

# SARS-CoV-2 genetic variations associated with COVID-19 pathogenicity

Pakorn Aiewsakun<sup>1,2,\*</sup>, Patrawee Nilplub<sup>2</sup>, Patompon Wongtrakoongate<sup>3,4</sup>, Suradej Hongeng<sup>5</sup> and Arunee Thitithanyanont<sup>1,2,\*</sup>

## Abstract

In this study, we performed genome-wide association analyses on SARS-CoV-2 genomes to identify genetic mutations associated with pre-symptomatic/asymptomatic COVID-19 cases. Various potential covariates and confounding factors of COVID-19 severity, including patient age, gender and country, as well as virus phylogenetic relatedness were adjusted for. In total, 3021 full-length genomes of SARS-CoV-2 generated from original clinical samples and whose patient status could be determined conclusively as either 'pre-symptomatic/asymptomatic' or 'symptomatic' were retrieved from the GISAID database. We found that the mutation 11 083G>T, located in the coding region of non-structural protein 6, is significantly associated with asymptomatic COVID-19. Patient age is positively correlated with symptomatic infection, while gender is not significantly correlated with the development of the disease. We also found that the effects of the mutation, patient age and gender do not vary significantly among countries, although each country appears to have varying baseline chances of COVID-19 symptom development.

## DATA SUMMARY

All sequence data used in this study were retrieved from the GISAID database. All supplementary information is available with the article as PDF files online. Supplementary material can be found in Figshare: <https://doi.org/10.6084/m9.figshare.16528950.v1> [1].

## INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19), was first reported in Wuhan, Hubei, China, in late December 2019 [2]. SARS-CoV-2 is a positive-sense ssRNA virus in the family *Coronaviridae* [3]. It is the seventh known coronavirus capable of infecting humans, after HCoV-229E, HCoV-OC43, HCoV NL63, HKU1, MERS-CoV and the original SARS-CoV. The first four typically cause non-lethal mild upper respiratory diseases, while the last two,

as well as SARS-CoV-2, can cause severe lethal respiratory illnesses [4, 5].

The virus was found to spread around the globe, and as of now (October 2021) has infected more than 240 million people globally [6], which has overwhelmed hospitals in many countries. Although the case-fatality ratio of SARS-CoV-2 (~1.4–2.29% [6, 7]) is lower than those of SARS-CoV (11%) [8, 9] and MERS-CoV (34–37%) [10], due to the greater number of infected cases, the number of deaths caused by SARS-CoV-2 is much greater than those caused by SARS-CoV and MERS-CoV. To date, at least 4.9 million deaths have been reported to be associated with SARS-CoV-2 infection [6]. A systematic review [11] showed that the serial interval time of COVID-19 (i.e. the time between illness onset in the primary case to the onset in the secondary case) is ~5.2 (mean range from 23 studies: 4.2–7.5) days while the incubation period (i.e. the time from infection to the onset of illness) is ~6.5 (mean range from 14 studies: 4.8–9) days. This shorter serial interval time

Received 09 June 2021; Accepted 28 October 2021; Published 06 December 2021

**Author affiliations:** <sup>1</sup>Department of Microbiology, Faculty of Science, Mahidol University, 272, Rama VI Road, Ratchathewi, Bangkok, 10400, Thailand; <sup>2</sup>Pornchai Matangasombut Center for Microbial Genomics, Department of Microbiology, Faculty of Science, Mahidol University, 272, Rama VI Road, Ratchathewi, Bangkok, 10400, Thailand; <sup>3</sup>Department of Biochemistry, Faculty of Science, Mahidol University, 272, Rama VI Road, Ratchathewi, Bangkok, 10400, Thailand; <sup>4</sup>Center for Neuroscience, Faculty of Science, Mahidol University, 272, Rama VI Road, Ratchathewi, Bangkok, 10400, Thailand; <sup>5</sup>Division of Hematology and Oncology, Department of Pediatrics, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, 270, Rama VI Road, Ratchathewi, Bangkok, 10400, Thailand.

\***Correspondence:** Pakorn Aiewsakun, pakorn.aie@mahidol.ac.th; Arunee Thitithanyanont, arunee.thi@mahidol.ac.th

**Keywords:** SARS-CoV-2; GWAS; nsp6.

**Abbreviations:** COVID-19, coronavirus disease 2019; GWAS, genome-wide association study; NSP6, non-structural protein 6; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

**Data statement:** All supporting data, code and protocols have been provided within the article or through supplementary data files. Seven supplementary tables, one supplementary figure and supplementary data are available with the online version of this article.

000734 © 2021 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License.

suggests a substantial role of pre-symptomatic/asymptomatic transmission [11]. Mathematical model analyses suggested that more than half of all transmissions may be attributable to pre-symptomatic/asymptomatic transmission [12]. In order to effectively control the virus, it is thus important to understand the factors that underlie the disease severity and pathogenesis.

Several host factors correlated with COVID-19 severity have been identified, including patient age [13–15], patient gender [15–18], pre-existing medical conditions such as diabetes and chronic liver disease [15, 19], levels of CD4 + and CD8+ T cell counts, levels of IL-6 and IL-8 [20], and genotypes of *human leucocyte antigen* genes [21]. A number of viral genetic factors associated with the increase in COVID-19 severity have also been characterized. For example, nucleotide mutations 14 408C>T (i.e. the mutation P323L in RNA-dependent RNA polymerase protein) and 23 403A>G (i.e. the mutation D614G in the virus spike protein), which tended to be found together, were reported to show significant positive correlations with death [21], and to be found more frequently in severe cases than mild cases [22]. The latter mutation was also reported to increase infectivity of the virus in multiple human cell types [23]. In addition, a study has shown that ORF3a mutations are associated with higher infection and mortality rates of SARS-CoV-2 [24]. However, viral mutations associated with pre-symptomatic/asymptomatic SARS-CoV-2 infection are still poorly characterized.

There are currently millions of SARS-CoV-2 genomes from around the world made publicly available on the database of the Global Initiative on Sharing All Influenza Data (GISAID) [25], and some of which have patient status made publicly available. While most of the sequences were from symptomatic cases, some were from asymptomatic cases, presenting us an opportunity to examine viral genetic factors that might be associated with the virus's pathogenicity on a large scale.

## METHODS

### SARS-CoV-2 genome sequences with patient status

In total, 18 705 full-length/almost full-length genome sequences of SARS-CoV-2 (>29 000 nt) with patient status were downloaded from the GISAID database on 28 December 2020 together with their metadata. A table of acknowledgements can be found in Table S1 (available in the online version of this article), and the metadata can be found in Table S2. To allow for accurate determination of the genetic factors associated with COVID-19 pathogenicity, we only analysed sequences whose patient status could be unambiguously determined as either 'pre-symptomatic/asymptomatic' or 'symptomatic'. We categorized 'Asymptomatic/Released', 'Asymptomatic', 'Asymptomatic, identified as positive during preoperation investigation', 'No clinical signs' and 'No clinical signs without hospitalization' as 'pre-symptomatic/asymptomatic' (257 sequences). If the patient status was 'Acute upper respiratory infection, unspecified', 'Death', 'Deceased', 'Hospitalized (Critical)', 'Hospitalized/Deceased', 'Hospitalized, deceased', 'Hospitalized in ICU', 'Hospitalized (Intensive

### Impact Statement

Coronavirus disease 2019 (COVID-19) is a major public health concern, caused by a novel coronavirus called SARS-CoV-2. As of now, more than 240 million cases of COVID-19 with more than 4.9 million deaths have been reported worldwide, and the rapid surge in the number of COVID-19 patients has overwhelmed hospitals' capacity around the world. It has been reported that up to 80% of SARS-CoV-2 infection might be asymptomatic, and this probably plays an important part in the rapid global transmission of the virus. To effectively contain the disease, it is thus imperative to understand the disease progression and the factors that drive the process. Here, we performed genome-wide association analyses on all SARS-CoV-2 genomes in the GISAID database, and found that a mutation at genomic position 11083, namely the 11083G>T mutation, located in the coding region of non-structural protein 6, is significantly associated with asymptomatic COVID-19, adjusted for various confounders and covariates, including virus phylogenetic relatedness, patient age, gender and country. Our results have implications for the development of better and more informative test kits, for example, allowing potentially symptomatic cases to be distinguished from asymptomatic ones, and this could lead to more effective disease management and control.

