Received: 10 February 2021 | Revised: 10 May 2021 | Accepted: 17 May 2021

DOI: 10.1002/gepi.22421

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

Genome-wide association analysis of COVID-19 mortality risk in SARS-CoV-2 genomes identifies mutation in the SARS-CoV-2 spike protein that colocalizes with P.1 of the Brazilian strain

Georg Hahn[1] | Chloe M. Wu[2] | Sanghun Lee[1,3] | Sharon M. Lutz[1,4] | Surender Khurana[5] | Lindsey R. Baden[6] | Sebastien Haneuse[1] | Dandi Qiao[7,8] | Julian Hecker[4,7] | Dawn L. DeMeo[7,8] | Rudolph E. Tanzi[9] | Manish C. Choudhary[7] | Behzad Etemad[7] | Abbas Mohammadi[7] | Elmira Esmaeilzadeh[7] | Michael H. Cho[7,8] | Jonathan Z. Li[7] | Adrienne G. Randolph[7,10] | Nan M. Laird[1] | Scott T. Weiss[7,8] | Edwin K. Silverman[7,8] | Katharina Ribbeck[2] | Christoph Lange[1,7,8]

[1]Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA

[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[3]Department of Medical Consilience, Graduate School, Dankook University, Yongin, South Korea

[4]PRecisiOn Medicine Translational Research (PROMoTeR) Center, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA

[5]Food and Drug Administration, Silver Spring, Maryland, USA

[6]Division of Infectious Diseases, Brigham and Women's Hospital, Boston, Massachusetts, USA

[7]Harvard Medical School, Harvard University, Boston, Massachusetts, USA

[8]Department of Medicine, Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA

[9]Genetics and Aging Research Unit, McCance Center for Brain Health, Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

[10]Department of Anesthesiology, Critical Care and Pain Medicine, Boston Children's Hospital, Boston, Massachusetts, USA

Correspondence
Georg Hahn, Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA.
Email: ghahn@hsph.harvard.edu

Funding information
National Heart, Lung, and Blood Institute, Grant/Award Numbers: 2U01HG008685, P01HL120839, P01HL132825, U01HL089856, U01HL089897; NIH Clinical Center, Grant/Award Numbers: 1R01AI154470-01, 2U01HG008685;

Abstract

SARS-CoV-2 mortality has been extensively studied in relation to host susceptibility. How sequence variations in the SARS-CoV-2 genome affect pathogenicity is poorly understood. Starting in October 2020, using the methodology of genome-wide association studies (GWAS), we looked at the association between whole-genome sequencing (WGS) data of the virus and COVID-19 mortality as a potential method of early identification of highly pathogenic strains to target for containment. Although continuously updating our analysis, in December 2020, we analyzed 7548 single-stranded SARS-CoV-2 genomes of COVID-19 patients in the GISAID database and associated variants with

Georg Hahn and Chloe M. Wu contributed equally to this study.

National Institutes of Health, Grant/Award Number: P30-ES002109; National Human Genome Research Institute, Grant/Award Number: R01HG008976; National Science Foundation, Grant/Award Numbers: NSF GRFP 1745302, NSF PHY 2033046

mortality using a logistic regression. In total, evaluating 29,891 sequenced loci of the viral genome for association with patient/host mortality, two loci, at 12,053 and 25,088 bp, achieved genome-wide significance (p values of 4.09e−09 and 4.41e−23, respectively), though only 25,088 bp remained significant in follow-up analyses. Our association findings were exclusively driven by the samples that were submitted from Brazil (p value of 4.90e−13 for 25,088 bp). The mutation frequency of 25,088 bp in the Brazilian samples on GISAID has rapidly increased from about 0.4 in October/December 2020 to 0.77 in March 2021. Although GWAS methodology is suitable for samples in which mutation frequencies varies between geographical regions, it cannot account for mutation frequencies that change rapidly overtime, rendering a GWAS follow-up analysis of the GISAID samples that have been submitted after December 2020 as invalid. The locus at 25,088 bp is located in the P.1 strain, which later (April 2021) became one of the distinguishing loci (precisely, substitution V1176F) of the Brazilian strain as defined by the Centers for Disease Control. Specifically, the mutations at 25,088 bp occur in the S2 subunit of the SARS-CoV-2 spike protein, which plays a key role in viral entry of target host cells. Since the mutations alter amino acid coding sequences, they potentially imposing structural changes that could enhance viral infectivity and symptom severity. Our analysis suggests that GWAS methodology can provide suitable analysis tools for the real-time detection of new more transmissible and pathogenic viral strains in databases such as GISAID, though new approaches are needed to accommodate rapidly changing mutation frequencies over time, in the presence of simultaneously changing case/control ratios. Improvements of the associated metadata/patient information in terms of quality and availability will also be important to fully utilize the potential of GWAS methodology in this field.

