scientific reports

OPEN

Unraveling the protective genetic architecture of COVID-19 in the Brazilian Amazon

Maria Clara Barros[1,12], Jorge Estefano Santana de Souza[2,9,12], Daniel Henrique F. Gomes[2], Catarina Torres Pinho[1], Caio S. Silva[1], Cíntia Braga-da-Silva[1], Giovanna C. Cavalcante[1], Leandro Magalhães[3], Jhully Azevedo-Pinheiro[1], Juarez Antônio Simões Quaresma[4,5], Luiz Fábio Magno Falcão[5], Patrícia Fagundes Costa[6], Cláudio Guedes Salgado[6], Thiago Xavier Carneiro[7], Rommel Rodrigues Burbano[7], José Ricardo dos Santos Vieira[8], Sidney Santos[10], Giordano Bruno Soares-Souza[3]✉, Sandro José de Souza[2,9,11]✉ & Ândrea Ribeiro-dos-Santos[10]✉

Despite all the efforts acquired in four years of the COVID-19 pandemic, the path to a full understanding of the biological mechanisms involved in this disease remains complex. This is partly due to a combination of factors, including the inherent characteristics of the infection, socio-environmental elements, and the variations observed within both the viral and the human genomes. Thus, this study aimed to investigate the correlation between genetic host factors and the severity of COVID-19. We conducted whole exome sequencing (WES) of 124 patients, categorized into severe and non-severe groups. From the whole exome sequencing (WES) association analysis, four variants (rs1770731 in *CRYBG1*, rs7221209 in *DNAH17*, rs3826295 in *DGKE*, and rs7913626 in *CFAP46*) were identified as potentially linked to a protective effect against the clinical severity of COVID-19, which may explain the less severe impact of COVID-19 on the Northern Region. Our findings underscore the importance of carrying out more genomic studies in populations living in the Amazon, one of the most diverse from the point of view of the presence of rare and specific alleles. To our knowledge, this is the first WES study of admixed individuals from the Brazilian Amazon to investigate genomic variants associated with the clinical severity of COVID-19.

Keywords *SARS-CoV-2*, COVID-19, WES, Amazon, Clinical severity

Throughout the four-year duration of the pandemic, extensive research has been dedicated to understanding COVID-19. As a result, non-genetic factors within the host, such as advanced age, male sex, and comorbidities such as hypertension, diabetes, and cardiovascular diseases, have already been associated with a worse prognosis in COVID-19.

However, these identified risk factors do not fully account for the diverse spectrum of clinical manifestations observed in the disease. It is noteworthy that severe forms of the disease, marked by poor oxygen saturation and lung damage, have not been confined solely to individuals within the aforementioned high-risk groups. Rather, they have also manifested in young individuals, with or without comorbidities[1–4]. These features suggest, according to Novelli et al., several hypotheses, including a breakdown of immunological tolerance, the viral load, an innate immune inefficiency, and the presence of common or rare risk alleles in protein-coding genes

[1]Laboratory of Human and Medical Genetics (LGHM) / Graduate Program Genetics and Molecular Biology (PPGBM), Federal University of Pará (UFPA), Belém 66075-110, PA, Brazil. [2]Graduate Program Bioinformatics, Federal University of Rio Grande do Norte (UFRN), Natal 59078-970, RN, Brazil. [3]Vale Technological Institute (ITV), Belém 66055-090, PA, Brazil. [4]Laboratory of Infectious Disease, School of Medicine, Federal University of Pará (UFPA), Belém 66075-110, PA, Brazil. [5]Department of Infectious Disease, School of Medicine, State University of Pará (UEPA), Belém 66087-670, PA, Brazil. [6]Dermatology and Immunology Laboratory, Federal University of Pará (UFPA), Marituba 67105-290, PA, Brazil. [7]Molecular Biology Laboratory, Ophir Loyola Hospital (HOL), Belém 66063-240, PA, Brazil. [8]Institute of Biological Sciences, Federal University of Pará (UFPA), Belém 66075- 110, PA, Brazil. [9]Multidisciplinary Bioinformatics Center (BiOMe), Federal University of Rio Grande do Norte (UFRN), Natal 59078-970, RN, Brazil. [10]Center of Oncology Research, Federal University of Pará (UFPA), Belém 66073- 005, PA, Brazil. [11]DNA-GTX Bioinformatics, Natal, RN, Brazil. [12]Maria Clara Barros and Jorge Estefano Santana de Souza have contributed equally to this work. ✉email: jwojwo@gmail.com; sandro@i2bio.org; akelyufpa@gmail.com

important for the biological cycle of the virus[5]. Therefore, genetic host factors may influence the complexity of the virus-host interaction.

In Brazil, there remains a scarcity of exomes-wide sequencing (WES) studies conducted on individuals infected by COVID-19 to date. From the exome analysis of 83 Brazilian couples of the Southeast region, with one partner infected and symptomatic, and the other uninfected, Castelli et al. (2021) demonstrated an association between alleles of *HLA-A* and *HLA-DRB1*genes and the susceptibility to symptomatic infection[6]. Another study by the same group observed missense variants in *MUC22*gene that may act as a protective factor against severe COVID-19. These variations were found with a higher prevalence in 87 individuals older than 90 years with mild symptoms or asymptomatic when compared to 55 individuals younger than 60 years who had a severe disease or died due to COVID-19[7].

Moreover, Secolin et al. (2021) detected rare variants in COVID-19-related genes, including *SLC6A20*, *LZTFL1*, *XCR1*, and *FURIN*, exclusive to the 88 exomes of Brazilians from the Southeast region[8]. Another Brazilian exome study for COVID-19 comes from Santos-Rebouças et al. (2022), which identified rare variants in *FREM1*, *MPO*, *POLG*, *C6*, *C9*, *ABCA4*, *ABCC6*, and *BSCL2*genes related to a higher risk of Multisystem Inflammatory Syndrome in Children (MIS-C) development, a complication of severe COVID-19, in 16 children living in Southeast Region[9].

Importantly, these exomes are restricted only to the South and Southeast regions (which have a high level of European genetic contribution). However, it is known that the Brazilian population has a high degree of ethnic admixture influenced mainly by three parent populations: Europeans, Africans and Native Americans, resulting in an extremely heterogeneous genetic structure that is unevenly distributed between regions, especially in the Northern region. Another challenge is the application of GWAS in this country, since due to the high rate of miscegenation, genetic findings associated with diseases may vary between different countries and within Brazil and within Brazilian regions[10–12].

Considering that *SARS-CoV-2* infection and the progression of COVID-19 are influenced by host proteins (and the exome precisely investigates protein-coding genes), the significance of using this technique to evaluate potential genetic factors linked to the severity of this pathology becomes evident.

Here, we emphasize the importance of identifying genetic protective and risk factors within the human host in the pathogenesis of *SARS-CoV-2*. Therefore, investigating the association between genetic factors and the severity of COVID-19 within an Amazonian population from Northern Brazil is imperative. Such studies are crucial for enhancing our understanding of the clinical progression of this disease, thereby enabling improvements in the healthcare system and facilitating efforts to reduce mortality rates, especially in populations that are underrepresented in genomic worldwide studies and databases, such as ours.

Results

Clinical results

The demographic and clinical profiles of the patients are outlined in Table 1. Our analysis comprised 124 individuals, categorized into two groups based on their clinical status: 68 with a severe form of COVID-19 (sCOV) and 56 with a non-severe form (nsCOV).

The average age of the sCOV group was 56 (±14.2), whereas in the nsCOV group was 39 (±14.6), with a significant association being observed between age and the clinical severity of COVID-19 ($p < 0.01$). Regarding

Variables	sCOV ($n = 68$)	nsCOV ($n = 56$)	OR (95% CI)	p-value
Age*	56 (±14.3)	39 (±14.6)		**< 0.01**
Sex**				0.28
Female	40	27		
Male	28	29		
Lineage***				
B.1.1	22	15	1.19 (0.50–2.8)	0.6
P.1	31	20	1.36 (0.61–3.08)	0.4
Comorbidities***	58	19	14.8 (5.4–46.3)	**< 0.01**
SAH	37	4	16.2 (5.49–59.7)	**< 0.01**
Obesity	15	5	2.91 (0.91–11.05)	0.05
Chronic Heart Disease	4	4	2.80 (0.78–12.7)	0.10
Overweight	26	3	19.5 (4.40–176.5)	**< 0.01**
Ancestry				
European	0.628	0.608		
African	0.109	0.110		
Native American	0.263	0.281		

Table 1. Bioanthropological and clinical data of patients with COVID-19. sCOV: severe form of COVID-19; nsCOV: non-severe form of COVID-19; SAH: systemic arterial hypertension. OR: Odds Ratio. *Wilcoxon test. ** Chi-squared test. *** Fisher test.

gender distribution, there was a predominance of females in the severe group (54,8%); however, no statistically significant difference was observed ($p=0.28$).

