scientific reports

Check for updates

OPEN

Host–pathogen dynamics in longitudinal clinical specimens from patients with COVID-19

Michelle J. Lin[1], Victoria M. Rachleff[1,2,3], Hong Xie[1], Lasata Shrestha[1], Nicole A. P. Lieberman[1], Vikas Peddu[1], Amin Addetia[1], Amanda M. Casto[4], Nathan Breit[1], Patrick C. Mathias[1], Meei-Li Huang[1,2], Keith R. Jerome[1,2,5]✉, Alexander L. Greninger[1,2,5]✉ & Pavitra Roychoudhury[1,2,5]✉

Rapid dissemination of SARS-CoV-2 sequencing data to public repositories has enabled widespread study of viral genomes, but studies of longitudinal specimens from infected persons are relatively limited. Analysis of longitudinal specimens enables understanding of how host immune pressures drive viral evolution in vivo. Here we performed sequencing of 49 longitudinal SARS-CoV-2-positive samples from 20 patients in Washington State collected between March and September of 2020. Viral loads declined over time with an average increase in RT-QPCR cycle threshold of 0.87 per day. We found that there was negligible change in SARS-CoV-2 consensus sequences over time, but identified a number of nonsynonymous variants at low frequencies across the genome. We observed enrichment for a relatively small number of these variants, all of which are now seen in consensus genomes across the globe at low prevalence. In one patient, we saw rapid emergence of various low-level deletion variants at the N-terminal domain of the spike glycoprotein, some of which have previously been shown to be associated with reduced neutralization potency from sera. In a subset of samples that were sequenced using metagenomic methods, differential gene expression analysis showed a downregulation of cytoskeletal genes that was consistent with a loss of ciliated epithelium during infection and recovery. We also identified co-occurrence of bacterial species in samples from multiple hospitalized individuals. These results demonstrate that the intrahost genetic composition of SARS-CoV-2 is dynamic during the course of COVID-19, and highlight the need for continued surveillance and deep sequencing of minor variants.

SARS-CoV-2 is the cause of coronavirus disease 2019 (COVID-19). There have been over 256 million COVID-19 cases and over 5.1 million total deaths due to COVID-19 worldwide, at time of writing[1]. Genomic analyses of longitudinal specimens within infected persons are critical to understanding the evolutionary trajectory of SARS-CoV-2. Sequencing of longitudinal samples from infected individuals allows examination of viral genetic diversity, host immune response, and dynamics of co-infecting pathogens over the course of infection and recovery. Within-host variants arise during viral replication and a number of processes shape their frequencies over time. These include selective pressures at different scales (molecular, immunological, epidemiological), host heterogeneity, spatial structure, population bottlenecks, and other stochastic processes[2]. Within-host variants may impact the success of vaccines and therapeutics, and a fraction of variants that arise will be transmitted between hosts and can eventually reach fixation in the population[3]. Recent studies of within-host diversity of SARS-CoV-2 have demonstrated the presence of low levels of minor variants and infrequent emergence of escape mutations[3–7]. Of particular note, deletions in the N-terminal domain of the spike glycoprotein have been observed in chronically infected immunocompromised patients that are associated with SARS-CoV-2 escape from sera[8–11], and are present in current circulating lineages of concern.

Here we examined longitudinal clinical specimens collected from 20 COVID-19-positive patients in Washington State. With metagenomic sequencing we identified changes in host gene expression and bacterial

[1]Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, WA 98102, USA. [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [3]Program in Molecular and Cellular Biology, University of Washington School of Medicine, Seattle, WA, USA. [4]Division of Allergy and Infectious Diseases, University of Washington School of Medicine, Seattle, WA, USA. [5]These authors jointly supervised this work: Keith R. Jerome, Alexander L. Greninger and Pavitra Roychoudhury. ✉email: kjerome@uw.edu; agrening@uw.edu; proychou@uw.edu

Characteristics	(N = 20)
Mean age, y (SD)	70 (18)
Male, n (%)	13 (65)
Race, n (%)	
White	12 (60)
Asian	4 (15)
Black or African American	2 (10)
American Indian or Alaska Native	1 (5)
Unknown or Unavailable	1 (5)
Comorbidities, n (%)	
Hypertension	10 (50)
Diabetes	7 (35)
Obesity	2 (10)
Asthma	1 (5)
Treatment, n (%)	
Convalescent Plasma	2 (10)
Hydroxychloroquine	2 (10)
Azithromycin	5 (25)
Tocilizumab	1 (5)
ACTT-1 Trial	2 (10)
No Treatment	8 (40)
Unknown	4 (20)
Hospital outcomes, n (%)	
Hospital admission	14 (70)
ICU admission for COVID-19	4 (20)
Survival to discharge	18 (90)

Table 1. Demographics and clinical characteristics of patients included in study. Different categories (in bold) and their subcategories are shown in the first column, with their respective number of patients in the second column. In parentheses, standard deviation is indicated in the first row, and percentages for all other rows.

co-occurrences, which may be associated with recovery[12,13]. We found negligible change in viral consensus sequences over time, but detectable changes in variant allele frequencies that are only weakly predictive of future consensus changes across the globe. We further observed rapid emergence of deletion variants in the N-terminus domain of the spike glycoprotein in one patient, potentially suggesting within-host SARS-CoV-2 evasion of NTD-directed antibodies. Taken together our results support the limited emergence and fixation of escape variants during a typical infection, and also highlight the need to monitor minor variants due to their potential impact on vaccine and therapeutic efficacy.

Results

Viral load dynamics in longitudinal samples from SARS-CoV-2 infected individuals. Residual clinical specimens were obtained from the University of Washington (UW) Virology Lab after testing for SARS-CoV-2[14]. By reviewing our laboratory information system, we identified 20 individuals who had two or more positive or inconclusive samples collected between March and September 2020 (Table 1, Supplementary Table S1). Inconclusive samples had one of two PCR targets detected. This commonly occurs in samples with small amounts of viral DNA, and thus these were considered presumptive positive for SARS-CoV-2, albeit with low viral load. A majority of samples came from inpatients who received care within the UW Medicine system, which includes UW Medical Center, Harborview Medical Center, and Northwest Hospital. Samples came from individuals with a mean age of 70 (range 42–99), many with severe disease given the availability of multiple samples from these patients. Consistent with other reports[15], we observed that viral load declined over time in most patients with two or more positive or inconclusive samples with an average increase in RT-QPCR cycle threshold (Ct) of 0.87 per day (Fig. 1, Supplementary Fig. S1).