care unit), 'Hospitalized (Mild)', 'Hospitalized (Moderate)', 'Hospitalized, oxygenotherapy, diarrhoea', 'Hospitalized (Severe)', 'Hospitalized, ICU, fully recovered', 'ICD-10 Disease: J00-J06 Acute upper respiratory infections', 'ICD-10 Disease: J06.9 Acute upper respiratory infection, unspecified', 'ICD-10 Disease: J18.1 Lobar pneumonia, unspecified organism', 'ICD-10 Disease: J18.9 Pneumonia, unspecified organism', 'ICU', 'ICU; Serious', 'Intensive Care Unit', 'Live, mild symptoms, at home', 'Mild', 'Mild, at home', 'Mild case', 'Mild clinical signs without hospitalization', 'Mild clinical signs without hospitalization. Diarrhoea', 'Mild clinical signs without hospitalization. Distorted ability to smell', 'Mild clinical signs without hospitalization. Distorted ability to smell and taste', 'Mild clinical signs without hospitalization. Distorted ability to taste', 'Mild symptoms (fever, cardiovascular disorders)', 'Mild symptoms inpatient for observation', 'Moderate/Outpatient', 'Paucisymptomatic', 'Severe/ICU', 'Severe', 'Symptomatic' or 'Symptoms indicative of upper respiratory infection', we annotated them as 'symptomatic' (10 172 sequences). Sequences that were not generated from original clinical samples or human samples were excluded (three asymptomatic samples and 10 symptomatic samples), leaving 10 416 sequences in the whole dataset. Virus genomes associated with symptomatic Japanese cases were subsampled to reduce the data redundancy (see main text). Fifteen sequences with unusually high sequence diversity were also removed. In total, our dataset comprised 3021 sequences,

252 of which were associated with pre-symptomatic/asymptomatic infections, while 2769 sequences were associated with symptomatic cases. Taxonomic and geographical distributions of the sequences can be found in Tables S3 and S4, respectively.

### Phylogenetic reconstruction

A manually curated full-length multiple sequence alignment of the 3021 SARS-CoV-2 genomes was constructed. Potential recombination within the alignment was checked by using the Phi test implemented in SplitsTree4 [26], but no evidence was found ( $P=0.91$ ). A maximum-likelihood phylogeny was estimated by using IQ-TREE [27]. ModelFinder [28] determined the general time reversible model (GTR) with empirical base frequencies (+F) and the 5-discrete-rate-category FreeRate model (+R5) to be the best-fit nucleotide substitution model under the Bayesian information criterion, and was used for tree reconstruction. Bootstrap clade support was computed based on 1000 pseudoreplicate datasets with ultrafast bootstrap approximation. The terminal branch leading to sample EPI\_ISL\_407976 was determined as a suitable root placement by comparing the estimated tree with the global audacity tree from the GISAID database. The maximum-likelihood tree in Newick format can be found in Data S1.

### Genetic variation detection

Polymorphic sites with less than 50% ambiguous bases and aggregated minor allele frequencies of more than 5% of the collected sequences were identified. For initial screening of candidate sites, genetic variations present in the reference genome (RefSeq accession number: NC\_045512.2) were considered 'reference' variations, or otherwise as 'non-reference' variations, and for each position, a binary profile of genetic variations was constructed. We believed this was reasonable as all positions had only one major alternative variant, with the frequencies of the third most common allele being less than 1% for all sites except one (Table 1). Uncertainty coefficients were computed for all pairs of positions using the *UncertCoeff* function in R (Table S5). Pairs with scores of  $>0.90$  (out of 1) were grouped together as sites with co-varying variants, and were analysed as strongly linked loci. Likewise, for each set of strongly linked loci, a sequence was considered to be a 'reference' variant if all of its genetic variations on the sites were those present in the reference genome, and otherwise as a 'non-reference' variant.

### Phylogenetic-based approach to GWAS

TreeWAS [29] was used for initial screening of SARS-CoV-2 genetic variations that might be associated with COVID-19 pathogenicity, defined as two discrete traits: 'symptomatic' and 'pre-symptomatic/asymptomatic', conditioning on the estimated virus phylogeny. Three separate tests of association were performed: the 'terminal', 'simultaneous' and 'subsequent' tests. Analysis-wide and separate Bonferroni multiple-testing corrections were applied. Ancestral states of both genetic and phenotypic states were inferred by using the TreeWAS software package [29] under the maximum-parsimony

framework. The analysis was applied to all of the 1000 trees in the bootstrap tree distribution obtained from the phylogenetic analysis described above to account for phylogenetic uncertainty. In each test, 1000 genetic loci were simulated to estimate the null distributions of association scores, fixing the tree topology and the distribution of phenotypic states but reassigning genetic substitutions to new branches. The direction of association was determined based on crude odds ratios, estimated using the *oddsratio.wald* function, implemented in the R software package *epitools*.

### Examination of the undetermined/ambiguous base found at site 11083

Our initial screening suggested that genetic variations at site 11083 are associated with COVID-19 pathogenicity, and examination revealed that 18 sequences had either an undetermined (N) or ambiguous (K) base at the position (see Results). To examine the nature of the genetic variations at this site, we were able to obtain raw sequencing data for 14 of them, and mapped them against the reference genome (NC\_045512.2) using *bwa* [30] and *samtools* [31]. Crude read mapping depths and base/indel calling statistics were computed by using *Integrative Genomics Viewer* [32] (Fig. S1).

### Generalized linear mixed model fitting

To further test the association detected by the initial screening, we fitted various binomial generalized linear mixed models to the data. In addition to the patient status and virus genetic variations, this analysis also incorporated patient age, gender and sequence country of origin, as well as virus phylogenetic structure to the models. The virus phylogenetic structure was simply the variance/co-variance matrix of the estimated phylogeny (Fig. 1), standardized to have a determinant of 1. The effects of the virus genetic mutation, patient age and gender were treated as fixed effects, while the effects of virus phylogenetic structure and patient country were treated as random effects. Four models were examined in total (see Table S6 for model specifications), and were fitted to the data by using the *relmatGLmer* function, implemented in the *lme4qtl* R package [33]. The *anova* function was used to perform the likelihood ratio test to identify the best-fit model. The estimated parameter values can be found in Table S7.

## RESULTS

### SARS-CoV-2 genome sequences with patient status

SARS-CoV-2 genomes with patient status data were retrieved from the GISAID database [25] on 28 December 2020. At the time of the study, 18705 genome sequences were full-length/almost full-length ( $>29000$  nt), and had patient status information (Table S1). The metadata of the sequences can be found in Table S2. Of the 18705 sequences, 10416 sequences were generated from original clinical samples, and had patient statuses that could be determined as either 'pre-symptomatic/asymptomatic' (254 sequences) or 'symptomatic' (10162 sequences). See the Methods section for more details about patient grouping. Further inspection revealed

**Table 1.** SARS-CoV-2 polymorphic sites under the study

Twenty-six polymorphic sites with <50% ambiguous bases and aggregated minor allele frequencies of >5% of the collected sequences were analysed in our analyses. Sites with co-occurring variants, i.e. those with pairwise uncertainty coefficients of >0.90 (Table S5), were grouped together for analysis.