KEYWORDS

GISAID database, logistic regression, mortality, SARS-CoV-2, spike protein, whole-genome sequencing

1 | INTRODUCTION

Viral mutations can cause increased virulence/transmissibility/immune evasion/pathogenicity (Long et al., 2020), both in animals (Brault et al., 2007; Geoghegan & Holmes, 2018), and in humans (Bae et al., 2018; Nogales et al., 2017). Especially for the SARS-CoV-2 virus, the discovery of potential links between viral mutations and disease outcome would have important implications for COVID-19 surveillance and containment (Lo & Jamrozy, 2020), diagnosis, prognosis, and treatment development. In this contribution, we probed each locus of the single-stranded RNA of the SARS-CoV-2 virus for direct association with host/patient mortality.

In our initial analysis (October 2020), we aimed to identify potential links between viral mutations and mortality by utilizing the GISAID database (Elbe & Buckland-Merrett, 2017; Shu & McCauley, 2017). Although continuously updating the analysis, in December 2020, GISAID contained data on 7548 COVID-19 patients from 86 countries for whom metadata was available, that is, age, sex, location, and patient status, and whose viral genomes are sequenced (see Table 1). The variable "patient status" indicates if the patient was alive or deceased at the time the virus sample was submitted to GISAID; we used it as a surrogate for mortality in our analysis. As non-deceased patients at enrollment could have died of Covid-19 later, such misclassifications can lead to reduced statistical power, but not to an inflated type-1 error. For the analysis, we repurposed the methodology of genome-wide association studies (GWAS) (Manolio, 2010). This approach is widely used in human

TABLE 1 Characteristics of all patients in the GISAID data set for whom complete meta-information and sequenced viral genomes were available

Region	#total	#females	#males	Deceased/ non- deceased	%deceased	Mean age	Mutation frequency in % at the following loci	
							12,053	25,088
Entire data set	7548	3313	4235	722/6826	9.6	47.6	1.2	2.2
Africa	1517	954	563	2/1515	0.1	38.8	0.0	0.2
Eastern Mediterranean	730	180	550	131/599	17.9	45.4	0.0	0.1
Europe	1872	896	976	70/1802	3.7	56.0	0.1	0.0
Pan American Health Organization	1505	637	868	435/1070	28.9	51.9	5.7	10.6
Brazil	430	223	207	192/238	44.7	55.1	20.0	37.0
South-East Asia	1116	367	749	83/1033	7.4	45.1	0.0	0.1
Western Pacific	808	279	529	1/807	0.1	41.6	0.0	0.2

Note: Total number of samples (as well as males/females), numbers of deceased/non-deceased, rate of deceased samples at enrollment, mean age, and mutation frequencies for 12,053 and 25,088 bp.

genetics and can test thousands of genetic loci for association in data sets such as the one of GISAID.

To identify potential confounding geographic factors in the sequencing data, we first conducted principal component analysis of the Jaccard similarity matrix (Figure 1) that was computed for the 7548 viral genomes available for our analysis. We utilized the Jaccard similarity matrix because its computation does not require estimates of the mutation frequency for each locus in the SARS-CoV-2 genome, in contrast to other similarity matrices such as the variance/covariance matrix (Prokopenko et al., 2016). We found that the virus genomes clustered in distinctive branches that correspond to the geographic regions from where their data was submitted to GISAID (Forster et al., 2020, Hahn, Lee, Weiss, et al., 2020) (see Figure 1). Both, the geographical clustering of the viral genomes and their similarity within regions, can cause bias in the association analysis if unaccounted for. Hence, we generated additional eigenvector plots to investigate the number of eigenvectors needed to eliminate bias caused by such clustering. Based on a visual inspection of these plots, we selected the first 10 eigenvectors of the Jaccard matrix as covariates for the following logistic regression analyses.

