Virus Genes
https://doi.org/10.1007/s11262-023-02011-0

ORIGINAL PAPER

Intra-host mutation rate of acute SARS-CoV-2 infection during the initial pandemic wave

Kim El-Haddad[1] · Thamali M. Adhikari[2] · Zheng Jin Tu[3] · Yu-Wei Cheng[3] · Xiaoyi Leng[2] · Xiangyi Zhang[2] · Daniel Rhoads[3] · Jennifer S. Ko[3] · Sarah Worley[4] · Jing Li[2] · Brian P. Rubin[3] · Frank P. Esper[1]

Received: 17 April 2023 / Accepted: 22 May 2023

Abstract

SARS-CoV-2 mutation is minimized through a proofreading function encoded by *NSP-14*. Most estimates of the SARS-CoV-2 mutation rate are derived from population based sequence data. Our understanding of SARS-CoV-2 evolution might be enhanced through analysis of intra-host viral mutation rates in specific populations. Viral genome analysis was performed between paired samples and mutations quantified at allele frequencies (AF) $\geq 0.25, \geq 0.5$ and ≥ 0.75. Mutation rate was determined employing F81 and JC69 evolution models and compared between isolates with (ΔNSP-14) and without (wtNSP-14) non-synonymous mutations in NSP-14 and by patient comorbidity. Forty paired samples with median interval of 13 days [IQR 8.5–20] were analyzed. The estimated mutation rate by F81 modeling was 93.6 (95%CI 90.8–96.4], 40.7 (95%CI 38.9–42.6) and 34.7 (95%CI 33.0–36.4) substitutions/genome/year at AF $\geq 0.25, \geq 0.5, \geq 0.75$ respectively. Mutation rate in ΔNSP-14 were significantly elevated at AF ≥ 0.25 vs wtNSP-14. Patients with immune comorbidities had higher mutation rate at all allele frequencies. Intra-host SARS-CoV-2 mutation rates are substantially higher than those reported through population analysis. Virus strains with altered NSP-14 have accelerated mutation rate at low AF. Immunosuppressed patients have elevated mutation rate at all AF. Understanding intra-host virus evolution will aid in current and future pandemic modeling.

Keywords SARS-CoV-2 · COVID-19 · Mutation rate · Allele frequency · NSP-14 · SNV

Background

Since the onset of the pandemic, evolution of SARS-CoV-2 has produced multiple variants associated with changes in disease severity and transmission dynamics [1]. The rate of SARS-CoV-2 mutation is commonly estimated using population based approach predicted through phylogenetic analysis of global databases comprised of unrelated virus sequences submitted ad hoc[2, 3]. This population-based rate began at a modest 21.9 substitutions/genome/year in the initial months and is now estimated at ~ 28.4 substitutions/genome/year [4]. However, few studies quantify the viral mutation rate within infected individuals. Our understanding of this virus' ability to mutate during the course of an infection and those viral and host factors which may impact this rate is minimal.

The mutation rate of SARS-CoV-2 genome is slower than most RNA viruses predominantly through the action of nonstructural protein 14 (NSP-14) [5]. NSP-14 is present in all coronaviruses and contains an *N*-terminal exonuclease (ExoN) domain providing replication fidelity for the RNA dependent RNA polymerase [6–8]. Mutagenesis or inactivation of ExoN was shown to increase genomic diversity and creates a "mutator phenotype," leading to a 15- to 21-fold rise in mutations during replication but adversely affects viral fitness [7, 9].

Edited by Juergen Richt.

✉ Kim El-Haddad
elhaddk@ccf.org

✉ Frank P. Esper
esperf@ccf.org

[1] Center for Pediatric Infectious Disease, Cleveland Clinic Children's, R3, 9500 Euclid Avenue, Cleveland 44195, OH, USA

[2] Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, USA

[3] Robert J. Tomsich Pathology and Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH, USA

[4] Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA

Published online: 13 June 2023

Additionally, viral mutation can be influenced by host factors [10, 11]. Descriptions of higher mutation rates occurring in immunosuppressed individuals with chronic SARS-CoV-2 infection have been reported [12–14]. Consequently, there is concern that novel variants may emerge within such hosts [15]. However, there is little data on SARS-CoV-2 mutation dynamics in patients with cardiovascular diseases, endocrine disorders or pulmonary conditions despite also having more severe and prolonged infections [16].

Here, we perform paired analysis of SARS-CoV-2 infected patients and calculate the intra-host mutation rate. Additionally we determine if changes to NSP-14 or host comorbidity alter this rate. Insight on this virus's ability to evolve has importance for accurate prediction modelling in current and future coronavirus pandemics [17].

Methods

Sample identification and collection

Patient samples were identified through The Cleveland Clinic Pathology and Laboratory Medicine Institute (PLMI) SARS-CoV-2 variant surveillance project[1]. Selected samples focused on the period of the initial pandemic wave between 3/17/2020 and 5/27/2020. By selecting this period, we aimed to minimize external factors which could influence the viral mutation rate. This selected period had limited treatment options beyond supportive care and preceded preventive measures such as immunization and monoclonal antibodies. Additionally, it was unlikely that individuals had prior immunity to SARS-CoV-2 during this time frame.

Adults age $\geq$ 18 years with multiple positive nasopharyngeal samples occurring within 5 to 60 days of initial screening were identified. This interval time frame was selected to prevent skewing of model results from short sampling intervals while further minimizing chance of re-infection [10, 18]. Only pairings where initial and subsequent samples had cycle threshold (CT) $\leq$ 30 were included to ensure high quality genomic sequencing. Children < 18 years were excluded as identification of SARS-CoV-2 in children during the initial pandemic wave was minimal. Those specimens with an indeterminate result, obtained from locations other than the nasopharynx, or whose samples contained discordant viral lineages (suggesting reinfection) were also excluded.

