliquots until the entire volume of the sample was extracted. RNA was eluted in 32 μL and stored at −80°C. Clinical samples from cases in severe ($n = 48$), mild ($n = 42$), and culture supernatants ($n = 34$) were extracted using the BioRobot EZ1 and EZ1 Virus Mini Kit v2.0 (Qiagen, Valencia, CA, United States) in PC4 facilities. An input volume of 100 μL was used and RNA was eluted into 60 μL as per the manufacturer's instructions. The culture supernatants were extracted prior to removal of RNA from the PC4 facility.

RT-qPCR of Severe Acute Respiratory Syndrome Coronavirus 2

The SARS-CoV-2 viral RNA was detected and quantified using a previously described RT-qPCR targeting the N-gene (Rahman et al., 2020; Lam et al., 2021). Ten-fold serial dilutions (1×10^6 to 1×10^2 copies/μL) of the commercially available synthetic RNA control reference strain (Wuhan-1 strain, TWIST Biosciences) containing six non-overlapping fragments of the SARS-CoV-2 reference sequence (NCBI GenBank accession MN908947.3) was used to generate a standard curve and quantify SARS-CoV-2 viral load. The N-gene SARS-CoV-2 RT-qPCR was employed to determine the viral load of positive specimens sequenced as part of this study. The absence of SARS-CoV-2 in negative samples was confirmed by RT-qPCR.

Virome Enrichment, Capture, and Sequencing

Viral enrichment of clinical extracts, *in vitro* culture isolates, and the synthetic SARS-CoV-2 positive control spiked into negative culture supernatant was performed using the Illumina RNA Prep and Enrichment with the Respiratory Viral Oligo Panel (RVOP) v2 (Illumina, United States). This probe-based capture technique was selected as it was designed to generate near full length SARS-CoV-2 genomic sequences with even coverage in mutagenic regions. RNA denaturation, first and second strand cDNA synthesis, cDNA tagmentation, library clean-up and normalisation were performed according to the manufacturer's instructions. Individual libraries were pooled in 3-plex reactions for probe hybridisation based on each samples SARS-CoV-2 viral load. The final probe hybridisation step was held overnight at 58°C. The enriched library was purified, and the concentration and fragment size were quantified using the Qubit™ 1x dsDNA High Sensitivity Assay (Thermo Fisher Scientific, United States), and Agilent High Sensitivity D1000 ScreenTape assay on the Agilent 4200 Tapestation (Agilent, Germany), respectively. The libraries were sequenced using 2 × 74 bp runs on the Illumina MiniSeq™ or iSeq (Illumina, United States) and multiplexed with the aim of producing 2×10^6 raw reads per library.

Bioinformatic Analysis and Clustering

The raw sequence reads were subjected to an in-house quality control procedure prior to downstream analysis. The reads were demultiplexed and quality trimmed using Trimmomatic version 0.36 (minimum read quality score of 20, leading/trailing quality of 5). Reference mapping and consensus calling was performed using iVar version 1.2.1. Reads were mapped to the reference SARS-CoV-2 Wuhan-Hu-1 genome (NCBI GenBank accession: MN908947) and unmapped reads discarded. Mapping coverage and depth across the genome and structural and accessory genes was determined using MOSDEPTH version 0.2.9. Only genomes with > 80% coverage over a 100× depth in all variant positions were included in further analyses. A consensus sequence was generated (map quality > 20), and the 5′ (first 55 nt) and 3′ (last 100 nt) UTR regions were masked due to the known suboptimal sequencing quality of these regions. All genomes that passed filtering were submitted to NCBI GenBank (PRJNA633948). Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN)[1] was used to infer SARS-CoV-2 lineages. In addition, isolates within lineages were assigned a genomic cluster denoting their position within NSW outbreak cases based on SNP distance from an epidemiological defined index case (Rockett et al., 2020).

Variant Filtering and Analysis

The frequency and positions of variants (SNPs and iSNVs) in all samples were determined using Varscan version 2.3.9. SNPs were defined as mutations with a read frequency of ≥ 0.9. Variants with a read frequency between 0.05 and 0.9 were defined as low frequency variants. Variants occurring in the TWIST control were highlighted as potential artefacts (MN908947 genome reference positions 5765, 5766, 1107, 11082, 12413, 12926, 23652, and 26433), and those associated with mis-mapping at the ends of insertion or deletion events in B.1.351 and B.1.617.2 (11288, 22029, 22034, 22287, 23598, 23607, 23609, 23616, 28248, and 28249) were identified and excluded from downstream analyses. Changes in the iSNV profiles were determined between longitudinal sampling from single cases, cases from the same household (known transmission events), within severe and mild cohorts, across lineages, and genomic clusters.

Detection of Severe Acute Respiratory Syndrome Coronavirus 2 Subgenomic RNAs

Subgenomic RNA analyses was conducted using Periscope (Parker et al., 2021). Briefly, Periscope distinguished sgRNA reads based on the 5′ leader sequences being directly upstream from each genes transcription. The sgRNA counts were then normalised into a measure termed sgRPTL, by dividing the sgRNA reads by the mean depth of the gene of interest and multiplying by 1,000.

[1] https://github.com/hCoV-2019/pangolin

Phylogenetic Analysis

Consensus SARS-CoV-2 genomes were processed using the Nextstrain Conda environment. Augur (bioinformatics tool) v13.0.0,[2] and Auspice (open-source visualisation tool) v2.30.0[3] were employed for analysis and visualisation (Hadfield et al., 2018). As a comparison, a representative global subset of SARS-CoV-2 genomes curated by Nextstrain between September 2019 and August 2021 was included in our phylogenetic analysis.

Statistical Analysis

Statistical significance ($p \leq 0.05$) was determined using the Mann–Whitney test for difference between means on variables which contained at least five data points (iSNV/SNP counts and read frequencies, sgRNA counts and read frequencies).

RESULTS

Phylogenetic Analysis

Overall, 84 high-quality SARS-CoV-2 genomes from the clinical cohorts and culture dilutions passed genome- and variant-level filtering and were included in the final analysis. The genomes included representatives from lineages A, A.2.2, B.1, B.1.1, B.1.617.2, B.6, and D.2, as illustrated in the SARS-CoV-2 global phylogeny (**Figure 1**).

Cohort Sequencing Results

Following quality control, 25 SARS-CoV-2 consensus genomes were recovered from the mild (n = 23, average sample per case = 1.1) and 26 from the severe (n = 16 cases, average sample per case = 1.6) cohorts (**Table 1**). There were no significant differences between age and sex across the mild and severe cohorts, however, a significant difference between the age ranges (p = 0.16) was noted (**Table 1**). Thirty-three SARS-CoV-2 consensus genomes were sequenced from 34 culture specimens of varying sample dilutions and time intervals (one genome did not pass quality filtering and was excluded). High depth genomes were produced across all cohorts and the median depth achieved was not significantly different. The median depth for the severe cohort was 2,021×, mild 928×, Lineage A 2,964×, Beta VOC 3,408×, and Delta VOC 2,529× (**Supplementary Figure 2B**).

Severe Acute Respiratory Syndrome Coronavirus 2 Viral Load

A range of SARS-CoV-2 viral loads were detected in each cohort (**Supplementary Figure 2A**). The median severe viral load was 516,643 copies (range: 151,246,026–2,512 copies), and the median mild viral load was 457,284 copies (range: 95,727,865–668.7 copies). Within the culture cohorts, lineage A median viral load was 1,408,340 copies (range: 18,976,383–29.4 copies), Beta median viral load was 560,453.7 copies (range: 9,026,044–130.2 copies), and Delta median viral load was 300,232.9 copies (range: 3,107,421–424.5 copies) (**Supplementary Figure 2A**). There was no significant difference between the viral loads across cohorts.