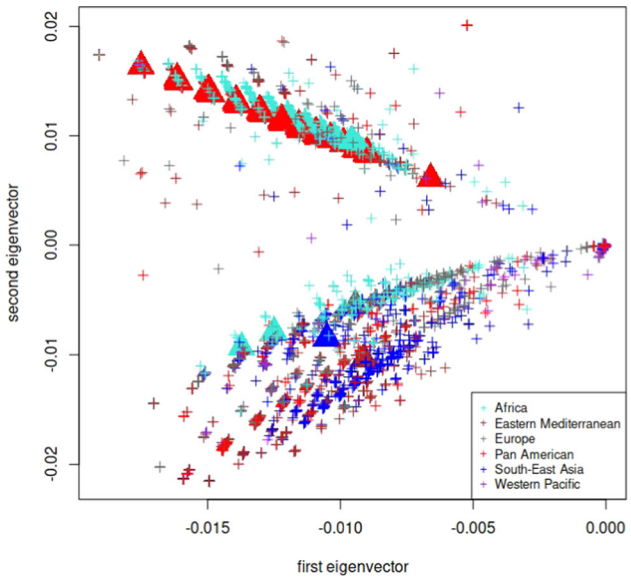

FIGURE 1 Geographic distribution of 7548 SARS-CoV-2 genomes. Genomes are depicted according to their first two eigenvectors of the Jaccard matrix and colored by geographic region. The eigenvector plot shows distinct grouping of SARS-CoV-2 genomes according to their geographic origin. Furthermore, genomes that carry a mutation at 12,053 or 25,088 bp are depicted by triangles. The majority of those are located in a subbranch whose samples come predominantly from Pan America

2 | METHODS

2.1 | Data acquisition

The analysis presented in this article is based on nucleotide sequences with accession numbers EPI_ISL_403962 to EPI_ISL_636981, downloaded from the GISAID database (Elbe & Buckland-Merrett, 2017; Shu & McCauley, 2017) as a file in "fasta" format on 06 December 2020. Only patients with additional metadata (age, sex, and hospitalization status as plain text

comments) were selected on GISAID, resulting in 8647 samples.

2.2 | Data cleaning

We filtered the 8647 samples for complete nucleotide sequences and aligned them to the SARS-CoV-2 reference sequence (published on GISAID under the accession number EPI_ISL_402124) using MAFFT (Katoh et al., 2002).

Using the location tag in the fasta file, we grouped all samples according to the WHO regional offices for Africa (AFRO, $N = 1517$), for the Eastern Mediterranean (EMRO, $N = 730$), for Europe (EURO, $N = 1872$), for South-East Asia (SEARO, $N = 1116$), for the Western Pacific (WPRO, $N = 808$), as well as the Pan American Health Organization (PAHO, $N = 1505$). In particular, the countries included in each group are as follows: (1) AFRO (Algeria, South Africa, Gambia, Nigeria, Senegal, as well as Congo, Madagascar, Mozambique, Tunisia, Ghana, Rwanda, Cameroon); (2) EMRO (Egypt, Morocco, Kuwait, Lebanon, Oman, Saudi Arabia, United Arab Emirates, as well as Iran, Iraq, Bahrain); (3) EURO (Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Faroe Islands, France, Germany, Hungary, Italy, Israel, Poland, Portugal, Romania, Russia, Slovakia, Spain, Sweden, Turkey, Kazakhstan, as well as Andorra, Georgia, Norway, Ukraine, Switzerland, Saint Barthelemy, Guadeloupe, Saint Martin, Mongolia, Greece, Finland, Moldova, Reunion); (4) PAHO (Canada, USA, Costa Rica, Mexico, Argentina, Brazil, Chile, Colombia, Ecuador, Peru, Venezuela, as well as Puerto Rico, Uruguay, Panama, Dominican Republic); (5) SEARO (Bangladesh, India, Indonesia, Myanmar, Nepal, Sri Lanka, Thailand); (6) WPRO (Cambodia, Japan, Malaysia, Vietnam, Australia, Guam, Hong Kong, China, Singapore, as well as South Korea, Taiwan, New Zealand, Philippines).

Finally, we matched the samples to the metadata information (age, sex, clinical outcome) available on GISAID. Filtering for those samples having complete metadata information resulted in $n = 7548$ samples.

2.3 | Data analysis

After alignment with MAFFT (Katoh et al., 2002), we compared all aligned sequences of length $p = 29{,}891$ entrywise to the SARS-CoV-2 reference sequence, and denoted in a matrix X with an entry $X_{ij} = 1$ that sequence i deviated from the reference sequence at position j. All other entries of X are zero.

We used the R-package "locStra" (Hahn, Lutz, Hecker, et al., 2020; Hahn, Lutz, & Lange, 2020) to calculate the Jaccard similarity matrix (Jaccard, 1901; Prokopenko et al., 2016; Schlauch et al., 2017; Tan et al., 2005) for the n viral genomes based on the matrix X. The Jaccard matrix $J(X)$ has n rows and n columns, and each entry (i,j) is the Jaccard similarity index between the binary vector of mismatches/mutations (with respect to the reference sequence) for the ith and jth SARS-CoV-2 genome in our data set. Computation of the first 10 eigenvectors of the Jaccard similarity matrix $J(X)$ allows us to visualize the geographic clustering of the viral genomes. We guard the logistic regression analysis against confounding by including the first eigenvectors in the regression analysis as covariates.

