Virus Evolution, 2021, 7(1): veab013

doi: 10.1093/ve/veab013
Research Article

SARS-CoV-2 genomic analyses in cancer patients reveal elevated intrahost genetic diversity

Juliana D. Siqueira,[1,*,†,‡] Livia R. Goes,[1,2,†] Brunna M. Alves,[1,†]
Pedro S. de Carvalho,[1] Claudia Cicala,[2] James Arthos,[2] João P.B. Viola,[3]
Andréia C. de Melo,[4] and Marcelo A. Soares[1,§];
on behalf of the INCA COVID-19 Task Force[¶]

[1]Programa de Oncovirologia, Instituto Nacional de Câncer, Rio de Janeiro, RJ 20.231-050, Brazil, [2]Laboratory of Immunoregulation, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20892, USA, [3]Programa de Imunologia e Biologia de Tumores, Instituto Nacional de Câncer, Rio de Janeiro, RJ 20.231-050, Brazil and [4]Divisão de Pesquisa Clínica e Desenvolvimento Tecnológico, Instituto Nacional de Câncer, Rio de Janeiro, RJ 20.231-050, Brazil

*Corresponding author: E-mail: sidoju@hotmail.com

†Authors contributed equally to this work.

‡https://orcid.org/0000-0002-4266-9795

§https://orcid.org/0000-0002-9013-2570

¶Participants of the INCA COVID-19 Task Force are listed at the end of the article.

Abstract

Numerous factors have been identified to influence susceptibility to Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) infection and disease severity. Cancer patients are more prone to clinically evolve to more severe COVID-19 conditions, but the determinants of such a more severe outcome remain largely unknown. We have determined the full-length SARS-CoV-2 genomic sequences of cancer patients and healthcare workers (non-cancer controls) by deep sequencing and investigated the within-host viral population of each infection, quantifying intrahost genetic diversity. Naso- and oro-pharyngeal SARS-CoV-2$^+$ swabs from 57 cancer patients and 14 healthcare workers from the Brazilian National Cancer Institute were collected in April to May 2020. Complete genome amplification using ARTIC network V3 multiplex primers was performed followed by next-generation sequencing. Assemblies were conducted in Geneious R11, where consensus sequences were extracted and intrahost single nucleotide variants were identified. Maximum likelihood phylogenetic analysis was performed using PhyMLv.3.0 and lineages were classified using Pangolin and CoV-GLUE. Phylogenetic analysis showed that all but one strain belonged to clade B1.1. Four genetically linked mutations known as the globally dominant SARS-CoV-2 haplotype (C241T, C3037T, C14408T and A23403G) were found in the majority of consensus sequences. SNV signatures of previously characterized Brazilian genomes were also observed in most samples. Another 85 SNVs were found at a lower frequency (1.4%–19.7%) among the consensus sequences. Cancer patients displayed a significantly higher intrahost viral genetic diversity compared to healthcare workers. This difference was independent of SARS-CoV-2 Ct values obtained at the diagnostic tests, which did not differ between the two groups. The most common nucleotide changes of intrahost SNVs in both groups were consistent with APOBEC and ADAR activities. Intrahost genetic diversity in cancer patients was not associated with disease severity, use of corticosteroids, or use of antivirals, characteristics that could influence viral diversity. Moreover, the presence of metastasis, either in general or specifically in the lung, was not associated with intrahost diversity among cancer patients. Cancer patients carried significantly higher numbers of minor variants compared to non-cancer counterparts. Further studies on SARS-CoV-2 diversity in especially vulnerable patients will shed light onto the understanding of the basis of COVID-19 different outcomes in humans.

Key words: SARS-CoV-2; COVID-19; cancer; single nucleotide variant; full-length genome.

1. Introduction

In December 2019, a new form of pneumonia was described in patients with severe acute respiratory syndrome in the city of Wuhan, province of Hubei, China (Li et al. 2020). Soon after, a new beta-coronavirus was identified as the causative agent of that disease (Wu et al. 2020). The new virus was named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), and the disease was called Coronavirus Disease 2019 (COVID-19) (Wu et al. 2020). Since its initial discovery, COVID-19 has become a pandemic of catastrophic proportions, with almost 100 million confirmed cases of viral infection and two million deaths worldwide (https://www.worldometers.info/coronavirus/, last accessed on 20 January 2021).

Numerous demographic, clinical, genetic, and behavioral factors have been identified to influence susceptibility to SARS-CoV-2 infection and, among those infected, the severity of the disease, including the risk of death. Those factors include age, sex (Asselta et al. 2020), genetic loci of certain cytokines/chemokines and the ABO blood system group (Ellinghaus et al. 2020; Kirtipal and Bharadwaj 2020), smoking history (Gallus et al. 2020), obesity and underlying comorbidities such as diabetes, hypertension, lung diseases (Singh et al. 2020; Tahvildari et al. 2020), and cancer (Lee et al. 2020; Yang et al. 2020; Zhang et al. 2020). Among cancer patients, those with malignancies of hematological origin have been reported as particularly vulnerable to COVID-19 (Willan et al. 2020).

SARS-CoV-2 is a single-stranded RNA virus that replicates using an RNA-dependent RNA polymerase. As such, the virus is subjected to nucleotide sequence changes and has evolved through molecular evolution and founder effects during its explosive spread throughout the globe. Virus replication rates directly impact the accumulation of mutations in the virus genome, enabling the existence of a viral quasispecies (a swarm of different, yet highly related, viral entities) within an infected host. Although within-host variations of SARS-CoV-2 have been documented (Jary et al. 2020; Shen et al. 2020), the impact of underlying comorbidities that promote persistent viral RNA detection and shedding on virus evolution remains to be elucidated. Moreover, viral genetic variation, as a source of novel mutations, may hinder future therapeutic antiviral and vaccine strategies targeting COVID-19, by the selection of drug-resistant and vaccine escape mutants (Fung and Liu 2019).

In the present work, we have determined the full-length SARS-CoV-2 genomic sequences of 57 cancer patients and 14 healthcare workers (HCW) (non-cancer controls) employing next-generation sequencing (NGS) and analyzed their epidemiological relatedness and lineage classification. This approach also allowed us to study the within-host viral population of each infection, quantify intrahost viral genetic diversity and characterize specific genetic changes with potential to impact SARS-CoV-2 biology. Finally, we have also assessed associations between viral diversity and patients' clinical and laboratory characteristics, thereby identifying determinant factors of viral evolution in this particular group of patients.

2. Materials and methods

2.1 Study population

Fifty-seven cancer patients followed at the Brazilian National Cancer Institute (INCA), Rio de Janeiro, Brazil, and 14 HCW diagnosed with COVID-19 between 7 April and 5 May 2020, early in the COVID-19 pandemic in Rio de Janeiro, were included in this study. SARS-CoV-2 infection was diagnosed through naso- and oropharyngeal swab specimens using real-time reverse-transcription polymerase chain reaction (RT-qPCR) following the U.S. Centers for Disease Control and Prevention protocol (Centers for Disease Control and Prevention 2020). All participants agreed to be enrolled in the study and signed an informed consent. Participants' data were treated anonymously. This study was approved by the Brazilian National Commission for Ethics in Research (approval number: CAAE 30608220.8.0000.5274).

