Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

EBioMedicine 74 (2021) 103695

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Genome-wide association study of hospitalized COVID-19 patients in the United Arab Emirates

Mira Mousa[a,b], Hema Vurivi[a], Hussein Kannout[a], Maimunah Uddin[c], Nawal Alkaabi[c], Bassam Mahboub[d], Guan K. Tay[e,f], Habiba S Alsafar[a,g,h,i,*], on behalf of the UAE COVID-19 Collaborative Partnership

[a] Center for Biotechnology, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[b] Nuffield Department of Women's and Reproduction Health, Oxford University, Oxford, United Kingdom
[c] Department of Pediatric Infectious Disease, Sheikh Khalifa Medical City, Abu Dhabi, United Arab Emirates
[d] Dubai Health Authority, Rashid Hospital, Dubai, United Arab Emirates
[e] Division of Psychiatry, Faculty of Health and Medical Sciences, the University of Western Australia, Crawley, Western Australia, Australia
[f] School of Medical and Health Sciences, Edith Cowan University, Joondalup, Western Australia, Australia
[g] Center for Biotechnology, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[h] Department of Biomedical Engineering, College of Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[i] Department of Genetics and Molecular Biology, College of Medicine and Health Sciences, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article History:
Received 2 July 2021
Revised 26 October 2021
Accepted 27 October 2021
Available online xxx

Keywords:
SARS-CoV-2
GWAS
Genetics
COVID-19

ABSTRACT

Background: The heterogeneity in symptomatology and phenotypic profile attributable to COVID-19 is widely unknown. The objective of this manuscript is to conduct a trans-ancestry genome wide association study (GWAS) meta-analysis of COVID-19 severity to improve the understanding of potentially causal targets for SARS-CoV-2.

Methods: This cross-sectional study recruited 646 participants in the UAE that were divided into two phenotypic groups based on the severity of COVID-19 phenotypes, hospitalized (n=482) and non-hospitalized (n=164) participants. Hospitalized participants were COVID-19 patients that developed acute respiratory distress syndrome (ARDS), pneumonia or progression to respiratory failure that required supplemental oxygen therapy or mechanical ventilation support or had severe complications such as septic shock or multi-organ failure. We conducted a trans-ancestry meta-analysis GWAS of European (n=302), American (n=102), South Asian (n=99), and East Asian (n=107) ancestry populations. We also carried out comprehensive post-GWAS analysis, including enrichment of SNP associations in tissues and cell-types, expression quantitative trait loci and differential expression analysis.

Findings: Eight genes demonstrated a strong association signal: *VWA8* gene in locus 13p14·11 (SNP rs10507497; p=9·54 x10^{-7}), *PDE8B* gene in locus 5q13·3 (SNP rs7715119; p=2·19 x10^{-6}), *CTSC* gene in locus 11q14·2 (rs72953026; p=2·38 x10^{-6}), *THSD7B* gene in locus 2q22·1 (rs7605851; p=3·07x10^{-6}), STK39 gene in locus 2q24·3 (rs7595310; p=4·55 x10^{-6}), *FBXO34* gene in locus 14q22·3 (rs10140801; p=8·26 x10^{-6}), *RPL6P27* gene in locus 18p11·31 (rs11659676; p=8·88 x10^{-6}), and *METTL21C* gene in locus 13q33·1 (rs599976; p=8·95 x10^{-6}). The genes are expressed in the lung, associated to tumour progression, emphysema, airway obstruction, and surface tension within the lung, as well as an association to T-cell-mediated inflammation and the production of inflammatory cytokines.

Interpretation: We have discovered eight highly plausible genetic association with hospitalized cases in COVID-19. Further studies must be conducted on worldwide population genetics to facilitate the development of population specific therapeutics to mitigate this worldwide challenge.

Funding: This review was commissioned as part of a project to study the host cell receptors of coronaviruses funded by Khalifa University's CPRA grant (Reference number 2020-004).

© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

* Corresponding author at: Center for Biotechnology, Khalifa University of Science and Technology, PO Box, 127788 Abu Dhabi, United Arab Emirates.
E-mail address: habiba.alsafar@ku.ac.ae (H.S. Alsafar).

https://doi.org/10.1016/j.ebiom.2021.103695
2352-3964/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Research in context

Evidence before this study

The heterogeneity in symptomatology and phenotypic profile attributable to COVID-19 disease may vary across different populations, suggesting a complex interaction between the host genetics and virus that influence the disease outcome. Multiple genome-wide association studies (GWAS) have demonstrated an association between host genetic architecture and vulnerability to COVID-19 outcome. Two loci that achieved genome-wide significance are identified in a gene cluster on chromosome 3p21·31 as a genetic susceptibility region, and in chromosome 9 confirming the potential involvement of the ABO blood-group system as a factor that contributes to severity of infection. GWAS studies in the United Kingdom have pointed to associations with loci on chromosome 19p13·2 and 21q22·1 and identified risk factors on chromosomes 2, 6, 7, 8, 10, 16, and 17.

Added value of this study

Our study presents a trans-ancestry GWAS meta-analysis of European (n=302), American (n=102), South Asian (n=99), and East Asian (n=107) ancestry populations to explore new genotypes associated to COVID-19 symptom severity, in a variety of ethnic populations that will yield to novel genetic association. We describe strong association signals from genes on chromosomes 2q22·1, 2q24·3, 5q13·3, 11q14·2, 13p14·11, 13q33·1, 14q22·3, and 18p11·31, which carry genes that are expressed in the lung, associated to tumour progression, emphysema, airway obstruction, and surface tension within the lung, as well as an association to T-cell-mediated inflammation and the production of inflammatory cytokines. Cell-type enrichment analysis demonstrated that there was an enrichment for lung tissue expression and donors. Identifying genetic variants associated to COVID-19 susceptibility and severity may uncover novel biological insights into disease pathogenesis and identify mechanistic targets for therapeutic and vaccine development.

Implications of all the available evidence

Performing GWAS across different racial and ethnic populations to identify genes and haplotypes associated with differential factors of infection and clinical outcome, as well as vaccine responses, is crucial. Some of these associations may lead to therapeutic approaches due to the expression in the lungs. Improving fundamental knowledge and underlying biological pathway of COVID-19 heterogenous phenotypes is critical to mitigating this disease.

1. Introduction

Coronavirus disease 2019 (COVID-19), the disease caused by the viral infection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has rapidly evolved into a global pandemic [4]. The heterogeneity in symptomatology, phenotypic profile, and comorbidities attributable to COVID-19 disease is widely unknown. Most of the infected persons (~80%) present with asymptomatic or mild symptoms, yet ~15% of cases are hospitalized with moderate and severe symptoms, and ~5% are fatal [5,6]. Severe COVID-19 disease has been responsible for ~5.00 million deaths worldwide, mainly due to bilateral interstitial pneumonia and acute respiratory distress syndrome [7]. While moderate and severe symptoms of persons infected with COVID-19 are associated with older age, male gender and pre-existing comorbid conditions, the pathogenesis of the disease remains poorly understood [8–10]. Severity of COVID-19 phenotypes vary across different populations, suggesting a complex interaction between the host genetics, and virus that influence the disease outcome [11]. Identifying genetic variants associated to COVID-19 susceptibility and severity may uncover novel biological insights into disease pathogenesis and identify mechanistic targets for therapeutic and vaccine development.

Although risk factors, disease management and health systems contribute to the heterogenous COVID-19 symptom phenotypes, multiple genome-wide association studies (GWAS) have demonstrated an association between host genetic architecture and vulnerability to COVID-19 outcome. Two loci that achieved genome-wide significance are located in chromosome 9 near the ABO gene, and a signal in chromosome 3 near a cluster of genes (SLC6A20, CXCR6, CCR1, CCR2, and CCR9) that play a role in immune function [12]. Recent papers have validated the two previous loci, and found novel risk variants at chromosome 1q22, 6p21·33, 12q24·13, 19p13·3, and 21q22·1 associated with severe COVID-19 [13–16]. Further investigation in independent datasets of various genetic populations is needed to investigate genetic variation among individuals infected with SARS-CoV-2. Genetic variability due to population-specific genetic characteristics, including variants, admixture and ancestral haplogroup distribution may lead to differential magnitudes of the variants' effect. Genetic heterogeneity contributes to clinical health disparities, where there is a lack of genetic studies conducted on the Middle Eastern population. Admixture and consanguinity have jointly affected the genetic diversity in the Middle East population [17,18]. Hence, our study provides the unique advantage of obtaining samples from the Middle Eastern population, an underrepresented region in genetic studies, and explore new genotypes in this population that will yield to novel genetic association.

We performed the first GWAS in the United Arab Emirates (UAE) to identify genetic variants that are associated with hospitalized cases of COVID-19 patients. The objective of this study is to conduct a trans-ancestry GWAS meta-analysis in a variety of ethnic populations to investigate host genetic factors associated to COVID-19 severity and further understand the biological pathways that are triggering these heterogenous phenotypes. We also utilized approaches and strategies to translate genetic risk loci of potential candidate genes and tissue-specific expression quantitative trait loci (eQTL) data to predict the causal genes. We performed gene-set enrichment analysis with GWAS summary statistics to identify the potential biological pathways for COVID-19 hospitalizations. This landscape of potential pleiotropic genes and biological pathways will help us to understand risk of severe COVID-19 disease and assist early detection and improve pharmaceutical therapy.

2. Methods

2.1. Ethics Statement

An informed written consent form was obtained from all participants, in accordance with the Declaration of Helsinki. Informed consent was obtained from a family member of patients who were on ventilators with a signed agreement by a supervising physician. All data was de-identified prior to use. This study has been approved by the local ethics committee at Abu Dhabi Health COVID-19 Research Ethics Committee (DOH/DQD/2020/538), Dubai Scientific Research Ethics Committee (DSREC-04/2020_09) and SEHA Research Ethics Committee (SEHA-IRB-005).