Patient comorbidities were identified through the COVID-19 registry [19]. Patients were classified into four comorbidity categories: Endocrine (obesity and diabetes mellitus), cardiac (hypertension and coronary artery disease), pulmonary (asthma, obstructive sleep apnea and COPD) and immunologic (autoimmune diseases, history of prior/current cancer and current immunosuppression therapy). Sample collection and medical review is approved by the Internal Review Board at Cleveland Clinic.

Library preparation and sequence data analysis

Following patient identification, initial and subsequent nasopharyngeal samples were retrieved from Biobank freezers housed at PLMI and processed for viral genome analysis though next generation sequencing (NGS). Total nucleic acids were purified from each specimen and subjected to reverse transcription (RT), NGS library preparation, sequencing, and data analysis according to the manufacturer's recommendation (Paragon Genomics, Hayward CA) [20]. Briefly: Total RNA from SARS-CoV-2 was converted into complementary deoxyribonucleic acid (cDNA) synthesis via RT in 20 µL reactions (10 min at 8 °C and 80 min at 42 °C). The derived panel of 343 amplicons utilized for SARS-CoV-2 enrichment covers 99.7% of the viral genome (MN908947/NC_045512.2) with 92 bases uncovered at each end. Purified cDNA was subject to multiplex PCR (10 min at 95 °C, followed by 10 cycles at 98 °C for 15 s each and 60 °C for 5 min). Excess primers and oligonucleotides were subsequently removed from the purified PCR products, after which a second round of PCR to append indexing primers was performed (initial denaturation, 10 min at 95 °C, followed by 24 cycles of 98 °C for 15 s and 60 °C for 75 s). Sequencing libraries were then prepared and quality was assessed visually using an Agilent® 2100 Bioanalyzer® (Agilent, Santa Clara CA). The presence of a ~ 275 bp peak indicated successful amplification and these libraries were then sequenced using a MiSeq instrument (Illumina, San Diego, CA). Raw fastq reads was extracted by Illumina bcl2fastq (v2.20.0) and mapped to the reference genome Wuhan-Hu- 1 (NC_045512.2) using BWA program [21]. Variants were called using FreeBayes program [22] and filtered at 5% and 10% allele fractions for insertion or deletion (INDEL) and single nucleotide variants (SNV), respectively. Amino acid changes were annotated using snpEff (v4.5) program [23]. All variant data was visually examined in Integrative Genome Browser (IGV, version 2.11.0) [24] to eliminate artifacts. Quality was ensured by monitoring mapping quality, phred score, and manual review. Sequences used in this study were submitted to GISAID database and accession numbers are available upon request.

Variant calling

Analysis of SARS-CoV-2 mutations within a host during the course of an infection have been highly variable and are affected by sequencing protocols and data analysis parameters (i.e. variant-calling) [15, 25]. Variant calling methodology is strongly dependent on the library protocol and

sequencing technology and requires tuning of parameters to distinguish true variants from false positive calls [26]. Variant calling was expanded from established WHO criteria [27] and was performed by manual review of each SNV by three independent investigators through IGV [24]. We used a minimum depth of ≥ 100 reads at each position for all samples and quantified SNV at 3 separate allele frequencies (AF ≥ 0.25, AF ≥ 0.5, and AF ≥ 0.75). AF was defined as the proportion of SNV in the sample reads. Mutation change represents the discordance in SNVs between initial and the subsequent samples at each AF. In addition, SNVs below 0.25 AF and those mutations where investigator consensus was not achieved were excluded from the analysis.

ΔNSP-14 and wt NSP-14 group assignment

Following viral genome analysis, patients with any isolates (either initial or subsequent) containing non-synonymous mutations of NSP-14 (ORF 1a/b amino acid position 5930 to 6450) were placed in the ΔNSP-14 group. As our understanding of SARS-CoV-2 NSP-14 is evolving, no weight was given to mutation types (Missense vs frameshift vs nonsense) or location within NSP-14 (active vs structural site).

Calculation of genome mutation rate

Changes in genome between initial and subsequent samples were quantified for each pair and used for calculation of mutation rate (standardized to mutations/genome/year) through both F81 and JC69 models. Given the limited sample size, we chose to employ two mutation models (F81 and JC69) for estimating the overall substitution rates among samples. These two models were selected because they assume uniform mutation rates across nucleotides (A,T,C,G), thereby reducing the number of parameters required [28, 29]. JC69 also assumes equal base frequencies indicating that mutation rate is assumed to be constant over time and across all nucleotide changes. Whereas F81 allows for variable base frequencies with equal over time providing a more realistic calculation of the mutation rate. For both models, mutation rates were estimated by the use of maximum likelihood algorithms. Hereafter, the results detail findings from the F81 model while results detailing findings from the JC69 analysis appear in the supplementary materials.