[2]https://github.com/nextstrain/augur

[3]https://github.com/nextstrain/auspice

Severe Acute Respiratory Syndrome Coronavirus 2 Pango Lineages and Epi-Clusters

A wide variety of pango lineages were defined across the clinical cohorts (**Table 1**). The majority of cases were designated to lineage B.1.617.2 or Delta VOC (n = 20, cases; P601, P603, P608, P611, P612, P614, P615, P622, P624, P626, P629, P604, P606, P618, P620, P621, PG27, and P628), followed by B.1 (n = 13 cases P0332, P0570, P0495, P0676, P1384, P1434, and P1494), D.2 genomes (n = 11, cases; P0340, P0341, P0417, P0642, P0858, P1149, P2099, P2152, P1498, and P1551) and B.6 (n = 4, case P0105). Single genomes from lineage A.2.2 (case P1727), lineage A genome (case P1811), and lineage B.1.1 (P1020) (**Figure 1** and **Table 1**). Within some lineages, genomes were designated distinct genomic epi-clusters based on SNP distances and known epidemiological links from public health investigations. Within Lineage B.1.617.2 there was one cluster (NSW 130, n = 20), B.1 had 3 clusters (NSW 17.5, n = 3; NSW 9, n = 7, and singletons n = 2), B.6 had 1 cluster (NSW 3.1, n = 4), and D.2 had 2 clusters (NSW 33.1, n = 6; NSW 33, n = 7). Within the epi-clusters known transmission events between members within a household were also captured; group 1 lineage B.1 (n = 3 cases, 3 samples), group 2 lineage D.2 (n = 2 cases, 2 samples), group 3 lineage B.1.617.2 (n = 2 cases, 3 samples), group 4 lineage B.1.617.2 (n = 2 cases, 2 samples), and group 5 lineage B.1.617.2 (n = 2 cases, 2 samples) (**Supplementary Figure 1B**).

Frequency of Intrahost Single Nucleotide Variants Across Clinical Cohorts and Cultured Genomes

Low frequency iSNVs were detected in 92% (24/26) of the severe (**Figure 2A**), 80% (20/25) of the mild (**Figure 2B**), and 100% (33/33) of the culture samples (**Figure 2C**). Longitudinal samples collected from the same patient were collected over a mean of 6.36 days (range: 0–11) post-symptom onset compared to 1–3 days after inoculation in cultured specimens. Overall, there were 91 iSNVs detected in the severe cohort (median number of iSNVs per sample per case = 3, median frequency = 0.106), and 129 iSNVs detected in the mild cohort (median number of iSNVs per sample per case = 4, median frequency = 0.085) (**Figure 2B**). The frequency of iSNVs per SARS-CoV-2 gene between severe and mild cohorts was significantly different only for the S gene (p = 0.0209) (**Figure 3**). Of the culture specimens there were 60 iSNVs (median frequency per specimen = 0.373) for the lineage A cultures, the Beta culture contained 39 iSNVs (median frequency = 0.345), and the Delta culture contained 39 iSNVs (median frequency = 0.214) (**Figure 2C**).

There was no significant difference between the median numbers of iSNVs per case when comparing the severe and mild cohorts. However, a significantly higher median read frequency of iSNVs between the severe and mild cohorts (p = 0.023) was observed. There was also a significant difference between the mean frequency of iSNVs between the severe and culture cohorts (p = 0.016), and the mild and culture cohorts (p = 0.00001). There was also a significantly higher

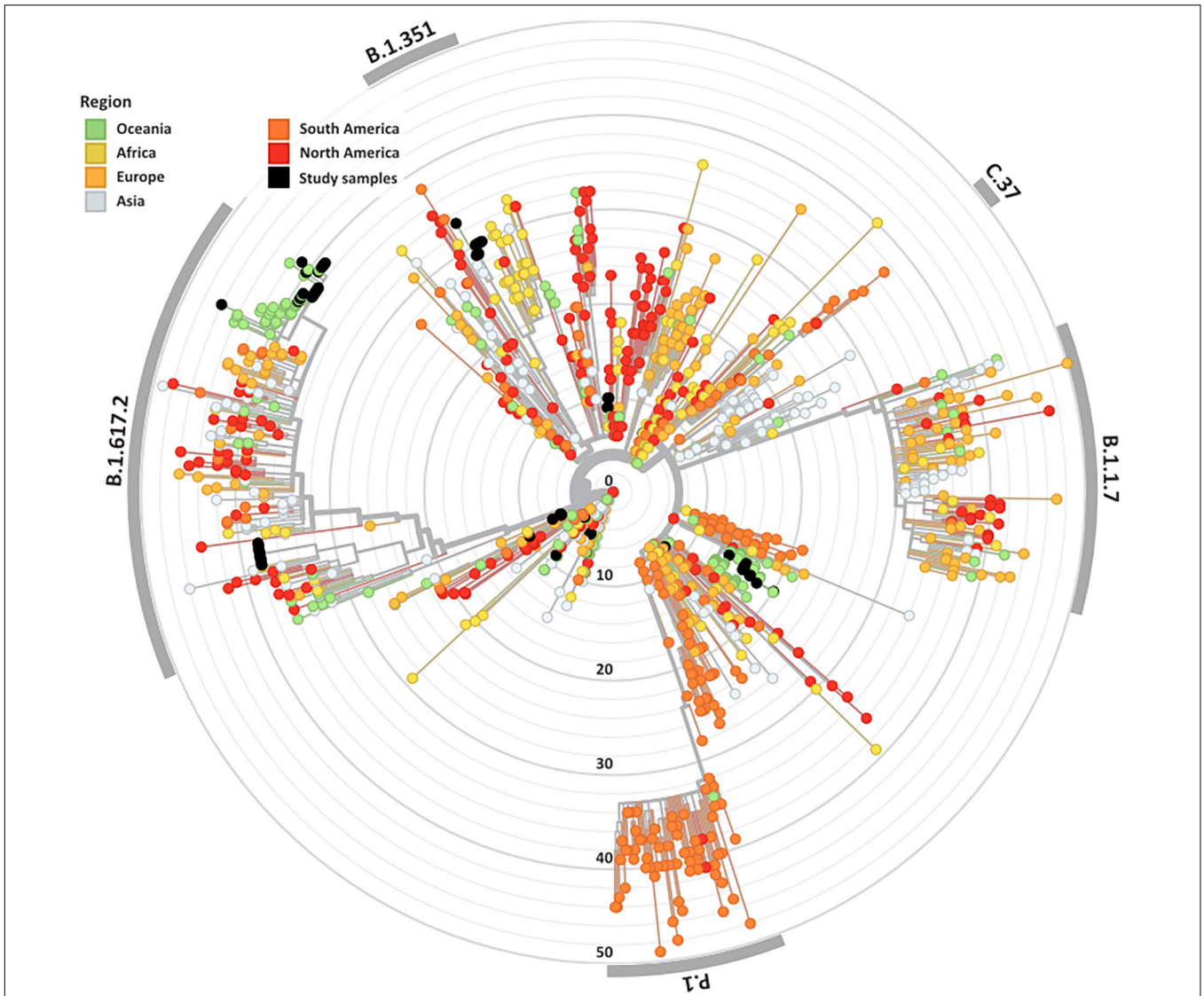

FIGURE 1 | Genomic and epidemiological diversity of clinical and culture SARS-CoV-2 sequences. The SARS-CoV-2 phylogeny was constructed in a local version of Nextstrain (auspice.us) and shows the evolutionary relationship of SARS-CoV-2 genomes sequenced within this study (black dots) and a subsample of publicly available SARS CoV-2 genomes collected globally (n = 1,079). The tree branch lengths represent divergence, with sequences rooted relative to early samples collected in Wuhan, China (Wuhan-Hu-1/2019).

number of iSNVs in the lineage A compared to Delta culture dilutions (p = 0.00001). The highest number of SNPs and iSNVs/per SARS-CoV-2 genome in cases for all cohorts were found in ORF1ab, S, ORF8, and N genes. Differences between the counts of iSNVs between severe and mild cohorts were significant only at the ORF1ab gene (p = 0.0357) (**Figure 4**). Across all cohorts, the number of non-synonymous SNP and iSNV mutations were greater than synonymous, indicating positive selection. The average non-synonymous/synonymous mutation ratio (Ka/Ks) per genome per cohort for iSNVs and SNPs was 2.30 and 3.33 in culture, 1.47 and 5.29 in severe cases, and 1.65 and 2.84 in the mild cohort. The SNP ka/ks ratio was highest across all cohorts when compared to the iSNV ka/ks ratio.