2.2 SARS-CoV-2 nucleic acid isolation, amplification and sequencing

Naso- and oropharyngeal swabs were collected and placed into a conical tube containing 2 ml of viral transport medium (VTM, Thermo Fisher Scientific, Waltham, MA). Viral DNA and RNA were extracted with the QIAamp MiniElute Virus Spin Kit (QIAGEN, Chatsworth, CA) according to manufacturer's instructions. All cDNAs were synthesized in duplicate using the SuperScript^{TM} III First-Strand Synthesis System (Thermo Fisher Scientific). The SARS-CoV-2 complete genome amplification was based on an openly available protocol developed by the ARTIC network (https://artic.network/ncov-2019, accessed March 26, 2020) using the V.3 multiplex primers scheme and Platinum Taq DNA Polymerase High Fidelity (Thermo Fisher Scientific). Positive PCR products were purified with the ReliaPrep^{TM} DNA Clean-Up and Concentration System (Promega, Madison, WI). Genomic libraries were constructed with the Nextera XT DNA Sample Preparation kit (Illumina Inc., San Diego, CA) according to the manufacturer's protocol, pooled with 1 per cent denatured PhiX DNA (sequencing control) and sequenced in a MiSeq platform (2 × 251 cycles paired-end run; Illumina). New PCR reactions using combinations of the primers described above were carried out to cover regions with low coverage for each sample. Positive products were purified and sequenced by Sanger using the *BigDye Terminator kit* (Thermo Fisher Scientific) in an automated 3130XL Genetic Analyzer (Thermo Fisher Scientific). Sequences were edited and assembled with SeqMan v.7.0.0 (DNAStar Inc., Madison, WI).

2.3 SARS-CoV-2 near-full-length consensus sequence and nucleotide variations

All analyses were conducted using Geneious R11 software (Biomatters, Auckland, New Zealand), where the reads were trimmed to achieve an error rate below 0.1 per cent and assembled to the Wuhan-Hu-1 reference sequence genome (GenBank number MN908947). A minimum mapping quality of 30 was required, providing a 99.9 per cent confidence level that the mapping is correct. Additionally, all assemblies were visually inspected to evaluate the mapped reads and consequently to ensure the quality of the consensus generated and single nucleotide variation (SNVs) analysis. Consensus sequences representing SARS-CoV-2 near-full-length genomes were extracted for each sample and aligned to the Wuhan-Hu-1 reference sequence genome. Nucleotide variations in relation to the reference sequence were identified and classified as SNVs. Intrahost SNV (iSNV) was defined as a variation with a frequency greater than 2 per cent and depth coverage by at least 500 reads. iSNVs were manually verified, and the intrahost viral genetic diversity rate was calculated as the number of nucleotide substitutions with a frequency greater than 2 per cent for the given sample divided by the number of positions with depth coverage greater than 500 times multiplied by 10^{-4} (substitutions/site x 10^{-4}). The 2 per cent threshold applied in this study was chosen based on previous studies on HIV, in which this threshold was able to distinguish between variants consistently detected in different sequencing replicates from spurious variants (Dudley et al. 2014; Alves et al. 2017).

2.4 SARS-CoV-2 classification and phylogenetic analysis

For SARS-CoV-2 lineage classification, consensus genomes were submitted to *Pangolin* software (https://github.com/cov-lineages/pangolin, downloaded on 10 June 2020) and to *CoV-GLUE* lineage system (http://cov-glue.cvr.gla.ac.uk/#/home, accessed on June 10th, 2020) (Singer et al. 2020), both based on the nomenclature proposed by Rambaut et al. (2020). An alignment including the consensus sequences generated and genomes from Brazilian sequences available on the GISAID Database classified as B1, B1.1 and the Brazilian clusters B1.1-BR/B1.1-EU/BR (Supplementary Table S1) were submitted to a maximum likelihood phylogenic reconstruction using PhyML v.3.0 and the best model of nucleotide substitution was defined with Model Generator (GTR) to investigate the sublineage classification of the study sequences (Keane et al. 2006; Guindon et al. 2010; Resende et al. 2020). Furthermore, a phylogenetic analysis that included the generated consensus sequences along with all SARS-CoV-2 sequences from Rio de Janeiro state (Brazil) presently available at GISAID (https://www.epicov.org/epi3/frontend, accessed on 27 July 2020, Supplementary Table S1) was performed in order to investigate epidemiological relatedness of sequences.

2.5 Statistical analyses

Clinical categorical variables were compared between cancer patients and HCWs using chi-square and Fisher's exact tests. Mann–Whitney two-tailed test was used to compare intrahost diversity (substitutions/site × 10^{-4}) between cancer patients and HCWs and between cancer patients' clinical categorical variable groups. The Benjamini–Hochberg method was applied for multiple comparison correction. False discovery rate (FDR) values were calculated with the Stats package v.4.1.0 for R v.4.0.3. Spearman's rank was employed to evaluate the correlation between intrahost diversity and continuous variables (such as age and SARS-CoV-2 RT-qPCR Ct values). All graphical representations and statistical analyses were performed using Geneious R11 (Biomatters) and GraphPad Prism v.8.0.1 (GraphPad Software Inc., San Diego, CA).

3. Results

3.1 Clinical characteristics of the studied population

Summarized demographic and clinical characteristics of the patients and HCW from whom SARS-CoV-2 sequences were studied is seen in Table 1. Among patients, the median age was 61 years and most of them (72%) had solid malignancies, 16 per cent of patients used corticosteroids and 14 per cent used oseltamivir previously or during COVID-19 diagnosis specimen collection. Among HCW, the median age was 40 years and most (86%) were female. The most prevalent COVID-19 symptoms among patients were cough, fever and dyspnea. Death from COVID-19 occurred in 33.3 per cent of the cases. For HCW, cough and coryza were the mainly reported COVID-19 symptoms (85.7% each), and all subjects recovered from the disease, with no deaths reported. No difference was found in sex distribution between the two groups ($P = 0.118$), but HCW had a lower median age when compared to cancer patients ($P < 0.001$). Death occurred in 38.6 per cent of the cancer patients, but no deaths occurred among HCWs ($P = 0.0029$). Some COVID-19-related symptoms at diagnosis also differed between the two groups (Table 1).

3.2 Sequence coverage, quality, and metrics

A total of 27,433,528 reads were obtained from sequencing, with an average of 382,118 reads per sample, ranging from 217,922 to 631,796 reads. Reads of each sample were assembled with Wuhan-Hu-1 reference genome with a minimum mapping quality of 30 Phred and the average depth coverage obtained was 1,468 (465–2,530). The coverage was heterogeneous across the genome but was similar among the samples (Supplementary Fig. S1). Consensus sequences containing more than 97.9 per cent of the SARS-CoV-2 complete genome were generated from all 57 cancer patients and 14 HCW samples.