F81 model derivation

For each of the n patients, we obtained two virus specimens at different time points and the time interval is denoted as t_k for patient k. To obtain the maximum likelihood estimate of the mutation rate based on the evolutionary model F81, we assume all the patients are independent. Therefore, the likelihood of the data (L) is the product of the likelihood (L_k) of each patient k, measuring the probability of observing the sequence evolving over time t_k. Because for each patient, both initial and subsequent sequences were available, under the assumption that all the nucleotides are independent, the probability L_k is the product of the probability over all nucleotides. Under the model F81, the probability that a nucleotide i ($i \in \{A, T, G, C\}$) remains unchanged over time t is

$$P_{ii}(\mu t) = e^{-\mu t} + p_i(1 - e^{-\mu t})$$

and the probability of a nucleotide i to change to a nucleotide j over time t is

$$P_{ij}(\mu t) = p_j(1 - e^{-\mu t})$$

where u is the mutation rate per nucleotide per year, and p_i is the frequency of nucleotide i. Let $l_{(\mathrm{ij}),\mathrm{k}}$ denote the number of nucleotide i changed to nucleotide j for patient k (in the case of i is the same as j, the nucleotide remains unchanged), the overall likelihood can thus be represented as

$$L = \prod_{k=1}^{n} L_k = \prod_{k=1}^{n} \prod_{i=A}^{T} \prod_{j=A}^{T} \left[p_{ik}.P_{ij}\left(\mu t_k\right)\right]^{l_{(ij),k}}$$

where p_{ik} is the frequency of nucleotide i in the first specimen of the kth patient (in practice, these frequencies are very similar to the frequencies from the SARS-CoV2 reference sequence). The log likelihood is

$$l = log(L) = C + \sum_{k=1}^{n} \sum_{i=A}^{T} \sum_{j=A}^{T} l_{(ij),k} log\left(P_{ij}(\mu t_k)\right)]$$

The maximum likelihood estimate cannot be obtained analytically. We relied on the Newton–Raphson method [30], which iteratively updates the new value of the mutation rate u until convergence.

The detailed derivations for both F81 and JC69 models can be found in the supplementary methods.

Statistical analysis

Continuous variables were described using median and range; categorical variables were described using frequency and percentage. Demographics and variant characteristics were compared between patients in different virus groups by using ANOVA or Wilcoxon rank sum tests for continuous variables and Fisher's exact or Pearson's chi-square tests for categorical variables. The estimated mutation rates from two different groups are compared using the t-test, assuming the maximum likelihood estimates follow approximately a

Table 1 Patient demographics of paired SARS-CoV-2 isolates

	Total	wt NSP-14	Δ NSP-14	p-value
Total pairs	40	28 (70.0%)	12 (30.0%)	
Median interval (days) [IQR]	13 [8.5, 20]	13 [8.5, 20]	14 [8.5, 20]	0.72 [b]
Demographics				
Median Age (yr) [IQR]	54 [31, 66]	56 [31, 69]	53 [32, 62]	0.65[b]
Males	20 (50.0%)	14 (50.0%)	6 (50.0%)	0.99[c]
Race*				0.46[d]
White	26 (67.0%)	16 (59.0%)	10 (83.3%)	
African American	10 (26.0%)	8 (30.0%)	2 (16.7%)	
Asian	3 (7.5%)	3 (11.0%)	0 (0%)	
Comorbidity				
Any	28 (70.0%)	19 (67.9%)	9 (75.0%)	0.72[d]
Endocrine	23 (57.5%)	14 (50.0%)	9 (75.0%)	0.14[c]
Cardiac	17 (42.5%)	12 (42.9%)	5 (41.7%)	0.94[c]
Pulmonary	8 (20.0%)	5 (17.9%)	3 (25.0%)	0.68[d]
Immune/oncologic	6 (15.0%)	4 (14.3%)	2 (16.7%)	0.99[d]

*Data not available for all subjects. Missing values: Race = 1.

Statistics presented as Median [P25, P75], N (column %).

P-values: b=Wilcoxon Rank Sum test, c=Pearson's chi-square test, d=Fisher's Exact test

normal distribution. The confidence interval of the estimated mutation rate is calculated based on the maximum likelihood estimate following approximately a normal distribution N(u, 1/I(u)), where u is the true value, and I(u) is the Fisher information. PRISM software (version 8.4.3, GraphPad Software, San Diego, CA) and Python (version 3.7.4) with statsmodel package (version 0.13.2, for construction of ML models) was used for analysis.

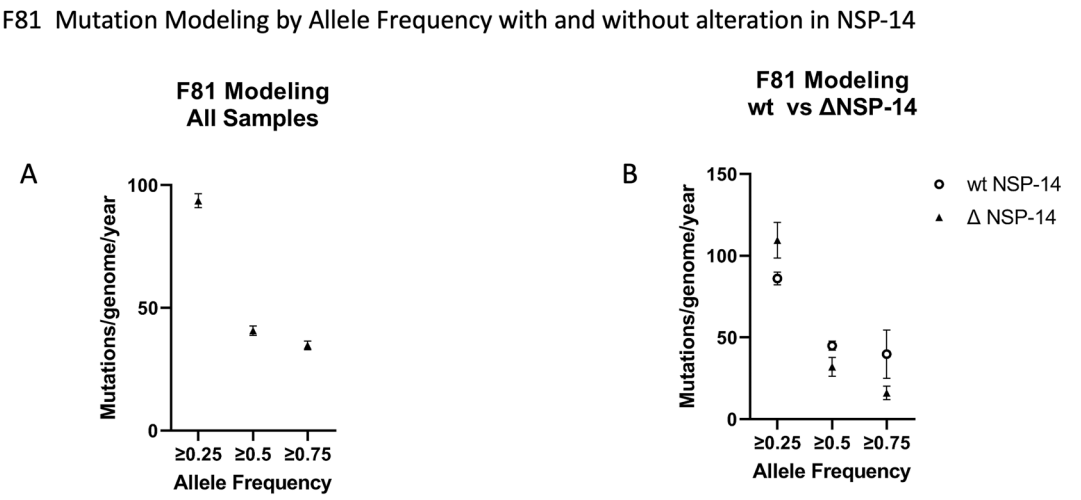
Results

From 3/17/2020 through 5/27/2020, a total of 40 paired nasopharyngeal samples (initial and subsequent) from acutely infected individuals with SARS-CoV-2 were identified and retrieved from the COVID19 biobank. Median days between paired tests was 13 days [IQR 8.5–20]. Median patient age was 54 years [IQR 31, 66], included 20/40(50.0%) males with 26/40 (67.0%) being white, and with 28/40 (70.0%) having at least one comorbidity (Table 1). Comorbidities included endocrine 23/40 (57.5%), cardiac 17/40 (42.5%), pulmonary 8/40 (20.0%) and Immune/Oncologic 6/40 (15.0%).