Among the 21 strains identified in our study, two emerged as more prevalent: P.1 ($p=0.4$; OR = 1.36; 95% CI = 0.61–3.08) and B.1.1 ($p=0.6$; OR = 1.19; 95% CI = 0.50–2.8). However, neither demonstrated a significant association with the clinical severity of COVID-19.

Regarding comorbidities, a significant association with pre-existing medical conditions was found, indicating a 14-fold increase in the risk of developing severe COVID-19 ($p<0.01$; OR = 14.8; 95% CI = 5.4–46.3) among individuals. Among the comorbidities analyzed, hypertension was associated with a 16-fold greater risk ($p<0.01$; OR = 16.2; CI 95% 5.49–59.7) while overweight exhibited a 19-fold greater risk ($p<0.01$; OR = 19.5; CI 95% = 4.40–176.5) of severe COVID-19.

Population structure analysis

Genomic data obtained through WES allowed for the evaluation of population structuring patterns present within the sample of COVID-19 patients. For the Admixture software analyses, a supervised approach was employed in a three-hybrid model. This model considered samples of individuals belonging to three distinct groups: Europeans, represented by the populations of the Iberian Peninsula (IBS – 1kG); Africans, represented by the Yoruba people (YRI/AFR – 1kG); and Native Americans, represented by indigenous populations from Brazil, Guatemala and Colombia (AMR - HGDP) (Fig. 1).

In addition, PCA was also performed to verify patterns of genetic diversity and population structuring of COVID-19 patients in relation to populations from different biogeographical regions. For this purpose, all the populations from the 1,000 Genomes, HGDP and COVID patients' datasets were used. As seen in Fig. 2, patients affected by COVID-19 (in orange/lilac) show a genetic diversity pattern similar to that of Latin American individuals deposited in the 1,000 Genomes panel database. Although there is great diversity in the contribution of Native Americans, European, and African populations in American populations, Brazilian individuals from the Amazon region affected by COVID-19 show a greater contribution from the indigenous populations (Native Americans) and Europeans, with a low African contribution, a phenomenon also observed in other Latin American populations.

Genomic association analyses

Variant calling in the exome data identified 2,291,530 variants among all subjects. We performed a quality filtering using vcftools filters --missing-site and --missing-indv lower than 10%. After this filtering, 20,546 variants were identified, comprising 15,369 SNPs and 5,177 INDELs. Table 2 illustrates the distribution of variants between patients with sCOV and nsCOV, as well as their corresponding putative impact, class, and types defined by SnpEff. While the overall variation in variant counts was minimal, few observations emerged: six high-impact variants were found only in the nsCOV group, and six missense and three frameshift variants were present only in the sCOV group.

Among these, 445 were classified as damaging according to the following computer prediction algorithms FATHMM, MutationAssessor, MutationTaster, SIFT, MetaSVM and PROVEAN and their characteristics regarding their putative impact, class, and types defined by SnpEff are described in Table 3. No difference was observed between clinical groups.

Due to our sample size, we considered a less restrictive p-value ($p<5\times10-4$) with FDR correction. Therefore, the single association analysis results represent candidate variants that will require further studies for validation of their role in the clinical severity of COVID-19. Regarding the exome-wide gene-based association analysis, no gene was found with a statistically significant p-value.

Through single association analyses, while considering population structuring, age and comorbidities as covariates, we identified four variants (rs1770731 in *CRYBG1*, rs7221209 in *DNAH17*, rs3826295 in *DGKE* and rs7913626 in *CFAP46*) potentially linked to a protective effect against the clinical severity of COVID-19 (as illustrated in Fig. 3 and detailed in Table 4). These associations reached a suggestive significance threshold, denoted by the red horizontal line in Fig. 3.

Moreover, in terms of metabolic processes, it becomes apparent that certain genes potentially implicated in the severe pathogenesis of COVID-19 are acting in the cilium movement involved in cell motility (*DNAH17* and *CFAP46*) and platelet activation (*DGKE*). These findings suggest an association with a pathogenic mechanism of *SARS-CoV-2* that has not been fully characterized.

To investigate the distribution of these variants within our study population, we selected 95 individuals from 13 Indigenous of America populations living in the Northern Region of Brazil. These included Araweté (ARW), Zo'é/Poturujara (ZOE), Wayãpy (WPI), and Awa-Guajá (AWA) from the Tupi-Guarani language group; Asurini do Koatinemo (AKW), and Asurini do Trocará (AST) from the Asurini language group, which belong to the

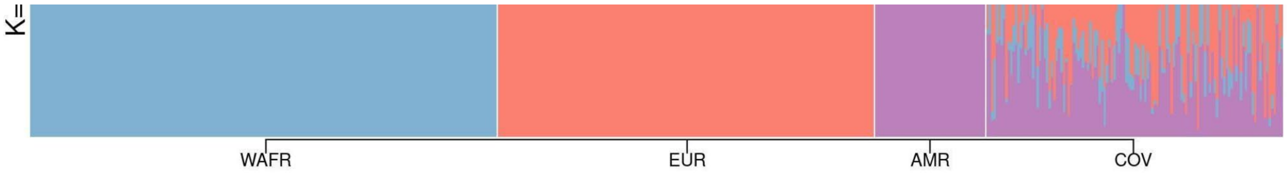

Fig. 1. Supervised analysis of the population structuring patterns present in the study sample. WAFR - West Africa (YRI), EUR - Europe (IBS), AMI - Indigenous of America and COV - COVID-19 affected patients.

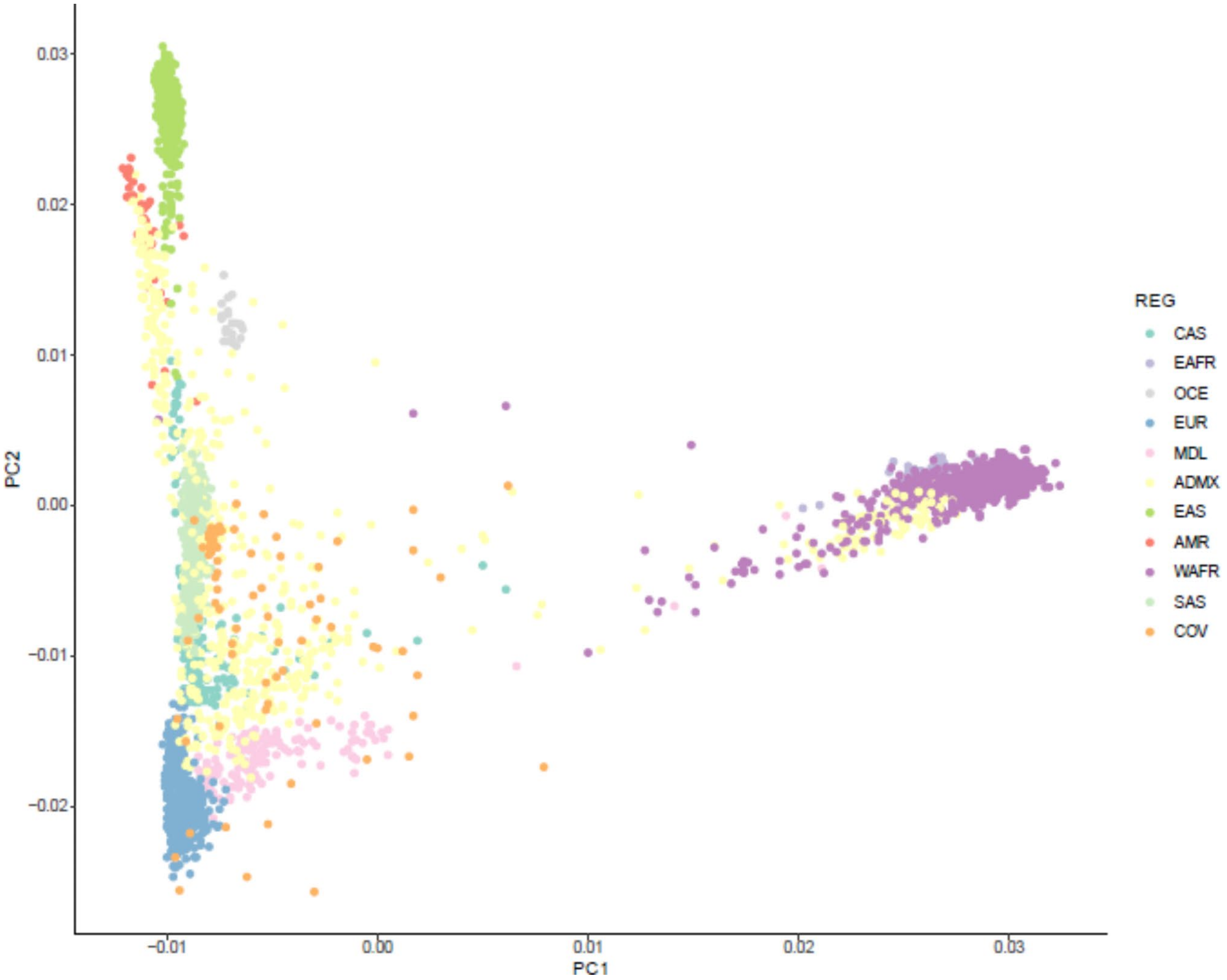

Fig. 2. PCA analysis of genomic ancestry based on allele frequency present in the study subjects and in the following subjects: CAS - Central Asia, EAFR - East Africa, OCE - Oceania, EUR - Europe, MDL - Middle East, ADMX - Mixed Populations of the Americas, EAS - East Asia, AMR - Amerindians, WAFR - West Africa, SAS - South/Southeast Asia, COV - COVID-19 patients.