3.3 Phylogenetic and epidemiological profile of SARS-CoV-2 sequences

SARS-CoV-2 genome sequence submission to the *Pangolin* and *CoV-GLUE* algorithms resulted in the same lineage classification in all cases, defining all but one virus belonging to clade B1.1, while the remaining sequence was classified as B.1. A phylogenetic analysis of the viruses together with sequences previously defined as Brazilian circulating strains B1.1-BR and B1.1-EU/BR showed that most B1.1 genomes generated in this study clustered with B1.1-BR sequences (Fig. 1A) (Resende et al. 2020). A phylogenetic tree including all local SARS-CoV-2 sequences isolated from patients residing in the state of Rio de Janeiro available at the GISAID database (accessed on 27 July 2020, Supplementary Table S1) was performed to investigate potential epidemiological linkage between samples (Fig. 1B). We noted that some of the viruses sequenced at INCA clustered in clades containing identical sequences, suggesting a transmission link between the study subjects. In some instances, both cancer patients and HCW were involved in those epidemiological clusters. Although in some cases sequences from outside the hospital were also identical to viruses from our series, therefore not excluding the possibility of community transmission, the most likely scenario for those cases is a nosocomial transmission between patients and/or HCW.

Table 1. Demographic and clinic characteristics of the cancer patients and healthcare workers studied.

Characteristic	Patients (%) n = 57	Healthcare workers (%) n = 14	P value
Age, years (median, range)	61 (9–79)	40.5 (33–57)	<0.001
Age			
<25 years	3.5	0	
25–64 years	57.9	100	
≥65 years	38.6	0	
Gender			0.118
Female	61.4	85.7	
Male	38.6	14.3	
Time in days between symptom onset and sample collection (median, range)[a]	3 (0–34)	NC	NT
Symptoms at COVID-19 diagnosis			
Cough	59.6	85.7	0.1956
Fever	57.9	57.1	0.7645
Dyspnea	56.1	7.1	<0.001
Fatigue	24.6	21.4	1
Diarrhea	14.0	7.1	0.6719
Nausea/vomiting	12.3	0	0.3301
Anorexia	7.0	0	0.5721
Sore throat	5.3	42.8	0.0018
Myalgia	3.5	0	1
Headache	3.5	42.8	<0.001
Anosmia	3.5	42.8	<0.001
Ageusia	3.5	0	1
Coryza	3.5	85.7	<0.001
None	0	0	
Missing	7.0	0	
Death			0.0029
Yes, from COVID-19	33.3	0	
Yes, other cause	5.3	0	
No	54.4	100	
Missing	7.0	0	
Smoking			
Past/current	21.0	NC	NT
Never	24.6	NC	NT
Missing	54.4	NC	NT
Primary cancer site			
Solid tumors	71.9	NA	NT
Hematological malignancies	28.1	NA	NT
Metastatic disease			
Yes, to the lung	14.0	NA	NT
Yes, to other organs	24.6	NA	NT
No	47.4	NA	NT
Missing	14.0	NA	NT
Use of corticosteroid			
Yes	19.3	NA	NT
No	87.2	NA	NT
Missing	3.5	NA	NT
Use of oseltamivir			
Yes	14.0	NA	NT
No	82.4	NA	NT
Missing	3.5	NA	NT

NC, not collected; NA, not applicable; NT, not tested.
[a] Data available for 45 patients.

3.4 SNVs across the SARS-CoV-2 genomes

Overall, 95 SNVs and three deletions were found across the SARS-CoV-2 consensuses analyzed (Supplementary Fig. S2). Four genetically linked mutations previously described as the globally dominant haplotype in April 2020 were found in the majority of our consensus sequences: C241T (100%; 5'UTR region), C3037T (98.6%; silent mutation), C14408T (100%; resulting in P4715L/P323L amino acid change in ORF1ab) and A23403G (100%; resulting in D614G amino acid change in S) (Korber et al. 2020). Additionally, SNV signatures of previously characterized Brazilian genomes were found in most samples, such as G28881A and G28882A (98.6%; resulting in R203K change in N),