We further investigated iSNVs encoding non-synonymous structural changes within the S gene. A total of 5, 10, and 3 non-synonymous iSNVs were detected in the severe, mild and *in vitro* cultured genomes respectively. Interestingly the mutation S:K129N was detected 4 times in different cases in the severe cohort. The iSNV S:T76I was detected in four culture genomes from the infection with the Beta variant and H655Y was also noted in culture genomes.

Clinical Longitudinal Samples Measuring Intra-Host Diversity

Our cohort contained four severe cases with longitudinal samples (P0105, P0332, P608, and P615), with an average of 3.25 samples

TABLE 1 | Demographics of SARS-CoV-2 positive cases within the severe and mild cohorts.

Cohort	Total cases	Total samples	Lineages (*n* samples)	Gender M:F	Age median range
Severe disease	16	26	B.1 (7), B.6 (4), B.1.617.2 (13), A.2.2 (1), A (1)	8M:8F	Median = 65; Range 27–94
Mild	23	25	D.2 (11), B.1 (6), B.1.617.2 (7), B.1.1 (1).	6M:15F	Median = 56; Range 12–71
Total	39	51	A (1), A.2.2 (1), B.1 (13), B.1.1 (1), B.1.617.2 (20), B.6 (4), D.2 (11)	14M:23F	Median = 58; Range 12–94

M, male; F, female.

per case (**Figure 2A**). The SNP profile of each patient remained relatively fixed over time, with an average of 1.66 SNPs gained or lost (range 1–2) compared to the initial genome collected for each case. There was a general trend observed with an increase in the number of iSNVs over the course of the earlier lineage infections. For severe case P0105, a single iSNV (genome position 6310 nt, read frequency < 0.12) remained consistent across three longitudinal samples which were collected 7–10 days post symptom onset. However, this iSNV was lost at 11 days post symptom onset. Despite the loss of the iSNV on day 11 (position 6,310 nt), this sample contained iSNVs (frequencies < 0.3) at eight new locations within the genome; five in ORF1ab, two in spike, and one in nucleocapsid. Severe case P0332 had no consistent iSNVs throughout their infection, although an average of three iSNVs (range: 0–13) were detected per sample. Days nine and ten post symptom onset (last two sampling points) contained the greatest diversity of iSNVs and were present at 12 positions on day nine (ORF1ab, $n = 9$; S, $n = 1$; M, $n = 1$; and NC, $n = 1$), and three positions on day ten (ORF1ab $n = 3$). For Delta genomes, there were fewer conserved iSNVs which tended to be lost rather than gained. For case P608 one iSNV and one low frequency deletion event were retained from 1 day to 11 days post symptom onset (NC, $n = 1$; ORF8, $n = 1$). In addition, one iSNV present at 1 day post symptom onset converted to a SNP at 11 days post symptom onset (NC, $n = 1$) and one high frequency deletion event at day one converted to a low frequency deletion event by day 11 (S, $n = 1$). The final longitudinal sample, severe case P0615, retained one iSNV and two low frequency deletion events from 3 days post symptom onset to 8 days post symptom onset (ORF1ab, $n = 1$; S $n = 1$; NC, $n = 1$) with one iSNV lost at day 8 (NC, $n = 1$).

Our cohort also contained five epi-linked family groups. There were no shared iSNVs between cases in groups 1 and 2 (Lineages B.1 and D.2). Group's three to five were all in lineage B.1.617.2, and of those groups, groups four and five had no shared iSNVs between cases. In group three there was one iSNV shared amongst all cases and samples (ORF1ab, $n = 1$), and one iSNV that was present in one case and gained later in the course of infection in the other case (NC, $n = 1$). However, in all three groups there were two shared deletion events (S, $n = 1$; NC, $n = 1$).

Culture Intra-Experiment Genomic Diversity

In culture, SNPs remained relatively stable over sampling time-points and across lineages and dilutions with zero SNPs in lineage A, 15 SNPs and four high frequency deletion events in lineage Beta, and one SNP and two high frequency deletion events in lineage Delta were lost or gained over the three-day experiment (**Supplementary Figures 3, 4**). There was overall greater diversity of iSNVs occurring within Beta and Delta when compared to lineage A (60 and 39 compared to 20 with median frequencies of 0.288 and 0.153 compared to 0.0663). There was a significant difference between the median frequencies of lineage A compared to Beta ($p = 0.00001$) and Delta ($p = 0.00214$) but no significant difference between Beta and Delta. Large indels were noted in the lineage A spike gene (22,203–22,213 nt and 23,595–23,585 nt) and remained at relatively high frequency (> 0.4) over time and dilutions. Lineage-specific deletions in Beta (position 22,270–22,280 nt) and Delta (position 28,240–28,247 nt) cultures were maintained over the time course of the experiment but remained at frequencies less than 0.9. Interestingly, iSNVs that occurred *in vitro* were not present at baseline (inoculum sample) 88% (8/9), 75% (6/8), and 71% (5/7) of the time in lineage A, Beta, and Delta variants, respectively. These iSNVs were often 96% (45/47) below a frequency of 0.3.

Within-Host Genomic Diversity Between Lineages and Epidemiologically Defined Transmission

Within the clinical cohort, there were higher numbers of iSNVs detected in the B.1, $n = 12$ genomes (39 positions along the genome, with an average of 1.025 iSNVs/position), D.2, $n = 11$ genomes (45 positions along the genome, with an average of 1.666 iSNVs/position), and B.1.617.2 lineages, $n = 13$ genomes (29 positions along the genome, with an average of 1.714 iSNVs/position). Lineages with < 5 genomes were excluded. Within those lineages iSNVs were most commonly shared between genomic clusters in the D.2 lineage and were only shared once between a singleton and cluster 9 lineage B.1. Within-host variants were shared between lineages B.1 and D.2 at eight positions (ORF1ab – 269 nt, 3,761 nt, 5,372 nt, 6,604 nt, 11,511 nt; ORF3a – 25,408 nt; M – 26,545 nt; and NC – 27,870 nt). There was one occasion where a SNP in lineage B.6 was a 0.05 frequency iSNV in another (D.2).

Subgenomic RNA

Subgenomic RNA was present in almost all genomes from the severe and mild cohorts (24/26, 25/25), although at low levels, median 1.7% of the read depth compared to gRNA (range; 0.02–52% of depth/average gene depth). The N-gene was the highest sgRNA transcript detected in the severe cohort measured by median sgRPTL (severe 41.916; mild 76.25) followed by the M gene (22.309; 22.549), ORF3 (24.873; 32.389), ORF7a (15.867; 24.259), ORF8 (6.227; 14.386), and S (4.353; 13.422) (**Figure 5**). Only a small number of genomes expressed sgRNA for ORF1ab (0.849 $n = 1$; 2.87 $n = 2$), ORF6 (0.386 $n = 1$; 6.623 $n = 11$) and E (0.311 $n = 1$; 7.335 $n = 6$), no sgRNA was detected for ORF10. There was a significant difference ($p = 0.00804$) between the sgRNA across all genes of the severe and mild cohorts – comparing individual genes where both had a sample size greater than 5, there were no significant