For the association analysis of the entire viral genome, we defined the response to be a binary indicator for the clinical outcome, where we only distinguish between all those patients/hosts whose hospitalization status tag at enrollment into the GISAID database was listed as "deceased" (outcome of 1) versus the remaining samples as non-deceased (outcome of 0). At this point, no other information regarding clinical outcome is available in GISAID.

We performed a logistic regression of the binary outcome variable for each of the $p = 29{,}891$ loci on the following covariates: the column vector $X_{\cdot i}$ encoding the mismatches/mutations of each sample at the ith location on the SARS-CoV-2 nucleotide sequence, the patient's age, sex, location (WHO region), and the first 10 eigenvectors of the Jaccard matrix. The WHO region was included as we observed in Figure 1 that the viral genomes cluster into distinct branches that correspond to the geographic regions. The logistic regression was carried out in R using the default "glm" command, where the parameter "family" was set to "family=binomial(link = "logit")." We tested the ith locus/location of the viral genome for association with mortality by testing whether the regression coefficient for column $X_{\cdot i}$ is equal to zero. We controlled for multiple tests using the Bonferroni correction at an uncorrected threshold of 0.05, resulting in the corrected threshold of $0.05/29{,}891 = 1.67\text{e}{-06}$.

To quantify unmeasured confounding, we computed E-values (VanderWeele & Ding, 2017) with the help of the function "evalues.OLS" of the R-package "EValue" (Mathur et al., 2021) on CRAN. Conditional on the measured covariates, the E-value is the minimum strength of association (with both treatment and outcome) required for an unmeasured confounder to fully explain a specific treatment–outcome association. The E-value is measured on the risk ratio scale. A small (large) E-value indicates that small (considerable) amounts of unmeasured confounding is needed to explain an effect estimate.

Finally, we also perform an analysis with a matched data set. For this, we match each sample in GISAID that is deceased at submission to the closest non-deceased one, measured in Euclidean distance in the eigenvector space of the Jaccard matrix (Figure 1). When running the logistic regression on the matched data set, we test each of the $p = 29{,}891$ loci on the column vector $X_{\cdot i}$ (encoding the mismatches to the reference genome), as well as the patient's age and sex only.

3 | RESULTS

After testing each locus (presence/absence of mutation) of the viral genome individually for association with the status indicator variable (deceased/non-deceased) of the host/patient at submission to GISAID, two loci of the SARS-CoV-2 genome achieved genome-wide significance: one at position 12,053 bp with p value 4.09e−09, and one at 25,088 bp with p value 4.41e−23 (Table 2). The E-values for both loci are 715 and 6696, respectively, hinting at the fact that a considerable unmeasured confounding would be needed to explain such an effect estimate.

To investigate the robustness of the highly significant association signals, we examined the data set at the individual patient and locus level. Our findings were enabled by two features specific to the data: (1) the Brazilian centers reported much larger numbers of deceased patients than the other centers world-wide. At enrollment, 44.7% of the Brazilian patients were deceased in contrast to 9.6% in the entire data set (including Brazil). (2) We also noticed that all genomes that carry at least one of the mutations either at 12,053 or 25,088 bp are located predominantly in the branch of the eigenvector plot (see Figure 1) that corresponds to the PAHO/South America region.

We conducted two different types of sensitivity analyses to minimize the chances that the observed associations are caused by confounding/GISAID data set composition (Table 2): (1) Our data set was restricted to genomes that were matched based proximity in the eigenvector plots (see Section 2 for details), called "matching" in Table 2. (2) As further examination of the deceased indicator variable revealed that all "deceased" carrier genomes came from Brazil, our second sensitivity analysis was restricted to genomes that were submitted from the PAHO region and Brazil, respectively. In both analyses, 25,088 bp maintained significance at 0.05/29,891 = 1.67e−06, but 12,053 bp ceased to be significant. The effect size estimates showed risk increases for mortality of a factor of 5–16 for carriers of a mutation at 25,088 (Table 2). The E-value for 25,088 bp in the Brazil analysis is 3.0, that is, to move the confidence interval to include the null, an unmeasured confounder that is associated with the Covid-19 mortality and the presence of the mutation at 25,088 by a risk ratio of 3.0-fold each could do so, but weaker confounding could not.