2.2. Study Participants and Recruitment

This cross-sectional study recruited 646 consenting participants that have tested positive for SARS-CoV-2 by Real-Time Polymerase Chain Reaction (RT-PCR) via nasopharyngeal swabs. Participants

were prospectively recruited from six collection sites across the UAE including medical centres, hospitals, and quarantine camps, from April 1, 2020, to January 31, 2021, as demonstrated in Supplementary Figure 1. The selection criteria were: [1] positive COVID-19 test for a single individual, [2] resident of the UAE, and [3] able to provide an informed consent and complete the survey. Information about the study was provided to potential participants. Consenting participants were provided with a questionnaire that includes details on the demographic characterization, risk factors, and symptomatology, which were cross-checked with the electronic health record. Reporting of the study design was done in adherence to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines [19].

After the participants were tested for the COVID-19 infection, twice in a 3-day waiting period, blood samples were collected in a sterile 5ml sample tube supplemented with ethylenediaminetetraacetic acid from the cubital vein. Samples were transported in a sealed biohazard bag using a cool transport container to Khalifa University Centre for Biotechnology in Abu Dhabi for genotyping and analysis.

2.3. Binary Phenotype Definition

The severity of COVID-19 classification can be grouped into five stages of illness categories: asymptomatic, mild, moderate, severe and critical [20]. Patients are classified as asymptomatic if they have no symptoms that are consistent with COVID-19. Patients with mild illness are individuals who have any of the various signs and symptoms of COVID-19 (e.g., fever, cough, sore throat, malaise, headache, muscle pain, nausea, vomiting, diarrhoea, loss of taste and smell) but do not have shortness of breath, dyspnoea, or abnormal chest imaging. Moderate COVID-19 patients were those with lower respiratory disease during clinical assessment or imaging and who have an oxygen saturation (SpO_2) ≥94% on room air at sea level and require hospitalization. Severe illness is defined as patients who have SpO_2 <94% on room air at sea level, a ratio of arterial partial pressure of oxygen to fraction of inspired oxygen (PaO_2/FiO_2) <300 mm Hg, respiratory frequency >30 breaths/min, or lung infiltrates >50%, and require hospitalization. Critical illness are patients who have respiratory failure, septic shock and/or multiple organ dysfunction, and require hospitalization.

The classifications of COVID-19 severity were taken as the worst classification during the patient's hospital stay. These assessments were done by at least two physicians and were further verified by the information derived from the participant questionnaire. In this study, to identify the susceptibility genetic loci contributing to hospitalized cases for COVID-19, the moderate, severe, and critical COVID-19 patients were defined as "cases", and the asymptomatic and mild COVID-19 patients were defined as "controls". Cases for the COVID-19 phenotype group were defined as hospitalized (n=482), where patients were hospitalized with COVID-19 as primary reason for admission, due to [1] lower respiratory disease, which was defined by the use of respiratory support or individuals with assisted respiratory distress syndrome (ARDS), pneumonia or progression to respiratory failure that requires supplemental oxygen therapy or mechanical ventilation support, [2] severe complications such as septic shock or multi-organ failure, or [3] mortality. These phenotypes were readily extracted from the electronic health records, recorded by two physicians, as well as information derived from the surveys. Controls for the COVID-19 group were defined as non-hospitalized (n=164) participants that did not require hospitalization, due to asymptomatic or mild symptoms. As per the UAE law, patients who are asymptomatic or have mild symptoms are advised to isolate at home, as opposed to visiting the hospital or a quarantine camp. Therefore, given that most of the patients were collected from a hospital site, the number of COVID-19 patients who showed asymptomatic or mild symptoms that did not require hospitalization were much lower than the number of patients who had moderate/severe/critical COVID-19 symptoms. Therefore, this study was not able to capture a higher number of control participants.

2.4. DNA extraction, genotyping and imputation

DNA was extracted using the automated MagPurix 12 system according to the manufacturer's protocol. DNA was quantified using the DS-11 Series of Spectrophotometer/Fluorometer (DeNovix). Genotyping was performed using the Infinium Global Screening Array-24 v3.0 BeadChip (Illumina, Inc., San Diego, CA, USA) according to the manufacturer's protocol, which contained 654,027 genetic markers and developed by Avera Institute for Human Genetics (South Dakota, USA). The Infinium Global Screening Array-24 v3.0 BeadChip would be best suited for this ethnically diverse cohort because it features a broad spectrum of diverse exonic content, including both cross population and population specific markers (EUR: 52,980 markers; EAS: 31,375 markers; AMR: 45,977 markers; AFR: 43,122 markers; SAS: 40,298 markers).

We then performed stringent quality controls (QC) for both samples and Single nucleotide polymorphisms (SNPs) to ensure subsequent robust association tests. Samples that failed to reach a 98.0% call rate were reanalysed. Samples were excluded for the following reasons: 1) duplicated samples, 2) discrepancies between reported sex and genetically inferred sex, 3) per-sample call rate <98%, 4) heterozygosity outlier (±3), 5) unusually high number of singleton genotypes, and 6) related individuals to second degree (pairwise identity-by-state (PI_HAT)>0.5, family sets with the lowest call rate were removed). SNPs were excluded due to: 1) low minor allele frequency (<0·01), 2) low genotyping rate (<95%), and 3) deviation from Hardy-Weinberg equilibrium ($p<10^{-6}$) significance level. After quality control, 438,617 variants passed filters. Subsequently, principal component analysis (PCA) was performed to stratify by calculating the first ten principal component (PCs) per individual, after merging with samples from the Phase 3 multi-ethnic 1000 Genomes Projects panel. Details on the number of samples and SNPs that did not pass quality control are outlined in Supplementary Figure 1.

Genotypes were prephased and imputed with untyped markers (~39M) using the Phase 3 multi-ethnic 1000 Genomes Projects panel, as the reference based on the human genome assembly hg19 (https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html), and was carried out by using BEAGLE, using standard protocols and recommended settings. Imputation was performed on both autosomal and sex chromosomes. A study conducted in the Kuwaiti-Arab population demonstrated that the 1000G phase 3 imputation panel is able to capture the appropriate genotype frequencies of the imputed markers with those genotypes in the population from the region, when comparing to the Kuwait Arab Exome Variant Database [21] and the Greater Middle East Variome project [22]. Post imputation quality control measures (genotyping rate < 95%, call rate < 98%, MAF > 1%, HWE $p > 1.0 \times 10^{-6}$) were conducted, and a total of 9,175,654 variants passed filters.

2.5. Statistical analysis

The statistical analysis was conducted as a case-control panel, with controls characterized as non-hospitalized symptoms and cases characterized as hospitalized symptoms, with the use of PLINK (version 1·9), R (version 3·4) and SPSS (version 16·0). Haploview v4·1 [23] was used to perform visualization maps, and regional plots were created using LDlink, with varying window size for optimal visualization [24].

To explore non-genetic associations with COVID-19 hospitalization, the following factors were investigated: [1] demographic variables (age, gender, nationality,); [2] risk factors (body mass index

(BMI), blood type, current smoking, alcohol intake); [3] symptoms (fever, cough, headache, fatigue, diarrhoea, loss of taste or smell, arthralgia, conjunctivitis, chills, malaise, nausea, sore throat, vomiting, sneezing and dyspnoea); and [4], pre-existing conditions (cardiovascular condition, chronic lung disease, diabetes mellitus, immunosuppressive condition, neurological disorder, metabolic disease, and renal disease). These variables will be investigated in association to hospitalization for COVID-19 using Pearson χ^2 for descriptive statistical test and multivariate logistic regression models. Multivariate regression models adjusted for age (continuous), sex, ancestry, current BMI (underweight, normal weight, overweight, obese), and presence of pre-existing conditions.

For case-control comparisons, we tested for association using logistic regression, assuming additive allelic effects for genotyped and imputed SNPs. Genotypes were exported, in Genome Reference Consortium human build 37 (GRCHb37) and Illumina 'source' strand orientation, using the Genotype Studio PLINK input report plugin v·2-1-4. To assess the association of a given SNP with COVID-19 severity, the allelic frequencies were compared by means of a χ^2 statistic, which yielded an individual p-value for each combination of SNP and allelic model. A quantile-quantile (Q-Q) plot analysis was carried out to check whether the distribution of the inflation *p*-values deviated from the expected distribution under the null hypothesis of no genetic association and investigate if the overall significance of the genome-wide associations is due to potential impact of population stratification. Systematic bias and the impact of population stratification was evaluated by calculating the genomic control inflation factor [λ_{GC}] and noted for each analysis. A Manhattan plot was generated with -$\log_{10}$ *p*-values.

To reduce technical confounders and population stratification, the participants were analysed separately by ethnic group depending on the genetic threshold. We performed GWAS by ancestry group followed by trans-ancestry meta-analysis across all ancestry groups. The 1000 genome v3 reference panel was used for genomic population assessment and divided into the following five super populations: African (AFR), admixed from the Americas (AMR), East Asian (EAS), South Asian (SAS), and European (EUR). Genetic ancestry was estimated using PCs performed based on uncorrelated SNPs. Using all 10 ancestry PCs, the median and variance for each population was calculated, followed by the Mahalanobis distance for each population sample. The sample in this study was then assigned to the 1000 genome v3 reference panel population with the minimum Mahalanobis distance. In each individual ethnic group, we performed association testing under an additive model for the effect of the risk allele while adjusting for age, sex, and population stratification. To adjust for population structure, we first used principal component analysis (PCA) after LD pruning from the initial pool. As a measure to control for population stratification, two eigenvectors were used in subsequent adjustment analyses. Association summary statistics from GWAS within each ancestry group were then combined by fixed-effects meta-analysis with inverse variance weights was used to obtain the combined effects for each SNP. We also computed the test statistics for heterogeneity of effect among studies for each variant using Cochran's Q-test. We removed variants with heterogeneity *p*-value$<$0.001 from the meta-analyses. Genome-wide association markers surpassed a conservative Bonferroni-corrected significance threshold of discovery $p<5\times10^{-8}$, whereas markers that identified associations that reached a suggestive association threshold of $p<5\times10^{-5}$.