A total of 49 samples (47 nasopharyngeal and 2 oropharyngeal swabs) were sequenced with sufficient reads to be included in this study (Supplementary Fig. S2A). Ct values for these samples ranged between 14.6 and 36.6. The length of time between collection dates for sequenced samples from the same individual ranged from 0 to 22 days. Samples collected on the same date were sequenced for three individuals. We sequenced nasopharyngeal and oropharyngeal samples from P004 collected at the time of autopsy, two samples from P006 collected at the same time during a hospital admission, and samples collected 9 h apart from P012 during an emergency room visit.

Negligible change in consensus sequences over time. We obtained full-length viral genome sequences with less than 2% unknown bases (Ns) for two or more time points in 14 out of 20 individuals,

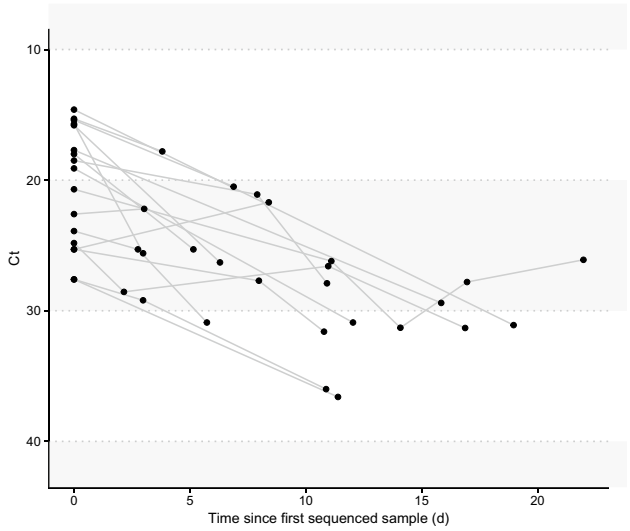

Figure 1. Viral load dynamics in sequenced samples. Dots represent a unique sequenced sample. Lines connect samples from a single patient. Same day samples are not shown (see Supplementary Fig. S1).

plus an additional two individuals with paired nasopharyngeal and oropharyngeal swabs collected at the same timepoint (Table 2, n = 38 sequences). After masking ambiguous sites and regions with sequencing or assembly errors, we found no differences between the first and subsequent consensus sequences in 15 out of 16 patients. In one patient (P001), two samples collected 3 days apart had 4 differences between their consensus sequences at reference positions 15,418 (G/T), 26,262 (G/T), 27,899 (T/A), and 27,944 (T/C). Two of these differences lead to coding changes (A660S in nsp12 and Q2K in ORF8). Variant alleles were observed for all four positions at low frequencies in the sample collected at the earlier timepoint. These mutations are not present in any currently characterized variants of concern.

Low frequency variants detected across the genome. We analyzed intrahost viral genetic variation by examining all sites with > 100 × locus depth, masking known problematic sites and filtering for high-confidence variants (see Methods). We examined sites in 47 samples from 20 different patients and found a total of 103 unique non-synonymous variants relative to the Wuhan-Hu-1 (NC_045512.2) reference genome present at frequencies between 5 and 95% (Fig. 2A). nsp12 had the highest number of variant sites (26, 25.24%), followed by nsp3 (16, 15.533%) and the spike glycoprotein (13, 12.62%). Even when adjusted for gene length, mutations were most prevalent in nsp12, at 0.009 variant sites per nucleotide. For samples that were sequenced multiple times (n = 10), variant frequencies were reproducible across replicates, particularly among samples sequenced with the same library preparation method (Fig. 2B). Intra-host mutations present in multiple longitudinal samples of the same patient were mostly fixed and highly clonal, with some low-level variants (Supplementary Fig. S2B).

Of the seven most commonly observed variants in our dataset (Table 3), the three most frequent define the Washington state outbreak clade[16,17] and the rest of the variants are clade-defining mutations in Nextstrain clades 20A and 20C. Nine out of 20 patients had the spike protein mutation D614G (A23403G), which has been associated with increased transmissibility and higher viral loads[18,19]. While this variant was rare at the beginning of the pandemic, it reached near fixation in the global SARS-CoV-2 population by June 2020[20]. This rapid rise in prevalence is reflected in our data, as this D614G mutation is present in all three patients with samples collected during or after June 2020. In 10 out of the 11 patients with the 614D variant, no alternate alleles were detected at this position. In the second sample from P007, 614G was detected with a variant allele frequency of 6.1%, but the read depth at this locus (82X) was insufficient to reach our QC standards.

Variants exhibiting intra-host evolution are limited in prediction of future global consensus changes and highlight SARS-CoV-2 antibody evasion. We further examined variants that underwent a maximum allele frequency change of ≥ 20% across timepoints within each patient. The derived alleles for all 25 non-synonymous amino acid changes meeting this criteria in our dataset were also observed among consensus genomes deposited in GISAID[21] by April 2021 (range: 1–2171, mean: 467.8, median: 86.5). Only 8 variants exhibited a maximum allele frequency change of ≥ 40%. The derived alleles for these variants were present in very few GISAID consensus sequences (range: 1–1751), which at the time of analysis represented a mere 0.0001–0.17% of all sequences deposited in GISAID (Fig. 3A). Though the number of consensus sequences with these derived alleles was relatively low, these sequences were diverse with respect to collection date and geographic origin (Fig. 3B).