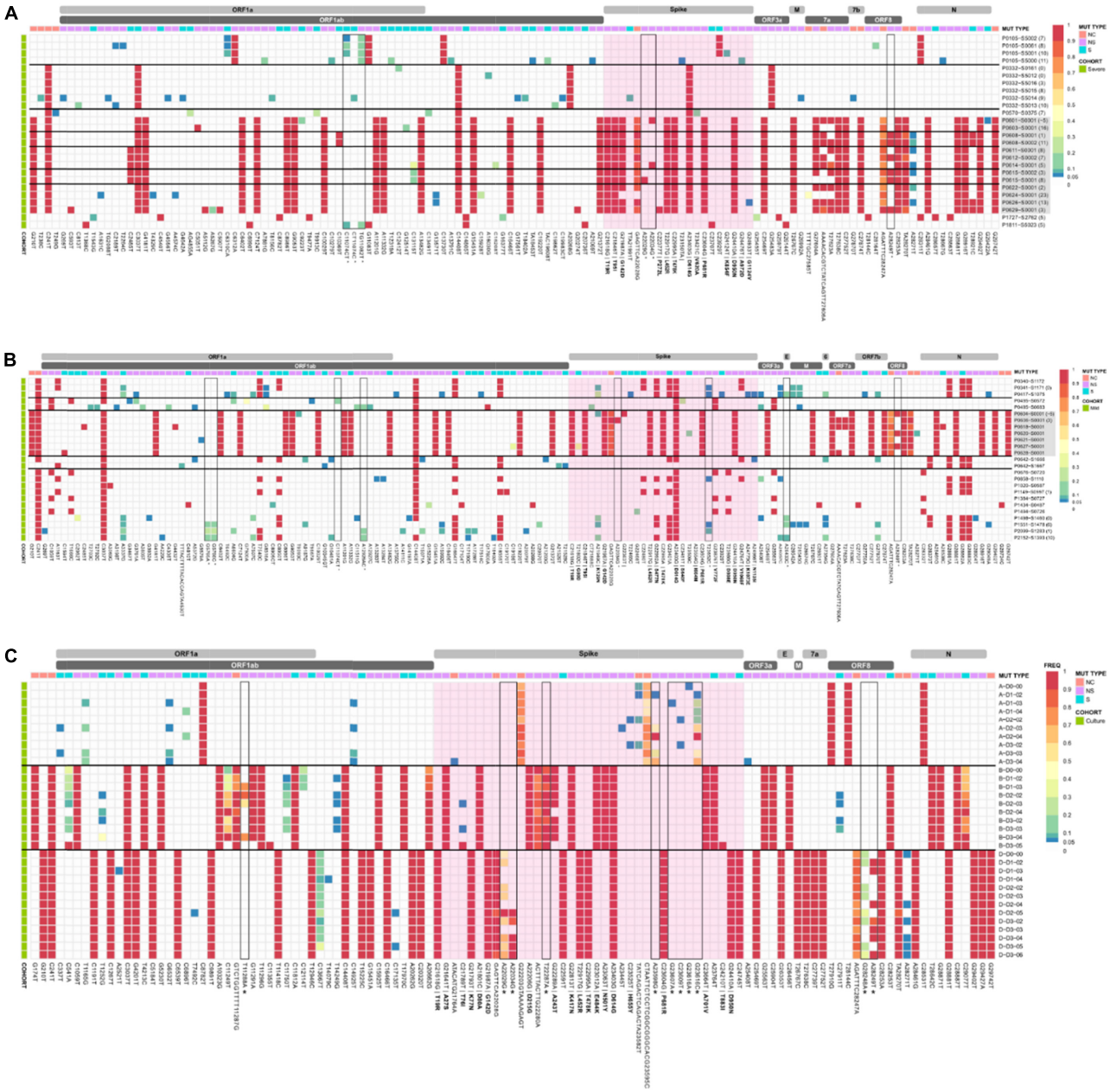

FIGURE 2 | iSNV and SNP frequencies and the corresponding synonymous and non-synonymous mutations across the SAR-CoV-2 genomes for the **(A)** severe and **(B)** mild cohorts, and **(C)** culture dilutions. A frequency of ≥ 0.9 was considered a SNP (red), frequencies below 0.05 were not included in the analysis. Identified problematic sites are outlined in black and denoted with an * on the *x*-axis. Amino acid changes for non-synonymous mutations in the spike region are in bold. The spike gene is highlighted in pink. SARS-CoV-2 Delta lineages are highlighted in grey. Horizontal black bars group samples from the same patient. Descriptors A, B, and D in panel **(C)** refer to Lineage A, B.1.351, and B.1.617.2, respectively. Bracketed numbers indicate the date of sample collection post-symptom onset.

differences except in the N (p = 0.0114) and S (p = 0.0128) genes (**Figure 5**). Although there were trends in sgRNA between lineages, the sample sizes per lineage were insufficient to determine significance (**Supplementary Table 1**). Within the culture cohort, sgRNA was also present in the majority of genomes (30/33) at low levels with a median of 0.8% of the total reads compared to gRNA (range 0.02–10.4% depth/average gene depth). Overall, sgRNA expression was significantly less than in both the severe (p = 0.0002) and mild (p = 0.0001) cohorts (**Supplementary Figure 5**).

There were some interesting trends in sgRNA expression during the three-day *in vitro* SARS-CoV-2 infections (**Supplementary Figure 6**). Within the Lineage A culture dilutions, the levels of sgRNA increase at each sampling time point except for a reduction in sgRNA expression on day three in the S and E genes The inoculum expressed higher levels

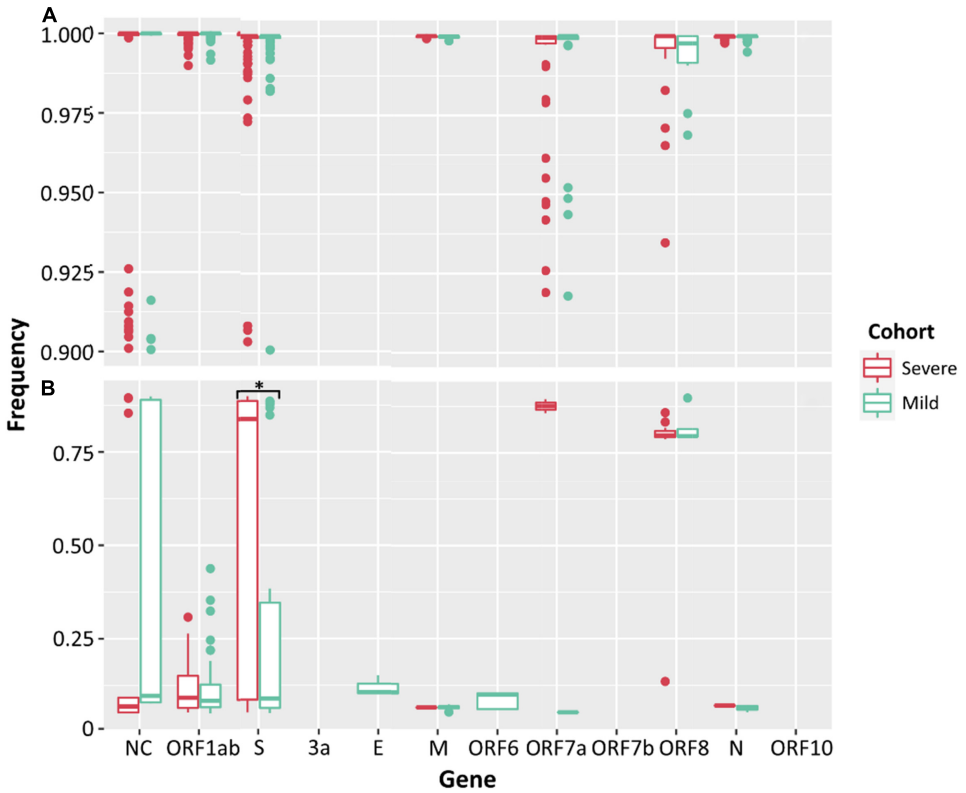

FIGURE 3 | Frequencies of SNPs **(A)** and iSNVs **(B)** by SARS-CoV-2 gene for severe (red) and mild (green) cohorts. Frequencies ≥ 0.9 were considered SNPs. Problematic sites are not included. NC signifies non-coding region of the genome. Statistical significance ($p \leq 0.05$) is denoted by (*). The frequency of iSNVs in the spike gene of SARS-CoV-2 were significantly different between the severe and mild cohorts.

of sgRNA than were seen at day one. However, the sgRNA expression exceeded that of the inoculum on days two and three in all but the N gene, where sgRNA expression only surpassed the inoculum levels at day three. For lineages Beta and Delta, the inoculum sgRNA levels remained higher than subsequent sampling points in all genes except E, ORF7a and ORF 8 for Beta and E and ORF7a for Delta (**Supplementary Figure 6**). Although sgRNA expression has shown to increase in culture over time, this theory requires further evaluation with biological replicates.