Tupi-Guarani language truck; Arara/Arara do Iriri (ARA), Tiriyó (TYR) and Hixkaryana (HIK) from the Karib language group; Kayapó (KAY) and Xikrin (XIK) from the Macro-Jê language group; Munduruku (MND) from the Munduruku language group, belonging to the Tupi-Guarani language truck; and Palikur (PL) from the parikwaki language group.

The allelic frequencies of the variants observed in the study populations (COVID-19 patients and Indigenous of America) and the five continental populations are presented in Table 5. We also performed a Fisher's Exact Test with Bonferroni Correction to verify whether there was a statistical difference between the frequencies observed.

The rs1770731 polymorphism exhibits a similar frequency in our cohort (87%) compared to other populations (AMR: 89%; EUR: 88%; EAS and EAS: 87%), except for AFR (77%) and our Indigenous of America population (AMI: 96%). The rs7221209 frequency is considerably higher in our cohort (COV: 82% and AMI: 81%) compared to other continental populations, especially EAS (64%) and varied significantly when compared to COVID patients and the EUR population (padj < 0.05).

The rs3826295 variant is higher in our cohort (COV: 84% and AMI: 94%) compared to AMR (79%), EUR (72%) and with a significant difference to AFR (60%, padj = 0.0004). Regarding the rs7913626 frequency, it is greater in COVID-19 patients (76%) and Indigenous of America (81%) compared to other continental populations and statistically different from the populations EUR (padj = 0.0002) and SAS (padj = 0.01).

As shown in Fig. 4, a varying distribution of the PRS of the four associated variants in different populations was found (oneway ANOVA, F (7, 3369) 18.19, $p < 2e\text{-}16$). Based on this, it was possible to observe that the PRS of the non-severe COVID patients (nsCOV) had a significant difference when compared to the other populations (Tukey, p adj < 0.01), including severe COVID patients (sCOV) (Tukey, padj = 0.002), except for the Indigenous of America population (AMI) (Tukey, padj = 0.99).

	All	sCOV	nsCOV
Impact			
High	170	51	57
Low	4.886	3.334	3.275
Moderate	3.177	2.514	2.502
Modifier	13.633	7.930	7.827
Class			
Missense	3.012	2.446	2.440
Nonsense	16	13	13
Silent	3.219	2.761	2.706
Type			
Exon	148	111	115
Frameshift	59	18	15
Intron	13.334	7.077	7.156
Non synonymous coding	3.003	2.431	2.437
Synonymous coding	3.218	2.705	2.760
Stop gained	17	13	13
Stop lost	7	7	7

Table 2. Distribution of variants in all COVID-19 patients and between the severe (sCOV) and non-severe (nsCOV) groups according to functional impact, functional class, and type defined by SnpEff.

	ALL
Impact	
High	19
Low	11
Moderate	393
Modifier	66
Class	
Missense	395
Nonsense	4
Silent	11
Type	
Exon	2
Intron	59
Non synonymous coding	393
Synonymous coding	11
Stop gained	4
Start lost	2

Table 3. Distribution of damaging variants in all COVID-19 patients according to functional impact, functional class, and type defined by SnpEff.

Discussion

In the Modern Age, few diseases have had such a profound impact as to isolate entire human populations across multiple fronts - socioeconomic, scientific, and within health systems -, unequivocally exposing the inequalities among countries worldwide. The COVID-19 pandemic stands as a poignant example. Despite the substantial and ever-expanding understanding of the virus´ biology and associated host risk factors, critical aspects of its pathogenesis and clinical progression remain to be uncovered. Furthermore, there is a pressing need to explore the host genetic factors that influence populations living in poorly represented regions.

Despite having great biological heterogeneity distributed between traditional and admixed populations, the Brazilian Amazon population is one of the least represented from a genomic point of view in the world's databases[13–15]. Hence, investigating individual and/or population genetic variability holds significant promise, particularly concerning the health outcomes of diverse populations. This approach identifies unique responses and metabolic pathways to drugs, critical in populations with heightened susceptibility to genetic, inflammatory, and infectious diseases. Moreover, it sheds light on potential resistance or adverse effects of treatment[16].

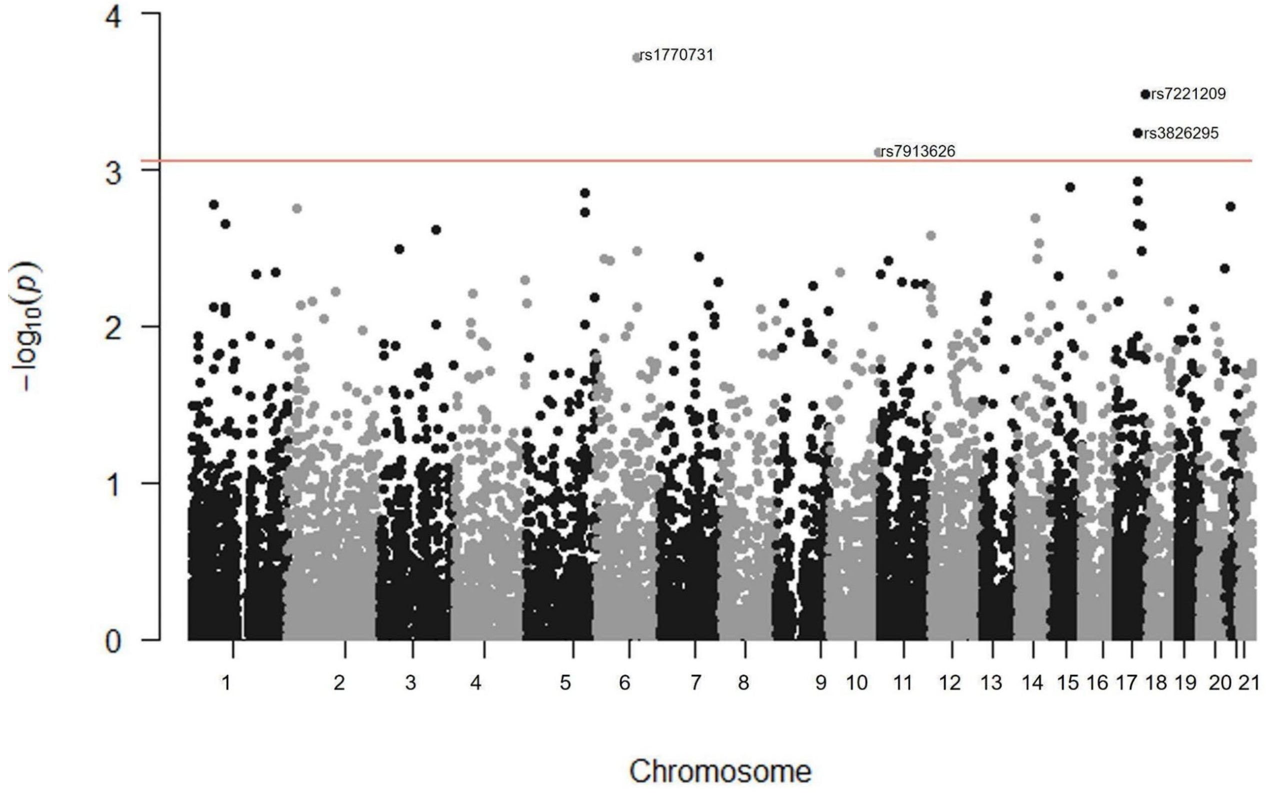

Fig. 3. Manhattan plot of the WES of 130 participants (non-severe ($n = 56$) vs. severe ($n = 68$)), highlighting four peaks with possibly association signals for severe cases of COVID-19. The WES analysis results are shown on the y-axis as -log10 (p-value), and on the x-axis is the chromosomal location. The red horizontal line illustrated the suggestive genome-wide association threshold ($p < 5 \times 10\text{-}4$).