2.6. Gene annotation and expression quantitative trait loci (eQTL)

To gain insight into the potential functional roles of the significantly associated loci, we performed enrichment to quantitative trait analysis, cell type transcriptomics and human tissue specific expression extensively by searching publicly available data from Human Protein Atlas (http://www.proteinatlas.org) [25–28] and the Genotype-Tissue Expression v7 (GTEx) tool were utilized [29–31]. Both Human Protein Atlas and GTEx have provided an atlas of the regulatory landscape of gene expression and splicing variation in a broad selection of primary human tissues. Nearly all protein-coding genes have one local variant associated with expression changes and the majority have common variants affecting alternative splicing (FDR$<$5%). To gain insight into the cell types that are important for COVID-19, we evaluated whether expression weights for lung tissues were investigated in association to gene enrichment variants. We interrogated coding variants for each lead SNP and its proxies ($R^2>0.8$) using Ensembl and HaploReg.

The GTEx was established to characterize human transcriptomes and has created a reference resource of gene expression levels from non-diseased tissues, including genotype, gene expression, and histological data. GTEx database captures the estimated effect size of an eQTL allele on gene expression, which allows for identifying genes, whose expression is affected by genetic variation, providing information on variant's potential involvement in COVID-19 phenotype. The output includes the levels of gene expression across all tissue types as well as within tissues, some of which are of interest as they are involved in lungs. eQTL-based and chromatin interaction-based mapping were used to aid the interpretation of variants identified.

2.7. Functional annotation

Functional annotation was performed with Functional Mapping and Annotation of genetic associations (FUMA) [32] software, an online platform for functional mapping of genetic variants from GWAS summary statistics. Independent significant SNPs are defined as SNPs that have a $p<1\text{x}10^{-5}$ and are independent from each other at the linkage disequilibrium threshold $r^2<0{\cdot}6$. Therefore, independent significant SNPs are essentially the same as SNPs that are contained after clumping GWAS tagged SNPs at the same *p*-value and r^2. Independent significant SNPs are used to select candidate SNPs that are in LD with the independent significant SNPs. Each locus is represented by the top lead SNP that has the minimum p-value at the locus. All SNP genomic locations are based on the NCBI37/hg19 assembly. Pairwise LD structure of SNPs was based on the 1000 Genomes Project Phase 3. Variant annotation for each locus was made on distinct significant SNPs and SNPs in LD with these, with a maximum distance of 250 kb and $p<0{\cdot}05$, referred to as candidate SNPs.

FUMA incorporates 18 biological data repositories and to test whether these expressions and polygenic signal measured in each GWAS loci clustered specific biological pathways, a gene-set analysis was produced in Multi-marker Analysis of GenoMic Annotation (MAGMA) [33]. MAGMA examines sets of biologically related genes that are more strongly associated with COVID-19 hospitalizations than other genes. This tool uses multiple linear regression models to assess whether genes in each gene set are more strongly associated with a polygenic trait, correcting for confounding factors such as linkage disequilibrium between variants and gene size. A gene-property analysis was conducted using MAGMA [33] in order to indicate a the role of particular tissue expression to gene associations, using the GWAS summary statistics. This will evaluate whether genes with high specificity in each cell-type are enriched for association with COVID-19 hospitalization. With a given polygenic trait than all other genes in this genome.

2.8. Role of the funding source

This review was commissioned as part of a project to study the host cell receptors of coronaviruses funded by Khalifa University's CPRA grant (Reference number 2020-004). The funders did not have

a role in study design, data collection, data analyses, interpretation or writing of report.

3. Results

3.1. Demographics of the participants

A total of 646 participants were included in this study, consisting of non-hospitalized patients (n=164) and hospitalized patients (n=482). Demographic data of the cohort are summarized in Table 1. The average age was 44·70 years (SD±15·53), and average BMI was 28·11 kg/m^2 (SD±15·53). Most of the participants were male (78·6%) of Asian (57·0%) or Middle Eastern (36·5%) ancestry. The distribution of gender and ethnicity indicate a patient cohort that is representative of the many segments of the UAE population (72% male; 59% Asian descent; 34% Middle Eastern descent)(34). The patients displayed several clinical presentations typical to COVID-19, which mainly involved cough (65·2%), pneumonia (52·0%), loss of taste/smell (47·7%), shortness of breath (44·1%), fatigue (41·0%), headache (38·0%), body ache (35·7%), sore throat (32·2%), acute respiratory distress syndrome (23·8%), chills (23·1%), malaise (20·6%), myalgia (19·6%), arthralgia (17·8%), and nausea (13·7%).

Consistent with previous reports, hospitalized patients tend to be older ($p<0·001$), higher BMI ($p<0·001$), and with a presence of one or more comorbidities ($p<0·001$). While smoking served as a protective factor from developing hospitalized COVID-19 symptoms ($p<0·001$), a multivariate regression analysis was performed by adjusting for age, gender, BMI, and presence of comorbid condition, demonstrating no association (OR: 0·53 (95% CI: 0·25, 1·11); p=0·091) between hospitalization for COVID-19 and smokers. Alcohol had no impact on disease severity (p=0·204), even after a multivariate regression analysis with coviariate adjustment (p=0·837). Contrasting to previous reports, male gender did not have a higher likelihood of developing severe/critical disease, in comparison to female gender (p=0·382). Blood type was not associated with symptom severity (p=0·136). Among the participants that reported a comorbid condition, hospitalized COVID-19 were more likely to have a previous diagnosis of cardiac condition ($p<0·001$), chronic lung disease ($p<0·001$), diabetes mellitus ($p<0·001$), immunosuppressive condition (p=0·004), neurological disorder (p=0·020), metabolic disease ($p<0·001$), and renal disease (p=0·018).

Table 1
Demographic details of the cohort.

	Non-Hospitalized (n=164)	Hospitalized (n=482)	*p*-value
Gender			
Male	125 (76.2%)	383 (79.3%)	0.382
Female	39 (23.8%)	99 (20.5%)	
Age			
<15	1 (0.6%)	7 (1.5%)	**<0.001**
16-30	66 (40.2%)	44 (9.1%)	
31-45	79 (48.2%)	169 (35.1%)	
46-60	14 (8.5%)	152 (31.5%)	
61-75	3 (1.8%)	91 (18.9%)	
>76	1 (0.6%)	19 (3.9%)	
BMI			
<18.50	1 (0.9%)	9 (1.9%)	**<0.001**
18.51-24.50	44 (39.3%)	95 (20.1%)	
24.51-29.99	50 (44.6%)	209 (44.3%)	
>30.00	17 (15.2%)	159 (33.7%)	
Self-reported ancestry			**<0.001**
Middle East	35 (21.3%)	201 (41.7%)	
South Asian	93 (56.7%)	179 (37.1%)	
East Asian	32 (19.5%)	64 (13.3%)	
African	4 (2.4%)	25 (5.2%)	
European	0 (0.0%)	8 (1.7%)	
Hispanic	0 (0.0%)	5 (1.0%)	
Smoker			
No	92 (80.7%)	436 (92.0%)	**<0.001**
Yes	22 (19.3%)	38 (8.0%)	
Alcohol consumer			
No	81 (91.0%)	399 (94.5%)	0.204
Yes	8 (9.0%)	23 (5.5%)	
Blood type			0.136
A	21 (22.8%)	139 (32.0%)	
AB	5 (5.4%)	30 (6.9%)	
B	22 (23.0%)	112 (25.7%)	
O	44 (27.8%)	154 (35.4%)	
Presence of comorbidities	135 (84.9%)	232 (48.9%)	**<0.001**
No	24 (15.1%)	242 (51.1%)	
Yes			
Cardiac condition			
No	156 (99.4%)	267 (80.9%)	**<0.001**
Yes	1 (0.6%)	63 (19.1%)	
Chronic lung disease			
No	158 (100.0%)	273 (91.0%)	**<0.001**
Yes	0 (0.0%)	27 (9.0%)	
Diabetes mellitus			
No	150 (95.5%)	207 (54.8%)	**<0.001**
Yes	7 (4.5%)	171 (45.2%)	
Immunosuppressive condition			**0.004**
No	99 (99.0%)	268 (90.2%)	
Yes	1 (1.0%)	29 (9.8%)	
Neurological disorder			**0.020**
No	158 (100.0%)	142 (96.6%)	
Yes	0 (0.0%)	5 (3.4%)	
Metabolic disease			
No	158 (100.0%)	272 (88.0%)	**<0.001**
Yes	0 (0.0%)	37 (12.0%)	
Renal disease			
No	97 (100.0%)	204 (94.4%)	**0.018**
Yes	0 (0.0%)	12 (5.6%)	

*p-value was measured using Pearson's χ^2

3.2. Identification of susceptibility loci by genome-wide association study

To further investigate genetic effects for the patient phenotypic heterogeneity, we performed a genome-wide association study of dichotomous classification of the defined "hospitalized group" that consists of the moderate, severe, and critical ill patients (n = 482), and the "non-hospitalized group" (n = 164) that consists of the asymptomatic and mild patients. An admixture informed principal component analysis (PCA) was plotted for the 646 participants from this study (represented as yellow circles in the PCA plot), merged with 1092 samples from the 1000 Genomes Project v3, suggesting a multi-ancestry cohort with significant population admixture and a diverse genetic pool across various ethnic groups (Supplementary Figure 2).

