Genetic variability in COVID-19-related genes in the Brazilian population

Rodrigo Secolin[1†], Tânia K. de Araujo[1†], Marina C. Gonsales[1], Cristiane S. Rocha[1], Michel Naslavsky[2], Luiz De Marco[3], Maria A.C. Bicalho[4], Vinicius L. Vazquez[5], Mayana Zatz[2], Wilson A. Silva[6], Iscia Lopes-Cendes[1*]

[1]Department of Translational Medicine, University of Campinas (UNICAMP), and the Brazilian Institute of Neuroscience and Neurotechnology (BRAINN), Campinas, SP, Brazil.

[2]Departament of Genetics and Evolutive Biology, Institute of Bioscience, University of São Paulo, (USP) and the Human Genome and Stem Cell Research Center, São Paulo, SP, Brazil.

[3]Department of Surgery, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG, Brazil.

[4]Department of Clinical Medicine, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG, Brazil.

[5] Molecular Oncology Research Center (CPOM), Barretos Cancer Hospital, Barretos, SP, Brazil.

[6]Department of Genetics, Ribeirão Preto Medical School, University of São Paulo at Ribeirão Preto (USP), Ribeirão Preto, SP, Brazil.

†These authors contribute equally to this work

***Correspondence to:** Iscia Lopes-Cendes, M.D., Ph.D.
Professor of Medical Genetics and Genomic Medicine
Department of Translational Medicine
School of Medical Sciences
University of Campinas (UNICAMP)
Tessália Vieira de Camargo, 126.
Campinas, SP, Brazil 13083-887
e-mail: icendes@unicamp.br

ABSTRACT

SARS-CoV-2 employs the angiotensin-converting enzyme 2 (ACE2) receptor and the transmembrane serine protease (TMPRSS2) to infect human lung cells. Previous studies have suggested that different host genetic backgrounds in *ACE2* and *TMPRSS2* could contribute to differences in the rate of infection or severity of COVID-19. Recent studies also showed that variants in 15 genes related to type I interferon immunity to influenza virus could predispose to life-threatening COVID-19 pneumonia. Additional genes (*SLC6A20, LZTFL1, CCR9, FYCO1, CXCR6, XCR1, IL6, CTSL*, *ABO*, and *FURIN*) and HLA alleles have also been implicated in response to infection with SARS-CoV-2. Currently, Brazil has recorded the third-highest number of COVID-19 patients worldwide. We aim to investigate the genetic variation present in COVID-19-related genes in the Brazilian population. We analysed 27 candidate genes and HLA alleles in 954 admixed Brazilian exomes. We used the information available in two public databases (http://www.bipmed.org and http://abraom.ib.usp.br/), and additional exomes from individuals born in southeast Brazil, the region with the highest number of COVID-19 patients in the country. Variant allele frequencies were compared with the 1000 Genomes Project phase 3 (1KGP) and the gnomAD databases. We found 395 non-synonymous variants; of these, 325 were also found in the 1000 Genome Project phase 3 (1KGP) and/or gnomAD. Six of these variants were previously reported as putatively influencing the rate of infection or clinical prognosis for COVID-19. The remaining 70 variants were identified exclusively in the Brazilian sample, with a mean allele frequency of 0.0025. *In silico* prediction of the impact in protein function revealed that three of these rare variants were pathogenic. Furthermore, we identified HLA alleles that were previously associated with

COVID-19 response at loci DQB1 and DRB1. Our results showed genetic variability common to other populations, but also rare and ultra-rare variants exclusively found in the Brazilian population. These findings could potentially lead to differences in the rate of infection or response to infection by SARS-CoV-2 and should be further investigated in patients with the disease.

Keywords: population genomics, admixed population, COVID-19, SARS-Cov-2

Introduction

COVID-19 disease, caused by the SARS-CoV-2 coronavirus, is currently a worldwide pandemic. To enter human lung cells, SARS-CoV-2 employs the spike protein, which is primed by the host serine protease (TMPRSS2), followed by angiotensin-converting enzyme 2 (ACE2) receptor binding, and proteolysis with activation of membrane fusion within endosomes by cathepsin L (CTSL)[1-4]. The main feature in SARS-CoV-2 infection is pre-activation of the spike protein by FURIN inside the host cell, which leads to increased SARS-CoV-2 spread into lung cells and increased virulence[5]. The rapid SARS-CoV-2 infection leads to an exacerbated immune reaction, and a few studies have shown that increased levels of IL-6 (an essential immune response mediator) are associated with increased inflammatory response, respiratory failure, increased probability of intubation, the presence of clinical complications, and higher mortality in patients with COVID-19[6-8]. Additional studies found the enrichment of rare variants predicted to be loss-of-function in genes related to type I interferon (IFN) immunity to influenza virus among patients with life-threatening COVID-19 pneumonia (*TLR3*, *TICAM1*, *TRIF*, *UNC93B1*, *TRAF3*, *TBK1*, *IRF3*, *NEMO*, *IKBKG*, *IFNAR1*, *IFNAR2*, *STAT1*, *STAT2*, *IRF7*, and *IRF9*)[9].

Specific variants in the *ACE2* and *TMPRSS2* genes have been reported among diverse populations worldwide, suggesting that different host genetic backgrounds could contribute to differences in COVID-19 infection and severity[2,10]. Ellinghaus et al.[11] performed a genome-wide association study (GWAS) including Italian and Spanish patients with confirmed COVID-19 and controls and identified six candidate genes associated with COVID-19 response on chromosome (chr) 3p21.31 (*SLC6A20*, *LZTFL1*, *FYCO1*, *CXCR6*,

XCR1, *CCR9*), and one on chr 9q34.2, the locus for genes encoding for the *ABO* blood group antigens. A subsequent, more extensive study replicated the association between the locus on chr 3p21.31 and COVID-19. It revealed a COVID-19 risk core haplotype ranging from 45,859,651bp to 45,909,024bp, which was inherited from Neanderthals and is currently carried by approximately 50% of people in South Asia and about 16% of people in Europe[12]. Interestingly, no evidence of association was found for the previously identified candidate genes that are potentially involved in the response to infection by SARS-CoV-2: *ACE2*, *TMPRSS2*, *FURIN*, and *IL6*.