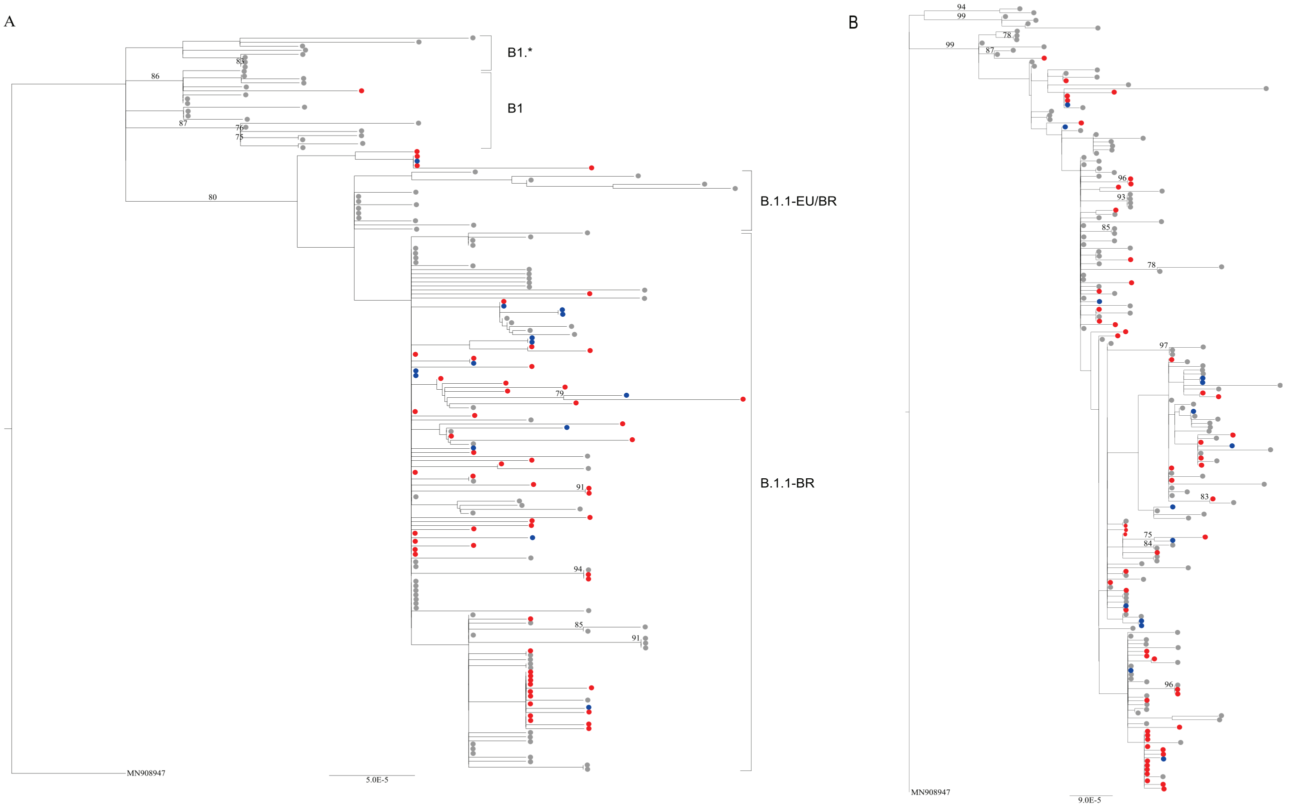


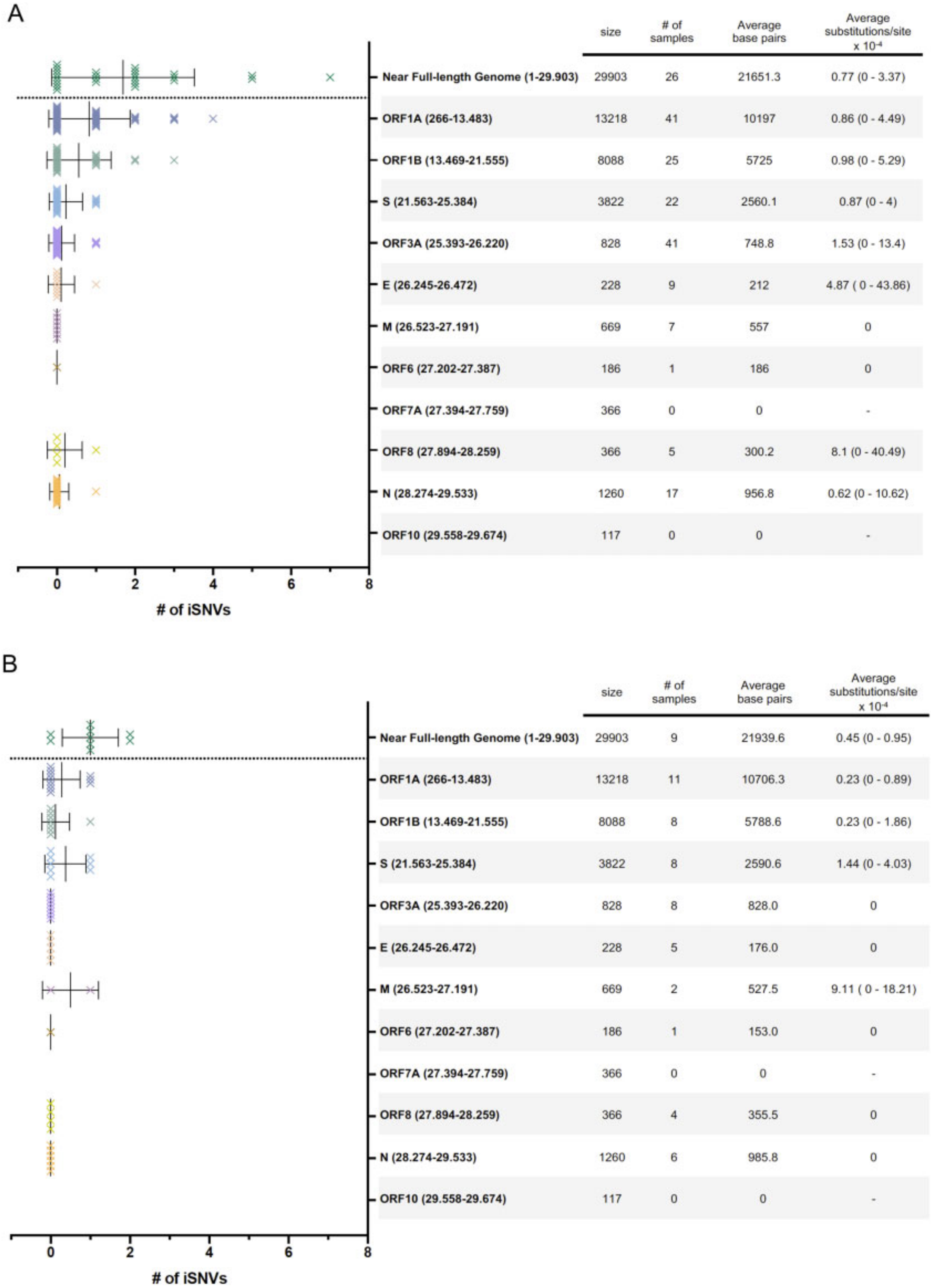
Figure 1. Maximum likelihood phylogenetic trees of near full-length SARS-CoV-2 genomes characterized. Tree including cancer patients (depicted in red circles), healthcare workers (in blue), Brazilian sequences classified as B1, B1.* and the Brazilian circulating strains B1.1-BR/B.1.1-EU/BR available on GISAID (in gray). (B) Tree showing epidemiological linkage of cancer patients (shown in red), healthcare workers (in blue) and all SARS-Cov-2 sequences from Rio de Janeiro state (in gray) available on GISAID Database. In both cases, GISAID was accessed on July 27th, 2020. Bootstrap values greater than 70 are shown in both trees.

G28883C (98.6%; resulting in G204R change in N), T27299C (91.6%; resulting in I33T change in ORF6), and T29148C (90.1%; resulting in I292T change in N) (Candido et al. 2020; Resende et al. 2020). The two latter SNVs are synapomorphic traits of the B1.1-EU/BR and B1.1-BR Brazilian circulating strains (Resende et al. 2020). Another 85 SNVs were observed in our consensus sequences at lower frequency (1.4%–19.7%; Supplementary Table S2), including nine non-synonymous mutations in S protein (V16F, V367L, K558N, Q675H, A879V, S939F, V1176F, K1191N, and G1219V). Deletions were found in three genomes: a 12-bp in-frame deletion in S (comprising positions 21,603–21,614), a 6-pb in-frame deletion in ORF3a (25,710–25,715) and a 244-pb frameshift deletion in ORF7 (27,508–27,751), resulting in a truncated protein. All deletions were confirmed by Sanger sequencing (data not shown).

3.5 SARS-CoV-2 intrahost genetic diversity

The NGS method used for studying viruses allowed us to assess the iSNVs that compose each subject's viral within-host population. iSNVs were distributed at 160 genome positions and all iSNVs present in overlapping regions of PCR fragments were concordant in both fragments. Five of them were observed in more than one sample, of which only one was found in epidemiologically linked samples. Of the 160 iSNVs, 140 were already observed in unrelated strains isolated from different countries/regions of the globe according to the Nexstrain, GESS and CovGLUE databases (Hadfield et al. 2018; Fang et al. 2020; Singer et al. 2020). The most frequent variations were missense (96 positions), silent variations were observed at 63 genome positions and one iSNV position was in a non-coding region (Supplementary Fig. S3). Nonsense changes were not observed. The absence of nonsense changes, coupled to the observations of missense mutations which appear in unrelated viruses across the globe suggests biological significance to these changes (e.g. immune escape or increase in fitness). The missense mutations also appear to have risen independently in different patients of the study and in those from abroad, as these viruses are not related by recent common ancestry. The ratio of nonsynonymous to synonymous intrahost variations was 1.49 and for most ORFs (ORF1a, ORF1b, S, ORF3a, M and N) this ratio was greater than 1.25. All but one iSNV with intrahost frequency greater than 20 per cent were found exclusively in cancer patients' samples (Supplementary Fig. S4 and Supplementary Table S3). The number of iSNVs across the viral genome can be visualized in Fig. 2 and their coverage is shown in Supplementary Fig. S5. Interestingly, patients displayed a significantly higher intrahost viral genetic diversity when compared to HCW ($P = 0.009$; Fig. 3A) and remained significant even after outlier subjects with higher virus diversity were excluded from the analysis ($P = 0.029$; Fig. 3B). Viral genetic diversity within each individual ORF was compared between the two groups, but no differences were found after correction for multiple comparisons.