To reduce technical confounders and population stratification, the participants were analysed separately by ethnic group depending on the genetic threshold. GWAS analysis was carried out separately for East Asian, South Asian, European, and American samples, followed by a trans-ancestry meta-analysis across the four population groups. The African ethnicity was not included in the analysis due to low sample size (n=34). The East Asian ancestry comprised 107 participants (72 cases; 35 controls) with 7,047,811 total variants that passed QC. The South Asian ancestry comprised 99 participants (81 cases; 18 controls) with 7,190,592 total variants that passed QC. The European ancestry comprised 302 participants (210 cases; 92 controls) with 8,615,040 total variants that passed QC. The American ancestry comprised 102 participants (90 cases; 12 controls) with 7,124,060 total variants that passed QC. All ancestry models were adjusted for age, gender, and population stratification (two eigenvectors of the principal component analysis). The adjusted and unadjusted analyses of the SNPs with a suggestive association signal (threshold of $p<5\times10^{-5}$) for each of the following ancestry groups are presented

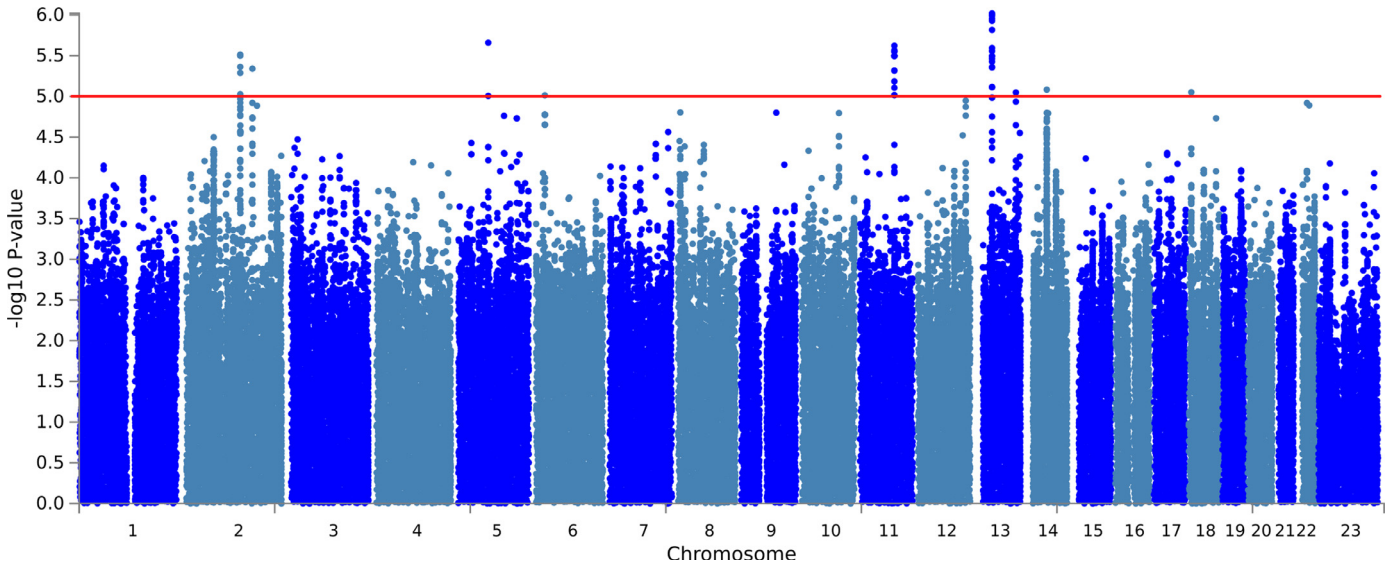

Fig. 1. Manhattan plot of the trans-ancestry GWAS meta-analysis of 610 participants (non-hospitalized (n=157) vs. hospitalized (n=453)), highlighting eight peaks with moderate association signal for hospitalized cases of COVID-19. The GWAS analysis results are shown on the y-axis as -log10 (*p*-value) and on the x-axis is the chromosomal location. The red horizontal line illustrated the suggestive genome-wide association threshold ($p<5x10^{-5}$).

in the Supplementary: East Asian population (Supplementary Figure 3, Supplementary Table 1), South Asian population (Supplementary Figure 4, Supplementary Table 2), European population (Supplementary Figure 5, Supplementary Table 3), and American population (Supplementary Figure 6, Supplementary Table 44). The results of the ethnic-specific, genomic control inflation factor were: European ancestry (λ_{GC} = 1.00), East Asian ancestry (λ_{GC} = 1.00), South Asian ancestry (λ_{GC} = 1.00), and American ancestry (λ_{GC} = 1.01). A quantile-quantile plot (Q-Q) plot of the meta-analyses of the four population groups (Supplementary Figure 7) demonstrates the associations in the tail distribution with minimal overall genomic inflation of the statistical results (λ_{GC} of 0·97 for meta-analysis).

The trans-ancestry genome-wide association meta-analysis of 610 participants (non-hospitalized (n=157) vs. hospitalized (n=453)) results identified eight loci that reached a suggestive association threshold of $p<5\times10^{-5}$ (red horizontal line in Fig. 1), but no variants surpassed a conservative Bonferroni-corrected significance threshold of discovery $p<5\times10^{-8}$. At each locus we identified the sentinel SNP (the leading SNP with the lowest *p* value), with the following eight loci that will be further elaborated in this manuscript: 2q22·1, 2q24·3, 5q13·3, 11q14·2, 13q21·33, 13q33·1, 14q22·3, and 18p11·31 (Table 2; Fig. 2 and Fig. 3). Regional association plots for the eight newly identified loci are shown in Fig. 2. There was no evidence for heterogeneity of effect (I^2=0.00) between the ancestry groups in the genome-wide association meta-analysis data. The complete list of genes, their functional annotations, and all SNPs in LD associated with hospitalized COVID-19 for the adjusted analyses are presented in Supplementary Table 5 and Supplementary Table 6.

3.3. *Gene prioritization of novel associations*

3.3.1. VWA8 *gene in locus 13p14·11*

Von Willebrand A Domain Containing Protein 8 (*VWA8*) gene located at the 13p14·11 locus (Fig. 2a) is localized in the mitochondria where it has a N-terminal mitochondrial targeting sequence that allows it to export to the mitochondria organelle. It plays a role in metabolic regulation or bioenergetic events [3]. A peak is demonstrated in this region, with the leading SNP rs10507497 (allele A; OR: 3·12 (95% CI: 1·98, 4·93); p=9·54 $x10^{-7}$) marking a suggestive association signal for the development of hospitalized COVID-19 phenotypes. Even though the reported lead SNP is common in all ethnic groups, the association signals are apparently specific to the populations of South Asian (p=6·05 $x10^{-3}$) and European ancestry (p=2·84 $x10^{-4}$). The *VWA8* gene is expressed in homogenized lung tissue and tracheal epithelium, and has previously been associated to Panlobular emphysema, a morphologic condition associated with kyphotic chest deformities, localized scleroses, deformities in the large bronchial airway and congenital emphysema [35,36]. The *VWA8* gene leads to airway obstruction and surface tension phenomena within the lung that progressively destroys the pulmonary contribution to homeostasis [37]. A SNP in proxy to rs10507497 have been expressed in the B-lymphocyte and lymphoblastoid (rs11842978, r^2=0·89).

3.3.2. PDE8B *gene in locus 5q13·3*

Phosphodiesterase 8B (*PDE8B*) gene located at the 5q13·3 locus (Fig. 2b) catalyses the hydrolysis of the second messenger cyclic AMP (cAMP) signalling pathway, and is involved with the thyroid function, and is seen in the epithelium of the apical portion of the lung bronchioles, the surface of cardiomyocytes, on the surface of the epithelial-cell apical portion of the stomach and the brush border surface of the presumptive kidney distal tubules [38]. The leading SNP of this peak is rs7715119 (allele A; OR: 0·34 (95% CI: 0·22, 0·54); p=2·19 $x10^{-6}$), predominantly present in participants of European (p=1·75 $x10^{-5}$) and American (p=4·80 $x10^{-2}$) ancestry. The expression of *PDE8* in the airway smooth muscle cells and T-lymphocytes indicate an additional target to T-cell-mediated inflammation and bronchoconstriction [39–41].

3.3.3. CTSC *gene in locus 11q14·2*

Cathepsin C (*CTSC*) gene located at the 11q14·2 locus (Fig. 2c) is a lysosomal cysteine protease essential for catalytic activation of many serine proteases. Loss-of-function mutations of *CTSC* result in inactivation of neutrophil serine proteases, hence leading to inflammatory disease, pneumonia and viral infection [42–44]. The leading SNP of this peak is rs72953026 (allele T; OR: 0·35 (95% CI: 0·23, 0·54); p=2·38 $x10^{-6}$), and has been associated with respiratory disorders, cardiomyopathy, emphysema or chronic bronchitis [45]. Although the risk allele has a consistent direction of effect and frequency across ethnicities, an association signal was only present in the East Asian ancestry (p=1.41 $x10^{-2}$), South Asian ancestry (p=1.31 $x10^{-2}$) and European ancestry (p=7·89 $x10^{-4}$). SNPs in proxy (rs7120867, r^2=0·87; rs72970468, r^2=0·93; rs55939583, r^2=0·93; rs11600428,

Table 2
Meta-analysed results of the susceptibility loci associated with hospitalized COVID-19 patient, stratified by ancestry (East Asian, South Asian, European and American Ancestry), associated with hospitalized COVID-19 patients. The total population included 610 participants (non-hospitalized (n=157) vs. hospitalized (n=453)). The SNPs that passed the association signal are included in this table. The *p*-value and corresponding odds ratio and 95% confidence intervals of the minor allele are presented.