Furthermore, one significant factor modulating resistance or susceptibility to viral infections is the human leukocyte antigens (HLAs). HLA polymorphism results from a set of amino-acid substitutions in the peptide-binding groove of the HLA molecules that produce variability in the peptide epitope binding-site and presentation to T cells, which may protect against epidemic infection[13]. Thus, genetic variability in the HLA alleles could influence the immune response in patients with COVID-19, modulating disease severity. Indeed, *in silico* analysis found that HLA-B*46:01 had the fewest predicted binding sites for SARS-CoV-2 peptides, and HLA-B*15:03 showed the greatest capacity to present highly conserved shared SARS-CoV-2 peptides to immune cells[14].

Brazil has reported the third-highest number of COVID-19 infections worldwide (updated on September 28th 2020; https://covid19.who.int/; https://coronavirus.jhu.edu/map.html), and the highest number of cases is concentrated in the south-eastern region of the country (updated on September 28th 2020; https://covid.saude.gov.br/). Brazilian individuals feature an admixed genome, encompassing European, sub-Saharan African, and Native Americans as the three main ancestry populations[15-17], and the distribution of ancestry components varies remarkably

throughout the genome[18]. Furthermore, it has been demonstrated that a significant proportion of genetic variability is still undiscovered in admixed Brazilians[19] and that genetic variability may lead to differential response to infection[20]. Therefore, we aim to investigate the genetic variation present in COVID-19 related genes in the Brazilian population.

Results

Exome analysis

We found 7,172 variants among the candidate genes analysed (Supplementary Table 1). Of these, 395 variants putatively impact protein function, including 354 non-synonymous variants, seven frameshift substitutions, three in-frame deletions, one in-frame insertion, 12 stop gains, two start losses, and 16 splice site variants (Supplementary Table 1). Three hundred and twenty-five variants were also present in the gnomAD and/or 1000 Genome Project phase 3 (1KGP) databases, including 56 common variants, with an alternative allele frequency (AAF) $\geq$0.01 and 269 rare variants (AAF<0.01) (Supplementary Data 1). Although AAF from the admixed Brazilian sample follows the distribution from NFE/EUR, AFR, and AMR in gnomAD and 1KGP databases (Fig. 1), we found differences in the AAF of these common and rare variants in the admixed Brazilian sample compared to gnomAD[21] and/or 1KGP[22] databases shown in Fig. 1 and Supplementary Data 1. Interestingly, we also observed some variability in the AAF among samples from different Brazilian towns and the two public databases of genomic information on the Brazilian population, BIPMed and ABraOM (Fig. 2).

More importantly, there were 70 variants which were exclusive to the Brazilian sample, including 11 variants in genes related to type I INF immunity to influenza virus[9], six in candidate genes for COVID response identified by GWAS[11], and five related to SARS-CoV-2 entry in lung cells and virus replication[2,10]. These are rare or ultra-rare variants, presenting a mean AF of 0.0025 (Supplementary Data 1). Among these, we found one in the dataset from Belo Horizonte and two in the ABraOM database for *ACE2* p.Arg219Cys; one in the dataset from Barretos and two in the ABraOM database for *ACE2* p.Leu731Phe; and the *TMPRSS2* p.Val160Met variant was present in samples from all the different Brazilian towns and the two public databases (BIPMed and ABraOM), with an AAF ranging from 0.1333 in Belo Horizonte to 0.2931 in Campinas. Among the reported variants in genes influencing type I INF immunity to influenza virus[9], we found three variants in the ABraOM database (one *TLR3* p.Pro554Ser, one *IFR3* p.Asn146Lys and one *IRF7* p.Pro246Ser) (Supplementary Data1 and 2).

In addition, we identified five variants (rs35044562, rs34326463, rs35508621, rs67959919, and rs35624553) which were previously described in the COVID-19 risk core haplotype and inherited from Neanderthals[12]. These were only present in samples from Ribeirão Preto and the BIPMed dataset (rs34326463), Campinas (rs35044562, and rs35508621), and the ABraOM dataset (rs35044562, rs35508621, rs67959919, and rs35624553) (Table 1).

In silico predictions

We identified seven variants that were predicted to affect protein function for the 12 algorithms used: p.Phe249Ser, p.Gly164Val, and p.Leu25Pro in *the SLC6A20* gene;

p.Leu96Arg in *LZTFL1*; p.Tyr287Ser in *XCR1*; and p.Gly146Ser and p.Asn414Ser in the *FURIN* gene (Table 2). Furthermore, the variant p.Gly146Ser in the *FURIN* gene was predicted to destabilise the protein (ΔΔG: -1.576 kcal/mol). We observed that the p.Phe249Ser variant is present in samples from Barretos, the BIPMed dataset, gnomAD, and 1KGP (NFE/EUR, AFR, AMR, and SAS populations), whereas the p.Gly164Val variant is present in the ABraOM dataset, gnomAD, and 1KGP (NFE/EUR populations), and the p.Gly146Ser variant is present in the ABraOM dataset, gnomAD, and 1KGP (NFE/EUR, AFR, AMR, EAS, and SAS populations). Notably, four of the variants predicted to be deleterious are found exclusively in admixed Brazilian individuals (p.Leu25Pro in Barretos; p.Leu96Arg in the ABraOM dataset; p.Tyr287Ser in Belo Horizonte; and p.Asn414Ser in the BIPMed dataset).