SARS-CoV-2 genomes of each pair were sequenced and mapped against the reference Wuhan strain (Wuhan-Hu-1, NC_045512.2). SNVs were identified for each pairing through IGV and filtered at allele frequencies (AF) ≥ 0.25, ≥ 0.5 and ≥ 0.75. A total of 120 SNVs changes between initial and subsequent samples were identified at AF ≥ 0.25, 53 at AF ≥ 0.5 and 33 at AF ≥ 0.75 (Table 2). The majority of SNV changes were gained over the course of the infection (93/120 (77.5%), 32/53 (60.4%), 18/33 (54.8%) at AF ≥ 0.25, ≥ 0.5, ≥ 0.75 respectively). Predominant SNVs were missense with most occurring in the ORF1a/b region and the spike protein region. While more SNVs were gained at low AF, there was no substantial difference between SNV types or gene location among different AF.

Table 2 Type and Location of SARS-CoV-2 Intra-host SNVs by Allele Fraction

	AF ≥ 0.25	AF ≥ 0.5 %	AF ≥ 0.75 %
SNV changes	120	53 (44.2)	33 (35.0)
Mutations Gained	93 (77.5%)	32 (60.4)	18 (54.8)
Mutations Lost	27 (22.5%)	21 (39.6)	15 (45.2)
Missense	71 (59.2%)	36 (67.9)	23 (69.7)
Silent	30 (25.0%)	11 (20.8)	7 (21.2)
INDEL	2 (1.6%)	2 (3.8)	1 (3.0)
Other	17 (14.2%)	4(7.5)	2 (6.1)
ORF1 a/b	82 (68.3%)	36 (67.9)	26 (61.9)
ORF3	4 (3.3%)	3 (5.7)	3 (7.1)
ORF6	2 (1.7%)	1 (1.9)	1 (2.4)
ORF7	1 (0.8%)	0 (0)	0 (0)
ORF8	3 (2.5%)	2 (3.8)	2 (4.8)
ORF10	1 (0.8%)	0 (0)	0 (0)
Spike	16 (13.3%)	6 (11.3)	5 (11.9)
Membrane	2 (1.7%)	1 (1.9)	1 (2.4)
Envelope	0 (0%)	0 (0)	0 (0)
Nucleocapsid	6 (5.0%)	4 (7.5)	4 (9.5)
Untranslated region (UTR)	3 (2.5%)	0 (0)	0 (0)



		Total (n=40)	wt NSP-14 (n= 28)	ΔNSP-14 (n=12)	
		Mutation rate (Subs/genome/year) [95% CI]	Mutation rate (Subs/genome/year) [95% CI]	Mutation rate (Subs/genome/year) [95% CI]	p-value
F81	AF ≥ ≥0.25	93.6 [90.8-96.4]	86.0 [82.1-89.9]	109.4 [99.7-119.1]	<0.001
	AF ≥ ≥0.5	40.7 [38.9-42.6]	44.9 [42.1-47.7]	32.0 [26.8-37.2]	<0.001
	AF ≥ ≥0.75	34.7 [33.0-36.4]	39.8 [25.0-54.5]	16.0 [7.0-25.1]	<0.001

Fig. 1 F81 Mutation Modeling by Allele Frequency with and without alteration in NSP-14. Graphic representation of F81 evolution modeling at AF ≥ 0.25, ≥ 0.5, ≥ 0.75 of **A** total patient sample and **B** comparison between wt and ΔNSP-14. Bars represent 95%CI. Table displaying data for F81 modeling is displayed below. P-values displayed represent comparison of wt and ΔNSP-14 groups

We identified 12/40 (30.0%) pairs with a non-synonymous mutation in NSP-14 (ΔNSP-14). Median age, gender, race and comorbidities were similar between both groups. Of NSP-14 mutations (n = 15), 11/15 (73.3%) were identified in the subsequent sample only while the remainder 4/15 (26.7%) occurred in both initial and subsequent samples. For both ΔNSP-14 and wtNSP-14 groups, the majority of SNVs were gained over the course of infection. Mutation types and locations were similar between groups (supplementary table 1 and 2).