As the within-host genetic diversity of viruses is commonly associated with viral replication, we have evaluated the correlation of the within-host population diversity in our subjects with the Ct values obtained in the RT-PCR swab tests of the same samples. Ct values work as a proxy for SARS-CoV-2 viral load in samples and are expected to be inversely correlated with viral



A

	size	# of samples	Average base pairs	Average substitutions/site x 10-4
Near Full-length Genome (1-29.903)	29903	26	21651.3	0.77 (0 - 3.37)
ORF1A (266-13.483)	13218	41	10197	0.86 (0 - 4.49)
ORF1B (13.469-21.555)	8088	25	5725	0.98 (0 - 5.29)
S (21.563-25.384)	3822	22	2560.1	0.87 (0 - 4)
ORF3A (25.393-26.220)	828	41	748.8	1.53 (0 - 13.4)
E (26.245-26.472)	228	9	212	4.87 (0 - 43.86)
M (26.523-27.191)	669	7	557	0
ORF6 (27.202-27.387)	186	1	186	0
ORF7A (27.394-27.759)	366	0	0	-
ORF8 (27.894-28.259)	366	5	300.2	8.1 (0 - 40.49)
N (28.274-29.533)	1260	17	956.8	0.62 (0 - 10.62)
ORF10 (29.558-29.674)	117	0	0	-

B

	size	# of samples	Average base pairs	Average substitutions/site x 10-4
Near Full-length Genome (1-29.903)	29903	9	21939.6	0.45 (0 - 0.95)
ORF1A (266-13.483)	13218	11	10706.3	0.23 (0 - 0.89)
ORF1B (13.469-21.555)	8088	8	5788.6	0.23 (0 - 1.86)
S (21.563-25.384)	3822	8	2590.6	1.44 (0 - 4.03)
ORF3A (25.393-26.220)	828	8	828.0	0
E (26.245-26.472)	228	5	176.0	0
M (26.523-27.191)	669	2	527.5	9.11 (0 - 18.21)
ORF6 (27.202-27.387)	186	1	153.0	0
ORF7A (27.394-27.759)	366	0	0	-
ORF8 (27.894-28.259)	366	4	355.5	0
N (28.274-29.533)	1260	6	985.8	0
ORF10 (29.558-29.674)	117	0	0	-

Figure 2. Number of iSNVs per ORF analyzed. Data for cancer patients (A) and healthcare workers (B) are shown. Each iSNV showed an intrahost frequency greater than 2% with a minimum depth coverage of 500x. The table on the right shows ORF name and genome coordinate based on SARS-CoV-2 Wuhan-Hu-1 reference sequence genome (GenBank number MN908947), ORF size in bp, number of samples, average base pairs analyzed and average (min–max) substitutions per site × 10^{-4}. The last three columns refer only to samples with minimum depth coverage of 500× for at least 60% of the given ORF region extension.

diversity and replication. Surprisingly, however, Ct of the samples did not inversely correlate with viral diversity, but rather showed a positive correlation, despite having a low r_s value, below 0.5 (Fig. 3C). This was also true when patients' samples were analysed separately ($r_s = 0.490$; $P = 0.001$; data not shown). Of note, no significant differences were observed when Ct values were compared between the two groups ($P = 0.175$). No correlation was observed when comparing viral genetic diversity

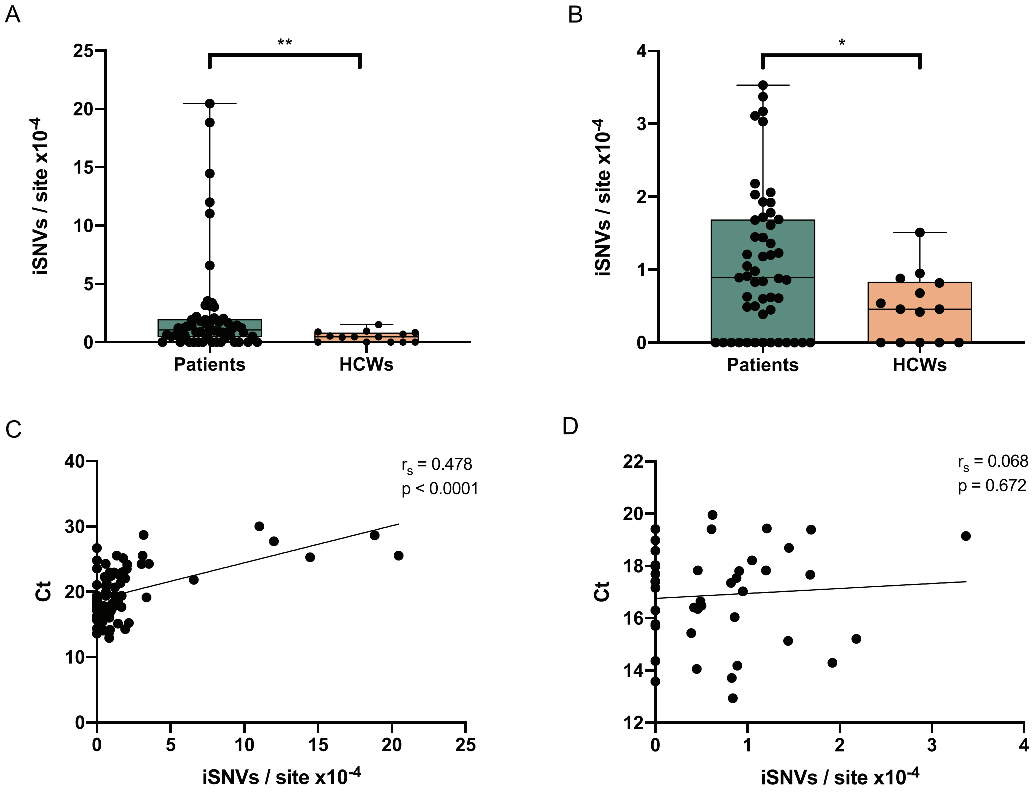


Figure 3. Viral genetic diversity in cancer patients and HCWs. Diversity was calculated by number of iSNVs per site × 10^{-4}. Tukey boxplots show the viral genetic diversity in cancer patients compared to HCW (Mann–Whitney test (two-tailed), **P = 0.0093) (A). The difference is still significant when outlier patients are removed from the analysis (Mann–Whitney test (two-tailed), *P = 0.0299) (B). Viral genetic diversity in cancer patients and healthcare workers samples (n = 71) showed a positive correlation with Ct values (C). No correlation between viral genetic diversity and Ct values is observed when analyzing samples with Ct < 20 (D). Spearman correlation analysis r_s and P-values are indicated.

and Ct when only Ct values lower than 20 were considered (Fig. 3D). Despite the above-mentioned age difference observed between HCW and cancer patients, age did not correlate with viral genetic diversity (Supplementary Fig. S6A, P = 0.844). The time between symptom onset and sample collection was neither associated with intrahost genetic diversity (*P* = 0.628) nor with the Ct values (*P* = 0.304) (Supplementary Figs. S6B and S6C, respectively).