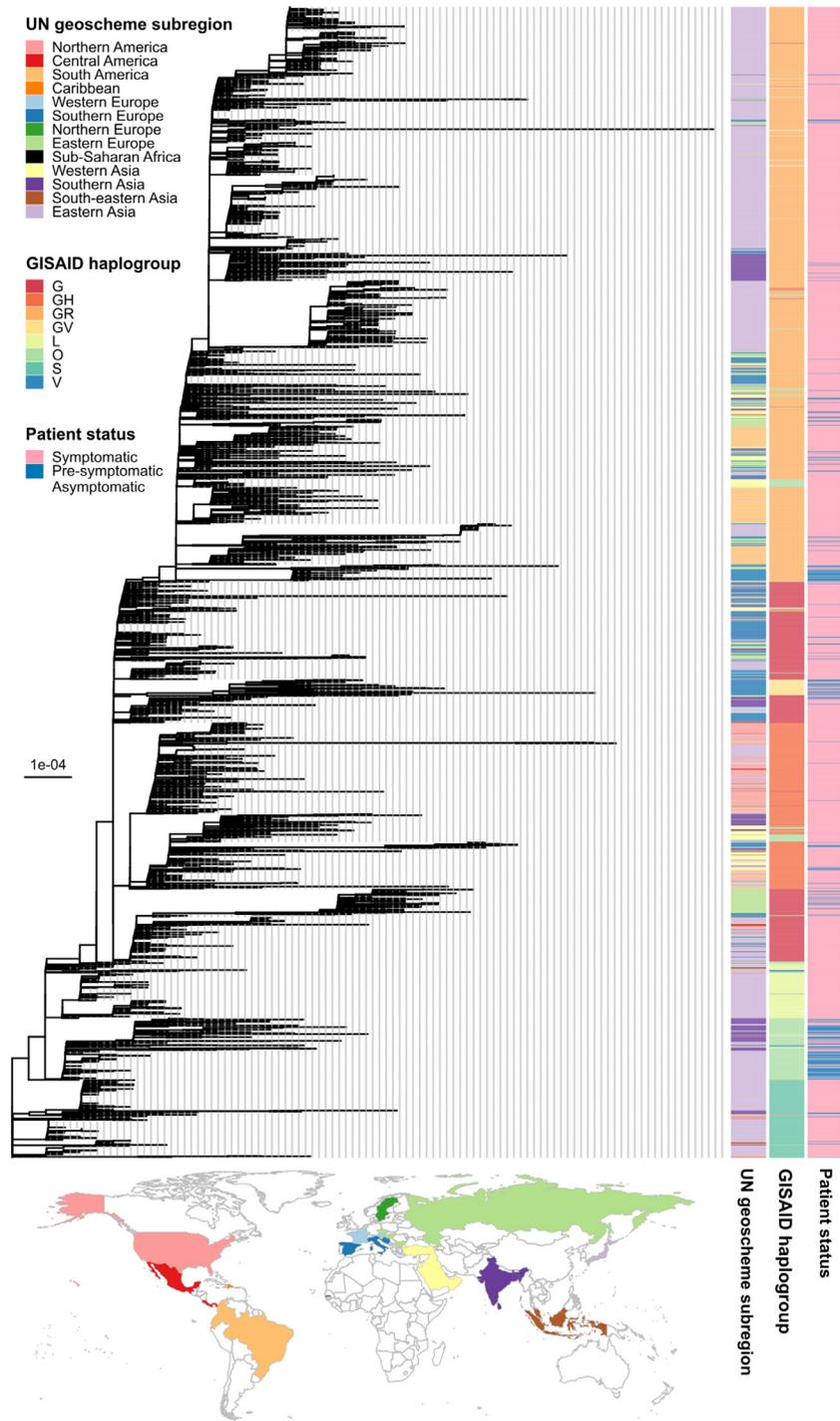
Nucleotide position*	Nucleotide grouping*	Reference variant [frequency (%)]*	Alternative variant [frequency (%)]	Gene location	Substitution type	Amino acid change†
313	313	C(70.18)	T(29.66), N(0.1), -(0.03), Y(0.03)	ORF1ab:NSP1	Synonymous	L16
1059	1059	C(91.43)	T(8.47), Y(0.07), N(0.03)	ORF1ab:NSP2	Non-synonymous	T85I
3037	3037	C(17.71)	T(82.16), N(0.07), Y(0.07)	ORF1ab:NSP3	Synonymous	F106
11083	11083	G(92.62)	T(6.79), N(0.56), K(0.03)	ORF1ab:NSP6	Non-synonymous	L37F
18877	18877	C(94.21)	T(5.76), N(0.03)	ORF1ab:NSP14	Synonymous	L280
20268	20268	A(93.02)	G(6.59), N(0.4)	ORF1ab:NSP15	Synonymous	L216
23403	23403	A(18.21)	G(81.53), N(0.2), R(0.07)	S	Non-synonymous	D614G
25563	25563	G(83.88)	T(14.96), C(1.06), N(0.1)	ORF3a	Non-synonymous	Q57H
26730	26730	G(94.64)	A(5.1), T(0.13), C(0.07), N(0.07)	M	Non-synonymous	V70L, V70E, V70L
26735	26735	C(94.87)	T(5.06), N(0.07)	M	Synonymous	Y71
28975	28975	G(92.98)	T(6.29), C(0.53), N(0.2)	N	Non-synonymous	M234I
29692	29692	G(92.65)	T(5.66), -(1.42), N(0.26)	3'UTR	Non-coding	-
241	241/14 408	C(16.95)	T(82.75), G(0.13), A(0.07), Y(0.07), -(0.03)	5'UTR	Non-coding	-
14408		C(17.28)	T(82.36), Y(0.2), N(0.17)	ORF1ab:RDRP	Non-synonymous	P323L
8782	8782/28 144	C(92.98)	T(6.92), N(0.1)	ORF1ab:NSP4	Synonymous	S76
28144		T(93.25)	C(6.75)	ORF8	Non-synonymous	L84S
18167	18167/21 518	C(94.14)	T(5.79), A(0.07)	ORF1ab:NSP14	Non-synonymous	P43L, P43H
21518		G(93.88)	T(5.83), N(0.3)	ORF1ab:NSP16	Non-synonymous	R287I
28881	28881/28 882/28 883	G(49.75)	A(49.85), N(0.3), R(0.1)	N	Non-synonymous	R203K, K204R
28882		G(49.95)	A(49.65), N(0.26), R(0.1), T(0.03)			
28883		G(50.05)	C(49.59), N(0.26), S(0.1)			
4346	4346/9286 /10 376/14 708/28 725	T(94.24)	C(5.76)	ORF1ab:NSP3	Non-synonymous	S543P
9286		C(94.11)	T(5.79), N(0.1)	ORF1ab:NSP4	Synonymous	N244
10376		C(93.88)	T(6.06), N(0.07)	ORF1ab:NSP5	Non-synonymous	P108S
14708		C(94.17)	T(5.79), N(0.03)	ORF1ab:RDRP	Non-synonymous	A423V
28725		C(94.11)	T(5.86), N(0.03)	N	Non-synonymous	P151L

\*With respect to the reference SARS-CoV-2 genome (RefSeq accession number: NC\_045512.2).

†Reported for non-ambiguous base changes only.

that the majority of the sequences associated with symptomatic infections were from Japan (8776/10 162=86.36%), and they were highly redundant – according to the classification scheme described previously [34] and the GISAID haplogroup assignment; 8136/10 162 (80.06%) of all symptomatic sequences comprised just five distinct lineages sampled from Japan, namely B.1.1.284 haplogroup GR ( $N=4005$ , 39.41%), B.1.1 haplogroup GR ( $N=2612$ , 25.70%), B.1.1.214 haplogroup GR ( $N=1020$ , 10.04%), B.1.1.48 haplogroup GR

( $N=343$ , 3.38%) and A haplogroup S ( $N=150$ , 1.48%). We thus randomly subsampled the Japanese data so that there were at most 150 sequences from each of these lineages. Preliminary phylogenetic analysis showed that 15 sequences had unusually long terminal branches and had unusually high sequence diversities ( $>9 \times 10^{-4}$  mutations per site) compared to the reference SARS-CoV-2 genome (RefSeq accession number: NC\_045512.2), probably due to primer contamination. Thirteen were associated with symptomatic cases, while two



**Fig. 1.** SARS-CoV-2 phylogeny. The tree was estimated under the maximum-likelihood framework implemented in IQ-TREE2 [27] based on a manually curated alignment of 3021 full-length SARS-CoV-2 genomes. Potential recombination within the alignment was checked by using the Phi test implemented in SplitsTree4 [26], but no evidence was found ( $P=0.91$ ). The best-fit nucleotide substitution model was determined to be GTR+F+R5 (the general time reversible model +empirical base frequencies+the 5-discrete-rate-category FreeRate model) by ModelFinder [28] under the Bayesian information criterion and was used for tree reconstruction. We compared our tree with the global tree obtained from GISAID, and determined the terminal branch leading to sample EPI\_ISL\_407976 as a suitable location for root placement. Bar, substitutions per site. The tree file in Newick format with bootstrap clade-support values, computed based on 1000 bootstrap trees, can be found in Data S1. The three columns on the right indicate the United Nations (UN) geoscheme subregion, the GISAID haplogroup assignment and the patient status of the sequences, respectively (see keys). The World map below the tree shows the countries from which the sequences were sampled, colored according to the UN geoscheme subregions.

were obtained from pre-symptomatic/asymptomatic cases, and they were removed from the dataset. In total, our curated dataset comprised 3021 sequences, 252 of which were associated with pre-symptomatic/asymptomatic infections while 2769 sequences were associated with symptomatic infections.