Three variants with a ≥ 20% within-host change in allele frequency localized to the spike glycoprotein (Fig. 4A). None of these had an allele frequency change of ≥ 50%. In patient P016, a S:143–145 6-nucleotide deletion was observed at an allele frequency of 20.7% in patient sample 2, but was not observed in samples collected

Patient	Sample #	Days since symptom onset	Ct value	%Ns	Clade (Nextclade/Pangolin)	Number of nt differences relative to first sample
P001	1	Asymptomatic	22.6	0.0	19B/A.1	–
	2	3**	22.2	0.0		4
P003	1	Unknown	18.0	0.0	19B/A.1	–
	2	5**	25.3	0.0		0
P005	1	0*	19.1	0.3	19B/A.1	–
	2	12	30.9	0.0		0
P006	1	Asymptomatic	25.6	0.0	19B/A.1	–
	2	0**	29.7	0.0		0
P007	1	0*	15.8	0.0	19B/A.1	–
	2	6	26.3	0.0		0
P008	1	0*	17.7	0.0	19B/A.1	–
	2	16	29.4	0.0		0
P009	1	0*	25.3	0.0	20C/B.1.21	–
	2	9	21.7	0.0		0
P010	1	−7	20.7	0.0	19B/A.1	–
	2	4	26.2	0.0		0
	3	7	31.3	0.0		0
	4	15	26.1	0.0		0
P011	1	0	25.3	0.5	20C/B.1.21	–
	2	8	27.7	0.0	20C/B.1.21	0
	3	11	31.6	0.1		0
P012	1	5	21.6	0.7	20C/B.1.21	–
	2	5	19.8	0.0		0
P014	1	Asymptomatic	27.6	0.0	19B/A.1	–
	2	3**	29.2	1.6		0
P015	1	3	18.5	0.0	19B/A.1	–
	2	11	21.1	0.0		0
	3	14	27.9	0.0		0
P016	1	16	23.9	0.0	20C/B.1	–
	2	19	25.3	0.0		0
	3	22	30.9	0.0		0
P017	1	10	24.8	0.0	19B/A.1	–
	2	13	28.6	0.0		0
	3	21	26.6	0.0		0
P018	1	Unknown	15.3	0.0	20B/B.1.1.77	–
	2	3**	17.8	0.0		0
P019	1	Unknown	14.6	0.0	20A/B.1	–
	2	19**	31.1	0.3		0

Table 2. Consensus sequence analysis of SARS-CoV-2 in longitudinal specimens. All patients with less than 2% unknown bases (Ns) are included. The last column indicates nucleotide differences compared to the first sample collected for each respective patient. One asterisk (*) indicates symptoms were present at first time point but exact date of symptom onset is unknown. Two asterisks (**) indicate days since first sample.

3 days prior and 3 days later. This patient was immunocompetent and not receiving any COVID-19 treatment. Interestingly, numerous other deletions arose at low frequencies in this patient, with the largest number present at day 0 (Fig. 4B). The most prevalent deletion variant in the day 0 sample (collected 16 days after initial symptom onset) was S:Δ141–144 at 1.87% allele frequency. This deletion was the second most common variant in the day 3 sample at 1.2% allele frequency, but was not observed at all in the day 6 sample. Deletions in this region, including S:Δ141–144, have previously been observed in chronically infected immunocompromised patients, and some are associated with escape from NTD-specific neutralizing antibodies or polyclonal sera[8–11,22]. Notably, a S:Δ142–144 deletion is a hallmark of the currently circulating Omicron variant.

Longitudinal RNAseq analysis illustrates loss of ciliated epithelium during infection. For samples that were sequenced metagenomically, we pseudo-aligned reads to the human transcriptome to perform differential expression analysis comparing initial (t = 0) timepoints to later timepoints. Samples with more than 900,000 pseudo-aligned reads (n = 7 initial, 3 later timepoints) were included in the analysis to determine vari-