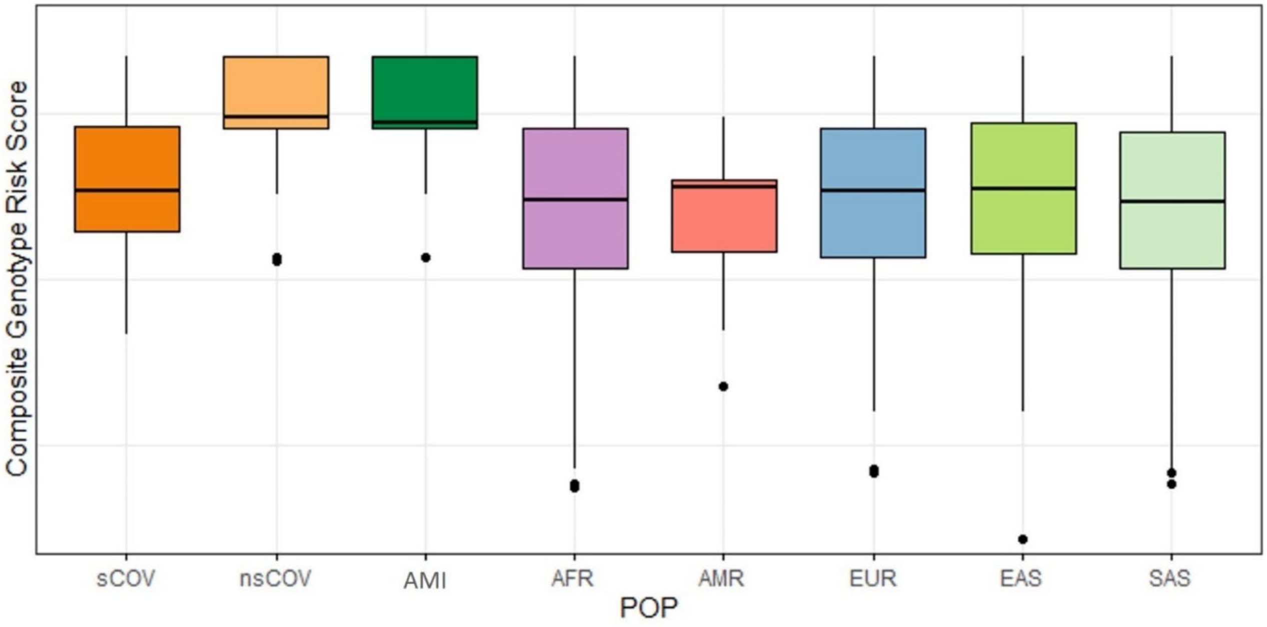

Fig. 4. The PRS distribution of the associated variants among the study populations (sCOV, nsCOV and AMI) and continental populations (AFR, AMR, EUR, EAS, and SAS).

pos	dbSNP variant	EA	Consequence type	genes	*p*-value*	AF	OR	gene's associated metabolic processes	gene's associated disorders
6:106553581	rs1770731	T	intron variant	CRYBG1	3,19E+4	0.876	-3.68	carbohydrate binding	Melanoma
17:78539866	rs7221209	C	synonymous variant	DNAH17	6,31E+4	0.827	-3.20	cilium movement involved in cell motility	Infertility
17:56862376	rs3826295	A	3 prime UTR variant	DGKE	5,03E+4	0.842	-3.5	platelet activation	Hemolytic uremic syndrome
10:132851371	rs7913626	G	intron variant	CFAP46	7,77E+4	0.767	-2.98	cilium movement involved in cell motility	B-Cell Adult Acute Lymphocytic Leukemia

Table 4. Genetic variants possibly associated with COVID-19 severity. *GLM; pos: position; EA: effect allele; OR: OddsRatio.

dbSNP variant	pos (hg38)	EA	COV	AMI	gnomAD AFR	gnomAD AMR	gnomAD EUR	gnomAD EAS	gnomAD SAS
rs1770731	6:106553581	T	0.876	0.969	0.77	0.88	0.90	0.85	0.86
rs7221209	17:78539866	C	0.827	0.810	0.78	0.71	0.76	0.78	0.73
rs3826295	17:56862376	A	0.842	0.94	0.62	0.81	0.70	0.88	0.61
rs7913626	10:132851371	G	0.767	0.818	0.84	0.72	0.66	0.44	0.65

Table 5. Comparison of allelic frequencies of variants observed between study populations (COV and AMI) and continental populations (AFR, AMR, EUR, EAS, and SAS) described in gnomAD database. *pos: position; EA: effect allele.

In agreement with the results of non-genetic factors associated with a worse COVID-19 prognosis obtained in other studies[3,17,18], a significant association was observed in the present study between age ($p < 0.01$) and the presence of comorbidities ($p < 0.01$; OR = 14.8; 95% CI = 5.4–46.3).

Among the various comorbidities that are associated with COVID-19, we observed that individuals with systemic arterial hypertension have a 16-fold greater risk of developing the severe form of this disease, probably due to the hyperactivation of the renin-angiotensin-aldosterone system (RAAS) which increases the inflammatory response and the recruitment of cytokines, causing endothelial damage[19,20].

Another comorbidity associated with severe COVID-19 was overweight, with a 19-fold greater risk ($p < 0.01$; OR = 19.5; CI 95% = 4.40–176.5). This is possibly due to obesity impairment of pulmonary function characterized by a decline in expiratory reserve volume and functional capacity, and due to the release of inflammatory cytokines TNF-α and IL-6 that already occurs in obesity and could exacerbate the severe cases of COVID-19[21,22].

The viral sequencing was performed to find out if the viral strains influenced clinical severity. No significant association was found between the most prevalent strains [P.1 ($p = 0.4$; OR = 1.36; 95% CI = 0.61–3.08) and B.1.1 ($p = 0.6$; OR = 1.36; 95% CI = 0.50–2.8)], suggesting that the host genetics may have a major influence on COVID-19 clinical severity.

Given the diverse genetic composition of the Brazilian population, shaped by multiple ancestral groups, it becomes imperative to assess population structuring patterns to control possible biases of miscegenation and to discover more about the genetic variability of these populations[10,11]. Therefore, a supervised analysis was carried out using the Admixture software, focusing on the three main ancestries that have contributed to the formation of the Brazilian population: Europeans, Africans, and Indigenous of America.

Figure 2 shows a notable prevalence of genetic ancestry tracing back to the original peoples of the Americas among COVID-19 patients, possibly due to the Brazilian formation process, which was highly heterogeneous and where indigenous communities are still prominently situated. Furthermore, a binomial GLM analysis was performed to investigate whether genetic ancestry has an influence on the clinical severity of *SARS-CoV-2* infection. However, no significant association was found (Table 1).

Nevertheless, studies such as Shelton et al. (2021) and Mathur et al. (2021) have already observed variations in the clinical progression of COVID-19 among individuals with different predominant ancestries. Specifically, individuals with a greater contribution of African American ancestry were more likely to be hospitalized, while those with Asian and African ancestry exhibited elevated risks of hospitalization, both compared to populations with greater European ancestry, after accounting for differences in sociodemographic, clinical, and household characteristics[23,24]. Despite these findings, studies associating individual genetic ancestry with COVID-19 susceptibility or severity remain scarce.

The genetic architecture of COVID-19 regarding pathogenicity mechanisms and host response is complex and remains unclear. In our study, we identified four variants potentially linked to a protective effect against the clinical severity of COVID-19, which have not been previously explored in GWAS studies of COVID-19.

Regarding their consequence type, we identified two intron variants with modifier impact (rs1770731 and rs7913626), one synonymous variant with low impact (rs7221209) and a 3 prime UTR variant with modifier impact (rs3826295). Because of this, it was not possible to predict the regulatory effects of the variants, however, they may be involved in the expression in some tissues, such as testis and hypothalamus (rs1770731); pancreatic islets and substantia nigra of the brain (rs7913626); cortex, lung left ventricle of the heart and Th1 memory (rs7221209) and colon, esophagus and small intestine (rs3826295).