Nearest Gene/ Cytoband	Lead SNP	Position (hg19)	Allele	Trans-ancestry Meta-analysis			East Asian Ancestry			South Asian Ancestry			European Ancestry			American Ancestry		
				RAF	OR (95% CI)	P-Value	RAF	OR (95% CI)	P-Value	RAF	OR (95% CI)	P-Value	RAF	OR (95% CI)	P-Value	RAF	OR (95% CI)	P-Value
VWA8/ 13q14.11	rs10507497	42320827	A	0.23	3.12 (1.98, 4.93)	9.54×10^{-7}	0.24	2.45 (0.94, 6.32)	0.063	0.29	7.37 (1.77, 30.7)	6.05×10^{-3}	0.19	3.11 (1.68, 5.74)	2.84×10^{-4}	0.25	2.42 (0.62, 9.38)	0.198
PDE8B/ 5q13.3	rs7715119	76675418	A	0.15	0.34 (0.22, 0.54)	2.19×10^{-6}	0.12	0.34 (0.09, 1.24)	0.103	0.10	1.01 (0.23, 0.43)	0.992	0.18	0.32 (0.18, 0.54)	1.75×10^{-5}	0.10	0.21 (0.04, 0.99)	4.80×10^{-2}
CTSC/ 11q14.2	rs72953026	88169599	T	0.18	0.35 (0.23, 0.54)	2.38×10^{-6}	0.21	0.28 (0.10, 0.77)	1.40×10^{-2}	0.18	0.25 (0.08, 0.75)	1.31×10^{-2}	0.17	0.36 (0.20, 0.65)	7.89×10^{-4}	0.18	0.64 (0.21, 2.01)	0.453
THSD7B/ 2q22.1	rs7605851	137706056	G	0.19	0.35 (0.22, 0.54)	3.07×10^{-6}	0.43	0.27 (0.11, 0.69)	6.76×10^{-3}	0.01	NA		0.21	0.40 (0.23, 0.67)	6.42×10^{-4}	0.07	0.21 (0.04, 1.03)	0.054
STK39/ 2q24.3	rs7595310	168810137	A	0.26	0.41 (0.27, 0.59)	4.55×10^{-6}	0.20	0.23 (0.08, 0.67)	7.06×10^{-3}	0.28	0.45 (0.17, 1.22)	0.116	0.24	0.46 (0.28, 0.76)	2.55×10^{-3}	0.35	0.32 (0.10, 1.00)	0.051
FBXO34/ 14q22.3	rs10140801	55730404	T	0.42	2.10 (1.51, 2.91)	8.26×10^{-6}	0.38	2.02 (0.88, 4.59)	0.093	0.38	1.81 (0.76, 4.32)	0.179	0.43	2.25 (1.47, 3.44)	1.78×10^{-4}	0.44	1.86 (0.69, 4.99)	0.22
RPL6P27/ 18p11.31	rs11659676	6461230	C	0.41	0.48 (0.34, 0.66)	8.88×10^{-6}	0.37	0.42 (0.19, 0.92)	2.99×10^{-2}	0.42	0.41 (0.16, 0.99)	4.80×10^{-2}	0.43	0.46 (0.30, 0.71)	4.29×10^{-4}	0.37	0.80 (0.32, 1.97)	0.63
METTL21C/ 13q33.1	rs599976	103348216	G	0.15	0.37 (0.24, 1.22)	8.95×10^{-6}	0.04	0.68 (0.13, 3.46)	0.643	0.22	0.27 (0.09, 0.75)	1.22×10^{-2}	0.16	0.38 (0.22, 0.65)	5.33×10^{-4}	0.16	0.36 (0.09, 1.42)	0.145

CI: Confidence Interval; NA: Not Applicable; RAF: Rare Allele Frequency; SNP: Single Nucleotide Polymorphism; OR: Odds Ratio.
Nearest gene indicates gene either harbouring the variant or nearest to it.
East Asian Ancestry: 107 participants (72 cases; 35 controls); 7047811 total variants that passed QC.
South Asian Ancestry: 99 participants (81 cases; 18 controls); 7190592 total variants that passed QC.
European Ancestry: 302 participants (210 cases; 92 controls); 8615040 total variants that passed QC.
American Ancestry: 102 participants (90 cases; 12 controls); 7124060 total variants that passed QC.
Trans-ancestry GWAS Meta-Analysis (fixed effect): 610 participants (453 cases; 157 controls); 7750527 total variants that passed QC; heterogeneity (I^2) in the reported SNPs was 0.00.
Additive effect of variant was applied for all models corrected for population stratification with adjustment to 2 principal components, age, and gender, and *p*-value was measured using Pearson's χ^2.
Monomorphic or minor allele frequency (<0.01) for the South Asian population in 1000 Genomes South Asian populations; http://useast.ensembl.org/index.html.

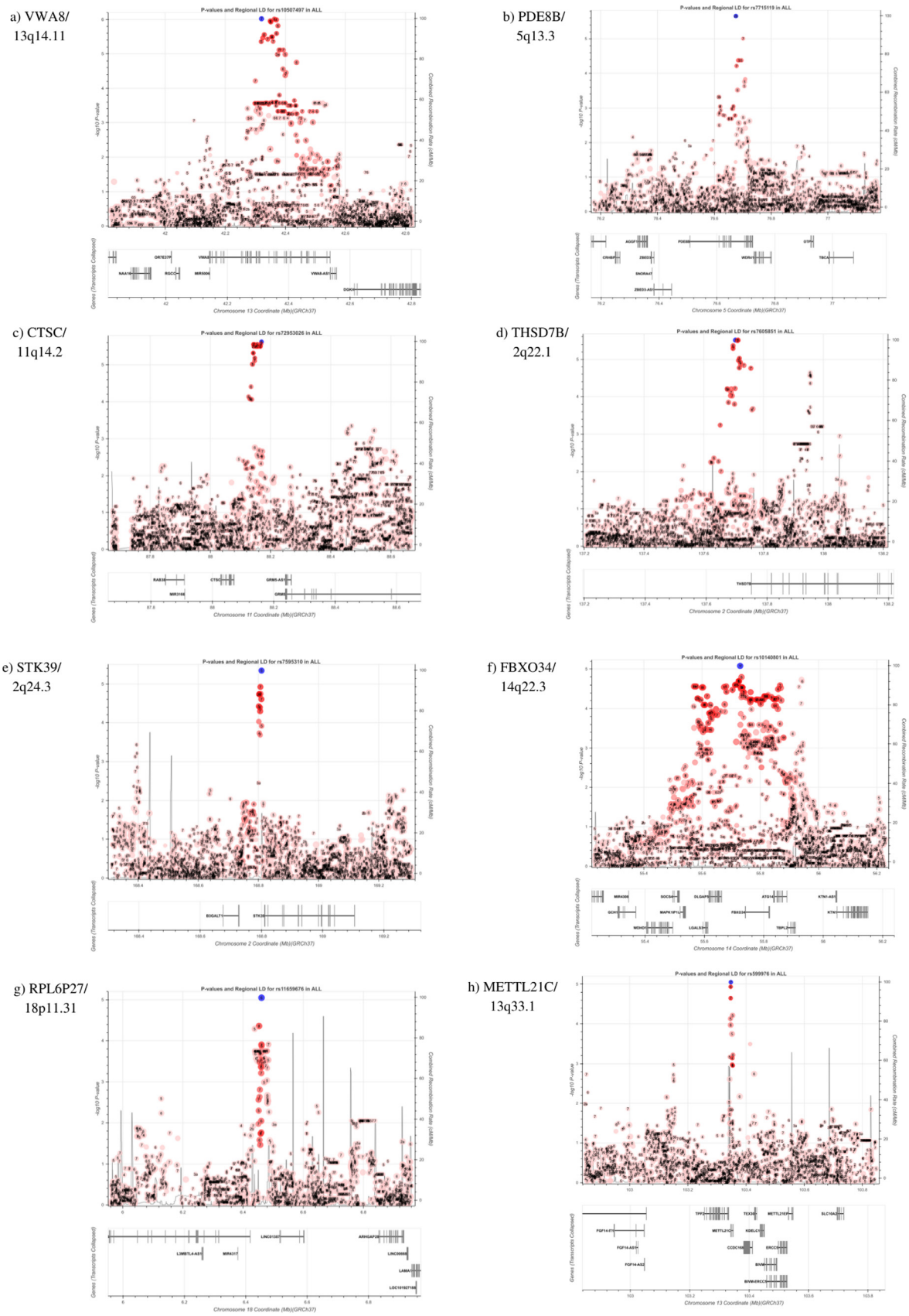

Fig. 2. Regional plots for association of genotyped and imputed SNPs in the trans-ancestry GWAS meta-analysis in association to hospitalized COVID-19. SNPs are plotted according to their chromosomal position (NCBI Build 37) with -log10 *p*-values on the y-axis, and the relative location of the annotated genes and the direction of transcription are shown in the lower portion of the Fig. The most strongly associated SNP is shown as a small purple circle. Linkage disequilibrium (LD; R^2 values) between the lead SNP and the other SNPs is indicated using red colours. The colour scheme indicated the LD displayed as r^2 values between all SNPs and the top-ranked SNP in each plot. Top-ranked SNPs are shown as purple