We did not find any predicted deleterious variants in *ACE2* and *TMPRSS2* based on our 12 algorithm criteria. However, Hou et al.[2] applied only Polyphen2 and CADD scores to variants in *ACE2* and *TMPRSS2* (Polyphen2 >0.96 and CADD >20 as the cut-off). Therefore, only variants defined as 'probably damaging' by Polyphen2 (http://genetics.bwh.harvard.edu/pph2/dokuwiki/overview) and CADD (>20) were included. We found 79 variants predicted to affect protein function, including the p.Val160Met variant in *TMPRSS2* reported by Hou et al.[2], and the p.Pro554Ser variant in *TLR3* previously reported by Zhang et al.[9] (Supplementary Data 2).

HLA analysis

Overall, we identified 331 different HLA alleles in the admixed Brazilian samples. Of these, three HLA alleles have been previously associated with COVID-19 response[14,23].

We compared the frequency of these HLA alleles in admixed Brazilians and in populations that occupy the top 10 positions with most cases of COVID-19 and the five populations less affected by the disease, including the United States, India, Russia, Colombia, Peru, Mexico, Spain, Argentina, South Africa, Japan, Australia, South Korea, Vietnam, and Taiwan. The frequency of these alleles is described in Supplementary Data 3. We noticed that the HLA-B*46:01, HLA-B*27:07, HLA-B*15:27, and HLA-C*07:29 alleles were absent in the Brazilian samples. The HLA-C*07:29 allele was also absent from other populations and is present at a low frequency (AF = 0.0003) in the Indian population. HLA-B*15:27 was identified in Vietnam, Taiwan, Japan, with an AF >0.01 and Spain with AF <0.0001. HLA-B*27:07 was detected at a low frequency in India, Colombia, Spain, and South Africa. The HLA-B*46:01 allele was detected in Russia, Mexico, Vietnam, Taiwan, and Japan.

Sixty-six Brazilian individuals (17.1%) presented the HLA-DQB1*06:02 allele (AF=0.08938), 47 individuals (12.2%) carry the HLA-DRB1*15:01 allele (AF=0.06477) and 32 individuals (8.29%) have both the HLA-DRB1*15:01 and HLA-DQB1*06:02 alleles. The population of other continents, except Oceania, also have these two HLA alleles (HLA-DRB1*15:01 and HLA-DQB1*06:02) with an AF >0.01. Also, 15 Brazilian individuals (3.88%) carry HLA-B*15:03 (AF = 0.02073), which is predicted to have the greatest capacity to present SARS-CoV-2 peptides to immune cells[14]. This allele was not found in the Asian population of Japan, South Korea, and Vietnam (Supplementary Data 3).

In silico analysis of viral peptide-HLA class I and II binding affinity

To verify the potential for an HLA allele type to produce an antiviral response to SARS-CoV-2, we performed HLA binding affinity analyses to the SARS-CoV-2 proteome.

We tested 42 HLA-A proteins, 77 HLA-B, 38 HLA-C, 60 HLA-DP (DPA1/DPB1), 145 HLA-DQ (DQA1/DQB1), 46 HLA-DRB1, 4 DRB3, 2 DRB4, and 6 DRB5.

The SARS-CoV-2 proteome was presented by a diversity of HLA alleles from classes I and II (Supplementary Table 2). The HLA proteins are predicted to bind a small proportion of all possible SARS-CoV-2 derived peptides with high affinity (on average 0.5% for HLA class I and 2% for HLA class II). Also, we found a small proportion of weak binders (on average, 1.5% for HLA class I and 8.2% for class II). Most of the HLA proteins do not bind either Class I (on average >96%) or class II (on average >89%) molecules (Supplementary data 4). Supplementary Data 5 shows a list of HLA strongest binders (>300 peptides bound at high affinity) of SARS-CoV-2 peptides. These are found in loci HLA-A, -B, -C, and DQ.

Discussion

Accessing the genomic sequences of the general population is relevant to identify the genetic variability involved in the molecular mechanisms of infection[20]. Also, it is known that the admixed Brazilian population is underrepresented in large public databases[21,22], and previous studies revealed variants present exclusively in Brazilian indivuduals[19,24].

We studied 27 human COVID-19-related genes and the HLA region in two public genomic databases of admixed Brazilians (BIPMed, www.bipmed.org[19]; ABraOM http://abraom.ib.usp.br/[24]), and additional samples from individuals born in three different towns of south-eastern Brazil. We reported the variants and HLA alleles found in these samples and compared them with worldwide populations. We also reported variants

constituting the COVID-19 risk core haplotype on locus 3p21.31, described as being inherited from Neanderthals[12].

Previous studies showed that the *ACE2*, *TMPRSS2*, *CTSL*, *FURIN*, and *IL6* genes, as well as the HLA region, may be involved in SARS-CoV-2 infection[1-5,10] and immune response[6-8,14,23,25]. Furthermore, variants on loci 3p21.31 and 9q34.2 (encompassing *SLC6A20*, *LZTFL1*, *FYCO1*, *CXCR6*, *XCR1*, *CCR9*, and *ABO*) have been associated with Spanish and Italian patients with COVID-19[11], and different variants were found to affect the predisposition to life-threatening illness in patients with COVID-19 from different ancestries[9].