Regarding patients' characteristics, intrahost virus diversity was not associated with disease severity (overall death [*P* = 0.722] or death due to COVID-19 [*P* = 0.934], ICU requirement [*P* = 0.722]), use of corticosteroids chronically or during COVID-19 course (*P* = 0.660), or use of oseltamivir prior to COVID-19 diagnosis (*P* = 0.384; Table 2). We also assessed the potential association of cancer patients with hematological malignancies compared to those with solid cancers, but no association was found (*P* = 0.722; Table 2). No intrahost diversity difference was found when comparing patients with metastatic cancer with cancer patients without metastasis (*P* = 0.279; Table 2). The same was observed when comparing patients with metastatic disease to the lung with all other cancer patients (*P* = 0.279; Table 2). No difference was found when comparing patients with pulmonary metastasis with patients with metastatic disease to other organs (*P* = 0.3562).

Nucleotide changes across the genome at intrahost level can be visualized in Fig. 4A. Cytidine-to-thymidine/uridine (C-to-T(U)) and adenosine-to-inosine/guanosine (A-to-I(G)) transitions that are characteristic of APOBEC and ADAR activities (Smith and Sowden 1996; Vieira and Soares 2013) are

Table 2. Association between SARS-CoV-2 genetic diversity and clinical outcomes.

Characteristic	Median iSNVs/site × 10^{-4}	*P*-value[a]
Death		
Yes	0.92	0.722
No	1.21	
Death from COVID-19		
Yes	1.05	0.934
No	1.04	
Intensive care unit		
Yes	0.93	0.722
No	1.05	
Use of corticosteroid		
Yes	0.91	0.660
No	1.11	
Use of oseltamivir		
Yes	1.92	0.384
No	0.89	
Haematological malignancies		
Yes	1.37	0.722
No	0.98	
Metastatic disease		
Yes	1.64	0.279
No	0.98	
Metastatic disease to the lung		
Yes	1.70	0.279
No	0.98	

[a] After Benjamini–Hochberg correction for multiple comparisons.

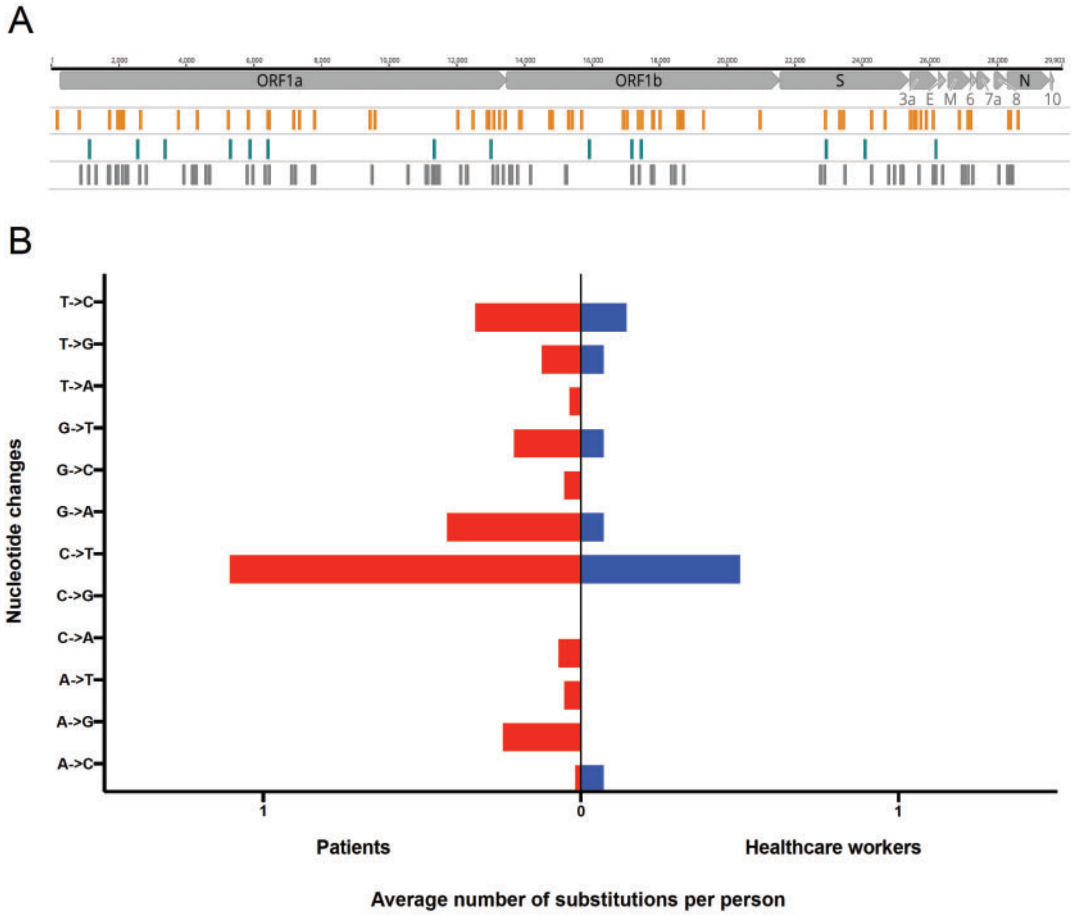


Figure 4. iSNVs according to the nucleotide change. Distribution of nucleotide changes across the genome for all samples (n = 71) is shown (A). C-to-T mutations are shown in orange, A-to-G are in green and the remaining changes are in grey. Genome coordinates are relative to the SARS-CoV-2 Wuhan-Hu-1 reference sequence (GenBank acc.# MN908947). Average frequency of each type of nucleotide change found for cancer patients (red) and healthcare workers (blue) (B).

highlighted. C-to-T was the most frequent iSNV observed in both cancer patients and HCWs (Fig. 4B). The distribution of these changes did not differ between the two groups studied (data not shown).

4. Discussion

The biology of SARS-CoV-2 infection in humans is striking to infectious disease clinicians worldwide, because no viral infection has been previously seen with such an enormous range of phenotypic outcomes, from no symptoms to severe respiratory distress and death. Most of this physiological variance, however, has been attributed to host genetic and behavioral factors. Numerous characteristics have been associated with susceptibility to SARS-CoV-2 infection and disease severity among infected subjects, and underlying comorbidities seem to play a major role in unfavorable disease outcomes. Chronic noncommunicable diseases such as cancer are among those conditions. Cancer patients have been reported to be more prone to SARS-CoV-2 infection and to clinically evolve to more severe conditions upon infection (Lee et al. 2020; Yang et al. 2020; Zhang et al. 2020), but the determinants of these severe outcomes remain largely unknown.

In this study we have evaluated the near full-length sequences of SARS-CoV-2 infecting cancer inpatients in one of the largest public cancer hospitals in South America, the Brazilian National Cancer Institute, and compared these sequences with those generated from healthcare professionals from the same institution. These complete SARS-CoV-2 genomes showed signatures characteristic of the virus that spread globally and is currently the predominant strain (Korber et al. 2020). All but one virus also belonged to clade B1.1, which is the clade primarily circulating in the Americas. The viral genomes also displayed sequence features of other already characterized Brazilian viruses, consistent with the hypothesis of local, community transmission rather than virus importation from abroad. In fact, the timeframe of the analyzed infections (from 7 April to 5 May 2020) is consistent with a period in Brazil where community virus transmission was already established and ongoing (Candido et al. 2020). Moreover, as a public and free hospital in the Brazilian Public Health System, INCA is also likely to admit patients with low socioeconomic resources who are mostly unable to travel abroad and most likely acquired viral infections from local sources.