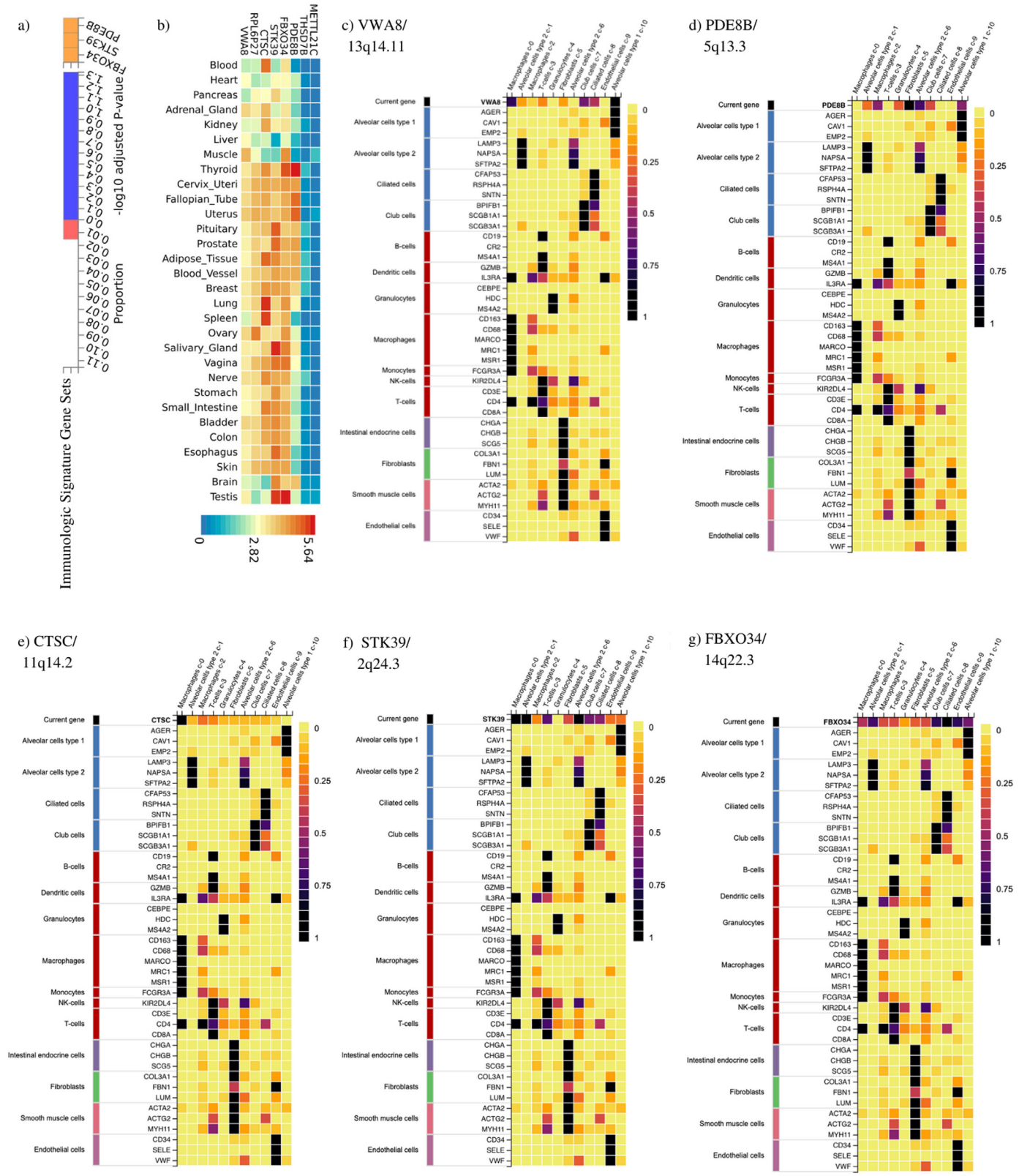

Fig. 3. Functional annotation and eQTL expression in the Lung tissue obtained from Human Protein Atlas database: (a) immunologic signature of the proportion of overlapping genes (*FBXO34*, *STK39*, *PDE8B*); b) heatmap of GTEx v8 30 general tissue types and the average expression per label (log2 transformed); c) *VWA8* gene in locus 13p14·11, (d) *PDE8B* gene in locus 5q13·3, (e) CTSC gene in locus 11q14·2, (f) STK39 gene in locus 2q24·3; (g) FBXO34 gene in locus 14q22·3.

diamonds in the top of each chromosomal locus: (***a***) 13q14·11, (***b***) 5q13·3, (***c***)11q14·2, (d) 2q22·1, (e) 2q24·3, (f) 14q22·3, (g) 18p11·31, and (h) 13q33·1. The blue lines represent the estimated recombination rates. Plots are generated using LDlink.

r^2=1·00; rs72953026, , r^2=1·00) to the lead SNP has been associated to lung fibroblasts and B-lymphocyte.

3.3.4. THSD7B *gene in locus 2q22·1*

Thrombospondin Type-1 Domain-Containing 7 (*THSD7B*) gene at the 2q22·1 locus (Fig. 2d) is a membrane-associated N-glycoprotein that mediates endothelial cell migration, cytoskeletal organization, and tube formation. This gene has been identified in GWAS in association with tumour progression such as pancreatic cancer [46], melanoma [47], and non-small lung cancer [1]. The leading SNP of this peak is rs7605851 (allele G; OR: 0·35 (95% CI: 0·22, 0·54); p=3·07 x10^{-6}). The *THSD7B* receptor is a target in the diagnosis and treatment of idiopathic membranous nephropathy(2). Translating this knowledge into the patho-mechanistic understanding of COVID-19 severity could be in association to immune-complex formation and subsequent molecular injury signalling. When assessing the association of this SNP between ancestry groups, this SNP was found to be monomorphic in the South Asian ancestry. While the magnitude and direction of odds ratio was consistent between the other ethnic groups, the association signals are apparently specific to the populations of East Asian (p=6·76 x10^{-3}) and European ancestry (p=6·42 x10^{-4}).

3.3.5. STK39 *gene in locus 2q24·3*

Serine/Threonine Kinase 39 (*STK39*) at the 2q24·3 locus (Fig. 2e) encodes a serine/threonine kinase that is thought to function in the cellular stress response pathway and acts as MAPK kinase and involved in stress response via activating p38 MAPK. The leading SNP of this peak is rs7595310 (allele A; OR: 0·41 (95% CI: 0·27, 0·59); p=4·55 x10^{-6}), predominantly present in the East Asian (p=7·06 x10^{-3}) and European ancestry (p=2·55 x10^{-3}). This gene has been strongly associated with cancer-related process and pathway, specifically non-small cell type lung cancer [48]. In addition, SNPs in linkage disequilibrium to the lead SNP in gene *STK39* has been associated to hypertension, mainly due to the involvement of *STK39* in sodium reabsorption and salt homeostasis [48–53].

3.3.6. FBXO34 *gene in locus 14q22·3*

The F-Box Protein 34 (*FBXO34*) gene in locus 14q22.3 (Fig. 2f) regulate proteins involved in differentiation and cell cycle progression including the Wnt signalling pathway [54]. Previous GWAS have identified the association of *FBXO34* gene to blood protein levels in cardiovascular risk, as well as leucocyte, reticulocyte and platelet count [55–58]. The leading SNP of this peak is rs10140801 (allele T; OR: 2.10 (95% CI: 1.51, 2.91); p=8.26 x10^{-6}). While the magnitude and direction of odds ratio was consistent between the other ethnic groups, the association signals are apparently specific to the population of European ancestry (p=1.78 x10^{-4}). SNPs in proxy (rs11547116, r^2=0.95; rs943590, r^2=0.96; rs7146752, r^2=0.97; rs13379169, r^2=0.96; rs72715769, r^2=0.91; rs58260073, r^2=0.90; rs10140869, r^2=0.89; rs11625423, r^2=0.86) to the lead SNP has been associated to lung fibroblasts and B-lymphocyte expression.

3.3.7. RPL6P27 *gene in locus 18p11·31*

Ribosomal Protein L6 Pseudogene 27 (*RPL6P27)* is in an intergenic region in locus 18p11·31 (Fig. 2g), and have been associated to ventricular arrhythmias, and circulating cytokine and growth levels. The leading SNP of this peak is rs11659676 (allele C; OR: 0·48 (95% CI: 0·34, 0·66); p=8·88 x10^{-6}), with a similar magnitude and direction of odds ratio present across the East Asian (p=2·99 x10^{-2}), South Asian (p=4·80 x10^{-2}) and European ancestry (p=4·29 x10^{-4}). The lead SNP has been expressed in lung fibroblast cells, T cells and B cells, which may lead to inflammation, alleviate tissue damage, mediate the cytokine storm and lead to a higher severity of symptoms associated to COVID-19.

3.3.8. METTL21C *gene in locus 13q33·1*

Methyltransferase 21C, AARS1 Lysine (*METTL21C*) in locus 13q33·1 (Fig. 2h) is a suggestive pleiotropic gene for bone and muscle, and the METTL21 family of proteins methylates chaperons that are involved in the aetiology of myopathy and regulation of the NF-kB signalling pathway, which is critical for bone and muscle homeostasis [59]. While the *METTL21C* gene may not have a direct effect on COVID-19 symptom severity, its association to the NF-kB signalling pathway may play a role in the expression of diverse cellular processes, including proliferation, survival, and immunity. In addition, the NF-kB family may also play a role in the production of inflammatory cytokines, enhancement of airway epithelium, increased airway neutrophilic and lymphocytic inflammation and goblet cell hyperplasia [60,61]. The leading SNP of this peak is rs599976 (allele G; OR: 0·37 (95% CI: 0·24, 1·22); p=8·95 x10^{-6}), with a predominant significant association signal in the South Asian (p=1·22 x10^{-2}) and European ancestry (p=5·33 x10^{-4}).

3.4. *eQTL and functional annotation in the lung domain*

A MAGMA-gene property analysis was conducted to investigate the association of the gene-property analysis with a full distribution of SNP p-values to differentially expressed genes (Supplementary Table 7). This demonstrated that a total of 56 genes are involved in B-cell receptor signalling, angiogenesis, and lymphomagenesis, leading to the association to COVID-19 hospitalization (p=2·87x10^{-5}). These genes play a key role in modifying the response of cells of lymphoid origin (such as B, T and NK cells) to self and tumour antigens, leading to B cell chronic lymphocytic leukaemia, as well as to viral infections like COVID-19 [62,63]. Given the potential importance of immunopathological complications in severe COVID-19 disease, regulation of these genes are critical to B cell receptor signalling, the chemokine signalling pathway, and decreases the influence of activated $CD4^+$ T cells to infiltrate and damage lung organs in the future [64]. This plays a significant role in the immune response, inflammation and cytokine storm through gene expression that leads to critical cases of COVID-19 infection.

After identifying the GWAS variants that have demonstrated a suggestive genetic association, SNP and gene enrichment analysis was conducted on cell types, specifically related to enrichment in gene expression across lung tissue. Multiple comprehensive eQTL mapping resources for profiling traits based on gene expression was conducted, including GTEx v7, Human Protein Atlas, and MAGMA analysis in FUMA. A significant (p=9·66x10^{-6}) overlap of the following immunologic gene sets (*FBXO34, STK39, PDE8B)* was present, demonstrating an association to COVID-19 hospitalization (Fig. 3a). Specifically, the immunologic gene set is involved in the expression of IFN-alpha and CD8 T-cells, and their response to antigen and costimulatory signals all play a significant role in the cytokine storm of the progression of symptoms associated to COVID-19. The identified COVID-19 gene associations were also enriched in the lung tissue, demonstrating an important role in lung function and pulmonary immune response (Fig. 3b). To investigate the overlap of individual gene expression to transcriptomics data integrated with lung expression and the immune system, a heatmap demonstrates that the following five genes were highly conserved across tissues and donors located in the lung (Fig. 2c-g): *VWA8, PDE8B, CTSC, STK39,* and *FBXO34,* respectively. Expression of these genes are associated to alveolar cells type 1 and 2, ciliated cells, club cells, B-cells, dendritic cells, granulocytes, macrophages, monocytes, NK-cells, T-cells, intestinal endocrine cells, fibroblasts, smooth muscle cells, and endothelial cells.