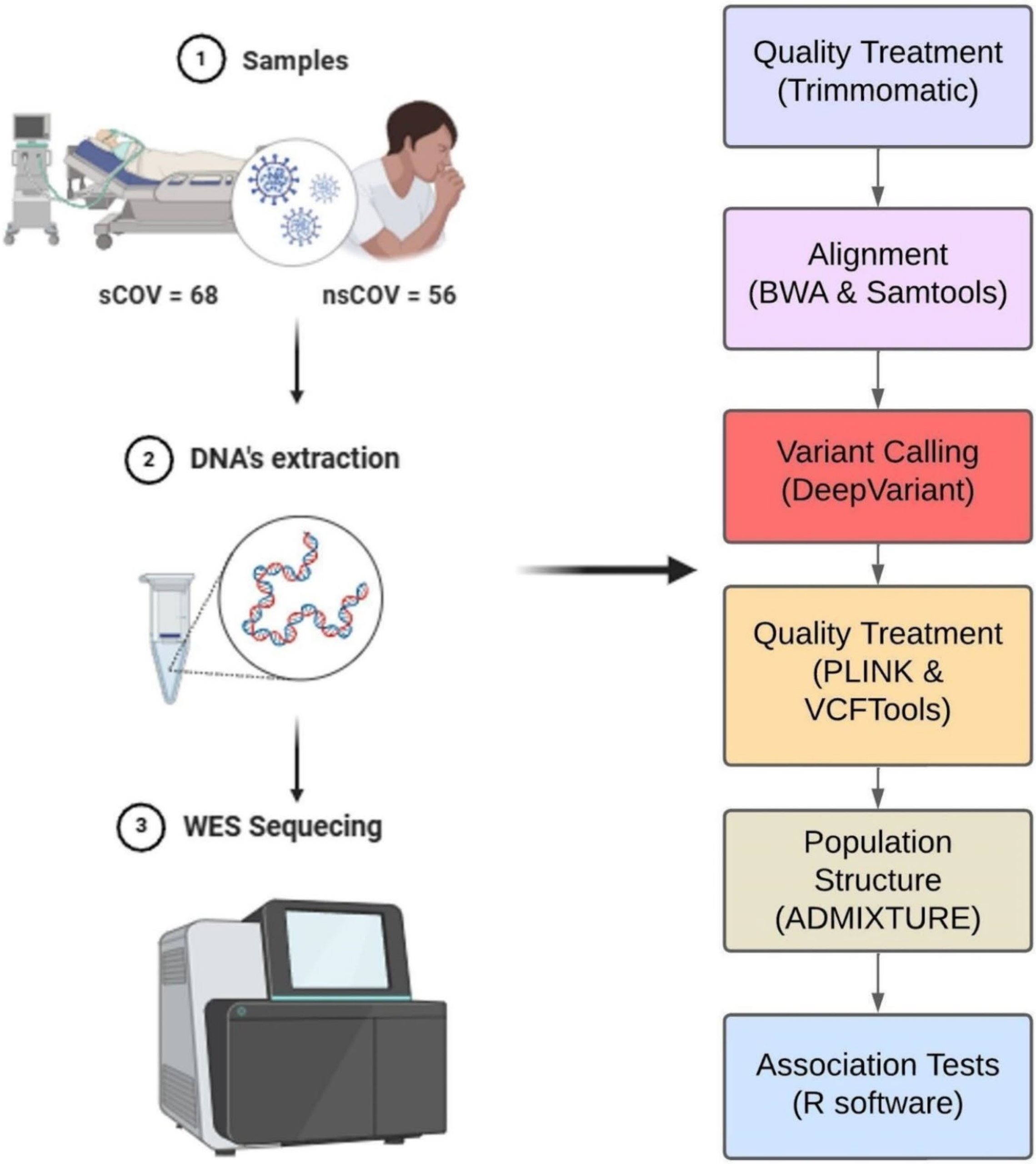
1 Samples
sCOV = 68
nsCOV = 56
2 DNA's extraction
3 WES Sequecing
Quality Treatment (Trimmomatic)
Alignment (BWA & Samtools)
Variant Calling (DeepVariant)
Quality Treatment (PLINK & VCFTools)
Population Structure (ADMIXTURE)
Association Tests (R software)

Fig. 5. Flowchart of the study methodology. DNA extraction and Whole Exome Sequencing (WES) were performed in 124 samples, divided in two groups: 68 patients with severe COVID-19 (sCOV) and 56 patients with non-severe COVID-19 (nsCOV). The figure also shows the main steps used in the exome's pipeline. In dark blue, reads's quality treatment by trimming and filtering. In lilac, mapping reads using a reference genome. In red, variant calling by DeepVariant pipeline. In orange, samples' quality treatment. In gray, population structure analysis. And in light blue, association tests analysis in R software.

To our knowledge, the genes *CRYBG1* and *CFAP46* have not been previously associated with COVID-19 or other respiratory diseases, apart from *DNAH17* and *DGKE*. Nevertheless, it is important to be aware of their roles in other disorders, so we can better understand their functions and pathways in COVID-19.

The rs177073 variant is in the *CRYBG1* (Crystallin Beta-Gamma Domain Containing 1, alias AIM1) gene, a tumor suppressor in melanoma and a potential oncogene in triple-negative breast cancer cases due to novel fusion genes[25]. Besides that, the *CRYBG1* gene was associated with the actin cytoskeleton and suppressed cytoskeletal remodeling and invasive properties in non-malignant prostate epithelial cells[26]. Therefore, a higher frequency of this gene can influence a non-aggressive form of certain diseases.

Besides this, COVID-19 has been increasingly recognized as a disease that primarily affects the elderly, resulting in higher mortality in this age group. It has been hypothesized that this happens especially because the coronavirus has a specific tropism for senescent cells in the lungs, which are more common in older people[27,28]. In our study, we found a variant in the *DNAH17* (Dynein Axonemal Heavy Chain 17) gene that protected our population from more serious forms of the disease. This gene encodes a dynein heavy chain protein that is normally observed in individuals with isolated male infertility due to several morphological anomalies of sperm cells since this gene is involved in cilium movement in cell motility[29].

Furthermore, the *DNAH17* gene is frequently mutated in hepatocellular carcinoma patients[30], and since the incidence of hepatic abnormalities significantly increases after COVID-19 infection due to *SARS-CoV-2* effects on liver and/or kidney, *deAndrés-Galiana et al.* (2022) observed *DNAH17* overexpressed in a genetic signature between healthy controls and COVID-19 patients[31].

Nevertheless, in 2024, Andrawus and collaborators stated that the expression of the *DNAH17* gene decreases with advancing age and that this gene is associated with longevity[32]. With this information in mind, it is reasonable to state that the effects of this gene on COVID-19 are still not fully understood. This gene may act differently in our study population, since it has a high indigenous genetic contribution and there are few genomic and epigenomic studies on its composition. Thus, further studies are needed to clarify the specific mechanisms by which *DNAH17* is involved in *SARS-CoV-2* infection.

The *DGKE* gene phosphorylates diacylglycerol to phosphatidic acid, ultimately activating protein kinase C that may induce a hypercoagulable state with platelet activation and developing thrombosis[33,34]. A novel likely pathogenic variant was observed in severe patients with COVID-19 from the WES of Bulgarian patients[35].

This gene is associated with atypical hemolytic uremic syndrome (aHUS), a disease similar to COVID-19 since both lead to venous thromboembolism, microvascular thrombosis, and multi-organ damage due to hyperactivation of the complement system[36]. The loss of the gene in aHUS may cause modifications of the vascular tone actin cytoskeleton, secretion of prothrombotic and antithrombotic factors, and the activation of platelets[37].

From this, we can hypothesize that the *DGKE* gene is preventing one of the most severe forms of COVID-19, characterized by thrombosis and organ damage due to hyperactivation of the complement system, since: (1) COVID-19 and aHUS have similar complications; (2) the lack of the *DGKE* gene aggravates aHUS and (3) the rs3826295 variant is quite frequent in our populations study, 84% in COVID-19 patients and 94% in Indigenous of America.

Another gene functionally impacted by the rs3826295 variant is the *TRIM25* (Tripartite Motif Containing 25) gene, which is involved in innate immune defense against viruses, including *SARS-CoV-2*, by mediating ubiquitination of RIG-I and subsequent type-I interferon production[38]. This gene has a direct interaction with *SARS-CoV-2* ORF6 protein, causing its degradation, and as a result Tavakoli et al. (2024) observed that when hypoexpressed, this gene is associated with increased disease severity of COVID-19 in individuals and when overexpressed in vitro the *TRIM25* gene partly counteracted viral inhibition, benefiting beta-interferon induction[39,40].

The *CFAP46* (cilia and flagella associated protein 46) gene is part of the central apparatus of the microtubule-based cytoskeleton of the cilium and is expressed in tissues that contain cilia, such as the testis, lung, and brain[41,42]. Thus, cilia in the lung are involved in lung repair processes and in regulating the production of cytokines and antimicrobials, as well as possibly having a functional role in innate immunity in the airways against bacterial infections by boosting innate immune defenses in response to bacterial antigens[43,44], which can be associated with the rs7913626 role in our study population.

Furthermore, the protective nature of these variants, whose allelic frequencies are higher in our region compared to other continental populations, may explain the relatively less severe impact of COVID-19 observed in our Indigenous population, even with the inadequate public policies implemented by the government.

In the wake of the emergence of the COVID-19 pandemic in 2020, initial predictions suggested a potentially devastating impact on Brazilian indigenous communities. This anticipation stemmed from the understanding that indigenous populations are more susceptible to various infectious diseases and pathogens because of their reduced genetic diversity at the major histocompatibility complex (MHC) locus[45–47].

Contrary to the initial predictions, data from the Ministry of Health and the Special Secretariat for Indigenous Health (SESAI) revealed that most COVID-19 cases of the people living in Indigenous territories in the state of Pará, in Northern Brazil, were asymptomatic or mild.

Moreover, even before the start of the vaccination campaign, a high prevalence of IgG anti-*SARS-CoV-2*antibodies was observed among indigenous groups, suggesting that this population may have reached a state of collective immunity in a relatively short period of time after the virus was introduced into the region[48,49].

In addition to the allele frequencies, another observation that could explain the impact of COVID-19 on the populations of the Northern Region is the PRS calculated for each individual in the study populations (sCOV, nsCOV, and NAT) and the continental ones (AFR, AMR, EUR, EAS, and SAS).