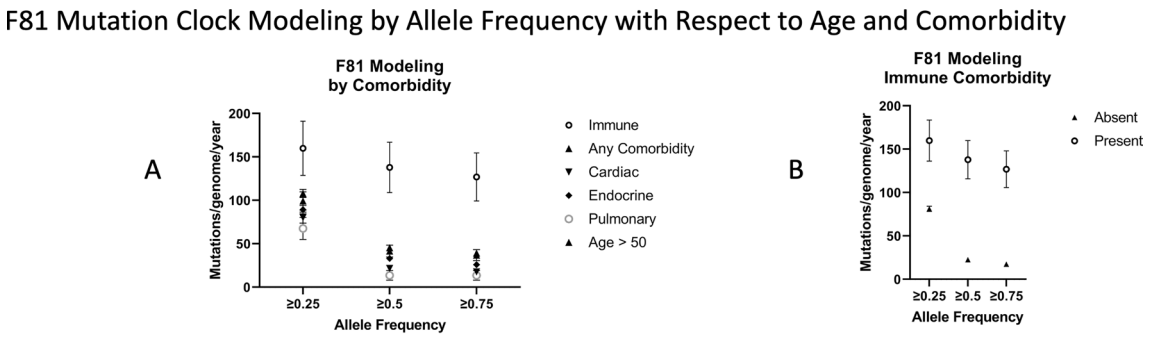
Mutation rates were calculated through the F81 and JC69 models (Fig. 1, supplementary Fig. 1 for JC69). Focusing on F81 modeling, the mutation rate from all samples was found to be 93.6 substitutions/genome/year [95%CI 90.8–96.4] at AF ≥ 0.25, 40.7 [95% CI 38.9–42.6] at AF ≥ 0.5 and 34.7 [95%CI 33.0–36.4] at AF ≥ 0.75. Mutation rate of ΔNSP-14 were significantly higher at low AF compared to wtNSP-14 group (109.4 [95%CI 99.7–119.1] vs 86.0 [95%CI 82.1–89.9] substitutions/genome/year, p-value < 0.001). Interestingly, mutation rates were lower in ΔNSP-14 compared to wtNSP-14 both at AF ≥ 0.5 (32.0 [95% CI 26.8–37.2] vs 44.9 [95% CI 42.1–47.7] substitutions/genome/year, p-value < 0.001) and at AF ≥ 0.75 (16.0 [95% CI 7.0–25.1] vs 39.8 [95% CI 25.0–54.5] substitutions/genome/year, p-value < 0.001).

Lastly, patients with underlying immunologic/oncologic comorbidities had substantially higher mutation rates than other comorbidities at all three AF (Fig. 2, supplementary Fig. 2 for JC69). Mutation rates in patients with immunologic/oncologic comorbidities were 160 [95% CI 136.2–183.7] vs 81.2 [95% CI 78.1- 84.2] substitutions/genome/year at AF ≥ 0.25, 137.9 [95% CI 115.8–160.0] vs 22.6 [95% CI 21.0–24.2] at AF ≥ 0.5 and 126.9[95% CI 105.7–148.0] vs 17.4 [95%CI 16.0–18.9] at AF ≥ 0.75.

Overall mutation rates calculated through JC69 modeling were comparable to those with F81 at all three AF (supplementary Fig. 3). Results based on JC69 modeling are presented in Supplementary Figs. 1 and 2.

Discussion

SARS-CoV-2 has lead to the emergence of new variants adversely affecting pandemic response [31]. The mutation rate commonly cited is calculated through analysis of unrelated regional and global sequences. These population based



	Allele Frequency	Mutation Rate (subs/genome/yr)	95% CI
Total (n=40)	AF ≥0.25	93.6	90.8-96.4
	AF ≥0.5	40.7	38.9-42.6
	AF ≥0.75	34.7	33.0-36.4
Age > 50 (n=22)	AF ≥ 0.25	99.3	88.8-109.8
	AF ≥ 0.5	41.6	34.8- 48.5
	AF ≥ 0.75	36.8	30.4- 43.3
Medical Comorbidity (n=28)	AF ≥ 0.25	107.9	103.4-112.3
	AF ≥ 0.5	45.4	42.5-48.2
	AF ≥ 0.75	39.2	36.6-41.2
Endocrine (n=23)	AF ≥ 0.25	89.2	84.6-93.6
	AF ≥ 0.5	33.1	30.3-35.9
	AF ≥ 0.75	25.9	23.4-28.4
Cardiac (n=17)	AF ≥ 0.25	80.1	74.2-86.1
	AF ≥ 0.5	21.5	18.4-24.6
	AF ≥ 0.75	17.6	14.8-20.4
Pulmonary (n=8)	AF ≥ 0.25	67.54	57.0-78.0
	AF ≥ 0.5	13.5	8.8-18.2
	AF ≥ 0.75	13.5	8.8-18.2
Immune/Oncologic (n=6)	AF ≥ 0.25	160.0	136.2-183.7
	AF ≥ 0.5	137.9	115.8-160.0
	AF ≥ 0.75	126.9	105.7-148.0

Fig. 2 F81 Mutation Clock Modeling by Allele Frequency with Respect to Age and Comorbidity. Graphic representations of mutation rates at AF ≥ 0.25, ≥ 0.5, ≥ 0.75 for **A** age and comorbidities and **B** those with and without immunologic/oncologic comorbidity. Bars represent 95%CI. Table displaying data for F81 modeling is displayed below

rates have ranged from 21.6 to 28.4 substitutions/genome/year [4]. The rate of evolution of SARS-CoV-2 was seemingly slow during the pandemics initial months with several reports suggesting the virus acquiring only two mutations per month [32, 33]. However, recently the viral mutation rate has accelerated and now lies at its fastest point with the emergence of the Omicron variant [34].

Here, we analyze intra-host viral mutation rates at multiple allele frequencies to better characterize and understand the capacity for SARS-CoV-2 to evolve following its initial introduction and prior to external influence by antivirals, vaccinations and prior immunity. While intra-host mutation dynamics have been previously described [35], the intra-host mutation rate over the course of an infection, needed for pandemic prediction has been poorly studied. We find the intra-host mutation rate was over 50% greater than estimated through population based surveillance at AF ≥ 0.75 (the WHO standard). Moreover, the intra-host mutation rate may be even higher if analyses include SNV with lower AF, 80% higher at AF ≥ 0.5 and nearly 350% greater at AF ≥ 0.25. Recognition that low frequency SNV contribute substantially to the estimated viral mutation rate, especially if these mutations serve as a reservoir for the generation of dominant mutations, may aid in our understanding of the evolutionary dynamics of this virus and could be useful in planning response to future coronavirus pandemics [36, 37].