4. Discussion

We have discovered eight genes with a suggestive association to hospitalized cases of COVID-19 by studying the genome, functional

annotation, eQTL enrichment analysis and clinical outcome of 646 patients in multiple designated hospitals and quarantine camps in the UAE. We evaluated the association of genetic variants with COVID-19 severity using GWAS to improve the understanding of potentially causal molecular targets for SARS-CoV-2. We designed our current approach of trans-ancestry meta-analysis to uncover genetic variants shared across ancestry groups. Although the discovery of the eight loci were largely driven by effects in the European ancestry populations, the magnitude and direction of allelic effect were similar in multiple ancestral populations, demonstrating the chances of those variants modulating the risk of infection and severe COVID-19 symptoms in different populations.

When investigating non-genetic effects, hospitalized patients for COVID-19 tend to be older ($p<0{\cdot}001$), higher BMI ($p<0{\cdot}001$), and with a presence of one or more comorbidities ($p<0{\cdot}001$), similar to published literature [9,10,14,65–67]. Mortality is predominantly driven by these subgroups of patients with comorbid conditions, which includes cardiac conditions, chronic lung disease, diabetes mellitus, immunosuppressive condition, neurological disorder, metabolic disease, and renal disease. Contrasting to previous reports, male gender did not have a higher likelihood of developing severe/critical disease, in comparison to female gender (p=0·382). After multivariate regression analysis, smoking (p=0·091) and alcohol (p=0·837) had no impact on COVID-19 disease severity.

Eight genes demonstrated a strong association signal: *VWA8* gene in locus 13p14·11 (SNP rs10507497; p=9·54 x10^{-7}), *PDE8B* gene in locus 5q13·3 (SNP rs7715119; p=2·19 x10^{-6}), *CTSC* gene in locus 11q14·2 (rs72953026; p=2·38 x10^{-6}), *THSD7B* gene in locus 2q22·1 (rs7605851; p=3·07x10^{-6}), *STK39* gene in locus 2q24·3 (rs7595310; p=4·55 x10^{-6}), *FBXO34* gene in locus 14q22·3 (rs10140801; p=8·26 x10^{-6}), *RPL6P27* gene in locus 18p11·31 (rs11659676; p=8·88 x10^{-6}), and *METTL21C* gene in locus 13q33·1 (rs599976; p=8·95 x10^{-6}). The genes are expressed in the lung, associated to tumour progression, emphysema, airway obstruction, and surface tension within the lung, as well as an association to T-cell-mediated inflammation and the production of inflammatory cytokines(1–3). The signals observed of these genes are associated with respiratory failure that requires invasive mechanical ventilation.

Functional annotation can help prioritise the associations of the genes of interest. When investigating the association of the GWAS summary statistics to the MAGMA-gene property analysis, a total of 56 genes are involved in B-cell receptor signalling, angiogenesis, and lymphomagenesis, leading to the association of COVID-19 hospitalization (p=2·87x10^{-5}). These genes play a key role in modifying the response of cells of lymphoid origin (such as B, T and NK cells) to self and tumour antigens, leading to B cell chronic lymphocytic leukaemia, as well as to viral infections like COVID-19 [62,63]. Given the potential importance of immunopathological complications in severe COVID-19 disease, regulation of these genes are critical to B cell receptor signalling, the chemokine signalling pathway, and decreases the influence of activated $CD4^+$ T cells to infiltrate and damage lung organs in the future [64]. This plays a significant role in the immune response, inflammation and cytokine storm through gene expression that leads to critical cases of COVID-19 infection. Consistent with the large degree of pleiotropy between disorders, cell-type enrichment analysis demonstrated that there was an enrichment for lung tissue expression and donors, specifically with genes *VWA8, PDE8B, CTSC, STK39,* and *FBXO34.*

There are several potential limitations in our study. There are methodological drawbacks with the small sample size and study design, as the power analysis indicated that sample size is barely sufficient to identify genome-wide significant variants. However, because of the urgency of completing and reporting the findings of this study, we have published the findings, as opposed to waiting to meet sample size requirement. It should be noted that given the small sample size, this study is a purely descriptive, hypothesis-generating phase of research that is required to be replicated in larger studies of Middle Eastern and South Asian ancestry. Therefore, a multi-ethnic cohort with a larger sample size would be more suited to investigate these genetic associations that have been identified in this manuscript. There is a clear predominance of males (78.6%) of Asian ancestry (57.0%) in this cohort, a limitation that the results are skewed for a specific demographic group. However, the demographic factors in this cohort is fairly distributed and representative of the UAE population, where 72% of the population are male of Asian descent (59%) [34]. While most of the population was of self-reported Middle Eastern and South Asian descent, the population composition is a mixed genetic pool, as demonstrated by the PCA plot (Supplementary Fig. 2). However, to reduce technical confounders and population stratification, the participants were analysed separately by ethnic group depending on the genetic threshold. GWAS analysis was carried out separately for East Asian, South Asian, European, and American samples, followed by a trans-ancestry meta-analysis across the four population groups that demonstrates the associations in the tail distribution with minimal overall genomic inflation of the statistical results (λ_{GC} of 0·97 for meta- analysis). To improve statistical power for genomic discovery, covariates, such as age, gender, and population stratification, are included to account for stratification and avoid confounding effects from demographic factors. Due to the inclusion of multiple data collection sites from across the country, it is possible that the included cases, although all were hospitalized COVID-19 patients, are not entirely homogeneous. It is possible that the criteria for hospitalization of COVID-19 patients are different across the collection sites, thus measurement errors may exist in this study. Given that most of the participants were recruited from the hospital, we had limited number of participants who had asymptomatic or mild symptoms as they were more likely to quarantine at home. This may raise the concern that the pathological differences between the cases and the controls may be relatively subtle. The strength of this study was limiting misclassification and ascertainment biases in the control group by only selecting COVID positive patients with asymptomatic or mild symptoms, whereas other GWAS studies have utilized population controls. The use of population controls might mean that the SNPs identified are associated with susceptibility of COVID-19, as opposed to the association with severe COVID-19, a limitation that has been identified by other authors using population controls. The use of Euro-centric GWAS arrays limits the possibility of including targeted SNPs in the genome. To limit this error, imputation of genotypes for genetic variants that are untyped in the arrays increases the information provided by each microarray by accurately evaluating the evidence for association at genetic markers that are not directly genotyped [68].

We have discovered eight highly plausible genetic association with hospitalized cases in COVID-19. Our findings implicate that the different lung response to the virus determined by host genetic diversity may be an important factor in determining the severity of COVID-19. Further functional studies are warranted to establish the roles of these eight loci in the pathogenesis of COVID-19. Some of these associations may lead to therapeutic approaches due to the expression in the lungs. Improving fundamental knowledge and underlying biological pathway of COVID-19 heterogenous phenotypes is critical to mitigating this disease. Performing GWAS studies across different racial and ethnic populations to identify genes and haplotypes associated with differential factors of infection and clinical outcome, as well as vaccine responses, is crucial. Allelic effect size differences could reflect interaction of the causal variant with environmental risk factors that differ in exposure between ethnicities. Further studies must be conducted on a large-scale multi-ethnic cohort to expand and substantiate our current knowledge, as well as facilitate the development of population specific therapeutics to mitigate this worldwide challenge.

5. Contributors

HSA and GKT conceived the project to study the role of the virus and host in COVID-19 in the United Arab Emirates. HSA and MM conceived the central research questions for the GWAS data. MM initiated the first draft of the manuscript. MM, GT, and HSA analysed, and constructed the Tables and Figures. BM, MU, and NA were responsible for the recruitment of the patients and collecting data for the study. HV and HK carried out the laboratory assays used in the study. MM, GT and HAS provided critical review during manuscript preparation. All authors on the primary list contributed to the data interpretation and critically reviewed the manuscript and approved the final manuscript for submission.

6. The UAE COVID-19 Collaborative Partnership

Juan Acuna, Eman Alefishat, Ernesto Damiani, Samuel F. Feng, Andreas Henschel, Abdulrahim Sajini, Ahmed Yousef (Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates); Bassam Ali (United Arab Emirates University, Al Ain, United Arab Emirates); Hiba Alhumaidan, Hala Imambabaccus, Amirtharaj Francis, Stefan Weber (Sheikh Khalifa Medical City and SEHA, Abu Dhabi, United Arab Emirates); Mohammad Tahseen Al Bataineh, Rabih Halwani, Rifat Akram Hamoudi (University of Sharjah, Sharjah, United Arab Emirates); Abdulmajeed Al Khajeh, Laila Salameh (Dubai Health Authority, Dubai, United Arab Emirates) for the COVID-19 Collaborative Partnership.

Declaration of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data sharing

The data that support the findings of this study are available on NCBI Gene expression omnibus database (GSE184150), as well as request from the corresponding author.

Acknowledgments

We thank the participants of the study for their generosity in providing samples to advance our understanding of Covid-19 infection. We acknowledge the assistance of the healthcare workers at the frontline of the COVID-19 pandemic, especially Ms Laila Salameh, Dr Amna Saeed, Dr Mawada Hussein, Dr Nimmi Kurien, Dr Mona Elhassan, Dr Eman Elzain, dr Salma Rahma, Dr Sarrosh Khan, Dr Omar Alassaf, Dr Muneeba Moin, Dr Hend Alsaidi, Dr Rufaida Osman, Dr Khalid Hassan, and Dr Afrah Joher without their assistance, this study would not have been possible. We are also grateful to Hema Vurivi, who assisted in processing the samples in the laboratory. This review was commissioned as part of a project to study the host cell receptors of coronaviruses funded by Khalifa University's CPRA grant (Reference number 2020-004).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2021.103695.