The curated dataset covered a wide diversity of SARS-CoV-2, comprising 148 distinct viral lineages and eight GISAID haplogroups (Table S3) sampled from 36 countries/territories and 13 United Nations (UN) geoscheme subregions (Table S4). Fig. 1 shows a maximum-likelihood phylogeny estimated from the SARS-CoV-2 genomes collected. We found that sequences of different geographical origins often clustered together phylogenetically, indicating substantial cross-country and cross-continent transmissions. Viruses of the same GISAID haplogroups tended to form distinct and tight clusters of sequences in the tree, with the exception of haplogroup G. Haplogroup G was estimated to be a paraphyletic group basal to haplogroups GR, GV and GH. This phylogenetic pattern is expected however, since the haplotype defining haplogroup G is parental/a subset of those defining haplogroups GR, GV and GH [35].

The distribution of SARS-CoV-2 genomes associated with pre-symptomatic/asymptomatic and symptomatic infections were significantly different among lineages ( $\chi^2$  test:  $\chi^2$  score=1094.3, degrees of freedom (d.f.)=147,  $P<2.2\times 10^{-16}$ ) and GISAID haplogroups ( $\chi^2$  test:  $\chi^2$  score=499.22, d.f.=7,  $P<2.2\times 10^{-16}$ ) as well as among countries/territories ( $\chi^2$  test:  $\chi^2$  score=406.33, d.f.=35,  $P<2.2\times 10^{-16}$ ) and broader geographical regions ( $\chi^2$  test:  $\chi^2$  score=230.17, d.f.=12,  $P<2.2\times 10^{-16}$ ). The majority of the genomes associated with pre-symptomatic/asymptomatic cases were from the Czech Republic (34/252=13.49%), India (52/252=20.63%), Italy (69/252=27.38%), Japan (61/252=24.21%) and Turkey (21/252=8.33%), while genomes associated with symptomatic cases were from around the globe (Table S4).

### Identification of polymorphic sites

Twenty-six polymorphic sites with less than 50% ambiguous bases and aggregated minor allele frequencies of more than 5% of the collected sequences were considered in our analyses, corresponding to sites 241, 313, 1059, 3037, 4346, 8782, 9286, 10376, 11083, 14408, 14708, 18167, 18877, 20268, 21518, 23403, 25563, 26730, 26735, 28144, 28725, 28881, 28882, 28883, 28975 and 29692 in the reference SARS-CoV-2 genome (RefSeq accession number: NC\_045512.2) (Table 1). Of these 26 sites, two were in non-coding regions, seven harbored synonymous changes and the rest harbored non-synonymous changes (Table 1). Analyses have shown that synonymous changes in viruses might not be neutral [36, 37], and since there are many ways that a pathogen and its host could interact, including through protein–protein interactions, protein–nucleotide interactions or even nucleotide–nucleotide interactions, to be as inclusive and as hypothesis-free as possible, we included all of these sites in our analyses.

For each position, a binary profile of genetic variations (reference variation vs. non-reference variation) was constructed, and pairwise uncertainty coefficients ( $U$ ) were computed

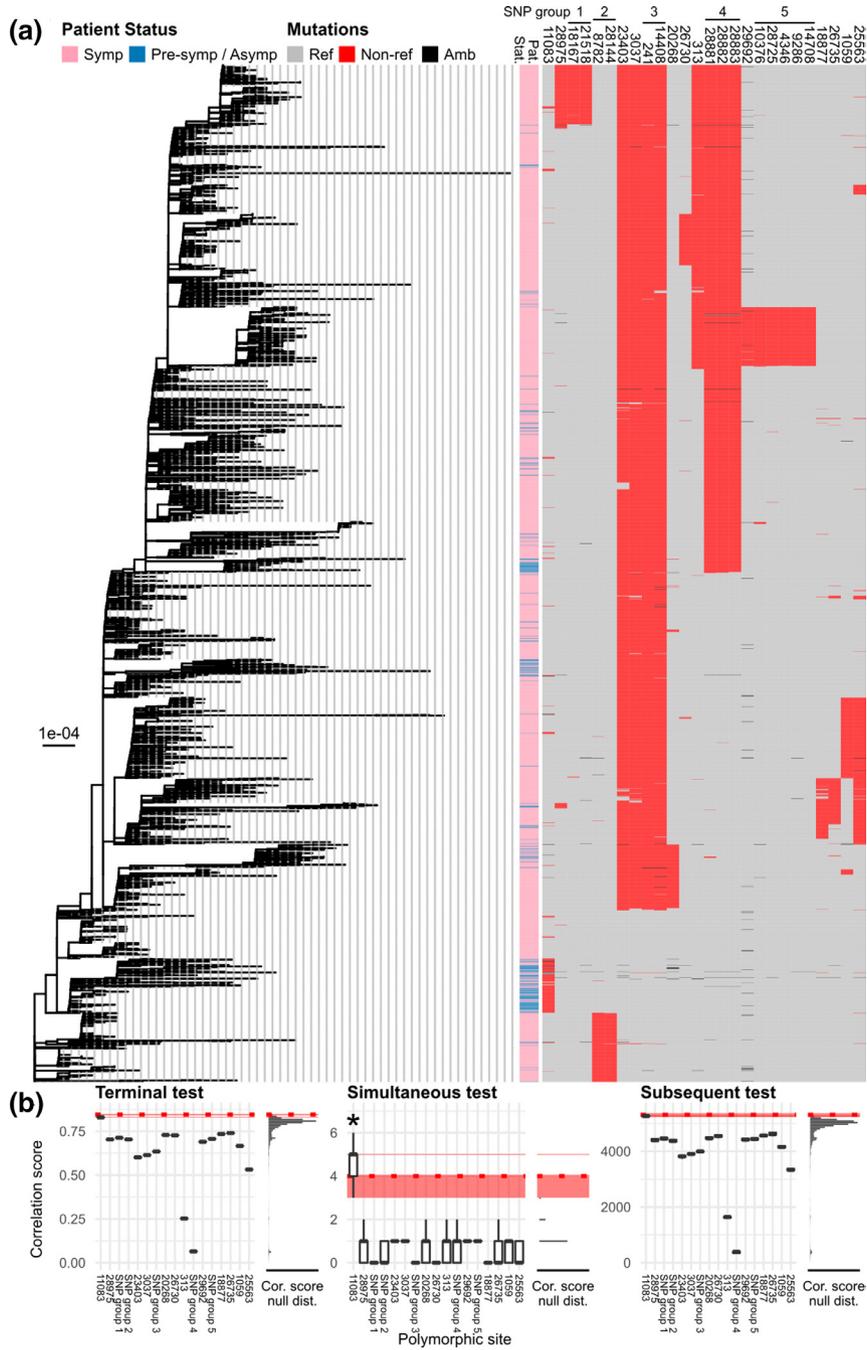
(Table S5). We considered pairs with  $U>0.90$  (out of 1) to be sites with co-occurring variations. Five sets of co-varying sites were detected under this criterion, namely (i) sites 241 and 14408, (ii) sites 8782 and 28144, (iii) sites 18167 and 21518, (iv) sites 28881, 28882 and 28883, and (v) sites 4346, 9286, 10376 14708 and 28725. Indeed, genetic variations of these five sets of sites were phylogenetically highly correlated, as could be seen pictorially in Fig. 2a, and they were thus analysed together in subsequent analyses as strongly linked sets of loci. Binary profiles of these linked loci were constructed in a similar manner (reference variant vs. non-reference variant). In total, 17 sets of polymorphic sites were analysed.

### Initial screening for candidate polymorphic sites with genetic variations associated with COVID-19 pathogenicity

To identify genetic mutations associated with COVID-19 pathogenicity, we first screened for potential candidates by using a phylogenetic-based approach implemented in TreeWAS [29], which was designed specifically to accommodate the great diversity of virus (and bacterial) genomes.

Three measurements of association between the observed patient status and virus genotype were computed: ‘terminal score’, ‘simultaneous score’ and ‘subsequent score’. The ‘terminal score’ indicates the degree of sample-wide association across the phylogeny’s leaves, the ‘simultaneous score’ is the number of times that the examined phenotype and genotype change in parallel together, and the ‘subsequent score’ measures the total tree length that the genotype and investigated phenotype are found to co-exist [29]. Ancestral states of both genetic and phenotypic data were inferred by using TreeWAS [29] under the maximum-parsimony principle. For each test, Bonferroni multiple-testing correction was applied (adjusted  $P$  value threshold=5%/17 sets of polymorphic sites analysed=0.294%). Three-test wide Bonferroni multiple-testing correction was also applied (adjusted  $P$  value threshold=5%/17/3=0.098%). To account for phylogenetic uncertainty, the analysis was applied to the entire distribution of the 1000 bootstrap trees.