By analyzing the genomic changes at lower AF, our study provides a better appreciation of intra-host SARS-CoV-2 diversity. We find the highest diversity at lowest AF (≥ 0.25) demonstrating that potential SNVs occur nearly 4 times higher than when analyses are performed at the AF utilized by the WHO (AF ≥ 0.75). Fitness of these low frequency SNVs and their effect on transmission remains poorly understood. Current literature is skeptical of significant person to person spread of low AF SNVs and report only rare transmission recognized among individuals within the same household [15, 25, 38]. However, it is suggested that accelerated intra-host episodic increases in mutation rate (~ fourfold higher than the background substitution rate) may drive the emergence of variants of concern [39]. We hypothesize that low AF SNVs could play a role in such a process.

Prior studies report that alteration in NSP-14 is associated with increased mutation load across the genome compared to other NSP changes [40]. NSP-14 is vital for survival of various coronaviruses including SARS-CoV-2 [41]. Inactivating NSP-14-ExoN in murine hepatitis virus (MHV-CoV) significantly altered recombination patterns and decreased recombination frequency compared with wild-type MHV-CoV [7]. While virus diversity has been found to contribute to disease severity in certain coronaviruses [36], further studies showed ExoN knockout mutants of MERS-CoV and SARS-CoV-2 are nonviable, suggesting excess mutation may have a deleterious effect [8, 42]. Our findings are consistent with this. While the mutation rate is significantly higher in ΔNSP-14, such change occurs only at low AF. This suggests SARS-CoV-2 viruses with altered NSP-14 may be less fit or are less likely to undergo purifying selection [41]. In this regard, SARS-CoV-2 NSP-14 is now being evaluated as a potential therapeutic target [7, 9].

Lastly, SARS-CoV-2 diversity and clinical outcome are influenced by host environment [37]. There is evidence that SARS-CoV-2 morbidity is worse in patients with underlying comorbidities. To date, there is little data on SARS-CoV-2 mutation rate in patients with cardiac, endocrine or pulmonary comorbidities whereas several reports describe elevated mutation rates in immunosuppressed individuals [12–14]. Prolonged viral shedding can occur in immunocompromised patients allowing increased time to for virus to acquire mutations [43]. In one example of a patient suffering from advanced lymphocytic leukemia and B-cell lymphoma, SARS-CoV-2 shedding was observed for 471 days. During this infection an unusually high number of mutations was detected and the mutation rate was calculated at 35.6 (95% CI: 31.6–39.5) substitutions per year through the Bayesian Skyline Model [44]. In our study, we included patients with several comorbidities and find that only those viruses originating from hosts with immune comorbidities had accelerated mutation rates [45]. This suggests that impaired host immune responses may contribute to intra-host viral evolution [44]. Better delineation of specific immune factors associated with alteration of evolutionary rate are needed.

There are several limitations to this study. First, while our pilot investigation of 40 SARS-CoV-2 patient pairs demonstrated substantially higher mutation rate than commonly reported, further analysis with larger cohorts would improve accuracy. Similarly, patients were grouped in broad comorbidity categories rather than by more specific underlying disease. Studies with greater characterization of underlying comorbidities, particularly immune, will provide a better picture of host factors associated with alteration in SARS-CoV-2 mutation [43, 46]. While a cutoff AF $\geq$ 0.75 was based on WHO guide for global variant surveillance, the significance of lower frequency SNVs remains unclear. This study sheds more light on the virus diversity present at lower AF thresholds. By analyzing viral isolates obtained from the initial pandemic wave, our study determined the intra-host mutation rate of SARS-CoV-2 in the absence of influence from external factors (e.g. antiviral medications, monoclonal antibody therapy, immunization, and natural immunity from prior infection). Determining the effect of pharmacologic interventions, immunization and previous infection on the mutation rate of subsequent SARS-CoV-2 isolates is a logical next step. Additionally, analysis of subsequent SARS-CoV-2 variants (Alpha, Delta, and Omicron) with parameter rich models such as HKY or GTR are currently being planned. Lastly, placement of patients within wt and ΔNSP-14 groups occurred without association to SNV location within the gene. It is possible that several NS mutations placed in this group did not substantially affect exonuclease function. Further study focusing on those SNVs with a defined effect on NSP-14 activity are needed [46].

Conclusion

Our study demonstrates the intra-host mutation rate of SARS-CoV-2 is substantially higher than previously reported through population-based analysis. In addition, low frequency intra-host mutations may be an important reservoir contributing to possible future variant emergence. When an NSP-14 variant was detected, an increased mutation rate was observed but only at low AF. In addition, among immunocompromised patients, but not patients with other co-morbidities, elevated mutation rates were observed at all AF. Understanding SARS-CoV-2 intra-host evolutionary dynamics may have important implications for pandemic planning, vaccine development and antiviral therapy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11262-023-02011-0.

Acknowledgements We appreciate Daniel H. Farkas, PhD and Charles Foster MD for their kind insight and thoughtful review of the project

Author contributions KEH, FE, and BR conceptualized and directed this research. TA, XL, XZ and JL, developed methodology, and performed evolutionary modeling and mutation statistics. TJ and YC assisted in sample acquisition, Illumina sequencing and pipeline development. DR and JK assisted in study design, sample identification and acquisition. SW assisted in statistics review. All authors contributed to discussions and manuscript preparation.