References

[1] Lee Y, Yoon KA, Joo J, Lee D, Bae K, Han JY, et al. Prognostic implications of genetic variants in advanced non-small cell lung cancer: A genome-wide association study. Carcinogenesis 2013.
[2] Thrombospondin type-1 domain-containing 7a in idiopathic membranous nephropathy. N Engl J Med 2015.
[3] Luo M, Mengos AE, Ma W, Finlayson J, Bustos RZ, Xiao Zhu Y, et al. Characterization of the novel protein KIAA0564 (Von Willebrand Domain-containing Protein 8). Biochem Biophys Res Commun 2017.
[4] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020.
[5] Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China. JAMA 2020.
[6] Stokes EK, Zambrano LD, Anderson KN, Marder EP, Raz KM, El Burai Felix S, et al. Coronavirus disease 2019 case surveillance — United States, January 22–May 30, 2020. MMWR Morb Mortal Wkly Rep 2020.
[7] Marini JJ, Gattinoni L. Management of COVID-19 Respiratory Distress. JAMA 2020.
[8] Berlin DA, Gulick RM, Martinez FJ. Severe Covid-19. N Engl J Med 2020.
[9] Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020.
[10] Li X, Xu S, Yu M, Wang K, Tao Y, Zhou Y, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol 2020.
[11] Gralinski LE, Ferris MT, Aylor DL, Whitmore AC, Green R, Frieman MB, et al. Genome wide identification of SARS-CoV susceptibility loci using the collaborative cross. PLoS Genet 2015.
[12] Severe Covid-19 GWAS Group. Genomewide association study of severe Covid-19 with respiratory failure. N Engl J Med 2020.
[13] Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. Nature 2021.
[14] Shelton JF, Shastri AJ, Ye C, Weldon CH, Filshtein-Somnez T, Coker D, et al. Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity. medRxiv 2020.
[15] Roberts GHL, Park DS, Coignet MV, McCurdy SR, Knight SC, Partha R, et al. AncestryDNA COVID-19 host genetic study identifies three novel loci. medRxiv 2020.
[16] Rouf Banday A, Stanifer Megan L, Florez-Vargas Oscar, Onabajo Olusegun O, Zahoor Muhammad A, Papenberg Brenen W, Ring Timothy J, Lee Chia-Han, Andreakos Evangelos, Arons Evgeny, Barsh Greg, Biesecker Leslie G, Boyle David L, Burnett-Hartman Andrea, Mary C LP-O. Genetic regulation of OAS1 nonsense-mediated decay underlies association with risk of severe COVID-19. medRxiv 2021.
[17] Daw Elbait G, Henschel A, Tay GK, Al Safar HS. Whole genome sequencing of four representatives from the admixed population of the united arab emirates. Front Genet 2020;11:681.
[18] Tay GK, Henschel A, Daw Elbait G, Al Safar HS. Genetic diversity and low stratification of the population of the United Arab Emirates. Front Genet 2020;11:608.
[19] Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbrouckef JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Bull World Health Organ 2007;85(11):867–72.
[20] National Institutes of Health. Clinical Spectrum of SARS-CoV-2 Infection. Radiol Cardiothorac Imaging 2020.
[21] John SE, Antony D, Eaaswarkhanth M, Hebbar P, Channanath AM, Thomas D, et al. Assessment of coding region variants in Kuwaiti population: implications for medical genetics and population genomics. Sci Rep 2018.
[22] Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of greater middle eastern genetic variation for enhanced disease gene discovery. Nature Genetics 2016.
[23] Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005.
[24] Machiela MJ, Chanock SJ. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015.
[25] Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science 2015(80-).
[26] Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science 2017(80-).
[27] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science 2017(80-).
[28] Digre A, Lindskog C. The Human Protein Atlas—Spatial localization of the human proteome in health and disease. Protein Sci 2021.
[29] Keen JC, Moore HM. The genotype-tissue expression (GTEx) project: linking clinical data with molecular analysis to advance personalized medicine. J Pers Med 2015.
[30] Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreservation and Biobanking. 2015.
[31] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nature Genetics 2013.
[32] Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017.
[33] de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol 2015.
[34] CORPORATE E. United Arab Emirates Demography and Statistics Through a Magnifying Glass. TGM Research FZE 2021.
[35] Castaldi PJ, Cho MH, Estépar RSJ, McDonald MLN, Laird N, Beaty TH, et al. Genome-wide association identifies regulatory loci associated with distinct local histogram emphysema patterns. Am J Respir Crit Care Med 2014.
[36] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: Integrating information about genes, proteins and diseases. Trends Genet 1997.

[37] WYATT JP, FISCHER VW, SWEET HC. Panlobular emphysema: anatomy and pathodynamics. Dis Chest 1962.
[38] Horvath A, Giatzakis C, Tsang K, Greene E, Osorio P, Boikos S, et al. A cAMP-specific phosphodiesterase (PDE8B) that is mutated in adrenal hyperplasia is expressed widely in human and mouse tissues: A novel PDE8B isoform in human adrenal cortex. Eur J Hum Genet 2008.
[39] Vang AG, Basole C, Dong H, Nguyen RK, Housley W, Guernsey L, et al. Differential expression and function of PDE8 and PDE4 in effector T cells: Implications for PDE8 as a drug target in inflammation. Front Pharmacol 2016.
[40] Johnstone TB, Smith KH, Koziol-White CJ, Li F, Kazarian AG, Corpuz ML, et al. PDE8 is expressed in human airway smooth muscle and selectively regulates cAMP signaling by b2-adrenergic receptors and adenylyl cyclase 6. Am J Respiratory Cell Molecular Biol 2018.
[41] Mokra D, Mokry J. Phosphodiesterase inhibitors in acute lung injury: What are the perspectives? Int J Mol Sci 2021.
[42] Andoniou CE, Fleming P, Sutton VR, Trapani JA, Degli-Esposti MA. Cathepsin C limits acute viral infection independently of NK cell and CD8+ T-cell cytolytic function. Immunol Cell Biol 2011.
[43] Korkmaz B, Lesner A, Letast S, Mahdi YK, Jourdan ML, Dallet-Choisy S, et al. Neutrophil proteinase 3 and dipeptidyl peptidase i (cathepsin C) as pharmacological targets in granulomatosis with polyangiitis (Wegener granulomatosis). Sem Immunopathol 2013.
[44] Sutherland RE, Barry SS, Olsen JS, Salantes DB, Caughey GH, Wolters PJ. Dipeptidyl peptidase i controls survival from Klebsiella pneumoniae lung infection by processing surfactant protein D. Biochem Biophys Res Commun 2014.
[45] Neale B, Churchhouse C, Abbott L, Anttila V, Aragam K, Baumann A, et al. Rapid GWAS of thousands of phenotypes for 337,000 samples in the uk biobank [Internet]. 2017. Available from: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank
[46] Low SK, Kuchiba A, Zembutsu H, Saito A, Takahashi A, Kubo M, et al. Genome-wide association study of pancreatic cancer in Japanese population. PLoS One 2010.
[47] Jiang J, Ding Y, Wu M, Chen Y, Lyu X, Lu J, et al. Integrated genomic analysis identifies a genetic mutation model predicting response to immune checkpoint inhibitors in melanoma. Cancer Med 2020.
[48] Li Z, Zhu W, Xiong L, Yu X, Chen X, Lin Q. Role of high expression levels of STK39 in the growth, migration and invasion of non-small cell type lung cancer cells. Oncotarget 2016.
[49] Cunnington MS, Kay C, Avery PJ, Mayosi BM, Koref MS, Keavney B. STK39 polymorphisms and blood pressure: an association study in British Caucasians and assessment of cis-acting influences on gene expression. BMC Med Genet 2009.
[50] Xi B, Shen Y, Zhao X, Chandak GR, Cheng H, Hou D, et al. Association of common variants in/near six genes (ATP2B1, CSK, MTHFR, CYP17A1, STK39 and FGF5) with blood pressure/hypertension risk in Chinese children. J Hum Hypertens 2014.
[51] Yang H, Ye L, Wang Q, Cai D, Chen Q, Pan H, et al. A meta-analytical assessment of STK39 three well-defined polymorphisms in susceptibility to hypertension. Sci Rep 2016.
[52] Persu A, Vikkula M. A genome-wide association study-derived candidate gene seeks replication: STK39. J Hypertension 2011.
[53] Xi B, Chen M, Chandak GR, Shen Y, Yan L, He J, et al. STK39 Polymorphism Is Associated with Essential Hypertension: A Systematic Review and Meta-Analysis. PLoS One. 2013.
[54] Randle SJ, Laman H. F-box protein interactions with the hallmark pathways in cancer. Seminars in Cancer Biology. 2016.
[55] Folkersen L, Fauman E, Sabater-Lleal M, Strawbridge RJ, Frånberg M, Sennblad B, et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genet 2017.
[56] Chen MH, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell 2020.
[57] Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet 2019.
[58] Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell 2020.
[59] Huang J, Hsu YH, Mo C, Abreu E, Kiel DP, Bonewald LF, et al. METTL21C is a potential pleiotropic gene for osteoporosis and sarcopenia acting through the modulation of the NF-κB signaling pathway. J Bone Miner Res 2014.
[60] Sheller JR, Polosukhin VV, Mitchell D, Cheng DS, Peebles RS, Blackwell TS. Nuclear factor kappa B induction in airway epithelium increases lung inflammation in allergen-challenged mice. Exp Lung Res 2009.
[61] Alvira CM. Nuclear factor-kappa-B signaling in lung development and disease: One pathway, numerous functions. Birth Defects Res Part A - Clin Mol Teratol 2014.
[62] McGettigan SE, Debes GF. Immunoregulation by antibody secreting cells in inflammation, infection, and cancer. Immunological Reviews 2021.
[63] Fu H, Kishore M, Gittens B, Wang G, Coe D, Komarowska I, et al. Self-recognition of the endothelium enables regulatory T-cell trafficking and defines the kinetics of immune regulation. Nat Commun 2014.
[64] Yu Z, Qin C, Cao M, He X, Ren H, Liu H. The Potential Genes Mediate the Pathogenicity of Allogeneic CD4+T Cell in aGVHD Mouse Model. Biomed Res Int 2021.
[65] The COVID-19 Host Genetics Initiative . Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. medRxiv 2021.
[66] Chen R, Liang W, Jiang M, Guan W, Zhan C, Wang T, et al. Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China. Chest 2020.
[67] Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Prospective observational cohort study. BMJ 2020.
[68] Yun L, Willer C, Sanna S, Abecasis G. Genotype imputation. Annual review of genomics and human genetics; 2009.