Our results showed that site 11083 had the highest association scores in all of the three tests (Fig. 2b). Conditioning on the maximum-likelihood tree (Fig. 2a), and under both analysis-wide and separate Bonferroni multiple-testing corrections, the simultaneous test detected the genomic position 11083 as the only site with variations significantly associated with COVID-19 patient status, while terminal and subsequent tests did not detect any significant signals, although the scores for site 11083 were very close to the thresholds. Analyses of bootstrap trees showed that site 11083 was detected as positive by the simultaneous test 37.6 and 58.5% of the times under the analysis-wide and separate Bonferroni multiple-testing correction, respectively. These relatively low bootstrap positive rates indicated substantial phylogenetic uncertainty, which perhaps was not too surprising given that the sequences were sampled from the same pandemic and were quite similar. These rates, nonetheless, were still much greater than expected had the



**Fig. 2.** Screening for candidate sites with genetic variations associated with COVID-19 pathogenicity by using TreeWAS [29]. (a) Maximum-likelihood tree of SARS-CoV-2 (as shown in Fig. 1) is shown on the left. Bar, substitutions per site. The viruses' patient status (pat. stat.) and mutational profiles of the 26 polymorphic sites investigated (Table 1) are shown on the right (see keys for details). Sites determined as strongly linked loci are indicated with black horizontal bars and numbers on the top. (b) Three separate tests of genotype-phenotype association implemented in the software TreeWAS [29] were performed, namely 'Terminal' (left), 'Simultaneous' (middle) and 'Subsequent' tests (right) with Bonferroni multiple-testing correction (adjusted  $P$  value threshold=5%/17 sets of polymorphic sites analysed=0.294%). To account for phylogenetic uncertainty, the tests were applied to the entire distribution of the 1000 bootstrap trees to obtain the distributions of correlation scores and null scores (Cor. score null dist.). The horizontal red strips indicate the 95% highest density intervals of the score cut-offs obtained from the 1000 bootstrap analyses. The horizontal red dotted lines indicate the score cut-off obtained from the maximum-likelihood tree analysis. All tests revealed that site 11 083 had the highest scores (horizontal red solid lines). Simultaneous tests suggested that site 11 083 was the only site with genetic variations significantly associated with COVID-19 patient status (marked with an asterisk, positive bootstrap testing rate=58.5%), while the other two tests did not detect significant signals.

**Table 2.** GISAID sequences with ambiguous bases at position 11 083

The sequence NC\_045512.2 is included as the reference sequence of the original strain. Ambiguous bases are in bold type. The base at position 11 083 is underlined.

Accession no.	Country	Lineage – GISAID haplogroup	Patient status	Sequence (11073–11093)
<i>Reference sequence</i>				
NC_045512.2	China	B – L	Symptomatic	TCTTTTTTTT <u>G</u> TATGAAAATG
<i>GISAID sequences with an ambiguous base at position 11 083</i>				
EPI_ISL_454602	Croatia	B.1.1 – GR	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_539777	Czech Republic	B.1 – GH	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>K</u> TATGAAAATG
EPI_ISL_626570	Czech Republic	B.1 – GH	Symptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_626613	Czech Republic	B.1.258 – G	Symptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_437454	India	B.6 – O	Symptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_479520	India	B.6 – O	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_436137	India	B.6 – O	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_436140	India	B.1.80 – G	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_436141	India	B.1.80 – G	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_436156	India	B.6 – O	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_436157	India	B.6 – O	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_486386	India	B.6 – O	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_486394	India	B.6 – O	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_486403	India	B.6 – O	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_486407	India	B.1 – S	Pre-symptomatic/Asymptomatic	TCTTTTTTTT <u>N</u> TATGAAAATG
EPI_ISL_486384	India	B.1 – O	Symptomatic	TCT <u>NNNNNNNNNNNNNNNNNNNN</u>
EPI_ISL_447776	Colombia	B.1 – GH	Symptomatic	<u>NNNNNNNNNNNNNNNNNNNN</u>
EPI_ISL_447801	Colombia	B.1.5 – G	Symptomatic	<u>NNNNNNNNNNNNNNNNNNNN</u>

scores been truly randomly sampled from the null distribution, in which there was no phylogenetic signal or genetic–phenotype association in the sequence data at all (0.098 and 0.294% per site per test, respectively). Biologically, these results support that no single site with genetic variations could fully explain the distinctions between pre-symptomatic/asymptomatic and symptomatic SARS-CoV-2 infections, but mutations at site 11 083 may cause changes in the disease symptoms through some complex complementary pathways involving many host and virus components.

### Genetic variations at site 11083 and the direction of association

Two major nucleotide variations were observed at genomic position 11 083, namely thymine (T11 083, 205/3021=6.78%) and guanine (G11 083, 2798/3021=92.62%). The reference genome (RefSeq accession number: NC\_045512.2), which is one of the earliest genomes from the COVID-19 outbreak in Wuhan, was identified as a G11 083 variant (Table 1). One of the closest relatives of SAR-CoV-2 currently known is the bat coronavirus RaTG13 (GenBank accession number:

MN996532.2) [28], and it also has a G at the homologous position, suggesting that the G11 083 variant is the wild-type/ancestral state while the T11 083 variant is a mutant variant. The rest (18 sequences, 18/3021=0.60%) had undetermined nucleotides at this site (Table 2).

Across the 2769 sequences associated with symptomatic cases, 95.88% (2655 sequences) were found to be the wild-type G11083 variants, while only 3.90% (108 sequences) were found to be the mutant T11083 variants. By contrast, of the 252 sequences associated with pre-symptomatic/asymptomatic infections, only 56.75% (143 sequences) were found to be the wild-type G11083 variants, while 38.49% (97 sequences) were the T11083 mutant variants. Based on this dataset, the crude odds ratio for causing a symptomatic infection of the mutant T11083 variant compared to the reference G11083 variant was estimated to be 0.060 by the Wald method (95% confidence interval=0.043–0.083; Fisher's exact test  $P$ -value= $2.11 \times 10^{-58}$ ). The association was robust to the inclusion of the Japanese sequences excluded from the dataset due to data redundancy (crude odds ratio=0.037;

95% confidence interval=0.028–0.050; Fisher's exact test  $P$ -value= $1.52 \times 10^{-82}$ ). Altogether, these results suggested that the 11 083G>T mutation is significantly associated with pre-symptomatic/asymptomatic infection.

Regarding the 18 sequences with undetermined or ambiguous bases at position 11 083 (Table 2), further inspection revealed that three of them had a very long stretch of 'N's spanning across the position, suggesting genuine sequencing failures. Fourteen sequences, however, either had only one or two N bases placed precisely at, or spanning across, site 11 083 amidst a long stretch of the T homopolymer, and one sequence had a K ambiguous base (either T or G) at the position. Among these 15 sequences, we found that most of them were from pre-symptomatic/asymptomatic cases (12/15=80%), and remarkably, they appeared to have an even stronger association with pre-symptomatic/asymptomatic infections compared to the T11 083 variant (crude odds ratio=0.225; 95% confidence interval=0.062–0.819; Fisher's exact test  $P$ -value=0.016). In addition, most of these sequences belonged to lineage B.6 (8/15=53.33%), which is the lineage that has a very high ratio of T11 083 to G11 083 variants (47:3 sequences).

It is well known that a long homopolymer stretch is difficult to sequence, and is prone to indel sequencing errors [38, 39], in particular deletions [40]. Coupled with the fact that site 11 083 is the only polymorphic site amidst the long homopolymer stretch, if such a deletion error were to occur in a T11 083 variant, it would be natural that a sequence aligner would place the gaps spanning the position to maximize the alignment score. Some of these sequences might also be a result of mixed viral populations – indeed, one of these sequences was reported to have an ambiguous base K at site 11 083, indicating that the original sample contained both the T11 083 and G11 083 variants. Alternatively, some of them might genuinely be the G11 083 (or other) variant, but sequencing errors just happened to occur precisely at site 11 083, resulting in undetermined bases. Among these three scenarios, we believed that the last scenario was the most unlikely.