DISCUSSION

Intra-Host Variation

Since the emergence of SARS-CoV-2 in 2019, the virus has been persistently acquiring polymorphisms. In some cases, these changes have resulted in VOCs that pose a greater threat to public health, in the form of increased transmissibility, more severe disease, and evasion of current prevention strategies (Wu et al., 2020). The evolutionary rate of SARS-CoV-2 has been estimated at 1.1×10^{-3} substitutions/site/year (Smith et al., 2014; Day et al., 2020), however, some variants, such as Alpha (Lineage B.1.1.7) and Delta (Lineage B.1.617.2) have established a higher percentage of nucleotide changes (Duchene et al., 2020; Rambaut et al., 2020). Within-host variants within the viral population of an affected host are thought to be a contributing factor to the emergence of mutations. These iSNVs have been described across the entirety of the genome and can affect both non-protein coding and protein coding genes (Karamitros et al., 2020; Armero et al., 2021; Lythgoe et al., 2021). In this study we documented that the majority of iSNVs detected in the clinical samples were present at low frequencies (average: 0.0795), and were not consistently present in longitudinal samples. To accurately quantify iSNVs we sampled longitudinal samples collected from six cases and three known transmission events. Cases P0105 and P0332 had severe disease and contained variants present on the day of symptom onset at reference position 12,412 nt (synonymous, gene ORF1a I4049I) that reverted, but subsequent samples 10 days post symptom onset detected iSNVs at position 19,862 nt (non-synonymous, gene ORF1b A2132V). Where consistent iSNVs were present across longitudinal samples, they were at low frequencies that were subsequently lost. For example, patient P0105 at position 6,310 nt (insertion, gene ORF1a 2015), iSNVs were present at seven, eight, and 10 days post symptom onset, but were no longer detectable at 11 days post symptom onset (final sampling timepoint). This appears to be consistent with previous studies that demonstrate transmission bottlenecks, where the majority of iSNVs are eventually lost and not transmitted onto new individuals (Lythgoe et al., 2021; Valesano et al., 2021). However, there may be implications for

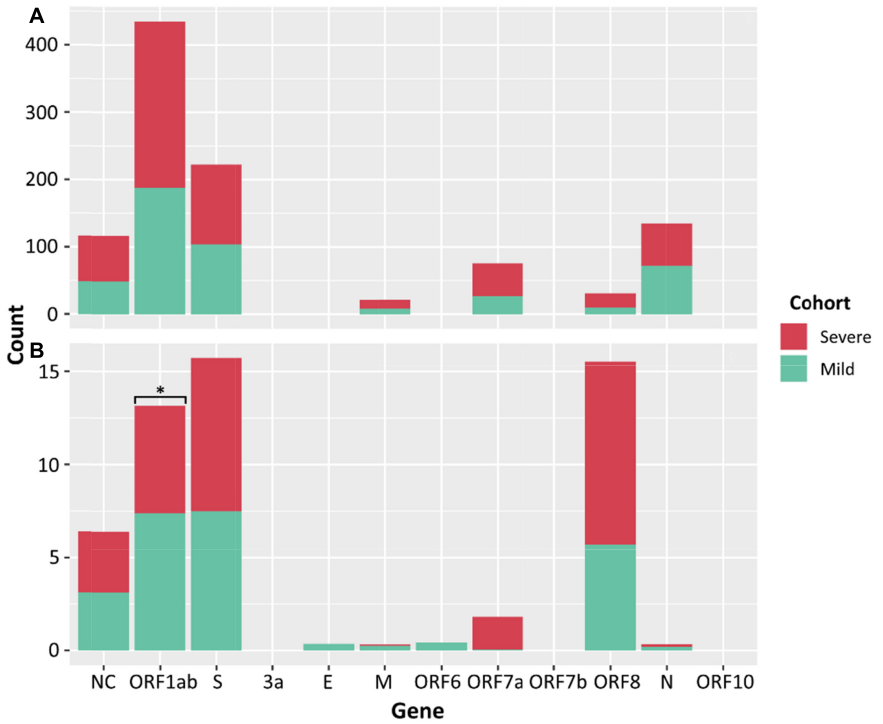

FIGURE 4 | Counts of SNPs **(A)** and iSNVs **(B)** by SARS-CoV-2 gene for severe (red) and mild (green) cohorts. Frequencies ≥ 0.9 were considered SNPs. Problematic sites are not included. Statistical significance ($p \leq 0.05$) is denoted by (*) and NC signifies non-coding region of the genome. The number of iSNVs in the ORF1ab gene of SARS-CoV-2 were significantly different between the severe and mild cohorts.

transmission and potential emergence of VOCs if iSNVs evolve into SNPs. Our findings indicate that the number of iSNVs tends to increase the longer the infection progresses, particularly in the ORF1ab, S, and N genes. It is therefore possible that the longer an individual remains infectious, likely in immunocompromised individuals, the higher the likelihood of the accumulation and transmission of functional iSNVs.

It is important to note that the presence and/or absence of iSNVs and their distribution across the genome changes from patient to patient and within cases over time. There are some consistently polymorphic positions, for example, position 23,929 nt in lineage B.6 is a consensus level change but is commonly an iSNV in the D.2 lineage. Lineage D.2 and B.1 in particular, had high levels of iSNVs respectively, and high levels of shared iSNV positions in the genome. All samples from lineages D.2 and B.1 were collected from cases of local transmission, whereas infection in cases associated with lineages A and B.6 were acquired overseas. This supports the recent report by Armero et al. (2021) of high levels of carryover iSNV diversity in the ORF1ab, S, and N genes within an outbreak in Victoria, Australia.

In contrast when SARS-CoV-2 was grown within *in vitro* culture systems and sampled at consistent timepoints, the location of iSNVs was conserved and present at a significantly higher read frequency than within the clinical cohorts. Within the Beta lineage culture, multiple iSNVs in roughly 50% of the reads at inoculation became a high frequency iSNV (in > 80% of reads) or a SNP (≥ 90% of reads) by day three. We also documented instances where iSNVs were present at low frequencies at inoculation (baseline sampling) and remained at a low frequency across the study period, except for one dilution in which a SNP developed at day two and persisted. The Delta culture contained an iSNV that was present at inoculation, lost in all dilutions at day one, returning in one dilution on day two and then became present in all but one dilution by day three. Additionally, there were no iSNVs within the S gene of the Delta lineage. Therefore, the presence of an iSNV early in infection does not ensure that the variant will remain throughout the patient's course of infection, consistent with what has been observed in prior work (Armero et al., 2021; Lythgoe et al., 2021). It is also evident that a lack of iSNVs early on in infection does not indicate that a mixed population will not arise at some point during the infection course. This observation has implications for interpreting relationships between genomes when iSNVs are used to trace chains of transmission. There was also an interesting change in the representation of deletion events at the sub-consensus level between culture lineages. Within lineage A there were two low frequency deletion events within the S gene that overlapped, positions 23,583–23,598 nt, and 23,596–23,617 nt where the first deletion was at a considerably lower frequency than the second. This is indicative of positive viral selection. These polymorphisms are concentrated near the furin cleavage site and occurred predominately when SARS-CoV-2 is grown in VeroE6 cells. This cell line lacks key proteinases that enable more efficient viral entry and fusion (Chaudhry et al., 2020). The lack