To summarize, all the results of the secondary analyses (Table 2) support the genome-wide significant association between the mutation at 25,088 bp and mortality. The large effect estimate and E-value for the mutation at 25,088 bp (Table 2) are substantial in support of the association, as it is difficult to imagine an unaccounted confounding mechanism that would affect this mutation among roughly 30k loci and that would be strong enough to cause such profound association signals in our analysis. Since the criteria for selection into the study likely varies by country, and may be related to the deceased indicator, the odds ratio estimate from the Brazil sample alone may be most interpretable. Among the samples from Brazil, 18.2% of the patients whose viral genome did not carry any mutation at either loci were deceased at enrollment, compared with 82.4% for patients whose viral genomes carried the mutation at 25,088 bp only.

As of December 2020, Table 1 also provides a regional breakdown of the "deceased-at-enrollment" rates and the mutation frequencies for both loci. The rarity of the mutations outside of Brazil in December 2020 means that there was virtually no power to detect any association (if they existed).

It is important to note that locus at 25,088 bp colocalizes with the P.1 variant that has become part of the CDC definition (precisely, substitution V1176F) of the Brazilian strain in April 2021 (UCSC Genome Browser on SARS-CoV-2, 2021).

TABLE 2 Sample size, number of deceased samples, as well as p values and odds ratios from the logistic regression on the two mutations: for the entire data set, for each WHO region, and for samples from Brazil only

Analysis	Sample size	Deceased	Locus	p Value	Odds ratio
Overall	7548	722	12,053	4.09e−09	6.4
			25,088	4.41e−23	12.9
Matched analysis	1452	722	12,053	5.53e−05	3.5
			25,088	4.91e−11	4.8
PAHO	1505	435	12,053	1.22e−09	7.3
			25,088	3.10e−24	15.9
Brazil	430	192	12,053	2.27e−04	3.5
			25,088	4.90e−13	9.2

4 | DISCUSSION

Single mutations in viruses can confer enhanced transmission and/or virulence associated with patient mortality (Bae et al., 2018; Brault et al., 2007). In our analysis of SARS-CoV-2, the mutation at 25,088 bp occurs in the spike glycoprotein, which mediates viral attachment and cellular entry.

The spike protein consists of two functional subunits: S1, which contains the receptor-binding domain, and S2, which contains the machinery needed to fuse the viral membrane to the host cellular membrane. The mutation at 25,088 bp is in the S2 subunit, and specifically occurs within the S2′ site, which is cleaved by host proteases to activate membrane fusion (Figure 2). The V1176F mutation in S2 is located in the Heptad repeat 2 domain, which is involved in the viral fusion machinery. In many viruses, membrane fusion is activated by proteolytic cleavage, an event which has been closely linked to infectivity—for instance, a multibasic cleavage site is a signature of highly pathogenic viruses including avian influenza (Walls et al., 2020). In coronaviruses, membrane fusion is known to depend on proteolytic cleavage at multiple sites, including the S1/S2 site, located at the interface between the S1 and S2 domains, and the S2′ site located within the S2 domain. These cleavage events can impact infection—in fact, a distinct furin cleavage site present in the SARS-CoV-2 S1/S2 site is not found in SARS-CoV (Vankadari, 2020), and it is thought to increase infectivity through enhanced membrane fusion activity (Vankadari, 2020; Walls et al., 2020; Xia et al., 2020). Consequently, mutations at these sites can alter virulence—for instance, a recent study reported that mutations disrupting the multibasic nature of the S1/S2 site affect SARS-CoV-2 membrane fusion and entry into human lung cells (Hoffmann et al., 2020). Several studies have also found that SARS-CoV mutants with an added furin recognition site at S2′ had increased membrane fusion activity (Belouzard et al., 2009; Watanabe et al., 2008). Although enhanced infectivity does not always cause a higher fatality rate, more infectious viruses can lead to a higher viral load, which can impact symptom severity and mortality (Pujadas et al., 2020).

All carriers of a mutation at 25,088 bp exhibit a G to T missense mutation (Table 3), which changes the encoded amino acid from valine to phenylalanine. Compared to the branched chain structure of valine, phenylalanine has a bulkier aromatic structure. Such a substitution may