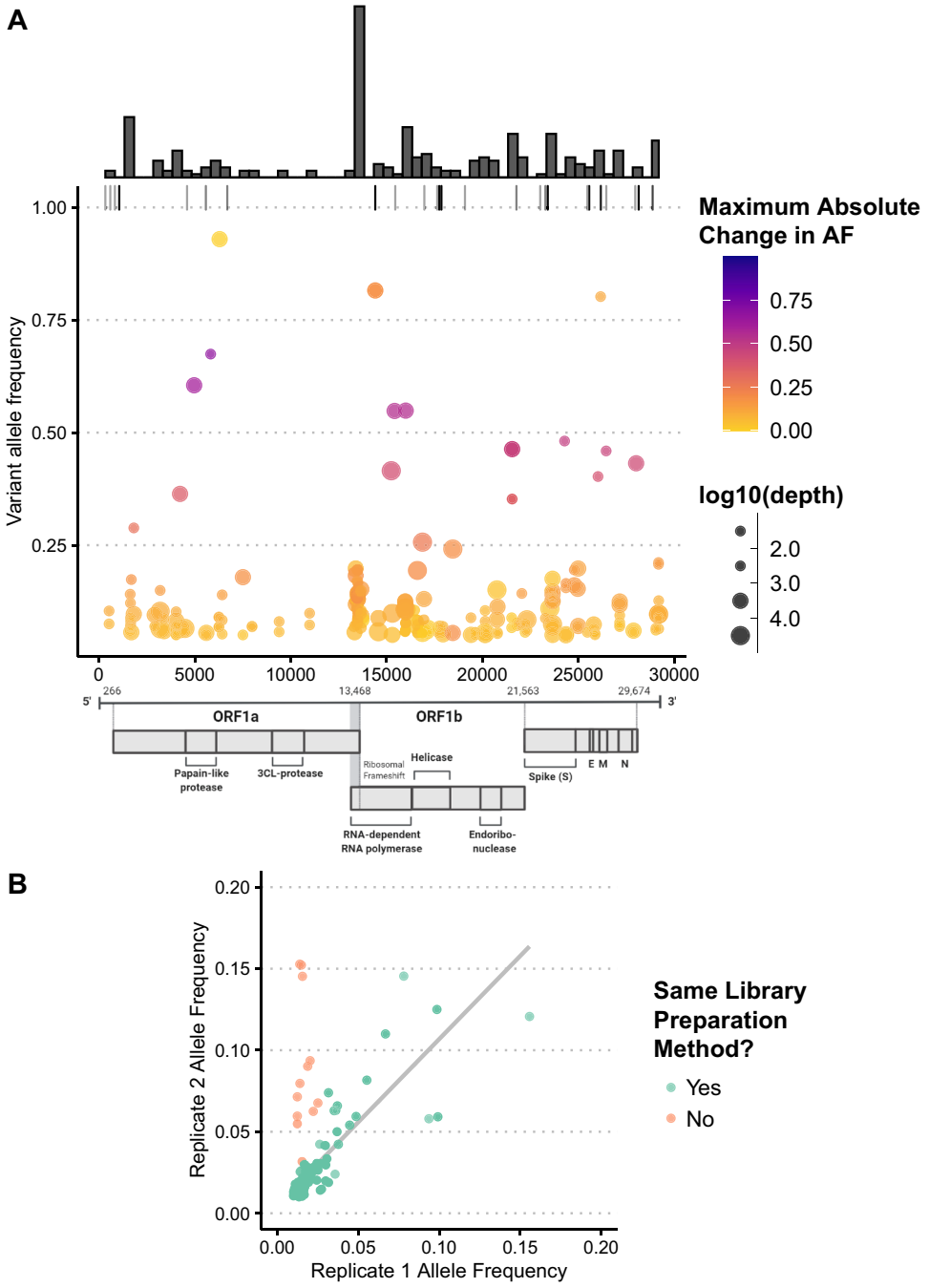

Figure 2. Low frequency variation is abundant but only a small number of variants exhibit a significant change in allele frequency over the course of infection. (**A**) Each dot represents a high-confidence coding change in a single sample relative to the Wuhan-Hu-1 (NC_045512.2) reference genome with variant allele frequency between 5 and 95%, at least ×100 coverage at the site, and reproducibility in multiple samples at lower frequencies (< 40%). Color scale represents the change in allele frequency across time points in the same patient with darker colors representing variants that had greater changes in frequency across samples. Small dark grey marks along the top margin shows positions with variant frequencies > 95% (fixed mutations relative to the reference). Size of circles indicates sequencing depth at the site. Marginal histogram shows distribution of variants using bin width of 500 nucleotides. (**B**) Comparison of allele frequencies of low-frequency variants (< 20%) across replicates of the same sample (n = 10 samples). Each dot represents a variant with ≥ 100 total depth and ≥ 10 allelic depth in each replicate. Line of best fit is shown in purple, and dots in orange represent replicates that were re-sequenced using a different library preparation method (amplicon sequencing vs. shotgun metagenomic sequencing).

Variant	AF range	# patients	# samples	ORF: effect
C17747T	0.98–1	11	24	ORF1ab: P5828L; helicase: P504L
A17858G	0.98–1	11	24	ORF1ab: Y5865C; helicase: Y541C
T28144C	0.02–1	12	23	ORF8: L84S
A23403G	0.99–1	9	19	S: D614G
C14408T	0.02–1	10	17	ORF1ab: P4715L; RdRp: P323L
G25563T	0.03–1	8	15	ORF3a: Q57H
C1059T	0.95–1	7	15	ORF1a: T265I; nsp2: T85I

Table 3. Frequent non-synonymous variants observed in ≥ 15 samples (n = 47). All variants called had at least 10 reads of support for the alternate allele. For the three variants with large ranges in allele frequency (T28144C, C14408T, G14408T), ≤ 3 outlier samples with variant AFs below 0.1 were present. When these samples are excluded, minimum AF increases to ≥ 0.99.

ation in host gene expression over time. We observed a dramatic downregulation of several cytoskeletal genes, particularly dynein heavy chain (*DNAH 2, 3, 5, 6, 7, 9, 10, 11, 12*), as well as *WDR*s, *MAP1A*, and others (Fig. 5A, Supplementary Table S3). Gene Ontology analysis (Fig. 5B) confirmed that downregulated genes are involved in biological processes associated with microtubule-based motility. This is consistent with the death of ciliated epithelial cells, which are enriched for transcripts encoding microtubule transport machinery[23], following SARS-CoV-2 infection. We observed upregulation of some actin cytoskeleton-related transcripts like VAV1, VASP, and RhoF[24]. We also observed downregulation of several interferon-stimulated genes, but these did not reach statistical significance in this small sample set.

Metagenomic analysis shows high levels of clinically relevant bacteria in three samples. We used a previously described metagenomic pipeline (CLOMP[25]) to perform taxonomic assignment of preprocessed reads. We excluded samples that had fewer than 10,000 reads after trimming and any samples that underwent enrichment via probe-capture or amplicon-sequencing for SARS-CoV-2. A total of 24 samples from 11 patients were included in further analysis (Supplementary Fig. S3, Supplementary Fig. S4). No viruses aside from SARS-related coronaviruses met the required cutoffs to be classified as co-infections (see "Methods"). Multiple samples had detectable numbers of bacterial reads, in particular *Staphylococcus aureus* and *Moraxella catarrhalis*. Samples from two patients (P004, P005) had high detectable levels (> 100 RPM) of both bacteria at one or more time point(s) (Supplementary Fig. S3B). In patient P006, who had paired nasopharyngeal (NP) and oropharyngeal (OP) swabs collected at the same time point, we detected *Capnocytophaga gingivalis*, *Capnocytophaga leadbetteri,* and *Streptococcus parasanguinis* in the OP swab at 21,960, 9475, and 4476 RPM, respectively. All three species of bacteria commonly colonize the oropharynx. In contrast, in the NP swab, the predominant species of bacteria was a common skin colonizer, *Cutibacterium acnes,* at 824 RPM. In P010, we found a large number of reads corresponding to *Corynebacterium* spp. at one time point. Upon review of medical records, we found no mention of bacterial co-infections or of positive bacterial cultures in the charts for P004 or P006. P005 had a nares culture that grew methicillin-resistant *Staphylococcus aureus* three months prior to SARS-CoV-2 infection.

Discussion

In this study, we performed high-throughput sequencing of longitudinal clinical specimens that were positive for SARS-CoV-2 by RT-QPCR. Most samples were sequenced using a metagenomic approach, which enabled us to simultaneously derive information about viral evolution, host transcription, and the presence of other organisms within patient samples.

We showed that although the viral consensus sequence remains largely unchanged over the course of infection, there is a relative abundance of genome-wide low-frequency variants. Similar to other studies, we saw a wide range in the number of variants detected across samples[26] and distribution of variants across the genome, though some positions appeared to be more prone to variation [4,26]. Studies of SARS-CoV-2 and other respiratory viruses[2,3,8,27] have demonstrated the transmission of minor variants and the role of these population bottlenecks on viral evolution, underscoring the importance of studying within-host viral variation. All variants demonstrating significant longitudinal evolution in our sample set collected March-September 2020 have been observed in consensus sequences from around the globe[21], albeit at relatively low prevalence.