The analysis of genetic variability in candidate genes for specific populations can help to identify individuals at a higher risk of infection or severe disease by constructing risk haplotypes, which can also provide therapeutic targets for the development of more effective treatments and the control of COVID-19[2,10]. Thus, in addition to investigating genetic variability in the 27 candidate genes, we extended our analysis to include HLA alleles, which influence immunological response to many infectious agents (updated on September 28th 2020; https://covid19.who.int/; https://coronavirus.jhu.edu/map.html). This is the first comprehensive study of genetic variability of COVID-19 genes in admixed individuals from Latin America, a hard-stricken population in the COVID-19 pandemic[26], both in terms of the number of infected individuals and the severity of disease leading to increased death rates. Indeed, in the USA, remarkable disparities of SARS-CoV-2 infection by ethnicity have been shown, with Hispanic/Latino and African American individuals presenting higher SARS-CoV-2 infection rates and risk mortality than 'non-Hispanic white' Americans[27-29]. Therefore, by looking at population genomics data, one may gain insights into disease-related variants, which could be disproportionally represented in

specific populations[18,30-32]. Furthermore, by evaluating individuals with unknown information on SARS-CoV-2 infection, one can achieve the random distribution of these variants, allowing better estimates of the distribution of population allele frequencies.

We identified small differences in AF in the 395 candidate variants identified among Brazilian samples, strengthening the hypothesis that different genetic backgrounds could influence SARS-CoV-2 infection and behaviour in human host cells[2,10]. Furthermore, this study and previous works[2,10] identified individuals who carry unique deleterious variants, which may influence gene function and could potentially lead to different responses to SARS-CoV-2 infection on an individual scale. However, the rather similar distribution of AFs among Brazilians and their ancestry populations (NFE/EUR and AFR), as well as other admixed Americans (AMR), and the fact that the unique variants identified in the Brazilian population are rare or ultra-rare, indicates that the admixed Brazilian genetic background is not sufficient to influence SARS-CoV-2 infection on a population scale.

Zeberg and Pääbo[12] have shown that the major genetic risk factor for severe COVID-19 was inherited from Neanderthals[12]. This finding is important on a regional scale, since 4% of admixed Americans analysed by Zeberg and Pääbo[12] (including 1533 Brazilian controls from the BRACOVID dataset) presented the core haplotype derived from Neanderthals. Interestingly, Campinas, Ribeirão Preto, and the BIPMed dataset showed only one risk allele, while Barretos and Belo Horizonte did not present any risk allele of the Neanderthal's core haplotype reported. Therefore, if further studies demonstrate that the Neanderthal-derived region confers a risk to COVID-19, this information should be carefully evaluated in additional admixed Brazilian samples from different geographic areas.

Currently, there is no consensus regarding a possible association of HLA alleles and susceptibility to COVID-19. Ellinghaus et al.[11] did not find any evidence of an association between HLA and COVID-19. On the other hand, HLA-DRB1*15:01, HLA-DQB1*06:02, and HLA-B*27:07 alleles were associated with Italian patients affected by an extremely severe or severe form of COVID-19[23], and an increased frequency of HLA-C*07:29 and HLA-B*15:27 was detected in Chinese patients with COVID-19 in comparison to the Chinese control population[25]. Interestingly, the HLA-C*07:29 allele is absent from the Brazilian admixed samples included in the present study and in all populations used in the comparisons, except for individuals from India, where this allele was found at a low frequency (0.0003). On the other hand, the HLA-B*15:27 allele was identified in individuals from three Asian countries (Vietnam, Taiwan and Japan) with AF >0.01, and at a low frequency in Spain (0.0001), but absent from Brazilian samples. The HLA-B*27:07 allele found in Italian individuals with a severe manifestation of COVID-19 was also identified in India, Colombia, Spain, and South Africa, but not in populations from Asia and Oceania (countries that are less affected by COVID-19) and from Brazil. In contrast, the HLA-DQB1*06:02 is present in all populations surveyed in this study, including Brazilian individuals (17.1%), with the exception of individuals from Australia. Also, the HLA-DRB1*15:01 allele is present in all populations investigated in this study, including Brazilian individuals (12.2%), but not in individuals from Australia and Peru. Interestingly, 8.29% of Brazilian individuals carry both the HLA-DRB1*15:01 and HLA-DQB1*06:02 alleles.

Furthermore, the HLA-A, -B, -C, and DQ loci show haplotypes that are strong binders of SARS-CoV-2 peptides in the Brazilian samples, especially for the HLA-A locus (20 alleles, Table 3). When comparing different populations, we found marked variability

in the frequency of the different HLA alleles putatively associated with the severe manifestation of COVID-19, such as HLA-DRB1*15:01, HLA-DQB1*06:02, and HLA-B*27:07 alleles[23]. Overall, 10% of Brazilian individuals carried at least two of the alleles associated with the severe manifestation of COVID-19. Interestingly, the same alleles were absent from individuals from Australia. Variability in the frequency of HLA alleles previously associated with COVID-19 highlights the importance of considering ethnic and geographic origin when performing studies investigating the role of HLA alleles and disease. Thus, it seems likely that different population-specific haplotypes may be associated with an increased risk of severe disease in different populations.

In conclusion, we found rare variants in three COVID-19-related genes that are present only in the Brazilian dataset and are predicted to affect protein function. Furthermore, we identified HLA alleles previously associated with COVID-19 immunological response and 31 HLA alleles predicted as strong binders to SARS-CoV-2 peptides at loci -A, -B, -C, and DQ, which indicates the importance of further investigation on the role of HLA haplotypes as modulators of response to infection to SARS-CoV-2. Although the variants predicted to affect protein function in COVID-19-related genes are rare in admixed Brazilians (varying from 0.0001 to 0.0032), these also emerge as candidates for modulating response to infection by the SARS-CoV-2 in the Brazilian population. Furthermore, our study suggests the utility of population genomic studies in the context of precision health to stratify risk for infection disorders.