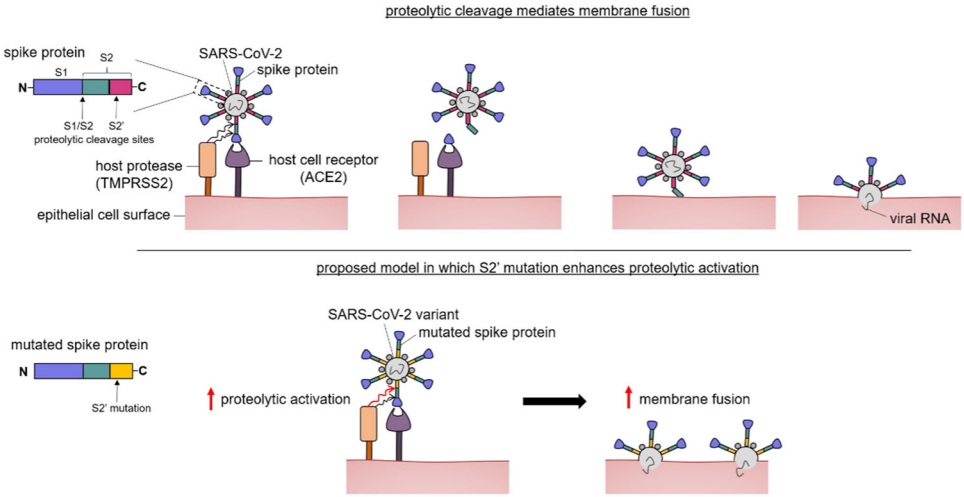

FIGURE 2 Proposed model showing how the S2 mutation may enhance proteolytic activation. The SARS-CoV-2 spike protein is colored by region (blue—S1, green—S2, magenta—S2′). The S2′ site is cleaved by host proteases, facilitating membrane fusion and viral entry into host cells. A mutation in this region, depicted in yellow, could theoretically increase proteolytic activity and membrane fusion, thereby causing greater infectivity

TABLE 3 Number of genomic variants at each locus, affected protein position, and corresponding amino acid change

Locus	A	C	G	T	Protein	Position	Primary substitution
12,053	0	**7453**	0	87	nsp7	71	Leu ⟶ Phe
25,088	0	0	**7331**	166	Spike	1176	Val ⟶ Phe

Note: Amino acid in the reference sequence in bold.

impose local structural constraints, stabilize particular secondary structures (Makwana & Mahalakshmi, 2015), or introduce specific interactions which lead to preferential binding. Therefore, a mutation in the S2′ domain which promotes proteolytic cleavage could theoretically enhance viral infectivity (Figure 2) and consequently, patient mortality. Although many current therapies primarily target the receptor binding domain within the S1 subunit of the SARS-CoV-2 spike protein, our findings suggest that the S2 domain may be an important additional target for therapeutic development. The emergence of a more aggressive P.1 lineage carrying this mutation was associated with a second wave of infection across Brazil (Faria et al., 2021). Several modeling approaches have estimated P.1 to have higher transmission and reinfection (Faria et al., 2021; Coutinho et al., 2021, preprint), and there is evidence suggesting that P.1 is less susceptible to therapeutic or vaccine-induced neutralizing antibodies (Hoffmann et al., 2021). Further experimental characterization of the biological effects of this mutations can have important implications for SARS-CoV-2 treatment and containment.

The mutation at 12,053 bp occurs within the ORF1ab gene, which expresses a polyprotein comprised of 16 nonstructural proteins (Yoshimoto, 2020). Specifically, 12,053 bp occurs in NSP7, which dimerizes with NSP8 to form a heterodimer that complexes with NSP12, ultimately forming the RNA polymerase complex essential for genome replication and transcription. Mutations causing enhanced viral polymerase activity have been linked to increased pathogenicity of influenza viruses. All carriers of a mutation at 12,053 bp exhibit a C to T missense mutation, which causes leucine to be substituted for phenylalanine (Table 3). Such a mutation may confer structural rigidity which could potentially alter interactions with other components of replication and transcription machinery, though experimental analysis is needed to test these hypotheses.

From a methodological perspective, there are potential strengths to our GWAS analysis approach to sequenced SARS-CoV-2 genomes. As the independent support of our association findings for the locus at 25,088 bp illustrates, GWAS methodology might be a well-suited tool for the early detection of new viral strains in global database systems such as GISAID, to which scientists submit their viral genomes during pandemics with minimal requirements regarding the meta/clinical information about the host/patient. In general, GWAS methodology would be suitable to analyze the highly correlated viral genomes in such data sets, as the GWAS approach can simultaneously handle different subpopulations with different proportions of cases/controls.

However, there are important limitations to applying GWAS methodology in a pandemic. More transmissible variants will alter mutation frequencies and increase the case/control ratio, as occurred for these two variants. Deployment of vaccines or targeted monoclonal antibodies may exert immunologic pressure on the virus leading to selective viral evolution. Standard GWAS methodology assumes a stable mutation frequency and is then no longer valid. Additional analytic methods are required to adjust for a time-changing variant frequency but to fully utilize all viral genome sequences, the availability and the quality of meta information/patient information must be robust, using consistent outcome definitions and accurate data capture.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributors, originating and submitting laboratories of the sequences from GISAID's EpiCoV™ Database (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017) on which this study is based. A detailed list of contributors is available in the Supplementary Information. Funding for this study was provided through the National Institutes of Health (1R01AI154470-01; 2U01HG008685; R01HG0-08976; U01HL089856, U01HL089897, P01HL120839, P01HL132825, 2U01HG008685) and the National Science Foundation (NSF PHY 2033046 and NSF GRFP 1745302) and NIH Center grant P30-ES002109.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

Sequence data that support the findings of this study are deposited in the GISAID database with accession numbers in the range of EPI_ISL_403962 to EPI_ISL_996048 (https://www.gisaid.org/).