Our results reveal that the PRS of the non-severe COVID-19 patients and the Indigenous of America are significantly similar, which indicates that the four associated variants observed in the study, when combined, act in a similar way among these populations, probably protecting them from the COVID-19 severity. Besides, a significant difference was observed between the severe and non-severe COVID-19 patients' groups (Tukey, padj = 0.002), corroborating our hypothesis.

Furthermore, the genes associated with COVID-19 best known in the world literature as *SARS-CoV-2* receptor genes *SLC6A20*, *LZTFL1*, and *FYCO1* and the chemokine receptors *CCR9* and *CXCR6*[50,51] and in the national literature *HLA-A*, *HLA-DRB1*, *XCR1* and *FURIN*[6,8] did not show a significant association in our sample, further reinforcing the idea that the Brazilian genetic architecture as a whole differs from other world populations and the genetic landscape of the Amazon has a strong substructure of indigenous peoples, which makes it different from other regions of Brazil.

Limitations

We encountered some limitations in our study, such as the sample size and the low coverage and depth obtained by the WES sequencing, resulting in a large amount of missing data. Due to this, we had to perform with more restricted quality parameters, in order to decrease the risk of false positives, and adopted a less stringent p-value in the genome-wide association analyses. This study also lacks a replication cohort to validate our identified suggestive associations, therefore, the next steps in our exploratory study are to increase the sample size and validate our findings in populations similar to ours in order to make future comparisons with results obtained by other studies and increase knowledge about the genomic architecture of individuals living in the Brazilian Amazon.

Conclusion

To characterize genomic markers linked to *SARS-CoV-2* infection and its clinical severity, we performed whole exome sequencing of individuals affected by COVID-19 spanning both severe and non-severe forms of the disease. Based on this, we identified four variants (rs1770731 in *CRYBG1*, rs7221209 in *DNAH17*, rs3826295 in *DGKE* and rs7913626 in *CFAP46*) potentially linked to a protective effect against the clinical severity of COVID-19 – a novel discovery not previously explored in the literature. These loci are hypothesized to represent specific markers within our Northern Brazilian population.

While the strides made during the four years of the COVID-19 pandemic are commendable, it is crucial to acknowledge the complexity and challenges inherent in unraveling the biological mechanisms involved in illness. Collaboration within the scientific community is imperative to bridge the gap between cosmopolitan populations and neglected communities.

To our knowledge, this is the first whole exome sequencing (WES) study of admixed individuals from the Brazilian Amazon, representing a pioneer investigation into genomic variants associated with the clinical severity of COVID-19. Our results reinforce the importance of further genomic studies in populations living in the Amazon region, renowned for their extraordinary genetic diversity and the presence of rare, specific alleles.

Methods

Design and samples

This is a population-based retrospective observational study using genomic and COVID-19 surveillance data collected from the state of Pará in North Brazil. We sequenced 124 naso-oropharyngeal swab and saliva samples, being 56 individuals with asymptomatic or mild cases (nsCOV) and 68 individuals with the severe form of COVID-19 (sCOV), from January 2021 to July 2022. The selection criteria for the samples in the severe group were in accordance with the World Health Organization (WHO), which classifies severe symptoms as difficult breathing (dyspnea), loss of speech or mobility, and chest pain; in addition to these, we included individuals who had oxygen saturation below 95% and hospitalized individuals (hospitalized in a normal bed and in the ICU). Individuals who did not meet the above criteria were classified as non-severe, whose main symptoms were fever, cough and sore throat. Swab samples were collected and transferred to viral transport media and saliva samples were collected in sterile plastic collection tubes. After collection, both sample types were stored at -80 °C until further analysis. Figure 5 describes the methodology workflow used in this study, from sample collection to the association tests carried out in the R software.

This study was approved by the Barros Barreto Hospital Research Ethics Committee (CAAE: 50865721.1.0000.0017). The patients/participants provided their written informed consent to participate in this study and all methods were performed in accordance with the relevant guidelines and regulations.

DNA isolation and WES

The biological material was isolated from naso-oropharyngeal swab and saliva samples using MagMAX Viral/Pathogen Nucleic Acid Isolation Kit (Thermo Fisher Scientific) at KingFisher System (Thermo Fisher Scientific). DNA purity was assessed by spectrophotometry (Nanodrop 1000—Thermo Fisher Scientific, Waltham, MA) and concentration was assessed by fluorometry (Qubit—Life Technologies, Foster City, CA, USA). Whole-exome sequencing (WES) libraries were constructed across 150 bp paired reads using Illumina DNA Prep with Exome v2 (Illumina) and checked for quality using 2200 TapeStation (Agilent Technologies), following manufacturer's recommendations. The libraries were sequenced on NextSeq 500 Sequencing System (Illumina) using NextSeq 500/550 High Output Kit (300 cycles - Illumina).

$$PRS_p = \sum e_i \cdot OR_i$$

Fig. 6. Formula used to calculate the population polygenic risk score. PRS: Polygenic Risk Score; p: population; e: effect allele; OR: OddsRatio; i: individual.

Read processing and variant calling

Initially, the quality and coverage of the generated reads were analyzed using the FASTQC tool. Low-quality sequences were removed using Trimmomatic v0.36[52], which was employed to filter and trim the sequences. Adapter sequences, reads shorter than 50 bp, and low-quality reads based on PHRED score (average $Q < 20$) were identified and removed. Only reads where both pairs met the quality criteria were retained for further analysis.

The filtered reads were then aligned to the GRCh38 reference genome using the BWA-MEM v0.7.12 aligner[53,54], generating SAM files. These files were sorted by genomic position using SAMtools v1.7[55] and then marked for duplicates using Picard-tools MarkDuplicates v2.27.3 (broadinstitute.github.io/picard/). The resulting SAM files were converted into BAM files using SAMtools and processed using the RealignerTargetCreator from GATK to create a target interval file for IndelRealigner (GATK), which directs the local realignment of reads. This realignment was performed to correct misalignments caused by indels.

The next step involved using BaseRecalibrator (GATK) to identify systematic errors in base quality scores exported by the sequencer, calculating a recalibration model to appropriately adjust these scores. After processing, the BAM files were input into DeepVariant for variant calling, resulting in the generation of VCF files.

We used DeepVariant v1.15.0 with the publicly available Whole Exome Sequencing (WES) model to perform single-sample variant calling. A single-line command to run DeepVariant on each sample, using a pre-built Docker container, is available in the public DeepVariant repository (https://github.com/google/deepvariant). The variant calls made by DeepVariant followed standard single-sample variant calling methods[56], using sequence reads aligned to a reference genome to identify and genotype positions that differ from the reference. Throughout the study, we used the human GRCh38 reference genome without ALT contigs.

The variants in the VCF files were annotated using SnpEff v5.1 and dbNSFP v4.2 to assess the impact and predict the function of the variants. Additionally, we used ClinVar[57]to identify associations with known diseases and determine the clinical relevance of the variants, while ExAC and gnomAD[58,59] were used to provide allele frequency annotations.

Damaging variants were classified according to the following computer prediction algorithms: Functional Analysis through Hidden Markov Models (FATHMM)[60], MutationAssessor[61], MutationTaster[62], Sort Intolerant from tolerant (SIFT)[63], MetaSVM[64] and PROVEAN[65]. The ClinVar database was also used for this classification[57]. A report of WES statistics is available in Supplementary File.

Data merging

Due to our higher missing data rate, we performed the following vcftools filters: --missing-site and --missing-indv lower than 10%. After quality control of reads, we merged our population data of 124 newly sequenced samples with previously published population dataset by 1,000 Genomes and HGDP projects, summing up to 4,281 individuals. Quality control of these data was performed using BCFTools v1.4, VCFTools v0.1.13 e PLINK v1.90b6.15 softwares with the variant filtering conditions -maf: 0.05, -hwe: 0.0001, -mind: 0.1, and -geno: 0.01. Variants and individuals with missing data above 0.1 were excluded. Therefore, in addition to quality control of variants, we chose the closest individuals that compose the Brazilian admixture, due to its influence on Brazilian colonization (Iberian – 1kG, Yoruba – 1kG and Native Americans - HGDP), obtaining a dataset containing 527 individuals with 2,172 SNPs to perform the following population genetic analysis.

Population structure

To further explore the ancestry composition and genetic similarity of our studied group, we carried out model-based clustering analysis using ADMIXTURE 1.2 and Eigenstrat v8.0 softwares. For Admixture analysis, we did an unsupervised ADMIXTURE approach (Supplementary Fig. 1), in which allele frequencies for unadmixed ancestral populations are unknown and are computed during the analysis with the populations from datasets 1,000 Genomes (1kG), HGDP and our samples (patients with COVID-19), varying the number of ancestral populations between $K = 2$ and $K = 16$; and a supervised approach with tri-hybrid model with the populations: Iberian (IBS – 1kG), Yoruba (YRI/AFR – 1kG) and Indigenous of America (AMI - HGDP).