Of these 18 sequences, we were able to obtain raw sequence data for 14, and mapped them against the reference genome to examine the underlying nature of the genetic variations at this site (Fig. S1). We were able to confirm that, among the three sequences with the long stretches of Ns, one was indeed a result of complete sequencing failure, and for the rest, it was often the case that majority of the raw sequence data (>50%) would support a deletion at the site. However, considering the remaining reads that resulted in base calling, all of them were either Ts or Ks, and none were Gs. Combining this information, we therefore deemed reasonable to group the remaining 15 sequences with other T11 083 variants for further analyses.

### Generalized linear mixed model-based method

One of the limitations of TreeWAS [29] was that, while it directly uses a pathogen's phylogenetic tree to compute genotype–phenotype association scores, it could not explicitly account for multiple host and virus covariates and confounding factors at the same time. Patient age, gender

and country data were available from the GISAID database for some of the sampled sequences, allowing us to further test the association detected by TreeWAS [29], accounting for the effects of these factors. Indeed, various studies have suggested that age and gender may be correlated with the disease severity [13–18], and the incidence of COVID-19 does vary from one country to another [6].

To examine if the association detected by TreeWAS [29] was robust to the inclusion of other covariates and confounding factors, we fitted four binomial generalized linear mixed models to the data (Table S6). In all of the four models, the effects of the mutation 11 083G>T, patient gender and age on disease outcome were treated as fixed effects, while the effects of country sampling and virus phylogenetic relatedness were considered random effects. In the most complex model, each individual virus and county was allowed to have varying baseline chances of symptom development while accounting for virus phylogenetic structure, and the effects of the mutation 11 083G>T, gender and age may also vary from country to country (Table S6; M1). In the second most complex model, each individual virus and country may still have varying baseline chances of symptom development adjusted for virus phylogenetic structure, but the effects of the mutation, gender, and age were 'fixed' (i.e. do not vary randomly) across countries (Table S6; M2). In the other two models, one excluded the effect of country sampling (Table S6; M3), and the other excluded the effect of individual virus and phylogenetic structure (Table S6; M4). Sequences with missing patient gender and/or age values were excluded from the model fittings, leaving 1546 sequences in the whole dataset. We also noted that sequence availability varied considerably among countries; most of the countries had fewer than ten sequences in each disease category (Table S4), and this might cause the model fittings and parameter estimates to be unreliable and have unnecessarily high uncertainty. We thus also performed the same analysis on a subsample dataset removing sequences from countries with fewer than ten samples in each disease category. Excluding those with missing patient gender and/or age values, the curated subsample dataset comprised 859 sequences from five countries, namely Czech Republic, India, Italy, Japan and Turkey.

Based on the analysis of the whole dataset, we found that M1 did not significantly better fit the data than M2 (Table S6;  $\chi^2$  test:  $\chi^2$  score=7.85, d.f.=9,  $P=0.55$ ), supporting that the effects of the mutation 11 083G>T, patient gender and age on COVID-19 pathogenicity do not vary significantly among countries. However, M2 was found to fit the data significantly better than M3 (Table S6;  $\chi^2$  test:  $\chi^2$  score=72.84, d.f.=1,  $P<2.2 \times 10^{-16}$ ), suggesting that populations of different counties may have varying baseline chances of developing COVID-19 symptoms. Similarly, M2 was found to better fit than M4 (Table S6;  $\chi^2$  test:  $\chi^2$  score=113.55, d.f.=1,  $P<2.2 \times 10^{-16}$ ), suggesting that phylogenetic relatedness among viruses plays a significant role in COVID-19 pathogenicity (i.e. closely related viruses tend to cause the same kind of infection). Estimated parameter values of the best-fit model can be found in Table S7, and their corresponding adjusted odds ratios are given in Fig. 3.

According to the best-fit model M2 (Fig. 3), we found that patient age is positively associated with symptomatic infection (adjusted odds ratio for increasing an age by 1 year=1.04; 95% confidence interval=1.02–1.05;  $P=7.3\times 10^{-7}$ ), but patient gender does not significantly correlate with the disease outcome (adjusted odds ratio for being female=0.99, 95% confidence interval=0.58–1.69;  $P=0.98$ ). The baseline odds for developing COVID-19 symptoms when getting infected was estimated to be (just) significantly different from 1 ( $P=4.2\times 10^{-2}$ ), computed to be 135.56 with an extremely large uncertainty (95% confidence interval=1.18–15516.82), and the value was estimated to vary substantially among countries (variance=49.66) and also among viruses, although to a much lesser extent (variance=0.38). At face value, this estimate is consistent with the overall baseline chance of developing symptoms being greater than that of not developing symptoms. Nonetheless, it is noteworthy that there might be a tendency for this value to be overestimated, due to the fact that viruses associated with symptomatic infection may be more likely to be sequenced and submitted to the GISAID database than those causing asymptomatic infection given typical healthcare seeking behaviour, making the database systematically biased towards viruses causing symptoms. Thus, this value should be interpreted with care. Nevertheless, taking all of these factors into account, the 11083G>T mutation was still found to be significantly associated with pre-symptomatic/asymptomatic infection (adjusted odds ratio=0.19, 95% confidence interval=0.05–0.75;  $P=1.70\times 10^{-2}$ ).

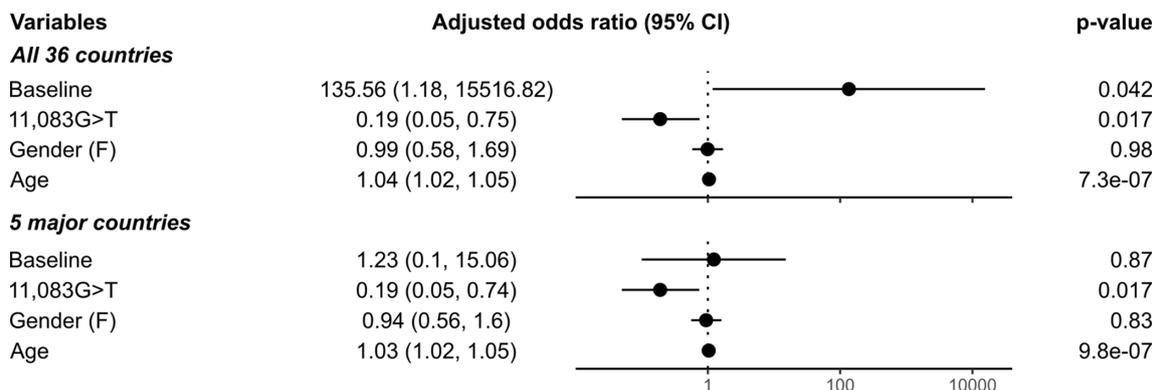
Analysis of the subsample data yielded largely consistent results. M1 was not found to significantly better fit than M2 (Table S6;  $\chi^2$  test: score=12.46, d.f.=9,  $P=0.19$ ), but M2 significantly better fitted the data than M3 (Table S6;  $\chi^2$  test: score=15.49, d.f.=1,  $P=8.31\times 10^{-5}$ ) and M4 (Table S6;  $\chi^2$  test: score=106.12, d.f.=1,  $P<2.2\times 10^{-16}$ ). According to the M2 model (Table S7, and Fig. 3), again, we found that age is positively associated with symptomatic infection (adjusted

odds ratio for increasing an age by 1 year=1.03; 95% confidence interval=1.02–1.05;  $P=9.8\times 10^{-7}$ ), while gender does not significantly correlate with the pathogenicity of the virus (adjusted odds ratio for being female=0.94, 95% confidence interval=0.56–1.60;  $P=0.83$ ). Unlike the results from the whole dataset analysis, the baseline odds was estimated to be insignificantly different from 1 (odds ratio=1.23, 95% confidence interval=0.10–15.06;  $P=0.87$ ), but the analysis still suggested that the value does vary considerably among countries (variance=2.04) and among viruses (variance=0.32). The overall strength of the random country-specific effect was estimated to be less than that estimated based on the whole dataset (variance=49.66), probably due to fewer countries being included in this dataset. Nonetheless, again, a significant (negative) effect of the 11083G>T mutation on SARS-CoV-2 pathogenicity could still be detected (adjusted odds ratio=0.19, 95% confidence interval=0.05–0.74;  $P=1.74\times 10^{-2}$ ).