Methods

Subjects

We evaluated exomes from 257 individuals from the BIPMed dataset[19], 609 from the ABraOM dataset[24], and an additional 88 exomes from individuals born in three towns in south-eastern Brazil: Barretos (N=30), Ribeirão Preto (N=30), located in the state of São Paulo, and Belo Horizonte (N=28) the capital of the state of Minas Gerais. Among the BIPMed individuals, 193 had information about their city of birth available. The HLA region was sequenced in 386 individuals, including the 257 from BIPMed, the 88 additional exomes, and an additional 41 individuals (22 from southeast Brazil). We signed terms of data privacy to obtain permission to use the raw data from BIPMed and ABraOM public databases and use raw data of the 88 exomes from Barretos, Ribeirão Preto, and Belo Horizonte. This study was approved by the University of Campinas Research Ethics Committee (UNICAMP, Campinas, São Paulo, Brazil). All methods were performed according to the relevant guidelines and regulations.

Exome analysis

Whole exome data were stored in variant call format (VCF) files built-in GRCh37/hg19 assembly. Gene regions were extracted by *vcftools*[33] based on the position reported in Ensembl GRCh37 Release 101[34] (Supplementary Table 1). Variant consequences were annotated from each gene region by ANNOVAR software (version 2019Oct24)[35], using the following flags: *-otherinfo* (to include Brazil AF); *-onetranscript*; *-buildver hg19*; *-remove*; *-protocol refGene,gnomad211_exome*; *ALL.sites.2015_08*; *EUR.sites.2015_08*; *AFR.sites.2015_08*; *AMR.sites.2015_08*; *EAS.sites.2015_08*;

SAS.sites.2015_08; *-operation g,f*; and *-nastring.* ANNOVAR software provides allele frequency (AF) information from African/African-American (AFR/AFA), Latino/admixed American (LAT/AMR), East Asian (EAS), non-Finish European (NFE), and South Asian (SAS) populations from gnomAD exome dataset[21], as well as sub-Saharan Africans (AFR), Europeans (EUR), admixed Americans (AMR), east Asians (EAS), and south Asians (SAS) from 1KGP phase 3 dataset[22]. In addition, we annotated variants which were not identified by ANNOVAR using Variant Effect Prediction (VEP) algorithm[36], with the following parameters: *--buffer_size 500*; *--canonical*; *--distance 5000*; *--species homo_sapiens*; *--symbol*.

To evaluate whether regional variability is observed among Brazilian samples, we separated the samples based on the city in which individuals were born, including 32 individuals from Campinas extracted from the BIPMed dataset.

In silico *prediction analysis*

To predict the impact on protein function of the non-synonymous variants identified, we applied the following computer algorithms, which is currently recommended by the ACMG/AMP guidelines: PANTHER[37], MutationTaster[38], Condel[39], PROVEAN[40], PolyPhen2[41], Sort Intolerant from tolerant (SIFT)[42], Align Grantham Variation/ Grantham Difference score (GVGD)[43], Combined Annotation Dependent Depletion (CADD)[44], PhD-SNPg[45], Functional Analysis through Hidden Markov Models (FATHMM)[46], SNPs&GO[47], and MutPred2 (http://mutpred.mutdb.org).

For Align-GVGD, we classified the variants based on the graded classifier with a cut-off of C35 or higher for deleterious classification. For CADD, we used the PHRED-like score with a cut-off of 20, below which the variants were classified as benign and otherwise

deleterious. For MutPred2, we considered a score threshold of 0.50 for pathogenicity. For all other algorithms, we considered the classification provided as an output.

To access the impact of mutations on protein dynamics and stability, we used the DynaMut server (http://biosig.unimelb.edu.au/dynamut/)[48]. The server requires an input file of protein structure in PDB format or by providing the four-letter accession code for any entry on the Protein Data Bank database (PDB; http://wwpdb.org). The code for the *FURIN* gene used was 5jxg. The other proteins are not available in the PDB database to be tested.

HLA analysis

We sequenced 11 HLA Loci (HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1, HLA-DPB1, HLA-DQA1, HLA -DPA1, HLA-DRB3, HLA-DRB4, HLA-DRB5) in 298 samples using NGSgo® panels (GenDx, Utrecht, The Netherlands). The DNA libraries were loaded onto a MiSeq Sequencer (Illumina Inc., San Diego, CA, USA), and the data were analysed with the NGSengine v.2.16.2 software (GenDx, Utrecht, The Netherlands). We determined the HLA alleles from the remaining 88 exomes using the HLA-HD (HLA typing from High-quality Dictionary) tool v.1.3.0 [49-51]. The IPD-IMGT/HLA database release 3.40.0 [52] was used as a reference. Even though we obtained results with six- and eight-digit precision, we restricted the results to four-digit accuracy to compare with published data. HLA allele frequencies were calculated by Arlequin v.3.5.2.2 software [53].

In silico analysis of viral peptide-HLA class I and II binding affinity

We performed *in silico* analysis of viral peptide-HLA class I and II binding affinity across HLA proteins found in our population for the entire SARS-CoV-2 proteome. All HLA-A, -B, -C alleles were selected to assess the peptide-binding affinity of their

corresponding proteins HLA-A, HLA-B, HLA-C, respectively. The HLA-DR is represented by HLA-DRA/DRB1 dimer. Since HLA-DRA is considered monomorphic, we just used the HLA-DRB1. The HLA-DP and DQ are represented by the HLA-DPA1/DPB1 dimer and HLA-DQA1/DQB1 dimer, respectively.