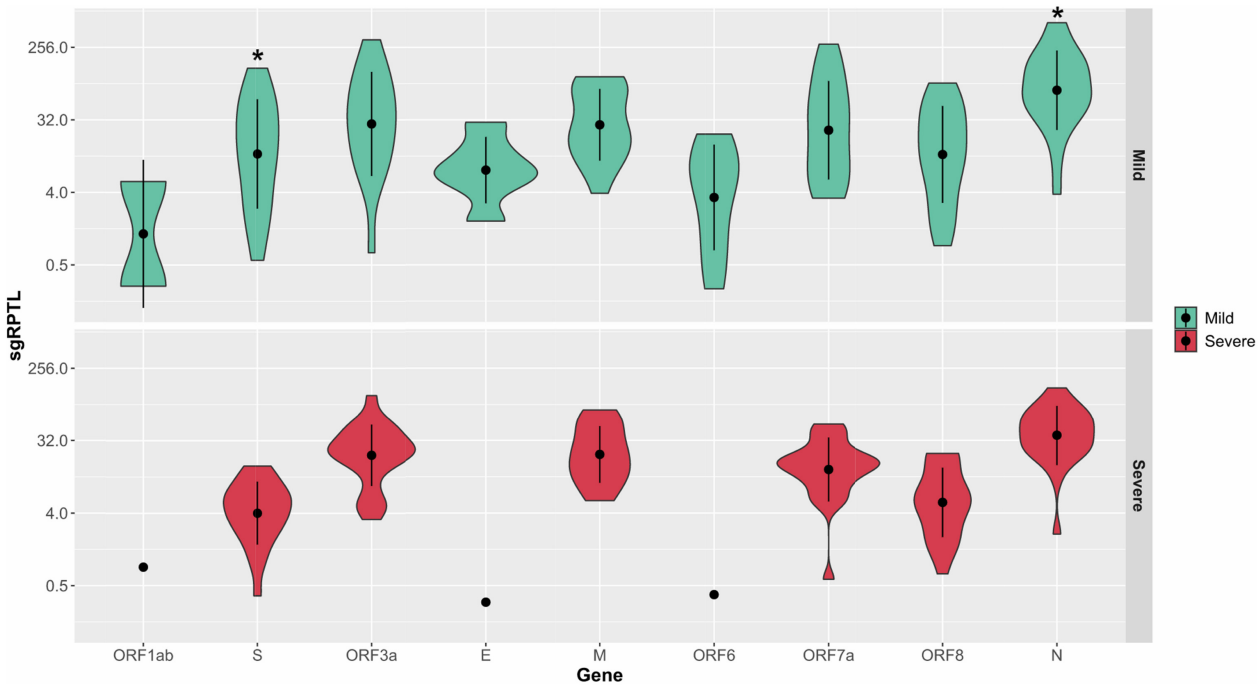

FIGURE 5 | Distribution of sgRPTL counts across all genes. sgRNA was detected in in the mild (red) and severe (green) clinical cohorts. Statistical significance ($p \leq 0.05$) is denoted by (*). S sgRNA detected in n = 24/26 severe and n = 21/25 mild genomes, ORF3a in 24/26 severe genomes and 17/25 mild genomes, E in 1/26 severe genomes and 6/25 mild genomes, M in 20/26 severe genomes and 15/25 mild genomes, ORF6 in 1/26 severe genomes and 11/25 mild genomes, ORF7a in 24/26 severe genomes and 18/25 mild genomes, ORF8 in 22/26 severe genomes and 20/25 mild genomes and N in 25/26 severe genomes and 24/25 mild genomes.

of proteinases in VeroE6 additionally explains the occurrence of the H655Y (C23525T) mutation in Lineage A culture samples days 2 and 3 (dilution 1×10^{-2}) at low frequencies (5.2 and 6.2%, respectively). This substitution has been found at high prevalence (> 98%) in the Gamma and Omicron VOCs, and at an extremely low prevalence (< 0.1%) in Delta, Alpha, and Beta lineages (Mullen et al., 2020). This mutation is proximal to the furin cleavage site (Escalera et al., 2022), and is associated with variations in antigenicity *via* conferring escape from human monoclonal antibodies (Baum et al., 2020).

While most mutations are purged or have no effect on the fitness of the virus, some may be selected for and alter transmissibility, infectivity, or pathogenicity (Plante et al., 2021). In both the clinical cohorts and culture dilutions, over 50% of the iSNVs resulted in a non-synonymous change, and between 21 to 37% iSNVs indicated a synonymous change. In all instances, positive selection was observed. Coronavirus mutations in the functionally important spike protein have the potential to affect virus infectivity, pathogenicity, and susceptibility to neutralising antibodies (Harvey et al., 2021). The spike gene encompasses positions 21,563–25,384 nt and iSNVs were seen within this range in the clinical cohorts and culture dilutions at the second highest frequency, with only ORF1ab being higher. This is consistent with studies that have observed positive pressure on protein coding genes, especially those associated with surface glycoproteins (Lo Presti et al., 2020).

Subgenomic RNA Variation

We uncovered low levels of sgRNA expression across all three cohorts, representing, on average less than 2% of the read depth of gRNA. However, the relative abundance of the eight sgRNA transcripts was similar to other investigations, where nucleocapsid sgRNA transcripts were the most abundant, and ORF10 sgRNA was not detected (Alexandersen et al., 2020; Parker et al., 2021). Interestingly, the pattern of sgRNA detection was similar across both the clinical cohorts and culture dilutions, dissimilar to Nomburg et al. (2020), where sgRNA was detected more frequently in culture. Instead, a significantly higher level of sgRNA was identified in patients with mild disease. This is an interesting finding as it has been reported that sgRNA transcripts are reduced in asymptomatic cases of COVID-19 (Wong et al., 2021). However, this may be a result of the lack of intervention in mild cases compared to interventions which would have been received by hospitalised severe disease cases. This similar pattern of expression also remained unchanged between viral lineages (D.2, B.1, and B.1.617.2). The presence of sgRNA transcripts in the E and N genes can be considered markers for increased replication (Zollo et al., 2021). However, it was postulated (Alexandersen et al., 2020) that levels of sgRNA may not be a reliable indicator of disease progression. Our data supports this assumption with levels of sgRNA in genes of importance, such as S and N remaining relatively even across the longitudinal clinical samples. Further to this concept, the

median levels of sgRNA detected in the severe cohort were less than detected in the mild cohort in all genes except M. This is inverse to the assumption that cases in the severe cohort are generally considered to have high levels of sgRNA. It is possible that the production of sgRNAs is more relevant for transmission, cell entry and less important in viral propagation. In addition, tri-nucleotide mutations have been identified in some lineages generating novel TRS which increases expression of sgRNA transcripts. This can be seen in the B.1 and D.2 lineages which expressed the highest levels of sgRNA transcripts for the nucleocapsid encoded by a GGG > AAC mutation (28,881–28,883 nt). It is still unknown how these new transcripts will impact pathology, but it is hypothesized that it could lead to diversification and adaptation to the host (Long, 2021).

We have established significant differences of iSNVs between severe and mild disease cohorts and SARS-CoV-2 genes, as well as distinct and consistent patterns of sgRNA. Our findings are also consistent with relative abundances of sgRNA described in S and N genes (Alexandersen et al., 2020). However, this study was limited in the available sample size, which was further complicated by low viral levels in later longitudinal samples. Strict lockdown procedures and border closures in NSW, Australia also greatly reduced or eliminated the proliferation of SARS-CoV-2 lineages, leading to low numbers of representative genomes per lineage. Further investigations with larger time frames and more longitudinal samples will be required to gain an understanding of the behaviour and contribution of iSNVs to COVID-19 disease and its transmission.