The identity-by-descent (IBD) analysis and the calculation of the genetic relationship matrix were obtained by the PC-relate method in the *GENESIS* v2.26.0 R package. From the genetic relationship matrix (GRM), corrected principal component analysis (PCA) was performed for COVID patients.

Variant associations analyses

The *GENESIS* v2.26.0 package was used to calculate kinship coefficients, inbreeding coefficients and IBD sharing probabilities. From the kinship analysis, three individuals of a pair with kinship values higher than 0.004 were excluded. To predict covariates associated with disease severity, a generalized linear model (GLM) involving sex, age, Indigenous of America, African and European ancestries was performed. To mitigate potential errors, we employed False Discovery Rate (FDR) correction.

A power calculation was performed considering a sample size of 124 participants, a significance level of 0.05, and an effect size of 0.5, resulting in a power of 0.80. Given our sample size, we were only able to identify variants with high impact and were unable to find rare variants with smaller effects or other variants previously reported in the literature as associated with COVID-19 severity. The limited sample size may have constrained our ability to detect more subtle genetic effects.

Exome-wide association analysis, both individual and aggregate by genes, was performed using GLM in the *GENESIS v2.26.0* package, where the first five PCAs, age and the comorbidities hypertension and overweight were added as covariates, and p-values were corrected with Bonferroni to minimize false positive results. All variants with their respective p-values are available in Supplementary Reports.

Polygenic risk score (PRS)

PRS of each participant was calculated by computing the sum of risk alleles of the associated variants weighted by the risk allele effect sizes and PRS for a population was calculated by taking the median PRS of all the individuals in that population as shown in the formula below, using R v.4.2.1. Population wise statistical significance was calculated using one-way ANOVA, with a post-hoc Tukey multiple comparisons means test. We decided to focus on population PRS rather than individual PRS due to our statistical and sampling limitations.

Statistical analyses

Clinical characteristics were analyzed using Fisher's or Chi-squared test for categorical variables, and Wilcoxon test for continuous variables (age). All graphs and statistical analyses were made using R (v.4.2.1). P-values ≤ 0.05 were considered to be statistically significant.

Data availability

The WES datasets have been deposited with links to project PRJEB75518, in ENA database from EBI (https://www.ebi.ac.uk/ena/browser/view/PRJEB75518). For better replication, clinical data is available in Supplementary Reports.

Received: 17 June 2024; Accepted: 29 October 2024
Published online: 09 November 2024

References

1. Wang, P. et al. Risk factors for severe COVID-19 in middle-aged patients without comorbidities: A multicentre retrospective study. *J. Transl Med.* **18**(1), 461. https://doi.org/10.1186/s12967-020-02655-8 (2020).
2. Ochani, R. K. et al. COVID-19 pandemic: from origins to outcomes. A comprehensive review of viral pathogenesis, clinical manifestations, diagnostic evaluation. *Manage.* **17**. (2021).
3. O'Driscoll, M. et al. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**(7844), 140–145. https://doi.org/10.1038/s41586-020-2918-0 (2021).
4. Brodin, P. Immune determinants of COVID-19 disease presentation and severity. *Nat. Med.* **27**(1), 28–33. https://doi.org/10.1038/s41591-020-01202-8 (2021).
5. Novelli, G. et al. COVID-19 one year into the pandemic: From genetics and genomics to therapy, vaccination, and policy. *Hum. Genomics* **15**(1), 27. https://doi.org/10.1186/s40246-021-00326-3 (2021).
6. Castelli, E. C. et al. MHC variants associated with symptomatic versus asymptomatic SARS-CoV-2 infection in highly exposed individuals. *Front. Immunol.* **12**, 742881. https://doi.org/10.3389/fimmu.2021.742881 (2021).
7. Castelli, E. C. et al. MUC22, HLA-A, and HLA-DOB variants and COVID-19 in resilient super-agers from Brazil. *Front. Immunol.* **13**, 975918. https://doi.org/10.3389/fimmu.2022.975918 (2022).
8. Secolin, R. et al. Genetic variability in COVID-19-related genes in the Brazilian population. *Hum. Genome Var.* **8**(1), 15. https://doi.org/10.1038/s41439-021-00146-w (2021).
9. Santos-Rebouças, C. B. et al. Host genetic susceptibility underlying SARS-CoV-2-associated multisystem inflammatory syndrome in Brazilian children. *Mol. Med.* **28**(1), 153. https://doi.org/10.1186/s10020-022-00583-5 (2022).
10. Rodrigues-Soares, F. et al. Genetic structure of pharmacogenetic biomarkers in Brazil inferred from a systematic review and population-based cohorts: A RIBEF/EPIGEN-Brazil initiative. *Pharmacogenomics J.* **18**(6), 749–759. https://doi.org/10.1038/s41397-018-0015-7 (2018).
11. Rodrigues, J. C. G. et al. da S,. Polymorphisms of ADME-related genes and their implications for drug safety and efficacy in Amazonian Amerindians. Sci Rep. **9**(1), 7201. (2019). https://doi.org/10.1038/s41598-019-43610-y
12. Secolin, R. et al. Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci. Rep.* **9**(1), 13900. https://doi.org/10.1038/s41598-019-50362-2 (2019).
13. Mills, M. C. & Rahal, C. The GWAS diversity monitor tracks diversity by disease in real time. *Nat. Genet.* **52**(3), 242–243 (2020).
14. Ribeiro-dos-Santos, A. M. et al. Exome sequencing of native populations from the Amazon reveals patterns on the Peopling of South America. *Front. Genet.* **11**, 548507. https://doi.org/10.3389/fgene.2020.548507 (2020).
15. Schaan, A. P. et al. New insights on intercontinental origins of paternal lineages in Northeast Brazil. *BMC Evol. Biol.* **20**(1), 15. https://doi.org/10.1186/s12862-020-1579-9 (2020).
16. Ribeiro-dos-Santos, A. M. et al. High-throughput sequencing of a South American Amerindian. Calafell F, editor. PLoS ONE. **8**(12), e83340. (2013). https://doi.org/10.1371/journal.pone.0083340
17. QueirozMAF et al. Cytokine profiles associated with acute COVID-19 and Long COVID-19 syndrome. *Front. Cell. Infect. Microbiol.* **12**, 922422. https://doi.org/10.3389/fcimb.2022.922422 (2022).
18. Angulo-Aguado, M. et al. Association between the LZTFL1 rs11385942 polymorphism and COVID-19 severity in Colombian Population. *Front. Med.* **9**, 910098. https://doi.org/10.3389/fmed.2022.910098 (2022).
19. Paz Ocaranza, M. et al. Counter-regulatory renin–angiotensin system in cardiovascular disease. *Nat. Rev. Cardiol.* **17**(2), 116–129. https://doi.org/10.1038/s41569-019-0244-8 (2020).
20. Ruiz-Sternberg, Á. M. et al. Genomic characterization of SARS-CoV-2 and its association with clinical outcomes: A 1-year longitudinal study of the pandemic in Colombia. *Int. J. Infect. Dis.* **116**, 91–100. https://doi.org/10.1016/j.ijid.2021.12.326 (2022).
21. De Leeuw, A. J. M., Oude Luttikhuis, M. A. M., Wellen, A. C., Müller, C. & Calkhoven, C. F. Obesity and its impact on COVID-19. *J. Mol. Med.* **99**(7), 899–915. https://doi.org/10.1007/s00109-021-02072-4 (2021).
22. Gasmi, A. et al. Interrelations between COVID-19 and other disorders. *Clin. Immunol.* **224**, 108651. https://doi.org/10.1016/j.clim.2020.108651 (2021).
23. Mathur, R. et al. Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: An observational cohort study using the OpenSAFELY platform. *Lancet* **397**(10286), 1711–1724. https://doi.org/10.1016/S0140-6736(21)00634-6 (2021).
24. Shelton, J. F. et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nat. Genet.* **53**(6), 801–808. https://doi.org/10.1038/s41588-021-00854-7 (2021).
25. Pommerenke, C. et al. Molecular characterization and subtyping of breast cancer cell lines provide novel insights into cancer relevant genes. *Cells* **13**(4), 301. https://doi.org/10.3390/cells13040301 (2024).
26. Haffner, M. C. et al. AIM1 is an actin-binding protein that suppresses cell migration and micrometastatic dissemination. *Nat. Commun.* **8**(1), 142. https://doi.org/10.1038/s41467-017-00084-8 (2017).
27. Sargiacomo, C., Sotgia, F. & Lisanti, M. P. COVID-19 and chronological aging: senolytics and other anti-aging drugs for the treatment or prevention of corona virus infection? *Aging (Albany NY).* **12**(8), 6511–6517. https://doi.org/10.18632/aging.103001 (2020).