FASTA-formatted protein sequence data was retrieved from the National Center of Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/Sars-cov-2). The follow eleven protein viral product was used in the analyses: ORF1ab (YP_009724389.1), Surface Glycoprotein (S) (YP_009724390.1), ORF3a (YP_009724391.1), Envelope (E) (YP_009724392.1), Membrane Glycoprotein (M) (YP_009724393.1), ORF6 (YP_009724394.1), ORF7a (YP_009724395.1), ORF7b (YP_009725318.1), ORF8 (YP_009724396.1), Nucleocapsid (N) (YP_009724397.2), and ORF10 (YP_009725255.1).

We k-merised these sequences into 8- to 12-mers to assess HLA class I-peptide binding affinity and into 15-mers to assess HLA class II binding affinity across the entire proteome. Predictions for HLA were performed using different HLA alleles found in our population with netMHCpan v4.1 for class I [54] and NetMHCIIpan - 3.2 for class II [55].

HLA allele and haplotype frequencies of other populations

HLA frequency data were obtained from the Allele Frequency Net Database (http://www.allelefrequencies.net/) [56] for 10 distinct populations that are most and least affected by COVID-19. We checked the HLA of the populations that occupy the top 10 positions (USA, India, Brazil, Russia, Colombia, Peru, Spain, Mexico, Argentina, South Africa) and those that were less affected (Australia, Vietnam, Taiwan, Japan, and South

Korea) (accessed on 04/24/2020, https://www.worldometers.info/coronavirus/) according to the availability of this data in the Allele Frequency Net Database.

Data availability

BIPMed raw dataset that supports this study's findings is available in EVA repository/PRJEB39251, https://www.ebi.ac.uk/eva/?eva-study=PRJEB39251. ABraOM raw dataset that supports the results of this study is available from ABraOM (http://abraom.ib.usp.br/).

Acknowledgments

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant number 2013/07559-3). RS was supported by FAPESP (grant number 2019/08526-8). TKA is supported by FAPESP (grant number 2017/01900-6). IL-C is supported by CNPq (grant number 311923/2019-4).

Competing interests

The authors declare no competing interests.

Author contributions

RS contributed with the study design, conceptualization, data acquisition, analysis, and paper writing; TKA contributed with HLA sequencing, analysis, *in silico* prediction analysis, and writing of the paper; MCG contributed with *in silico* prediction analysis and paper writing; CSR contributed with public data acquisition and processing; MN and MZ

contribute with public data acquisition; LD and MACB contributed with Belo Horizonte data acquisition and sample information; VLV contributed with Barretos data acquisition and sample information; WAS contributes with Ribeirão Preto data acquisition and sample information; ILC contributed with project conceptualization and served as principal investigators. All authors reviewed the manuscript.

References

1 Shang, J. *et al.* Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **117**, 11727-11734, doi:10.1073/pnas.2003138117 (2020).

2 Hou, Y. *et al.* New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. *BMC Medicine* **18**, 216, doi:10.1186/s12916-020-01673-z (2020).

3 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565-574, doi:10.1016/S0140-6736(20)30251-8 (2020).

4 Simmons, G. *et al.* Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11876-11881, doi:10.1073/pnas.0505577102 (2005).

5 Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A Multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular Cell* **78**, 779-784.e775, doi:10.1016/j.molcel.2020.04.022 (2020).

6 Ye, Q., Wang, B. & Mao, J. The pathogenesis and treatment of the `Cytokine Storm' in COVID-19. *Journal of Infection* **80**, 607-613, doi:https://doi.org/10.1016/j.jinf.2020.03.037 (2020).

7 Herold, T. *et al.* Elevated levels of IL-6 and CRP predict the need for mechanical ventilation in COVID-19. *Journal of Allergy and Clinical Immunology* **146**, 128-136.e124, doi:10.1016/j.jaci.2020.05.008 (2020).

8 Aziz, M., Fatima, R. & Assaly, R. Elevated interleukin-6 and severe COVID-19: A meta-analysis. *Journal of Medical Virology* **n/a**, doi:10.1002/jmv.25948 (2020).

9 Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science*, doi:10.1126/science.abd4570 (2020).

10 Cao, Y. *et al.* Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov* **6**, 11, doi:10.1038/s41421-020-0147-1 (2020).

11 Ellinghaus, D. *et al.* Genome-wide association study of severe covid-19 with respiratory failure. *New England Journal of Medicine*, doi:10.1056/NEJMoa2020283 (2020).

12 Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*, doi:10.1038/s41586-020-2818-3 (2020).

13 Shi, Y. *et al.* COVID-19 infection: the perspectives on immune responses. *Cell Death Differ* **27**, 1451-1454, doi:10.1038/s41418-020-0530-3 (2020).

14 Nguyen, A. *et al.* Human leukocyte antigen susceptibility map for severe acute respiratory syndrome coronavirus 2. *Journal of Virology* **94**, e00510-00520, doi:10.1128/JVI.00510-20 (2020).

15 Rodrigues de Moura, R., Coelho, A. V. C., de Queiroz Balbino, V., Crovella, S. & Brandão, L. A. C. Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. *American Journal of Human Biology* **27**, 674-680, doi:10.1002/ajhb.22714 (2015).

16 Kehdy, F. S. G. *et al.* Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences* **112**, 8696-8701, doi:10.1073/pnas.1504447112 (2015).

17 Lima-Costa, M. F. *et al.* Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative). *Sci Rep* **5**, 9812, doi:10.1038/srep09812 (2015).

18 Secolin, R. *et al.* Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci Rep* **9**, 13900, doi:10.1038/s41598-019-50362-2 (2019).