In one patient, we observed rapid turnover of multiple deletion variants in the N-terminal domain of the spike glycoprotein, which has previously been seen in persistent infection in immunocompromised individuals and has been associated with viral escape of neutralizing antibodies[8,10,22,28]. Deletions in the NTD are of particular significance due to their presence in currently circulating lineages of concern. Here we show the emergence of a deletion in this genomic region in an immunocompetent background. It is unclear if the absence of this mutation at day 6 is due to successful clearance of the NTD variant or lack of detection of the minor allele associated with lower copy numbers. In addition, while some of these low frequency deletion alleles have been previously shown to arise independently in different patients in response to similar selection pressures[8], the presence of multiple low frequency deletion alleles within the same patient may be the product of parallel within-host microevolutionary processes suggestive of some selective advantage, particularly with its similarity to the S:Δ142–144 deletion in the currently circulating Omicron variant. Notably, we did not find any evolving

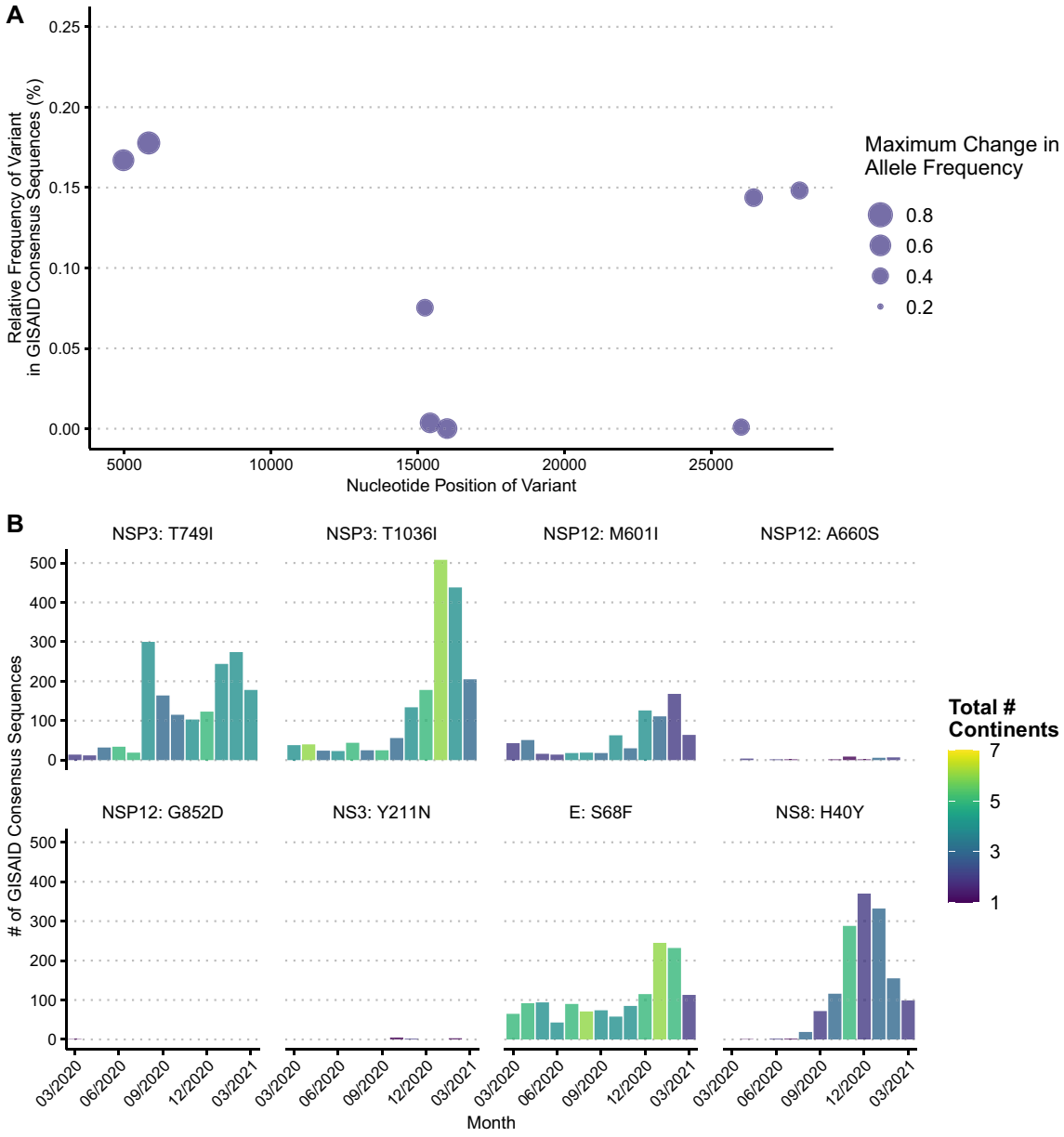

Figure 3. Variants that exhibit ≥ 40% maximum change in allele frequency in the individuals profiled here in summer 2020 show limited ability to predict future GISAID consensus sequences as of April 2021. (**A**) Relative frequencies of the derived allele found in GISAID consensuses across the genome. Dots represent each unique variant with size indicating the maximum intra-host change in allele frequency found in our study. (**B**) Number of GISAID consensuses with the derived allele for each variant. Height of vertical bars represents the total number of consensuses with the derived allele collected for each month from March 2020 to March 2021 and bar color represents the number of continents of origin for these consensuses.

variants selected for in the RBD, the main target for neutralizing activity of human plasma [29,30]. Additionally, we found two highly dynamic mutations within the Ubl2 and PL2Pro domains of the multifunctional nsp3 protein[31] and three mutations within the nsp12 polymerase. While we and others have previously linked a specific nsp12 mutation to remdesivir resistance[32], each of these mutations has no phenotypic associations to date and awaits further biochemical and virological characterization.