Funding This research was supported through the Ellen and Steven Ross Fellowship Research Award, Cleveland Clinic Children's IF-110077. This project was supported in part by NSF IIS-2027667 and NSF CCF-2200255 (JL and FE), NSF CCF-2006780 (JL), NSF CCF-1815139 (JL) and through unrestricted funds from the Robert J. Tomsich Pathology and Laboratory Medicine Institute. Ellen and Steven Ross Fellowship Research Award, IF-110077, IF-110077, IF-110077, IF-110077, IF-110077,IF-110077, National Science

Foundation,CCF-1815139, National Science Foundation,United States, CCF-2006780, CCF-2200255

Declarations

Competing interests DDR performs collaborative research that is sponsored by industry collaborators: BD, bioMerieux, Cepheid, Cleveland Diagnostics, Hologic, Luminex, Q-Linea, Qiagen, Roche, Specific Diagnostics, Thermo Fisher, and Vela. DDR is or has been on advisory boards for Luminex, Talis Biomedical, and Thermo Fisher. FE has served as a consultant to Proctor & Gamble. The remaining authors have or do not have an association that might pose a conflict of interest.

References

1. Esper FP, Cheng Y-W, Adhikari TM, Tu ZJ, Li D, Li EA, Farkas DH, Procop GW, Ko JS, Chan TA, Jehi L, Rubin BP, Li J (2021) Genomic epidemiology of SARS-CoV-2 infection during the initial pandemic wave and association with disease severity. JAMA Netw Open 4:e217746
2. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121
3. Mercatelli D, Holding AN, Giorgi FM (2021) Web tools to fight pandemics: the COVID-19 experience. Brief Bioinform 22:690
4. Nextstrain. https://nextstrain.org/ncov/gisaid/global/6m?l=clock. Accessed April 2023
5. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, Rocchi P, Ng W-L (2020) Coronavirus RNA proofreading: molecular basis and therapeutic targeting. Mol Cell 79:710
6. Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, Lou Z, Yan L, Zhang R, Rao Z (2015) Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. Proc Natl Acad Sci U S A 112:9436
7. Tahir M (2021) Coronavirus genomic nsp14-ExoN, structure, role, mechanism, and potential application as a drug target. J Med Virol 93:4258
8. Ogando NS, Zevenhoven-Dobbe JC, van der Meer Y, Bredenbeek PJ, Posthuma CC, Snijder EJ (2020) Coronavirus genomic nsp14-ExoN, structure, role, mechanism, and potential application as a drug target. J Virol 94:e01246
9. Hsu JC-C, Laurent-Rolle M, Pawlak JB, Wilen CB, Cresswell P (2021) Translational shutdown and evasion of the innate immune response by SARS-CoV-2 NSP14 protein. Proc Natl Acad Sci U S A 118:e2101161118
10. Li W, Su Y-Y, Zhi S-S, Huang J, Zhuang C-L, Bai W-Z, Wan Y, Meng X-R, Zhang L, Zhou Y-B, Luo Y-Y, Ge S-X, Chen Y-K, Ma Y (2020) Virus shedding dynamics in asymptomatic and mildly symptomatic patients infected with SARS-CoV-2. Clin Microbiol Infect 26:1556.e1
11. Wang R, Hozumi Y, Zheng Y-H, Yin C, Wei G-W (2020) Host immune response driving SARS-CoV-2 evolution. Viruses 12:E1095
12. Avanzato VA, Matson MJ, Seifert SN, Pryce R, Williamson BN, Anzick SL, Barbian K, Judson SD, Fischer ER, Martens C, Bowden TA, de Wit E, Riedo FX, Munster VJ (2020) Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. Cell 183:1901
13. Sonnleitner ST, Prelog M, Sonnleitner S, Hinterbichler E, Halbfurter H, Kopecky DBC, Almanzar G, Koblmüller S, Sturmbauer C, Feist L, Horres R, Posch W, Walder G (2022) Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. Nat Commun 13:2560
14. Leung WF, Chorlton S, Tyson J, Al-Rawahi GN, Jassem AN, Prystajecky N, Masud S, Deans GD, Chapman MG, Mirzanejad Y, Murray MCM, Wong PHP (2022) COVID-19 in an immunocompromised host: persistent shedding of viable SARS-CoV-2 and emergence of multiple mutations: a case report. Int J Infect Dis 114:178
15. Braun KM, Moreno GK, Wagner C, Accola MA, Rehrauer WM, Baker DA, Koelle K, O'Connor DH, Bedford T, Friedrich TC, Moncla LH (2021) Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. PLoS Pathog 17:e1009849
16. Pei L, Chen Y, Zheng X, Gong F, Liu W, Lin J, Zheng R, Yang Z, Bi Y, Chen E (2023) Comorbidities prolonged viral shedding of patients infected with SARS-CoV-2 omicron variant in Shanghai: a multi-center, retrospective, observational study. J Infect Public Health 16:182
17. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, Boerwinkle E, Fu Y-X (2004) Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol 4:21
18. Shrestha NK, Marco Canosa F, Nowacki AS, Procop GW, Vogel S, Fraser TG, Erzurum SC, Terpeluk P, Gordon SM (2020) Distribution of transmission potential during nonsevere COVID-19 illness. Clin Infect Dis 71(11):2927–2932
19. Jehi L, Ji X, Milinovich A, Erzurum S, Rubin BP, Gordon S, Young JB, Kattan MW (2020) Individualizing risk prediction for positive coronavirus disease 2019 testing. Chest 158:1364
20. Paragon Genomics. https://www.paragongenomics.com/product/cleanplex-sars-cov-2-flex-panel/. Accessed April 2023
21. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, Zhu A, Huang Y, Xiao F, Yao J, Gan M, Li F, Luo L, Huang X, Zhang Y, Wong S-S, Cheng X, Ji J, Ou Z, Xiao M, Li M, Li J, Ren P, Deng Z, Zhong H, Xu X, Song T, Mok CKP, Peiris M, Zhong N, Zhao J, Li Y, Li J, Zhao J (2021) Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. Genome Med 13:30
22. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. ArXiv preprint arXiv:12073907. https://doi.org/10.48550/arXiv.1207.3907
23. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6:80
24. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24
25. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Andersson M, Otecko N, Wise EL, Moore N, Lynch J, Kidd S, Cortes N, Mori M, Williams R, Vernet G, Justice A, Green A, Nicholls SM, Ansari MA, Abeler-Dörner L, Moore CE, Peto TEA, Eyre DW, Shaw R, Simmonds P, Buck D, Todd JA, Connor TR, Ashraf S, da Silva Filipe A, Shepherd J, Thomson EC, Bonsall D, Fraser C, Golubchik T, Oxford Virus Sequencing Analysis Group (OVSG) COVID-19 Genomics UK (COG-UK) Consortium (2021) SARS-CoV-2 within-host diversity and transmission. Science. https://doi.org/10.1126/science.abg0821
26. Koboldt DC (2020) Best practices for variant calling in clinical sequencing. Genome Med 12:91
27. World Health Organization (2021) Genomic Sequencing of SARS-CoV-2: A Guide to Implementation for Maximum Impact on Public Health, 8 January 2021. World Health Organization, Geneva
28. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368

29. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolism Acadamic press. Elsevier, Amsterdam
30. Nocedal J, Wright SJ (2006) Numerical Optimization, 2nd edn. Springer, New York, NY
31. Thakur S, Sasi S, Pillai SG, Nag A, Shukla D, Singhal R, Phalke S, Velu GSK (2022) SARS-CoV-2 mutations and their impact on diagnostics, therapeutics and vaccines. Front Med 9:815389
32. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G (2020) Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol. https://doi.org/10.1093/ve/veaa061
33. Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, Rambaut A, Suchard MA, Wertheim JO, Lemey P (2020) The emergence of SARS-CoV-2 in Europe and North America. Science 370:564
34. Kim S, Nguyen TT, Taitt AS, Jhun H, Park H-Y, Kim S-H, Kim Y-G, Song EY, Lee Y, Yum H, Shin K-C, Choi YK, Song C-S, Yeom SC, Kim B, Netea M, Kim S (2021) Omicron mutation is faster than the chase: multiple mutations on spike/ACE2 interaction residues. Immune Netw 21:e38
35. A. L. Valesano, K. E. Rumfelt, D. E. Dimcheff, C. N. Blair, W. J. Fitzsimmons, J. G. Petrie, E. T. Martin, and A. S. Lauring, BioRxiv 2021.01.19.427330 (2021).
36. Al Khatib HA, Benslimane FM, Elbashir IE, Coyle PV, Al Maslamani MA, Al-Khal A, Al Thani AA, Yassine HM (2020) Within-host diversity of SARS-CoV-2 in COVID-19 patients with variable disease severities. Front Cell Infect Microbiol 10:575613
37. Li J, Du P, Yang L, Zhang J, Song C, Chen D, Song Y, Ding N, Hua M, Han K, Song R, Xie W, Chen Z, Wang X, Liu J, Xu Y, Gao G, Wang Q, Pu L, Di L, Li J, Yue J, Han J, Zhao X, Yan Y, Yu F, Wu AR, Zhang F, Gao YQ, Huang Y, Wang J, Zeng H, Chen C (2022) Two-step fitness selection for intra-host variations in SARS-CoV-2. Cell Rep 38:110205
38. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, Zhou Z, Yang J, Zhong J, Yang D, Guo L, Zhang G, Li H, Xu Y, Chen M, Gao Z, Wang J, Ren L, Li M (2021) Corrigendum to: genomic diversity of severe acute respiratory syndrome–coronavirus 2 in patients with coronavirus disease 2019. Clin Infect Dis 73:2374
39. Tay JH, Porter AF, Wirth W, Duchene S (2022) The emergence of SARS-CoV-2 variants of concern is driven by acceleration of the substitution rate. Mol Biol Evol. https://doi.org/10.1093/molbev/msac013
40. Eskier D, Suner A, Oktay Y, Karakülah G (2020) Mutations of SARS-CoV-2 nsp14 exhibit strong association with increased genome-wide mutation load. PeerJ 8:e10181
41. Takada K, Ueda MT, Watanabe T, Nakagawa S (2020) Genomic diversity of SARS-CoV-2 can be accelerated by a mutation in the Nsp14 gene. Iscience 2(1):265
42. Niu X, Kong F, Hou YJ, Wang Q (2021) Crucial mutation in the exoribonuclease domain of nsp14 of PEDV leads to high genetic instability during viral replication. Cell Biosci 11:106
43. V. Nussenblatt, A. E. Roder, S. Das, E. de Wit, J.-H. Youn, S. Banakis, A. Mushegian, C. Mederos, W. Wang, M. Chung, L. Pérez-Pérez, T. Palmore, J. N. Brudno, J. N. Kochenderfer, and E. Ghedin, MedRxiv 2021.10.02.21264267 (2021).
44. Chaguza C, Hahn AM, Petrone ME, Zhou S, Ferguson D, Breban MI, Pham K, Peña-Hernández MA, Castaldi C, Hill V, Schulz W, Swanstrom RI, Roberts SC, Grubaugh ND, Yale SARS-CoV-2 Genomic Surveillance Initiative (2022) Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. Cell Reports Medicine 4(2):32
45. Choudhary MC, Crain CR, Qiu X, Hanage W, Li JZ (2022) severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequence characteristics of coronavirus disease 2019 (COVID-19) persistence and reinfection. Clin Infect Dis 74:237
46. Becares M, Pascual-Iglesias A, Nogales A, Sola I, Enjuanes L, Zuñiga S (2016) Mutagenesis of coronavirus nsp14 reveals Its potential role in modulation of the innate immune response. J Virol 90:5399