19 Rocha, C. S., Secolin, R., Rodrigues, M. R., Carvalho, B. S. & Lopes-Cendes, I. The Brazilian Initiative on Precision Medicine (BIPMed): fostering genomic data-sharing of underrepresented populations. *npj Genomic Medicine* **5**, 42, doi:10.1038/s41525-020-00149-6 (2020).

20 Chapman, S. J. & Hill, A. V. Human genetic susceptibility to infectious disease. *Nat Rev Genet* **13**, 175-188, doi:10.1038/nrg3114 (2012).

21 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

22 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

23 Novelli, A. *et al.* HLA alleles frequencies and susceptibility to COVID-19 in a group of 99 Italian patients. *HLA* **n/a**, doi:10.1111/tan.14047 (2020).

24 Naslavsky, M. S. *et al.* Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum Mutat* **38**, 751-763, doi:10.1002/humu.23220 (2017).

25 Wang, W., Zhang, W., Zhang, J., He, J. & Zhu, F. Distribution of HLA allele frequencies in 82 Chinese individuals with coronavirus disease-2019 (COVID-19). *HLA* **96**, 194-196, doi:10.1111/tan.13941 (2020).

26 Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat Rev Immunol* **18**, 325-339, doi:10.1038/nri.2017.143 (2018).

27 Alcendor, D. J. Racial disparities-associated COVID-19 mortality among minority populations in the US. *J Clin Med* **9**, doi:10.3390/jcm9082442 (2020).

28 Kaufman, H. W., Niles, J. K. & Nash, D. B. Disparities in SARS-CoV-2 positivity rates: Associations with race and ethnicity. *Popul Health Manag*, doi:10.1089/pop.2020.0163 (2020).

29 Poulson, M. *et al.* National disparities in COVID-19 outcomes between black and white Americans. *J Natl Med Assoc*, doi:10.1016/j.jnma.2020.07.009 (2020).

30 Deng, L., Ruiz-linares, A., Xu, S. & Wang, S. Ancestry variation and footprints of natural selection along the genome in Latin American populations. *Scientific Reports* **6**, 1-7, doi:10.1038/srep21766 (2016).

31 Cáceres, A. & González, J. R. Following the footprints of polymorphic inversions on SNP data: from detection to association tests. *Nucleic acids research* **43**, e53-e53, doi:10.1093/nar/gkv073 (2015).

32 Homburger, J. R. *et al.* Genomic insights into the ancestry and demographic history of South America. *PLoS Genetics* **11**, doi:10.1371/journal.pgen.1005602 (2015).

33 Danecek, P. *et al.* The variant call format and VCF tools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

34 Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Research* **48**, D682-D688, doi:10.1093/nar/gkz966 (2020).

35 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164-e164, doi:10.1093/nar/gkq603 (2010).

36 McLaren, W. *et al.* The Ensembl variant effect predictor. *Genome Biol* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

37 Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modelling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* **41**, D377-386, doi:10.1093/nar/gks1118 (2013).

38 Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575-576, doi:10.1038/nmeth0810-575 (2010).

39 González-Pérez, A. & López-Bigas, N. Improving the assessment of the outcome of non-synonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88**, 440-449, doi:10.1016/j.ajhg.2011.03.004 (2011).

40 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**, e46688, doi:10.1371/journal.pone.0046688 (2012).

41 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7.20, doi:10.1002/0471142905.hg0720s76 (2013).

42 Sim, N. L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40**, W452-457, doi:10.1093/nar/gks539 (2012).

43 Tavtigian, S. V. *et al.* Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* **43**, 295-305, doi:10.1136/jmg.2005.033878 (2006).

44 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894, doi:10.1093/nar/gky1016 (2019).

45 Capriotti, E. & Fariselli, P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res* **45**, W247-W252, doi:10.1093/nar/gkx369 (2017).

46 Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536-1543, doi:10.1093/bioinformatics/btv009 (2015).

47 Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* **30**, 1237-1244, doi:10.1002/humu.21047 (2009).

48 Rodrigues, C. H., Pires, D. E. & Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* **46**, W350-W355, doi:10.1093/nar/gky300 (2018).

49 Kawaguchi, S. & Matsuda, F. High-definition genomic analysis of HLA genes via comprehensive HLA allele genotyping. *Methods Mol Biol* **2131**, 31-38, doi:10.1007/978-1-0716-0389-5_3 (2020).

50 Kawaguchi, S., Higasa, K., Yamada, R. & Matsuda, F. Comprehensive HLA typing from a current allele database using next-generation sequencing data. *Methods Mol Biol* **1802**, 225-233, doi:10.1007/978-1-4939-8546-3_16 (2018).

51 Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum Mutat* **38**, 788-797, doi:10.1002/humu.23230 (2017).

52 Robinson, J. *et al.* IPD-IMGT/HLA database. *Nucleic Acids Res* **48**, D948-D955, doi:10.1093/nar/gkz950 (2020).

53 Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**, 564-567, doi:10.1111/j.1755-0998.2010.02847.x (2010).

54 Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **48**, W449-W454, doi:10.1093/nar/gkaa379 (2020).

55 Nielsen, M., Justesen, S., Lund, O., Lundegaard, C. & Buus, S. NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res* **6**, 9, doi:10.1186/1745-7580-6-9 (2010).

56 Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res* **48**, D783-D788, doi:10.1093/nar/gkz1029 (2020).