Individual host factors, such as the immune response and respiratory tract microbiome, may play an important role in viral persistence. In particular, because SARS-CoV-2 infection is slow to resolve, the adaptive immune response could drive within-host viral evolution as variants that can escape T-cell and antibody responses develop. Although we were underpowered to see specific evidence of an adaptive immune response being mounted against SARS-CoV-2 in the nasopharynx, detailed studies evaluating antibody and T-cell receptor repertoire changes throughout the course of infection could shed light on the role of immune pressure in the development of minor variants. Similarly, the relationship between bacterial colonization of the nasopharynx

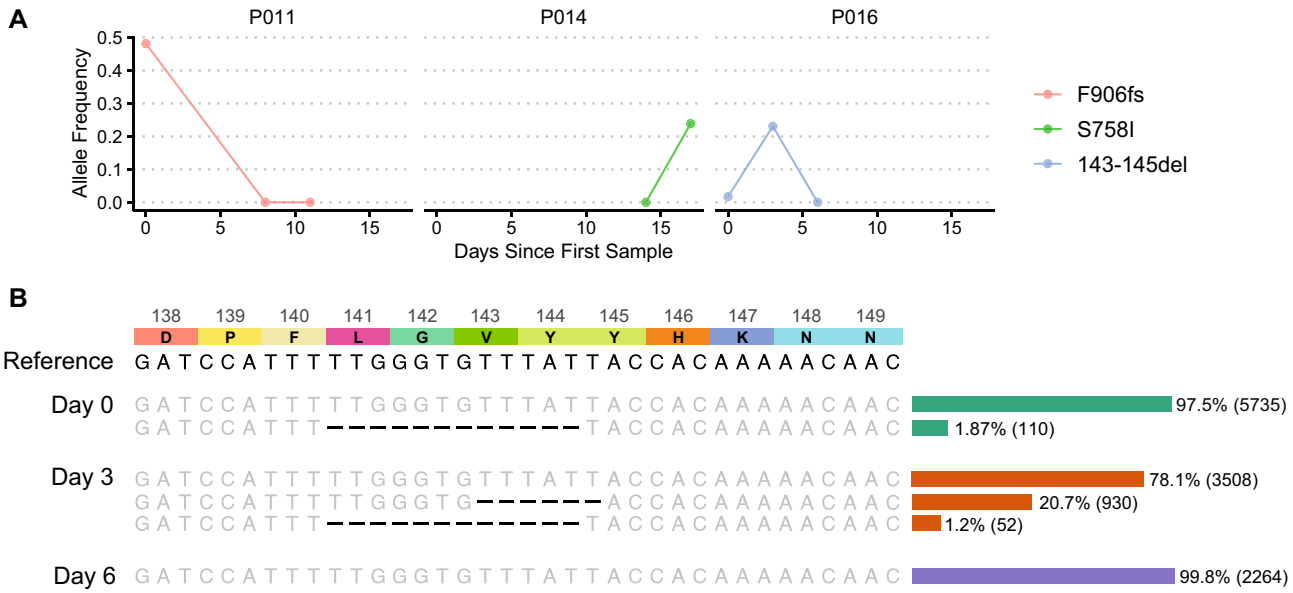

Figure 4. Variants that exhibit intra-host evolution in the spike protein across all patients. (**A**) All non-synonymous variants located in the spike protein with a ≥ 20% change in allele frequency among timepoints for any patient. (**B**) Enumeration of deletions that arose between residues 138–149 of the spike protein in P016 at ≥ 1% relative frequency reveals a rapidly changing complement of low frequency alleles present over a 6-day period. The reference nucleotide sequence (NC_045512) is located at the top of the sequence alignment. Above the reference is the corresponding amino acid sequence with associated residue numbers. Alleles that match the reference are in gray, and deletions are shown in black. To the right of the sequence alignment is a bar graph showing the square root of the relative frequency of each variant, for visualization purposes, labeled with the allele frequency in percentage and read depth in parentheses.

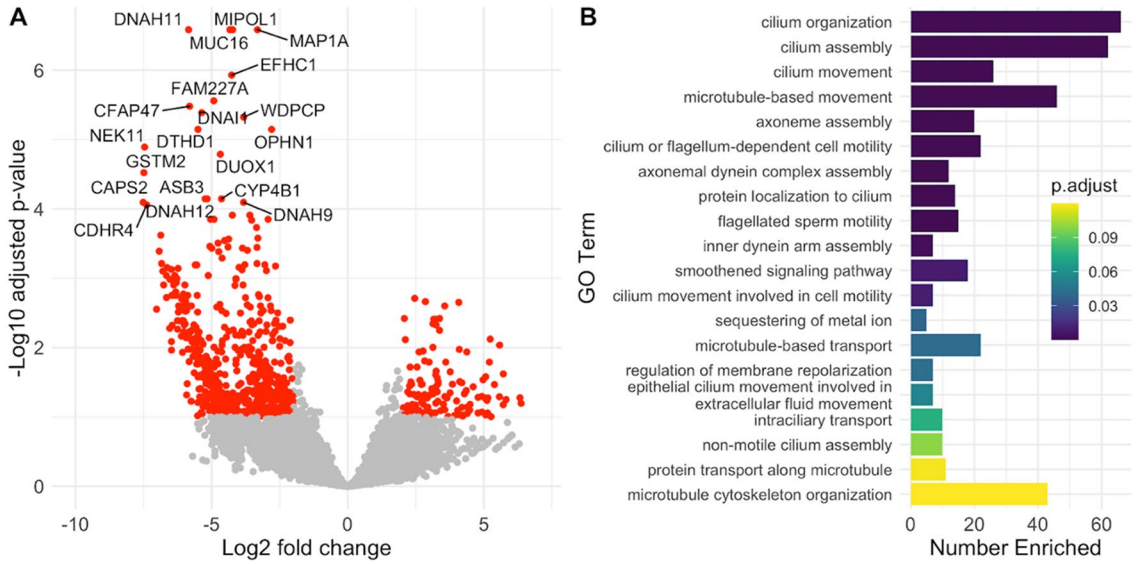

Figure 5. Differentially expressed genes during SARS-CoV-2 infection. (**A**) Twenty differentially expressed genes with lowest adjusted *p* value. Fold changes are of later samples relative to initial samples. Genes highlighted in red have a log2 fold change > 2 and an adjusted *p* value < 0.1. (**B**) Gene Ontology analysis reveals that differentially expressed genes are significantly enriched in biological processes related to microtubule-based motility. The twenty biological processes with the lowest adjusted *p* values are shown. The length of the horizontal bars corresponds to the number of DE genes in each GO category ("Number Enriched). Bar color corresponds to the adjusted *p* value for enrichment of DE genes in each pathway.

and the development or suppression of inflammation in response to SARS-CoV-2 infection remains poorly understood.