## DISCUSSION

An unprecedented number of SARS-CoV-2 genomes have been generated at a rapid rate and made publicly available in near real-time. Millions of sequences of SARS-CoV-2 genomes have been made publicly available on the GISAID database [25], and some sequences have patient status available, as well as patient age, gender and country. This allowed us to investigate viral genetic factors that might be associated with SARS-CoV-2 pathogenicity.

In this study, we performed GWAS analyses on 3021 SARS-CoV-2 genomes, and identified variations at genomic position 11083 to be associated with SARS-CoV-2 pathogenicity. Specifically, we found that the mutation 11083G>T is associated with pre-symptomatic/asymptomatic cases (Fig. 3,  $P=1.70$ – $1.74\times 10^{-2}$ ), adjusted for various factors that might also be associated with disease severity, including patient age,



**Fig. 3.** Adjusted odds ratios and 95% confidence intervals of various potential risk factors for COVID-19 symptom development. The values were estimated based on the best-fit binomial generalized linear-mixed model M2, in which the effects of the mutation 11083G>T, patient gender and age on the disease outcome were treated as fixed effects, and the effects of country sampling and virus phylogenetic relatedness were considered random effects. The model allowed each individual virus and country to have varying baseline chances of symptom development while adjusting for the virus phylogenetic structure. See model specification in Table S6 and estimated parameter values in Table S7.

gender and country of origin, as well as virus phylogenetic relatedness and structure.

Our analysis detected a positive correlation between age and COVID-19 symptom development, consistent with results from previous studies [13–15]. A significant association between gender and pathogenicity could not be detected by our study. In contrast, several other studies have detected a gender effect on COVID-19 severity. For example, an analysis of a global survey of hospital records suggested that male patients have higher odds of requiring intensive treatment unit admission and death compared to females [18]. A similar trend was observed in the USA, adjusted for various potential covariates including age, race, ethnicity, marital status, insurance type, median income, body mass index, smoking and other 17 comorbidities [17]. It has been suggested that this could be due to differences in social and cultural behaviours between men and women, and differences in the sex hormone influencing, for example, virus entry and priming, immune and inflammatory response, and coagulation and thrombosis diathesis (see [16] for review). The lack of signal in this study could be because we focused on virus pathogenicity, grouping death, severe and mild cases together as symptomatic. Nonetheless, it is of note that there are also studies that did not find a significant association between gender and COVID-19 severity. For example, Raimondi *et al.* found that gender did not play a significant role as an independent predictor of death in an Italian cohort after adjusting for various confounding factors such as age and severity of the disease at hospital presentation [41]. This suggested that such a trend may vary from country to country.

Paralleling this, we found that patient country is, indeed, a significant random effect, suggesting that the baseline chance of COVID-19 symptom development varies from country to country (although country-specific random effects of the virus mutation, patient age and gender are not significant). In part, this may be because different countries may differ in the way they recorded COVID-19 symptoms and/or how the disease surveillances or treatments were performed, resulting in different distributions of pre-symptomatic/asymptomatic vs. symptomatic cases. This result is also consistent with different human populations having varying degrees of susceptibility to the virus infection. Several studies have revealed that background genetic makeup does, indeed, associate with susceptibility to SARS-CoV-2 infection and COVID-19 severity. For example, a study of two case-control cohorts of Italian and Spanish patients identified a 3p21.31 gene cluster on chromosome 3 to be a genetic susceptibility locus in COVID-19 patients, and showed a higher risk in blood group A than in other blood groups, corresponding to the rs657152 variant at locus 9q34.2 on chromosome 9 [42]. A separate analysis of a dataset from the COVID-19 Host Genetics Initiative confirmed the former region to be significantly associated with COVID-19 severity, and determined that the risk variant was inherited from Neanderthals, carried by around 50% of South Asians and around 16% of Europeans [43]. There are several more host genetic factors that have been suggested as COVID-19 severity predictors,

such as genetic variability in *human leucocyte antigen* genes [21], *major histocompatibility complex class I* genes [44] and *TLR7* gene on chromosome X [45], for example. Different human populations have different genetic backgrounds, and this could potentially partly explain our finding.

A study identified the 11083G>T mutation to be associated with asymptomatic cases by Fisher's exact test ( $P=8.45 \times 10^{-35}$ ) and examination of Pearson's correlation coefficient (Pearson correlation=0.61,  $P=5.42 \times 10^{-56}$ ) [46]; however, the analysis neither accounted for phylogenetic relatedness among viruses nor potential host factors that might also correlate with COVID-19 severity. In this study, we identified the genetic mutation 11083G>T to be associated with pre-symptomatic/asymptomatic cases by using a phylogenetic approach, and after correcting for patient gender, age, country of origin and virus phylogenetic relatedness, the association still remained significant, suggesting that this association is robust. An earlier study of SARS-CoV-2 from a Shanghai cohort identified the T11083 variant to be more prevalent in pre-symptomatic/asymptomatic cases (nine in 91 cases=9.89%) compared to symptomatic cases (one in 21 cases=4.76%), but the association was not significant [20]. This could be due to their relatively small data set ( $N=112$ ), and different groupings of the disease outcome (mild symptomatic and pre-symptomatic/asymptomatic cases vs. severe and critical cases), which could potentially mask the effect we observed. Epidemiological data examination showed that, indeed, countries with a high prevalence of the T11083 variant tended to have lower rates of COVID-19 mortality than those that had low prevalence of the mutation [46], further supporting the association.

This mutation is located in the coding region of non-structural protein 6 (NSP6, nt 10973–11842), coded by ORF1ab (nt 266–21555). NSP6 is a multi-functional protein, and can be found in many coronaviruses. It has been demonstrated that, located to the endoplasmic reticulum, SARS-CoV NSP6 can form a protein complex with NSP3 and NSP4 to induce double-membrane vesicles, crucial for the formation of the virus replication/transcription complex [47]. Studies of an avian gamma-coronavirus, infectious bronchitis virus, showed that NSP6 can activate formation of autophagosomes [48], but ones with small and restricted sizes [49], which may interfere with the host's ability to deliver viral components to lysosomes for degradation. A recent study [50] showed that SARS-CoV-2 NSP6 can bind TANK binding kinase 1, and suppress phosphorylation of IFN regulatory factor 3, which in turn antagonizes type I IFN production. The study also showed that the protein can suppress STAT1 and STAT2 phosphorylation, inhibiting type I IFN signalling, which might mitigate the host immune response even further. In addition, the study found that SARS-CoV NSP6 could not inhibit type I IFN signalling, at least not as efficiently as that of SARS-CoV-2, potentially explaining the relatively more delayed disease onset and less systemic and severe clinical manifestations of COVID-19 compared to SARS.

The 11083G>T mutation confers an amino acid change from leucine (L) to phenylalanine (F) at the 37<sup>th</sup> position

in the NSP6 protein (L37F). *In silico* 3D structure analyses suggested that the mutation may reduce the stability of the protein structure [46, 51], and at the same time may increase the rigidity of the protein, reducing interactions with the endoplasmic reticulum, and interfering with the protein functions in turn [46]. As discussed above, the functions of NSP6 are to antagonize host immune responses, interacting with many virus and host components. Nonetheless, the physical interactions between them are yet to be characterized, and it would be interesting to see if the mutation affects or alters the interactions or not. Our results warrant further experimental confirmations to validate the biological significance of this mutation and its consequences.

The mutation 11 083G>T has occurred independently many times over the course of this on-going pandemic. In fact, this genomic position has been quantified as having one of the highest rates of homoplasmy [52]. Some previous studies suggested that this might suggest on-going adaptation/positive selection [51, 53]; however, concrete evidence supporting this notion is lacking [54]. In fact, according to examination of the sequences in the GISAID database, it was reported that the frequencies of the T11 083 variant relative to the wild-type G11 083 variant decreased globally, supporting that the mutation may hinder the virus transmissibility [46]. However, an alternative possibility is that, as the numbers of COVID-19 patients have surged very rapidly and overwhelmed public healthcare systems in many countries around the globe, limited central disease surveillance and medical resources may tend to be allocated to search for and treat patients with severe clinical manifestations of COVID-19. Asymptomatic cases (perhaps with the 11 083G>T mutation) may thus be overlooked and less likely to be tested, and in turn have fewer virus sequences deposited in public databases. Continual surveillance of COVID-19 should monitor this genomic region as it might affect the virus's pathogenicity and control of COVID-19. These results have potential applications for the development of better, and more informative test kits, potentially allowing for asymptomatic cases to be distinguished from symptomatic ones.