We explored the evolutionary and phylogenetic relationships between the SARS-CoV-2 sequences of the studied samples. Upon a phylogenetic inference with viral sequences isolated from other infected subjects residing in the state of Rio de Janeiro (the same geographic location of the study site), we found that almost half of the sequences from our subjects lie in clusters with sequences from other patients and/or from HCW. Some of the consensus sequences within each cluster were identical, suggesting a direct epidemiological link between those groups of patients/HCW. Some sequences retrieved from the database representing subjects from the community outside the hospital were also identical to some hospital-based sequences, ruling out the possibility of completely excluding transmission from outside the hospital. However, the most parsimonious explanation is nosocomial transmission in those cases. Indeed, the subjects' samples were collected at a time in Brazil when tests for SARS-CoV-2 infection were not easily accessible, and inpatients and HCW had to wait several days for a test result, thus presenting a risk for further transmission.

SNVs were found across the entire SARS-CoV-2 genome. The spike (S) D614G mutation, found in all samples analyzed, has been associated with higher viral titers, suggesting increased viral infectivity (Korber et al. 2020). Other variations were also found in different regions of the spike protein, including a 12-bp in-frame deletion that harbors part of the signal peptide and the predicted cleavage site in the beginning of S. As expected, the P323L change in the RNA-dependent RNA polymerase (RdRp), genetically linked to D614G, was also found in all our samples. *In silico* analysis showed that P323L may impact the protein secondary structure, leading to a reduction in its molecular flexibility (Begum et al. 2020). However, the phenotypic impact of these mutations is still poorly understood. Numerous other missense mutations were found that warrant further investigation concerning their phenotypes.

The most striking observation of our intrahost population variation analysis was that cancer patients carried significantly higher numbers of minor variants when compared to non-cancer counterparts. This difference was independent of, and unrelated to the Ct values obtained at the diagnostic tests, which did not differ between the two groups. Despite cancer cases with metastatic sites have been associated with COVID-19-related death (de Melo et al. 2020), we did not find any association between intrahost diversity and metastasis or disease severity (requirement for ICU, death by any cause or COVID-19 related). Unexpectedly, this difference was also not related to the use of corticosteroids (which could lower their immunity status), use of oseltamivir (which was used by some patients to overcome a potential H1N1 infection before the COVID-19 diagnosis), neither associated with the type of primary malignancy (solid tumor vs. hematologic tumors). Despite conflicting data existing in the literature, the hematologic cancer patients infected with SARS-CoV-2 herein analyzed did not show an increased chance of COVID-19 severe outcomes when compared to those with solid tumors (de Melo et al. 2020).

Surprisingly, Ct values (as a proxy to viral load) not only did not inversely correlate with virus diversity, but showed a weak positive correlation, albeit with a low r_s coefficient and did not remain significant when comparing only Ct values below 20. It is well established that naso- and oropharyngeal swabs are not the best types of sample for detecting SARS-CoV-2, compared to sputum for example, which contain a larger amount of viral genetic material (Mohammadi et al. 2020). Our data underscore the possibility that the variation in the viral population that we see is not generated in the naso- or oropharynx, but rather more distally in the respiratory tract (lungs) or even in other tissues such as the gut. Reports on the comparative expression of the virus' cellular receptor ACE2 support the idea that those other tissues might be relevant sources of viral replication and, consequently, sites where diversity emerges (Hikmet et al. 2020; Lamers et al. 2020).

C-to-T intrahost transition was the most prevalent iSNV found in both groups studied. This change is characteristic of RNA editing by APOBEC enzymes (Smith and Sowden 1996; Vieira and Soares 2013) and has been reported by other groups when comparing SARS-CoV-2 strains (Simmonds 2020) and other coronaviruses (Di Giorgio et al. 2020).

Despite the fact we found SARS-CoV-2 within-host population variation in cancer patients, we do not know the mechanism(s) by which, or the anatomical site(s) where this variation is generated. In addition, other limitations of our study are evident, such as the age and disease severity differences between the two studied groups. Nevertheless, by generating a higher number of distinct variants, the virus can explore wider areas of the sequence landscape and test variants with different regulatory and structural changes. Variation may impact tissue tropism, protein expression and function, stability, immune escape, drug resistance, and pathogenicity. Further studies on SARS-CoV-2 diversity, especially in vulnerable patients with underlying comorbidities will shed light on our understanding of the wide spectrum of disease outcomes associated with COVID-19 in humans.

Acknowledgments

We would like to thank all the participants of the INCA COVID-19 Task Force, clinical staff and patients from the Brazilian National Cancer Institute for providing conditions and samples that enabled the conduction of this study. We also thank PhD. Renata Olício for providing support to Sanger DNA sequencing. We thank Thiago Muramatsu for the computational analysis support. We kindly acknowledge GISAID Database (https://www.gisaid.org/), the authors and laboratories for the SARS-CoV-2 genomes data shared. A table containing all genome sequences used in this article and the respective information can be found on Supplementary Table S1.

Data availability

Access to sequencing data files generated in this study is available in the Sequence Read Archive (SRA) database under project number PRJNA657032. SARS-CoV-2 complete genome sequences are available the GISAID database under IDs EPI_ISL_513513-513583.

Funding

This work was supported by grants to MAS from Brazilian Research Concil (305765/2015-9) and Carlos Chagas Filho Rio de Janeiro State Science Foundation (E-26/202.894/2017); to JPBV from Brazilian Research Concil (307042/2017-0) and Carlos Chagas Filho Rio de Janeiro State Science Foundation (202.640/2019, 211.562/2019 and 010.000162/2020); and to JA from National Institute of Allergy and Infectious Diseases (ZIA AI001273). JDS, LRG and BMA have received postdoctoral fellowships by the Brazilian Ministry of Health while conducting this study.

Members of INCA COVID-19 Task Force

Luiza M. Abdo, Maria Theresa Accioly, Lucas R. Almendra, Rodrigo O.C. Araujo, Elisa Bouret C. Barroso, Marcelo A. Bello, Anke Bergmann, Ricardo S. Bigni, Martin H. Bonamino, Franz S. Campos, Samuel Z.B. Cordeiro, Susanne

Crocamo, Magda S. da Conceição, Jesse L. da Silva, Carolina S. Dantas, Lucas Z. de Albuquerque, Roberto R.M. de Araujo Lima, Renata de Freitas, Fernando L. Dias, Jorge L.A. Dias, Michelle M.Q. dos Santos, André F. Duarte, Sima E. Ferman, Vanessa C. Fernandes, Erico L. Ferreira, Priscila S. Ferreira, Kelly M. Fireman, Carolina Furtado, Marianne M. Garrido, Renan G. Gomes Junior, Bruno A.A. Gonçalves, Juliana G. Gonçalves, Gustavo H.C. Guimarães, Nelson J. Jabour Fiod, Ana Cristina M. Leão, Décio Lerner, Valdirene S. Lima, Eduardo Linhares, Monique S.A. Lopes, Ianick S. Martins, Bruna P. Matta, Amanda S. Medeiros, Ana Cristina P. Mendes Pereira, Paulo A. Mora, Miguel A.M. Moreira, Daniela P. Oliveira, Alexandre M. Palladino, Diego J.G. Paula, Ana C. Pecego, Bárbara C. Peixoto, Patrícia A. Possik, Gelcio L. Quintella Mendes, Matheus A. Rajão, Maria D.M. Rocha, Fernando L.B. Rocha Gutierrez, Luciana de O.R. Rodrigues, Giovani B. Santos, Marcelo R. Schirmer, Karina L. Silva, Lilian S. Silva, Antonio A.D. Souto, Leandro S. Thiago, Luiz C.S. Thuler, Fabiana Tonellotto, Gisele M. Vasconcelos, Dolival L. Veras Filho.