As viral load decreases during recovery, it becomes more challenging to recover viral genomes. As a result, one of the limitations of our study is the variability in sequencing depth across samples and the difficulty in ensuring similar sequencing depth for samples from different time points. We used an amplicon sequencing-based approach described previously[33] to obtain near full-length genomes from low viral load samples (up to Ct values of 36). We also used multiple library preparations and performed re-sequencing to ensure the accuracy of variant calls.

Taken together, our results suggest that low frequency genomic variants emerge in immunocompetent individuals, but that these variants are unlikely to reach fixation. Given the emergence of rapidly spreading variants of concern over the past several months, the limited intra-host evolution observed in our dataset highlights the critical impact that a select few individual intra-host evolutionary events may have on the course of the global pandemic and the need for continual genomic surveillance.

Methods

Sample collection and clinical testing for SARS-CoV-2. Specimens were obtained as part of clinical testing for SARS-CoV-2 ordered by local healthcare providers or collected at drive-through testing sites. RNA was extracted and the presence of SARS-CoV-2 was detected by RT-QPCR as previously described using either the emergency use-authorized UW CDC-based laboratory-developed test, Hologic Panther Fusion or Roche cobas SARS-CoV-2 tests[34,35]. Supplemental Table 1 contains clinically relevant details of these specimens.

Chart review and ethics approval. We received approval from the University of Washington Institutional Review Board (UW IRB) to use residual clinical specimens for sequencing and to review clinical records of patients who received care within the UW network. We also received a waiver of informed consent from the UW IRB for the use of residual clinical specimens and retrospective chart review to perform this work. Information obtained from medical records included sex, age, comorbidities, medication, hospital or critical care admission, and discharge status. All methods were carried out in accordance with relevant guidelines and regulations. 18 of the 20 total patients had one or more known comorbidities (Supplementary Table 1).

Sequencing and bioinformatic analysis. Sequencing was attempted on all samples with a positive RT-QPCR assay result that had a Ct ≤ 36 using one of three methods available at the time: (1) a shotgun metagenomic approach using Illumina Nextera XT described previously[36] for samples with a Ct less than 24; (2) an oligo-nucleotide probe capture-based approach from IDT (xGen NGS hybridization capture, Integrated DNA Technologies) similar to previous work[37] for samples with Ct between 24 and 28, or (3) using Swift Biosciences' Normalase Amplicon Panel library preparation workflow[38] for samples with Ct between 28 and 36. Libraries were sequenced on Illumina MiSeq, NextSeq, or NovaSeq instruments using 300, 150, 100, or 75 bp reads. Consensus sequences were assembled using TAYLOR[38], a custom bioinformatics pipeline (https://github.com/greninger-lab/covid_swift_pipeline) with or without an additional primer clipping step depending on library preparation method.

Consensus sequences from each individual were aligned with the reference sequence NC_045512 using MAFFT v7[39]. Clade assignments were generated using Pangolin (http://github.com/cov-lineages/pangolin) and Nextstrain[40] in December 2020. Consensus sequences with < 5% Ns across the length of the genome were considered for further analysis.

Variants were also called with TAYLOR from aligned reads. Variants leading to coding changes with a sequencing depth of > 100 and an allele frequency > 0.01 were subjected to further analysis. For variants that had an allele frequency of < 0.4, we included an additional quality-control step to only include mutations that were present across multiple samples. We also excluded mutations in the first 100 and last 50 bases, as well as variants determined to be due to sequencing error. Most samples were re-prepped and sequenced multiple times to ensure accuracy of variant calls. Variants at positions 6700, 11081-83, 19989, and 29056 were observed in a large number of samples but were determined to be the result of homopolymer sequencing error and were excluded.

RNAseq analysis. For samples that were prepared with no target enrichment step (metagenomic sequencing), reads were adapter and quality trimmed with Trimmomatic v0.39[41] using the call "leading 3 trailing 3 slidingwindow:4:15 minlen 20", then pseudoaligned to the hg38-derived human transcriptome using Kallisto v0.46[42]. Only samples with more than 900,000 reads pseudo-aligned to the human genome were used for analysis. Differential expression analysis using the Wald test was performed using DEseq2[43] and deemed significant at a Benjamini–Hochberg adjusted p value < 0.1. Statistical enrichment of Gene Ontology Biological Processes was performed on all significant genes using the R package clusterProfiler[44]. Raw counts were submitted to the Gene Expression Omnibus, accession number GSE173310.

Metagenomic analysis. Raw FASTQ files were analyzed using CLOMP v0.1.4 (https://github.com/FredHutch/CLOMP) as previously described[25]. Samples with more than 10 million reads were randomly down-sampled to 10 million reads before analysis using the "sample" command in seqtk (https://github.com/lh3/seqtk). The pipeline output was visualized using the Pavian metagenomic explorer[45], and reads per million (RPM) calculations were done using a custom R script. Results were filtered to highlight RPM counts for a shortlist of clinically relevant taxa (Supplementary Table S4). Samples were determined to be positive if the species level RPM was at least 30 for viruses, and 100 for bacteria.

Data availability

Consensus sequences were deposited to GISAID, and raw reads to SRA under BioProject PRJNA610428. GISAID accessions are available in Supplementary Table S2. Raw counts in RNASeq analysis were submitted to the Gene Expression Omnibus, accession number GSE173310. Figure generation code is available on GitHub (github.com/greninger-lab/longitudinal_sarscov2).