Tables

Table 1. Alternative allele frequency (AAF) of variants encompassing the COVID-19 risk core haplotype and the alleles present in Neanderthal samples.

dbSNP	Chr	Pos	ALT	Brazil AF	Risk allele*	Campinas	Barretos	Ribeirão Preto	Belo Horizonte	BIPMed	ABraOM
rs35044562	3	45909024	G	0.0011	G	0.0862	0.0000	0.0000	0.0000	0.0279	0.0311
rs34326463	3	45899651	G	0.0011	G	0.0000	0.0000	0.0357	0.0000	0.0000	0.0000
rs35508621	3	45880481	C	0.0011	C	0.0345	0.0000	0.0000	0.0000	0.0000	0.0039
rs67959919	3	45871908	A	0.0011	A	0.0000	0.0000	0.0000	0.0000	0.0000	0.0039
rs35624553	3	45867440	G	0.0262	G	0.0000	0.0000	0.0000	0.0000	0.0000	0.0039

*data extracted from Zeberg and Pääbo, 2020 (doi:10.1038/s41586-020-2818-3)

Table 2. Alternative allele frequency of deleterious variants according to 12 different prediction algorithms

		Alternative Allele Frequency										
Gene	Variant	Campinas	Barretos	Ribeirão Preto	Belo Horizonte	ABraOM	BIPMed	NFE	AFR	AMR	EAS	SAS
SLC6A20	Phe249Ser	0	0.0167	0	0	0.0008	0.0019	0.0005	0.0615	0.0003	0	0.0084
SLC6A20	Gly164Val	0	0	0	0	0.0008	0	0.0088	0	0	0	0
SLC6A20	Leu25Pro	0	0.0167	0	0	0	0	0	0	0	0	0
LZTFL1	Leu96Arg	0	0	0	0	0.0008	0	0	0	0	0	0
XCR1	Tyr287Ser	0	0	0	0.0167	0	0	0	0	0	0	0
FURIN	Gly146Ser	0	0	0	0	0.0008	0	0.0004	0.0002	0.0005	0.0544	0.0327
FURIN	Asn414Ser	0	0	0	0	0	0.0019	0	0	0	0	0

NFE=non-Finish European; AFR=sub-Saharan African/African American; AMR=admixed Americans/Latinos; EAS=east Asians; SAS=south Asians.

Table 3. List of HLA strongest binders (>300 peptides bound at high affinity) of SARS-CoV-2 peptides and frequency in the Brazilian sample.

HLA	HLA alleles	Allele frequency
HLA -A	A*01:01	0.10233
	A*11:01	0.04145
	A*11:67	0.00130
	A*23:01	0.03756
	A*23:17	0.00389
	A*24:02	0.10104
	A*24:03	0.00259
	A*24:05	0.00130
	A*25:01	0.00389
	A*26:01	0.02979
	A*26:02	0.00130
	A*26:08	0.00130
	A*29:01	0.00259
	A*29:02	0.04534
	A*29:119	0.00130
	A*30:02	0.02591
	A*30:04	0.00259
	A*34:02	0.00777
	A*36:01	0.00389
	A*80:01	0.00130
HLA-B	B*15:08	0.00130
	B*15:11	0.00130
HLA-C	C*03:02	0.00389
	C*07:02	0.05699
	C*07:50	0.00130
	C*14:02	0.02979
	C*14:03	0.00259
HLA-DQ	DQA1*0201-DQB1*0402	0.00259
	DQA1*0301-DQB1*0402	0.00130
	DQA1*0303-DQB1*0401	0.00130
	DQA1*0303-DQB1*0402	0.00130

Figures legend

Figure 1. Distribution of alternative allele frequency of common variants (AAF≥0.01) from samples of admixed Brazilians and worldwide public datasets. The x-axis shows the 56 variants found in common between the Brazil sample and the gnomAD and 1KGP dataset. (A) Comparison between Brazilians and gnomAD, and (B) including non-Finland Europeans (NFE), sub-Saharan Africans/African Americans (AFR) Venn diagrams that show the overlap between samples (A) and variants (B) in the WES and the SNP array datasets from BIPMed reference samples.

Figure 2. Distribution of alternative allele frequency of common variants (AAF>0.01) separated by Brazilian cities. The x-axis shows the 56 variants found in common with gnomAD and 1KGP. This barplot also includes the two public Brazilian datasets (BIPMed and ABraOM) and the frequency of all samples combined (Brazil).

Figure 1

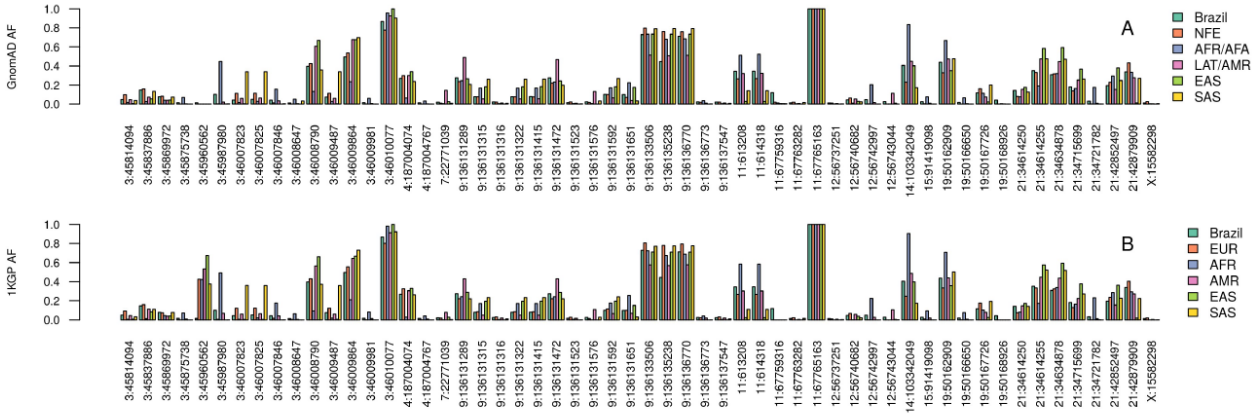

Figure 2