ORCID

Georg Hahn http://orcid.org/0000-0001-6008-2720

REFERENCES

Bae, J.-Y., Lee, I., Kim, J. I., Park, S., Yoo, K., Park, M., Kim, G., Park, M. S., Lee, J.-Y., Kang, C., Kim, K., & Park, M.-S. (2018). A single amino acid in the polymerase acidic protein determines the pathogenicity of influenza B viruses. *Journal of Virology*, *92*(13), e00259-18.

Belouzard, S., Chu, V. C., & Whittaker, G. R. (2009). Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 5871–5876.

Brault, A. C., Huang, C., Langevin, S. A., Kinney, R. M., Bowen, R. A., Ramey, W. N., Panella, N. A., Holmes, E. C.,

Powers, A. M., & Miller, B. R. (2007). A single positively selected West Nile viral mutation confers increased virogenesis in American crows. *Nature Genetics*, *39*, 1162–1166.

Centers for Disease Control and Prevention. (2021). *SARS-CoV-2 variant classifications and definitions*. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html

Coutinho, R. M., Marquitti, F. M. D., Ferreira, L. S., Borges, M. E., da Silva, R. L. P., Canton, O., Portella, T. P., Lyra, S. P., Franco, C., da Silva, A. A. M., Kraenkel, R. A., de Sousa Mascena Veras, M. A., & Prado, P. I. (2021). Model-based evaluation of transmissibility and reinfection for the P.1 variant of the SARS-CoV-2. *MedRxiv*. https://doi.org/10.1101/2021.03.03.21252706

Elbe, S., & Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges, 1*, 33–46.

Faria N. R., Mellan T. A., Whittaker C., Claro I. M., Candido D. da S., Mishra S., Crispim M. A. E., Sales F. C. S., Hawryluk I., McCrone J. T., Hulswit R. J. G., Franco L. A. M., Ramundo M. S., de Jesus J. G., Andrade P. S., Coletti T. M., Ferreira G. M., Silva C. A. M., … Sabino E. C. (2021). Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, *372*(6544), 815–821. https://doi.org/10.1126/science.abh2644

Forster, P., Forster, L., Renfrew, C., & Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9241–9243.

Geoghegan, J. L., & Holmes, E. C. (2018). The phylogenomics of evolving virus virulence. *Nature Reviews Genetics*, *19*, 756–769.

Gómez, C. E., Perdiguero, B., & Esteban, M. (2021). Emerging SARS-CoV-2 variants and impact in global vaccination programs against SARS-CoV-2/COVID-19. *Vaccines (Basel)*, *9*(3), 243.

Hahn, G., Lee, S., Weiss, S. T., & Lange, C. (2020). Unsupervised cluster analysis of SARS-CoV-2 genomes reflects its geographic progression and identifies distinct genetic subgroups of SARS-CoV-2 virus. *Genetic Epidemiology*, *45*(3), 316–323.

Hahn, G., Lutz, S. M., Hecker, J., Prokopenko, D., Cho, M. H., Silverman, E., Weiss, S. T., & Lange, C. (2020). locstra: Fast analysis of regional/global stratification in whole genome sequencing (WGS) studies. *Genetic Epidemiology*, *45*(1), 82–98.

Hahn, G., Lutz, S. M., & Lange, C. (2020). *locStra: Fast implementation of (local) population stratification methods (v1.3)*. https://cran.r-project.org/package=locStra

Hoffmann, M., Arora, P., Groß, R., Seidel, A., Hörnich, B. F., Hahn, A. S., Krüger, N., Graichen, L., Hofmann-Winkler, H., Kempf, A., Winkler, M. S., Schulz, S., Jäck, H.-M., Jahrsdörfer, B., Schrezenmeier, H., Müller, M., Kleger, A., Münch, J., & Pöhlmann, S. (2021). SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. *Cell*, *184*(9), 2384–2393. https://doi.org/10.1016/j.cell.2021.03.036

Hoffmann, M., Kleine-Weber, H., & Pöhlmann, S. (2020). A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular Cell*, *78*(4), 779–784.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société vaudoise des sciences naturelles*, *37*, 547–579.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Lo, S. W., & Jamrozy, D. (2020). Genomics and epidemiological surveillance. *Nature Reviews Microbiology*, *18*, 478.