Received: 13 September 2021; Accepted: 16 March 2022
Published online: 07 April 2022

References

1. World Health Organization. *COVID-19 Weekly Epidemiological Update-67* (WHO, 2021).
2. Xue, K. S., Moncla, L. H., Bedford, T. & Bloom, J. D. Within-host evolution of human influenza virus. *Trends Microbiol.* **26**, 781–793 (2018).
3. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* https://doi.org/10.1126/science.abg0821 (2021).
4. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).
5. Sashittal, P., Luo, Y., Peng, J. & El-Kebir, M. *Characterization of SARS-CoV-2 Viral Diversity Within and Across Hosts* (2020). https://doi.org/10.1101/2020.05.07.083410.
6. Ramazzotti, D. *et al.* VERSO: A comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples. *Patterns* **2**, 100212 (2021).
7. Rose, R. *et al. Intra-Host Site-Specific Polymorphisms of SARS-CoV-2 is Consistent Across Multiple Samples and Methodologies* (2020). https://doi.org/10.1101/2020.04.24.20078691.
8. McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371**, 1139–1142 (2021).
9. Chen, L. *et al. Emergence of Multiple SARS-CoV-2 Antibody Escape Variants in an Immunocompromised Host Undergoing Convalescent Plasma Treatment* (2021). https://doi.org/10.1101/2021.04.08.21254791.
10. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* https://doi.org/10.1016/j.cell.2021.03.028 (2021).
11. Andreano, E. *et al. SARS-CoV-2 Escape In Vitro from a Highly Neutralizing COVID-19 Convalescent Plasma* (2020). https://doi.org/10.1101/2020.12.28.424451.
12. Lieberman, N. A. P. *et al.* In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age. *PLOS Biol.* **18**, e3000849 (2020).
13. van der Sluijs, K. F., van der Poll, T., Lutter, R., Juffermans, N. P. & Schultz, M. J. Bench-to-bedside review: Bacterial pneumonia with influenza—Pathogenesis and clinical implications. *Crit. Care Lond. Engl.* **14**, 219 (2010).
14. Nalla, A. K. *et al.* Comparative performance of SARS-CoV-2 detection assays using seven different primer-probe sets and one assay kit. *J. Clin. Microbiol.* **58**, e00557-e620 (2020).
15. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
16. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020).
17. Worobey, M. *et al.* The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
18. Korber, B. *et al.* Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812-827.e19 (2020).
19. Müller, N. F. *et al. Viral Genomes Reveal Patterns of the SARS-CoV-2 Outbreak in Washington State* (2020). https://doi.org/10.1101/2020.09.30.20204230.
20. Yurkovetskiy, L. *et al.* Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**, 739-751.e8 (2020).
21. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data—From vision to reality. *Eurosurveillance* **22**, 30494 (2017).
22. Avanzato, V. A. *et al.* Case study: Prolonged Infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. *Cell* **183**, 1901-1912.e9 (2020).
23. Maiti, A. K. *et al.* Identification, tissue specific expression, and chromosomal localisation of several human dynein heavy chain genes. *Eur. J. Hum. Genet.* **8**, 923–932 (2000).
24. Zhu, N. *et al.* Morphogenesis and cytopathic effect of SARS-CoV-2 infection in human airway epithelial cells. *Nat. Commun.* **11**, 3910 (2020).
25. Peddu, V. *et al.* Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral superinfection and colonization. *Clin. Chem.* **66**, 966–972 (2020).
26. Shen, Z. *et al.* Genomic diversity of severe acute respiratory syndrome-coronavirus 2 in patients with coronavirus disease 2019. *Clin. Infect. Dis.* **71**, 713–720 (2020).
27. Lythgoe, K. A. *et al. Within-Host Genomics of SARS-CoV-2* (2020). https://doi.org/10.1101/2020.05.28.118992.
28. Choi, B. *et al.* Persistence and evolution of SARS-CoV-2 in an immunocompromised host. *N. Engl. J. Med.* **383**, 2291–2293 (2020).
29. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e6 (2021).
30. Piccoli, L. *et al.* Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell* **183**, 1024-1042.e21 (2020).
31. Lei, J., Kusov, Y. & Hilgenfeld, R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antivir. Res.* **149**, 58–74 (2018).
32. Gandhi, S. *et al.* De Novo *Emergence of a Remdesivir Resistance Mutation During Treatment of Persistent SARS-CoV-2 Infection in an Immunocompromised Patient: A Case Report* (2021). https://doi.org/10.1101/2021.11.08.21266069.
33. Addetia, A. *et al.* Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates. *J. Clin. Virol.* **129**, 104523 (2020).
34. Perchetti, G. A. *et al.* Validation of SARS-CoV-2 detection across multiple specimen types. *J. Clin. Virol.* **128**, 104438 (2020).
35. Perchetti, G. A. *et al.* Pooling of SARS-CoV-2 samples to increase molecular testing throughput. *J. Clin. Virol.* **131**, 104570 (2020).
36. Greninger, A. L. *et al.* Rapid metagenomic next-generation sequencing during an investigation of hospital-acquired human parainfluenza virus 3 infections. *J. Clin. Microbiol.* **55**, 177–182 (2017).
37. Greninger, A. L. *et al.* Ultrasensitive capture of human herpes simplex virus genomes directly from clinical samples reveals extraordinarily limited evolution in cell culture. *mSphere* **3**, e00283-e318 (2018).
38. Addetia, A. *et al.* Sensitive recovery of complete SARS-CoV-2 genomes from clinical samples by use of swift biosciences' SARS-CoV-2 multiplex amplicon sequencing panel. *J. Clin. Microbiol.* **59**, e02226-e2320 (2020).
39. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

40. Hadfield, J. *et al.* Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
41. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
42. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
43. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
44. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **16**, 284–287 (2012).
45. Breitwieser, F. P. & Salzberg, S. L. Pavian: Interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics* **36**, 1303–1304 (2020).

Acknowledgements

We gratefully acknowledge the contributions of all originating and submitting laboratories of deposited sequences on GISAID. This work is supported by the National Institutes of Health [ORIP Grant S10OD028685 to Scientific Computing at Fred Hutch, and UW-Fred Hutch CFAR AI027757 to PR].

Author contributions

M.J.L. and P.R. managed and conceived the project. V.M.R., H.X., L.S., N.A.P.L., and A.A. performed the experiments. M.J.L., V.M.R., N.A.P.L., VP, and P.R. performed bioinformatics analyses. M.J.L., V.M.R., A.A., A.M.C., N.B., P.C.M., M.-L.H., K.R.J., A.L.G., and P.R. contributed to project discussion and interpreted the results. All authors read and approved the final manuscript.

Competing interests

ALG reports contract testing from Abbott Laboratories and research support from Gilead and Merck, outside of the described work. All other authors report no other financial or non-financial competing interests for this work.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-09752-2.

Correspondence and requests for materials should be addressed to K.R.J., A.L.G. or P.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.