#### Funding information

This research project was partly supported by Mahidol University (MRC-IM 02/2563).

#### Acknowledgements

The authors would like to thank Yuttapong Thawornwattana for helpful discussion.

#### Author contributions

P.A., A.T., conceived the study. P.A., P.N., performed data curation, analysis, and interpretation of the results. P.A., drafted the manuscript. All authors discussed the results. All authors revised and approved the final manuscript.

#### Conflicts of interest

The authors declare that there are no conflicts of interest.

#### References

- Aiewsakun P, Nilplub P, Wongtrakongate P, Hongeng S, Thitithanyanont A. SARS-CoV-2 genetic variations associated with COVID-19 pathogenicity. *Figshare* 2021.
- Zhu N, Zhang D, Wang W, Li X, Yang B, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;382:727–733.
- Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, et al. The species Severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020;5:536–544.
- Huang C, Wang Y, Li X, Ren L, Zhao J, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;395:497–506.
- Chen N, Zhou M, Dong X, Qu J, Gong F, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 2020;395:507–513.
- Worldometers. COVID-19 coronavirus pandemic. Internet; 2021. <https://www.worldometers.info/coronavirus/>
- Guan W-J, Ni Z-Y, Hu Y, Liang W-H, Ou C-Q, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med* 2020;382:1708–1720.
- Chan-Yeung M, Xu R-H. SARS: Epidemiology. *Respirology* 2003;8 Suppl:S9–14.
- World Health Organization. Consensus document on the epidemiology of severe acute respiratory syndrome (SARS); 2003. <https://apps.who.int/iris/handle/10665/70863>
- World Health Organization. Middle East respiratory syndrome coronavirus (MERS-CoV) [Internet]; 2020. <https://www.who.int/emergencies/mers-cov/en/>
- Alene M, Yismaw L, Assemie MA, Ketema DB, Gietaneh W, et al. Serial interval and incubation period of COVID-19: A systematic review and meta-analysis. *BMC Infect Dis* 2021;21:257.
- Johansson MA, Quandelacy TM, Kada S, Prasad PV, Steele M, et al. SARS-CoV-2 transmission from people without COVID-19 symptoms. *JAMA Netw Open* 2021;4:e2035057.
- Poletti P, Tirani M, Cereda D, Trentini F, Guzzetta G, et al. Association of age with likelihood of developing symptoms and critical disease among close contacts exposed to patients with confirmed SARS-CoV-2 infection in Italy. *JAMA Netw Open* 2021;4:e211085.
- Liu Y, Mao B, Liang S, Yang J-W, Lu H-W, et al. Association between age and clinical characteristics and outcomes of COVID-19. *Eur Respir J* 2020;55:2001112.
- Biswas M, Rahaman S, Biswas TK, Haque Z, Ibrahim B. Association of sex, age, and comorbidities with mortality in COVID-19 patients: A systematic review and meta-analysis. *Intervirology* 2021;64:36–47.
- Pivonello R, Auriemma RS, Pivonello C, Isidori AM, Corona G, et al. Sex disparities in covid-19 severity and outcome: Are men weaker or women stronger? *Neuroendocrinology* 2021;111:1066–1085.
- Vahidy FS, Pan AP, Ahnstedt H, Munshi Y, Choi HA, et al. Sex differences in susceptibility, severity, and outcomes of coronavirus disease 2019: Cross-sectional analysis from a diverse US metropolitan area. *PLoS One* 2021;16:e0245556.
- Peckham H, de Groot NM, Raine C, Radziszewska A, Ciurtin C, et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ICU admission. *Nat Commun* 2020;11:6317.
- Liu X, Xue S, Xu J, Ge H, Mao Q, et al. Clinical characteristics and related risk factors of disease severity in 101 COVID-19 patients hospitalized in Wuhan, China. *Acta Pharmacol Sin* 2021.
- Zhang X, Tan Y, Ling Y, Lu G, Liu F, et al. Viral and host factors related to the clinical outcome of COVID-19. *Nature* 2020;583:437–440.
- Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *J Hum Genet* 2020;65:1075–1082.
- Biswas SK, Mudi SR. Spike protein D614G and RdRp P323L: The SARS-CoV-2 mutations associated with severity of COVID-19. *Genomics Inform* 2020;18:e44.

23. Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, et al. The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *Elife* 2021;10:e65365.
24. Majumdar P, Niyogi S. ORF3a mutation associated with higher mortality rate in SARS-CoV-2 infection. *Epidemiol Infect* 2020;148:e262.
25. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveill* 2017;22:13.
26. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006;23:254–267.
27. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;37:1530–1534.
28. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017;14:587–589.
29. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput Biol* 2018;14:e1005958.
30. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589–595.
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
32. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, et al. Integrative genomics viewer. *Nat Biotechnol* 2011;29:24–26.
33. Ziyatdinov A, Vázquez-Santiago M, Brunel H, Martínez-Pérez A, Aschard H, et al. lme4qtl: Linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics* 2018;19:68.
34. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020;5:1403–1407.
35. GISAID. Clade and lineage nomenclature aids in genomic epidemiology studies of active hCoV-19 viruses [Internet]; 2021. <https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/> [accessed 22 October 2021].
36. Novella IS, Zárate S, Metzgar D, Ebendick-Corpus BE. Positive selection of synonymous mutations in vesicular stomatitis virus. *J Mol Biol* 2004;342:1415–1421.
37. Nougaiyre A, De Fabritus L, Aubry F, Gould EA, Holmes EC, et al. Random codon re-encoding induces stable reduction of replicative fitness of chikungunya virus in primate and mosquito cells. *PLoS Pathog* 2013;9:e1003172.
38. Seneca S, Vancampenhout K, Van Coster R, Smet J, Lissens W, et al. Analysis of the whole mitochondrial genome: translation of the Ion Torrent Personal Genome Machine system to the diagnostic bench? *Eur J Hum Genet* 2015;23:41–48.
39. Ivády G, Madar L, Dzsudzsák E, Koczok K, Kappelmayer J, et al. Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. *BMC Genomics* 2018;19:158.
40. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012;30:434–439.
41. Raimondi F, Novelli L, Ghirardi A, Russo FM, Pellegrini D, et al. Covid-19 and gender: lower rate but same mortality of severe disease in women—an observational study. *BMC Pulm Med* 2021;21:96.
42. Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, Bujanda L, Buti M, et al. Genomewide association study of severe Covid-19 with respiratory failure. *N Engl J Med* 2020;383:1522–1534.
43. Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* 2020;587:610–612.
44. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, et al. Human leukocyte antigen susceptibility map for Severe Acute Respiratory Syndrome Coronavirus 2. *J Virol* 2020;94:e00510-20.
45. van der Made CI, Simons A, Schuurs-Hoeijmakers J, van den Heuvel G, Mantere T, et al. Presence of genetic variants among young men with severe COVID-19. *JAMA* 2020;324:663–673.
46. Wang R, Chen J, Hozumi Y, Yin C, Wei GW. Decoding asymptomatic COVID-19 infection and transmission. *J Phys Chem Lett* 2020;11:10007–10015.
47. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *mBio* 2013;4:e00524-13.
48. Cottam EM, Maier HJ, Manifava M, Vaux LC, Chandra-Schoenfelder P, et al. Coronavirus nsp6 proteins generate autophagosomes from the endoplasmic reticulum via an omegasome intermediate. *Autophagy* 2011;7:1335–1347.
49. Cottam EM, Whelband MC, Wileman T. Coronavirus NSP6 restricts autophagosome expansion. *Autophagy* 2014;10:1426–1441.
50. Xia H, Cao Z, Xie X, Zhang X, Chen J-C, et al. Evasion of type I interferon by SARS-CoV-2. *Cell Rep* 2020;33:108234.
51. Benvenuto D, Angeletti S, Giovanetti M, Bianchi M, Pascarella S, et al. Evolutionary analysis of SARS-CoV-2: How mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy. *J Infect* 2020;81:e24–e27.
52. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* 2020;83:104351.
53. Phelan J, Deelder W, Ward D, Campino S, Hibberd ML, et al. Controlling the sars-cov-2 outbreak, insights from large scale whole genome sequences generated across the world. *bioRxiv*; 2020.04.28.066977 2020.
54. van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, et al. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat Commun* 2020;11.

### Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at [microbiologyresearch.org](https://microbiologyresearch.org).