Author contributions

JDS, LRG, BMA, and MAS conceived and designed the study. JDS, LRG, and BMA optimized all reagents and performed the sequencing and analysis. ACM collected clinical data. CC, JA, JPBV, ACM, and MAS provide expert advice on experimental planning and data interpretation. JDS, LRG, BMA, and MAS wrote the article. JDS, LRG, BMA, CC, JA, JPBV, ACM, and MAS revised and edited the article.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Alves, B. M. et al. (2017) 'Characterization of HIV-1 near Full-Length Proviral Genome Quasispecies from Patients with Undetectable Viral Load Undergoing First-Line HAART Therapy', *Viruses*, 9: 392.

Asselta, R. et al. (2020) ' 'ACE2 and TMPRSS2 Variants and Expression as Candidates to Sex and Country Differences in COVID-19 Severity in Italy', *Aging*, 12: 10087–98.

Begum, F. et al. (2020) 'Specific mutations in SARS-CoV2 RNA Dependent RNA Polymerase and Helicase Alter Protein Structure, Dynamics and Thus Function: Effect on Viral RNA Replication', *bioRxiv*.

Candido, D. S., Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic Network. et al. (2020) 'Evolution and Epidemic Spread of SARS-CoV-2 in Brazil', *Science*, 369: 1255–60.

Centers for Disease Control and Prevention (2020), 'CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel', <https://www.fda.gov/media/134922/download> accessed 30 April.

de Melo, A. C., et al.; on behalf of the Brazilian National Cancer Institute COVID-19 Task Force. (2020) 'Cancer Inpatient with COVID-19: A Report from the Brazilian National Cancer Institute', *PLoS One*, 15: e0241261.

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

Dudley, D. M. et al. (2014) 'Cross-Clade Simultaneous HIV Drug Resistance Genotyping for Reverse Transcriptase, Protease, and Integrase Inhibitor Mutations by Illumina MiSeq', *Retrovirology*, 11: 122.

Ellinghaus, D. et al. (2020) 'Genomewide Association Study of Severe Covid-19 with Respiratory Failure', *N Engl J Med*, 383: 1522–34.

Fang, S. et al. (2021) ' 'GESS: A Database of Global Evaluation of SARS-CoV-2/hCoV-19 Sequences', *Nucleic Acids Research*, 49: D706–D14.

Fung, T. S., and Liu, D. X. (2019) ' 'Human Coronavirus: Host-Pathogen Interaction', *Annual Review of Microbiology*, 73: 529–57.

Gallus, S., Lugo, A., and Gorini, G. (2020) 'No Double-Edged Sword and No Doubt about the Relation between Smoking and COVID-19 Severity', *European Journal of Internal Medicine*, 77: 33–5.

Guindon, S. et al. (2010) 'New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3', *Systematic Biology*, 59: 307–21.

Hadfield, J. et al. (2018) 'Nextstrain: Real-Time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.

Hikmet, F. et al. (2020) 'The Protein Expression Profile of ACE2 in Human Tissues', *Molecular Systems Biology*, 16: e9610.

Jary, A. et al. (2020) 'Evolution of Viral Quasispecies during SARS-CoV-2 Infection', *Clinical Microbiology and Infection*, 26: 1560.e1–4.

Keane, T. M. et al. (2006) 'Assessment of Methods for Amino Acid Matrix Selection and Their Use on Empirical Data Shows That Ad Hoc Assumptions for Choice of Matrix Are Not Justified', *BMC Evolutionary Biology*, 6: 29.

Kirtipal, N., and Bharadwaj, S. (2020) 'Interleukin 6 Polymorphisms as an Indicator of COVID-19 Severity in Humans', *J Biomol Struct Dyn*, 1–3.

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

Lamers, M. M. et al. (2020) 'SARS-CoV-2 Productively Infects Human Gut Enterocytes', *Science*, 369: 50–4.

Lee, L. Y. W. et al. (2020) 'COVID-19 Mortality in Patients with Cancer on Chemotherapy or Other Anticancer Treatments: A Prospective Cohort Study', *The Lancet*, 395: 1919–26.

Li, Q. et al. (2020) 'Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia'', *New England Journal of Medicine*, 382: 1199–207.

Mohammadi, A. et al. (2020) 'SARS-CoV-2 Detection in Different Respiratory Sites: A Systematic Review and Meta-Analysis', *EBioMedicine*, 59: 102903.

Rambaut, A. et al. (2020) ' 'A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology', *Nature Microbiology*, 5: 1403–7.

Resende, P. C. et al. (2020), 'Genomic surveillance of SARS-CoV-2 reveals community transmission of a major lineage during the early pandemic phase in Brazil', *bioRxiv*.

Shen, Z. et al. (2020) 'Genomic Diversity of Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients with Coronavirus Disease 2019', *Clinical Infectious Diseases*, 71: 713–20.

Simmonds, P. (2020) 'Rampant C–>U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and

Consequences for Their Short- and Long-Term Evolutionary Trajectories', *mSphere*, 5: e00408-20.

Singer, J. et al. (2020), 'CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation', *Preprints* 2020060225.

Singh, A. K. et al. (2020) 'Prevalence of co-Morbidities and Their Association with Mortality in Patients with COVID-19: A Systematic Review and Meta-Analysis', *Diabetes, Obesity and Metabolism*, 22: 1915–24.

Smith, H. C., and Sowden, M. P. (1996) 'Base-Modification mRNA Editing through Deamination — the Good, the Bad and the Unregulated', *Trends in Genetics*, 12: 418–24.

Tahvildari, A. et al. (2020) 'Clinical Features, Diagnosis, and Treatment of COVID-19 in Hospitalized Patients: A Systematic Review of Case Reports and Case Series', *Frontiers in Medicine*, 7: 231.

Vieira, V. C., and Soares, M. A. (2013) 'The Role of Cytidine Deaminases on Innate Immune Responses against Human Viral Infections', *BioMed Research International*, 2013: 1–18.

Willan, J. et al. (2020) 'Care of Haematology Patients in a COVID-19 Epidemic', *British Journal of Haematology*, 189: 241–3.

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

Yang, K. et al. (2020) 'Clinical Characteristics, Outcomes, and Risk Factors for Mortality in Patients with Cancer and COVID-19 in Hubei, China: A Multicentre, Retrospective, Cohort Study', *The Lancet Oncology*, 21: 904–13.

Zhang, L. et al. (2020) 'Clinical Characteristics of COVID-19-Infected Cancer Patients: A Retrospective Case Study in Three Hospitals within Wuhan, China', *Annals of Oncology*, 31: 894–901.