Long, S. W., Olsen, R. J., Christensen, P. A., Bernard, D. W., Davis, J. J., Shukla, M., Nguyen, M., Saavedra, M. O., Yerramilli, P., Pruitt, L., Subedi, S., Kuo, H.-C., Hendrickson, H., Eskandari, G., Nguyen, H. A. T., Long, J. H., Kumaraswami, M., Goike, J., Boutz, D., … Musser, J. (2020). Molecular architecture of early dissemination and massive second wave of the SARS-CoV-2 virus in a major metropolitan area. *mBio*, *11*, e02707-20. https://doi.org/10.1101/2020.09.22.20199125

Makwana, K. M., & Mahalakshmi, R. (2015). Implications of aromatic-aromatic interactions: From protein structures to peptide models. *Protein Science*, *24*(12), 1920–1933.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, *363*, 166–176.

Mathur, M. B., Smith, L. H., Ding, P., & VanderWeele, T. J. (2021). *EValue: Sensitivity analyses for unmeasured confounding and other biases in observational studies and meta-analyses (v4.1.2)*. https://cran.r-project.org/package=EValue

Nogales, A., Martinez-Sobrido, L., Topham, D. J., & DeDiego, M. L. (2017). NS1 protein amino acid changes D189N and V194I affect interferon responses, thermosensitivity, and virulence of circulating H3N2 human influenza A viruses. *Journal of Virology*, *91*(5), e01930-16.

Prokopenko, D., Hecker, J., Silverman, E., Pagano, M., Nöthen, M., Dina, C., Lange, C., & Fier, H. (2016). Utilizing the Jaccard index to reveal population stratification in sequencing data: A simulation study and an application to the 1000 Genomes Project. *Bioinformatics*, *32*(9), 1366–1372.

Pujadas, E., Chaudhry, F., McBride, R., Richter, F., Zhao, S., Wajnberg, A., Nadkarni, G., Glicksberg, B. S., Houldsworth, J., & Cordon-Cardo, C. (2020). SARS-CoV-2 viral load predicts COVID-19 mortality. *Lancet Respiratory Medicine*, *8*(9), e70.

Schlauch, D., Fier, H., & Lange, C. (2017). Identification of genetic outliers due to sub-structure and cryptic relationships. *Bioinformatics*, *33*(13), 1972–1979.

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data—From vision to reality. *EuroSurveillance*, *22*(13), 30494.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (1st ed.). Pearson

UCSC Genome Browser on SARS-CoV-2. (2021). *Substitution V1176F*. https://genome.ucsc.edu/cgi-bin/hgTracks?db=wuhCor1%26lastVirtModeType=default%26lastVirtModeExtraState=%26virtModeType=default%26virtMode=0%26nonVirtPosition=%26position=NC_045512v2%3A1%2D29903%26hgsid=1105354641_9ijaxoab6aags7jVnBFnnFMgXfdC

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, *167*, 268–274.

Vankadari, N. (2020). Structure of furin protease binding to SARS-CoV-2 spike glycoprotein and implications for potential targets and virulence. *Journal of Physical Chemistry Letters*, *11*(16), 6655–6663.

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, *181*(2), 281–292.

Watanabe, R., Matsuyama, S., Shirato, K., Maejima, M., Fukushi, S., Morikawa, S., & Taguchi, F. (2008). Entry from cell surface of SARS coronavirus with cleaved S protein as revealed by pseudotype virus bearing cleaved S protein. *Journal of Virology*, *82*(23), 11985–11991.

Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., Qi, F., Bao, L., Du, L., Liu, S., Qin, C., Sun, F., Shi, Z., Zhu, Y., Jiang, S., & Lu, L. (2020). Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Research*, *30*(4), 343–355.

Yoshimoto, F. K. (2020). The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. *Protein Journal*, *39*(3), 198–216.

How to cite this article: Hahn, G., Wu, C. M., Lee, S., Lutz, S. M., Khurana, S., Baden, L. R., Haneuse, S., Qiao, D., Hecker, J., DeMeo, D. L., Tanzi, R. E., Choudhary, M. C., Etemad, B., Mohammadi, A., Esmaeilzadeh, E., Cho, M. H., Li, J. Z., Randolph, A. G., Laird, N. M., ... Lange, C. (2021). Genome-wide association analysis of COVID-19 mortality risk in SARS-CoV-2 genomes identifies mutation in the SARS-CoV-2 spike protein that colocalizes with P.1 of the Brazilian strain. *Genetic Epidemiology*, 45, 685–693. https://doi.org/10.1002/gepi.22421