CONCLUSION

We demonstrate that iSNVs in SARS-CoV-2 genomes can accumulate over the course of COVID-19 disease and were predominately sporadic across cases with severe or mild disease. There were lineage-specific hot spots associated with persistent and low level iSNVs within diverse samples. sgRNA expression appears relatively consistent across both severe and mild disease, with the exception of significantly higher expression of sgRNA S and N transcripts in the mild disease cohort. The levels of sgRNA were, on average, less than two percent of the total reads for any gene in any clinical sample, indicating that SARS-CoV-2 sgRNA may not be a major contributor to the severity of clinical presentations of COVID-19. The ongoing surveillance and monitoring of subpopulations and iSNVs within lineages over time can improve our understanding of the underlying SARS-CoV-2 host adaptation. In addition, monitoring of sgRNA levels, especially associated with severity of disease may be important in understanding their impact on the spread of COVID-19.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the NCBI Sequence Read Archive online repository which can be accessed through (https://www.ncbi.nlm.nih.gov/sra/). Data pertaining to BioSample accession numbers SAMN22208473 to SAMN22208556 can be found in the BioProject PRJNA633948 and SRA SRP262661.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Governance and human ethics approval for clinical metadata and use of specimens from cases positive for SARS-CoV-2 in New South Wales was obtained by the Western Sydney Local Health District Human Research Ethics Committee (2020/ETH02426 and 2020/ETH02282). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JJ-M, JA, RR, and VS: designed experiments. JJ-M, JA, RR, WF, and CL: performed experiments. JJ-M, JA, RR, WF, MG, and AA: data analysis and visualisation and performed bioinformatic analyses. RR, VS, JK, AA, and KB: logistics and resources. JJ-M and JA: drafting the initial manuscript. JJ-M, JA, RR, WF, KB, VS, and CL: critical manuscript revision and supervision of research. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Centre for Infectious Diseases and Microbiology – Public Health (CIDM-PH), WSLHD, Prevention Research Support Program funded by the NSW Ministry of Health, NSW Health COVID-19 priority funding, Medical Research Future Fund Coronavirus Research Response 2020 – Tracking COVID-19 using Genomics (MRF9200006).

ACKNOWLEDGMENTS

We thank the staff from the Virology Laboratory and Microbial Genomics Reference Laboratory at the Institute of Clinical Pathology and Medical Research, NSW Health Pathology for their logistical support, and the Sydney Informatics Hub, University of Sydney Core Research Facility at the University of Sydney for providing research computing services. We also thank partner laboratories within the NSW Health Pathology network for their participation in the COVID-19 genomic surveillance, and GISAID team for their efforts in collating global data on the SARS-CoV-2 pandemic.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.824217/full#supplementary-material

Supplementary Figure 1 | COVID-19 cohorts and sampling timeline. **(A)** Representation of the initial patient pool considered for this study. Patients excluded (grey) and their respective samples (yellow circle) did not pass quality filtering. The pink and light blue blocks represent the severe and mild cohorts respectively. Within cohorts, the blue silhouettes represent longitudinally sampled patients and green represents patients where only one sample was taken. The number of longitudinal samples included in the study are indicated but the numerical values below the silhouette. The number of samples in the cohort that failed quality filtering are illustrated by the black cross. **(B)** The cohort type and duration of onset days of enrolled patients. Samples of patients with known symptom onset (green circle) and unknown symptom onset (green triangle) are illustrated. When more than one sample was taken on the same day, the shape is outlined in black. Severe and mild cases are shaded with red and blue, respectively. Patients infected with SARS-CoV-2 Delta lineage are represented by grey vertical bars. Patients within the same genomic epi-cluster are grouped by purple (Lineage B.1.617.2, cluster NSW 130), yellow (Lineage D.2, cluster NSW 33.1), orange (Lineage D.2, cluster NSW 33), and green (Lineage B.1, cluster NSW 17.5) shaded boxes. Black diamonds denote household contacts within the genomic epi-cluster.

Supplementary Figure 2 | Boxplots depicting the SARS-CoV-2 copy number **(A)** and median genome depth **(B)** for all cohorts. The bold line indicates the median, the interquartile range (25th to 75th percentile) is represented by the white shading, the whiskers represent the minimum and maximum values, and the outliers are shown by the black circles.

Supplementary Figure 3 | Counts of SNPs **(A)** and iSNVs **(B)** by SARS-CoV-2 gene for the culture cohort. Frequencies $\geq$ 0.9 were considered SNPs. Problematic sites are not included. NC signifies non-coding region of the genome.

Supplementary Figure 4 | Frequencies of SNPs **(A)** and iSNVs **(B)** by SARS-CoV-2 gene for culture cohort. Frequencies $\geq$ 0.9 were considered SNPs. Problematic sites are not included. NC signifies non-coding region of the genome.

Supplementary Figure 5 | Violin plots depicting the sgRPTL normalized counts for sgRNA abundance at each gene from top to bottom of Culture A, Culture Beta, and Culture Delta. There was significantly higher sgRNA across all genes in Delta compared to A and Beta. A and Beta were not significantly different. Individually, N was significantly higher in Delta compared to Beta and ORF 7a and 8 were significantly high than both A and Beta. The boxplots within the violins indicate the median and the interquartile ranges (25th to 75th percentile).

Supplementary Figure 6 | Median sgRPTL by SARS-CoV-2 gene from inoculum (day 0) to day 3 for culture dilutions (top) lineage A, (middle) Beta, and (bottom) Delta.

REFERENCES

Al Khatib, H. A., Benslimane, F. M., Elbashir, I. E., Coyle, P. V., Al Maslamani, M. A., Al-Khal, A., et al. (2020). Within-host diversity of SARS-CoV-2 in COVID-19 patients with variable disease severities. *Front. Cell. Infect. Microbiol.* 10:575613. doi: 10.3389/fcimb.2020.575613

Alexandersen, S., Chamings, A., and Bhatta, T. R. (2020). SARS-CoV-2 genomic and subgenomic RNAs in diagnostic samples are not an indicator of active replication. *Nat. Commun.* 11:6059. doi: 10.1038/s41467-020-19883-7

Armero, A., Berthet, N., and Avarre, J.-C. (2021). Intra-host diversity of SARS-Cov-2 should not be neglected: case of the state of Victoria, Australia. *Viruses* 13:133. doi: 10.3390/v13010133

Arya, R., Kumari, S., Pandey, B., Mistry, H., Bihani, S. C., Das, A., et al. (2021). Structural insights into SARS-CoV-2 proteins. *J. Mol. Biol.* 433:166725.

Baum, A., Fulton Benjamin, O., Wloga, E., Copin, R., Pascal Kristen, E., Russo, V., et al. (2020). Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* 369, 1014–1018. doi: 10.1126/science.abd0831

Chaudhry, M. Z., Eschke, K., Grashoff, M., Abassi, L., Kim, Y., Brunotte, L., et al. (2020). SARS-CoV-2 quasispecies mediate rapid virus evolution and adaptation. *bioRxiv* [Preprint]. doi: 10.1101/2020.08.10.241414

Davidson, A. D., Williamson, M. K., Lewis, S., Shoemark, D., Carroll, M. W., Heesom, K. J., et al. (2020). Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome Med.* 12:68. doi: 10.1186/s13073-020-00763-0

Day, T., Gandon, S., Lion, S., and Otto, S. P. (2020). On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* 30, R849–R857. doi: 10.1016/j.cub.2020.06.031

Domingo, E., and Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* 51, 151–178. doi: 10.1146/annurev.micro.51.1.151

Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., and Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* 6:veaa061. doi: 10.1093/ve/veaa061

Escalera, A., Gonzalez-Reiche, A. S., Aslam, S., Mena, I., Laporte, M., Pearl, R. L., et al. (2022). Mutations in SARS-CoV-2 variants of concern link to increased spike cleavage and virus transmission. *Cell Host Microbe* 30, 373–387.e7. doi: 10.1016/j.chom.2022.01.006

Graepel, K. W., Lu, X., Case, J. B., Sexton, N. R., Smith, E. C., and Denison, M. R. (2017). Proofreading-deficient coronaviruses adapt for increased fitness over long-term passage without reversion of exoribonuclease-inactivating mutations. *Mol. Biol.* 8, e01503–17. doi: 10.1128/mBio.01503-17

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123. doi: 10.1093/bioinformatics/bty407

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* 19, 409–424.