28. Ying, K. et al. Genetic and phenotypic analysis of the causal relationship between aging and COVID-19. *Commun. Med.* **1**(1), 35. https://doi.org/10.1038/s43856-021-00033-z (2021).
29. Whitfield, M. et al. Mutations in DNAH17, encoding a sperm-specific axonemal outer dynein arm heavy chain, cause isolated male infertility due to Asthenozoospermia. *Am. J. Hum. Genet.* **105**(1), 198–212. https://doi.org/10.1016/j.ajhg.2019.04.015 (2019).
30. Fan, X. et al. The association between methylation patterns of DNAH17 and clinicopathological factors in hepatocellular carcinoma. *Cancer Med.* **8**(1), 337–350. https://doi.org/10.1002/cam4.1930 (2019).
31. deAndrés-Galiana, E. J. et al. Analysis of transcriptomic responses to SARS-CoV-2 reveals plausible defective pathways responsible for increased susceptibility to infection and complications and helps to develop fast-track repositioning of drugs against COVID-19. *Comput. Biol. Med.* **149**, 106029. https://doi.org/10.1016/j.compbiomed.2022.106029 (2022).
32. Andrawus, M. et al. Genome integrity as a potential index of longevity in Ashkenazi Centenarian's families. *Geroscience*. **46**(5), 5391–5392. https://doi.org/10.1007/s11357-024-01253-6 (2024).
33. Bezdíčka, M., Pavlíček, P., Bláhová, K., Háček, J. & Zieg, J. Various phenotypes of disease associated with mutated DGKE gene. *Eur. J. Med. Genet.* **63**(8), 103953. https://doi.org/10.1016/j.ejmg.2020.103953 (2020).
34. Raina, R. et al. *Pediatr. Atyp. Hemolytic Uremic Syndrome Adv. Cells* ;**10**(12), 3580. https://doi.org/10.3390/cells10123580 (2021).
35. Kamenarova, K. et al. Rare host variants in ciliary expressed genes contribute to COVID-19 severity in Bulgarian patients. *Sci Rep.* **14**(1), 19487. https://doi.org/10.1038/s41598-024-70514-3 (2024).
36. Conway, E. M. & Pryzdial, E. L. G. Is the COVID-19 thrombotic catastrophe complement-connected? *J. Thromb. Haemost.* **18**(11), 2812–2822. https://doi.org/10.1111/jth.15050 (2020).
37. Zhu, J. et al. Loss of diacylglycerol kinase epsilon in mice causes endothelial distress and impairs glomerular Cox-2 and PGE 2 production. *Am. J. Physiology-Renal Physiol.* **310**(9), F895–908. https://doi.org/10.1152/ajprenal.00431.2015 (2016).
38. Hu, Y. et al. The severe acute respiratory syndrome coronavirus nucleocapsid inhibits type I interferon production by interfering with TRIM25-mediated RIG-I ubiquitination. *J. Virol.* **94**(20), e01378-20. https://doi.org/10.1128/JVI.01378-20 (2020).
39. Khatun, O., Sharma, M., Narayan, R. & Tripathi, S. SARS-CoV-2 ORF6 protein targets TRIM25 for proteasomal degradation to diminish K63-linked RIG-I ubiquitination and type-I interferon induction. *Cell. Mol. Life Sci.* **80**(12), 364. https://doi.org/10.1007/s00018-023-05011-3 (2023).
40. Tavakoli, R. et al. Exploring the impression of TRIM25 gene expression on COVID-19 severity and SARS-CoV-2 viral replication. *J. Infect. Public Health.* **17**(8), 102489. https://doi.org/10.1016/j.jiph.2024.102489.
41. Ziegler, C. et al. The DNA methylome in panic disorder: a case-control and longitudinal psychotherapy-epigenetic study. *Transl Psychiatry* **9**(1), 314. https://doi.org/10.1038/s41398-019-0648-6 (2019).
42. Cassuto, N. G. et al. Molecular Profiling of Spermatozoa Reveals Correlations between Morphology and Gene Expression: A Novel Biomarker Panel for Male Infertility. Lin YH, editor. BioMed Research International. **2021**1–14. doi: (2021). https://doi.org/10.1155/2021/1434546
43. McFie, M. et al. Ciliary proteins specify the cell inflammatory response by tuning NFκB signaling, independently of primary cilia. *J. Cell Sci.* jcs.239871 (2020).
44. Kuek, L. E. & Lee, R. J. First contact: The role of respiratory cilia in host-pathogen interactions in the airways. *Am. J. Physiology-Lung Cell. Mol. Physiol.* **319**(4), L603–L619. https://doi.org/10.1152/ajplung.00283.2020 (2020).
45. Fellows, M. et al. Under-reporting of COVID-19 cases among indigenous peoples in Brazil: A new expression of old inequalities. *Front. Psychiatry* **12**, 638359. https://doi.org/10.3389/fpsyt.2021.638359 (2021).
46. Mendes, M. F. et al. COVID-19 pandemic evolution in the Brazilian indigenous population. *J. Racial Ethnic Health Disparities* **9**(3), 921–937. https://doi.org/10.1007/s40615-021-01031-6 (2022).
47. Putira Sacuena, E. R. et al. Host genetics and the profile of COVID-19 in indigenous people from the Brazilian Amazon: A pilot study with variants of the ACE1, ACE2 and TMPRSS2 genes. *Infect. Genet. Evol.* **118**, 105564. https://doi.org/10.1016/j.meegid.2024.105564 (2024).
48. Lima, C. N. C. et al. Anti-SARS-CoV-2 antibodies among indigenous populations of the Brazilian Amazon: A cross-sectional study. *BMJ Open.* **12**(2), e054271. https://doi.org/10.1136/bmjopen-2021-054271 (2022).
49. Rodrigues, E. P. S. et al. High prevalence of anti-SARS-CoV-2 IgG antibody in the xikrin of Bacajá (Kayapó) indigenous population in the Brazilian Amazon. *Int. J. Equity Health.* **20**(1), 50. https://doi.org/10.1186/s12939-021-01392-8 (2021).
50. Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from neanderthals. *Nature* **587**(7835), 610–612. https://doi.org/10.1038/s41586-020-2818-3 (2020).
51. Downes, D. J. et al. Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. *Nat. Genet.* **53**(11), 1606–1615. https://doi.org/10.1038/s41588-021-00955-3 (2021). Epub 2021 Nov 4.
52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics.* **30**(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 (2014).
53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 (2009).
54. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv; [cited 2022 Sep 1]. (2013). http://arxiv.org/abs/1303.3997
55. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
56. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**(10), 983–987. https://doi.org/10.1038/nbt.4235 (2018).
57. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153 (2018).
58. Exome Aggregation Consortium et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**(7616), 285–291. https://doi.org/10.1038/nature19057 (2016).
59. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7 (2020).
60. Rogers, M. F. et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics.* **34**(3), 511–513. https://doi.org/10.1093/bioinformatics/btx536 (2018).
61. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic. Acids. Res.* **39**(17), e118. https://doi.org/10.1093/nar/gkr407 (2011).
62. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. Mutationtaster evaluates disease-causing potential of sequence alterations. *Nat. Methods.* **7**(8), 575–576. https://doi.org/10.1038/nmeth0810-575 (2010).
63. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic. Acids Res.* **31**(13), 3812–3814. https://doi.org/10.1093/nar/gkg509 (2003).
64. Dong, C. et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**(8), 2125–2137. https://doi.org/10.1093/hmg/ddu733 (2015).
65. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One.* **7**(10), e46688. https://doi.org/10.1371/journal.pone.0046688 (2012).

Acknowledgements

The authors want to thank all participants involved in the study for their participation. We are also grateful to the Federal University of Pará and Federal University of Rio Grande do Norte for their technical assistance and generous support.

Author contributions

MCB, GBSS, JESdS and ÂR-d-S: conceived and designed the study. MCB, GBSS, JESdS, DHFG: bioinformatic analysis. CTP, MCB, CSS, CB-d-S, GCC, LM, JAP: sample processing and sequencing. MCB and GBSS: statistical analysis. JASQ, LFMF, PFC, CGS, TXC, RRB, JRdSV and ÂR-d-S: sample acquisition. SS and SJS: funding. ÂR-d-S: study conception and supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded by Fundação de Amparo a Estudos e Pesquisas no Estado do Pará (FAPESPA) to support the COVID-19 Genomic Surveillance Network based on New Strategies for Diagnosis, Prevention, and Prognosis (Rede de Vigilância Genômica da COVID-19 baseada em Novas Estratégias de Diagnóstico, Prevenção e Prognóstico). ÂR-d-S is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Brazil) with Produtividade (304413/2015–1), also by Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Pró-Reitoria de Pesquisa e Pós-Graduação da Universidade Federal do Pará (PROPESP/UFPA).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Human and animal rights

The studies involving human participants were reviewed and approved by the Barros Barreto Hospital Research Ethics Committee (CAAE: 50865721.1.0000.0017). The patients/participants provided their written informed consent to participate in this study.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-78170-3.

Correspondence and requests for materials should be addressed to G.B.S.-S., S.J.S. or Â.R.-d.-S.

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.