Johnson, J. A., Li, J. F., Wei, X., Lipscomb, J., Irlbeck, D., Craig, C., et al. (2008). Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naïve populations and associate with reduced treatment efficacy. *PLoS Med.* 5:e158. doi: 10.1371/journal.pmed.0050158

Karamitros, T., Papadopoulou, G., Bousali, M., Mexias, A., Tsiodras, S., and Mentis, A. (2020). SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. *J. Clin. Virol.* 131:104585. doi: 10.1016/j.jcv.2020.104585

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J. W., Kim, V. N., and Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. *Cell* 181, 914–921.e10. doi: 10.1016/j.cell.2020.04.011

Kumar, S., Nyodu, R., Maurya, V. K., and Saxena, S. K. (2020). "Host immune response and immunobiology of human SARS-CoV-2 infection," in *Coronavirus Disease 2019 (COVID-19) Medical Virology: From Pathogenesis to Disease Control*, ed. S. Saxena (Berlin: Springer), 43–53. doi: 10.1007/978-981-15-4814-7_5

Lam, C., Gray, K., Gall, M., Sadsad, R., Arnott, A., Johnson-Mackinnon, J., et al. (2021). SARS-CoV-2 genome sequencing methods differ in their abilities to detect variants from low-viral-load samples. *J. Clin. Microbiol.* 59:e0104621. doi: 10.1128/JCM.01046-21

Lo Presti, A., Rezza, G., and Stefanelli, P. (2020). Selective pressure on SARS-CoV-2 protein coding genes and glycosylation site prediction. *Heliyon* 6:e05001. doi: 10.1016/j.heliyon.2020.e05001

Long, S. (2021). SARS-CoV-2 subgenomic RNAs: characterization, utility, and perspectives. *Viruses* 13:1923. doi: 10.3390/v13101923

Lythgoe, K. A., Hall, M., Ferretti, L., De Cesare, M., Macintyre-Cockett, G., Trebes, A., et al. (2021). SARS-CoV-2 within-host diversity and transmission. *Science* 372:eabg0821.

Mullen, J. L., Tsueng, G., Latif, A. A., Alkuzweny, M., Cano, M., Haag, E., et al. (2020). *Outbreak.info*. Available online at: https://outbreak.info/ (accessed April 7, 2022).

Nakagawa, S., and Miyazawa, T. (2020). Genome evolution of SARS-CoV-2 and its virological characteristics. *Inflamm. Regen.* 40:17.

Naqvi, A. A. T., Fatima, K., Mohammad, T., Fatima, U., Singh, I. K., Singh, A., et al. (2020). Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: structural genomics approach. *Biochim. Biophys. Acta Mol. Basis Dis.* 1866:165878. doi: 10.1016/j.bbadis.2020.165878

Nomburg, J., Meyerson, M., and Decaprio, J. A. (2020). Pervasive generation of non-canonical subgenomic RNAs by SARS-CoV-2. *Genome Med.* 12:108. doi: 10.1186/s13073-020-00802-w

Nowak, M. A., Anderson, R. M., Mclean, A. R., Wolfs, T. F., Goudsmit, J., and May, R. M. (1991). Antigenic diversity thresholds and the development of AIDS. *Science* 254, 963–9. doi: 10.1126/science.1683006

Parker, M. D., Lindsey, B. B., Leary, S., Gaudieri, S., Chopra, A., Wyles, M., et al. (2021). Subgenomic RNA identification in SARS-CoV-2 genomic sequencing data. *Genome Res.* 31, 645–658. doi: 10.1101/gr.268110.120

Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., et al. (2021). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 592, 116–121.

Popa, A., Genger, J.-W., Nicholson, M. D., Penz, T., Schmid, D., Aberle, S. W., et al. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci. Transl. Med.* 12:eabe2555.

Rahman, H., Carter, I., Basile, K., Donovan, L., Kumar, S., Tran, T., et al. (2020). Interpret with caution: an evaluation of the commercial AusDiagnostics versus in-house developed assays for the detection of SARS-CoV-2 virus. *J. Clin. Virol.* 127:104374. doi: 10.1016/j.jcv.2020.104374

Ramazzotti, D., Angaroni, F., Maspero, D., Gambacorti-Passerini, C., Antoniotti, M., Graudenzi, A., et al. (2020). Characterization of intra-host SARS-CoV-2 variants improves phylogenomic reconstruction and may reveal functionally convergent mutations. *bioRxiv* [Preprint]. doi: 10.1101/2020.04.22.044404

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., et al. (2020). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Genomic Epidemiol. Virol.* Available online at: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563

Rockett, R. J., Arnott, A., Lam, C., Sadsad, R., Timms, V., Gray, K.-A., et al. (2020). Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nat. Med.* 26, 1398–1404. doi: 10.1038/s41591-020-1000-7

Sapoval, N., Mahmoud, M., Jochum, M. D., Liu, Y., Elworth, R. A. L., Wang, Q., et al. (2020). Hidden genomic diversity of SARS-CoV-2: implications for qRT-PCR diagnostics and transmission. *bioRxiv* [Preprint]. doi: 10.1101/2020.07.02.184481

Smith, E. C., Sexton, N. R., and Denison, M. R. (2014). Thinking outside the triangle: replication fidelity of the largest RNA viruses. *Annu. Rev. Microbiol.* 1, 111–132. doi: 10.1146/annurev-virology-031413-085507

Sola, I., Almazán, F., Zúñiga, S., and Enjuanes, L. (2015). Continuous and discontinuous RNA synthesis in coronaviruses. *Annu. Rev. Virol.* 2, 265–288. doi: 10.1146/annurev-virology-100114-055218

Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., et al. (2019). From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses* 11:59. doi: 10.3390/v11010059

Stone, J. K., Andino, R., Arnold, J. J., Cameron, C. E., and Vignuzzi, M. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439, 344–348. doi: 10.1038/nature04388

Tonkin-Hill, G., Martincorena, I., Amato, R., Lawson, A. R. J., Gerstung, M., Johnston, I., et al. (2021). Patterns of within-host genetic diversity in SARS-CoV-2. *Elife* 10:e66857. doi: 10.7554/eLife.66857

Valesano, A. L., Rumfelt, K. E., Dimcheff, D. E., Blair, C. N., Fitzsimmons, W. J., Petrie, J. G., et al. (2021). Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLoS Pathog.* 17:e1009499. doi: 10.1371/journal.ppat.1009499

Wang, Y., Wang, D., Zhang, L., Sun, W., Zhang, Z., Chen, W., et al. (2021). Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. *Genome Med.* 13:30. doi: 10.1186/s13073-021-00847-5

Wells, H. L., Letko, M., Lasso, G., Ssebide, B., Nziza, J., Byarugaba, D. K., et al. (2021). The evolutionary history of ACE2 usage within the coronavirus subgenus Sarbecovirus. *Virus Evol.* 7:veab007.

Wong, C. H., Ngan, C. Y., Goldfeder, R. L., Idol, J., Kuhlberg, C., Maurya, R., et al. (2021). Reduced subgenomic RNA expression is a molecular indicator of asymptomatic SARS-CoV-2 infection. *Commun. Med.* 1:33.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.

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

Zollo, M., Ferrucci, V., Izzo, B., Quarantelli, F., Domenico, C. D., Comegna, M., et al. (2021). SARS-CoV-2 Subgenomic N (sgN) transcripts in oro-nasopharyngeal swabs correlate with the highest viral load, as evaluated by five different molecular methods. *Diagnostics (Basel)* 11:288. doi: 10.3390/diagnostics11020288

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.